Exploratory Analysis (NYC Job Openings)

Introduction

Brief description about the topic and the project.

1. How does your project help those who are interested in this topic?

Ans.1

Our project completely revolves around the NYC jobs. So, anyone graduating or seeking for jobs in any technical or non-technical field can look for the job opportunities in NYC, and can easily find the job based on the work location and the type of job - IT or Non-IT. Moreover, they can see the total number of opening in every field. They can also go through the salary structure both maximum and minimum salary offered on annual, daily, hourly basis.

2. Why data analysis is needed for this?

Ans.2

As our data is about the job opening in NYC boroughs, data analysis is necessary for easy approach to every possible detail of data. Data Analysis helps to extract information and gives certain pattern of data that is exploited in certain fruitful decisions. Data analysis is needed for the following reasons-

1. Inorder to filter the job categories based on Technical and Nontechnical fields in contrast to the job locations so as to get the number of part time and full time.

2. As our data has the details of salary on annual, daily and hourly basis, we analyze the data inorder to get the average maximum and minimum salaries in each location for a better understanding.

3. Inorder to get that which work Location has the most job openings provided with salary we need data analysis.

4. It gives information about minimum and maximum average salary range in IT and Non-IT fields according to location. Also, it provides number of total job openings of IT / Non-IT Job Openings based on locations.

3. How do you expect your analysis to improve decision making in this area?

Ans.3

Our analysis based on the dataset NYC_JOBS is useful in providing an overview of how jobs in different categories (IT/Non-IT) are distributed throughout different work locations. We can analyze the relationship between salary range based on job categories (IT/Non-IT) with different work locations, different working hour (Full-Time/ Part-Time/ Other-Time), which can provide job hunters to apply for more suitable jobs based on their interest and desired salary. Using the dataset, we can also do analysis on the relationship between job categories (IT/Non-IT) and jobs openings. We can not only provide the number of job openings available currently but also predict the number of job openings available in the future.

Data

Describe your data;

1.What is the source?

Ans.1

We acquire NYC_JOBS data set from NYC Open Data. NYC Open Data makes the wealth of public data generated by various New York City agencies and other City organizations available for public use. As part of an initiative to improve the accessibility, transparency, and accountability of City government, this catalog offers access to a repository of government-produced, machine-readable data sets. Anyone can use these data sets to participate in and improve government by conducting research and analysis or creating applications, thereby gaining a better understanding of the services provided by City agencies and improving the lives of citizens and the way in which government serves them. The database we choses, NYC_JOBS dataset, contains current job postings available on the City of New York��s official jobs site (https://opendata.cityofnewyork.us/). Internal postings available to city employees and external postings available to the general public are included.

2. What are your variables? If there are too many, focus on the ones that you will explore.

Ans.2

Our columns are as follows using ‘colnames’ command -

```
colnames(nyc_data)

##  [1] "Work Location"      "IT_Salary_From"     "IT_Salary_To"
##  [4] "NonIT_Salary_from"  "NonIT_Salary_To"    "Annual_salary_from"
##  [7] "Annual_Salary_to"   "Daily_Salary_from"  "Daily_Salary_to"
## [10] "Hourly_Salary_from" "Hourly_Salary_to"   "Annual_Salary_freq"
## [13] "Daily_salary_freq"  "Hourly_salary_freq" "Total_Opening"
## [16] "Non_IT"             "IT"                 "Full_Time"
## [19] "Part_Time"
```

```
## [1] "Work Location"      "IT_Salary_From"     "IT_Salary_To"
"NonIT_Salary_from"  "NonIT_Salary_To"    "Annual_salary_from"
"Annual_Salary_to"
## [8] "Daily_Salary_from"  "Daily_Salary_to"    "Hourly_S_from"
"Hourly_Salary_to"   "Annual_Salary_freq" "Daily_salary_freq"
"Hourly_salary_freq"
## [15] "Total_Opening"      "Non_IT"             "IT"
"Full_Time"          "Part_Time"          "Other_time_NA"
```

We have explored almost every variable but our main focus was on the ‘Work Location’,‘Non_IT and IT’ variable as we can calculate and explore all other variables in contrast with our ‘Work Location’ and “IT and Non-IT’ variables. We worked with Work Location, IT_Salary_From, IT_Salary_To, NonIT_Salary_from, NonIT_Salary_To, Annual_salary_from, Annual_Salary_to, Daily_Salary_from, Daily_Salary_to, Hourly_Salary_from, Hourly_Salary_to, Annual_Salary_freq, Daily_salary_freq, Hourly_salary_freq, Total_Opening, Non_IT, IT, Full_Time, Part_Time. We tried to find the relations between ‘Work Location’ and ‘All

Average Salaries available’. We also explored variables like ‘Full_Time and Part_Time’ that helped in finding relations between the ‘Full_Time and Part_Time jobs’ based on available locations.

