Teachers Institute of Philadelphia

Analysis and Report

Overview

This report summarizes the findings from a study that assesses whether a teacher's participation in the Teachers Institute of Philadelphia affected his/her probability of retention within the School District of Philadelphia (SDP). The study uses data on teachers in SDP over an eight-year period from 2010 through 2018. The analytic file includes demographic variables, as well as title, years spent in the district, and age. The original data file is supplemented with school-level information obtained from open data made available by the District Performance Office at the School District of Philadelphia (www.philasd.org/performance).

University of Pennsylvania

Katherine Wilson and Wendy Chan

August 10, 2020

Data Cleaning

The initial data set consisted of 16,013 unique, 7,014 of whom we have information on exiting the school system. A total of 356 unique teachers who participated in TIP. After checking for missing data, 4 observations in teacher exits were coded as duplicates, and 2 observations in the TIP participation variables set were duplicates. Of the demographic variables, 43 observations were NA (missing). Of the total teachers set, 32 observations were NA. After removing the NA values, the analytic file included 16,005 unique teachers.

Upon merging the datasets, it became apparent that some unique teacher IDs were present in certain datasets but missing from others. Specifically, using the anti-merge function in our code, we can see the observations that are included in the information on exits but not in the information on total teachers. There are 1,067 teachers who exited, and who were also missing from the total teachers data. We removed these missing values from the merge of teacher exits with the total teachers.

We proceeded to merge the deidentified TIP IDs with the complete teachers file. Of the TIP participants, 44 unique teachers were missing from the total teachers set. This brings our final number of unique teachers who participated in TIP (with complete covariate information) down from 358 to 314. After removing two duplicates, the final number of unique TIP participants in the dataset is 312.

In the descriptive statistics and statistical analyses below, we use a final analytic file that consists of 16,005 teachers, of which 312 participated in TIP. This information is based on the 8-year period from 2010 – 2018.

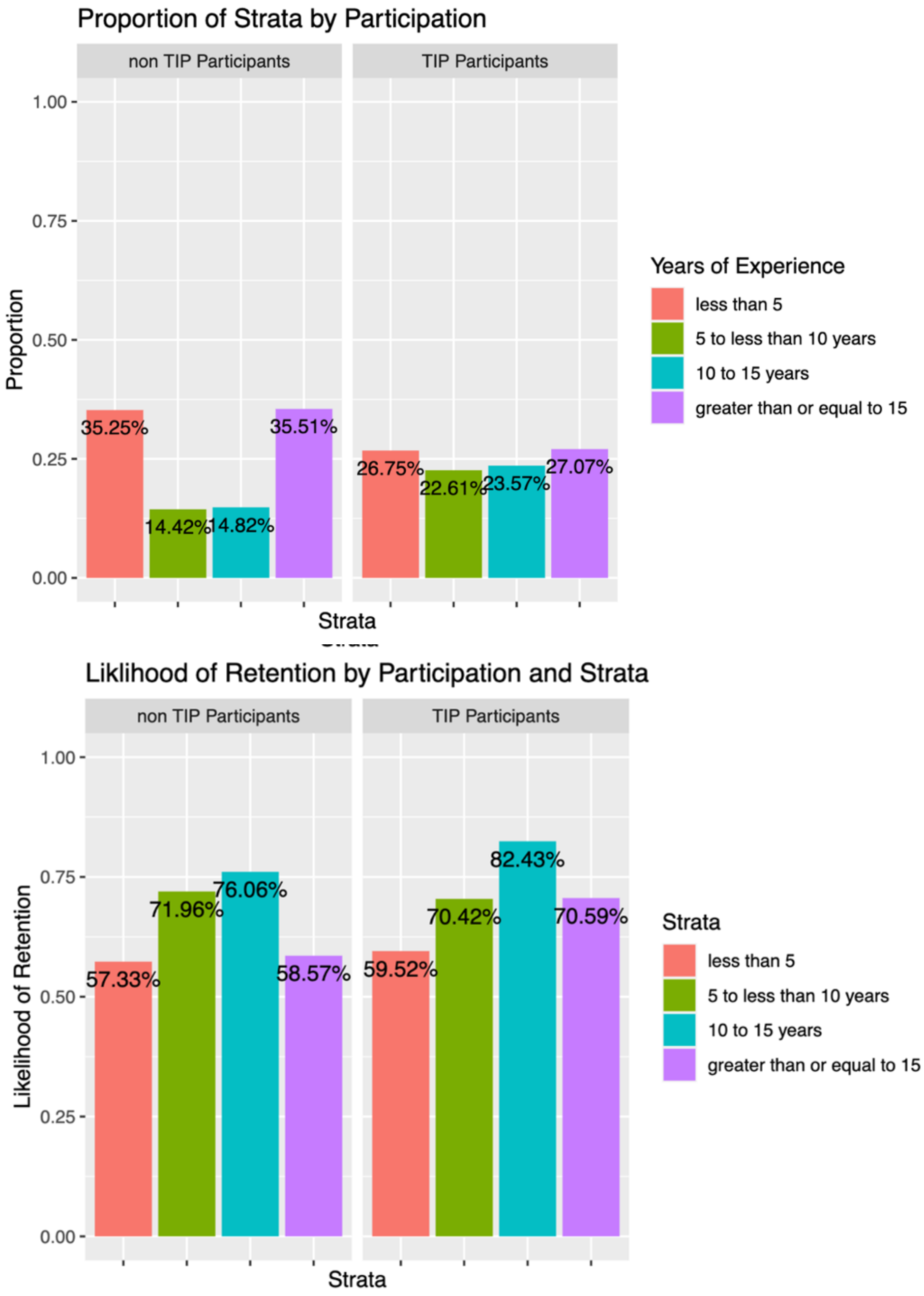
Variable Definitions

For each teacher in the dataset, we created the indicator variable, "Participated," that equals 1 for a TIP participant and 0 otherwise. The outcome variable of interest is the binary indicator, "Retained," which is equal to 1 if the teacher remained in SDP for the duration of the study period and 0 otherwise. Additionally, we created variables that tracked the number of times a teacher changed schools, changed titles, etc., but remained within SDP. Because there were 9 observations with missing gender information, the following statistical models were run on a sample of 16,003 teachers, of whom 312 participated in TIP.

Descriptive Statistics

In the first report (presented in January), we sorted teachers alphabetically, and thus their years of service was the first year in the dataset. In this report, we use the maximum years in the dataset as an indicator of the years spent and the maximum age of the teacher in the dataset. To keep in line with the analyses results, the descriptive statistics for these maximum ages are presented here. The proportions are relatively similar to the proportions in the original analysis, with a relatively even split of strata among tip and non-TIP participants. Likewise, teachers with less than 5 years of experience are less likely to be retained, among both TIP and non-TIP groups.

Figure 1.



Phase 1- Exploratory Analysis and Logistic Regression Models

Phase 1 began with a visualization and descriptive analysis of the breakdown of teacher strata and retention. After that, we analyzed the significance of participation in TIP on retention through a logistic regression. Participating in TIP is associated with an increase in the log odd likelihood of retention. Specifically, the coefficient was .35, meaning that participating in TIP was associated with a .35 increase in retention.

Certain ethnicities, when ethnicity is added as a second predictor, are also associated with increases in likelihood of retention.

Statistical Analyses

Because the outcome variable, retention, is binary, we use the following logistic regression model for our inferences.

$$log\left(\frac{r}{1-r}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_p x_p$$

The variable r is the retention indicator and each x represents a predictor variable that explains variation in the log odds of retention, such as gender or years of service in the district.

Model 1. Does participation in TIP affect the log odds of retention?

First, we test the model using participation in TIP as the predictor and retention as the outcome variable. The model and the results are given below.

$$log\left(\frac{r}{1-r}\right) = \beta_0 + \beta_1 x_1$$
$$x_1 = Participation$$

	Estimate	**Standard Error**	**P value**
(Intercept)	0.518	0.017	<.05*
Participated in TIP	0.353	0.125	< .05*

The output of Model 1 shows that the log odds for those who participate in TIP are 0.353 greater than those who did not participate in TIP. This value is statistically significant α = 0.05.

Model 2. Does the effect of TIP on the log odds of retention differ when controlling for gender and ethnicity?

In this analysis, we extended Model 1 by controlling for the effects of Gender and Ethnicity. However, we found (results not shown) that Gender was not statistically significant in predicting the log odds of retention and so the results for Model 2 below are only given for Ethnicity.

$$log\left(\frac{r}{1-r}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$
$$x_1 = Participation$$
$$x_2 = Ethnicity$$

	Estimate	**Standard Error**	**P value**
(Intercept)	0.575	0.033	<.05*
Participated in TIP	0.349	0.125	< .05*
Ethnicity (Asian)	-0.006	0.114	0.960
Ethnicity (Caucasian)	-0.097	0.039	<.05*
Ethnicity (Latina)	0.200	0.102	<.05*
Ethnicity (Nat. Am)	-0.005	0.349	0.989
Ethnicity (Other)	0.023	0.164	0.887

In the table above, the reference ethnicity group (intercept term) refers to teachers who identified as African-American. As a result, all coefficients for the ethnicity variables are comparisons between the specified ethnicity group and African-American teachers. The table above shows that, among African-American teachers, participating in TIP improved their log odds of retention by a factor of 0.349 on the log odds scale.

Among the other ethnicity groups, the coefficients associated with Caucasian and Latina teachers are also statistically significant. Interestingly, for Caucasian teachers who participated in TIP, the log odds of retention are actually *lower* compared to non-TIP African-American teachers. This is seen by the -0.097 coefficient. On the other hand, the log odds is higher among Latina TIP participants compared to the reference group of non-TIP African-American teachers.

Model 3. Does the effect of TIP on the log odds of retention differ when controlling for years of service?

Model 3 includes the “Strata” variable that accounts for the years of service. Note that the reference group (given by the intercept) is associated with teachers who have worked in SDP for less than 5 years and who did not participate in TIP.

$$log\left(\frac{r}{1-r}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$
$$x_1 = Strata$$
$$x_2 = Participation$$

	Estimate	**Standard Error**	**P value**
(Intercept)	0.671	0.023	<.05*
Strata 2 (5-10 years)	0.360	0.057	< .05*
Strata 3 (10 to 15 years)	0.378	0.055	<.05*
Strata 3 (over 15 years)	-1.011	0.041	<.05*
Participated	0.256	0.120	<.05*

For teachers who have worked in SDP for less than 5 years, participation in TIP is still statistically significant in terms of improving the log odds of retention. Each of the coefficients for Strata is also statistically significant, which implies that years of experience, coupled with participation in TIP, also help predict the log odds of retention. Interestingly, the coefficient for Strata 3, which refers to teachers who have more than 15 years of experience, is negative at -1.011. This implies that TIP participants wither over 15 years of experience had a lower log odds of retention compared to non-TIP teachers with less than 5 years of experience.

Discussion/Limitations

The first question of interest is to determine whether participation in TIP is associated with the retention of teachers in SDP. The analyses show that participation in TIP is a significant predictor of retention among teachers and that the effect of participation may be larger for some groups. For instance, the effect of TIP on retention was significantly higher for teachers who identified as African American or Latina, and the same effect was significantly lower for teachers who identified as Caucasian.

We also considered interaction effects (analyses not shown) between the potential moderators of “years of experience” and “gender” and found no significant effects. Thus, our analyses focus on the potential confounding effects of these variables on the relationship between TIP participation and retention.

