*ALY6010*

*Capstone-Report*

*Northeastern University, Vancouver, BC, Canada*

*Mohsen Soltanifar, PhD, AStat*
*Faculty Lecturer*

*16th, December 2022*

Northeastern University
Vancouver

# Earning analysis of San Francisco

Milan Prajapati
Master of Professional Studies in Analytics

The report is created based on the job market condition of San Francisco between the years 2011 to 2014, The dataset is taken from the open-source government websites of San Francisco to make the analysis more reliable and our aim is to find the best professionals in terms of earning and government fund spending on various departments and optimize the government spending as well as a finding profession with highest financial freedom and for succeeding in our purpose we are using the various methods such linear regression, hypotheses testing, t-tests and many other data analysis concepts

Base Pay, Overtime Pay, other Pay, Police Department, San Francisco

## 1 Introduction

The dataset contains important information about the job market of San Francisco and it's important to look out at the past data on the job market and government spending on various job files by analyzing this data we can get information about how the future trend will work and in which filed of jobs had more capital gain and had support the economy of the city and country

There are many questions that we can answer by analyzing the dataset and a couple of questions we are answering in the report are:

1) **What was the Average earning of people working in the city of San Francisco, Is the minimum average earning of workers was above or below 35,000 $? and did the people working more hours earn more money or the total earnings of individuals is dependent on their base salaries?**
2) **What was the average earning of police officers between the years 2011 to 2014 based on the position? how much extra payment police officers were paid for various facilities and do this other payment have any relation with their base salary?**

3) **What is the relation between employees' base pay and the total amount they get paid including overtime and other allowance in various professions, do the overtime earnings have any significant dependence on base salary?**

The report provides the detailed answer to all the questions listed above with strong and clear statements followed by mathematical explanation and visualization

Solution structure:

- Explanation and Assumptions
- Mathematical Solutions
- Visualization
- Conclusion

# 2 Materials & Methods

## 2.1 Dataset

The dataset contains information such as Id, Employee Name, Job Title, Base Pay, Overtime Pay, Other Pay, Benefits, Total Pay, Benefits, Year, Notes, Agency, and Status and for making accurate predication and analysis data is chosen from government websites, along with data set also contain many null and empty value, to eliminate that data cleaning and modeling is also required in the chosen dataset

## 2.2 Statistical Analysis

### 1. Hypothesis tests

2.2.1.1 Hypothesis test for Question No.1

To identify the Average base pay of people working in San Francisco the one-sided t-test is used with an α value of 0.05 which indicates a 95 percent of the confidence interval and by making an assumption that the average base pay of employees is around 66,000 $ a year by using the formula $t = ( \bar{x} – \mu) / (s / \sqrt{n})$

Where,

- $\bar{x}$ = Observed Mean of the Sample
- $\mu$ = Theoretical Mean of the Population
- s = Standard Deviation of the Sample
- n = Sample Size

2.2.1.2 Hypothesis test for Question No.2

To calculate base pay and other pay together, the two-sided t-test is used by making the assumption that base pay and other pay both are the same for every police officer irrespective of their post and which is 18500 $, and 95 percent confidence interval is used for calculation by using the formula $t = ( \bar{x}1 – \bar{x}2) / \sqrt{} [(s21 / n 1 ) + (s22 / n 2 )]$

Where,

- $\bar{x}1$ = Observed Mean of 1st Sample
- $\bar{x}2$ = Observed Mean of 2nd Sample
- s1 = Standard Deviation of 1st Sample
- s2= Standard Deviation of 2nd Sample
- n 1 = Size of 1st Sample
- n 2 = Size of 2nd Sample

2.2.1.3 Hypothesis test for Question No.3

To find the average base pay for employees working in San Francisco we perform a statistical t-test by taking 95 percent confidence intervals and considering the average base pay in San Francisco is around 75,000 $ a year with the same formula used in first t test