3. Describe the data cleaning process if you needed

Ans.3

Data is usually not in the form that is needed to perform analysis. So, it needs to be cleaned so that it is technically correct to be worked on and is consistent. Uncleaned data is not in a form that the data mining models will accept Fields that are obsolete or redundant. There are missing values or values are not consistent. These are the data processing steps which we needed to process data - Finding and Searching data, cleaning the data as per our requirements, analyzed the data.

a. We found and segregated the data that was related to the jobs in NYC. We did this on the basis of the requirements that were needed for our analysis. E.g.- Filtered the data based on departments in different work locations. As our data was related to jobs, we divided the departments into 2 categories - IT and Non-IT and set as separate variables.

b. We changed the work locations to the 6 NYC Boroughs as they were scattered in the data according to addresses in all the boroughs. However, we took NYC as 6th category because there are some data in the dataset that did not give us location explicitly. They give location information like the Office of the Director, Office of the Director, Office of Public Information etc. Also, they have their locations all over NYC boroughs or some boroughs. So We took this kind of data as separate NYC category**.**

c. We removed many unnecessary variables as they were not at all relevant to our analysis.

In addition, we set salary range as separate variables. we also separated variable Salary Frequency which salary of job opening is Annual/Daily/Hourly. Likewise, we separated number of Full time/ Part time Job openings as different variables. We set these all variables based on work locations (which is NYC boroughs).

4. Provide descriptive statistics. Try to write a paragraph that gives descriptive statistics instead of just giving numbers in a table.

Ans.4

```
summary(nyc_data)
```

```
##  Work Location      IT_Salary_From   IT_Salary_To    NonIT_Salary_from
##  Length:217         Min.   :    0   Min.   :     0   Min.   :     0
##  Class  :character  1st Qu.:    0   1st Qu.:     0   1st Qu.: 35683
##  Mode   :character  Median :    0   Median :     0   Median : 52836
##                     Mean   :12430   Mean   : 17590   Mean   : 50296
##                     3rd Qu.:    0   3rd Qu.:     0   3rd Qu.: 61386
##                     Max.   :93756   Max.   :137577   Max.   :185000
##  NonIT_Salary_To  Annual_salary_from Annual_Salary_to Daily_Salary_from
##  Min.   :     0   Min.   :     0     Min.   :     0   Min.   :  0.00  
##  1st Qu.: 43079   1st Qu.: 47569     1st Qu.: 59058   1st Qu.:  0.00  
##  Median : 70113   Median : 57560     Median : 77242   Median :  0.00  
##  Mean   : 71129   Mean   : 56050     Mean   : 79074   Mean   : 23.01  
##  3rd Qu.: 90795   3rd Qu.: 66390     3rd Qu.: 95012   3rd Qu.:  0.00  
##  Max.   :224749   Max.   :185000     Max.   :224749   Max.   :385.00  
##  Daily_Salary_to  Hourly_Salary_from Hourly_Salary_to Annual_Salary_freq
##  Min.   :  0.00   Min.   : 0.000     Min.   : 0.000   Min.   :  0.00  
##  1st Qu.:  0.00   1st Qu.: 0.000     1st Qu.: 0.000   1st Qu.:  2.00  
##  Median :  0.00   Median : 0.000     Median : 0.000   Median :  3.00  
##  Mean   : 23.79   Mean   : 6.106     Mean   : 7.207   Mean   : 15.42  
##  3rd Qu.:  0.00   3rd Qu.: 0.000     3rd Qu.: 0.000   3rd Qu.:  8.00  
##  Max.   :385.00   Max.   :56.000     Max.   :64.000   Max.   :322.00  
##  Daily_salary_freq Hourly_salary_freq Total_Opening        Non_IT      
##  Min.   :0.0000    Min.   : 0.000     Min.   :  1.00   Min.   :  0.00  
##  1st Qu.:0.0000    1st Qu.: 0.000     1st Qu.:  2.00   1st Qu.:  2.00  
##  Median :0.0000    Median : 0.000     Median :  4.00   Median :  3.00  
##  Mean   :0.1659    Mean   : 1.346     Mean   : 16.94   Mean   : 14.83  
##  3rd Qu.:0.0000    3rd Qu.: 0.000     3rd Qu.: 10.00   3rd Qu.:  8.00  
##  Max.   :4.0000    Max.   :32.000     Max.   :354.00   Max.   :340.00  
##        IT           Full_Time         Part_Time     
##  Min.   : 0.000   Min.   : 0.000   Min.   :  0.00  
##  1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.:  2.00  
##  Median : 0.000   Median : 0.000   Median :  3.00  
##  Mean   : 2.074   Mean   : 2.143   Mean   : 14.76  
##  3rd Qu.: 0.000   3rd Qu.: 2.000   3rd Qu.:  8.00  
##  Max.   :83.000   Max.   :70.000   Max.   :295.00
```

```
sapply(select(nyc_data, 2:19), sd) # This gives Standerd deviation except work location
```

```
##     IT_Salary_From       IT_Salary_To  NonIT_Salary_from 
##       2.660978e+04       3.744854e+04       2.573115e+04 
##    NonIT_Salary_To Annual_salary_from   Annual_Salary_to 
##       4.183042e+04       2.399050e+04       3.951272e+04 
##  Daily_Salary_from    Daily_Salary_to Hourly_Salary_from 
##       8.440376e+01       8.661848e+01       1.223167e+01 
##   Hourly_Salary_to Annual_Salary_freq  Daily_salary_freq 
##       1.396916e+01       4.092976e+01       6.597827e-01 
## Hourly_salary_freq      Total_Opening             Non_IT 
##       4.019097e+00       4.352506e+01       4.047626e+01 
##                 IT          Full_Time          Part_Time 
##       7.572740e+00       6.991586e+00       3.829234e+01
```

