**Analysis of New York City Current Job Postings**

IST652 – Scripting for Data Analysis

Himadri Tewari – hstewari@syr.edu

School of Information Studies

The Syracuse University

09/04/2019

1. **About the Project:**

In today’s world, due to the rapid growth of technology and development of the Internet community, a large amount of valuable and complex data is growing exponentially and is expected not to slow down anytime soon. There are two aspects of utilizing big data. One would be to use the data to design new processes for data modeling and prediction using predictive models, and another would be to explore the data and identify the trends, hidden pattern, design charts, and visuals and help the organization to take correct strategic decision.

In general, data analysis refers to the process of inspecting the raw data, converting it by cleaning, transforming and modeling with the objective of identifying useful information, suggesting collections and supporting decision making. The data is the main requirement which can be collected as a sample from the population. All the related variables regarding the population need to be specified. Data can be structured or unstructured, numerical or categorical. The data can be collected from a variety of sources like the internet, documents, social media platforms. The data collected need to be processed and organized by adding or modifying to include rows and columns in a tabular structure. Once the data is processed, then it is cleaned by correcting any incomplete, duplicate or error related to the data.

The goal of this work is to apply data analysis techniques to the job posting data. The job posting is one of the platforms to communicate to the public about any vacancy that an organization needs to fill up. The posting provides the applicants a good perspective about the job, which includes the job description which contains broad details of what the position is for. The posting also communicates to the public what is needed in the person to fulfill the job. To fulfill the job-seeking position and to find the correct person, many organizations hired employment agencies who help with its staffing needs. Employment agencies post the job vacancies on the job posting platforms like Glassdoor, Indeed, Dice or LinkedIn, mentioning the job details. The job details include the job duties which details the roles and responsibilities or tasks, the person needs to perform in the work. The detail varies from different posting and organization to organization. A job description can be mentioned in broadly or in specific details based on the job criteria and the type of person, the organization is looking for. It also contains the knowledge and skills required for the person to apply for the job. This is the basic feature that the person seeking for the job need to know. These features are the thing that an organization does not have the time and resource to teach the new hire. The next feature is the minimum educational qualifications and experience requirements, which talks about formal education and work experience the applicant needs to apply for the job. This a flexible feature and also varies from organization to organization based on the need. It helps the organization to cluster the posting to entry-level, mid-level or higher-level applicants. The next job detail is the salary range and salary frequency let the job seeker know how much the organization is willing to pay for the position. Most of the times, the job posting will have to provide the salary details, so that the applicant can make a decision based on that. The last part of the job details in the application instruction which let the applicant how and when to apply for the job. It includes what the additional and supporting documents needed to apply for the job.

1. **Analysis**
   1. **About the Data**

**2.1.1 Choosing the Data Set**

The dataset is collected from the Kaggle website (<https://www.kaggle.com/new-york-city/new-york-city-current-job-postings>) which contains job posting available in the City of New York’s official job site. The data is hosted in open data platform – NYC Open Data which contains over 1200 datasets with new data added every day (<https://opendata.cityofnewyork.us/>). This dataset is updated weekly. Figure 1, shows the first four rows from the dataset, some of the attributes and the corresponding values. The complete list of all the features in the initial dataset is given in Figure 2. The initial dataset has twenty-eight attributes. It contains the name of the agency, posting type, number of positions, job category, full time or part-time, salary range and frequency. It also contains additional important attributes which focuses more on the posting details like minimum qualification required, preferred skills, work location, and the posting dates. It also contains the instruction to apply and additional information.

Figure 1. Tabular Form Summary

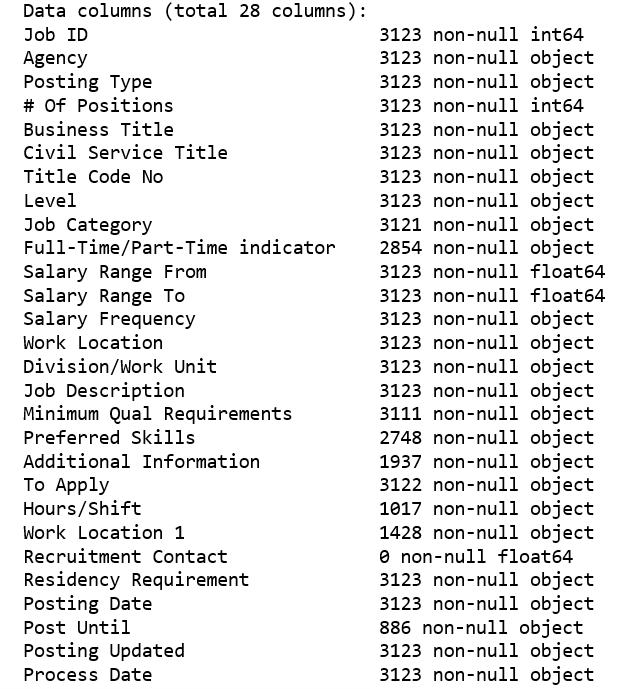


Figure 2. Dataset 28 Attributes

**2.1.2 Complete Metadata**

**Job ID:** The job opening identification number.

**Agency:** Name of the New York City agency where a job vacancy exists.

**Posting Type:** Identifies whether a job posting is an Internal or External posting. Internal postings are available to City employees only and external postings are available to the general public.