Next, we plan to look more closely at the differences in teacher retention rates between strata and number of years served for TIP teachers. As mentioned in the original proposal, “non-retention of high-performing teachers is a serious problem for urban school districts.” A better understanding of the patterns of retention based on years of experience is needed. This can potentially shed light on the specific groups in which TIP has the strongest impact.

There are several limitations to the current analyses. First, the sample size of TIP teachers is a fraction of the sample size of non-TIP teachers. Such sample size differences may result in complications for future interaction analyses or causal interpretations. Additionally, teacher retention is calculated as a teacher who stays in the school district dataset from year to year. This fails to capture teachers who change titles or who move to different schools within the district.

Lastly, an additional variable that would help to answer the original research questions posed in the proposal would be an indicator of the date that teachers began the TIP program. As it stands, the only information that we have from TIP is the Deidentified IDs of participants, which were matched with administrative data from the 8-year period. If given an indicator of the start date and/or end date that teachers participated in TIP, we could analyze trends like amount of time participated in TIP, and its effect on retention.

Phase 2- Causal Inference Analysis

In Phase 1 of the analysis, teachers who participated in TIP were more likely to be retained. Additionally, the effect of participation may have been larger for some groups. For instance, the effect of TIP on retention was significantly higher for teachers who identified as African American or Latinx, and lower for teachers who identified as Caucasian. On the other hand, prior years of experience and gender were not significant.

The logistic regression in Phase 1 performs well as a classification method in predicting which teachers will be retained based on participation in TIP. The logistic regression also allows us to explore associative relationships among variables related to retention. However, the logistic regression is limited in its ability to express a causal relationship between participating in TIP and retention.

In Phase 2, we add rigor to the logistic regression by asking if there is a causal relationship between participation in TIP and retention. We match teachers based on available and measured covariates, such as years in the classroom, gender, or title (classroom teacher or administration). By matching, we reduce the chance that these factors (whether or not the teacher is a male or female, for instance) explain retention. Instead, the variable of causal interest (whether or not the teacher participated in the intervention) is isolated.

We spent ample time exploring the differences in groups for each measured covariates, since the goal of balancing will depend on the raw differences in the sample size representation of each of these covariates. Appendix A includes differences in covariate groups among TIP and non-TIP teachers. The impacts of each covariate on the percent balance improvement among TIP and non-TIP teachers was considered, and in the end we chose to trim the sample to include only the ages of teachers who participated in TIP.

To be more specific, certain ages of teachers are represented in the non-TIP sample, yet not represented in the Treatment (TIP) sample. The ages that are not represented in the TIP samples are: 32, 38, and 40+. Especially in the context of our study, where the question of interest is teacher retention, then an important confounder of staying in the classroom is age of the teacher. Therefore, matching on age without taking into account these non-representative ages may affect the balance allowed on other variables (specifically gender and ethnicity). 300 non-participating teachers were not matched in the age of the TIP participating teachers.

Trimming the sample has the effect of improving percent balance improvement for Gender and the other category of Ethnicity. In fact, all of the variables now have sufficient percent balance improvement. Appendix A includes the original matching model with all covariates included, and describes how the percent balance improvement informed the selection of covariates in our final matching model, presented below. This final model includes the trimmed TIP sample so that we only include TIP teachers with Years of experience that are matched in the other sample of non-TIP teachers.

Matching Model: Participated ~ Gender + Ethnicity + Strata + Employee Age + Title Status
Outcome Model: Retained ~ Participated

	Estimate	Std. Error	p-value
Intercept	0.4162	0.1157	0.000***
Participated	0.4557	0.1697	0.0073**

Discussion/Limitations

In conjunction with Phase 1 of the study, Phase 2 implies that participation in TIP may have a causal effect on the retention of teachers in the Philadelphia school district, when teachers are matched on the measured covariates. All else things being equal that are associated with retention (gender, ethnicity, age, years in the classroom, and title), then those who participate in TIP are more likely to be retained, within the time period of our study. Limitations to this conclusion include the small sample size of TIP participants, and a myriad of other confounders that could have been measured.

In broader implications of the study, we first consider the associations in ethnicity found in Phase 1 that may be of interest to certain minority groups. From a psychological standpoint, students who have a teacher who “looks like them” (such as similar gender or Ethnicity) are predicted to build stronger relationships in the classroom, and relationship building is frequently tied to academic performance and interest. One study looked at this so called “Teacher Match” on students’ academic perceptions and attitudes. The study found that demographically similar teachers, “especially in gender matches” are significant in the quality of student-teacher communication and college aspiration.[1] If retaining those groups of teachers is a driver in reducing disparities in education among certain ethnic and gender groups, then better understanding how different programs lead to disparate outcomes of retention for different groups of teachers (gender/ethnicity) would be an important next step for TIP.

Further research would likewise do well to explore this interaction between ethnicity, age, and gender. White teachers are more represented in the 20s and 30s group of teachers than non-white teachers (about 23% of teachers in their 20s are nonwhite, and 26% of teachers in their 20s are

[1] Egalite, Anna J., and Brian Kisida. "The effects of teacher match on students’ academic perceptions and attitudes." *Educational Evaluation and Policy Analysis* 40.1 (2018): 59-81.

non-white, while 75% of teachers in their 50s are non-white), but the limitations of this study in sample size of ethnicity and gender prevent that question from being directly addressed.

A second broader implication is from the teacher, rather than the student, perspective. Teacher retention is a much studied topic[2][3], and programs like TIP, which enable teachers to pursue university-level study, thereby bringing new content to students and increasing teachers' morale, are rooted in the mission to keep teachers helping those students who need them most. Both authors of this study were public school teachers themselves, and know that "intellectual engagement" is certainly as important as "purpose" in a career. One possible explanation behind the increased retention of teachers who participate in TIP may by this intellectual engagement that the program offers.

There remains unanswered questions in this domain that the TIP organization can continue to pursue and work to address. Specifically, what are the outcomes of teachers who do leave the district, after participating in TIP? Our study was limited to a binary variable of retained or not retained. This lacks a very important other possibility, which is that teachers left their role in SDP to pursue other intellectually engaging and purpose driven opportunities, such as a doctorate in education or starting their own computer programming boot-camp for minority students in Philadelphia. It is hard to argue that these teachers are contributing to educational inequity by leaving their classroom positions. TIP can continue to reach out to teachers who leave to learn qualitatively about the specifics of roles that former teachers are pursuing outside of the classroom role, and whether or not participating in TIP actually encouraged them to seek these roles that are equally as influential as the work of a classroom teacher.

[2] Holmes, B., Parker, D., & Gibson, J. (2019). Rethinking teacher retention in hard-to-staff schools.
[3] Cochran-Smith, M. (2004). Stayers, leavers, lovers, and dreamers: Insights about teacher retention.

Appendix A

Descriptive Differences in Groups

The mean Years of Service for non-TIP teachers is 11.9, and the mean years of service for TIP teachers is 11.1. The mean age for non-TIP teachers 43.8, and the mean age for TIP teachers is 44.4. Strata is also relatively evenly split among participants and non-participants.

Participated	Count	Mean Years of Service	Mean Age
0	15695	11.9	43.8
1	312	11.1	44.4

Strata	0 (non participants)	1 (Participants
< 5	5530	83
5 to less than 10 years	2267	70
10-15 years	2324	74
Greater than 15	5574	85

Categorical variables are a little trickier, since there are many positions not represented in the TIP sample. For instance, 24 non-TIP participants are "Demonstration Special Ed" teachers but 0 TIP teachers identify as "Demonstration Special Ed" teachers. On the other hand, no TIP teachers had positions that were not present in the non-TIP sample. Because of this, I recode teachers into "classroom teachers" or "non-classroom teachers".

Title Status	0 (non Participants)	1 (Participants)
Non-classroom teacher	1062	7
Classroom teacher	14633	305

As for gender, 26.40% of non-TIP teachers were male, and 24.36% of TIP teachers were male.

Gender	0 (non Participants)	1 (Participants)
Female	11552	236

Male	4143	76

Ethnicity is also relatively equally distributed across tip and non-tip samples, with the majority Caucasian followed by African American. The ratio switched for Asian and Latina, and some groups were not represented. I created a catch all group “Other”, which is the aggregate of Prefer Not Disclose and Other.

Ethnicity	**0 (non Participants)**	**1 (Participants)**
African Am	3860	102
Asian	358	13
Caucasian	10729	188
Latinx	502	4
other	246	5

Causal Matching Models

We spend time here exploring the balance in the matched comparisons, to ensure that variables are representative in the unmatched and matched group, and that matching does not overrepresented a certain variable, for instance. Then, we build a logistic regression model using these propensities, or likelihoods, that units belong to a certain group, on the outcome of Retention.

We ran multiple different matching models, and one by one removed covariates from each model. The first Matching Model includes all of the five measured covariates in our study: Gender, Ethnicity, Strata (Years in the classroom), Employee Age, and Teacher Status.

Initial Matching Model: Participated ~ Gender + Ethnicity + Strata + Employee Age + Title Status

The Percent Balance Improvement for matched data is shown below:

Percent Balance Improvement

	Mean Diff.
distance	99.203
Gender(F)	-10.09
Gender(M)	-10.09
Ethnicity (As.)	66.01
	96.05

Ethnicity	66.551
(Cauca.)	-810.9112
Ethnicity (Lat.)	59.895
Ethnicity (other)	71.224
Strata (5-10)	96.125
Strata(10-15)	69.834
Strata(>15)	85.827
Employee age	
Teacher_Status	

Initial Outcome Model: Retained ~ Participated

	Estimate	Std. Error	p-value
Intercept	0.664	0.120	0.000***
Participated	0.207	0.172	0.229

Balance is a representation of similarity in the distribution of a covariate among the treatment and the control group. The standardized bias is a quantified measure of bias, represented by the difference in the means of the covariate between the treated group and the comparison divided by the standard deviation.

Before investigating the balance, note the raw differences in means for treated and control groups before and after matching (Consult Appendix B). Prior to matching, the TIP sample is 24.36% Male, and the non-TIP sample 26.40% Male. After matching we have 24.36% males in the TIP sample and 22.12% males in the non-TIP sample. The absolute difference between the percentage of males in the respective samples has increased from the unmatched to matched sample. The percent balance improvement (-10.09) reflects this undesired (negative) change in balance. The story is the same for females: note the negative percent balance improvement for females (-10.09).

A decrease in balance from unmatched to matched is also evident in the Ethnicity variable, where 1.6% of TIP teachers identify as Other, and 1.57 identify as Other in the non-TIP sample. The matching algorithm actually increases this difference, and over-represents the "Other" ethnicity in the matched data. Again, the large negative balance improvement reflects this undesirable matching.

We hypothesize that the negative percent balance improvement in these variables is a result of the skewed sample size. Males make 26.4% of the non-TIP sample. When matched, however, the percent of Male non-TIP teachers decreases to 22%, and Male teachers are no longer well represented in the Control sample. Females, on the other hand, represent a larger majority of both the TIP and nonstop samples. After matching, females are overrepresented in the Control sample.

Likewise, ethnicity is overrepresented in the control group after matching, widening the gap between these two representations. One approach is to coarsen the ethnicity variable into Caucasian or non-Caucasian. This resulted in a negative percent balance improvement for employee age, perhaps because ethnicity and age are related. Indeed, age and non-white status are associated: white teachers are more represented in the 20s and 30s group of teachers than non-white teachers (about 23% of teachers in their 20s are nonwhite, and 26% of teachers in their 20s are non-white, while 75% of teachers in their 50s are non-white).

