Exploratory Data Analysis

Sonal, Joyce, Abhay and Vishwesh

Department of Computer Science
Seidenberg School of CSIS
Pace University

Introduction

- In statistics, exploratory data analysis (EDA) is an approach analyzing data sets to summarize their main characteristics, often with visual methods.
  - NYC Jobs Data from NYC Open Data
  - Explored Work Location and Number of Job Openings based on IT / Non-IT Category

How does our project help those who are interested in this topic?

- 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.

Why data analysis is needed for our data ?

- 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.
- Extract these information from this project:

✓ Full time and part time job opening per locations

✓ Minimum and maximum salary according to work location based on salary frequency like Annually, Daily or Hourly

✓ Which work Location has the most job

✓ 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.

How our our analysis to improve decision making in this area?

- Providing an overview of how jobs in different categories (IT/Non-IT) are distributed throughout different work locations.
- 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 provides job hunters to apply for more suitable jobs based on their interest and desired salary.
- 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

Source

- Acquired 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.
- Anyone can use these data sets to participate in and improve government by conducting research and analysis or creating applications.
- (https://opendata.cityofnewyork.us/)

Data Cleaning Process

- Data is usually not in the form that is needed to perform analysis. It did not have enough numerical variable for analysis.
- There are missing values or values are not consistent. These are the data processing steps which we needed to process data.

1. 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. Eg- Filtered the data based on departments in different work locations.
2. As our data was related to jobs ,we divided the departments into 2 categories - IT and Non-IT and set as separate variables.
3. We changed the work locations to the 6 NYC category as they were scattered in the data according to addresses in all the boroughs.

Data Cleaning Process conti...

- **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 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.**

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

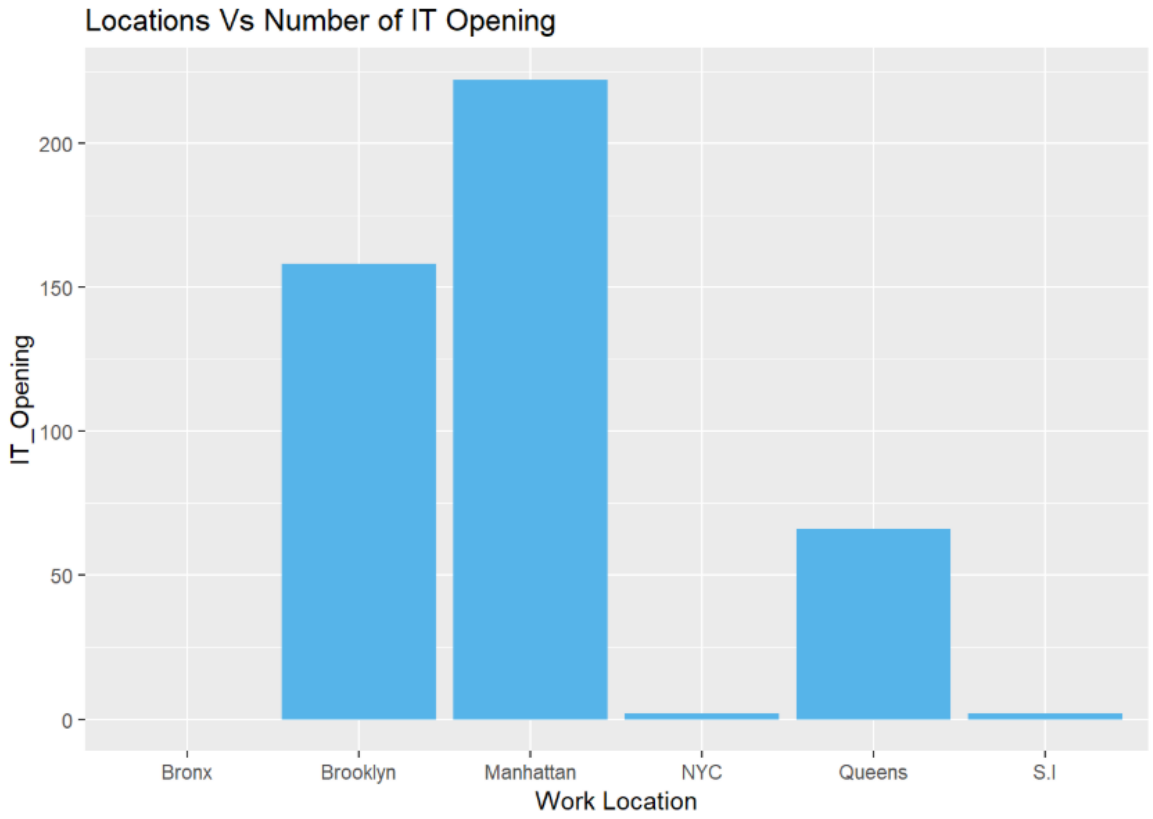
5. We tried to make the data as compact and informative as possible by removing the ‘NA’ values.

