Illinois Public School Teacher Salaries

A Deep Dive into a Decade of Data

Lukasz Filipek, Ki Young Han, Li Huang, Robin Wilcox
DePaul, CSC672

ABSTRACT

We take a deep dive into 10 years of Illinois public school system data, spanning years 2003 to 2012. We built 7 models to predict teacher salary within this dataset, and achieved the best results using XGBoost. We are able to predict salary within 6,700 RMSE. We also looked at salary in terms of fairness using various features like gender, ethnicity, position, location and highest degree. Like Simpson's paradox, while examining the fairness of teacher's salary, we found different aggregation methods applied to this dataset significantly change the underlying story. We present the underlying stories we discovered, and details of our predictive models, in this paper.

1. INTRODUCTION

Salary across genders appears to be a topic of great interest and controversy recently. We see examples such as Google analyzing salaries paying out $9.7million, to even the playing field, of additional compensation to 10,677 employees where a higher percentage of the money went to men[4]. If Google found something interesting in their dataset of candidates, it is compelling to analyze other datasets containing salary information for inequity. What makes such an analysis interesting, yet challenging, is that inequity itself is not necessarily anything meaningful, and that is why an analysis has to take into consideration any variable available to make sure one is comparing like populations and isolating only one variable, in this case, gender and ethnicity.

2. DATA PREPROCESSING

Original Dataset

We have been provided a collection of public school employment records from the state of Illinois, years 2003-2012 (see Figure 1) that serve as our dataset. The dataset was provided as ten separate csv files, one for each year. The dataset contains 61 variables; 9 of them are continuous variables, the rest a combination of nominal and ordinal variables. The dataset contains many paired redundant features where one feature has a numerical code while the other a description of that coded numerical variable. For example, *high_degre_cd* is a coded equivalent to *high_degre_desc*. See Figure 1.

Column	Definition
fy	Fiscal Year (school year)
rcdt	Region-County-District-Type Code
dst_name	District Name
dst_addr	District Street Address
dst_city	District City
dst_st	District State
dst_zip	District ZIP Code
dst_zip_plus4	District ZIP Code plus 4 digits
sch_num	School Code
sch_name	School Name
sch_addr	School Street Address
sch_city	School City
sch_st	School State
sch_zip	School ZIP Code
sch_zip_plus4	School ZIP Code plus 4 digits
last_name	Educator Last Name
first_name	Educator First Name
middle_init	Educator Middle Initial
gender	Gender
race_ethnicity_cd	Race/Ethnicity Code
race_ethnicity_desc	Race/ethnicity description
tsr_status_cd	TSR Status Code
tsr_status_desc	TSR status description
location_cd	Location Code
location_desc	Location description
emply_type	Employment Type
emply_desc	Employment type description
salary	Salary (total creditable earnings)
months_employed	Months Employed
pct_emp	Percent Time Employed
fte	Full-Time Equivalent
dist_exp	District Years Experience
state_exp	IL Years Experience
out_of_state_exp	Out-of-State Years Experience
bacc_coll	Baccalaureate College Code
bacc_coll_desc	Baccalaureate college description

Column	Definition
high_degre_cd	Highest College Degree Code
high_degre_desc	Highest college degree description
adv_coll	Advanced College Code
adv_coll_desc	Advanced college description
pos_cd	Position Code
pos_desc	Position description
low_grade	Lowest Grade Taught Code
low_grd_desc	Lowest grade taught description
high_grade	Highest Grade Taught Code
high_grd_desc	Highest grade taught description
assignment_1	Main Assignment Code
assign1_desc	Main assignment description
assignment_2	Other Assignment 1 Code
assign2_desc	Other assignment 1 description
assignment_3	Other Assignment 2 Code
assign3_desc	Other assignment 2 description
assignment_4	Other Assignment 3 Code
assign4_desc	Other assignment 3 description
assignment_5	Other Assignment 4 Code
assign5_desc	Other assignment 4 description
assignment_6	Other Assignment 5 Code
assign6_desc	Other assignment 5 description
assignment_7	Other Assignment 6 Code
assign7_desc	Other assignment 6 description
pct_admin	Percent Administration

FIGURE 1 - ORIGINAL DATASET DEFINITION

Data Splitting

Before modeling, we begin by splitting the dataset 80/20 training/ testing. Stratification across years is unnecessary as records are fairly uniformly distributed across the years.

The 20% is put aside as a final test set, splitting with a random seed (1234). The other 80% of the data is used to build models and tune hyperparameters through a 5-fold grid search. We evaluate our final model on the testing data using the metrics such as RMSE and MAE.

Data Reduction

In order to focus specifically on teachers, we reduced the original dataset by eliminating all non-teaching positions (retaining only *pos_cd* = 18,19,20,22). We further reduced the dataset by eliminating teacher names, school and district names and addresses. We did, however, retain the first 3 digits of the school zip codes, which effectively created school location clusters.

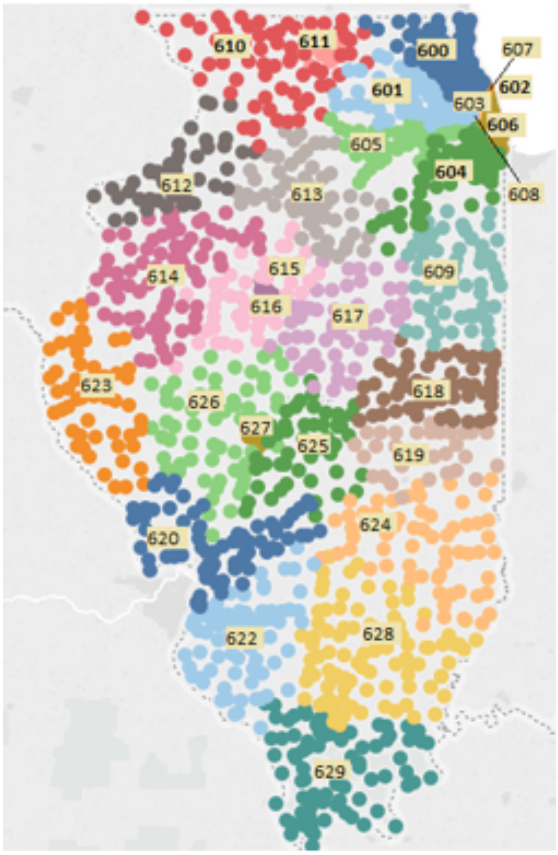

FIGURE 2 - SCHOOL LOCATION CLUSTERS USING 3 DIGITS OF ZIP CODE

We replaced advanced college description with a binary to indicate whether or not the teacher pursued a post bachelor education. We added *state experience* with *out of state experience* and substituted those two fields with a single field *years experience*. We added a new field, *assignment count*, to indicate the number of assignments listed in fields assign1_desc through assign7_desc. We grouped the assignments listed in assign1_desc through assign7_desc (originally over 260 unique assignments) to 21 definitions, and dropped the original fields.