Because the negative balance improvement may be due to the skewed/off kilter sample sizes in these respects, we keep these variable (Ethnicity and Gender) in the original matching model and investigate the pattern of missingness on other variables to understand how missingness in certain representations of variables affect balance.

Specifically, the Years of Experience variable is an indicator of the amount of years a teacher has spent in the classroom. Certain ages of teachers are represented in the non-TIP sample, yet not represented in the Treatment (TIP) sample. The ages that are not represented in the TIP samples are: 32, 38, and 40+. Especially in the context of our study, where the question of interest is teacher retention, then a possible confounder of staying in the classroom is age of teacher. Therefore, matching on age without taking into account these non-representative ages may be affected the balance allowed on other variables (specifically gender and ethnicity).

Various approaches are available to treat missing values on levels of measured covariates, such as imputing the missing values or investigating a pattern of missingness. However, in our case, the unrepresented ages only amount to 300 teachers, a fraction of the total sample size of 16,007. Therefore, we decide to exclude subjects with ages not represented in the TIP sample. In this case, missing data is not a broad problem (less than 2%), and our sample size is not reduced markedly.

In response, we trim the TIP sample so that we only have TIP in terms of Years of experience that are matched in the other. This reduces the observations down to 15,707 from the 16,007 in the prior data set. This means that 300 non-participating teachers were not matched in the age of the TIP participating teachers.

Matching Model: Participated ~ Gender + Ethnicity + Strata + Employee Age + Title Status

Percent Balance Improvement

	Mean Diff.

distance	96.49
Gender(F)	53.59
Gender(M)	53.59
Ethnicity (As.)	65.43
Ethnicity	68.91
(Cauca.)	83.75
Ethnicity (Lat.)	100.00
Ethnicity (other)	87.53
Strata (5-10)	88.85
Strata(10-15)	81.85
Strata(>15)	42.17
Employee age	100
Teacher Status	

Appendix A

Overview

Upon cleaning and merging the TIP data, the final merged TIP data file is used to find summary statistics and compare percentages of completion between TIP participants and non TIP participants. The merged file is brought in from the script, “TIP merge and data clean” which produces one final data set for analysis. For more information on the nature of missing data and the merge, consult the prior script.

The merged TIP data file we use for the below figures and analyses has complete variables for 16,012 teachers, 312 of which participate in TIP, and 15700 who did not participate in TIP. Information on these teachers is collected across an 8 year period, from 2010 through 2018.

Visualization

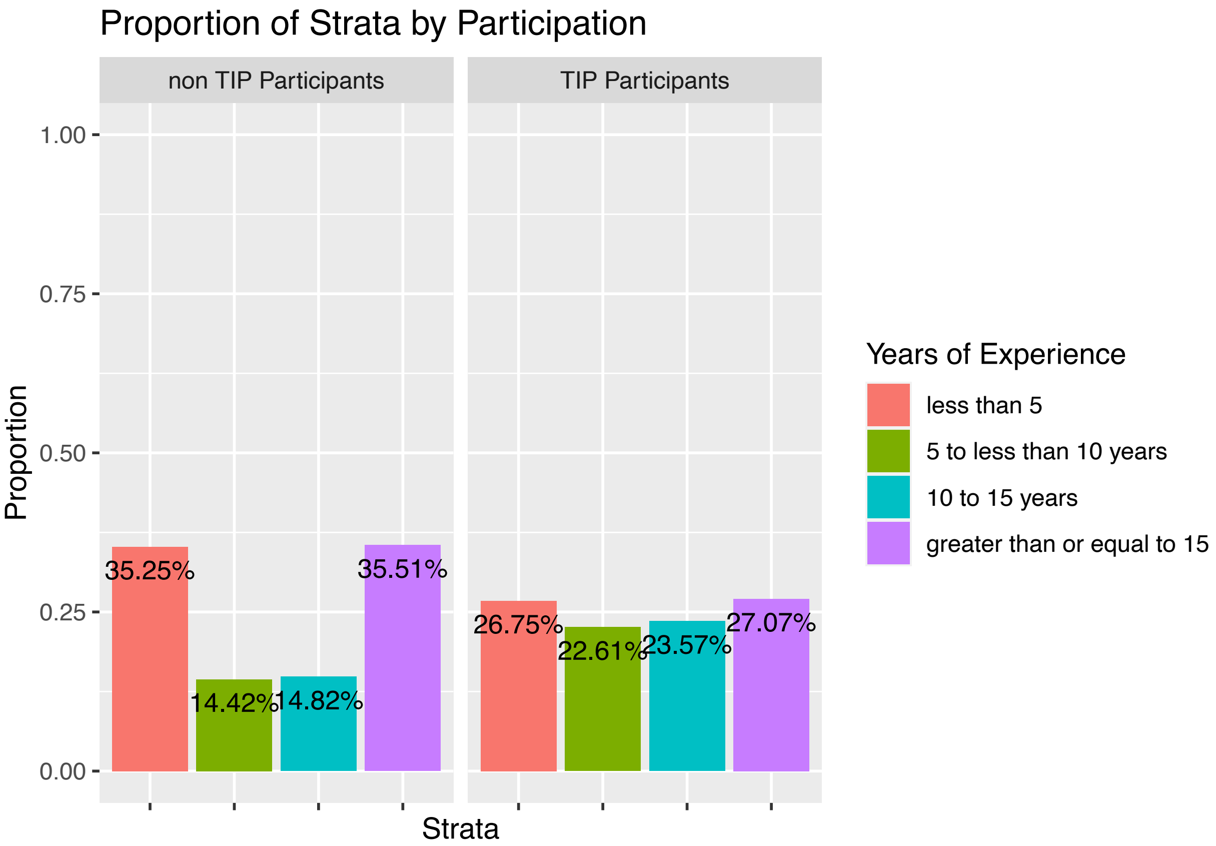
In order to evaluate the effectiveness of TIP participation on retention, we first look to get a picture of the average length of time spent in the district by teachers, as well as how retention patterns regularly play out in the Philadelphia school district.

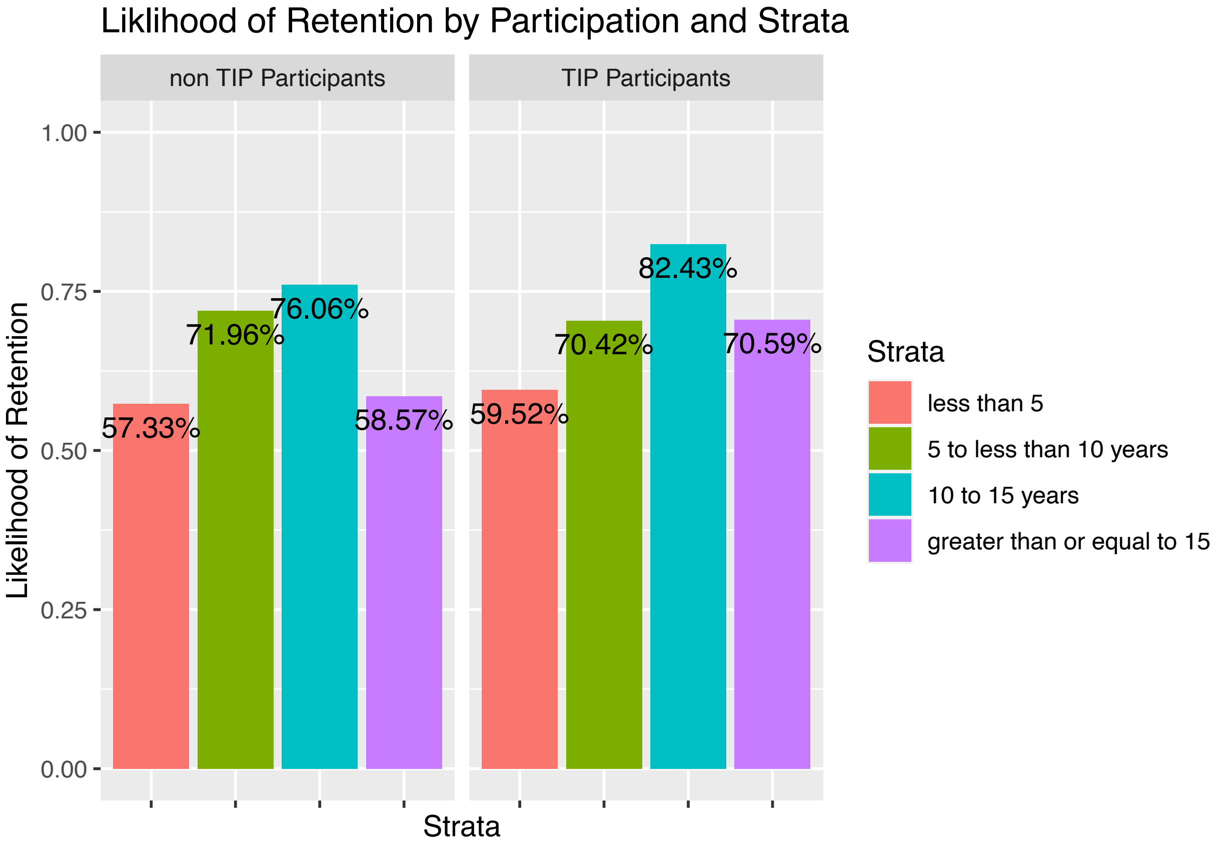
How long do TIP and non TIP participants stay in teaching profession, on average? Figure 1 breaks down these differences. TIP participation rates are relatively evenly split among four strata of teacher groups. These strata divide teachers into one of four groups, based on how long each teacher has served in the district.

A majority of teachers have less than 5 years of service in the district. Around 30% of both TIP and non-TIP teachers have been in the school district for 5-15 years.

Figure 2 breaks down the liklihood of retention for teachers based on their strata. The table sorts teachers by strata to compare how groups of teacher strata differ in retention rates.

For Strata 1 (less than 5 years), the likelihood of retention is very similar for TIP and non TIP teachers, at around 64%. However, as we move into stratas 2 and 3, then the liklihood of retention increases for TIP teachers. This trend also holds for the final strata, where TIP teachers are more likely to stay in the profession than non-TIP teachers.


Proportion of Strata by Participation


Liklihood of Retention by Participation and Strata

Constructing a dataset for analysis

In order to continue with a statistical analysis, we create an analysis dataset with a row for each Deidentified ID that has all of its unique variables. Recall that the merged_dataset has multiple rows for the same Deidentified ID. For instance, teacher ID #72 may have changed schools over the time we have the data. Or, teacher ID #72 may have changed their title (example: a Special Education teacher moving to a General Education teacher). In order to capture that change, we can come up with a new variable that records the number of times the teacher changed schools, titles, etc.

Variables that change for each ID, such as gender, or school, will be gathered and then re-coded. For instance, the original data set had a unique entry for each year in the data. Here, that is recoded as "total years in TIP data set". Or, some teachers changed gender over the time frame. This is recoded as "number of genders". Or, some teachers changed schools during the time frame. This is recoded as "number of schools".

These new variables, named "number of titles", or "numer of schools", are then merged back with merged_participation, to give us one complete dataset for analysis. Limitations to this approach are discussed later.

We end up with a data set of total of 16012 unique teachers, 312 of whom participanted and 15700 of whom did not participate. An additional note is that the analysis data set here includes a total of 16003 observations, 9 were removed because the variable for Gender was missing.