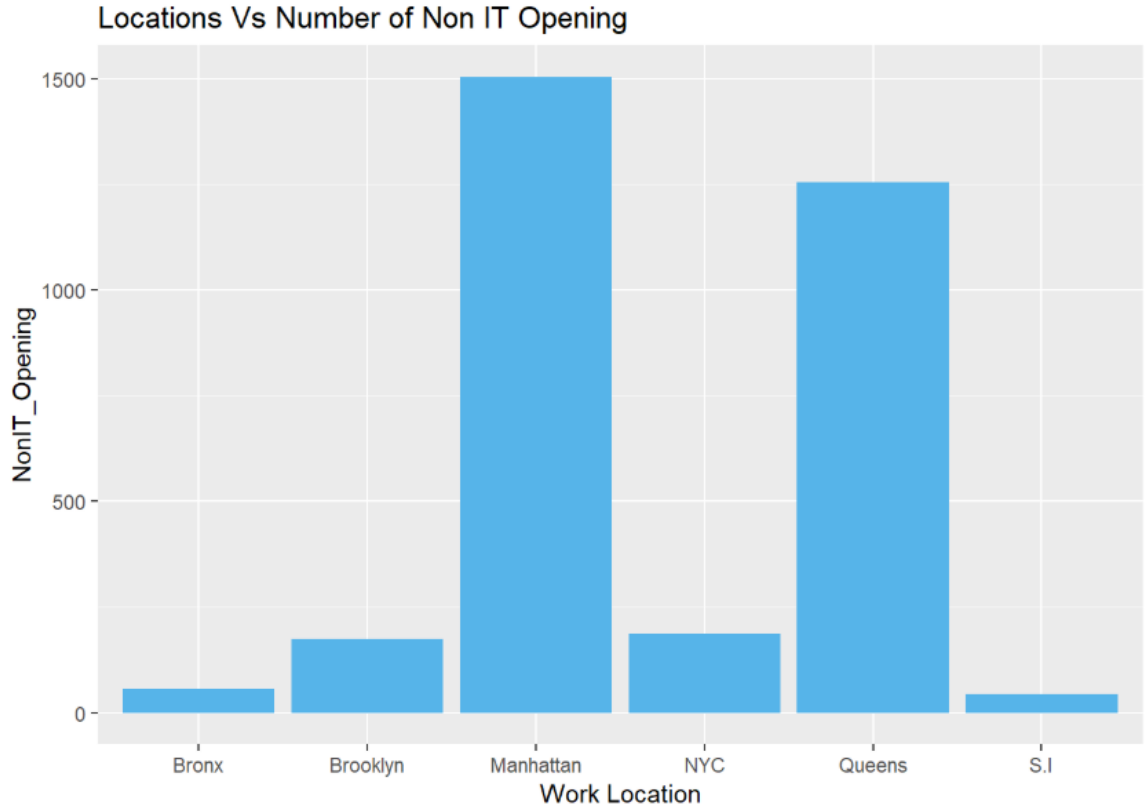
Focusing Variables:

- 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.
- 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.
- 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.

