IST652 Scripting for Data Analysis
Fall 2020
12/4/2020

School of Information Studies
SYRACUSE UNIVERSITY

IST 652 Final Project Report

Submitted by:
Xiwei Shen | Sihan Yang

Introduction

The dataset we are using for this final project is called the “MCG Personnel Management Review (PMR)” dataset, which contains a summary of the County Government and composition by generational category, age, race, ethnicity, gender, years of service, and job class. The dataset contains nine variables and 9244 entries in total. The source of our dataset is: https://catalog.data.gov/dataset/mcg-personnel-management-review-pmr. The secondary dataset we are using is called "Average Salary by Job Classification", which can be found at: https://catalog.data.gov/dataset/average-salary-by-job-classification. This dataset contains five variables, which are position title, position class code, grade, average of base salary, and the number of employees. The supplement dataset contains the average salary by position title and grade for full-time regular employees. It is like a summary of our original dataset. For this project, we are planning to first use the descriptive analysis to summarize our data and find interesting patterns. For categorical variables, we want to regroup the data in a different unit of analysis than is present in the original dataset. For this step, we can use pivot tables and plots to interpret our analysis results. At last, we want to try to conduct a regression analysis to analyze the relationship between each variable and the salary level.

Data Information and Preprocessing

Our dataset contains nine columns in total, which are "Generation", "Age", "Ethnic Origin", "Gender", "Length Of Service", "Job Class", "Grade", "Assignment Category" and "Salary Range". The "Generation" column contains the generational category of an employee and there are five unique values in total, which are "Generation X", "Baby Boomers", "Millennial Generation", "Traditionalist/Silent Generation" and "Post Millennials". The "Ethnic Origin" column contains the race of an employee and there are eight unique values in total, which are "White", "Hispanic or Latino", "Black or African American", "Asian", "Unreported", "Two or More Races", "American Indian or Alaska Native" and "Native Hawaiian/Other Pacific Islander". The "Salary Range" column splits the annual salary of an employee into 16 different groups. The lowest annual salary is less than 20K, and the highest annual salary is greater than 150K. For the data preprocessing steps, we first checked the missing values in our dataset. We found that there are no missing values in our dataset. Next, we add a dummy column in our dataset which contains all values 1. This column is very useful when using pivot tables to perform our analysis of different categories of data. We can use this column in the pivot table to count the number of categorical variables.

Data Analysis

For this project, we first conduct our analysis based on the salary range column of our dataset. We first plot the number of people in each salary range. Based on the plot (Figure 1), we can see that the majority of people have a salary of 50K-80K. There are only around 200 people who have a salary greater than 150K, and there are only around 100 people who have a salary less than 20K.

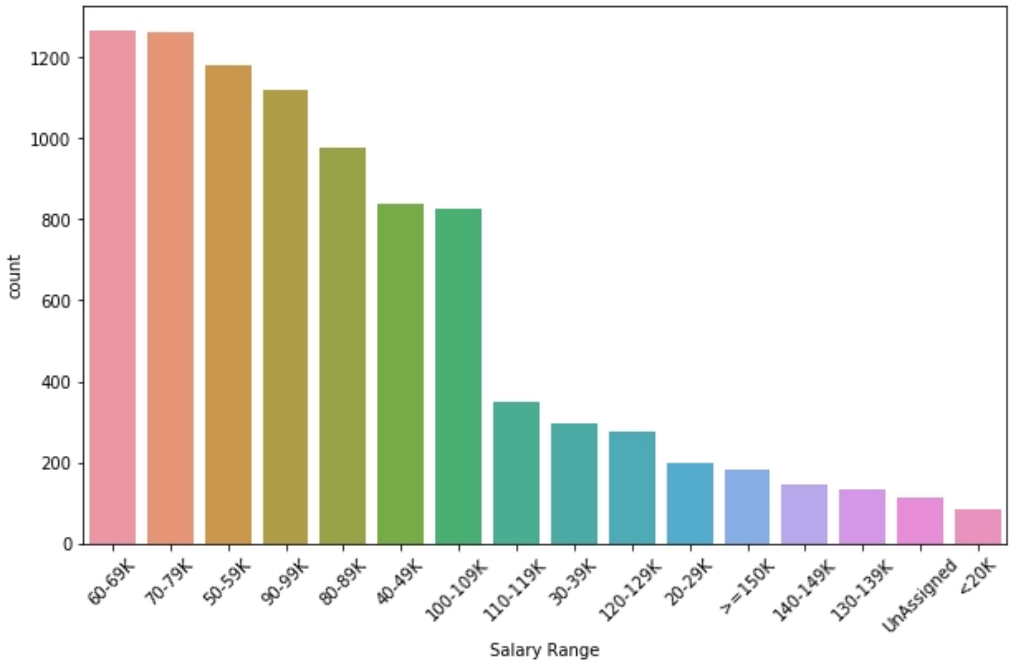

Figure. 1

Next, we focused on the Ethnic Origin variable of our dataset. We check the distribution of the Ethnic Origin variable through a bar plot.

