Project 02 - Data Exploration and Design

EN 605.427 Data Visualization

For this project, I have chosen the Kaggle San Francisco Salaries dataset. It describes the salaries of all San Francisco City employees from the year 2011 to 2014. Let's dive into an exploration of the data.

Explore and Analyze the Data

A basic description from pandas of what the data looks like when loaded in memory is as follows:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 148654 entries, 0 to 148653
Data columns (total 13 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   Id                148654 non-null  int64
 1   EmployeeName      148652 non-null  object
 2   JobTitle          148654 non-null  category
 3   BasePay           148045 non-null  float64
 4   OvertimePay       148650 non-null  float64
 5   OtherPay          148650 non-null  float64
 6   Benefits          112491 non-null  float64
 7   TotalPay          148654 non-null  float64
 8   TotalPayBenefits  148654 non-null  float64
 9   Year              148654 non-null  int64
 10  Notes             0 non-null       float64
 11  Agency            148654 non-null  category
 12  Status            38119 non-null   category
dtypes: category(3), float64(7), int64(2), object(1)
memory usage: 12.0+ MB
```

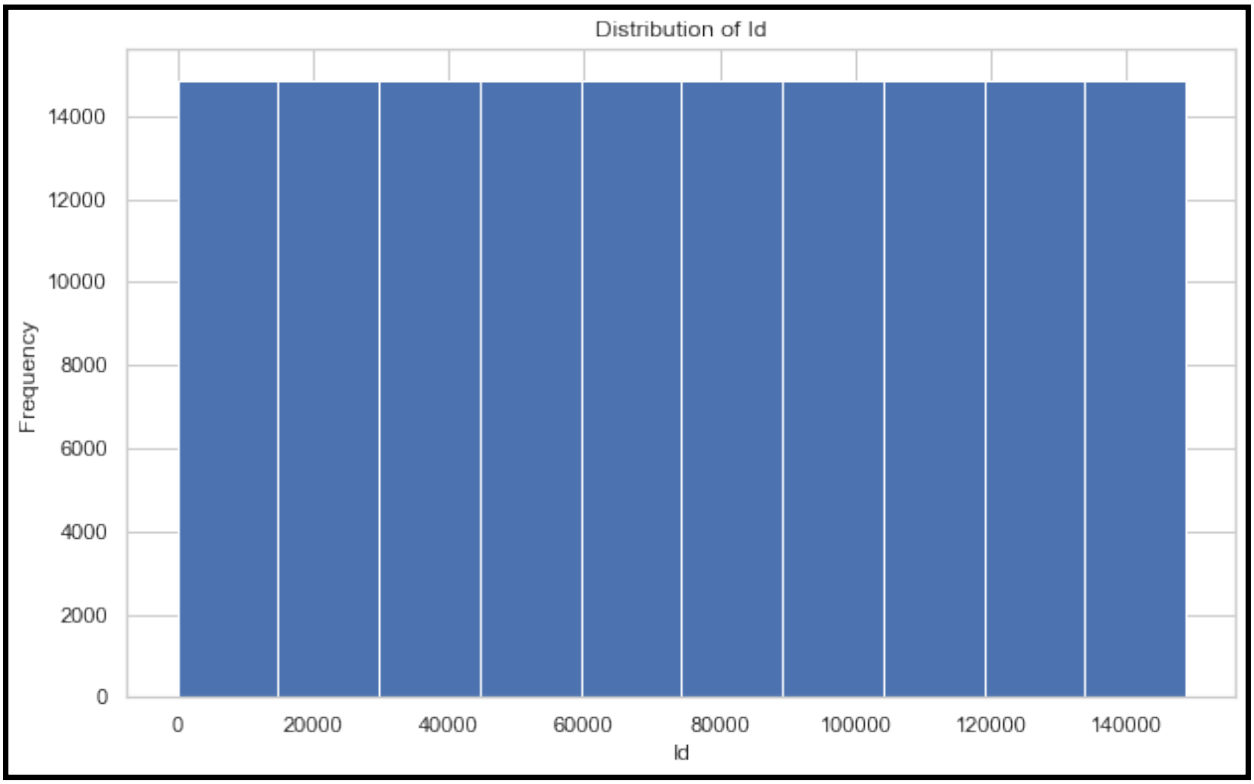
This puts the data in the following categories

Variable	Data Type
ID	Stored as integer, but representing nominal information
EmployeeName	Nominal

Variable	Data Type
JobTitle	Nominal (Categorical)
BasePay	Quantitative
OvertimePay	Quantitative (Ratio)
OtherPay	Quantitative (Ratio)
Benefits	Quantitative (Ratio)
TotalPay	Quantitative (Ratio)
TotalPayBenefits	Quantitative (Ratio)
Year	Stored as integer, but representing nominal information
Notes	Nominal (Arbitrary)
Agency	Nominal (Categorical)
Status	Nominal (Categorical)

Now that we've seen their categories, let's also examine some basic plots of all this data.