## 2. Correlation tests

2.2.2.1 Correlation test for Question No.1

For finding the relation between employees' salaries and their income by working extra hours we took some samples from the dataset and run a Pearson correlation test with the formula $r=\sum(x-mx)(y-my)\sqrt{}\sum(x-mx)2\sum(y-my)2$ $r = \sum ( x – m x ) ( y – m y ) \sum ( x – m x ) 2 \sum ( y – m y ) 2$

where,

- mx = mean distribution of x
- my= mean distribution of y

2.2.2.2 Correlation test for Question No.2

For finding the dependence between base pay and other income of police officers of various departments we run a correlation test between base pay and other pay using the same formula $r=\sum(x-mx)(y-my)\sqrt{}\sum(x-mx)2\sum(y-my)2$ $r = \sum ( x – m x ) ( y – m y ) \sum ( x – m x ) 2 \sum ( y – m y ) 2$ where mx and my is mean distribution of x and y respectively

2.2.2.2 Correlation test for Question No.3

To check whether base pay and total pay have any dependence on each other we run the same correlation test by using the base variable and total pay, the result of the correction function will provide the dependence count

### 3. Linear regression models

The regression model is used to determine the strength of the dependence and in order to answer all the above questions linear and nonlinear regression models are being used, from the visualization we can determine the type of regression and strength of dependence based on the regression line

### 4. Statistical software used.

In the report mainly R programing language and Rstudio are used along with that ggplot2, dplyr, scales, data.table , mass, stargazer libraries are used for visualization and calculation purpose

# 3 Results

## 3.1 Exploratory Data Analysis (EDA)

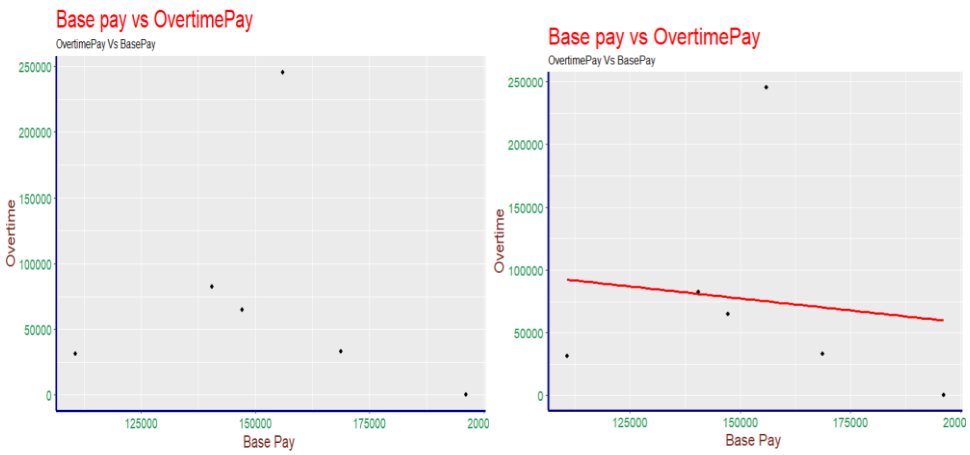
The Data set contain 13 columns and 1,48,654 rows and each column has a different data type and structure for providing a solution to the given question majorly base pay, total pay, and overtime pay are being used while other columns such as department, status, are used for grouping purpose and in dataset preparation

## 3.2 Average earning of people working in the city of San Francisco,

To answer the first question and to identify the average base salary of employees working in San Francisco, one side t-test is performed by making the assumption that the average salary will be around 66,000 $ per year and the result of our hypothesis testing prove our assumption wrong and provide that average **base salary range of 66107.61$ to 66543.29 $ per years**.