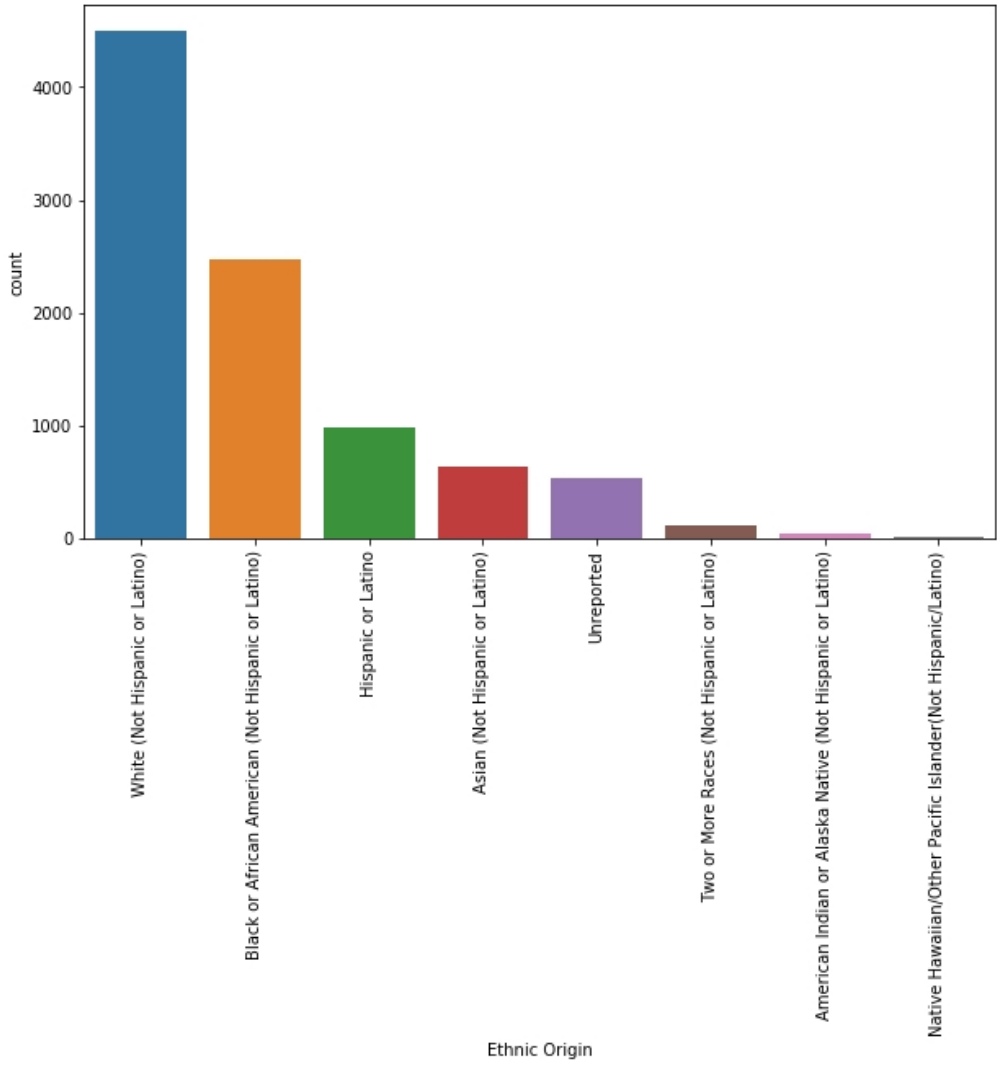

Figure. 2

Based on the plot (Figure 2), we can see that the majority of the employees are white origin and black origin. Other groups of people, like Asian, only take a small proportion of the total employee.

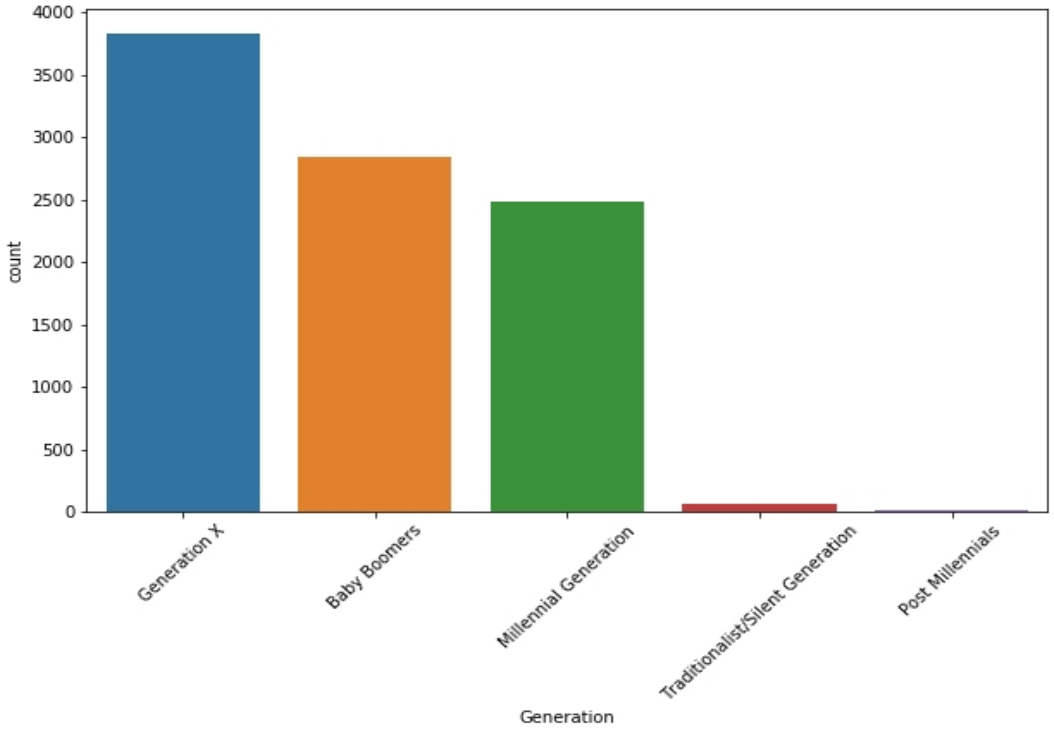

Figure. 3

Next, we plot the number of employees in each generational category (Figure 3). We found that the majority of employees were born during Generation X, which is from 1965 to1980, and Baby Boomers, which is from 1946 to 1964, and Millennials, which is from 1981 to 1996.

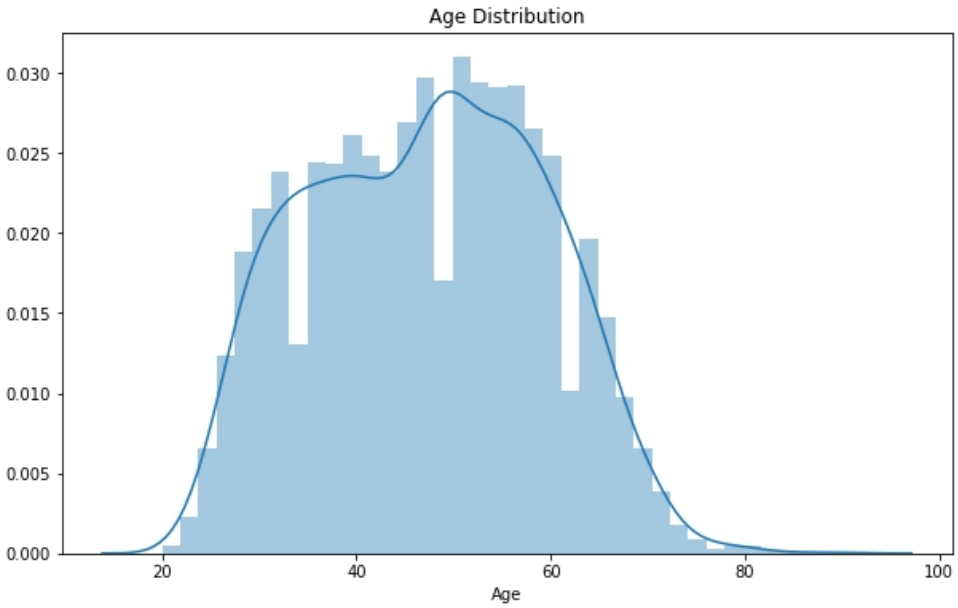

Figure. 4

The field we analyzed next is age. We draw a density plot to check the age distribution of employees (Figure 4). We can see that the age distribution of our dataset roughly follows a normal distribution. Most employees were included in the age range from 20 to 80. Age range from 50 to 60 seems to contain the largest number of employees.