ID Histogram



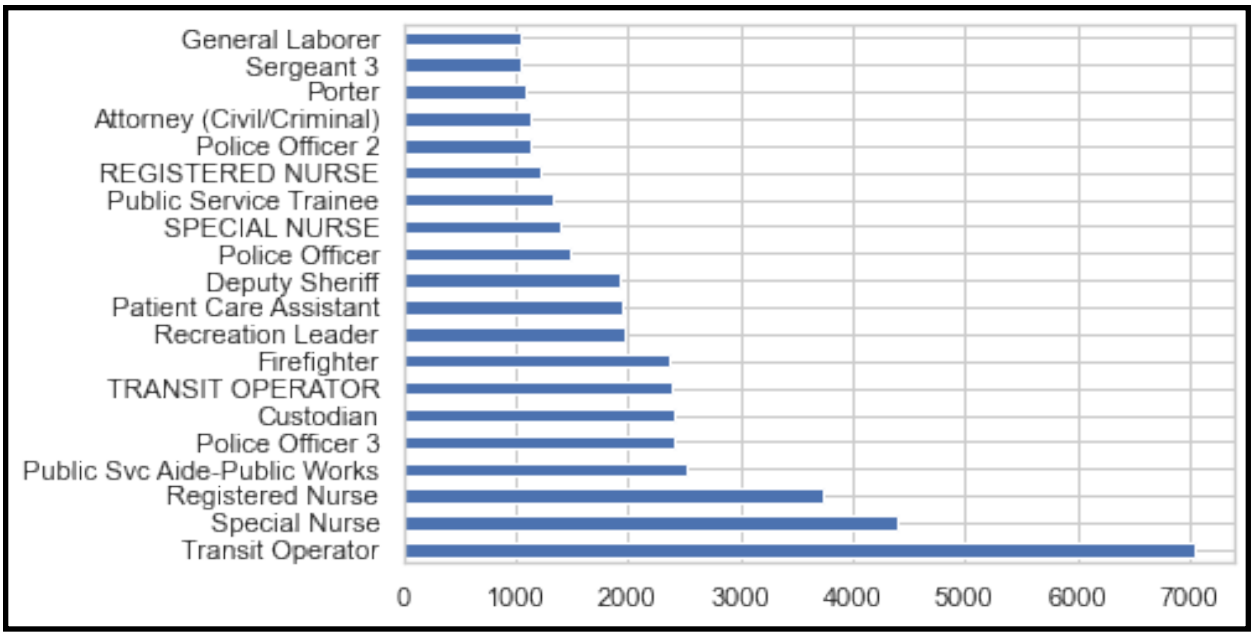
As we can see, the IDs are distributed evenly across all the employees.

EmployeeName
First let's look at the top 20 most common names

Name	Count
KEVIN LEE	22
RICHARD LEE	19
WILLIAM WONG	18
STEVEN LEE	18
DAVID WONG	16
STANLEY LEE	16
MICHAEL LEE	15
JOHN CHAN	15
WILLIAM LEE	14
MICHAEL WONG	14
MICHAEL BROWN	14
VICTOR LEE	12
ALAN WONG	11
JOHN LEE	11
VINCENT WONG	11
SANDY WONG	11
JOHN MILLER	11
THOMAS SMITH	10
STEPHEN LEE	10
ROBERT WONG	10

How often do names re-occur in the dataset? This is a question we might want to ask later.

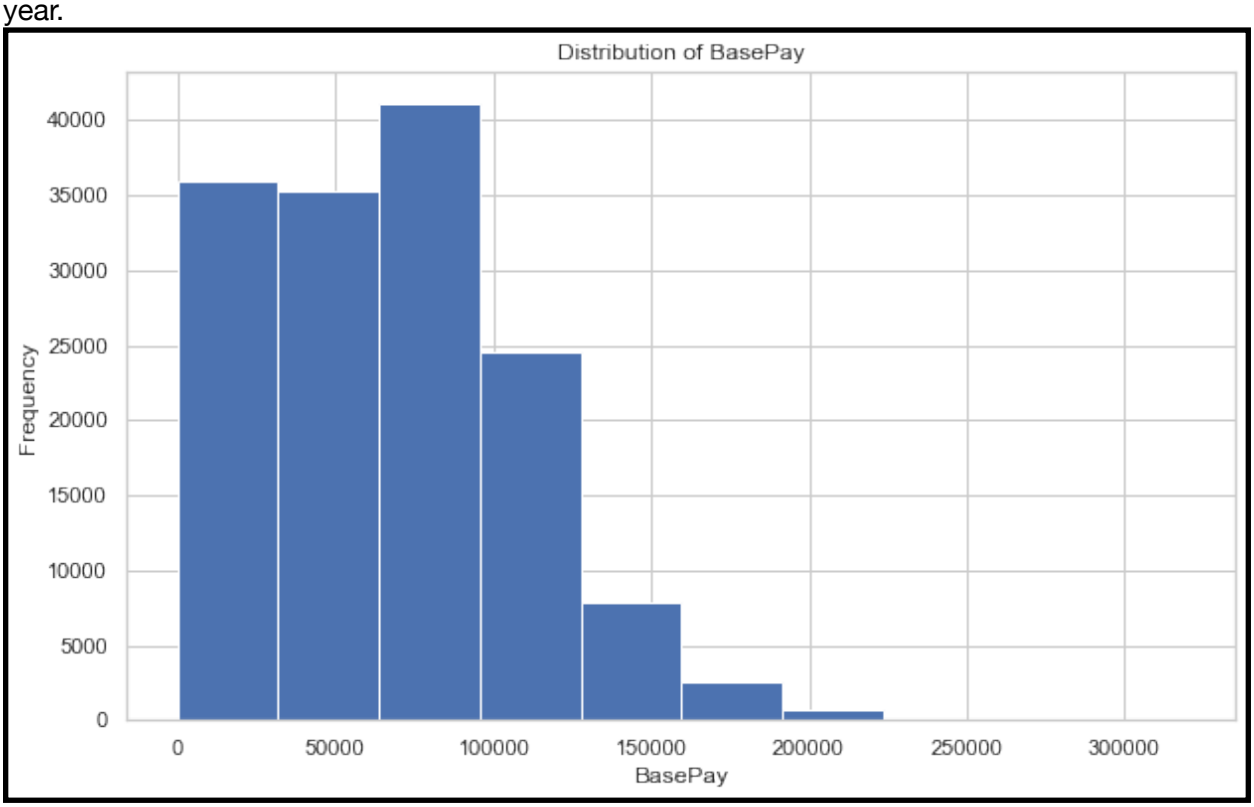
JobTitle
This is a categorical variable with a number of unique values. Let's take a look at the count for the top 20



It looks like the top three most common jobs employed by the city of San Francisco are Transit Operator, Special Nurse, and Registered Nurse.