From the above data summary. Check Work Location, it gives length:217 which tell us this dataset has 217 rows. It means we have a total 217 workplaces in the data.

From Total_Opening, there are maximum 340 and minimum 1 job opening in NYC. And Average job opening is 17.

In addition, for IT Job Opening, there are maximum 83. For Non-IT companies, its maximum job openings are 340.

We can see the average salary range for IT and non-IT, which is from 12430 to 17590 and from 50296 to 71129 respectively. Likewise, we can see an average of the annual salary range, that is from 56050 to 79074, an average of Daily salary range which is from 23.01 to 23.79 and average of hourly salary which is from 6.106 to 7.207.

Further, we can know that how many job openings per location that pay annually, daily or hourly. From the above summary, see Annual_Salary_freq, there are 15 average job openings that pay annually. Also, Check Daily_salary_freq, it tells us there is almost no job opening that pay on daily basis. It means most companies do not pay daily. See Hourly_salary_freq, there are average 1 job opening which pays hourly.

Also, there are average around 2 and 15 job openings which are Full time and Part time respectively. Part time Job Opening are more than full time based on maximum Job openings of full time and part time that has maximum job 70 and 295 respectively.

Exploratory Analysis (Complete analysis in .rmd file)

Conclusion

We worked on the dataset NYC_Jobs. After going through all the variables and details of the project and dataset we analysed all the available material and tried simplifying it as much as possible to make it understandable. We explored most of the variables which were relevant to our analysis. After dividing the job categories into IT and Non-IT fields we used the variables in relevance to the jobs available. We got to know about what are the salaries paid in 6 boroughs of NYC may it be maximum or minimum. Here are our findings –

1. We got to know the relationship between Work Location and IT and Non IT Job Opening with graph and explanation.

2. We explored the average salary in each location both maximum and minimum on annual and hourly basis.

3. We analysed jobs based on timings - both part-time and full time location wise.

4. Explored the relationship between salary frequency based on location.

After exploring and analyzing our data, and finding our results, we are still interested in predicting the outcomes related to all possible variables. For example, we can predict the salary range and IT/Non-IT Job Openings based on work locations.