Analysis of Salaries in San Francisco

by

Gaurav Kshirsagar & Ajinkya Ingle

Under the guidance of

Distinguished Prof. Amirhossein Gandomi
Stevens Institute of Technology
Class of Fall 2017

Analysis of Salaries in San Francisco

Abstract

We are proposing an analysis by classification and clustering of the citizens based on the salaries they are earning in San Francisco. One way to recognize how a city government works is by looking at who it employs and how its employees are compensated. This data contains the names, job title, and compensation for San Francisco city employees on an annual basis from 2011 to 2014. The positions studied span a wide range of duties and responsibilities. We have tried and tested different models of the dataset to classify the different jobs into some generalized categories. We have plotted some graphs from the dataset which gives us the basic idea about where to focus on in the entire dataset. By using classification techniques like KNN and Logistic regression we tried classifying the employees in male or female based on their income in various categories. Principle Component analysis was used to do dimension reduction on the dataset. By doing this analysis we have managed to highlight some trends in the salaries distributed amongst the employees of a metropolitan city. The focus of this analysis was to showcase this trends and stats for further study of appropriately compensating the employees of any metropolitan city.

Introduction

A new study by the finance site GOBankingRates discovers that of the 50 most populous cities in the country, San Francisco requires the most income to reach a comfort level determined by the 50-30-20 budgeting rule. Under the rule, 50 percent of income covers necessities, 30 percent covers discretionary items and 20 percent is for savings. If your income is sufficient to cover your cost-of-living expenses, you can live comfortably.

The study found that you would need to earn $110,357 to achieve that goal in San Francisco — more than New York, Honolulu or Washington, D.C. In fact, the three largest cities in the Bay Area ranked in the top four: San Jose was second ($87,153); New York third ($86,446) and Oakland fourth ($80,438).

Another study, by Attom Data Solutions in January, calculated that renters in the Marin County/San Francisco metro area will spend more than 77 percent of their salary, on average, to pay rent in 2017. The national average is 38.7 percent.

New York City joined Philadelphia and Massachusetts in passing legislation that will ban employers from asking job applicants about their salary history, in an attempt to narrow the wage gap between women and men. More than 20 other cities and states including San Francisco and California have similar legislation in the works. The goal is to prevent gender discrimination from being passed from one workplace to the next by basing an employee's pay on his or her prior salary.

These statistics inspired us to study on the actual data for getting into those intricate details and discovering some of the undiscovered facts from the dataset.

Analysis of Salaries in San Francisco

Problem Description

- We have a taken a database which consists of employee names, Job titles, base pay and various aspects related to them. It includes the overtime pay, benefits and date.
- These variables are further cleaned and categorized for further classification.
- Before starting the actual analysis, a lot of cleaning needs to be done on the data.
- The cleaning process includes removing the variables which are unnecessary, and removing the null values.
- There are also some missing values which needs to be replaced.
- Dimensional Reduction was performed to select the specific variables which showed high Eigen values.
- The job titles given in the data are too specific and varied which cannot be used to classify the data, so we have to categorize it further.
- Similarly, our analysis is mainly based on the Gender of the employees so we need to find the Gender of the employees. We have used the first names of the employees to find their Gender.
- Using the Job Title, Gender and salaries of the Employees we have applied KNN and Logistic Regression algorithms to them.
- In KNN the half of the data was used to train the system and half of the data was used to test the accuracy of the system.
- Among KNN and Logistic Regression the algorithm which shows the most accuracy will be chosen.