Data Joined

We obtained demographic data by county (ideally, it would by zip, but we were unable to obtain this), which we then mapped to our school zip code clusters. The additional data we obtained included: Per Capita Income, Population Per Square Mile, Percent of Population Age 25+ with a HS Diploma, and Median Household Income. Our final dataset contained 25 features.

Data Modification

Since our dataset spanned ten years, the target variable, salary, is sensitive to inflation. Note, $100.00 in 2003 has the same value as $124.78 in 2012. By using the inflation rate associated with each year, we adjusted salary[2]. The boxplot figures below show the raw data(figure 3A) vs inflation adjusted salary (figure 3B).

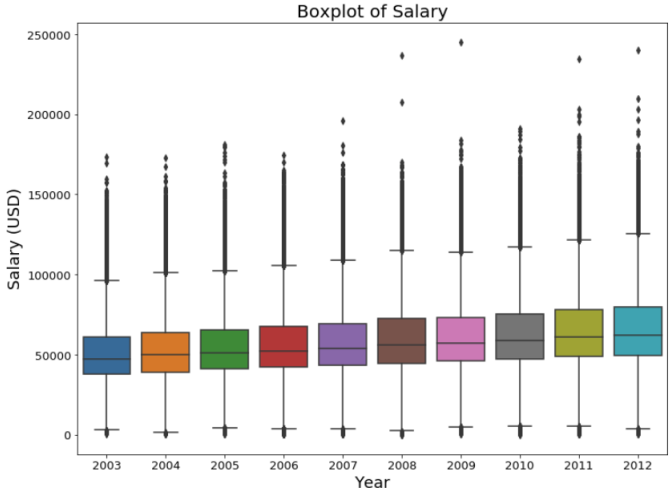

FIGURE 3A - ORIGINAL/RAW SALARY BY YEAR

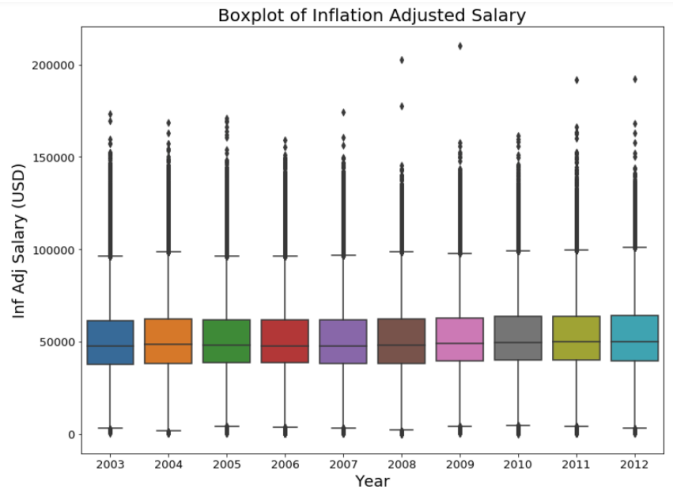

FIGURE 3B - INFLATION ADJUSTED SALARY BY YEAR

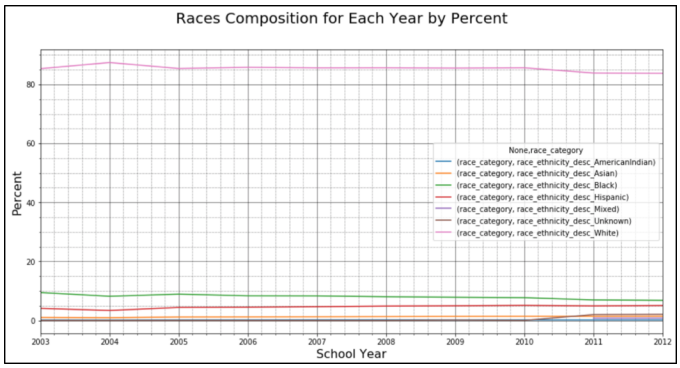

FIGURE 5 - RACE COMPOSITION BY PERCENT FROM 2003 TO 2012

The following graph, see Figure 6, propelled further analysis into race and salary.

3. EXPLORATORY ANALYSIS

Analysis of Distribution of Salary

Our target feature Salary have right skewed distribribution as shown in below figure 4. The salary has mean of $51,670.77, median value of $48,551.41, standard deviation of $17,601.4 and range of $210,298.5.

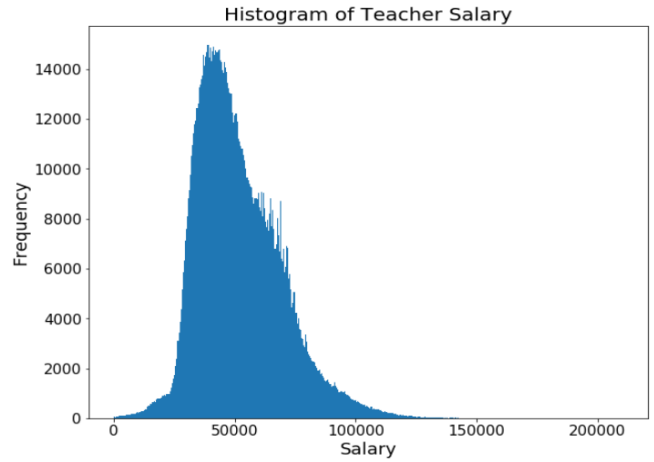

FIGURE 4 - TEACHER SALARY BY RACE AND GENDER

Analysis of Distribution by Race

Race white is consistently the majority of school teachers in our dataset, each year. The other races are far below, see Figure 5.

Race White, Black and Hispanic races comprise 98% of all teachers with domination over the years.