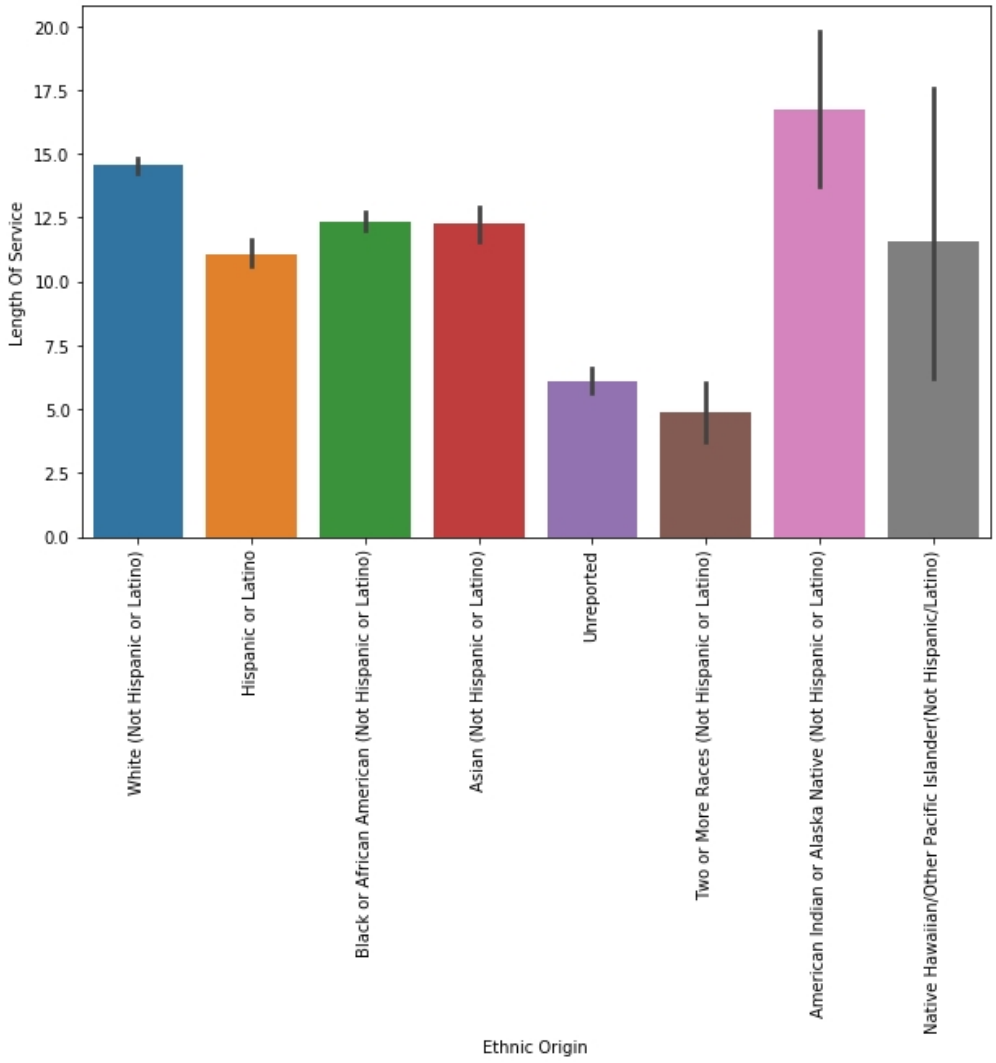

Figure. 5

The next fields we analyzed are ethnic origin and length of service (Figure 5). Since the length of service is a numerical attribute, we computed its average value for each ethnic group. We can see that the employees of American Indian or Alaska Native have the longest average length of services then followed by employees of White origin, while employees of Two or More Races have the shortest average length of services. For other ethnic origins, the average length of services is similar.

Gender	Female	Male
Ethnic Origin		
American Indian or Alaska Native (Not Hispanic or Latino)	46.666667	45.080000
Asian (Not Hispanic or Latino)	51.801205	48.572391
Black or African American (Not Hispanic or Latino)	49.185756	48.546154
Hispanic or Latino	46.193684	43.907445
Native Hawaiian/Other Pacific Islander(Not Hispanic/Latino)	43.000000	45.375000
Two or More Races (Not Hispanic or Latino)	39.407407	36.020000
Unreported	43.411765	40.816199
White (Not Hispanic or Latino)	51.176434	45.040972

Figure. 6

Next, we use the pivot table to conduct our analysis of fields age, gender, and ethnic origin (Figure 6). For each ethnic origin, we compute the average age of female employees and male employees. Basically, for every ethnic origin, the average age of female employees is greater than the average age of male employees, except for Native Hawaiian/Other Pacific Islander. Asian employees are the oldest one, while employees of Two or More Races are the youngest group.

