## COMP4925 - Data Mining

Daniel Lin

Norman Lim

Manveer Bhangu

Subproject 1

## Introduction

A breakdown of the city of San Francisco employment show that roughly 25% are professional business services, 16% in government services, 15% in leisure and hospitality, 11% in education and health care, and 9% in financial services. San Francisco's GDP continues on a steady increase at approximately 11.92 billion each year from 2001 to 411.97 billion in 2014.

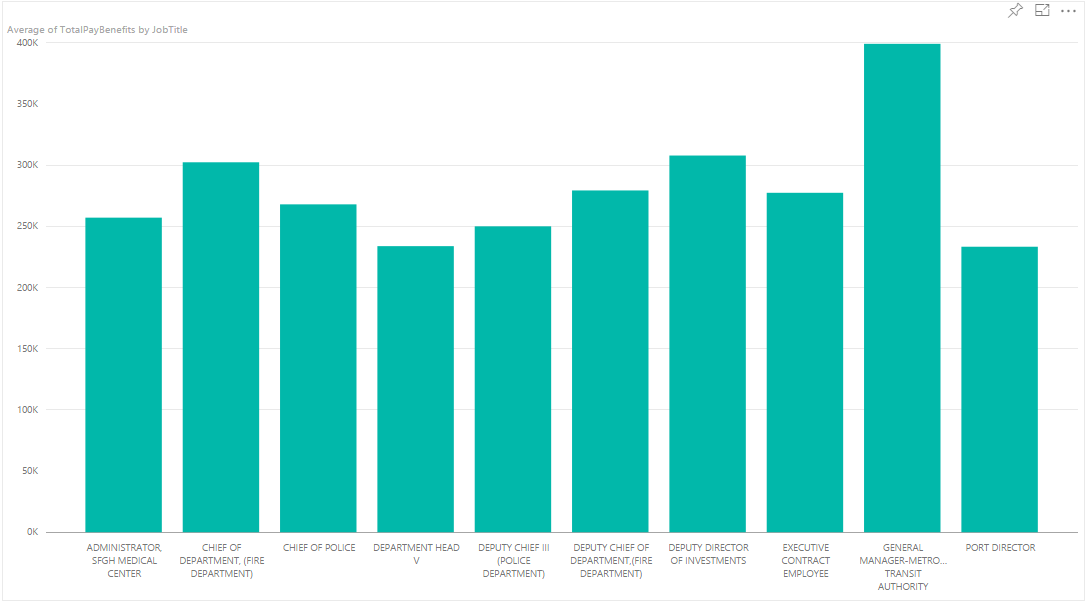
Our dataset focuses on the year from 2011 to 2014. San Francisco's GDP those 4 years increased by approximately 16.7 billion each year.

|  |
| --- |
| Excel Insights |
| We hypothesized that the increase in GDP would reflect an increase in salary. We want to find if the apparent increase of the salary is at all significant.   |  |  |  |  | | --- | --- | --- | --- | | **Row Labels** | **Average of BasePay** | **Average of OvertimePay** | **Average of OtherPay** | | 2011 | 63595.95652 | 4531.065429 | 3617.081926 | | 2012 | 65436.40686 | 5023.417824 | 3653.437583 | | 2013 | 69630.03022 | 5281.64198 | 3819.969007 | | 2014 | 66564.42192 | 5401.993737 | 3505.421251 |   Using t-Test (unequal variance): Two Sample for Means, we examined total salary from 2011 and 2014 to see if there are any significant difference over 4 years. We found that there is a significant probable difference between the year 2011 and 2014 in terms of salary.   |  |  |  | | --- | --- | --- | | t-Test: Two-Sample Assuming Unequal Variances | |  | |  |  |  | |  | *Variable 1* | *Variable 2* | | Mean | 71678.18572 | 75335.15144 | | Variance | 2249389889 | 2668368509 | | Observations | 33962 | 35712 | | Hypothesized Mean Difference | 0 |  | | df | 69586 |  | | t Stat | -9.740602164 |  | | P(T<=t) one-tail | 1.04569E-22 |  | | t Critical one-tail | 1.644875525 |  | | P(T<=t) two-tail | 2.09139E-22 |  | | t Critical two-tail | 1.959998076 |  | |