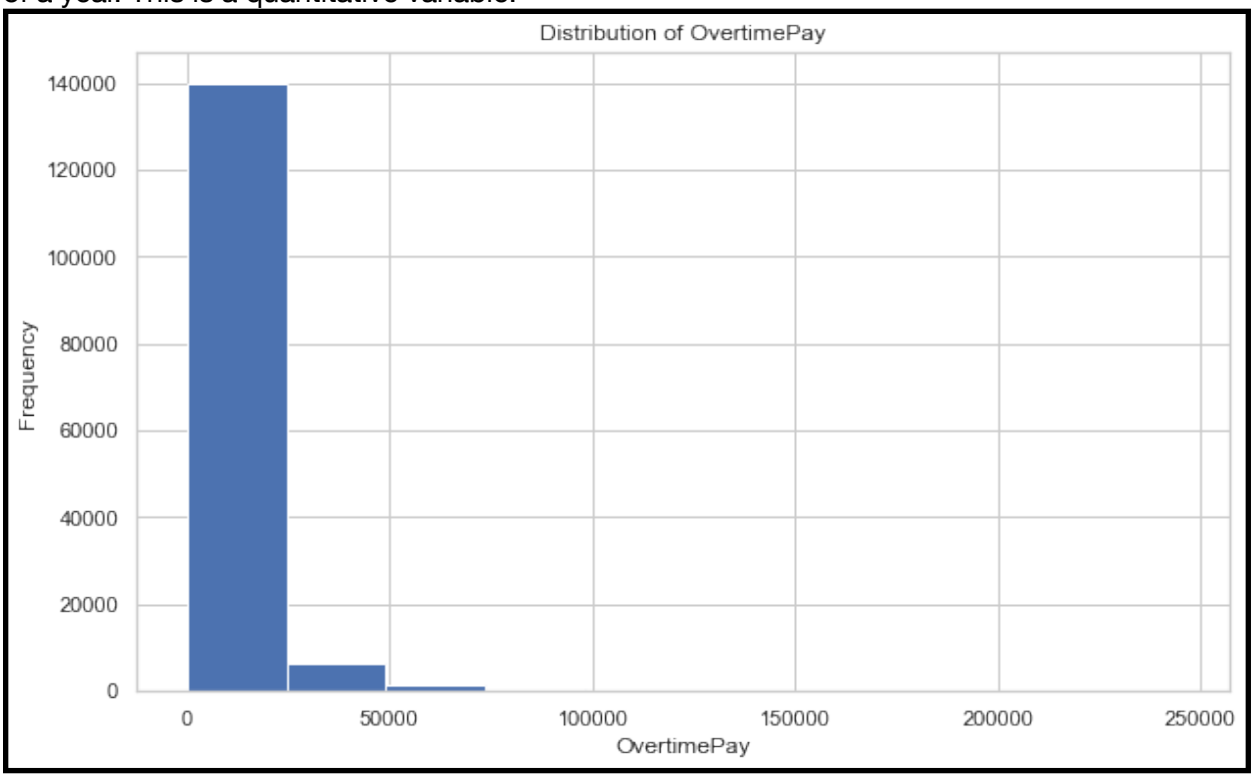
BasePay

I would guess that BasePay is the employee's total regular income from the city for the given year.



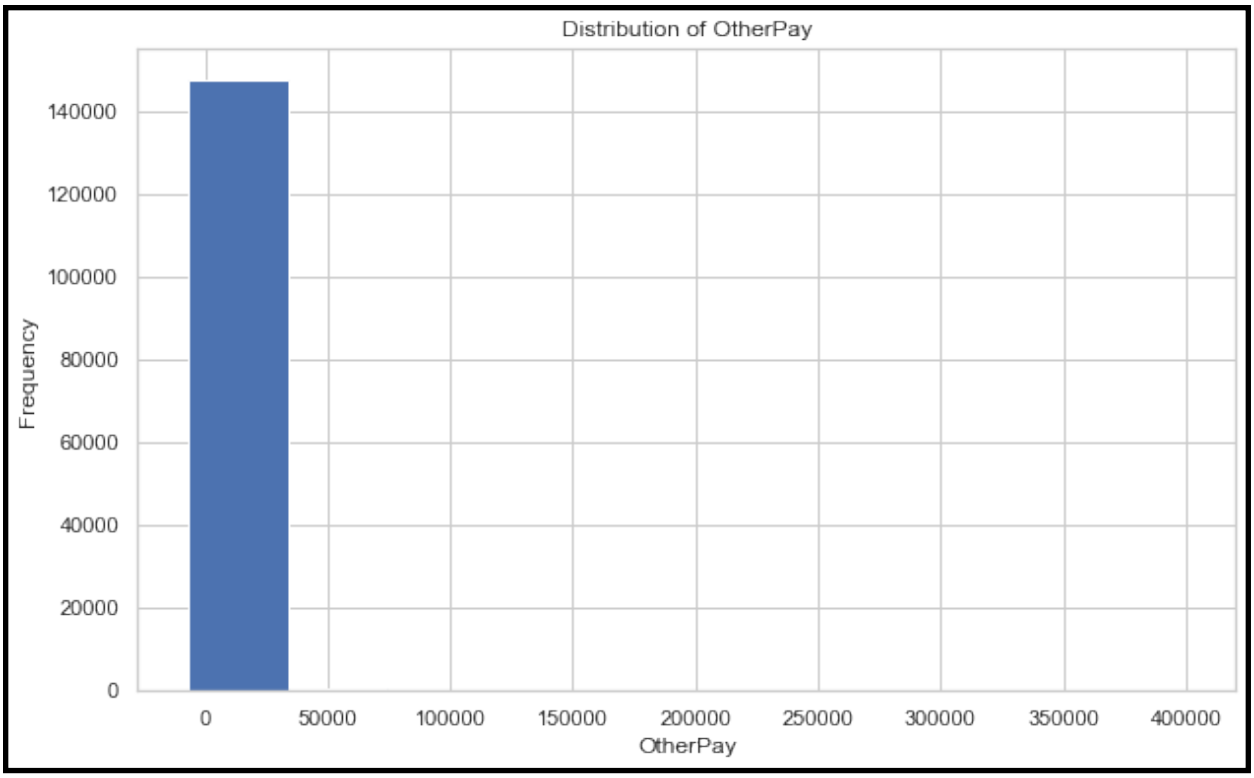
It looks like the majority of the employees are making somewhere between $50k and $100k. This makes sense, because the median is $65k. The number of employees who get paid more than $100k is rather small.