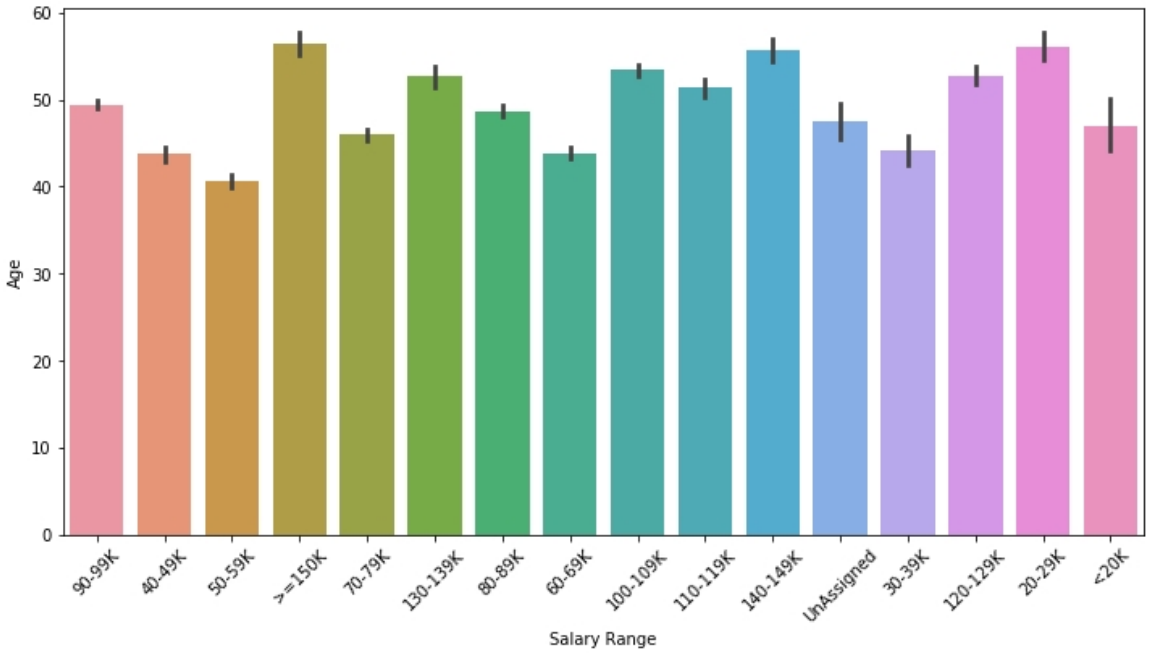

Figure. 7

Next, we use the bar plot to conduct our analysis of fields age, and salary range. We plot the average age for each salary range (Figure 7). We were expecting that older people have a higher salary, but surprisingly, the difference in average age among each salary range is not significant. It is true that the average age of people who have a salary greater than 150K is nearly 60. But if we look at the people who have a salary less than 20K, and within 20-29K, their average age is also very high.

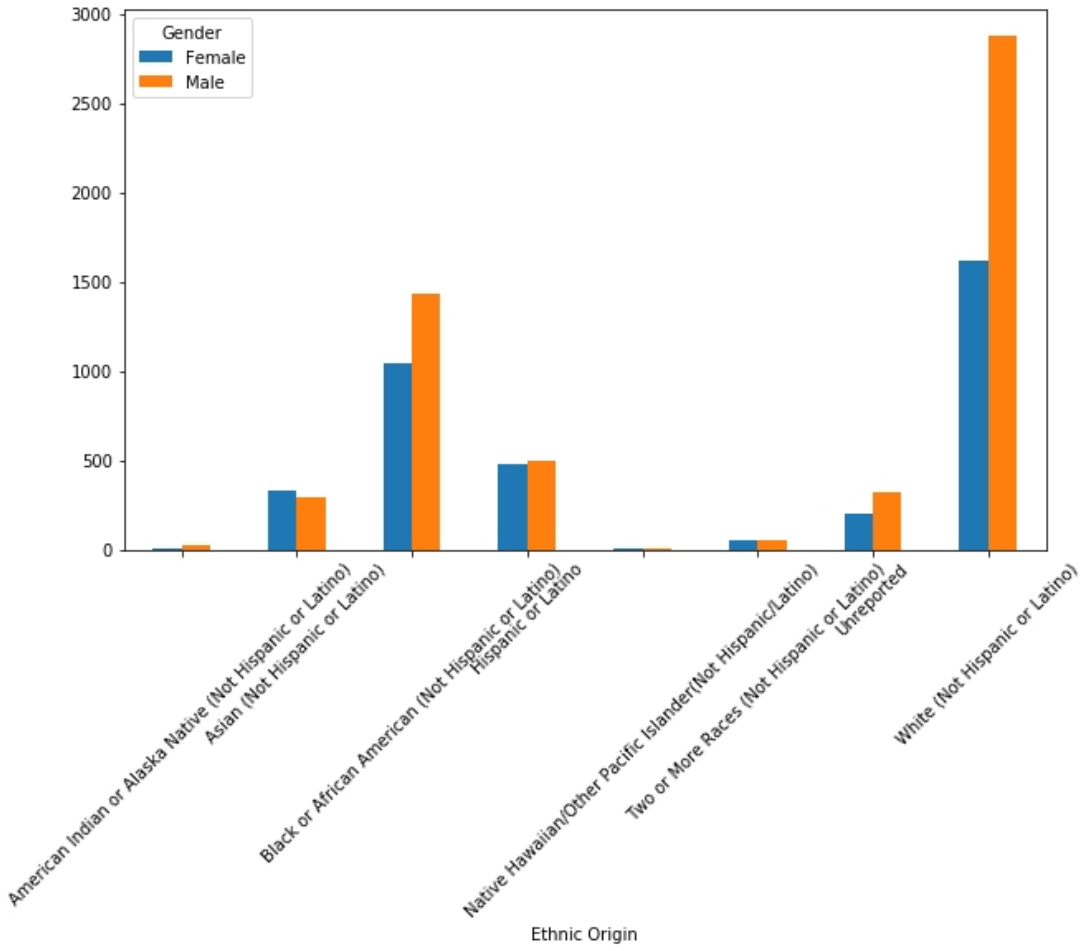

Figure. 8

The next analysis we did is based on gender and ethnic origin fields. We use the pivot table to get the distribution of gender for each ethnic group, store the result into a data frame, and then display the output (Figure 8). We can see that for two major ethnic origins, White and Black or African American, there are more male employees than female employees. For Asian and Hispanic or Latino, the number of male employees and female employees is overall balanced.