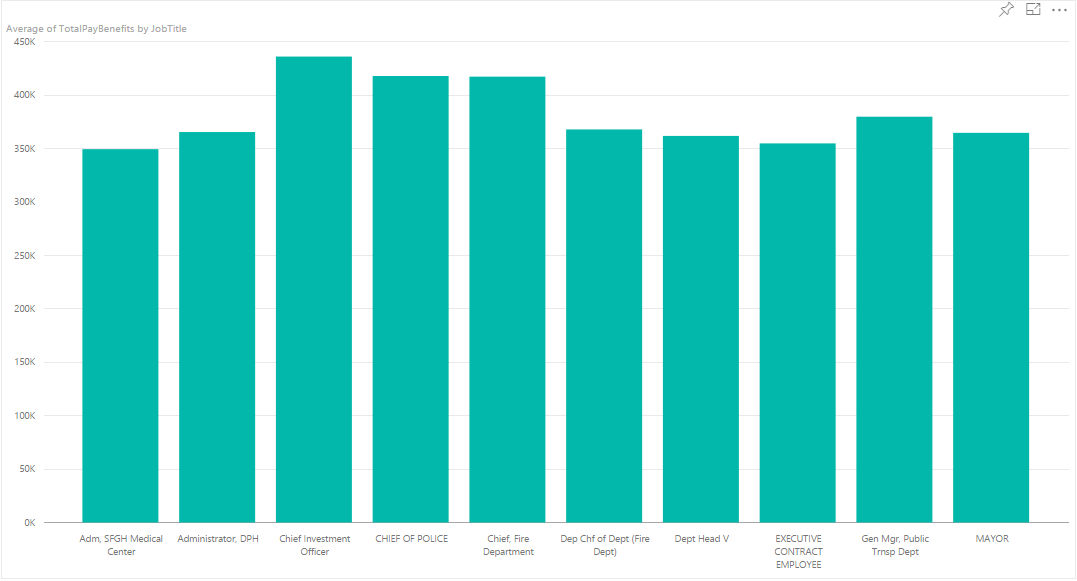
Next, we want to see what the top paid occupations are and if they are seeing any increase in salary from the year 2011 to 2014. We hypothesized that they will see an increase in salary.

***Top 10 paying job in 2011 compared to 2014***

**Top paying jobs in 2011**



**Top paying jobs in 2014**



Top paid jobs are the following:

Administrator San Francisco General Hospital Medical Center

Chief of Department (Fire Department)

Chief of Police

Department Head

Deputy Chief

General Manager of Transit Authority

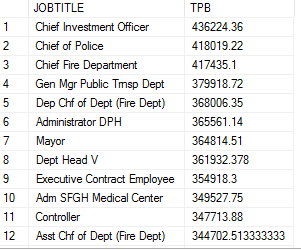
Port Director

Executive Contract Employee

Mayor

Chief Investment Officer

We wanted to verify BI against SQL. Results are the same.



Chief Investment Officer is the highest paying job. We were interested in finding out who they are. If you have the term “investment” in SF, you are making the big bucks, only four people have it.

Top 10 paying jobs do correspond to some of what we know about San Francisco’s high number of government jobs.

**Overall pay increase**

From the two graph we can see that top paying jobs didn’t really change a whole lot but one interesting thing we noticed from this data is that there was an average of 50K increase in the top jobs from 2011 to 2014. This made us curious. From this test, we decided to test if lower paying jobs saw any type of large increase over 3 years.

**Lower paying jobs test**

Hypothesis:

Since we know there was a significant increase in pay overall in the top paying jobs we believe that there will also be an increase in lower paying jobs from 2011 to 2014. It may just be a slight increase but we think there will be one.