Evaluation of Database

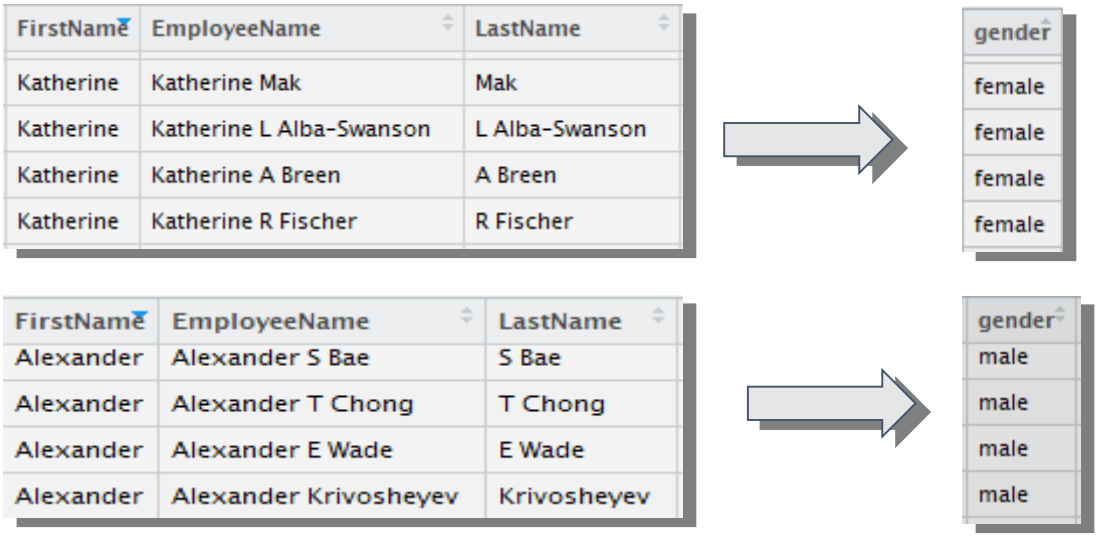
Id	EmployeeName	JobTitle	BasePay	Overtime	OtherPay	Benefits	TotalPay	TotalPayB	Year	Agency	Status
1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSI	167411.2	0	400184.3		567595.4	567595.4	2011	San Francisco	
2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966	245131.9	137811.4		538909.3	538909.3	2011	San Francisco	
3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.1	106088.2	16452.6		335279.9	335279.9	2011	San Francisco	
4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916	56120.71	198306.9		332343.6	332343.6	2011	San Francisco	
5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT,(FIRE DEPART	134401.6	9737	182234.6		326373.2	326373.2	2011	San Francisco	
6	DAVID SULLIVAN	ASSISTANT DEPUTY CHIEF II	118602	8601	189082.7		316285.7	316285.7	2011	San Francisco	
7	ALSON LEE	BATTALION CHIEF, (FIRE DEPARTMENT)	92492.01	89062.9	134426.1		315981.1	315981.1	2011	San Francisco	
8	DAVID KUSHNER	DEPUTY DIRECTOR OF INVESTMENTS	256577	0	51322.5		307899.5	307899.5	2011	San Francisco	
9	MICHAEL MORRIS	BATTALION CHIEF, (FIRE DEPARTMENT)	176932.6	86362.68	40132.23		303427.6	303427.6	2011	San Francisco	
10	JOANNE HAYES-WHIT	CHIEF OF DEPARTMENT, (FIRE DEPARTMENT)	285262	0	17115.73		302377.7	302377.7	2011	San Francisco	
11	ARTHUR KENNEY	ASSISTANT CHIEF OF DEPARTMENT, (FIRE DEP	194999.4	71344.88	33149.9		299494.2	299494.2	2011	San Francisco	
12	PATRICIA JACKSON	CAPTAIN III (POLICE DEPARTMENT)	99722	87082.62	110804.3		297608.9	297608.9	2011	San Francisco	
13	EDWARD HARRINGTO	EXECUTIVE CONTRACT EMPLOYEE	294580	0	0		294580	294580	2011	San Francisco	
14	JOHN MARTIN	DEPARTMENT HEAD V	271329	0	21342.59		292671.6	292671.6	2011	San Francisco	
15	DAVID FRANKLIN	BATTALION CHIEF, (FIRE DEPARTMENT)	174872.6	74050.3	37424.11		286347.1	286347.1	2011	San Francisco	
16	RICHARD CORRIEA	COMMANDER III, (POLICE DEPARTMENT)	198778	73478.2	13957.65		286213.9	286213.9	2011	San Francisco	
17	AMY HART	DEPARTMENT HEAD V	268604.6	0	16115.86		284720.4	284720.4	2011	San Francisco	
18	SEBASTIAN WONG	CAPTAIN, EMERGENCYCY MEDICAL SERVICES	140546.9	119397.3	18625.08		278569.2	278569.2	2011	San Francisco	
19	MARTY ROSS	BATTALION CHIEF, (FIRE DEPARTMENT)	168692.6	69626.12	38115.47		276434.2	276434.2	2011	San Francisco	
20	ELLEN MOFFATT	ASSISTANT MEDICAL EXAMINER	257510.6	880.16	16159.5		274550.3	274550.3	2011	San Francisco	
21	VENUS AZAR	ASSISTANT MEDICAL EXAMINER	257510.5	0	16679.79		274190.3	274190.3	2011	San Francisco	
22	JUDY MELINEK	ASSISTANT MEDICAL EXAMINER	257510.4	377.21	15883.56		273771.2	273771.2	2011	San Francisco	
23	GEORGE GARCIA	CAPTAIN, FIRE SUPPRESSION	140546.9	93200.58	39955.25		273702.7	273702.7	2011	San Francisco	
24	VICTOR WYRSCH	BATTALION CHIEF, (FIRE DEPARTMENT)	168692.6	77896.14	24083.86		270672.6	270672.6	2011	San Francisco	

- This is the sample of data which we are using. The data consists of 11 variables, from which we have used the variables which were the most important for our classification.
- The most important variables were the Employee name, Job Title and the Total Pay.