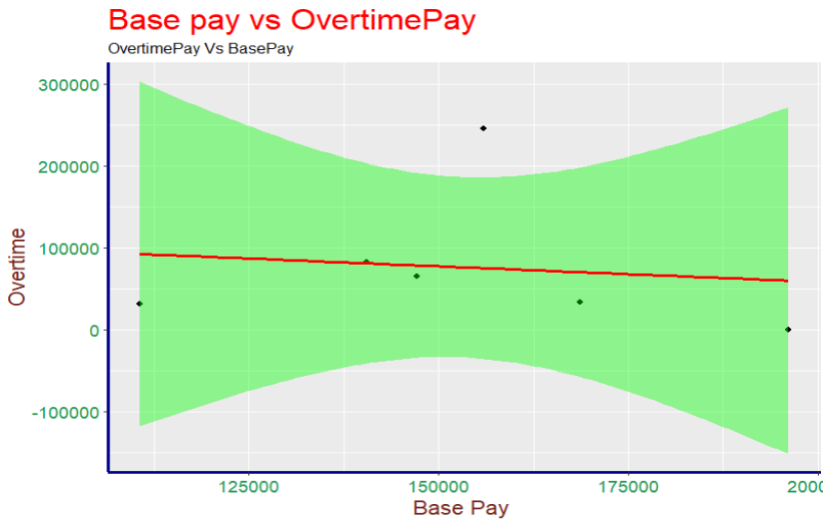
To Identify the relation between Base Pay and overtime we run the co-relation test by taking a few random samples and the results prove **that overtime earnings are 0.26 percent positively dependent upon the base pay** and working more hours will help an individual to earn more but comparatively, people who had more base salary can earn more even by working fewer hours comparatively.

**Visual representation of data:**



**Regression model:**

Here the regression line shows negative linear relation and which contradicts our hypothesize test the reason behind this is that the overtime pay is dependent upon the base pay but the employees with high salaries do less overtime work so base pay and overtime pay are positive liners to each other but for entire data with an employee with high salary do less overtime work.



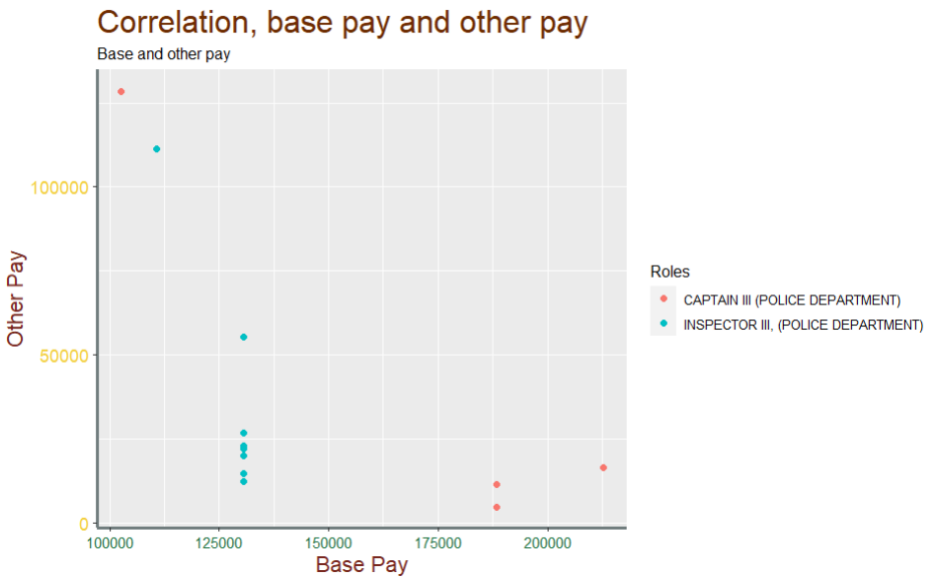

## 3.3 The average earning and fund allocation for police of department

To find the average earnings of police officers, the data is first filtered by using the department, and a two-sided statical test is taken used by making the assumption the average base salary and other extra funds allocated for various facility provided to police offices is cost same to the government which is 18,500 $ per year irrespective of the department in which they work