Race composition distribution over years

FIGURE 6 - TEACHER SALARY BY RACE AND GENDER

We see that Blacks appear to have a very different relationship between genders when it comes to salary. Further analysis is done by taking the total population, filtering it out to get a group of people as similar to each other as possible. This means, filtering for full-time employees that were employed for 9 months and were elementary school teachers, the biggest group of teachers in the dataset, and then calculating the mean of 100 individuals for 100 samples to see whether we can infer anything about the populations.

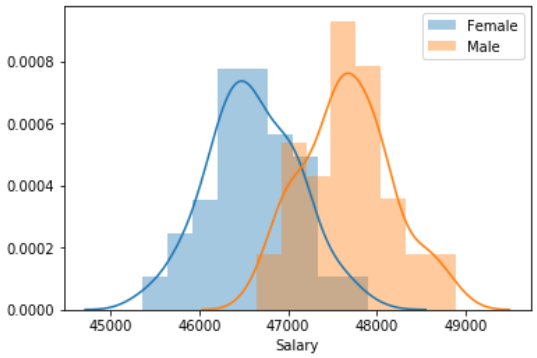

FIGURE 7A - TEACHER SALARY BY WHITE RACE AND GENDER

We see a statistically significant difference of ~1000 with a p-value < 1E-23 for race white.

We also see a statistically significant difference with a p-value < 1E-30 for race black. However, the relationship is flipped between men and women and the difference is ~2500.

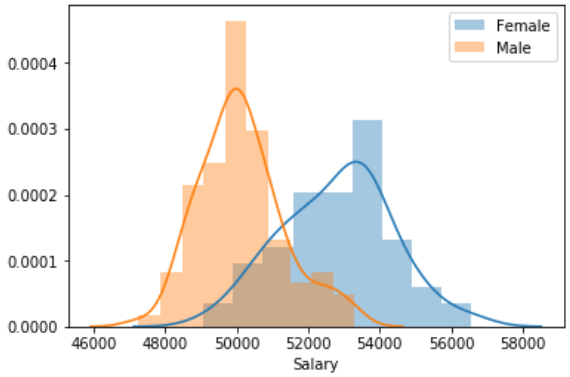

FIGURE 7B - TEACHER SALARY BY BLACK RACE AND GENDER

Details show that this relationship is not constant for race white across time. For the most recent year in the dataset, 2012, we see no statistically significant difference between the male and female populations. See Figure 7C.

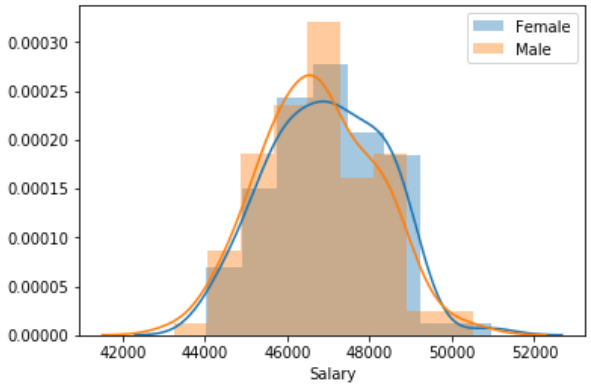

FIGURE 7C - TEACHER SALARY BY WHITE RACE AND GENDER, YEAR 2012

The black race, however, maintains a consistent relationship across all years, with the only variable being the degree of difference between means. In 2012, we see a difference of ~5000. See Figure 7D.

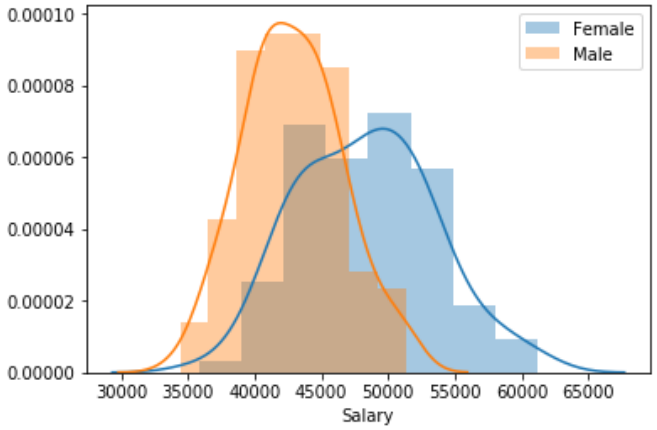

FIGURE 7D - TEACHER SALARY BY BLACK RACE AND GENDER, YEAR 2012

Analysis of Distribution of salary by Gender

To explore the salary distribution by gender, the salary was binned by 20 equal width as well as 20 equal frequency. For equal width, each bin has size of $10,514 salary. For equal frequency, each bin has size of 63,831 instances. Upon examining the male percentage in 20 equal width binned salaries, we saw the male percentage exponentially increases as the number of bins increases. We also noticed the exponential increases start from salaries greater than 84,000. Even after we binned with equal frequency, we observed that male percentage slowly increases as bin number increases. The yellow highlight on each top and bottom plot on figure 8 indicates similar salary range from $30,000 ~ $84,000.