OvertimePay
OvertimePay is probably how much the employee was paid for overtime work over the course of a year. This is a quantitative variable.



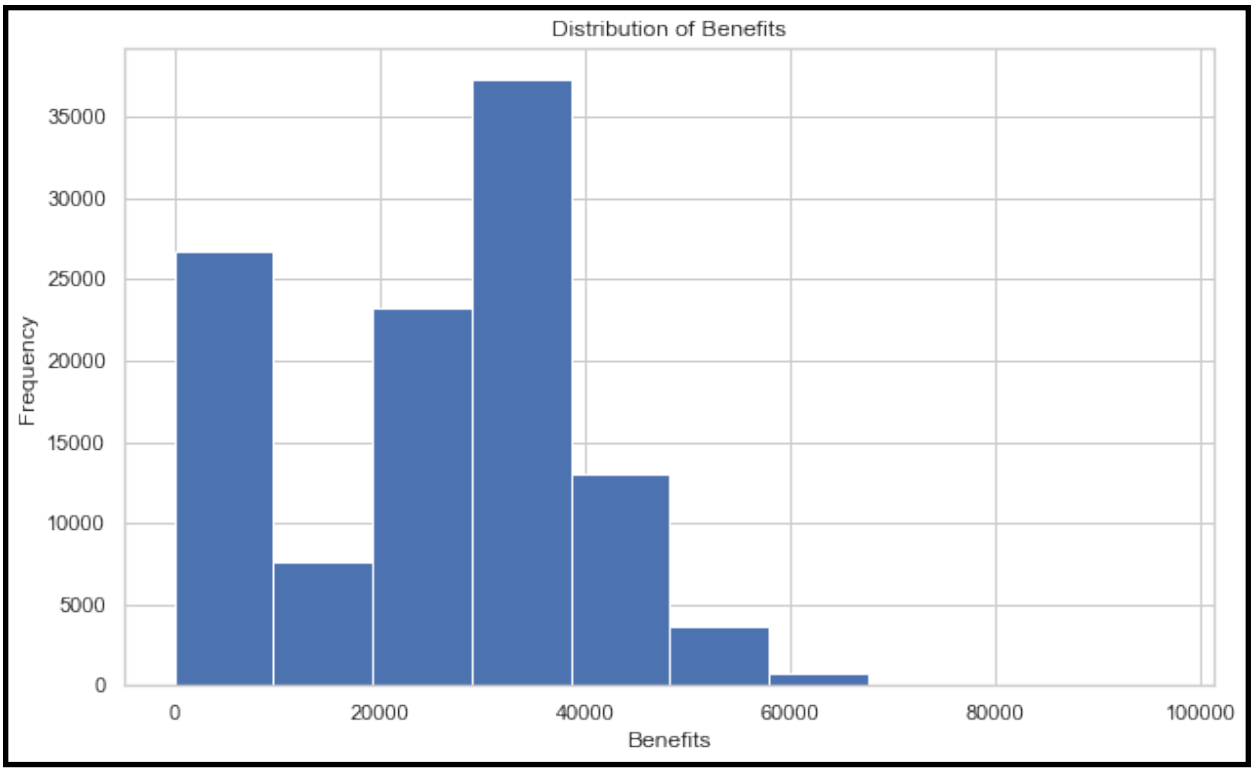
This shows that the vast majority of overtime pay did not amount to large sums. However, there were a few cases of people who were paid up to and over $200k

OtherPay
OtherPay is probably income that did not come from the city.