- We used the Employee name to find the Gender of the employees which was further used to classify the employees based on gender.
- We have used the Job Title and further categorized the it into more specific Jobs. Like all the Jobs in the Fire fighter and Museum guard were put in the Public Services.
- We have removed some of the columns in the data as it was not that important and the missing values.

Data Processing and Preparation

- In the data, the Gender of the Employees was not available, so to find the Gender of the Employees we have used the Gender package in R.
- In this the name of the Employee is split in First and Last name. Based on the First name the package assigns the Gender to the name.
- Like from the example provided we can see that the name Katherine Mak is split in two parts "Katherine" and "Mak". Here "Katherine" is the first name, from this name the Gender of the person is predicted and Katherine is assigned as female.
- Similarly, in the second example based on the name Alexander the gender is assigned as male.



FirstName	EmployeeName	LastName
Katherine	Katherine Mak	Mak
Katherine	Katherine L Alba-Swanson	L Alba-Swanson
Katherine	Katherine A Breen	A Breen
Katherine	Katherine R Fischer	R Fischer

gender
female
female
female
female

FirstName	EmployeeName	LastName
Alexander	Alexander S Bae	S Bae
Alexander	Alexander T Chong	T Chong
Alexander	Alexander E Wade	E Wade
Alexander	Alexander Krivosheyev	Krivosheyev

gender
male
male
male
male

- Here from the example we see that there are many different Job Titles which needs to be categorized further.
- There are Job Titles as Automotive Mechanic, IS Engineer-Senior, etc which are further categorized into STEM.
- Similarly, the other remaining Jobs are categorized into Healthcare, Police/Law and Security, Public Services and Transit/Transportation.
- This has to be done as it is difficult to classify the Employees with their current Job Titles which are too specific and varied.

Analysis of Salaries in San Francisco

EmployeeName	LastName	JobTitle
Aaron M Del Tredici	M Del Tredici	Physician Specialist
Aaron S Duran	S Duran	Automotive Mechanic
Aaron W Bjorkquist	W Bjorkquist	Police Officer
Aaron Stevenson	Stevenson	Lieutenant, Fire Suppression
Aaron P Smith	P Smith	IS Engineer-Senior
Aaron I Fisher	I Fisher	Firefighter
Aaron C Wilson	C Wilson	Public Service Trainee
Aaron Brinkerhoff	Brinkerhoff	Biologist III
Aaron A Hipolito	A Hipolito	Museum Guard
Aaron T Dunn	T Dunn	Transit Operator
Aaron C Ballonado	C Ballonado	Police Officer 3
Aaron Del Tredici	Del Tredici	Physician Specialist

Sector
HEALTHCARE
STEM
POLICE/LAW n SECURITY
PUBLIC SERVICES
STEM
PUBLIC SERVICES
PUBLIC SERVICES
STEM
PUBLIC SERVICES
TRANSIT/TRANSPORTATION
POLICE/LAW n SECURITY
HEALTHCARE

- For the analysis of the data we need the numeric values of the variables, so we have assigned numbers to the Job Titles.
- This was done using the grep function in R
- For example, the STEM job title was assigned number 3 and similarly Healthcare was assigned number 2 and so on.

```
#Assigning Numerical values to sectors
salaries_g2[grep("POLICE/LAW n SECURITY",salaries_g2$Sector, ignore.case = TRUE), "SectorN"] <- "1"
salaries_g2[grep("HEALTHCARE",salaries_g2$Sector, ignore.case = TRUE), "SectorN"] <- "2"
salaries_g2[grep("STEM",salaries_g2$Sector, ignore.case = TRUE), "SectorN"] <- "3"
salaries_g2[grep("EDUCATION",salaries_g2$Sector, ignore.case = TRUE), "SectorN"] <- "4"
salaries_g2[grep("TRANSIT/TRANSPORTATION",salaries_g2$Sector, ignore.case = TRUE), "SectorN"] <- "5"
salaries_g2[grep("RETAIL",salaries_g2$Sector, ignore.case = TRUE), "SectorN"] <- "6"
salaries_g2[grep("REAL ESTATE",salaries_g2$Sector, ignore.case = TRUE), "SectorN"] <- "7"
salaries_g2[grep("SERVICES",salaries_g2$Sector, ignore.case = TRUE), "SectorN"] <- "8"
salaries_g2[grep("ENERGY",salaries_g2$Sector, ignore.case = TRUE), "SectorN"] <- "9"
salaries_g2[grep("PUBLIC SERVICES",salaries_g2$Sector, ignore.case = TRUE), "SectorN"] <- "10"
salaries_g2[grep("WHITE COLLAR",salaries_g2$Sector, ignore.case = TRUE), "SectorN"] <- "11"
```

Sector2	SectorN
WHITE COLLAR	11
PUBLIC SERVICES	10
RETAIL	6
RETAIL	6
TRANSIT/TRANSPORTATION	5
TRANSIT/TRANSPORTATION	5
TRANSIT/TRANSPORTATION	5
SERVICES	8
PUBLIC SERVICES	10
HEALTHCARE	2
STEM	3
PUBLIC SERVICES	10

Methods Used

PCA – Principle Component Analysis is a technique used to emphasized variation and bring out strong patterns in a dataset. It's often used to make data easy to explore and visualize.