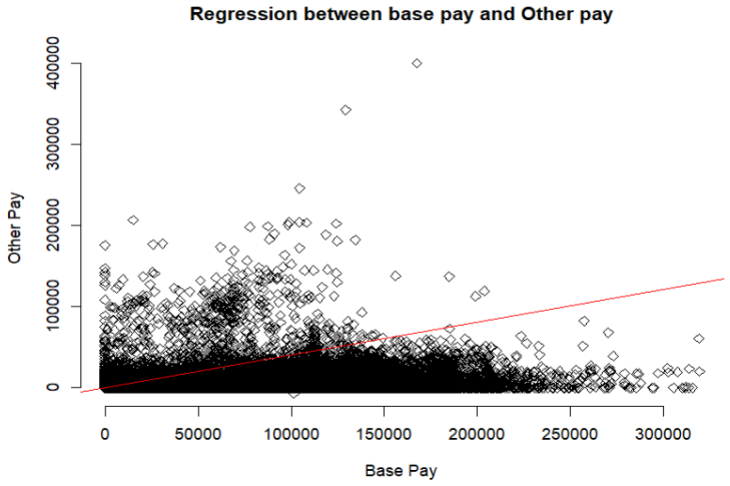
The result of our hypothesis proves our assumption wrong and with results proving that the average cost to the government for **providing salary to individual police officers is around 1,03,534 $ per year while the average other costs spent on individual police officers are around 1,82.685 $ per year.**

To check whether the funds provided to individual police officers depend on their yearly income or not we run a correlation test between base pay and other earning pay and the results show that other funds provided to the individual officer **are 0.28 percent positively dependent on their yearly income**, which shows that police officer which has high base salary received comparatively

**Visual representation:**




**Regression model:**




The visualization shows **no linear regression is exist** between base pay and other pay

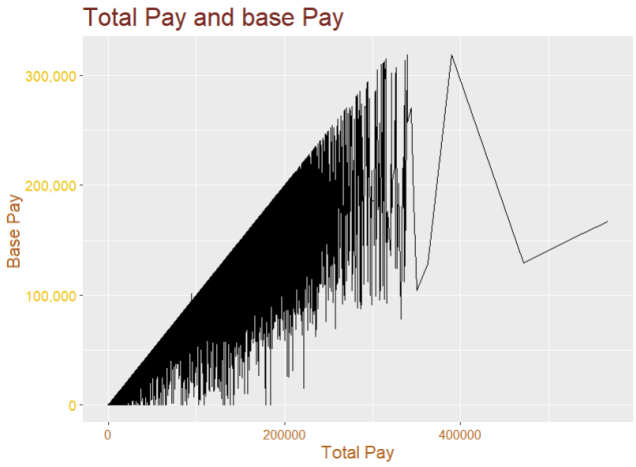
## 3.4 Relation between base pay and total pay for employees

To find the total average earnings of employees working in San Francisco we run one-sided hypothesis testing by making assumptions that every individual working in the city earns 75,000 $ per year, The result of hypothesis testing rejects our assumption and gave the **average earnings of people in San Francisco in between 74,511 $ to 75,025 $ per years.**

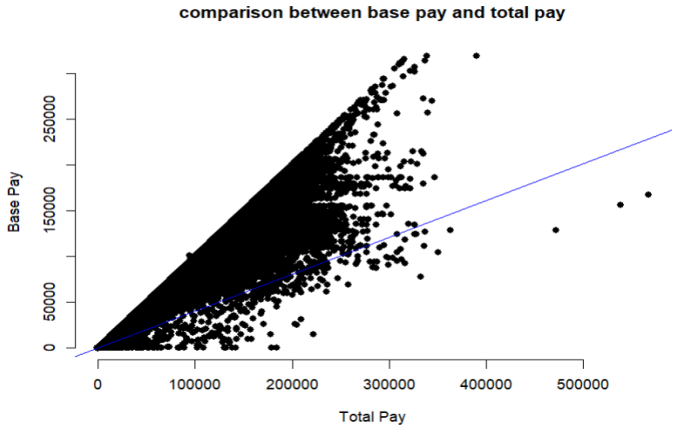
To find the relation between Base Pay and total earnings of individual employees working in the city of San Francisco we run the co-relation test by taking a few random samples and the results prove that the total **earnings of the individual are 0.95 percent dependent on**

**their base salary.**

**Visual representation:**




**Regression model:**




The visualization besides shows **Positive linear regression** between total pay and base pay, with the increment the base pay of employee total pay is increased.