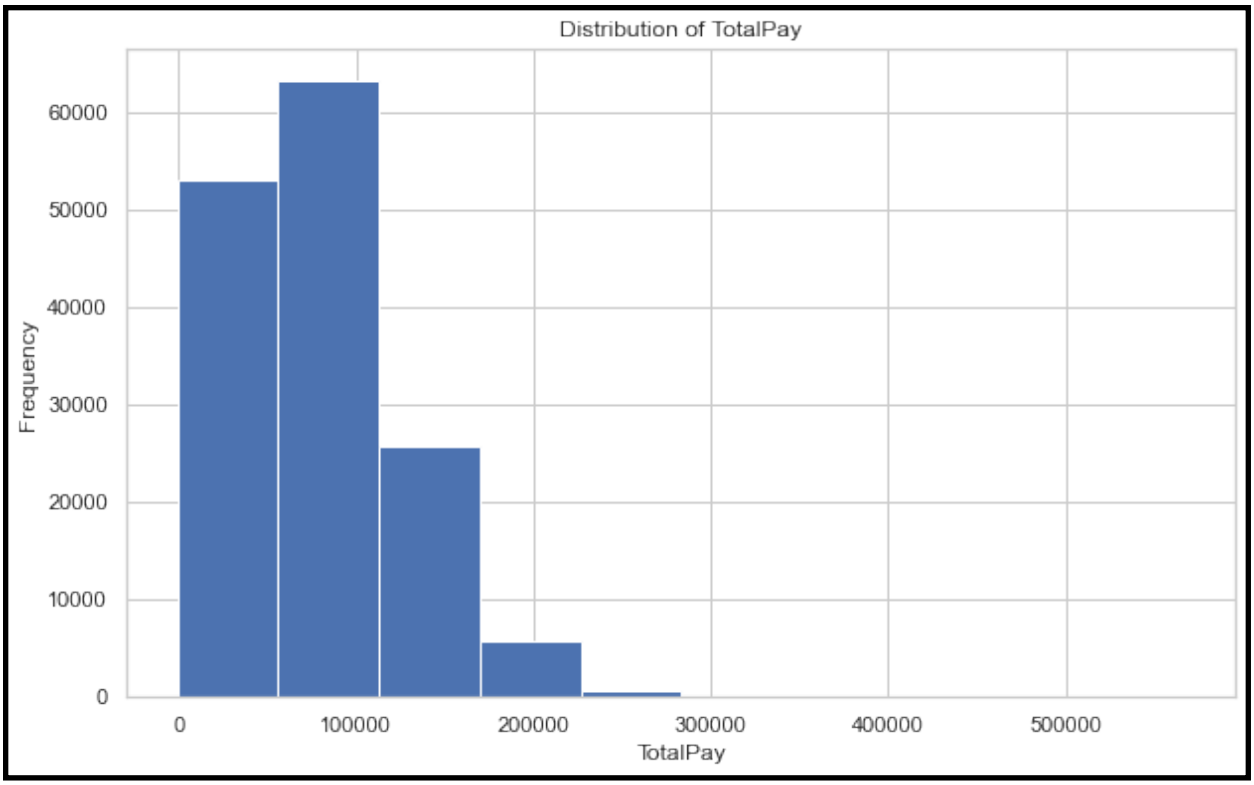
This shows that very few people make much money outside of their role as a city employee. And once again, there are a very few people who made a considerable amount of money outside of their job for San Francisco.

Benefits
Benefits is likely the monetary value of benefits that each employee received.



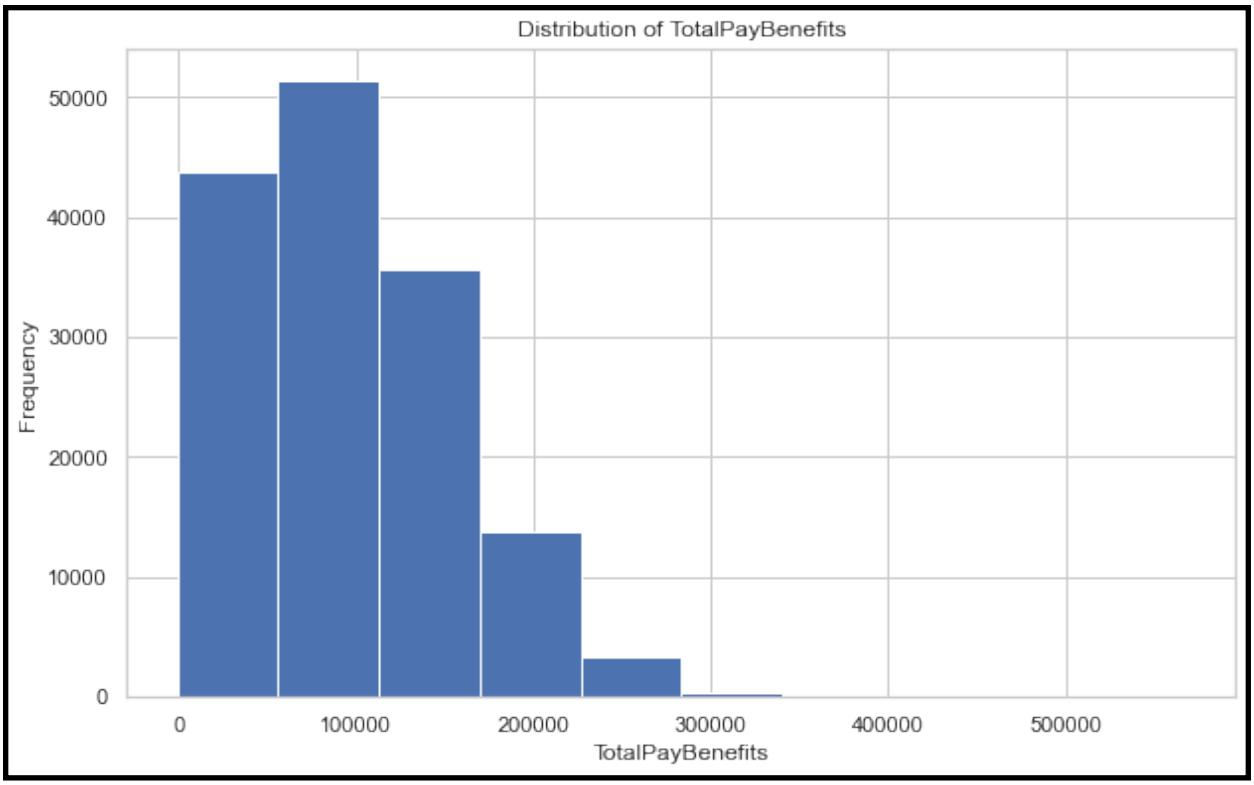
This shows a bimodal distribution. There are some people who receive <$5k, and a larger number of people who receive around $30k in benefits.