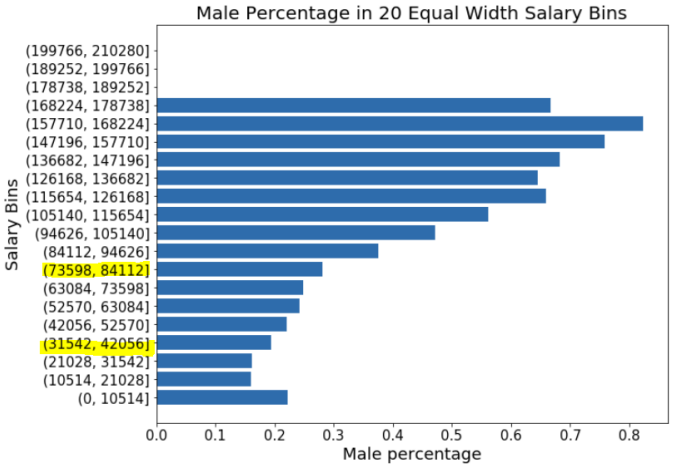

FIGURE 8A - MALE PERCENTAGE ON BINNED SALARY BY 20 EQUAL WIDTH BINNING

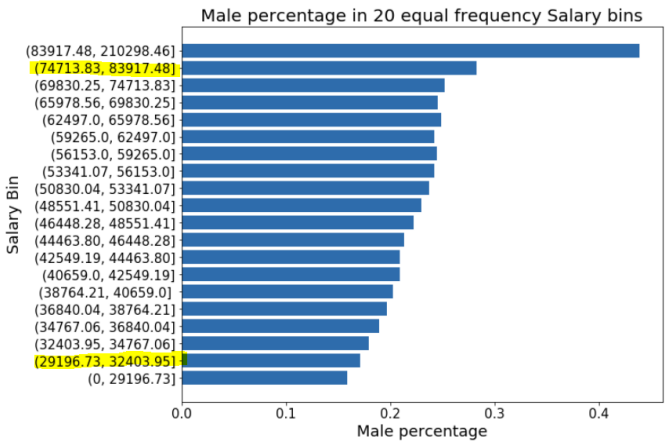

FIGURE 8B - MALE PERCENTAGE ON BINNED SALARY BY 20 EQUAL FREQUENCY BINNING

Bin Range	Frequency	Bin range	Frequency
(0, 10514]	2994	(-0.001, 29196.728]	63831
(10514, 21028]	14088	(29196.728, 32403.945]	63831
(21028, 31542]	90484	(32403.945, 34767.06]	63826
(31542, 42056]	322424	(34767.06, 36840.038]	63829
(42056, 52570]	316870	(36840.038, 38764.208]	63829
(52570, 63084]	221071	(38764.208, 40659.0]	63837
(63084, 73598]	169319	(40659.0, 42549.192]	63822
(73598, 84112]	76350	(42549.192, 44463.804]	63830
(84112, 94626]	34463	(44463.804, 46448.276]	63826
(94626, 105140]	17510	(46448.276, 48551.406]	63829
(105140, 115654]	7373	(48551.406, 50830.036]	63829
(115654, 126168]	2524	(50830.036, 53341.069]	63828
(126168, 136682]	774	(53341.069, 56153.0]	63830
(136682, 147196]	246	(56153.0, 59265.0]	63841
(147196, 157710]	58	(59265.0, 62497.0]	63817
(157710, 168224]	17	(62497.0, 65978.559]	63829
(168224, 178738]	9	(65978.559, 69830.246]	63829
(178738, 189252]	0	(69830.246, 74713.83]	63828
(189252, 199766]	2	(74713.83, 83917.482]	63829
(199766, 210280]	1	(83917.482, 210298.456]	63829

FIGURE 8C - SALARY RANGE AND INSTANCE COUNT FOR EQUAL WIDTH BINNING(LEFT) AND EQUAL FREQUENCY BINNING(RIGHT)

Impact of Recession

To discover how the 2008 recession has impacted teachers' salaries, we looked in two directions.

First, we looked at 2008 recession impact on teachers' annual average salary increase rate.

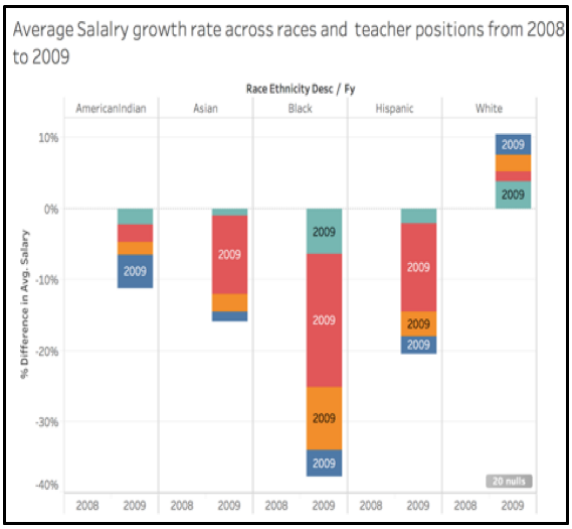

FIGURE 9A - ANNUAL AVERAGE SALARY INCREASE RATE ACROSS DIFFERENT RACES AND POSITION

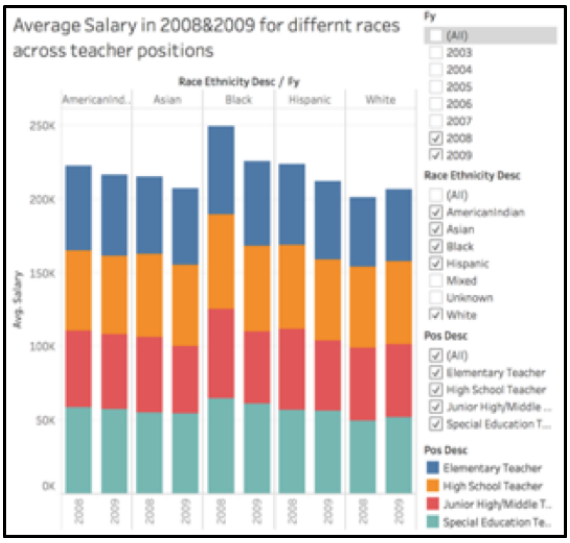

FIGURE 9B - INCREASE RATE FROM 2008 TO 2009 ACROSS RACES AND POSITIONS

From Figures 9A and 9B, we see race black has the biggest decrease of average salary in 2009, but still retains the highest average salary, as it has a higher base from previous years.

All races, except race White, experienced a decreased average salary in 2009. The decrease rate is highest in junior High teacher position for race Asian, Hispanic and Black.

Second, we looked at how the recession impacted teacher's salary in and out of Chicago.

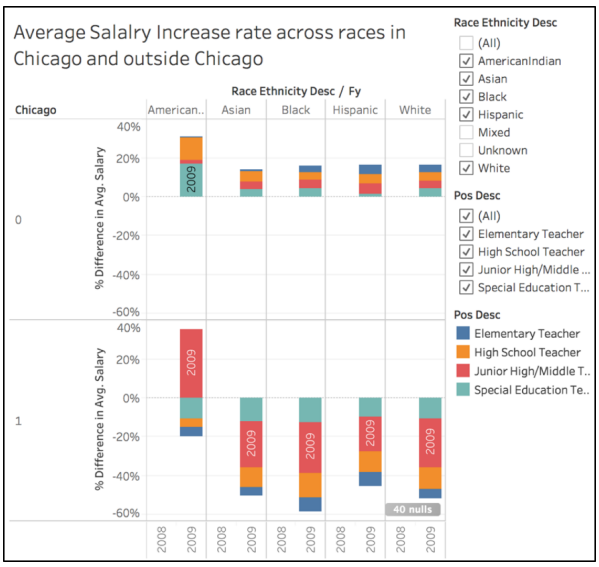

FIGURE 9C - AVERAGE SALARY FOR 2008 AND 2009 ACROSS RACES AND POSITIONS INSIDE & OUTSIDE CHICAGO

From figure 9C, we see in 2009, the average salary has all increased for teachers outside of Chicago, but decreased for teachers in Chicago.