We used PCA to apply dimension reduction on our dataset which will help us determine the most significant variables among the 12 variables we had for representing the variations.

1. We used 'prcomp' package in R for implementing the PCA analysis on the data.
2. It performs a principal components analysis on the given data matrix and returns the results as an object of class.
3. We get the following result after the applying it on the normalized dataset.

```
> prn_comp$rotation
                        PC1        PC2         PC3         PC4         PC5         PC6         PC7         PC8
BasePay          0.444282206  0.2708983 -0.079127718  0.10145911 -0.01095877  0.387026509 -0.747467329 -0.063680769
OvertimePay      0.273031311 -0.5717201  0.062795229 -0.39505783  0.65129969 -0.009776054 -0.119477436 -0.003795278
OtherPay         0.307316237 -0.5324469  0.142350522 -0.19890352 -0.74434678 -0.056596349 -0.070094506 -0.007884955
Benefits         0.438370005  0.2732849 -0.007921134  0.02709980  0.04649084 -0.836316965 -0.082766055  0.154635906
TotalPay         0.464784477  0.1097385 -0.051685397  0.03119924  0.01809313  0.358524955  0.455095192  0.657502049
TotalPayBenefits 0.465342977  0.1395706 -0.042976219  0.03178004  0.03191460  0.111776258  0.456031644 -0.734603090
Year            -0.006046783 -0.2219382 -0.967659767  0.08817132 -0.04706053 -0.065899919 -0.003092759  0.000435887
SectorN         -0.094076746  0.4009822 -0.169072138 -0.88520100 -0.12615266  0.045921933  0.012969784 -0.002255320
```

4. By using the '$rotation' we get to see the principal components for each variable.
5. By interpreting this output, we try to determine the most significant principle components.

```
#Using inbuilt R package for PCA
prn_comp <- prcomp(training1,scale. = T)
names(prn_comp)

prn_comp$rotation

#Computing Standard Deviation of each principal component
stdev <- prn_comp$sdev

#Computing Variance (Eigen values)
p_var <- stdev^2
p_var
```

6. To see the Eigen values, we print 'p_var' and get following values

```
> p_var
[1] 4.4865090883 1.0414763674 1.0036362874 0.9668305647 0.4008383632 0.0833249642 0.0165092054 0.0008751594
```

7. By observing the Eigen values, we get a clue how many principle components should be used for the final analysis.
8. We consider first 5 values since they seem to be the deciding factors.
9. We calculate proportion of variance and get the following results

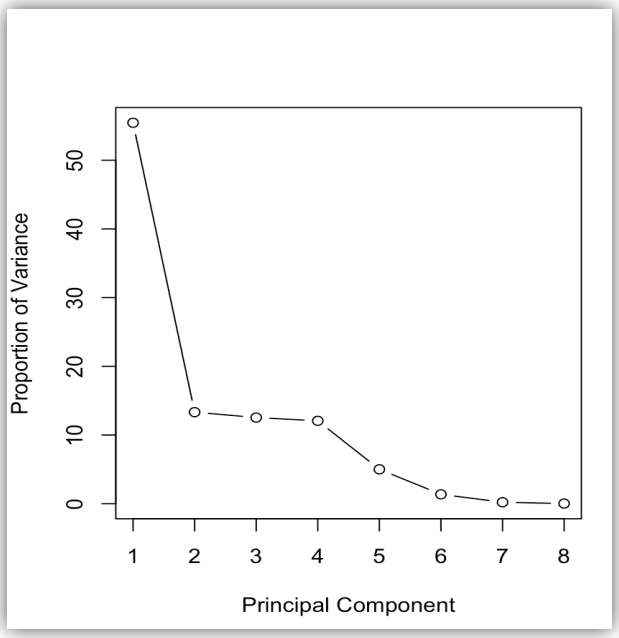
```
> prop_var[1:8]
[1] 55.4504556 13.3279540 12.5422054 12.0681324  5.0075855  1.3655979  0.2150451  0.0230240
```

10. Here we can clearly see the distribution of the proportion of variance.
11. By getting Cumulative Proportion of variance we get the following results.