We also assumed the pay increase will be skewed toward the top getting a larger percentage increase than those closer to the minimum wage.

Data:

For our data we randomly choose 10 jobs between the pay of $25,000 and $32,000 in 2011 and compared them to 2014. We choose a double bar graph to help represent our findings.

2011:

Accountant Intern 29031

Assistant Recreation Director 29442

Beautician 31301

Clerk 28968

Crafts Instructor 31213

Home Health Aide 28066

Museum Guard 29352

Museum Preparator 25896

Recreation Specialist 31016

Sheriff’s Cadet 29312

2014:

Accountant Intern 32406

Assistant Recreation Director 19801

Beautician 70411

Clerk 55279

Crafts Instructor 53383

Home Health Aide 57535

Museum Guard 42266

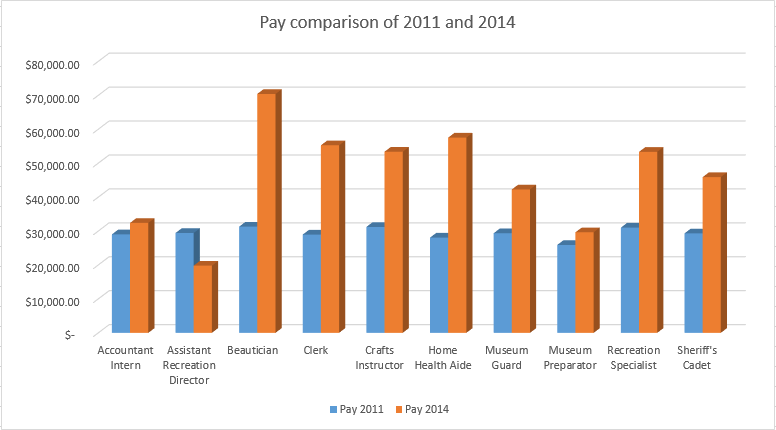
Museum Preparator 29636

Recreation Specialist 53355

Sheriff's Cadet 45913

Conclusion:

From the graph below we can see that aside from Assistant Recreation Director, the pay increase for some of the jobs went up drastically such as a Beautician who saw an increase of 30K from 2011 to 2014. Overall our hypothesis was correct and we can conclude that most jobs, even lower paying jobs, saw an increase in pay. Contrary to our instinct, the pay increase was not exclusive to the top.



|  |
| --- |
| Miscellaneous |
| Are public servants providing essential services working more overtime hours?  We hypothesized that they would not work more because the public sector lacks the same competitive atmosphere compared to the private.  It turns out that public employees that is essential to the smooth operation of a city occupies a significant portion of the overtime pay. With the transit operator and firefighter to be the top two, followed by police officer.    We can look at the instance (count) of actual overtime pay between these professions.   |  |  | | --- | --- | | JobTitle | TRANSIT OPERATOR | |  |  | | **Count of OvertimePay** |  | | 9424 |  |  |  |  | | --- | --- | | JobTitle | Police Officer | |  |  | | **Count of OvertimePay** |  | | 1476 |  |   Police officer and Transit operator being some of the most common professions to be paid to work overtime. In comparison, a common profession like a junior clerk have far less occurrences of a paid overtime.   |  |  | | --- | --- | | JobTitle | JUNIOR CLERK | |  |  | | **Count of OvertimePay** |  | | 755 |  | |
| SQL scripts we ran to expand on our results from BI.  SELECT DISTINCT  JOBTITLE,  AVG(TOTALPAYBENEFITS) AS TPB  FROM SALARIES  WHERE  YEAR = 2014  GROUP BY JOBTITLE  ORDER BY TPB DESC  SELECT  EMPLOYEENAME,  JOBTITLE,  YEAR,  AVG(TOTALPAYBENEFITS) AS TPB  FROM SALARIES  WHERE  --YEAR = 2014  JOBTITLE LIKE '%Investment%'  GROUP BY JOBTITLE, EMPLOYEENAME, YEAR  ORDER BY TPB DESC |
|  |