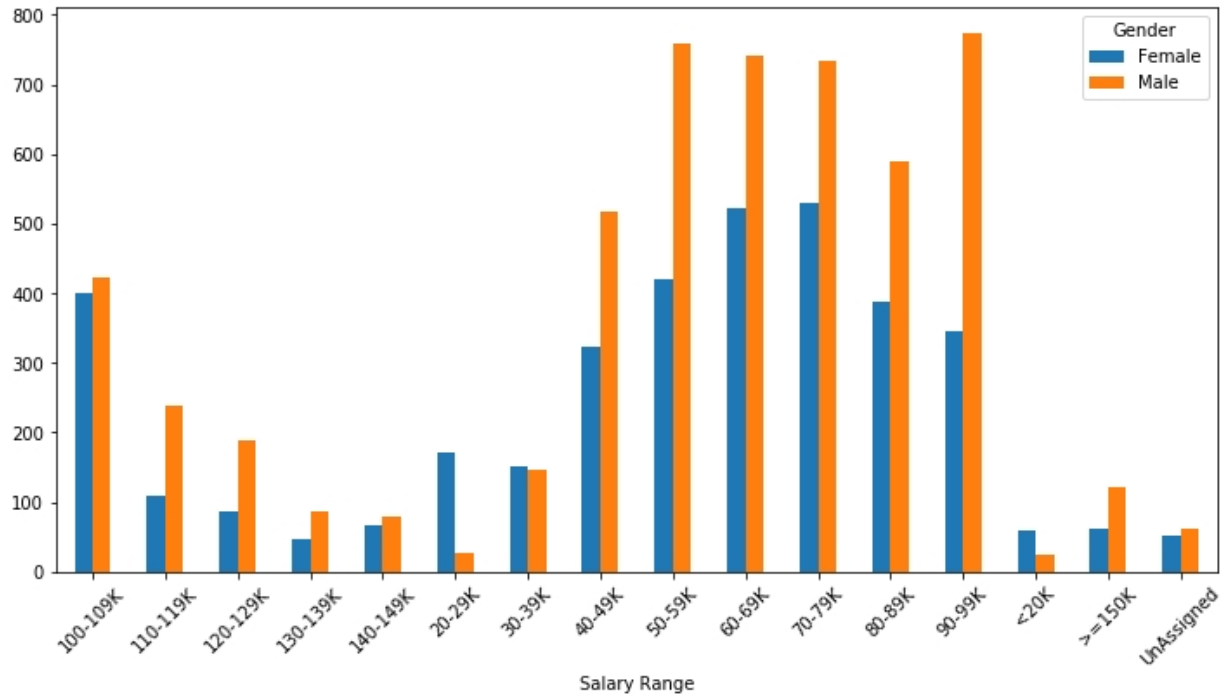

Figure. 9

Next, we follow the same approach to analyze the gender and salary range fields (Figure 9). We can see that for the majority of the salary range, the number of male employees is greater than the number of female employees. But for the salary range of 30-39K, 100-109K and 140-149K, the number of male and female employees are relatively balanced. For the salary range of 20-29K, the number of female employees is much greater than the number of male employees.

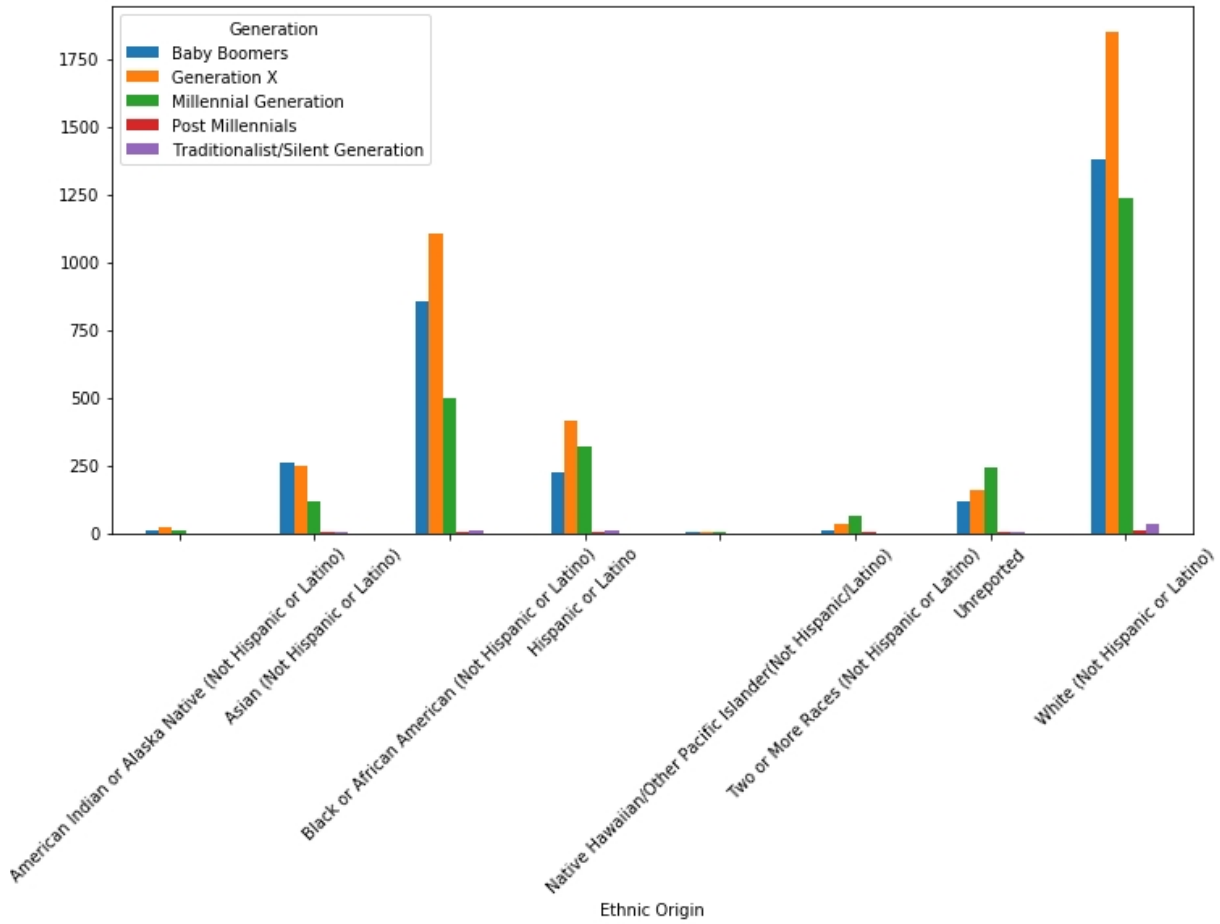

Figure. 10

We then use the pivot table to analyze the generation and ethnic origin fields and then plot the result data frame (Figure 10). For two major ethnic origins, White and Black or African American, most of the employees were born during Generation X, then followed by Baby Boomers. For Hispanic or Latino, most of the employees were born during Generation X, then followed by Millennials. For Asians, the number of employees that were born during Baby Boomers and Generation X is almost the same.