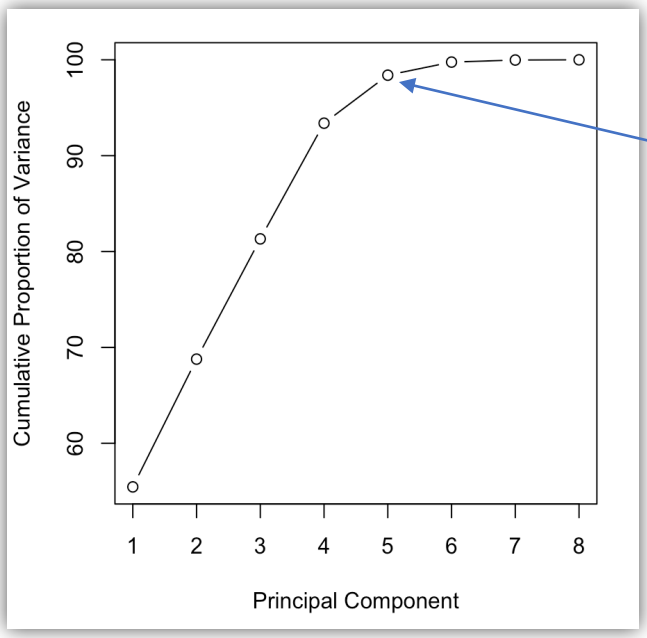
```
> cvar
[1]  55.45046  68.77841  81.32062  93.38875  98.39633  99.76193  99.97698 100.00000
```

12. Here we finalize that the first 5 principle components represent 98.76 % of the variations in the data.

- Scree Plot of the Proportion of Variance



- Scree Plot of Proportion of Variance(Cumulative).



The curve stabilizes after 5th PC

Classification

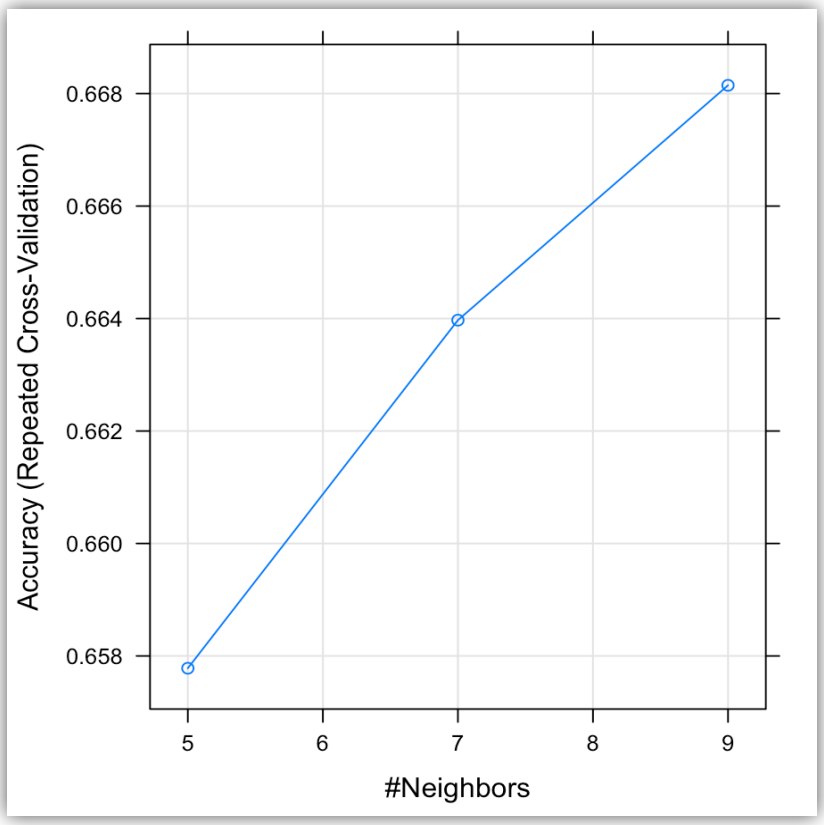
- ***KNN classification*** –

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970's as a non-parametric technique.

1. We used KNN as our first method to classify the workers in Male or Female based on the salaries they were paid in the different categories.
2. 'caret' Library of R was used to apply KNN classification on the trained data.

```
#Classification and Prediction using KNN with PCA
pred_model_pca <- train(GenderN~., data=training, method = "knn", preProcess = "pca",
                        trControl = trainControl(preProcOptions = list(pcaComp = 5),
                        method = "repeatedcv", number = 5))
```

3. We used 'pca' as preprocess as we have applied PCA on the dataframe earlier.
4. We also used method to be "repeatedcv" i.e. repeated Cross- Validation for greater accuracy.
5. We got the following graph after running the code.



6. We get the following confusion matrix of KNN

```
> table(knn.probs,testing$GenderN)

knn.probs     1     2
        1 10378  4249
        2  4184  7016
```

7. By calculating the accuracy, we get
   Accuracy = ((10378+7016)/25827) *100 = ***67.34 %***

- ***Logistic Regression Classification*** –

Logistic regression is another technique borrowed by machine learning from the field of statistics. It is the go-to method for binary classification problems (problems with two class values).

Logistic regression is named for the function used at the core of the method, the logistic function.
The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

1 / (1 + e^-value)

Where e is the base of the natural logarithms (Euler's number or the EXP() function in your spreadsheet) and value is the actual numerical value that you want to transform. Below is a plot of the numbers between -5 and 5 transformed into the range 0 and 1 using the logistic function.