TotalPay
Total Pay is probably the sum of BasePay, OvertimePay, and OtherPay. Like those, this is a quantitative variable.



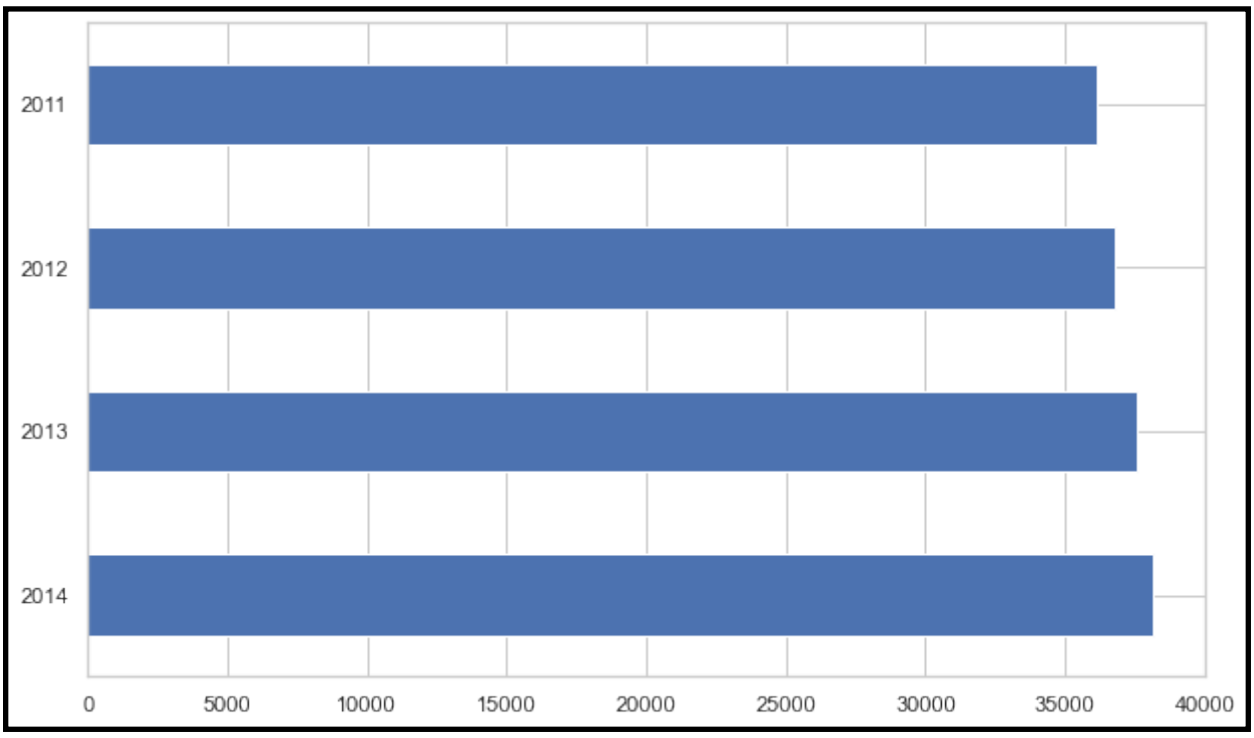
It seems like the majority of people make <$100k, and some few make up to and over $500k. This also looks a bit like a lognormal distribution

TotalPayBenefits
TotalPayBenefits is probably the sum of TotalPay and Benefits. It is a quantitative variable.



Again, this looks somewhat lognormal. The scale is shifted slightly to the right as well, which is as expected.

Year
Year is the year for which a row of data is recorded.

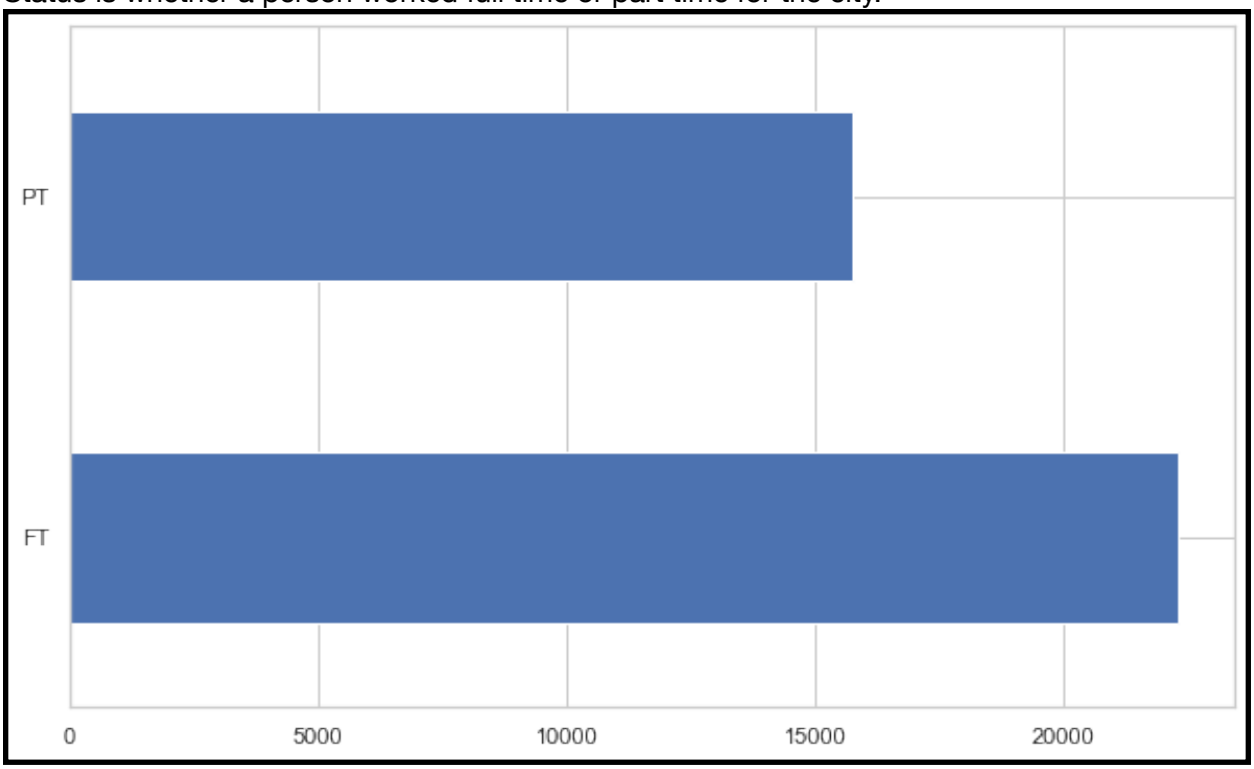


Here we can see that the city of San Francisco hired successively more people year on year from 2011 to 2014. The changes were not large though.