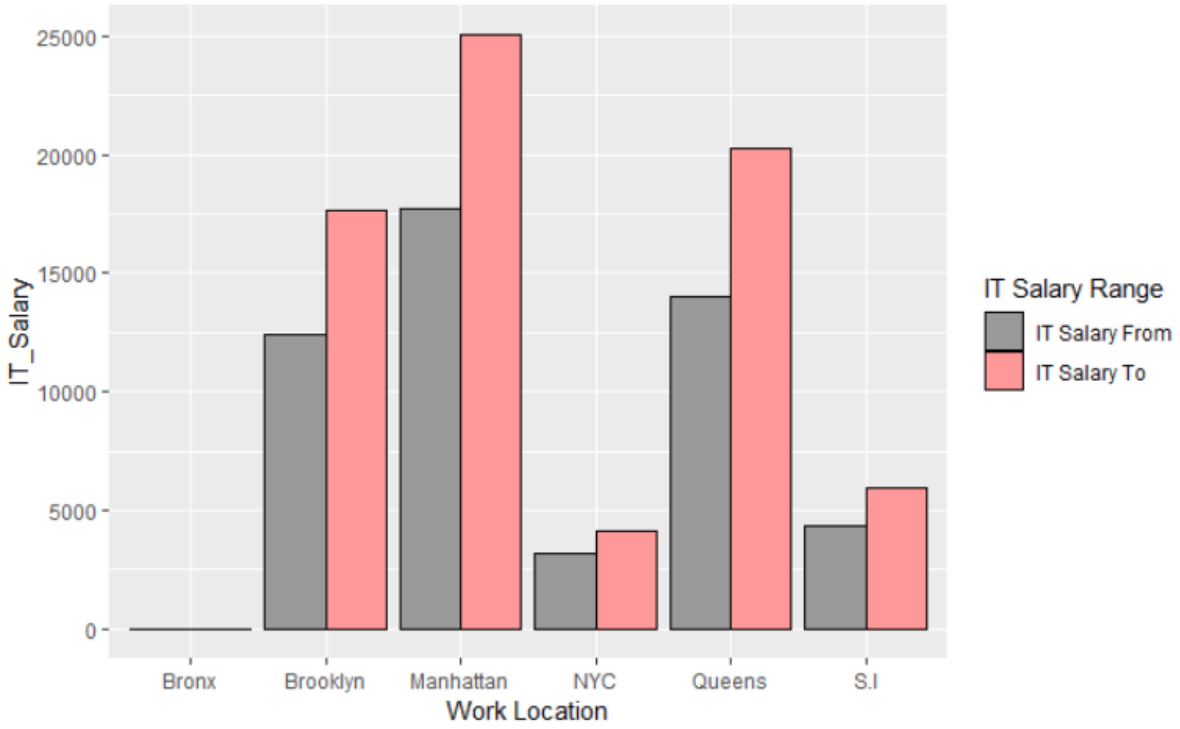
Exploratory Analysis

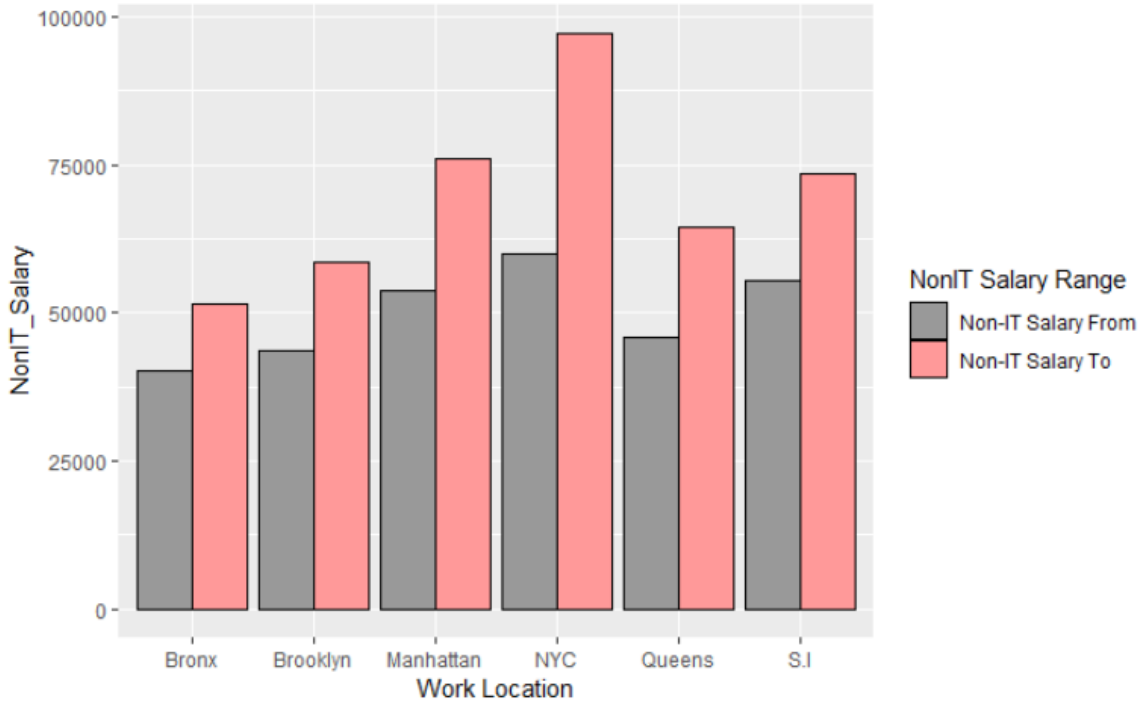
IT / Non-IT Job Openings based on Work Location:



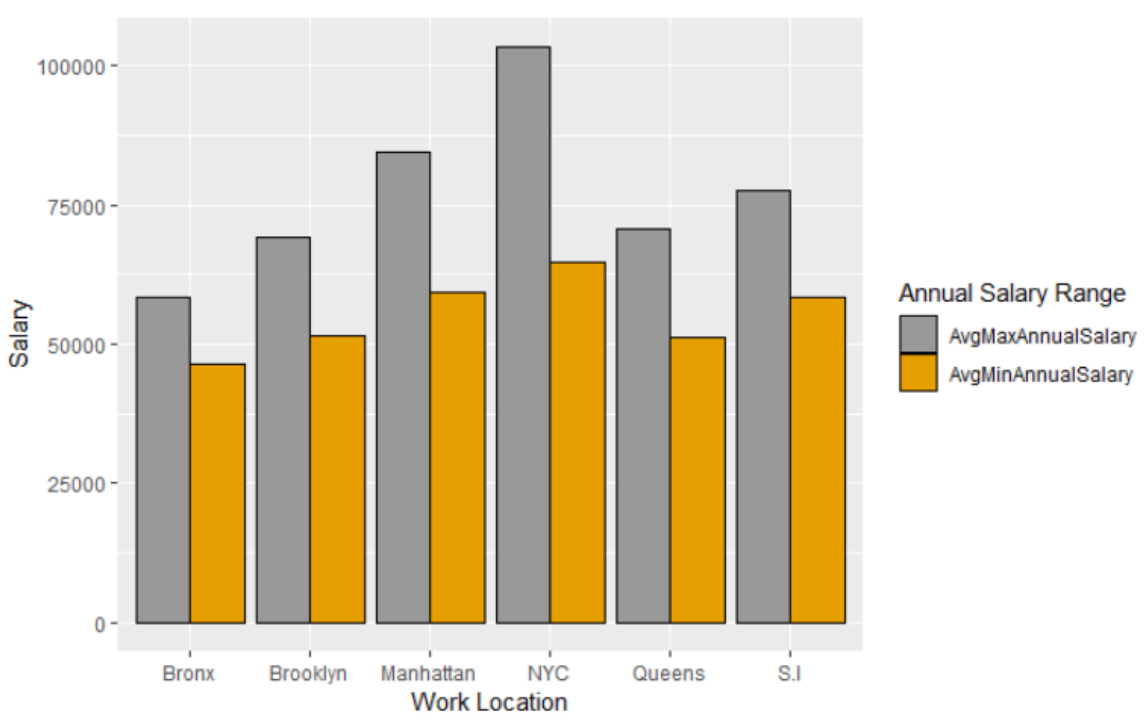


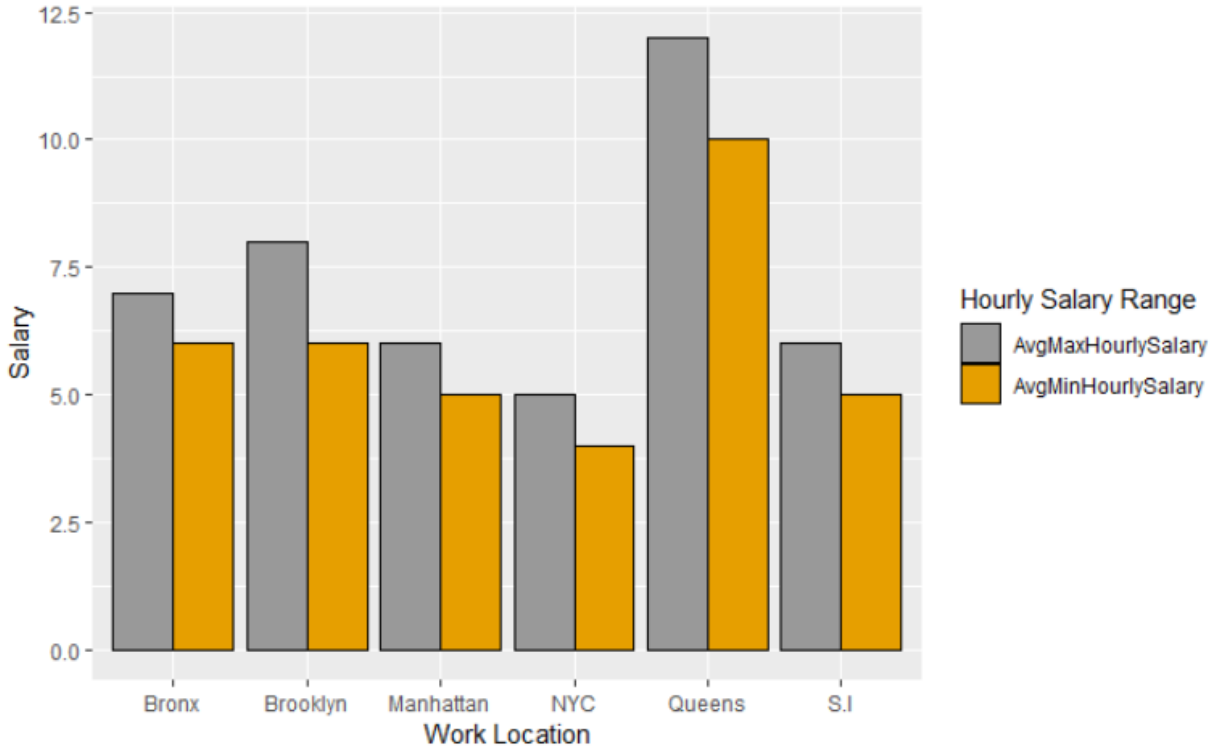
IT / Non-IT Salary based on Work Location:



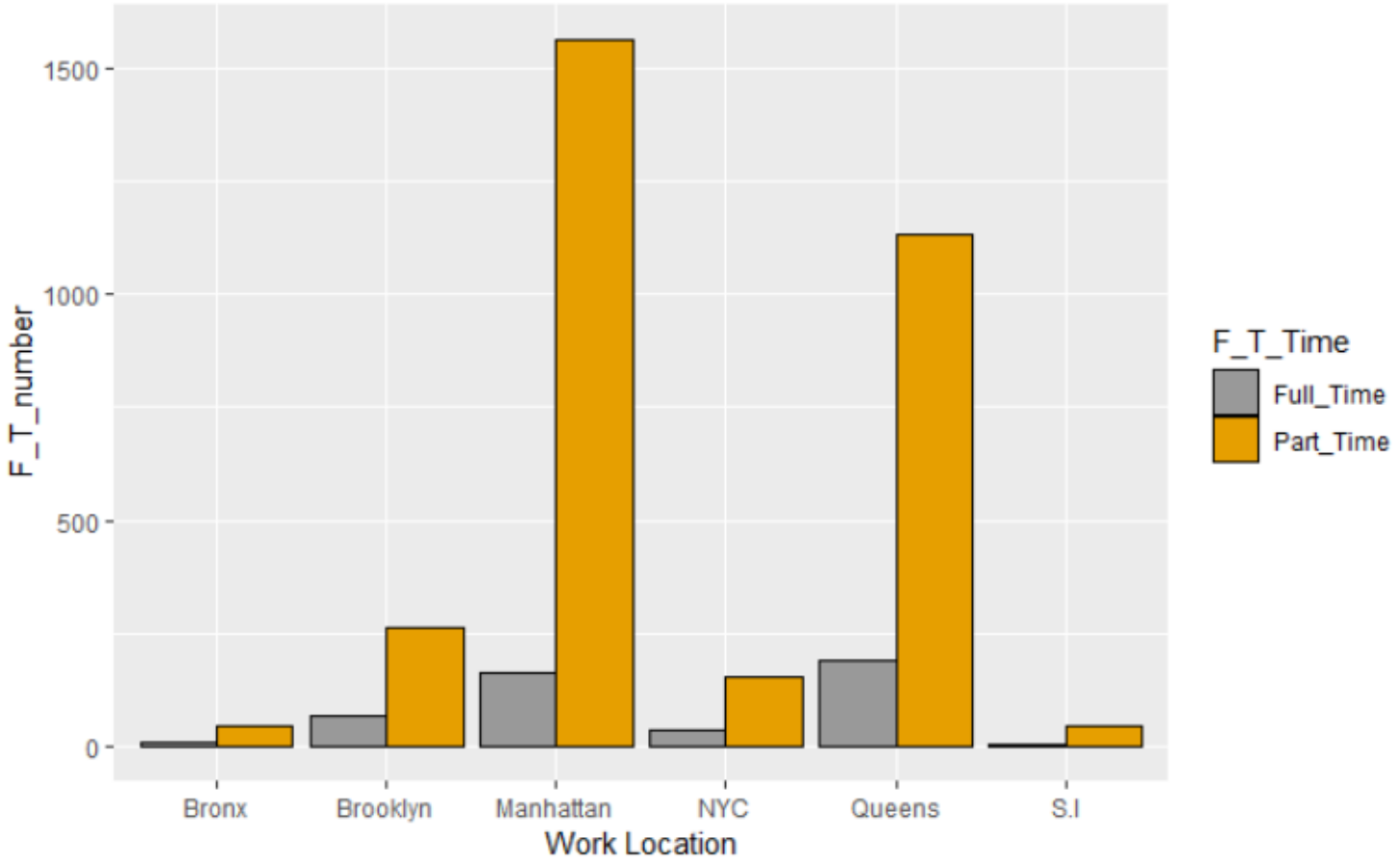


Average maximum and minimum annual and hourly salary in each locations:

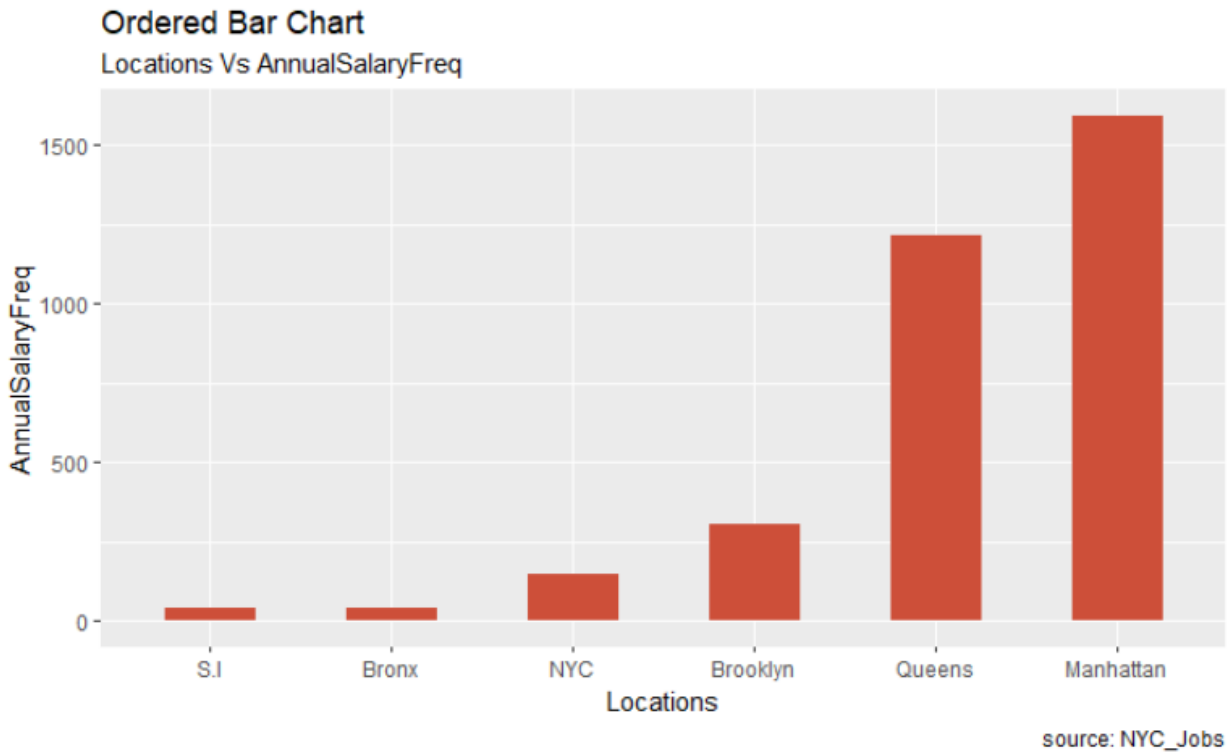


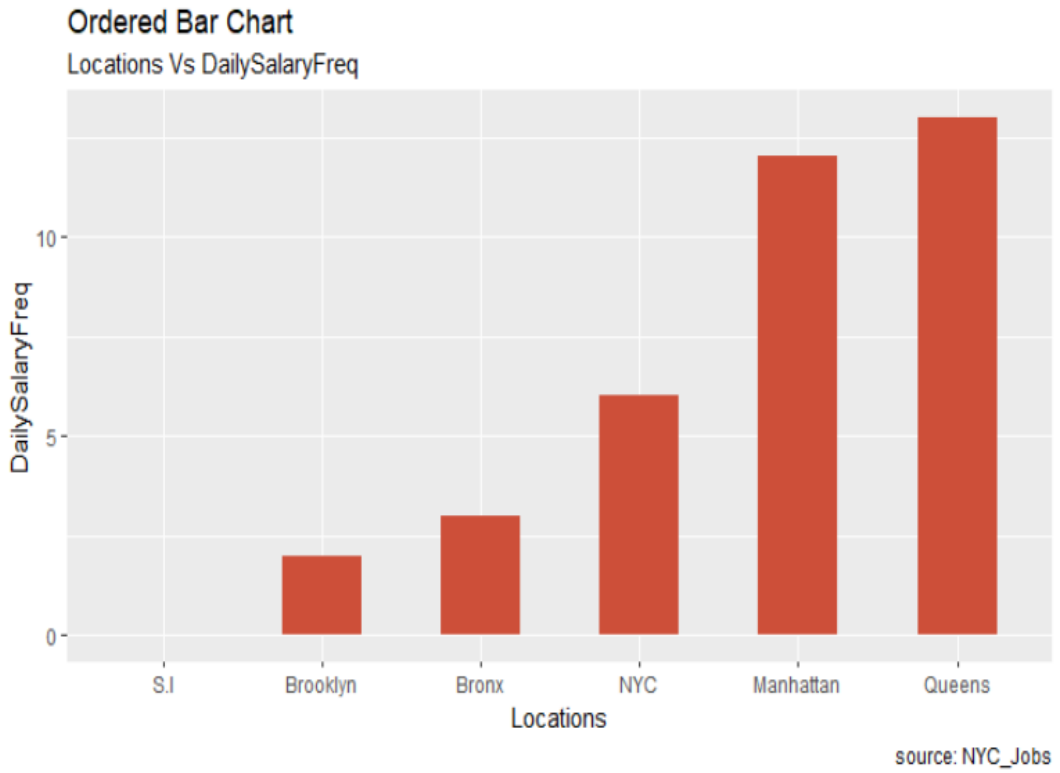


Part-time and full time jobs are