Outside Chicago, American Indian race received the highest increase in 2009 for High school and special education teachers. In Chicago, all the teachers' salary decreased except American Indian in Junior High/Middle teachers;

Analysis of Salary by Gender

In an effort to compare gender as fairly as possible, we reduced the dataset to only those teachers reporting 100% employed, 100% full time, exactly 9 months.
We analyze each teaching position separately, in year 2003 (left side) and again in year 2012 (right side). Each male teacher is plotted in blue, and female in magenta. The leftmost column in each year contains teachers with a bachelor degree. The rightmost column in each year contains teachers with a graduate degree.

We verified the statistical significance of the data highlighted by randomly selecting 10,000 or 1,000 (when < 10,000 samples were available) males and females from the filtered dataset explained above, and performing a t-test on their average salaries. The dataset was randomly shuffled and repeated ten times.

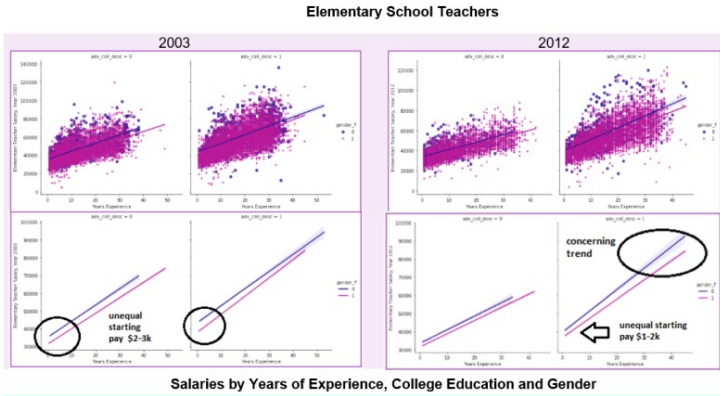

FIGURE 10A - ELEMENTARY SCHOOL TEACHERS, SALARY BY YEARS OF EXPERIENCE, SEPARATED BY ADVANCED EDUCATION AND GENDER

In 2003, we find male **Elementary School** teachers to earn a starting salary of $2-3k more than female elementary school teachers, for teachers with both graduate and bachelor degrees. This salary gap closes with years of experience for teachers with graduate degrees. The salary gap remains unchanged with years of experience for teachers with bachelor degrees. See figure 10A.
In 2012, male and female **Elementary School** teachers with a bachelor degree earn the same starting salary, and their average salaries remain the same with years of experience. Male elementary school teachers with a graduate degree earn a starting salary of $1-2k higher than females and this salary gap increases with years of experience. See Figure 10A.

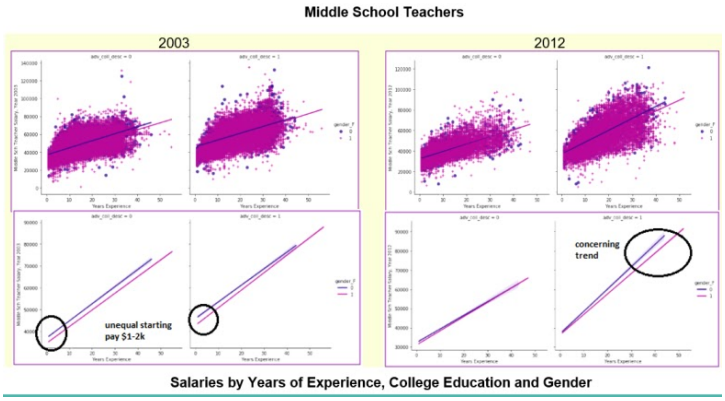

FIGURE 10B - MIDDLE SCHOOL TEACHERS, SALARY BY YEARS OF EXPERIENCE, SEPARATED BY ADVANCED EDUCATION AND GENDER

In 2003, we find male **Middle School** teachers to earn a starting salary of $1-2k more than female elementary school teachers, for teachers with both graduate and bachelor degrees. This salary gap closes with years of

experience for teachers with graduate degrees. The salary gap remains unchanged with years of experience for teachers with bachelor degrees. See figure 10B.
In 2012, male and female **Middle School** teachers with bachelor and graduate degrees earn the same starting salary. Average salaries remain the same for teachers with bachelor degrees. A salary gap begins, with males earning more than females, as years of experience increase, for teachers with graduate degrees. See Figure 10B

.

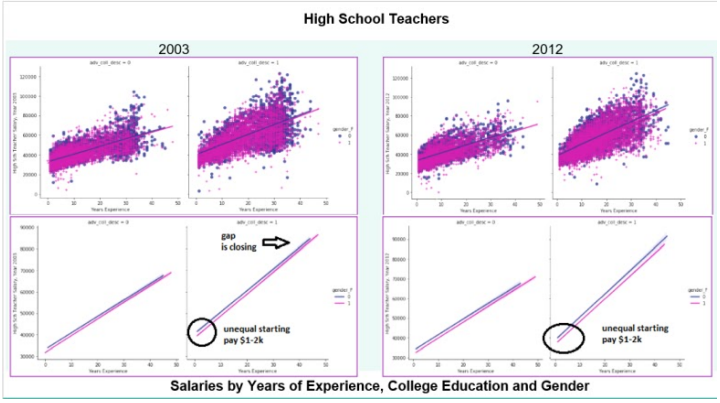

FIGURE 10C - HIGH SCHOOL TEACHERS, SALARY BY YEARS OF EXPERIENCE, SEPARATED BY ADVANCED EDUCATION AND GENDER

In 2003 and 2012, male and female **High School** teachers with bachelor degrees earn the same average salary, and average salaries remain the same with increased years of experience. See Figure 10C.
In 2003 and 2012, male **High School** teachers with graduate degrees earn a starting salary $1-2k higher than females. In 2003, this salary gap closes with increased years of experience. In 2012, this salary gap remains constant with years of experience. See Figure 10C.