In addition, we want to separate our data into different categories and provide summary statistics on those categories. Therefore, we defined a function to separate all data into different categories based on the salary range field of the dataset. If an employee has a salary lower than 60K, the person will be classified as a "Low Salary" employee. If an employee has a salary between 60K and 110K, the person will be classified as a "Medium Salary" employee. Employees with a salary greater than 110K will be classified as "High Salary" employees. We then added a new column to the data frame for this income category field.

	Salary_category	Average_age
0	Low Salary	43.406623
1	Medium Salary	47.814094
2	High Salary	53.311808

Figure. 11

The first analysis we did is to check the average age of employees in each salary category (Figure 11). Here the trend is clearer and more reasonable. Older employees tend to have a higher salary.

Next, we focused on the ethnic origin attribute. For each salary category, we would like to find out the most frequent ethnic origin of that category (Figure 12). Since we already have three sub data frames for three salary categories, simply counting the number of ethnic origin variables will provide us the desired output.

	Most Common Ethnic Origin of Low Salary Group	Most Common Ethnic Origin of Medium Salary Group	Most Common Ethnic Origin of High Salary Group
0	Black or African American (Not Hispanic or Lat...	White (Not Hispanic or Latino)	White (Not Hispanic or Latino)
1	White (Not Hispanic or Latino)	Black or African American (Not Hispanic or Lat...	Black or African American (Not Hispanic or Lat...
2	Hispanic or Latino	Hispanic or Latino	Asian (Not Hispanic or Latino)
3	Unreported	Asian (Not Hispanic or Latino)	Hispanic or Latino
4	Asian (Not Hispanic or Latino)	Unreported	Unreported
5	Two or More Races (Not Hispanic or Latino)	Two or More Races (Not Hispanic or Latino)	American Indian or Alaska Native (Not Hispanic...
6	American Indian or Alaska Native (Not Hispanic...	American Indian or Alaska Native (Not Hispanic...	Two or More Races (Not Hispanic or Latino)
7	Native Hawaiian/Other Pacific Islander(Not His...	Native Hawaiian/Other Pacific Islander(Not His...	Native Hawaiian/Other Pacific Islander(Not His...

Figure. 12

We can see that there are some differences between different income categories. Black and African American employees are the most common one in the low salary group, while White employees are the most common one in the medium and high salary group. Employees of these

two ethnic origins are the most common two in all salary groups due to the large amount of data points. The number of Asian employees rises up from low salary group to high salary group.

Based on the same logic, we then focused on the generation attribute. We want to find out the most frequent generational category of each income category (Figure 13).

	Most Common Generation of Low Salary Group	Most Common Generation of Medium Salary Group	Most Common Generation of High Salary Group
0	Millennial Generation	Generation X	Generation X
1	Generation X	Baby Boomers	Baby Boomers
2	Baby Boomers	Millennial Generation	Millennial Generation
3	Post Millennials	Traditionalist/Silent Generation	Traditionalist/Silent Generation

Figure. 13

The most common generational category for the low salary group is Millennials. For the medium and high salary group, the most common generational category is Generation X, and the Baby Boomers category is more common than the low salary group.

Analysis of Second Dataset

Next, we conducted some analysis on our supplement dataset. We were trying to combine these two datasets together since the datasets that we used are from the same source. But since the second dataset only contains general information of our original dataset and there is no specific information of the employee, such as gender and age. Therefore, we are not able to merge two datasets together. Instead, we just performed some analysis on these two datasets separately. Our secondary dataset contains position title, position class code, job grade, average of base salary for each position, and number of employees for each position.

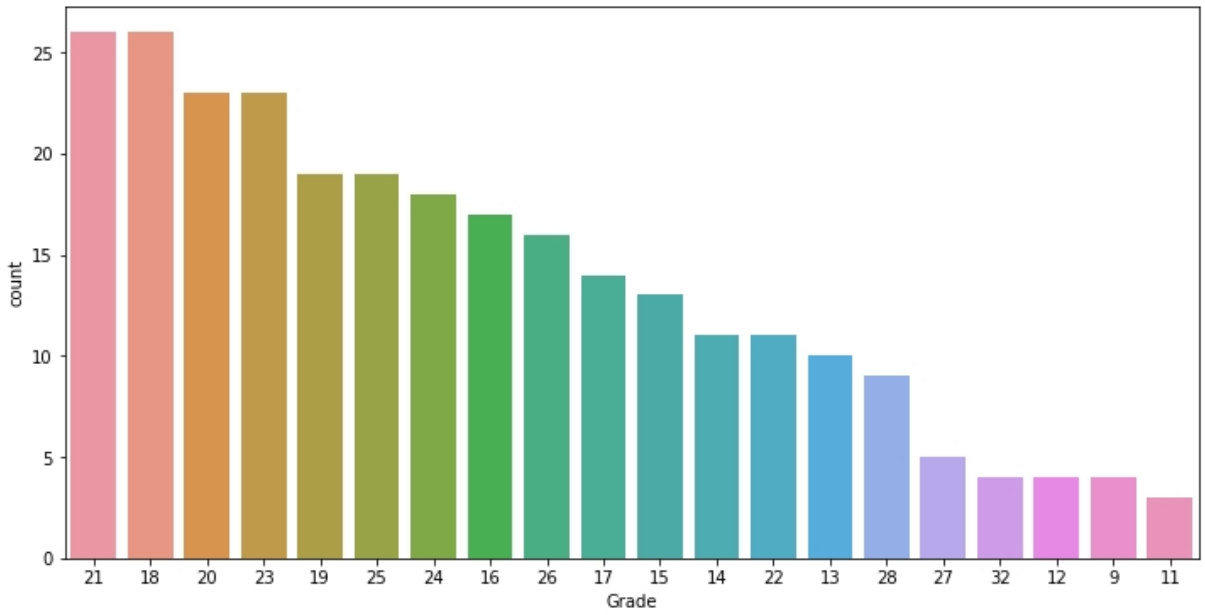

Figure. 14

We first checked the distribution of the job grade variable by using a plot (Figure 14). Based on our research, the higher the job grade value, the higher the position level. Based on the plot, we can see that the majority of the positions are in the medium level. The high level and low level positions only take a small proportion of the total position.