1. We used Logistic regression as a second method to classify our results.
2. "caret" library we described earlier also contains the package to apply logistic regression on the trained dataset.
3. We use 'glm' method to do Logistic regression analysis on our data.

```
#Classification and Prediction using Logistic regression with PCA
pred_model_glm <- train(GenderN~., data=training, method = "glm", preProcess = "pca",
                  trControl = trainControl(preProcOptions = list(pcaComp = 5),
                  method = "repeatedcv", number = 5))
```

4. Here also we used repeated cross validation technique for better results.
5. We get the following confusion matrix after the process

```
> table(glm.probs,testing$GenderN)

glm.probs     1     2
        1 10864  5471
        2  3698  5794
```

6. We calculate the accuracy as before for Logistic regression
   Accuracy = ((10864+5794/25827) *100 = ***64.49 %***

Comparing Accuracy's

1. After performing the KNN classification and Logistic regression classification techniques on the data we compared their accuracies side by side to highlight the recommended method.

```
> table(knn.probs,testing$GenderN)

knn.probs     1     2
        1 10378  4249
        2  4184  7016
```

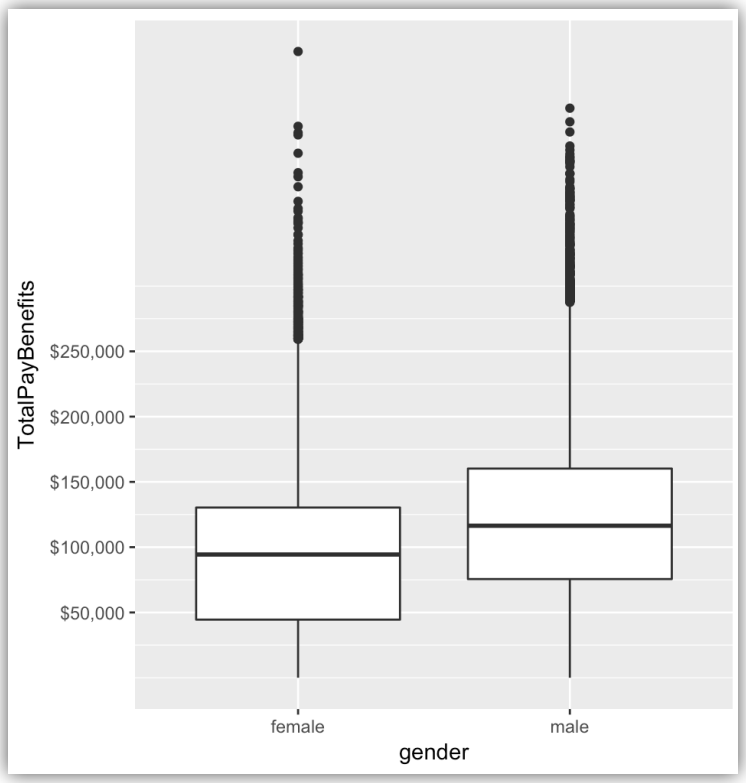
```
> table(glm.probs,testing$GenderN)

glm.probs     1     2
        1 10864  5471
        2  3698  5794
```

2. ***Accuracy*** = ((10378+7016)/25827) *100 = ***67.34 % (KNN)***
3. ***Accuracy*** = ((10864+5794/25827) *100 = ***64.49 % (LR)***
4. We can clearly see that accuracy of KNN is greater than that of the Logistic regression and hence it is recommended method for classification.

Plots and Graphs

1. ***Gender box plot*** –

   - This is a box plot representing the distribution of salaries between males and females.
   - Male employees earn significantly higher salaries in all 4 quartiles.
   - Female employees are concentrated in 48000-127000 salary category whereas majority of the male employees are located in 75000-160000 category of annual salaries.
   - The top most points represent the outliers in the plot.