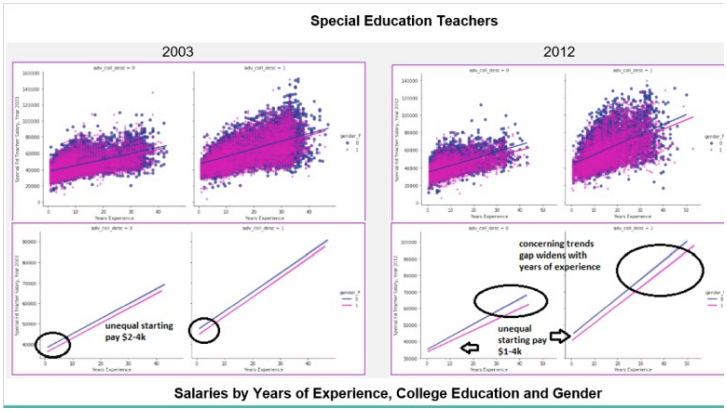

FIGURE 10D - SPECIAL EDUCATION SCHOOL TEACHERS, SALARY BY YEARS OF EXPERIENCE, SEPARATED BY ADVANCED EDUCATION AND GENDER

In 2003, we find male **Special Education School** teachers to earn a starting salary of $2-4k more than female elementary school teachers, for teachers with both graduate and bachelor degrees. This salary gap remains constant with increased years of experience. See Figure 10D.
In 2012, male **Special Education School** teachers with bachelor and graduate degrees earn a starting salary of $1-4k more than female special education teachers. This salary gap widens with increased years of experience. The gap widens at a faster rate for special education teachers with bachelor degrees. See Figure 10D.

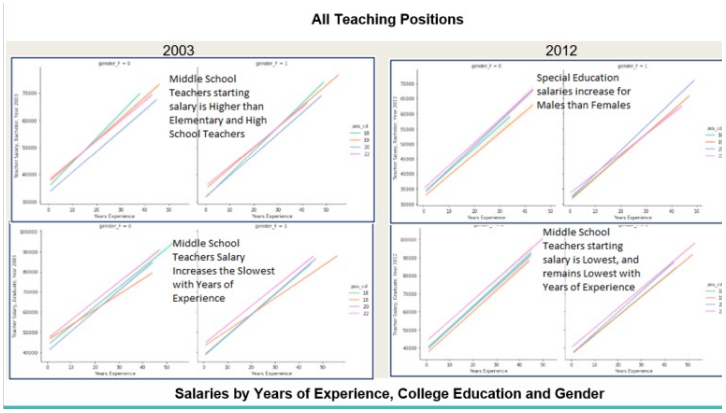

FIGURE 11 - SALARY VS. YEARS OF EXPERIENCE. ELEMENTARY TEACHERS IN GREEN, MIDDLE SCHOOL IN ORANGE, HIGH SCHOOL BLUE, SPECIAL ED IN RED.. 2003 ON LEFT, 2012 ON RIGHT. NO GRADUATE DEGREE (LEFTMOST COLUMN) AND GRADUATE DEGREE (RIGHTMOST COLUMN)

We also group all the positions together in one view to compare across positions. In 2003, we find **Middle School** teachers starting salary is higher than Elementary School teachers and High School Teachers, but becomes the lowest salary with increased years of experience. In 2012, **Middle School** teachers start with the lowest salary across the positions, and remain the lowest salary with years of experience.

4. PREDICTIVE MODELING

Model Exploration

K-means clustering application

In our project, clustering algorithm K-modes was applied on all the independent features: numeric and categorical, and the clustering result is added as a new feature, used together with the original features to train the random forest model and make prediction. This method has improved our model, but just lightly.

From a research paper[1] in a similar salary prediction problem, we got inspiration to use k-means clustering on the numeric variables and use the result as a new feature, then train the model, this method has improved our model more than k-modes clustering. The clustering is performed on all the 12 numeric variables in our dataset, and the k selected is 6.

Grouping & Assess Salary related features

By clustering on the 7 significant numeric features that are correlated with salary, we divided the data into 5 groups, And the groups are separated by different combination of the important features. For example, salary is highly correlated with combination of Population and Experience: salary is highest for Teacher with more experience in a populated area, and lowest for Teacher with less experience in a less populated area.

Hybrid Model

As our data is a combination of numeric and categorical variables. A hybrid model combining a decision tree with another regression algorithm is applied to analyze mixed data[2]. In the proposed model, the portions explained by all the categorical variables of the dataset are estimated by random forest and the remaining parts are predicted by linear regression algorithm trained by all the numerical variables.

The hybrid model consists of two models, after building the first model and calculating the difference of target variable and the prediction from first model, then the hybrid Model is used to predict.

Since we cannot perform cross validation on hybrid model, we divided the data into train-test(80/20). After building a hybrid model on training data, we make prediction on the test data, which is added value of the results of two
models, one for categorical and one for numeric. The RMSE & MAE are as following:

Hybrid Model	RMSE	MAE
training data	9346	6960
testing data	10289	7626

As we can see, the hybrid model did not improve our prediction with a higher RMSE and MAE on training and testing. In summary, the proposed algorithm does not require a coding system or dissimilarity/similarity functions for mixed data, but it does not consider the interaction between the categorical and numeric variables. This could be the reason why the algorithm is not improving our prediction.

Random Forest and XGB

Tree based ensemble algorithms are fantastic because they do not require much feature preprocessing and achieve great results relatively quickly. This is why no analysis would be complete without a random forest and an xgboost comparison.

Decision Tree, Bagging Tree, AdaBoost

These are tree based algorithms that most frequently used to compare the model performances. The models were tested on various configuration of hyperparameters then we came up with best performing configuration.

Ridge Regression

The predictive outcome of slices of data using Ridge Regression highlights specifically where variation exists. By using this model, we found we can better predict Female Teaching salaries than male Teaching salaries. We can better predict teachers with a Bachelor degree than teachers with a Graduate degree. See Figure 12A. We can better predict salaries within regions of similar wealth and

population than regions with variation of wealth and population. See Figure 12B.