In choosing the variables to analyze from the analysis dataset, we must make a choice about which variables to choose. For instance, some teachers have more than 1 role, or more than 1 school worked at. In this example code, I simply choose the first observation that appears for that teacher. Since the information is organized by employee age, this is the youngest age at which the teacher started. We can later choose to organize this information differently if need be.

Model 1

$$log(r/(1-r)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...\beta_p x_p$$

The logistic regression model (above) forms the basis for the following statistical tests. Where r is proability of a teacher being retained or not retained, and each x represents a possible predictor, such as gender or years of service in the district, then the model tests the significance of each regression coefficient in predicting the log odds of retention.

First, we test the model using participation in TIP as the predictor and retention as the outcome variable.

$$log(r/(1-r)) = \beta_0 + \beta_1 x_1$$

$$\beta_1 = Participation$$

The output of Model 1 shows that participation in TIP increases the log odds of retention by 0.35. In other words, TIP teachers are 35% more likely to be retained than non-TIP teachers. This value is statistically significant, given that p <.05.

```
##
## Call:
## glm(formula = Retained ~ Participated, family = binomial, data = analysis_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5628  -1.4041   0.9665   0.9665   0.9665
##
```

```
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)    
## (Intercept)    0.51868    0.01651   31.43  < 2e-16 ***
## Participated1  0.35315    0.12525    2.82  0.00481 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 21121  on 16004  degrees of freedom
## Residual deviance: 21113  on 16003  degrees of freedom
## AIC: 21117
## 
## Number of Fisher Scoring iterations: 4
```

	x
(Intercept)	0.518684
Participated1	0.353155

Model 2

Next, we test the model controlling for the effects of Gender and Ethnicity. Controlling for ethnicity and participation in TIP, Gender is not a significant predictor of retention. We compared this full model (with the three predictors) against models that only included Gender and Ethnicity, or Participation and Gender, or Participation and Ethnicity, or one of each of the variables. The results from the F test comparing these models indicate that the model with Participation and Gender explains the most variation in Retention Rate, rather than the model with all three predictor variables.

Ethnicity, on the other hand is a significant predictor for 4 of the 8 listed ethnicities. GO ON HERE

$$log(r/(1-r)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

$$log(r/(1-r)) = \beta_0 + \beta_1 x_1$$

$$\beta_1 = Participation$$

$$\beta_2 = Gender$$

$$\beta_3 = Ethnicity$$

```
## 
## Call:  glm(formula = Retained ~ Participated + Ethnicity, family = binomial, 
##     data = analysis_data)
## 
## Coefficients:
##           (Intercept)          Participated1     EthnicityASIAN/PAC  
##              0.575070               0.349143              -0.005711  
##    EthnicityCAUCASIAN      EthnicityLATINA/O  EthnicityNAT AM/INUIT  
##             -0.096267               0.200386              -0.004525  
##        EthnicityOTHER  EthnicityPREF NO DISC  
##              0.023313               2.057279  
## 
## Degrees of Freedom: 16004 Total (i.e. Null);  15997 Residual
## Null Deviance:      21120
## Residual Deviance: 21070     AIC: 21090
```

```
##
## Call:
## glm(formula = Retained ~ Participated + Ethnicity, family = binomial,
##     data = analysis_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3246  -1.3863   0.9449   0.9820   0.9820
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)            0.575070   0.033244  17.298  < 2e-16 ***
## Participated1          0.349143   0.125413   2.784 0.005370 **
## EthnicityASIAN/PAC    -0.005711   0.113672  -0.050 0.959929
## EthnicityCAUCASIAN    -0.096267   0.038554  -2.497 0.012527 *
## EthnicityLATINA/O      0.200386   0.101680   1.971 0.048751 *
## EthnicityNAT AM/INUIT -0.004525   0.348578  -0.013 0.989642
## EthnicityOTHER         0.023313   0.163919   0.142 0.886904
## EthnicityPREF NO DISC  2.057279   0.597487   3.443 0.000575 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 21121  on 16004  degrees of freedom
## Residual deviance: 21075  on 15997  degrees of freedom
## AIC: 21091
##
## Number of Fisher Scoring iterations: 4
```

```
##   (Intercept) Participated1
##      0.518684      0.353155
```

	x
(Intercept)	0.518684
Participated1	0.353155

Model 3

Next, we examine the interaction effects of Strata and Participation. While strata itself was a signficant predictor of retention, the interaction between strata and participation in TIP was not significant.

$$log(r/(1-r)) = \beta_0 + \beta_1 x_1 * \beta_2 x_2$$

$$\beta_1 = Strata$$

$$\beta_2 = Participation$$

```
##
## Call:
## glm(formula = Retained ~ Strata * Participated, family = binomial,
##     data = analysis_data)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.9103  -1.0371   0.7851   0.9084   1.3245  
## 
## Coefficients:
##                                                 Estimate Std. Error z value
## (Intercept)                                      0.67194    0.02308  29.114
## Strata5 to less than 10 years                    0.34691    0.05768   6.014
## Strata10 to 15 years                             0.38121    0.05569   6.845
## Stratagreater than or equal to 15               -1.01141    0.04159 -24.317
## Participated1                                    0.21651    0.16794   1.289
## Strata5 to less than 10 years:Participated1      0.41329    0.38760   1.066
## Strata10 to 15 years:Participated1              -0.13023    0.44246  -0.294
## Stratagreater than or equal to 15:Participated1 -0.06810    0.35428  -0.192
##                                                 Pr(>|z|)    
## (Intercept)                                      < 2e-16 ***
## Strata5 to less than 10 years                   1.81e-09 ***
## Strata10 to 15 years                            7.67e-12 ***
## Stratagreater than or equal to 15                < 2e-16 ***
## Participated1                                      0.197    
## Strata5 to less than 10 years:Participated1        0.286    
## Strata10 to 15 years:Participated1                 0.769    
## Stratagreater than or equal to 15:Participated1    0.848    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 21121  on 16004  degrees of freedom
## Residual deviance: 20207  on 15997  degrees of freedom
## AIC: 20223
## 
## Number of Fisher Scoring iterations: 4
```

```
## % latex table generated in R 4.0.2 by xtable 1.8-4 package
## % Thu Aug  6 13:28:49 2020
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrrrr}
##   \hline
##  & Estimate & Std. Error & z value & Pr($>$$|$z$|$) \\ 
##   \hline
## (Intercept) & 0.6719 & 0.0231 & 29.11 & 0.0000 \\ 
##   Strata5 to less than 10 years & 0.3469 & 0.0577 & 6.01 & 0.0000 \\ 
##   Strata10 to 15 years & 0.3812 & 0.0557 & 6.84 & 0.0000 \\ 
##   Stratagreater than or equal to 15 & -1.0114 & 0.0416 & -24.32 & 0.0000 \\ 
##   Participated1 & 0.2165 & 0.1679 & 1.29 & 0.1973 \\ 
##   Strata5 to less than 10 years:Participated1 & 0.4133 & 0.3876 & 1.07 & 0.2863 \\ 
##   Strata10 to 15 years:Participated1 & -0.1302 & 0.4425 & -0.29 & 0.7685 \\ 
##   Stratagreater than or equal to 15:Participated1 & -0.0681 & 0.3543 & -0.19 & 0.8476 \\ 
##    \hline
## \end{tabular}
## \end{table}
##
## Call:
## glm(formula = Retained ~ Strata * Participated + Ethnicity +
##     Gender, family = binomial, data = analysis_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3328  -1.0970   0.8003   0.9204   1.4089
##
## Coefficients:
##                                                 Estimate Std. Error z value
## (Intercept)                                      0.85783    0.04164  20.600
## Strata5 to less than 10 years                    0.33447    0.05789   5.778
## Strata10 to 15 years                             0.35593    0.05604   6.351
## Stratagreater than or equal to 15               -1.04989    0.04261 -24.638
## Participated1                                    0.20247    0.16831   1.203
## EthnicityASIAN/PAC                              -0.28392    0.11657  -2.436
## EthnicityCAUCASIAN                              -0.21802    0.04079  -5.345
## EthnicityLATINA/O                               -0.01796    0.10453  -0.172
## EthnicityNAT AM/INUIT                           -0.34314    0.35154  -0.976
## EthnicityOTHER                                  -0.28507    0.16555  -1.722
## EthnicityPREF NO DISC                            1.79508    0.59794   3.002
## GenderM                                         -0.05354    0.03820  -1.402
## Strata5 to less than 10 years:Participated1      0.43137    0.38802   1.112
## Strata10 to 15 years:Participated1              -0.14739    0.44312  -0.333
## Stratagreater than or equal to 15:Participated1 -0.09570    0.35492  -0.270
##                                                 Pr(>|z|)
## (Intercept)                                      < 2e-16 ***
## Strata5 to less than 10 years                   7.58e-09 ***
## Strata10 to 15 years                            2.14e-10 ***
## Stratagreater than or equal to 15                < 2e-16 ***
## Participated1                                    0.22900
## EthnicityASIAN/PAC                               0.01486 *
## EthnicityCAUCASIAN                              9.03e-08 ***
## EthnicityLATINA/O                                0.86355
## EthnicityNAT AM/INUIT                            0.32900
## EthnicityOTHER                                   0.08508 .
## EthnicityPREF NO DISC                            0.00268 **
## GenderM                                          0.16098
## Strata5 to less than 10 years:Participated1      0.26625
## Strata10 to 15 years:Participated1               0.73943
## Stratagreater than or equal to 15:Participated1  0.78743
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 21121  on 16004  degrees of freedom
## Residual deviance: 20152  on 15990  degrees of freedom
## AIC: 20182
##
## Number of Fisher Scoring iterations: 4
## % latex table generated in R 4.0.2 by xtable 1.8-4 package
## % Thu Aug  6 13:28:49 2020
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrrrr}
##   \hline
##  & Estimate & Std. Error & z value & Pr($>$$|$z$|$) \\
##   \hline
## (Intercept) & 0.8578 & 0.0416 & 20.60 & 0.0000 \\
##   Strata5 to less than 10 years & 0.3345 & 0.0579 & 5.78 & 0.0000 \\
##   Strata10 to 15 years & 0.3559 & 0.0560 & 6.35 & 0.0000 \\
##   Stratagreater than or equal to 15 & -1.0499 & 0.0426 & -24.64 & 0.0000 \\
##   Participated1 & 0.2025 & 0.1683 & 1.20 & 0.2290 \\
##   EthnicityASIAN/PAC & -0.2839 & 0.1166 & -2.44 & 0.0149 \\
##   EthnicityCAUCASIAN & -0.2180 & 0.0408 & -5.35 & 0.0000 \\
##   EthnicityLATINA/O & -0.0180 & 0.1045 & -0.17 & 0.8635 \\
##   EthnicityNAT AM/INUIT & -0.3431 & 0.3515 & -0.98 & 0.3290 \\
##   EthnicityOTHER & -0.2851 & 0.1655 & -1.72 & 0.0851 \\
##   EthnicityPREF NO DISC & 1.7951 & 0.5979 & 3.00 & 0.0027 \\
##   GenderM & -0.0535 & 0.0382 & -1.40 & 0.1610 \\
##   Strata5 to less than 10 years:Participated1 & 0.4314 & 0.3880 & 1.11 & 0.2662 \\
##   Strata10 to 15 years:Participated1 & -0.1474 & 0.4431 & -0.33 & 0.7394 \\
##   Stratagreater than or equal to 15:Participated1 & -0.0957 & 0.3549 & -0.27 & 0.7874 \\
##    \hline
## \end{tabular}
## \end{table}
```

Discussion/Limitations

The first question of interest is the retention of teachers in SDP (the School District of Philadelphia) who participated or did not particpate in TIP. The analyses show that TIP is a significant predictor of retention for teachers in SDP.