	Position Title	Number of Employees
158	Police Officer III	811
238	Transit Bus Operator	623
307	Firefighter/Rescuer III	409
319	Manager III	252
74	Office Services Coordinator	210
171	Master Firefighter/Rescuer	210
268	Program Manager II	175
75	Police Sergeant	157
272	Community Health Nurse II	148
249	Correctional Officer III (Corporal)	142
229	Fire/Rescue Captain	142
77	Income Assistance Program Specialist II	129
239	Fire/Rescue Lieutenant	121
91	Program Specialist II	115
28	Manager II	115
164	Principal Administrative Aide	101
188	Equipment Operator I	96
140	Mechanic Technician II	95
167	Social Worker III	92
58	Therapist II	85

Figure. 15

Next, we want to find out the top 20 positions that have the largest number of employees. We just select the position title and number of employee columns from the dataframe and sort the data by number of employees (Figure 15). We can see that the police officer, bus operator and firefighter are the positions that have the largest number of employees. Some other positions, such as nurse and program manager, also have a large number of employees.

	Position Title	Average of Base Salary
278	Medical Doctor Psychiatrist IV	223953.00
155	Medical Doctor Psychiatrist III	203593.00
9	Manager I	182281.86
176	Information Technology Project Manager	182075.00
131	Senior Investment Officer	171350.49
212	Fire/Rescue Division Chief	168617.00
220	Chief Veterinarian	162707.55
28	Manager II	160257.48
258	Enterprise Technology Expert	159475.00
284	Chief Deputy Sheriff (Colonel)	159332.00
317	Police Captain	156780.64
117	Public Health Dentist	153375.00
259	Fire/Rescue Assistant Chief	148712.00
0	Information Technology Expert	146925.34
55	Psychologist Supervisor	145122.13
111	Senior ERP Functional Business Analyst	143615.33
104	Police Lieutenant	138968.30
33	Assistant County Attorney III	138905.08
299	Information Technology Supervisor	138517.48
152	Fire/Rescue Battalion Chief	135022.66

Figure. 16

Then, we want to find out the top 20 positions that have the highest average base salary. We just select the position title and average base salary columns from the dataframe and sort the data by average base salary (Figure 16). Without surprise, medical doctors and various managers have the highest average base salary. Some other positions like the senior investment officer, fire division chief, and chief veterinarian also have a high enough average base salary.

Linear Regression

In this part, we want to understand the relationship between each column in this data set through linear regression analysis, understand the characteristics of all employees by analyzing their personal information, and then understand all the factors affecting the salary level of employees through linear regression.

In order to achieve this goal, we will carry out the following steps: first, introduce the data, convert it into a format capable of data analysis, then classify each column, and finally introduce the regression equation to understand its correlation.

1. Import data:

We decided to use the json format of data, from the website https://data.montgomerycountymd.gov/api/views/bedm-7sqa/rows.json?AccessType=DOWNLOAD

imports JSON-formatted data to Jupyter and then decodes JSON-formatted data to Jupyter through Python JSON packages.

```
In [105]:
#Import the required packages
import urllib
import json
import numpy as np
import statsmodels.api as sm
```

```
In [10]:
mcg_url = "https://data.montgomerycountymd.gov/api/views/bedm-7sqa/rows.json?accessType=DOWNLOAD"#paste the website and store as mcg_url

#decode the json file and store in jupyter
mcg = urllib.request.urlopen(mcg_url)
json_string = mcg.read().decode('utf-8')
mcg_parsed_json1 = json.loads(json_string)
mcg_data = mcg_parsed_json1['data']
#check how many rows it contains in this json file
len(mcg_data)
```

```
Out[10]: 9243
```

2. Transform data:

For the next step, we use the panda package to select nine useful columns from the JS-formatted data and convert them to panda Dataframe.

```
In [24]:
#use panda package to transform this json file into dataframe
import pandas as pd

mcgall = pd.DataFrame(mcg_data, columns=['1', '2', '3','4','5','6','7','8','9','10','11','12','13','14','15','16','17'])
```

```
In [38]:
#Since we only need those useful columns in our analysis, we keep those columns and store in mcg
mcg = mcgall[['9','10','11','12','13','14','15','16','17']]
mcg.columns = ['Generation', 'Age', 'Ethnic','Gender','Length_of_Service','Job_Class','Grade','Assignment_Category','Salary']
```

```
In [39]:
#check this dataframe
mcg
```

```
Out[39]:
```

	Generation	Age	Ethnic	Gender	Length_of_Service	Job_Class	Grade	Assignment_Category	Salary
0	Generation X	42	White (Not Hispanic or Latino)	Female	14	NON_MLS	25	Fulltime-Regular	90-99K
1	Generation X	44	Hispanic or Latino	Male	1	NON_MLS	15	Fulltime-Regular	40-49K
2	Generation X	41	Black or African American (Not Hispanic or Lat...	Female	11	NON_MLS	15	Fulltime-Regular	50-59K
3	Generation X	47	White (Not Hispanic or Latino)	Male	12	NON_MLS	15	Fulltime-Regular	50-59K
4	Baby Boomers	59	Black or African American (Not Hispanic or Lat...	Female	2	NON_MLS	14	Fulltime-Regular	40-49K
...	...	...	...	...	...	...	...	...	...
9238	Millennial Generation	29	White (Not Hispanic or Latino)	Male	5	NON_MLS	P4	Fulltime-Regular	60-69K
9239	Millennial Generation	29	White (Not Hispanic or Latino)	Male	5	NON_MLS	P4	Fulltime-Regular	60-69K
9240	Millennial Generation	29	White (Not Hispanic or Latino)	Male	5	NON_MLS	P4	Fulltime-Regular	60-69K
9241	Millennial Generation	29	White (Not Hispanic or Latino)	Male	5	NON_MLS	P4	Fulltime-Regular	60-69K
9242	Generation X	51	Hispanic or Latino	Male	23	NON_MLS	P4	Fulltime-Regular	90-99K

9243 rows × 9 columns

3. Create dummy variables:

In this dataset, we can easily notice that there are many columns that contains categorical variables, like ethnic and gender, in order to do the linear regression, we need to create dummy variables for those categorical variables:

```
In [40]:
#Because some of the columns only contains catagorical variables, we decided to create dummy variables for each catagorical variables
mcg = mcg.join(pd.get_dummies(mcg.Ethnic))
mcg = mcg.join(pd.get_dummies(mcg.Gender))
mcg = mcg.join(pd.get_dummies(mcg.Job_Class))
mcg = mcg.join(pd.get_dummies(mcg.Assignment_Category))
mcg = mcg.join(pd.get_dummies(mcg.Generation))
```

```
In [119]:
#check the forst ten rows to see if the dummy variables are successfully created
mcg.head(10)
```