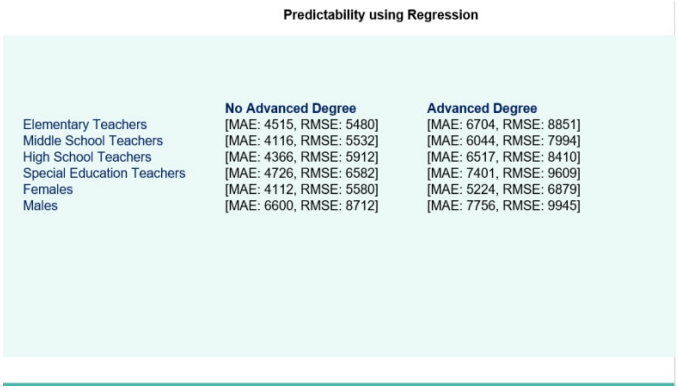

FIGURE 12A - MAE AND RMSE FOR SLICES OF DATA USING RIDGE REGRESSION

We have the most difficulty predicting males with Graduate degrees and Special Education teachers with Graduate degrees (Figure 12A) and regions inside and surrounding chicago (Figure 12B.

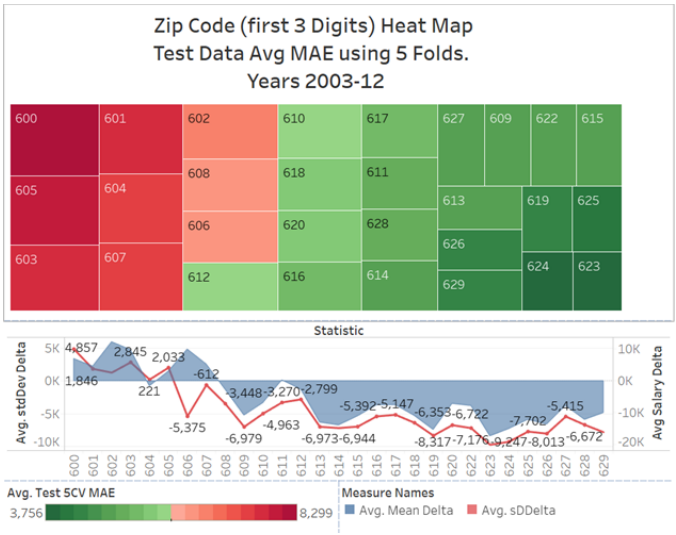

FIGURE 12B - MAE USING RIDGE REGRESSION, AND SALARY STD DEV BY LOCATION

5. VALIDATION and TESTING

Approach

We used 5-fold cross validation on the train set split for tuning of the models and choosing the best results based on the distribution of scores generated from the test splits in cross validation. We then do a final sanity check on the test dataset split to make sure that our model is indeed generalizing well.

6. RESULTS

Optimal Parameters for each Models

Optimal configurations for each model are shown as following:

- Decision Tree - min_samples_split: 5, min_samples_leaf: 10, max_leaf_nodes: 2000
- Bagging Tree - N estimators: 10, max_samples: 1.0, max_feature: 1.0, bootstrap: True, bootstrap_features: False
- Adaboost - n estimators: 10, learning rate: 0.1, loss: exponential
- Ridge Regression - Alpha: 21
- Random Forest & K-means Clustering - N_estimators: 20, k: 6,
- Random Forest - N_estimators: 500, max depth: 10
- **XGBoost - N estimators: 1000, learning rate: 0.01, max depth: 10**

Comparison of Models

Model	R-sq	Train RMSE	Test RMSE	Run Time
Decision Tree	0.82	7374	7339	1.5 min
Bagging	0.83	7196	7166	4.2 min
AdaBoost	0.72	7257	7236	21.6 min
Ridge Regr	0.84	8997	9109	2 min
Random Forest	0.87	5032	6817	5 min
RF & K-mean	0.84	3258	6831	61 min
XGBoost	**0.89**	**4505**	**6710**	**15 min**

FIGURE 13 - COMPARISON OF MODELS

Best Models

XG boost performed best. RF & K-mean have lowest train RMSE but it was overfitting. The Test RMSE of RF & K-mean model was higher than XG boost and the run time took quadruple more than XG Boost's run time although the numbers should be taken with a grain of salt as these were run on different tiers of machines. The beauty though about XGBoost is that it is coded to take advantage of GPU's, therefore, if it was useful, we could have built a much bigger model in a fraction of the time of the other models.

We see that among the top 5 most important features for predictability as:

- Advanced college degree (binary)
- % Population with HS diploma
- Years of experience
- District experience (years)
- Highest grade taught

The topic of fairness also creeps into algorithms. How does a model perform for different subsections of a population in this particular case. Does this model for example predict just as effectively for males and for females.

We look particularly at ethnicity and gender; however, this type of analysis could be and we would argue, should be, applied to the output of any model that makes important decisions.

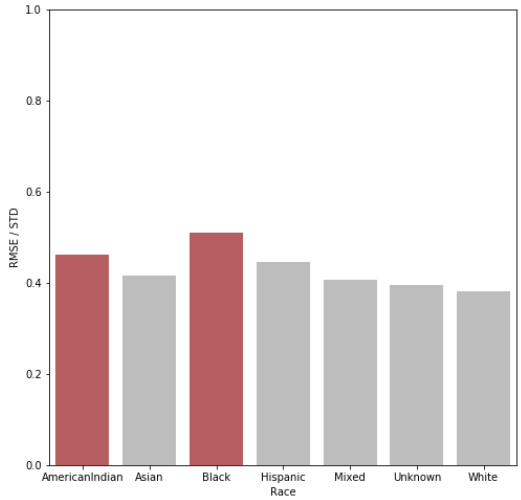

FIGURE 14A - RMSE/STD Ratio For Race/Ethnicity

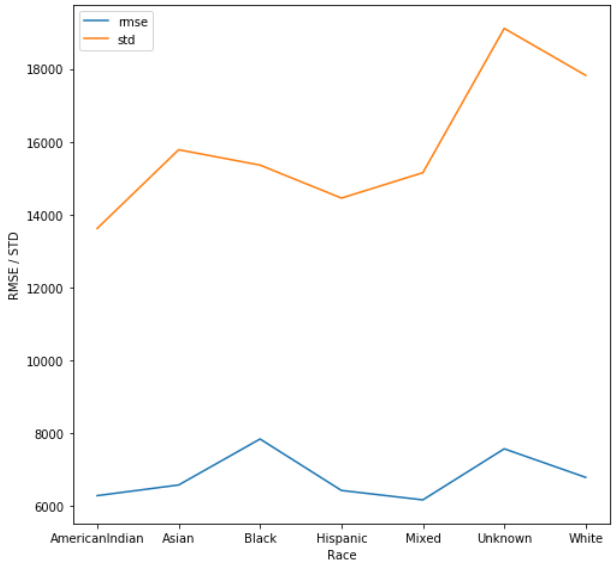

FIGURE 14B - RMSE and STD by Race

Looking at a metric such as RMSE alone is not enough. We have to take into consideration the variance for a particular ethnicity. We attempt to remedy this by taking a ratio of RMSE and the standard deviation of the salary. We see that gender Black and gender American Indian appear to have a bit higher ratio than the other ethnicities. This leads us to keep a note on these ethnicities to do further analysis of the predictive power of the model.