Model 1 shows us that participating in TIP did have a positive effect on teacher retention.

Model 4 demonstrates how this effect varied for different strata and year levels. This resonates with the visual nature of the plots and graphs, which show that the liklihood of retention for TIP teachers changes in each Strata.

In model 3.5, we also analyze the predictor variables on years of service as predictors of teacher retention. Teachers who served for 1,3, and 4 years were more likely to leave than teachers who stayed for 1 years, as were teachers who stayed for 13,21, and 22 years.

Next steps will be to look more closely at the differences in teacher retention rates betweeen strata and number of years served for TIP teachers. As mentioned in the original proposal, "non-retention of high-performing teachers is a serious problem for urban school districts". Better understanding the patterns of retention based on strata and years worked can shed light the specific groups that a program like TIP has the most effect on, and better yet, why, such a program impacts retention rates.

A few limitations on the current analysis exist. First, the sample size of TIP teachers is a fraction of the sample size of non_TIP teachers. Such sample size differences may result in complications for future interaction analyses or causal interpretations. Additionally, teacher retention is calculated as a teacher staying in the school district dataset from one year to another. This fails to capture differencees in changed Titles of teachers, or where specifically teachers move throughout the system.

Lastly, an additional variable that would help to answer the original research questions posed in the proposal would be an indicator of the date that teachers began the TIP program. As it stands, the only information that we have from TIP is the Deidentified IDs of participants, which were matched with administrative data from the 8 year period. If given an indicator of the start date and/or end date that teachers participated in TIP, we could analyze trends like amount of time participated in TIP, and it's effect on retention.

Appendix B

Kat Wilson

7/28/2020

Matching Model 1

You can also embed plots, for example:

```
### Matching 1- All covariates in the matching model
##set seed
set.seed(1731)
class(analysis_data$Strata)
```

```
## [1] "factor"
```

```
analysis_data$Gender <- as.factor(analysis_data$Gender)
analysis_data$employeeage <- as.numeric(analysis_data$`Employee Age`)
analysis_data$title_status <- as.factor(analysis_data$title_status)
analysis_data$Ethnicity <- as.factor(analysis_data$Ethnicity)
analysis_data$HomeOrgCode <- as.factor(analysis_data$`HOME ORG CODE`)
analysis_data$YearsOfService <- as.numeric(analysis_data$`Years of Service`)
nearest <- matchit(Participated ~
                     Gender+ Ethnicity + Strata + employeeage+ title_status
                     ,
                   family = "binomial",
                   method = "nearest",
                   caliper = 0.25,
                   data = analysis_data)
summary(nearest)
```

```
##
## Call:
## matchit(formula = Participated ~ Gender + Ethnicity + Strata +
##     employeeage + title_status, data = analysis_data, method = "nearest",
##     family = "binomial", caliper = 0.25)
##
## Summary of balance for all data:
##                                    Means Treated Means Control SD Control
## distance                                 0.0244        0.0194     0.0097
## GenderF                                  0.7564        0.7360     0.4408
## GenderM                                  0.2436        0.2640     0.4408
## EthnicityASIAN/PAC                       0.0417        0.0228     0.1493
## EthnicityCAUCASIAN                       0.6026        0.6836     0.4651
## EthnicityLATINA/O                        0.0128        0.0320     0.1760
## Ethnicityother                           0.0160        0.0157     0.1242
## Strata5 to less than 10 years            0.2244        0.1444     0.3515
## Strata10 to 15 years                     0.2372        0.1481     0.3552
## Stratagreater than or equal to 15        0.2724        0.3551     0.4786
## employeeage                             44.4006       43.8269    13.3040
## title_statusteacher                      0.9776        0.9323     0.2512
##                                   Mean Diff eQQ Med eQQ Mean eQQ Max
## distance                             0.0050  0.0055   0.0050   0.014
## GenderF                              0.0204  0.0000   0.0224   1.000
## GenderM                             -0.0204  0.0000   0.0224   1.000
## EthnicityASIAN/PAC                   0.0189  0.0000   0.0160   1.000
## EthnicityCAUCASIAN                  -0.0810  0.0000   0.0801   1.000
## EthnicityLATINA/O                   -0.0192  0.0000   0.0192   1.000
## Ethnicityother                       0.0004  0.0000   0.0000   0.000
## Strata5 to less than 10 years        0.0799  0.0000   0.0801   1.000
## Strata10 to 15 years                 0.0891  0.0000   0.0865   1.000
## Stratagreater than or equal to 15   -0.0827  0.0000   0.0833   1.000
## employeeage                          0.5738  1.0000   0.7212  14.000
## title_statusteacher                  0.0452  0.0000   0.0481   1.000
##
##
## Summary of balance for matched data:
##                                   Means Treated Means Control SD Control
## distance                                 0.0244        0.0244     0.0113
## GenderF                                  0.7564        0.7788     0.4157
## GenderM                                  0.2436        0.2212     0.4157
## EthnicityASIAN/PAC                       0.0417        0.0353     0.1847
## EthnicityCAUCASIAN                       0.6026        0.6058     0.4895
## EthnicityLATINA/O                        0.0128        0.0064     0.0799
## Ethnicityother                           0.0160        0.0192     0.1376
## Strata5 to less than 10 years            0.2244        0.1923     0.3947
## Strata10 to 15 years                     0.2372        0.2628     0.4409
## Stratagreater than or equal to 15        0.2724        0.2692     0.4443
## employeeage                             44.4006       44.5737    13.2129
## title_statusteacher                      0.9776        0.9712     0.1676
##                                   Mean Diff eQQ Med eQQ Mean eQQ Max
## distance                             0.0000   2e-04   0.0003  0.0024
## GenderF                             -0.0224   0e+00   0.0224  1.0000
## GenderM                              0.0224   0e+00   0.0224  1.0000
## EthnicityASIAN/PAC                   0.0064   0e+00   0.0064  1.0000
## EthnicityCAUCASIAN                  -0.0032   0e+00   0.0032  1.0000
## EthnicityLATINA/O                    0.0064   0e+00   0.0064  1.0000
## Ethnicityother                      -0.0032   0e+00   0.0032  1.0000
## Strata5 to less than 10 years        0.0321   0e+00   0.0321  1.0000
## Strata10 to 15 years                -0.0256   0e+00   0.0256  1.0000
## Stratagreater than or equal to 15    0.0032   0e+00   0.0032  1.0000
## employeeage                         -0.1731   0e+00   0.5705 14.0000
## title_statusteacher                  0.0064   0e+00   0.0064  1.0000
##
## Percent Balance Improvement:
##                                   Mean Diff.  eQQ Med eQQ Mean eQQ Max
## distance                              99.2034  96.5816  94.2965 82.9651
## GenderF                              -10.0896   0.0000   0.0000  0.0000
## GenderM                              -10.0896   0.0000   0.0000  0.0000
## EthnicityASIAN/PAC                    66.0057   0.0000  60.0000  0.0000
## EthnicityCAUCASIAN                    96.0445   0.0000  96.0000  0.0000
## EthnicityLATINA/O                     66.5509   0.0000  66.6667  0.0000
## Ethnicityother                      -810.9112   0.0000     -Inf    -Inf
## Strata5 to less than 10 years        59.8948   0.0000  60.0000   0.0000
## Strata10 to 15 years                 71.2244   0.0000  70.3704   0.0000
## Stratagreater than or equal to 15    96.1248   0.0000  96.1538   0.0000
## employeeage                          69.8343 100.0000  20.8889   0.0000
## title_statusteacher                  85.8271   0.0000  86.6667   0.0000
##
## Sample sizes:
##           Control Treated
## All         15695     312
## Matched       312     312
## Unmatched   15383       0
## Discarded       0       0
```

```
###plotting these
#plot(nearest)
nearest_matched <- match.data(nearest)
#now perform the statistical analysis
nearest_matched$Participated <- as.factor(nearest_matched$Participated)
model <- glm(Retained ~ Participated, data = nearest_matched,
             family = binomial)
summary(model)
```

```
##
## Call:
## glm(formula = Retained ~ Participated, family = binomial, data = nearest_matched)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5628  -1.4694   0.8359   0.9112   0.9112
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.6644     0.1195   5.559 2.72e-08 ***
## Participated1   0.2074     0.1723   1.203    0.229
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 779.78  on 623  degrees of freedom
## Residual deviance: 778.33  on 622  degrees of freedom
## AIC: 782.33
##
## Number of Fisher Scoring iterations: 4
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Trimming

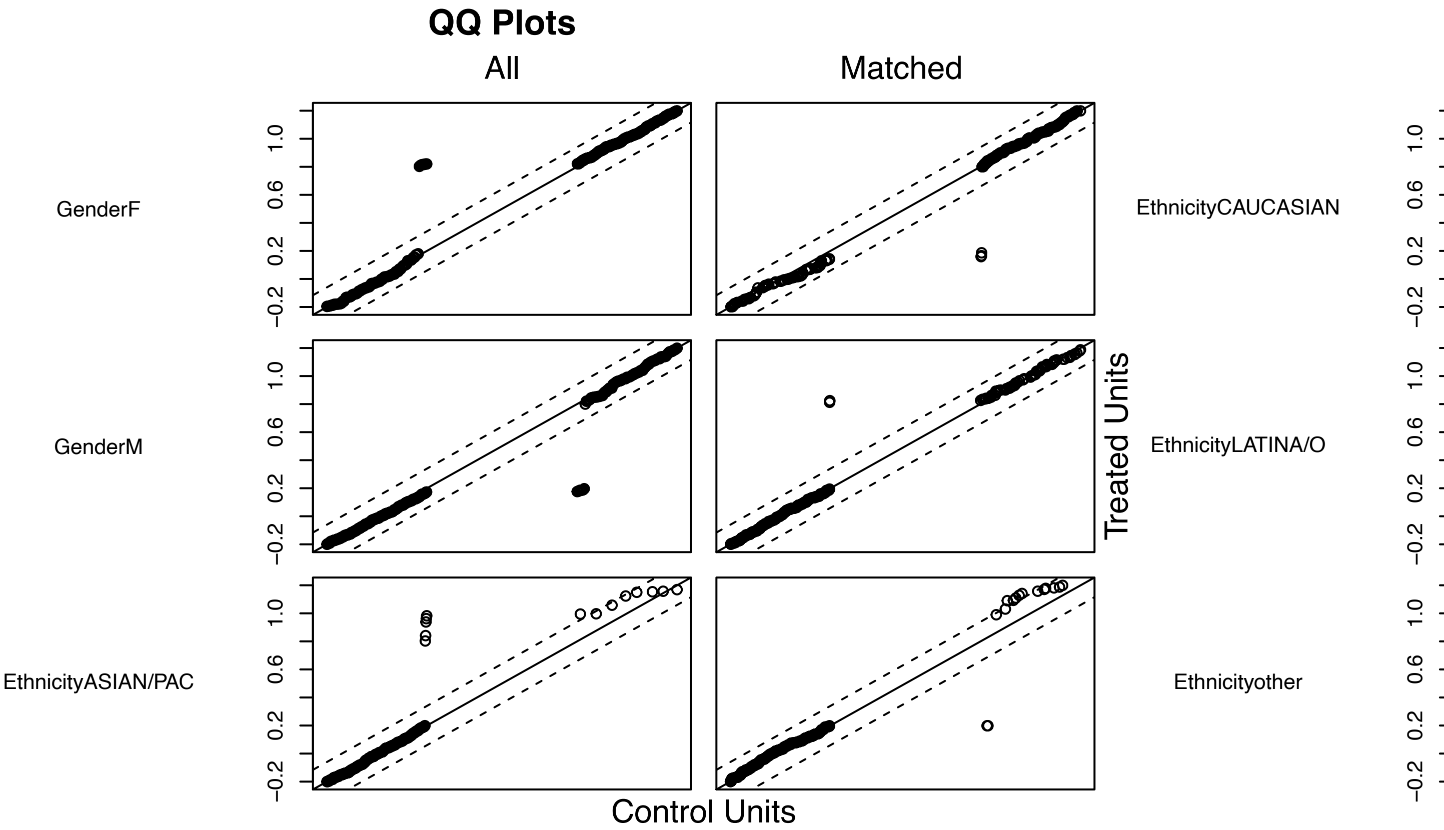
Trim the non TIP sample so that it is only tip in terms of Years of Experience. There are no 32 years olds, 38 year olds, or 40 or older TIP teachers. However, these grouped ARE represented in the non-TIP sample.

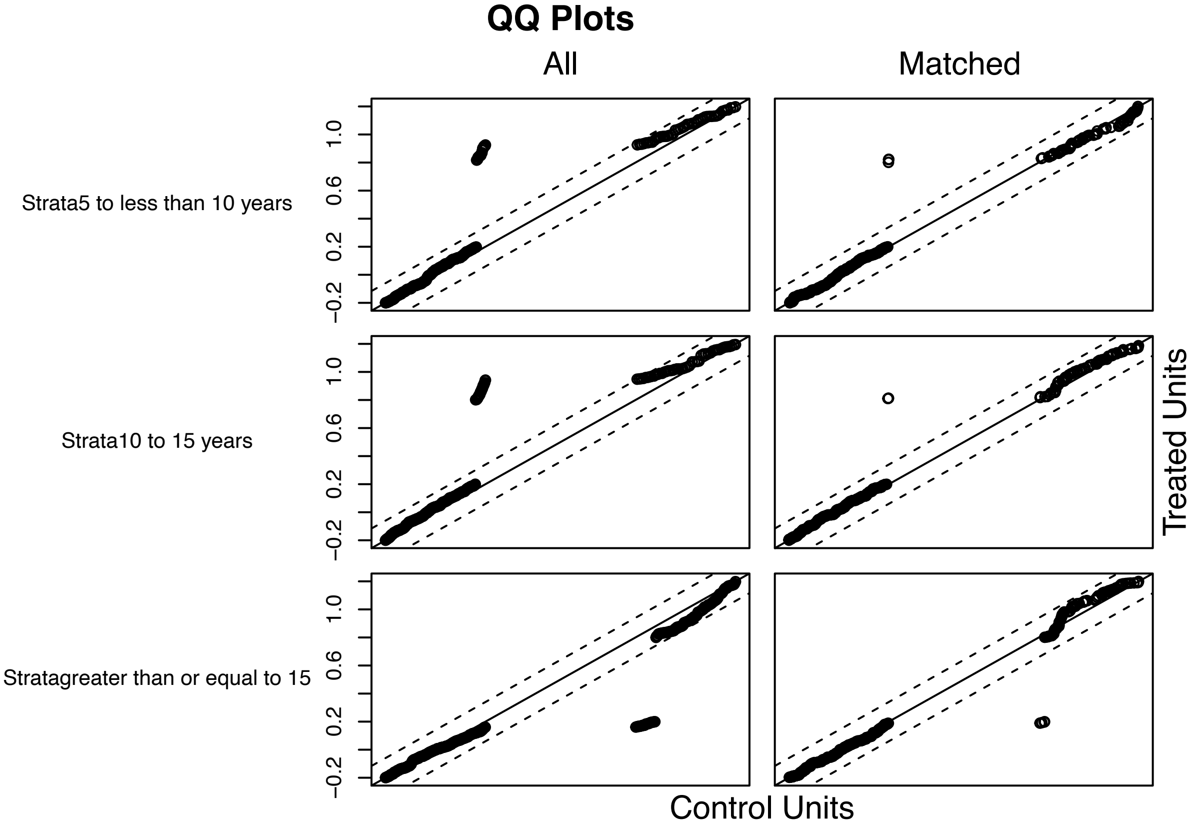
```
set.seed(1731)
# trimming the sample
```

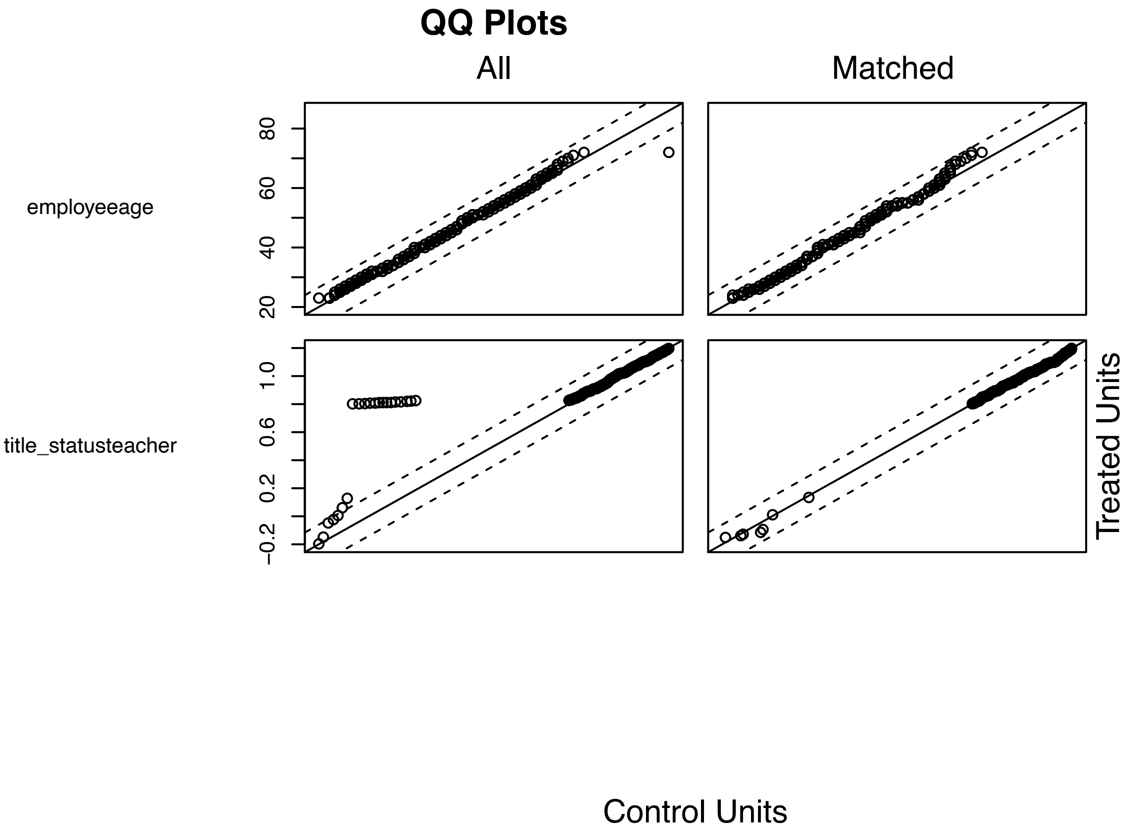
```r
analysis_data_trimmed <- analysis_data %>%
  filter(YearsOfService %in% c(0:31, 33,34,35,36,37,39))