# 1 Discussion

The result shows that earning from overtime is dependent upon the base salary but people with higher base pay earn less income by doing extra work while for the police department base salary and other income is very less dependent and the government spends more funds in providing other services to police individuals as compared to spending on their base salary while base pay and total pay are positively dependent for every job filed, and any change in base pay linearly impact the total pay

The dataset used contains many null and empty values which affect the results along with that

the other pay field dataset does not provide detailed information about what kind of other pays are considered, so this anonymous factor can also impact the final results

In the next part of the project, we are planning to implement a predictive model based on this data and the study will try to predict upcoming opportunities for San Francisco and also find out which job has good growth in upcoming years in San Francisco city

# 2 References


*What Is a Regression Model?* (2021, June 16). IMSL by Perforce. https://www.imsl.com/blog/what-is-regression-model

Author Removed At Request Of Original Publisher. (2016, December 1). *Nonlinear Relationships and Graphs without Numbers – Principles of Macroeconomics*. Pressbooks. https://open.lib.umn.edu/macroeconomics/chapter/nonlinear-relationships-and-graphs-without-numbers/

taylor.curley@gatech.edu. (n.d.). *T-Tests: One and Two-sample*. https://rstudio-pubs-static.s3.amazonaws.com/326285_978dc80a48de439f86cb0ac56925ddf6.html


# 3 Appendix

**Question 1 (R code Output)**

One Sample t-test

```
data:  .
t = 2.9, df = 148044, p-value = 0.003
alternative hypothesis: true mean is not equal to 66000
95 percent confidence interval:
 66108 66543
sample estimates:
mean of x
   66325


Call:
lm(formula = OvertimePay ~ BasePay, data = relationdf)

Residuals:
    2    38    40    45    55    95
170043 -61216 -13464 -59779   1456 -37040

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 134608.588 236101.293    0.57     0.60
BasePay        -0.382      1.519   -0.25     0.81

Residual standard error: 97200 on 4 degrees of freedom
Multiple R-squared:  0.0155,           Adjusted R-squared:  -0.231
F-statistic: 0.0631 on 1 and 4 DF,  p-value: 0.814

> cor(df$BasePay,df$OvertimePay, use = "complete.obs") # corelation between
      variables
      [1] 0.2667
```

**Question 2 (R code Output)**

```
            Welch Two Sample t-test

data:  dfGr$BasePay and dfGr$OtherPay
t = 5.2, df = 4.5, p-value = 0.005
alternative hypothesis: true difference in means is not equal to 18500
95 percent confidence interval:
  77346 202117
sample estimates:
mean of x mean of y
   165499     25767

Call:
lm(formula = BasePay ~ OtherPay, data = dfGr)

Residuals:
    2    37    45    55
 34085  -7071   6724 -33738

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 201890.777  29377.491    6.87    0.021 *
OtherPay       -1.412      0.921   -1.53    0.265
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34600 on 2 degrees of freedom
```

```
Multiple R-squared:  0.54, Adjusted R-squared:  0.31
F-statistic: 2.35 on 1 and 2 DF,  p-value: 0.265
```

**Question 3 (R code Output)**

**One Sample t-test**

```
data:  .
t = -1.8, df = 148653, p-value = 0.08
alternative hypothesis: true mean is not equal to 75000
95 percent confidence interval:
 74512 75025
sample estimates:
mean of x
   74768

Call:
lm(formula = BasePay ~ TotalPay, data = relationdf)

Residuals:
    2     38     40     45     55     95
 1370 -42290  -5783  43149 -12335  15891

Coefficients:
             Estimate   Std. Error t value Pr(>|t|)
(Intercept) 151492.65793  37232.53656    4.07    0.015 *
TotalPay       0.00576      0.11870    0.05    0.964
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32000 on 4 degrees of freedom
Multiple R-squared:  0.000588,       Adjusted R-squared:  -0.249
F-statistic: 0.00235 on 1 and 4 DF,  p-value: 0.964

> cor(df$BasePay,df$TotalPay, use = "complete.obs") # corelation between
      variables
[1] 0.9545
```

# THANK YOU!

Special Thanks to
Professor Mohsen Soltanifar, PhD, AStat
For the guidance and Instruction.