Out[119]:

	Generation	Age	Ethnic	Gender	Length_of_Service	Job_Class	Grade	Assignment_Category	Salary	American Indian or Alaska Native (Not Hispanic or Latino)	...	MLS	NON_MLS	Fulltime-Regular	Part Re
0	Generation X	42	White (Not Hispanic or Latino)	Female	14	NON_MLS	25	Fulltime-Regular	90-99K	0	...	0	1	1	
1	Generation X	44	Hispanic or Latino	Male	1	NON_MLS	15	Fulltime-Regular	40-49K	0	...	0	1	1	
2	Generation X	41	Black or African American (Not Hispanic or Lat...	Female	11	NON_MLS	15	Fulltime-Regular	50-59K	0	...	0	1	1	
3	Generation X	47	White (Not Hispanic or Latino)	Male	12	NON_MLS	15	Fulltime-Regular	50-59K	0	...	0	1	1	
4	Baby Boomers	59	Black or African American (Not Hispanic	Female	2	NON_MLS	14	Fulltime-Regular	40-49K	0	...	0	1	1	

At the same time, when we check the column Grade, we can notice that this column has many variables that contain non-numeric variables, so we delete those variables and only use numerical variables.

```
In [56]:
#Because the Grade column has many variables that contain non-numeric variables
#so we delete those variables and only use numerical variables.

mcg = mcg[mcg.Grade.apply(lambda x: x.isnumeric())]
```

4. Create dependent and independent variable:

For this step, we use the average salary instead of salary range, assign average salary for each range, then. We use all numeric variables columns and dummy variables as independent variables, and average salary as dependent variables, creating two new dataframe for further use.

In [57]:
```
#we use a loop to get the mean value of Salary, then store in a new column called Average_Salary
result = []
for value in mcg["Salary"]:
    if value == "<20K":
        result.append(20000)
    elif value == "20-29K":
        result.append(25000)
    elif value == "30-39K":
        result.append(35000)
    elif value == "40-49K":
        result.append(45000)
    elif value == "50-59K":
        result.append(55000)
    elif value == "60-69K":
        result.append(65000)
    elif value == "70-79K":
        result.append(75000)
    elif value == "80-89K":
        result.append(85000)
    elif value == "90-99K":
        result.append(95000)
    elif value == "100-109K":
        result.append(105000)
    elif value == "110-119K":
        result.append(115000)
    elif value == "120-129K":
        result.append(125000)
    elif value == "130-139K":
        result.append(135000)
    elif value == "140-149K":
        result.append(145000)
    else:
        result.append(150000)


mcg["Average_Salary"] = result
mcg
```

Out[57]:

In [59]:
```
#we use all the variable in mcg except average salary as independent variable
#Since we already have all the dummy variable, so we delete those catagorical variables
independent=mcg.iloc[:, 1:28]
independent =independent.drop(labels='Ethnic',axis=1)
independent =independent.drop(labels='Gender',axis=1)
independent =independent.drop(labels='Job_Class',axis=1)
independent =independent.drop(labels='Assignment_Category',axis=1)
independent =independent.drop(labels='Salary',axis=1)
independent
```

Out[59]:

	Age	Length_of_Service	Grade	American Indian or Alaska Native (Not Hispanic or Latino)	Asian (Not Hispanic or Latino)	Black or African American (Not Hispanic or Latino)	Hispanic or Latino	Native Hawaiian/Other Pacific Islander(Not Hispanic/Latino)	Two or More Races (Not Hispanic or Latino)	Unreported	...	Male	MLS	NON_MLS	Fulltin Regu
0	42	14	25	0	0	0	0	0	0	0	...	0	0	1	
1	44	1	15	0	0	0	1	0	0	0	...	1	0	1	
2	41	11	15	0	0	1	0	0	0	0	...	0	0	1	
3	47	12	15	0	0	0	0	0	0	0	...	1	0	1	
4	59	2	14	0	0	1	0	0	0	0	...	0	0	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
9233	48	4	15	0	0	1	0	0	0	0		1	0	1	

In [62]:
```
#use Average_Salary column as the dependent variable
Dependent = mcg['Average_Salary']
Dependent
```

Out[62]:
```
0        95000
1        45000
2        55000
3        55000
4        45000
         ...
9233     45000
9234     45000
9235     75000
9236    125000
9237     65000
Name: Average_Salary, Length: 5882, dtype: int64
```

5. Create regression model:

In the last step, we apply those two dataframe into the regression model through sklearn.linear_model package, use all the numeric variables and dummy variables as independent variables, and average salary as dependent variables.

```
In [ ]:
#ues liner model from the sklearn package to do the regression model
from sklearn.linear_model import LinearRegression
model = LinearRegression().fit(independent,Dependent)
```

```
In [103]:
#check the result of the liner model
model.score(independent,Dependent)
```

```
Out[103]: 0.7094593259275835
```

```
In [131]:
#print out all the result and score of this model
r_sq = model.score(independent,Dependent)

print('coefficient of determination:', r_sq)

print('intercept:', model.intercept_)

print('coefficients:', model.coef_)
```

```
coefficient of determination: 0.7094593259275835
intercept: -14885.605307606325
coefficients: [ 1.19017663e+02  5.52944519e+02  3.74594457e+03 -5.71377788e+02
 -1.12176122e+03 -3.51521992e+03 -2.94743092e+03  1.17233533e+04
 -1.04424106e+03  7.34098546e+01 -2.59673227e+03 -8.86281128e+02
  8.86281128e+02 -1.13686838e-12  0.00000000e+00  7.73509678e+03
 -7.73509678e+03  2.41824802e+03  1.87340363e+03 -2.35148800e+03
  1.69561360e+03 -3.63577724e+03]
```

```
In [ ]:
```

Through the linear regression analysis, we obtained a regression equation, and through the analysis of this equation, we obtained a score, namely 0.7094, which is relatively not very high, so this data set is actually not very suitable for the prediction by regression equation.

Conclusion

The goal and questions that we listed in our project proposal were all properly solved by our analysis work. We do get some insights of our dataset based on our analysis results. We also conducted the linear regression analysis on our dataset, but since the dataset contains many categorical variables, linear regression might not be an optimal choice. We finished this project by a group of two. About the work distribution, Sihan Yang did the works that related to the regression analysis and I did all the rest of the works, and we finished the report together. Our team collaborated well on this project.