there in each location:

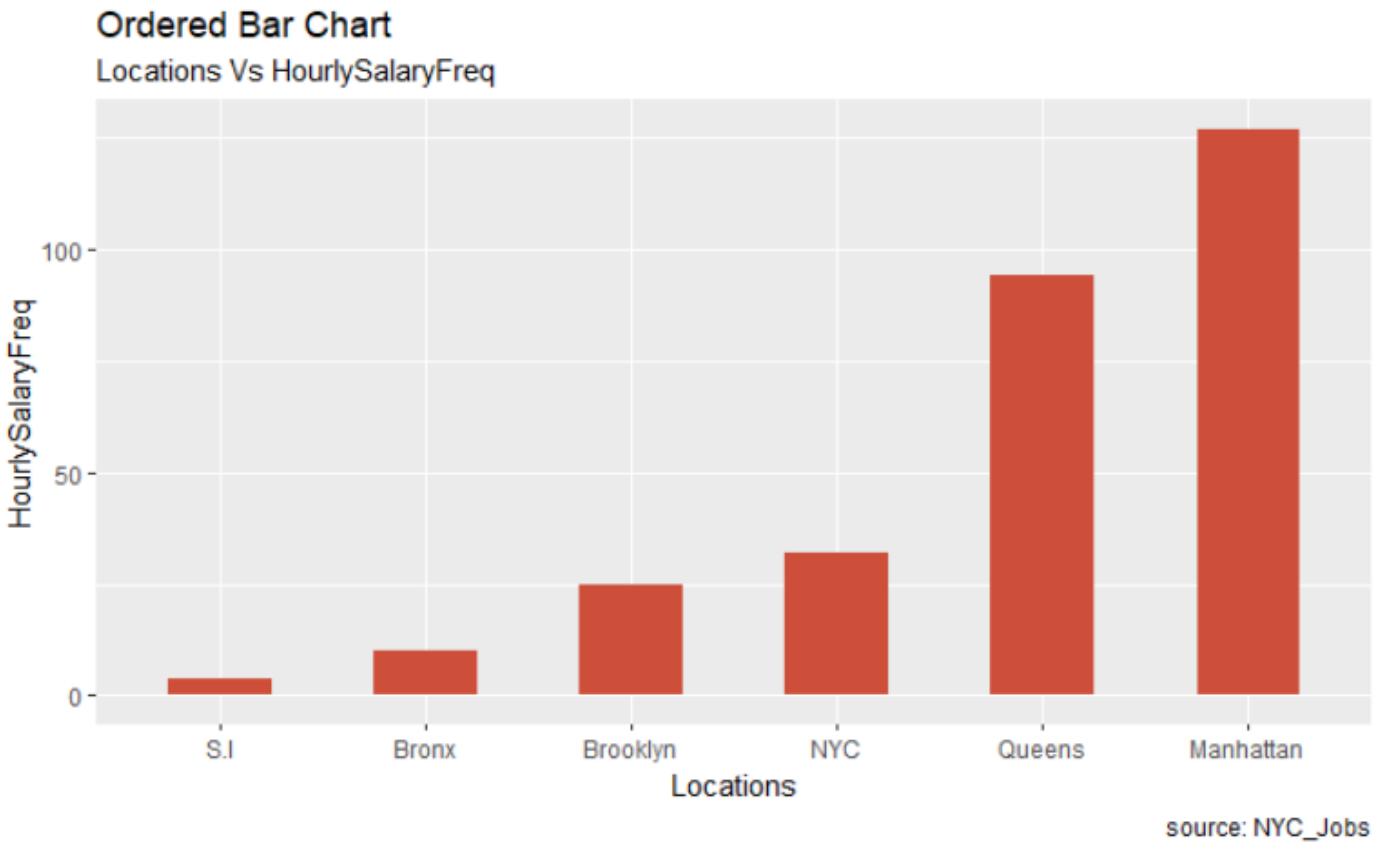


Locations Vs Annual Salary Freq and Daily Salary Freq:

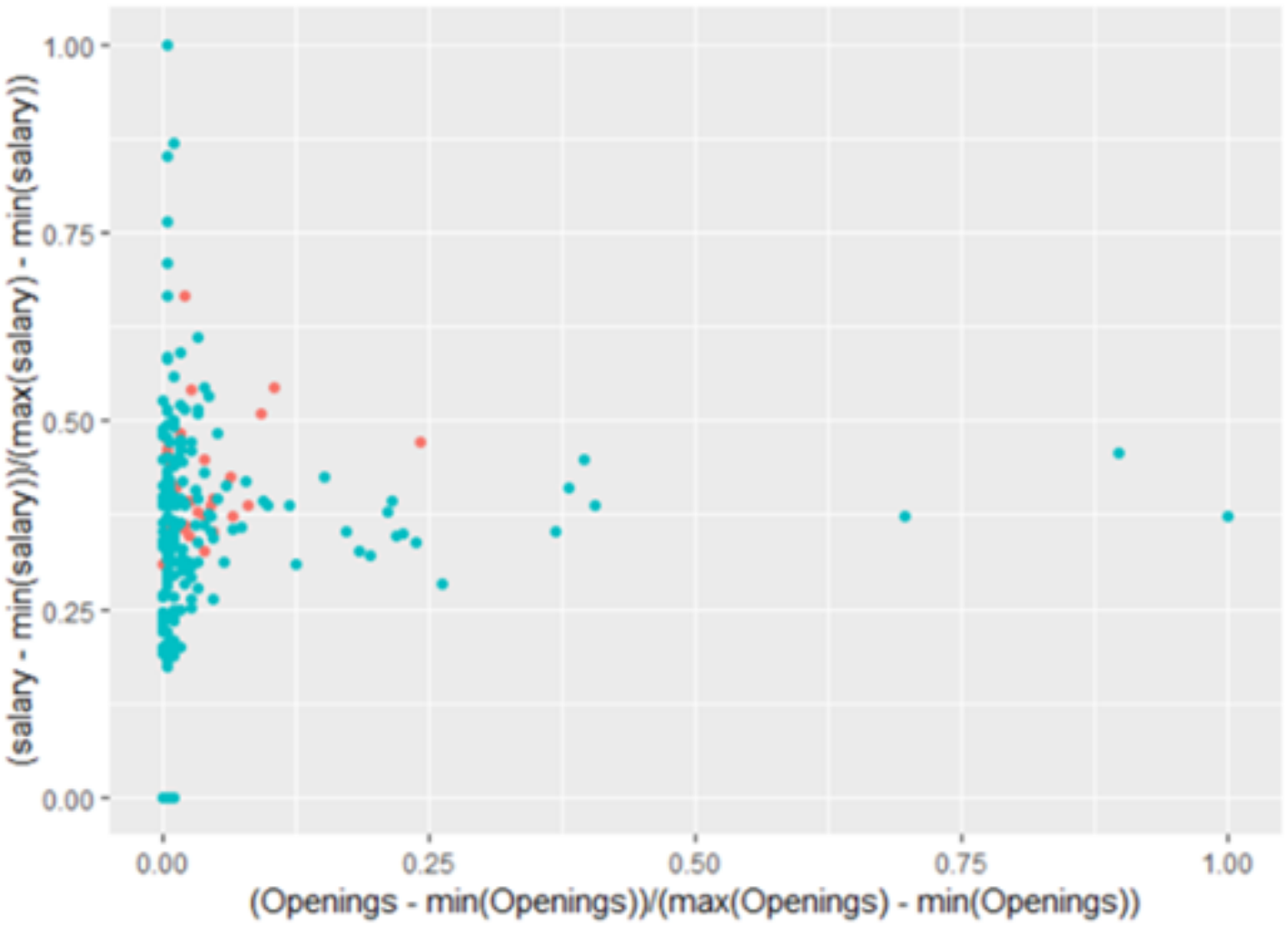




Locations Vs Hourly Salary Freq:



Prediction Analysis




```
[1] 0.8510638
             testing_label
predications IT Non_IT
       IT     1      0
       Non_IT 7     39
```

Thank you