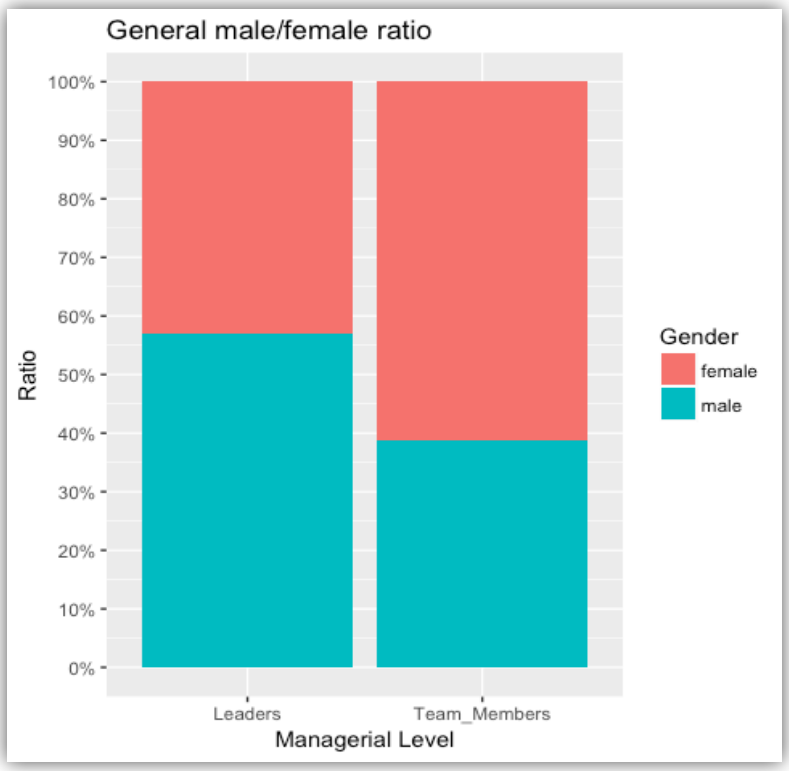


2. ***Distribution of Managerial Jobs –***

This plot describes the managerial jobs that are distributed amongst the male and female employees in San Francisco. We used strings like 'supervisor', 'manager', 'chief', 'director' to separate leader's jobs from that of other workers or team members.

We used following code to get the plot

```
## Salaries Summary by Managerial Level via Gender
sal_grp <- sal_grp %>%
  mutate(JobTitle = tolower(JobTitle)) %>%
  mutate(Leaders = ifelse(grepl("supervisor|manager|chief|head|mayor|director", JobTitle),
                                  "Leaders", "Team_Members")) %>%
  mutate(Leaders = as.factor(ifelse(grepl("assistant", JobTitle),
                                        "Team_Members", "Leaders")))
```



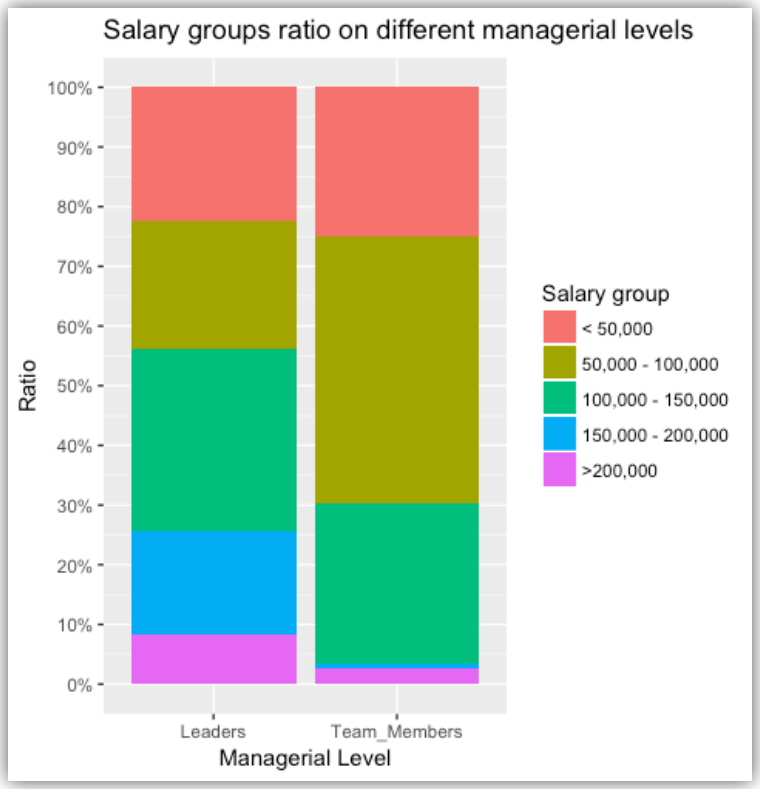
- There are more leaders in male population of the workforce.
- Whereas, more females are observed in the "Team members" category.
- This demonstrates uneven distribution of leadership roles in the city.

3. *Distribution of Managerial Jobs (Salary Groups) -*

Here we categorized the employees based on the salary groups as 0-50k, 100k-150k, and so on. And we plotted a graph that demonstrates the distribution of this salary groups.