nearest <- matchit(Participated ~
                     Gender+ Ethnicity + Strata + employeeage+ title_status
                     ,
                   family = "binomial",
                   method = "nearest",
                   caliper = 0.25,
                   data = analysis_data_trimmed)
summary(nearest)
```

```
##
## Call:
## matchit(formula = Participated ~ Gender + Ethnicity + Strata +
##     employeeage + title_status, data = analysis_data_trimmed,
##     method = "nearest", family = "binomial", caliper = 0.25)
##
## Summary of balance for all data:
##                                   Means Treated Means Control SD Control
## distance                                 0.0247        0.0198     0.0097
## GenderF                                  0.7564        0.7357     0.4410
## GenderM                                  0.2436        0.2643     0.4410
## EthnicityASIAN/PAC                       0.0417        0.0231     0.1503
## EthnicityCAUCASIAN                       0.6026        0.6850     0.4645
## EthnicityLATINA/O                        0.0128        0.0325     0.1774
## Ethnicityother                           0.0160        0.0160     0.1254
## Strata5 to less than 10 years            0.2244        0.1473     0.3544
## Strata10 to 15 years                     0.2372        0.1510     0.3580
## Stratagreater than or equal to 15        0.2724        0.3426     0.4746
## employeeage                             44.4006       43.4807    13.1841
## title_statusteacher                      0.9776        0.9339     0.2485
##                                   Mean Diff eQQ Med eQQ Mean eQQ Max
## distance                             0.0049  0.0048   0.0049  0.0136
## GenderF                              0.0207  0.0000   0.0224  1.0000
## GenderM                             -0.0207  0.0000   0.0224  1.0000
## EthnicityASIAN/PAC                   0.0185  0.0000   0.0160  1.0000
## EthnicityCAUCASIAN                  -0.0825  0.0000   0.0833  1.0000
## EthnicityLATINA/O                   -0.0197  0.0000   0.0224  1.0000
## Ethnicityother                       0.0000  0.0000   0.0000  0.0000
## Strata5 to less than 10 years        0.0771  0.0000   0.0769  1.0000
## Strata10 to 15 years                 0.0862  0.0000   0.0865  1.0000
## Stratagreater than or equal to 15   -0.0701  0.0000   0.0705  1.0000
## employeeage                          0.9200  1.0000   0.9647 14.0000
## title_statusteacher                  0.0437  0.0000   0.0449  1.0000
##
##
## Summary of balance for matched data:
##                                   Means Treated Means Control SD Control
## distance                                 0.0247        0.0245     0.0112
## GenderF                                  0.7564        0.7660     0.4240
## GenderM                                  0.2436        0.2340     0.4240
## EthnicityASIAN/PAC                       0.0417        0.0481     0.2143
## EthnicityCAUCASIAN                       0.6026        0.5769     0.4948
## EthnicityLATINA/O                        0.0128        0.0160     0.1258
## Ethnicityother                           0.0160        0.0160     0.1258
## Strata5 to less than 10 years            0.2244        0.2147     0.4113
## Strata10 to 15 years                     0.2372        0.2276     0.4199
## Stratagreater than or equal to 15        0.2724        0.2853     0.4523
## employeeage                             44.4006       43.8686    12.5905
## title_statusteacher                      0.9776        0.9776     0.1483
##                                   Mean Diff eQQ Med eQQ Mean eQQ Max
## distance                             0.0002   2e-04   0.0003  0.0023
## GenderF                             -0.0096   0e+00   0.0096  1.0000
## GenderM                              0.0096   0e+00   0.0096  1.0000
## EthnicityASIAN/PAC                  -0.0064   0e+00   0.0064  1.0000
## EthnicityCAUCASIAN                   0.0256   0e+00   0.0256  1.0000
## EthnicityLATINA/O                   -0.0032   0e+00   0.0032  1.0000
## Ethnicityother                       0.0000   0e+00   0.0000  0.0000
## Strata5 to less than 10 years        0.0096   0e+00   0.0096  1.0000
## Strata10 to 15 years                 0.0096   0e+00   0.0096  1.0000
## Stratagreater than or equal to 15   -0.0128   0e+00   0.0128  1.0000
## employeeage                          0.5321   1e+00   0.8590  5.0000
## title_statusteacher                  0.0000   0e+00   0.0000  0.0000
##
## Percent Balance Improvement:
##                                   Mean Diff. eQQ Med eQQ Mean  eQQ Max
## distance                              96.4903 96.4995  93.7170  82.8878
## GenderF                               53.5866  0.0000  57.1429   0.0000
## GenderM                               53.5866  0.0000  57.1429   0.0000
## EthnicityASIAN/PAC                    65.4290  0.0000  60.0000   0.0000
## EthnicityCAUCASIAN                    68.9062  0.0000  69.2308   0.0000
## EthnicityLATINA/O                     83.7489  0.0000  85.7143   0.0000
## Ethnicityother                       100.0000  0.0000   0.0000   0.0000
## Strata5 to less than 10 years         87.5292  0.0000  87.5000   0.0000
## Strata10 to 15 years                  88.8480  0.0000  88.8889   0.0000
## Stratagreater than or equal to 15     81.7223  0.0000  81.8182   0.0000
## employeeage                           42.1662  0.0000  10.9635  64.2857
## title_statusteacher                  100.0000  0.0000 100.0000 100.0000
##
## Sample sizes:
##           Control Treated
## All         15395     312
## Matched       312     312
## Unmatched   15083       0
## Discarded       0       0
```