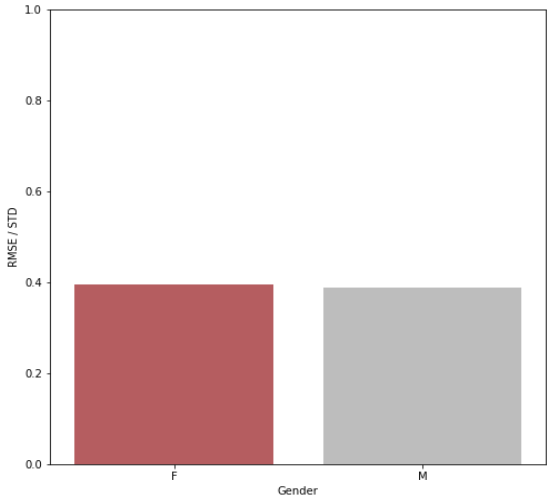

FIGURE 14C - RMSE/STD Ratio For Gender

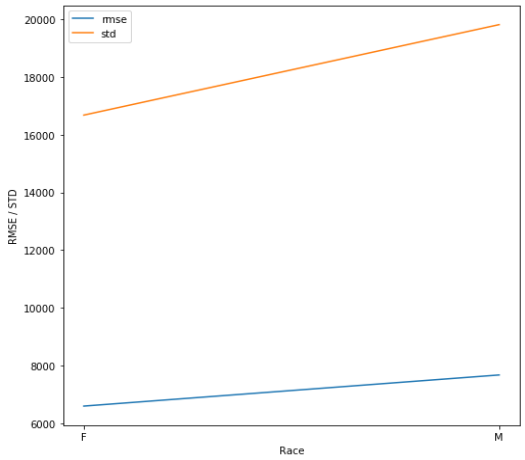

FIGURE 14D - RMSE and STD by Gender

We see that the RMSE is different for the two genders; however, variance is also different with the same relationship. The ratio shows us that there is not much a difference, therefore, we feel comfortable that the model is predicting well for both genders.

7. DISCUSSION

It may be worth noting, as mentioned previously, we did not use names of schools. We did, however, explore the use of them by grouping ino the following categories: elementary school, primary school, middle school, junior high school, high school, academy, center, grade school, intermediate school, magnet school. We also initially grouped assignments by applying a bag of words technique to group. Unfortunately, we did not find either of these were predictive for modeling purposes.

As we have demonstrated, a predictive model can be built for predicting salaries. Yet every model hit a stand still. What this is telling us is that there are not enough features, and frankly, we are surprised of the performance given only the handful of features that were useful for the model. For further analysis, it would be great to have more information about teachers' performance. It would also be great to demographic data specific to the school zip code, and have a key to link the data across years and within years. Fuzzy matching across first and last names, and the varying colleges attended, has shown evidence for further investigation. We saw that teachers had multiple jobs, in many cases with salaries there were quite broad. This would give us a story of total compensation and also it would show how much a particular teacher actually works in the field.

We also saw that it was more difficult to predict chicago proper salaries as compared to certain suburban and rural salaries. Searching for causes at this point would be speculation, but further analysis diving into specific zip codes would be interesting.

Beyond modeling, we found the exploratory work to tackle "fairness" the most interesting. There is not necessarily one story to take away from this analysis; however, there are definitely a few clear lessons to be learned. Relationships change over time. When looking at a dataset, one must be careful not to take a look at everything and draw conclusions without exploring the granularities as a relationship that may have been statistically significant at one point, doesn't necessarily hold in further years. Categorical features can expose interesting relationships. For example, like our example of race/ethnicity, gender and salary, one cannot look at gender or ethnicity alone, as these additional features, expose very interesting stories..

Finally, one must be very careful when doing these types of analyses. Diving into the details and looking at every cross section of a feature can potentially tell a cool story; however, this is also susceptible to p-value hacking and the issue of running many statistical tests and only reporting those findings that were found to be significant. The more tests one runs on a particular dataset, especially when running tests on different slices of one feature, the more likely that there is at least one test that shows up statistically significant out of pure chance.

8. CONCLUSION

In conclusion, we found that a model can be built to predict salary, although, the use of such a model in this particular case should be used with caution as a model gives the output of the training dataset, biases etc. included. Analysis of the data showed that there is a story to be told; however, different slices of the data tell different stories, and one has to be very careful into

drawing conclusions. Finally, more analysis in this problem should be had, as there was evidence to show more interesting findings may be discovered..

References

[1] Ignacio Martín, Andrea Mariello, Roberto Battiti, José Alberto Hernández, “Salary Prediction in the IT Job Market with Few High-Dimensional Samples: A Spanish Case Study”, International Journal of Computational Intelligence Systems, Volume 11, Issue 1, January 2018, Pages 1192 – 1209.

[2] “Inflation Rate between 2003-2013 | Inflation Calculator.” *2003 Dollars in 2013 | Inflation Calculator*, Alioth LLC, www.in2013dollars.com/2003-dollars-in-2013.

[3] K.Kim and J.S.Hong, “ A hybrid decision tree algorithm for mixed numeric and categorical data in regression analysis”, Pattern Recognition Letters, vol.98, pp. 39-45, 2017.

[4] Wakabayashi, Daisuke. “Google Finds It's Underpaying Many Men as It Addresses Wage Equity.” *The New York Times*, The New York Times, 4 Mar. 2019, www.nytimes.com/2019/03/04/technology/google-gender-pay-gap.html.

Appendix