Project 3 - Interactive Visualization using Tableau

Nathan McIntosh

Due October 13, 2020

1 Introduction

The data set used for this project is the Kaggle San Francisco city employee salary data. It represents data on all the employees of the city from 2011 to 2014, some 148k rows. Each row contains a name, job title, and compensation in various forms for a given year. **Explain the importance of the topic and data**

2 Dataset

	Id	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits	Year
count	148654	148045	148650	148650	112491	148654	148654	148654
mean	74328	66325	5066	3649	25008	74768	93693	2013
std	42913	42765	11454	8057	15402	50517	62794	1
min	1	-166	0	-7059	-34	-618	-618	2011
25%	37164	33588	0	0	11535	36169	44066	2012
50%	74328	65007	0	811	28629	71427	92404	2013
75%	111491	94691	4658	4236	35567	105839	132876	2014
max	148654	319275	245132	400184	96571	567595	567595	2014
category	nominal	quant	quant	quant	quant	quant	quant	quant

3 Analytical Questions

1. Which jobs get the most and least base pay, as ordered by median base pay?
2. Which jobs get the most overtime pay?
3. Which jobs have the largest pay/benefits disparity?
4. Do the number of employees in a job roll affect the median pay of that job?
5. Which jobs had the greatest change in compensation from 2011 to 2014?

4 Design - design process, screenshot(s) of the different visualizations, and explanation of overall dashboard

In creating this dashboard, I wanted to examine questions 1, 2, 3, and 4. To create the first, I set the job title in the "columns" spot and $median(BasePay)$ in the "rows", sorted by $median(BasePay)$, and used a line as the marker to produce the plot in the top left. To examine count of job title against median total pay, I set $cnt(JobTitle)$ as the columns, $median(BasePay)$ as the rows, and set Job Title as a "detail", which created the scatter plot seen in the top right. Because most of the jobs have less than 100 people performing

them, and a few have almost 10,000, I used a log scale better spread the data out. To see the top ten jobs by overtime pay, I set job title as columns, and median overtime pay as the rows, and sorted the data. Finally, to visualize the pay/benefits disparity, I created a calculated field named BasePayBenefitsDiff, calculated by $median((BasePay - Benefits)/BasePay)$. This calculated field is 0 when there is no difference between Base Pay and Benefits, and it is 1 when Benefits is 0, or is negligibly small in comparison to Base Pay. I then placed this newly calculated field as the columns, and job title as the rows, and sorted. The final result is the dashboard seen below.

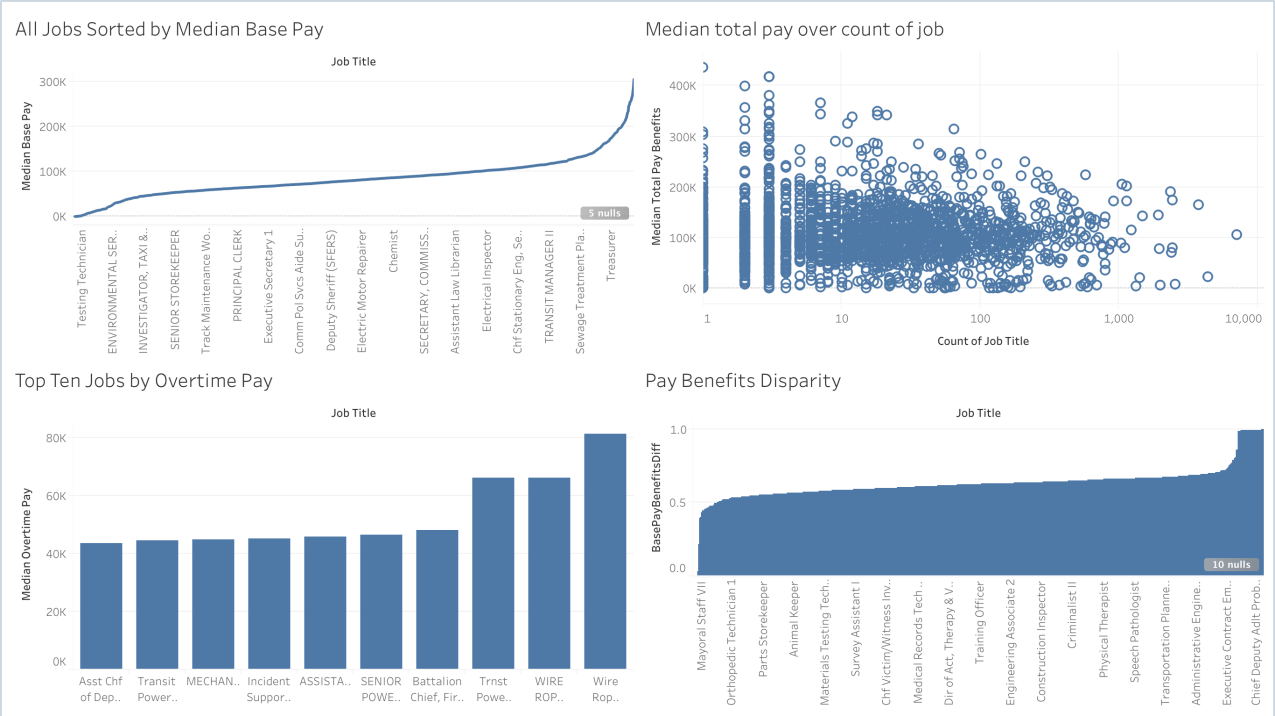

Figure 1: Screen Shot of the Dashboard

5 Discussion - explain how the dashboard answers some of the analytical questions

Question 1 was which jobs get the most and least base pay, as ordered by median base pay? The top left plot answers this question by showing us the total range of base pays that employees of the city receive. The plot shows that some receive nothing for their work, most receive less than $100k, and a few receive over $300k.

Question 2 was which jobs get the most overtime pay? The bottom left graphic clearly answers this question with its bar chart of said top ten jobs as ranked by overtime pay. The names are listed at the bottom of each bar, and the bar heights show the median overtime pay for that profession.

Question 3 was which jobs have the largest pay/benefits disparity? The bottom right plot answers this with a bar chart of BasePayBenefitsDiff for all professions. The interactive nature allows the user to selectively view jobs.

Question 4 was do the number of employees in a job roll affect the median pay of that job? The top right plot shows that this is unlikely to be true. There is no clear trend here one way or another.