```
###plotting these
plot(nearest)
```


QQ Plots
All
Matched
GenderF
GenderM
EthnicityASIAN/PAC
EthnicityCAUCASIAN
EthnicityLATINA/O
Ethnicityother
−0.2
0.2
0.6
1.0
Treated Units
Control Units


QQ Plots
All
Matched
Strata5 to less than 10 years
Strata10 to 15 years
Stratagreater than or equal to 15
−0.2
0.2
0.6
1.0
Treated Units
Control Units



```
nearest_matched <- match.data(nearest)
#now perform the statistical analysis
nearest_matched$Participated <- as.factor(nearest_matched$Participated)
model <- glm(Retained ~ Participated, data = nearest_matched,
             family = binomial)
summary(model)
```

```
##
## Call:
## glm(formula = Retained ~ Participated, family = binomial, data = nearest_matched)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5628  -1.3585   0.8359   1.0065   1.0065
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.4162     0.1157   3.597 0.000322 ***
## Participated1     0.4557     0.1697   2.685 0.007249 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 805.00  on 623  degrees of freedom
## Residual deviance: 797.73  on 622  degrees of freedom
## AIC: 801.73
##
```

```
## Number of Fisher Scoring iterations: 4
```

Coursen the Ethnicity Variable

Coursen the Ethnicity variable. Match it by Caucasian and non-Caucasian

```
analysis_data <- analysis_data %>%
  mutate(Caucasian_binary = ifelse(Ethnicity == "CAUCASIAN", 1, 0))
analysis_data$Caucasian_binary <- as.factor(analysis_data$Caucasian_binary)
set.seed(111731)
nearest <- matchit(Participated ~
                     Gender+ Caucasian_binary + Strata + employeeage+ title_status
                     ,
                   family = "binomial",
                   method = "nearest",
                   caliper = 0.25,
                   data = analysis_data)
summary(nearest)
```

```
##
## Call:
## matchit(formula = Participated ~ Gender + Caucasian_binary +
##     Strata + employeeage + title_status, data = analysis_data,
##     method = "nearest", family = "binomial", caliper = 0.25)
##
## Summary of balance for all data:
##                                     Means Treated Means Control SD Control
## distance                                   0.0237        0.0194     0.0091
## GenderF                                    0.7564        0.7360     0.4408
## GenderM                                    0.2436        0.2640     0.4408
## Caucasian_binary1                          0.6026        0.6836     0.4651
## Strata5 to less than 10 years              0.2244        0.1444     0.3515
## Strata10 to 15 years                       0.2372        0.1481     0.3552
## Stratagreater than or equal to 15          0.2724        0.3551     0.4786
## employeeage                               44.4006       43.8269    13.3040
## title_statusteacher                        0.9776        0.9323     0.2512
##                                     Mean Diff eQQ Med eQQ Mean eQQ Max
## distance                               0.0043  0.0044   0.0043  0.0108
## GenderF                                0.0204  0.0000   0.0224  1.0000
## GenderM                               -0.0204  0.0000   0.0224  1.0000
## Caucasian_binary1                     -0.0810  0.0000   0.0801  1.0000
## Strata5 to less than 10 years          0.0799  0.0000   0.0801  1.0000
## Strata10 to 15 years                   0.0891  0.0000   0.0865  1.0000
## Stratagreater than or equal to 15     -0.0827  0.0000   0.0833  1.0000
## employeeage                            0.5738  1.0000   0.7212 14.0000
## title_statusteacher                    0.0452  0.0000   0.0481  1.0000
##
##
## Summary of balance for matched data:
##                                     Means Treated Means Control SD Control
## distance                                   0.0237        0.0237     0.0099
## GenderF                                    0.7564        0.7692     0.4220
## GenderM                                    0.2436        0.2308     0.4220
## Caucasian_binary1                          0.6026        0.6058     0.4895
## Strata5 to less than 10 years             0.2244        0.2308     0.4220
## Strata10 to 15 years                      0.2372        0.2276     0.4199
## Stratagreater than or equal to 15         0.2724        0.3269     0.4698
## employeeage                              44.4006       45.7500    12.2871
## title_statusteacher                       0.9776        0.9808     0.1376
##                                  Mean Diff eQQ Med eQQ Mean eQQ Max
## distance                            0.0000   1e-04   0.0003  0.0018
## GenderF                            -0.0128   0e+00   0.0128  1.0000
## GenderM                             0.0128   0e+00   0.0128  1.0000
## Caucasian_binary1                  -0.0032   0e+00   0.0032  1.0000
## Strata5 to less than 10 years      -0.0064   0e+00   0.0064  1.0000
## Strata10 to 15 years                0.0096   0e+00   0.0096  1.0000
## Stratagreater than or equal to 15  -0.0545   0e+00   0.0545  1.0000
## employeeage                        -1.3494   2e+00   1.8942  4.0000
## title_statusteacher                -0.0032   0e+00   0.0032  1.0000
##
## Percent Balance Improvement:
##                                  Mean Diff.   eQQ Med  eQQ Mean eQQ Max
## distance                            98.8392   96.6484   93.9630 83.3064
## GenderF                             37.0917    0.0000   42.8571  0.0000
## GenderM                             37.0917    0.0000   42.8571  0.0000
## Caucasian_binary1                   96.0445    0.0000   96.0000  0.0000
## Strata5 to less than 10 years       91.9790    0.0000   92.0000  0.0000
## Strata10 to 15 years                89.2092    0.0000   88.8889  0.0000
## Stratagreater than or equal to 15   34.1219    0.0000   34.6154  0.0000
## employeeage                       -135.1810 -100.0000 -162.6667 71.4286
## title_statusteacher                 92.9135    0.0000   93.3333  0.0000
##
## Sample sizes:
##           Control Treated
## All         15695     312
## Matched       312     312
## Unmatched   15383       0
## Discarded       0       0
```

```
###plotting these
#plot(nearest)
nearest_matched <- match.data(nearest)
#now perform the statistical analysis
nearest_matched$Participated <- as.factor(nearest_matched$Participated)
model <- glm(Retained ~ Participated, data = nearest_matched,
             family = binomial)
summary(model)
```

```
##
## Call:
## glm(formula = Retained ~ Participated, family = binomial, data = nearest_matched)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5628  -1.3764   0.8359   0.9907   0.9907
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.4565     0.1162   3.929 8.54e-05 ***
## Participated1   0.4154     0.1700   2.443   0.0146 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 801.12  on 623  degrees of freedom
## Residual deviance: 795.11  on 622  degrees of freedom
## AIC: 799.11
##
## Number of Fisher Scoring iterations: 4
```

```r
table(analysis_data$Age, analysis_data$Caucasian_binary)
```

```
##
##          0    1
##   20s  549 2372
##   30s 1034 2895
##   40s 1154 2032
##   50s 1410 1883
##   60s  918 1689
##   70s   25   46
```

Coursen the Ethnicity Variable and use with the trimmed data

Coursen the Ethnicity variable. Match it by Caucasian and non-Caucasian

```r
analysis_data_trimmed <- analysis_data_trimmed %>%
  mutate(Caucasian_binary = ifelse(Ethnicity == "CAUCASIAN", 1, 0))
analysis_data_trimmed$Caucasian_binary <- as.factor(analysis_data_trimmed$Caucasian_binary)
set.seed(1731)
nearest <- matchit(Participated ~
                     Gender+ Caucasian_binary + Strata + employeeage+ title_status
                     ,
                   family = "binomial",
                   method = "nearest",
                   caliper = 0.25,
                   data = analysis_data_trimmed)
summary(nearest)
```

```
##
## Call:
## matchit(formula = Participated ~ Gender + Caucasian_binary +
##     Strata + employeeage + title_status, data = analysis_data_trimmed,
##     method = "nearest", family = "binomial", caliper = 0.25)
##
## Summary of balance for all data:
##                                     Means Treated Means Control SD Control
## distance                                   0.0240        0.0198     0.0091
## GenderF                                    0.7564        0.7357     0.4410
## GenderM                                    0.2436        0.2643     0.4410
## Caucasian_binary1                          0.6026        0.6850     0.4645
## Strata5 to less than 10 years              0.2244        0.1473     0.3544
## Strata10 to 15 years                       0.2372        0.1510     0.3580
## Stratagreater than or equal to 15          0.2724        0.3426     0.4746
## employeeage                             44.4006       43.4807    13.1841
## title_statusteacher                      0.9776        0.9339     0.2485
##                                   Mean Diff eQQ Med eQQ Mean eQQ Max
## distance                             0.0042  0.0046   0.0042  0.0101
## GenderF                              0.0207  0.0000   0.0224  1.0000
## GenderM                             -0.0207  0.0000   0.0224  1.0000
## Caucasian_binary1                   -0.0825  0.0000   0.0833  1.0000
## Strata5 to less than 10 years        0.0771  0.0000   0.0769  1.0000
## Strata10 to 15 years                 0.0862  0.0000   0.0865  1.0000
## Stratagreater than or equal to 15   -0.0701  0.0000   0.0705  1.0000
## employeeage                          0.9200  1.0000   0.9647 14.0000
## title_statusteacher                  0.0437  0.0000   0.0449  1.0000
##
##
## Summary of balance for matched data:
##                                   Means Treated Means Control SD Control
## distance                                0.0240        0.0238     0.0098
## GenderF                                 0.7564        0.7308     0.4443
## GenderM                                 0.2436        0.2692     0.4443
## Caucasian_binary1                       0.6026        0.6058     0.4895
## Strata5 to less than 10 years           0.2244        0.2340     0.4240
## Strata10 to 15 years                    0.2372        0.2340     0.4240
## Stratagreater than or equal to 15       0.2724        0.2692     0.4443
## employeeage                            44.4006       43.5833    12.2873
## title_statusteacher                     0.9776        0.9776     0.1483
##                                   Mean Diff eQQ Med eQQ Mean eQQ Max
## distance                             0.0001   2e-04   0.0003  0.0015
## GenderF                              0.0256   0e+00   0.0256  1.0000
## GenderM                             -0.0256   0e+00   0.0256  1.0000
## Caucasian_binary1                   -0.0032   0e+00   0.0032  1.0000
## Strata5 to less than 10 years       -0.0096   0e+00   0.0096  1.0000
## Strata10 to 15 years                 0.0032   0e+00   0.0032  1.0000
## Stratagreater than or equal to 15    0.0032   0e+00   0.0032  1.0000
## employeeage                          0.8173   1e+00   1.2212  4.0000
## title_statusteacher                  0.0000   0e+00   0.0000  0.0000
##
## Percent Balance Improvement:
##                                   Mean Diff. eQQ Med eQQ Mean  eQQ Max
## distance                             97.4034  95.468  92.7165  84.9578
## GenderF                             -23.7689   0.000 -14.2857   0.0000
## GenderM                             -23.7689   0.000 -14.2857   0.0000
## Caucasian_binary1                    96.1133   0.000  96.1538   0.0000
## Strata5 to less than 10 years        87.5292   0.000  87.5000   0.0000
## Strata10 to 15 years                 96.2827   0.000  96.2963   0.0000
## Stratagreater than or equal to 15    95.4306   0.000  95.4545   0.0000
## employeeage                          11.1589   0.000 -26.5781  71.4286
## title_statusteacher                 100.0000   0.000 100.0000 100.0000
##
## Sample sizes:
##           Control Treated
## All         15395     312
## Matched       312     312
## Unmatched   15083       0
## Discarded       0       0
```