Notes
In this case, there are only NaN values, representing no notes, in all cases.

Agency
In this case, since everyone worked for the city of San Francisco, that is listed in all cases.

Status
Status is whether a person worked full time or part time for the city.



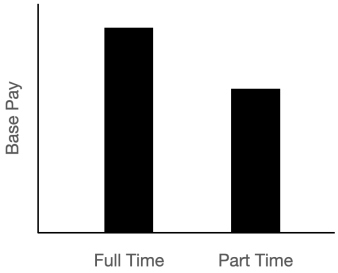
It looks like there were more people who worked for the city full time than part time. However, it is clear that part time workers make up a sizable portion.

List Five Analytical Questions that Users Might be Wondering

1. How does full time employee pay compare with part time employee pay?
2. How does full time employee benefits compare with part time employee benefits?
3. Which jobs get the most overtime pay?
4. Which jobs get the most benefits?
5. Which jobs have the greatest disparities between base pay and benefits?

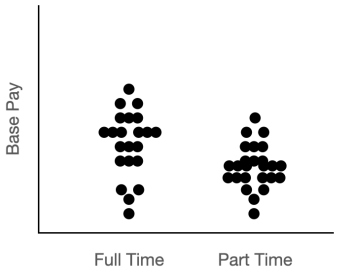
Select three of the five questions above, and create three visualizations for each question selected using tools such as PowerPoint, Excel, etc. Write one paragraph for each drawing explaining the design, purpose, and logic behind the design.

1. How does full time employee pay compare with part time employee pay?
    1. Bar chart of mean base pay for the two categories



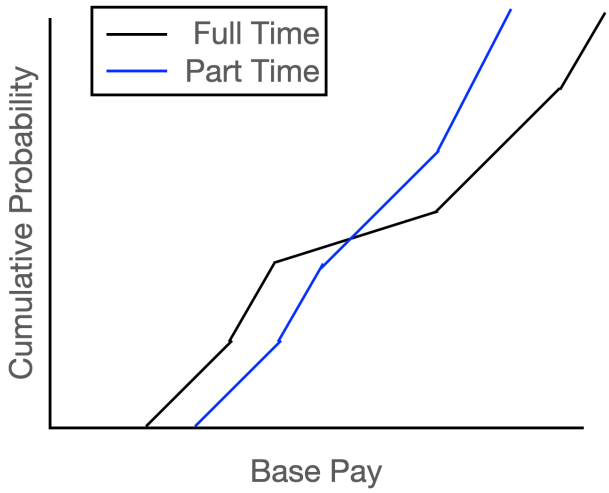
- This is designed to be as simple as possible: just two bars that are easily comparable. The purpose is to make starkly clear where the mean value for the two categories of workers is. Line length is something that humans are good at decoding into quantitative value, so this will be an easy chart for the viewer to understand.

2. Bee swarm of base pay for the two categories



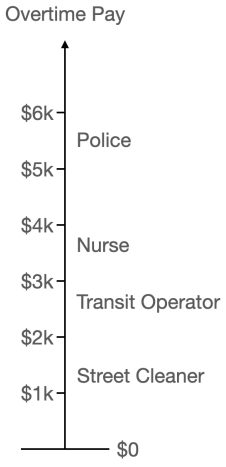
- This chart will show all of the data points for each category. Doing it in this fashion allows the viewer to see all the points, and get an intuitive feel for where points fall on the scale. The previous plot showed only the mean value, whereas with this, no data point is hidden.

3. Empirical cumulative distribution function (ECDF)



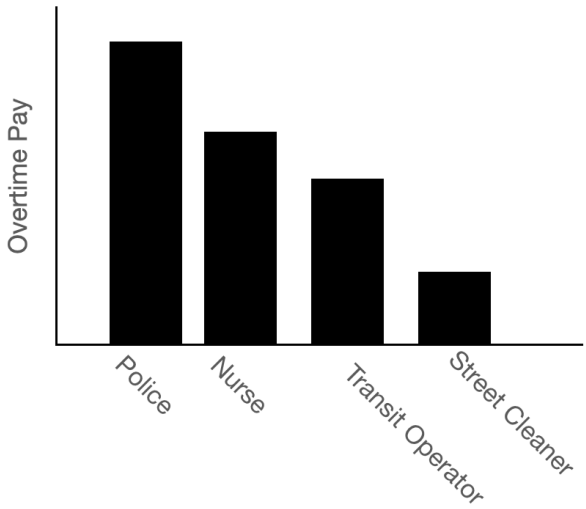
- Similar to the bee swarm plot, the point is show the user all of the data. This plot is very useful for getting statistical information in an easy manner. From it, the viewer can easily find information like maximum, minimum, and any percentile that is resolvable on the scale.