We used 'cut' function of R which divides the range of x into intervals and codes the values in x according to which interval they fall. The leftmost interval corresponds to level one, the next leftmost to level two and so on.

```
## Data Enhancement Salary Groups (0-50K-100K-150K-200K)
sal_grp <- salaries_g2
sal_grp$SalaryGroup <- cut(salaries_g2$TotalPayBenefits,
                           breaks = c(-Inf, 50000, 100000, 150000, 200000, Inf),
                           labels = c("< 50,000", "50,000 - 100,000", "100,000 - 150,000"
                                      , "150,000 - 200,000", ">200,000"),
                           right = FALSE)
```

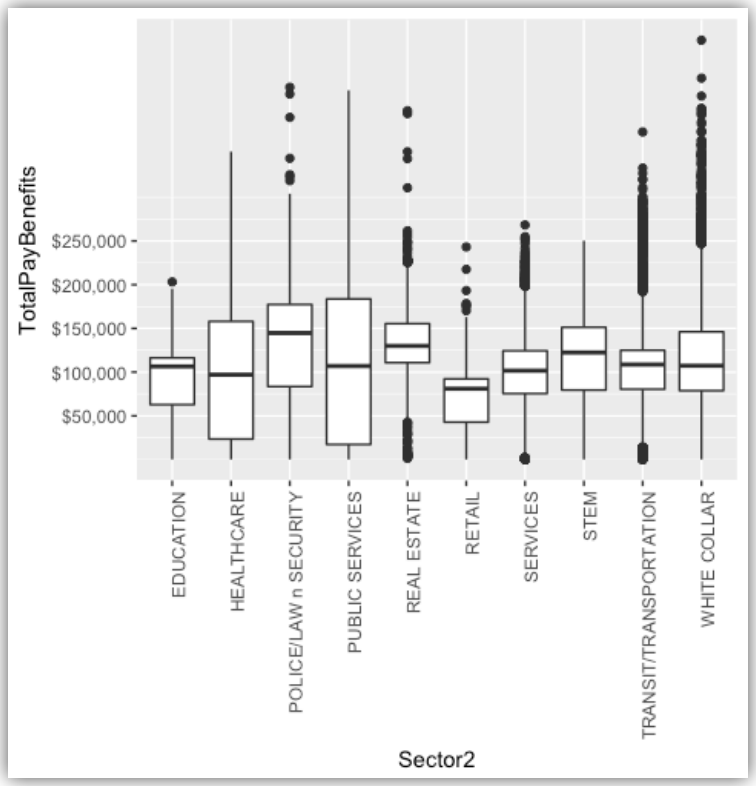


1. Here we can see managers are considerably paid higher than team member or lower designation employees.
2. Most managers earn between 100k-150k salary group whereas the team members earn 50k-100k category.
3. This highlights the fact how unevenly the managers are paid in the organizations.

4. Sector Box Plot –

We had categorized the employees in 8 sectors based on their profession. We had very diverse job titles so we had to categorize them into some sectors so as to carry out some legitimate analysis. Following is the box plot generated from the data.

```
## Salaries Summary by Sector and histogram
by(sal_grp$TotalPayBenefits,sal_grp$Sector2,summary)
q <- qplot(x=Sector2, y=TotalPayBenefits, data=sal_grp, geom='boxplot') +
  scale_y_continuous(labels = scales::dollar, breaks = c(50000, 100000, 150000, 200000, 250000))
q + theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



1. It is observed that 'White collar' jobs appear to be highest paying jobs in among all other sectors.
2. "Public Services" people have more stable distribution throughout.
3. "Real Estate" people have most diverse distribution because of probably very unstable real estate price variations.
4. Most "Retail" sector jobs are under 100k which is also a noticeable fact.

5. ***Word Cloud of Jobs –***

   - We developed word cloud of the people from the top 80% quartile of the total income and displayed the job titles that come under that quartile.

   - Second word cloud represents the lowest 20% quartile of the total income and displayed the job titles that come under that quartile.

Conclusion

After studying and analyzing different trends and distribution of the salaries from the dataset, we have come to the conclusion that the distribution of salaries is quite uneven in the city of San Francisco. We observed how there is considerable difference in the salaries of the male and female employees in the same sectors. Female employees were highly underpaid as compared to their male counterparts in the same profession. We also saw the pay-gap between the managerial positions and the team members of different professions which shows uneven payment in an organization. Our final conclusion being that San Francisco needs to regulate the way their city's employees are compensated in order for everyone to be able to survive in the city peacefully.

Future Research

We have tried and highlighted the areas that the city of San Francisco needs to focus on and try to improve the underpaid employees.

1. This will help them improve the overall economic condition of their administration. Better lives for employees with appropriate salaries can be expected with some legislative changes from the administration's side.
2. Also, the gender-pay-gap can be reduced by proposing new rules for equal payment for both genders.
3. Managing jobs salaries can be regulated to divert some of the cash flow towards the underpaid sectors of the economy
4. The dataset can also be further used to discover how much bonuses the employees are receiving from their respective sectors.
5. The facts like 'How the employees are paid for overtime' can also be explored from the given dataset.
6. If we somehow get state tax data on these salaries we can probably get clearer picture of the overall economic factors affecting the incomes.

References

https://www.kaggle.com/kaggle/sf-salaries
https://machinelearningmastery.com
http://www.sfchronicle.com/
http://www.sfgate.com/
https://www.glassdoor.com