```
###plotting these
#plot(nearest)
nearest_matched <- match.data(nearest)
#now perform the statistical analysis
nearest_matched$Participated <- as.factor(nearest_matched$Participated)
model <- glm(Retained ~ Participated, data = nearest_matched,
             family = binomial)
summary(model)
```

```
##
## Call:
## glm(formula = Retained ~ Participated, family = binomial, data = nearest_matched)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5628  -1.5219   0.8359   0.8684   0.8684
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.78101    0.12197   6.403 1.52e-10 ***
## Participated1  0.09083    0.17405   0.522    0.602
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 767.04  on 623  degrees of freedom
## Residual deviance: 766.77  on 622  degrees of freedom
## AIC: 770.77
##
## Number of Fisher Scoring iterations: 4
```

Assocation of Ethnicity and Strata

```
ethnicity_strata <- analysis_data %>%
  group_by(Ethnicity, Strata) %>%
  summarise(n =n())
```

```
## `summarise()` regrouping output by 'Ethnicity' (override with `.groups` argument)
```

```
ethnicity_strata_spread <- spread(ethnicity_strata, key = "Strata", value = "n")
class(ethnicity_strata_spread$`less than 5`)
```

```
## [1] "integer"
```

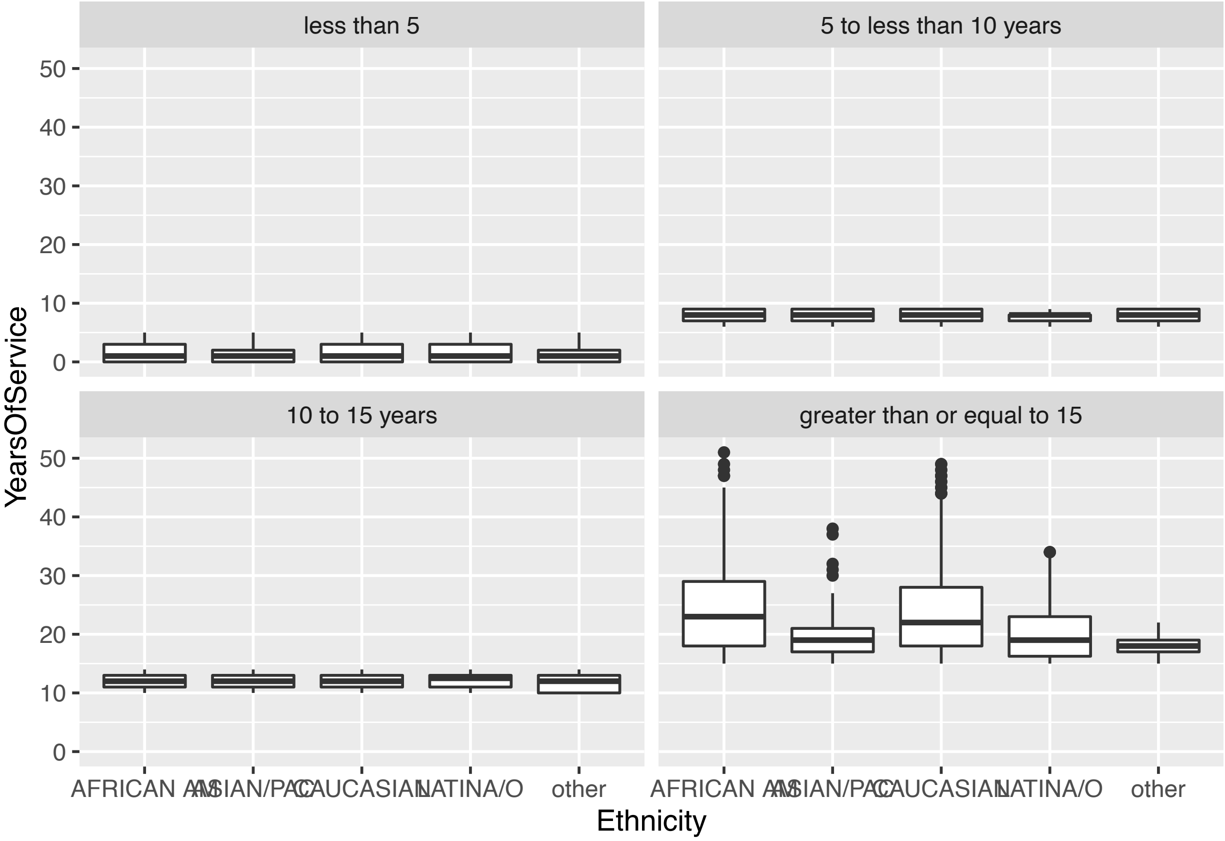
```
ethnicity_strata_spread <- ethnicity_strata_spread %>%
  mutate(total = `less than 5` + `5 to less than 10 years`+
          `10 to 15 years` +`greater than or equal to 15`) %>%
  mutate(percent_less_than5= `less than 5`/total,
         percent_5_to_10 = `5 to less than 10 years`/total,
         percent_10_15 = `10 to 15 years`/total,
         percent_greater_15 = `greater than or equal to 15`/total)
ethnicity_strata_spread <- ethnicity_strata_spread %>%
  select(Ethnicity, percent_less_than5, percent_5_to_10,
         percent_10_15, percent_greater_15)
library(scales)
```

```
##
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':
##
##     discard

## The following object is masked from 'package:readr':
##
##     col_factor
```

```
ethnicity_strata_spread$percent_less_than5 <- percent(ethnicity_strata_spread$percent_less_than5)
ethnicity_strata_spread$percent_5_to_10 <- percent(ethnicity_strata_spread$percent_5_to_10)
ethnicity_strata_spread$percent_10_15 <- percent(ethnicity_strata_spread$percent_10_15)
ethnicity_strata_spread$percent_greater_15 <- percent(ethnicity_strata_spread$percent_greater_15)
ethnicity_strata_spread
```

```
## # A tibble: 5 x 5
## # Groups:   Ethnicity [5]
##   Ethnicity  percent_less_than5 percent_5_to_10 percent_10_15 percent_greater_15
##   <fct>      <chr>              <chr>           <chr>         <chr>
## 1 AFRICAN AM 24.9%              11.96%          15.346%       47.78%
## 2 ASIAN/PAC  39.4%              13.21%          21.294%       26.15%
## 3 CAUCASIAN  37.7%              15.68%          14.711%       31.95%
## 4 LATINA/O   40.9%              12.45%          15.415%       31.23%
## 5 other      64.5%              15.54%          10.757%       9.16%
```

```
ethnicity_strata <- analysis_data %>%
  group_by(Ethnicity, Strata)
ggplot(ethnicity_strata, aes(x = Ethnicity, y = YearsOfService)) +
  geom_boxplot() +
  facet_wrap(~Strata)
```



Assocation of Ethnicity and Age

```
ethnicity_age <- analysis_data %>%
  group_by(Ethnicity, Age) %>%
  summarise(n =n())
```

```
## `summarise()` regrouping output by 'Ethnicity' (override with `.groups` argument)
```

```
ethnicity_age_spread <- spread(ethnicity_age, key = "Age", value = "n")
class(ethnicity_age_spread$`less than 5`)
```

```
## Warning: Unknown or uninitialised column: `less than 5`.

## [1] "NULL"
```

```
ethnicity_age_spread <- ethnicity_age_spread %>%
  mutate(total = `20s` + `30s`+
           `40s` +`50s` +`60s`+`70s`) %>%
  mutate(percent_20s= `20s`/total,
         percent_30s = `30s`/total,
         percent_40s = `40s`/total,
         percent_50s = `50s`/total,
         percent_60s = `60s`/total,
         percent_70s = `70s`/total)
ethnicity_strata_spread <- ethnicity_age_spread %>%
  select(percent_20s, percent_30s, percent_40s, percent_50s,
         percent_60s, percent_70s)
```

```
## Adding missing grouping variables: `Ethnicity`
```

```
library(scales)
ethnicity_age_spread$percent_20s <- percent(ethnicity_age_spread$percent_20s)
ethnicity_age_spread$percent_30s <- percent(ethnicity_age_spread$percent_30s)
ethnicity_age_spread$percent_40s <- percent(ethnicity_age_spread$percent_40s)
ethnicity_age_spread$percent_50s <- percent(ethnicity_age_spread$percent_50s)
ethnicity_age_spread$percent_60s <- percent(ethnicity_age_spread$percent_60s)
ethnicity_age_spread$percent_70s <- percent(ethnicity_age_spread$percent_70s)
ethnicity_age_spread
```

```
## # A tibble: 5 x 14
## # Groups:   Ethnicity [5]
##   Ethnicity `20s` `30s` `40s` `50s` `60s` `70s` total percent_20s percent_30s
##   <fct>     <int> <int> <int> <int> <int> <int> <int> <chr>       <chr>
## 1 AFRICAN ~   300   733   878  1235   798    18  3962 7.6%        18.50%
## 2 ASIAN/PAC    93    95    73    62    44     4   371 25.1%       25.61%
## 3 CAUCASIAN  2372  2895  2032  1883  1689    46 10917 21.7%       26.52%
## 4 LATINA/O     82   123   154    88    58     1   506 16.2%       24.31%
## 5 other        74    83    49    25    18     2   251 29.5%       33.07%
## # ... with 4 more variables: percent_40s <chr>, percent_50s <chr>,
## #   percent_60s <chr>, percent_70s <chr>
```

Assocation of nonWhite and Age

```
caucasian_binary_age <- analysis_data %>%
  group_by(Caucasian_binary, Age) %>%
  summarise(n =n())
```

```
## `summarise()` regrouping output by 'Caucasian_binary' (override with `.groups` argument)
```

```
caucasian_binary_age_spread <- spread(caucasian_binary_age, key = "Age", value = "n")
caucasian_binary_age_spread
```

```
## # A tibble: 2 x 7
## # Groups:   Caucasian_binary [2]
##   Caucasian_binary `20s` `30s` `40s` `50s` `60s` `70s`
##   <fct>            <int> <int> <int> <int> <int> <int>
## 1 0                  549  1034  1154  1410   918    25
## 2 1                 2372  2895  2032  1883  1689    46
```