2. Which jobs get the most overtime pay?
    1. A line with job labels, sorted from most overtime to least overtime, using median value



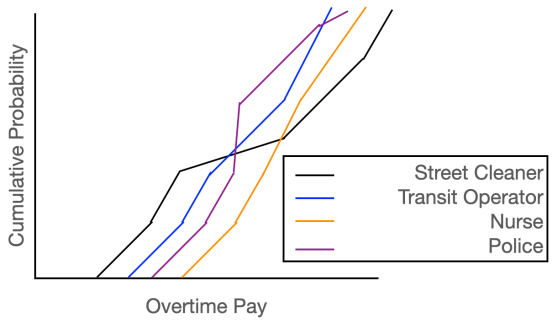
- This single line shows the amount of overtime pay on a single linear scale. This is very similar to a bar chart where all the bars are reduced into a single bar, and a label is applied to the location of each former bar's end. Humans are pretty good at interpreting position, so seeing all the jobs on a single scale should be pretty good at representing the data in a graphical fashion.

    2. A bar chart of median overtime pay as the length of the bar, and each bar represents one job type



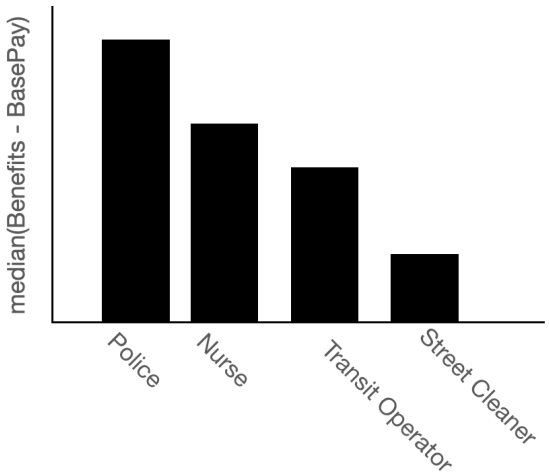
- This is a classical bar chart using length to encode quantitative information. The bars are ranked from largest to smallest, perhaps skipping most of the ones in the middle, showing only the largest and smallest. This design makes it easy to compare relative scale of one job to another.

3. Empirical cumulative distribution function (ECDF) plot, where each line represents one job.



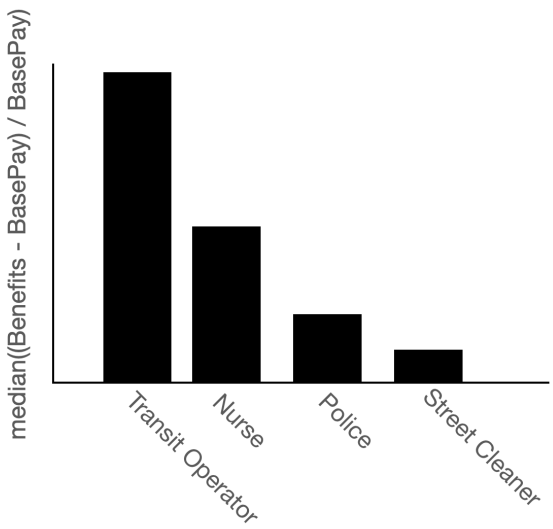
- In this ECDF, I image that the majority of lines are mostly transparent and gray, with a few highlighted and labeled in other colors. The ECDF format makes it easy to spot disparity within jobs: is a curve bifurcated? Because it uses linear scale on the x-axis, it is also easy to compare certain percentiles to one another between jobs.

3. Which jobs have the greatest disparities between base pay and benefits?
    1. A bar chart of median numerical difference, where each bar represents one job.



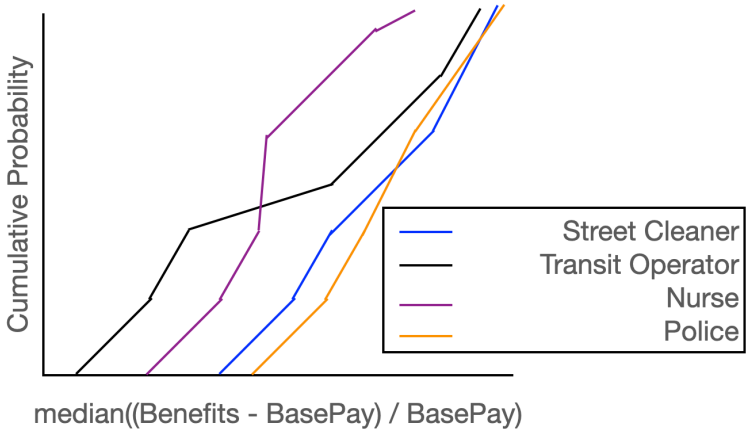
- This bar chart has median(Benefits - BasePay) as bar length, with one bar for each job. By using the raw difference in pay, the viewer get's a good sense of the raw dollar difference in each job.

2. A bar chart of median percent difference, where each bar represents one job.



- This chart is almost the same as the one before, but now we are looking at the difference as a percent of BasePay. This would be calculated as median((Benefits - BasePay) / BasePay). Showing the data in this form emphasizes the relative helpfulness of benefits to the employee. Employees with smaller base pays will be helped much more by a given amount of benefits than an employee with a high base pay.

3. An ECDF of percent difference, where each line represents one job.



- Instead of only showing a single bar of percent difference, the viewer can now see all of the data points. This makes more obvious any disparities within jobs. It also shows how job A might have a higher mean than job B, but the vast majority job B is in fact higher than that of job A.