**# Of Positions:** The total number of vacancies to be filled under the job ID listed.

**Business Title:** The "business title", or "office title", for the job posting listed.

**Civil Service Title:** The civil service title for the job posting listed.

**Title Code No:** The Title Code Number (“Title Code No.”) that corresponds to the civil service title posted.

**Level:** The civil service title assignment level that the posted position is being filled at.

**Job Category:** The occupational group in which the posted job belongs.

**Full-Time/Part-Time Indicator:** This denotes whether the job is a full time or part-time employment. F - Full time, P - Part-time

**Salary Range From:** The lowest salary on a job posting for a position within the salary band for the related civil service title.

**Salary Range To:** The highest salary on a job posting for a position within the salary band for the related civil service title.

**Salary Frequency:** The frequency of the proposed salary. Possible salary frequency values include “hourly”, “daily”, and “annual”.

**Work Location:** The physical address for the agency listing the job posting.

**Division/Work Unit:** The department/area within the hiring agency for the job position listed in a job posting.

**Job Description:** The description of the job responsibilities for the position listed in the job posting.

**Minimum Qual Requirements:** Minimum qualifications (“minimum qual”) required for the position listed in the job posting.

**Preferred Skills:** Additional, preferred skills desired by the hiring agency for a given position.

**Additional Information:** Additional information provided by the hiring agency.

**To Apply:** Instructions on how to apply for a given job vacancy.

**Hours/Shift:** Projected working hours, working days and shift information.

**Work Location 1:** Where applicable, additional work location details for the job vacancy.

**Recruitment Contact:** Where applicable, recruitment contact information is provided.

**Residency Requirement:** Residency requirements for a given job vacancy.

**Posting Date:** The date and time that a job vacancy was posted.

**Post Until:** The last date that a job vacancy will be posted.

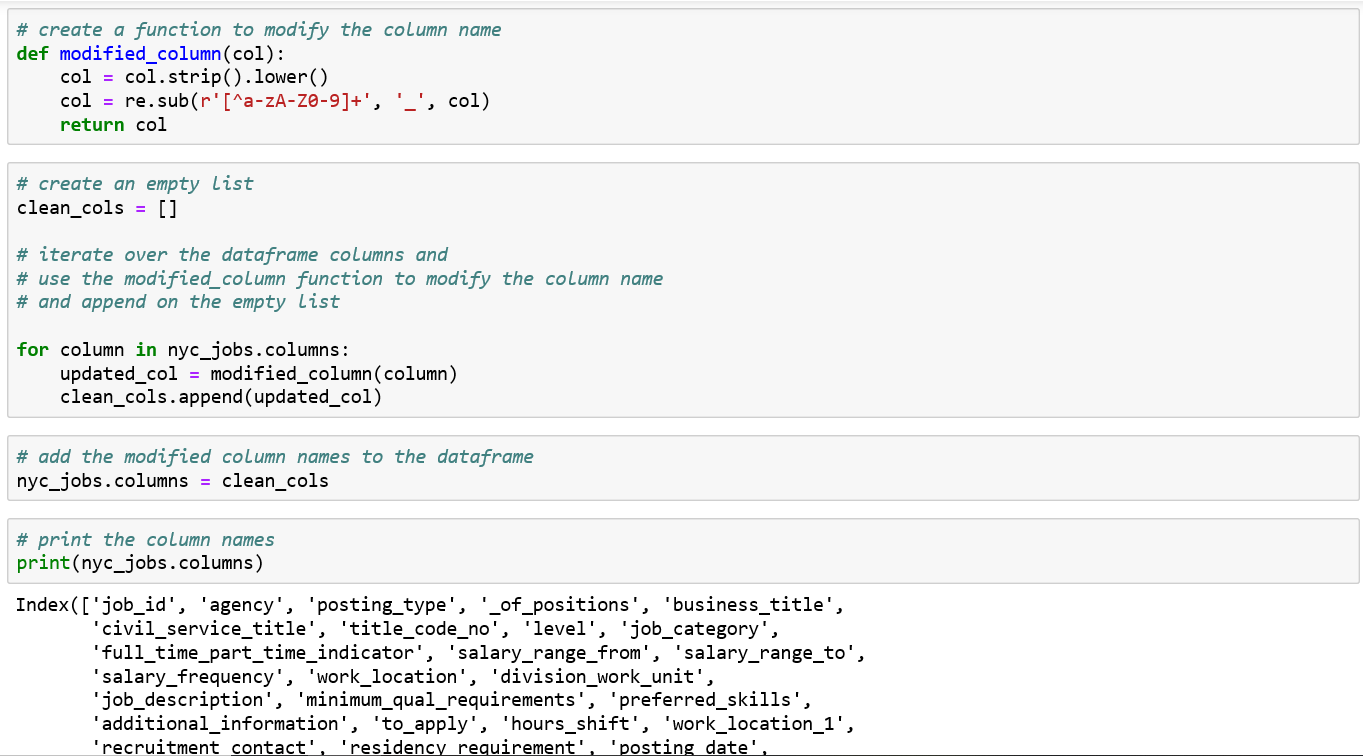
**Posting Updated:** The last date and time a job vacancy posting was modified.

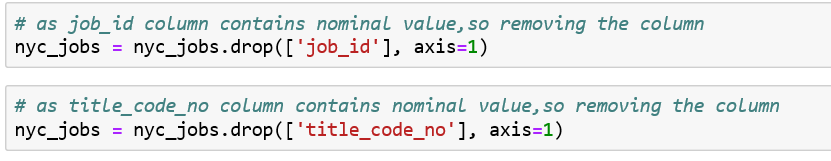
**Process Date:** The date and time that the dataset created date to include the corresponding job vacancy posting.

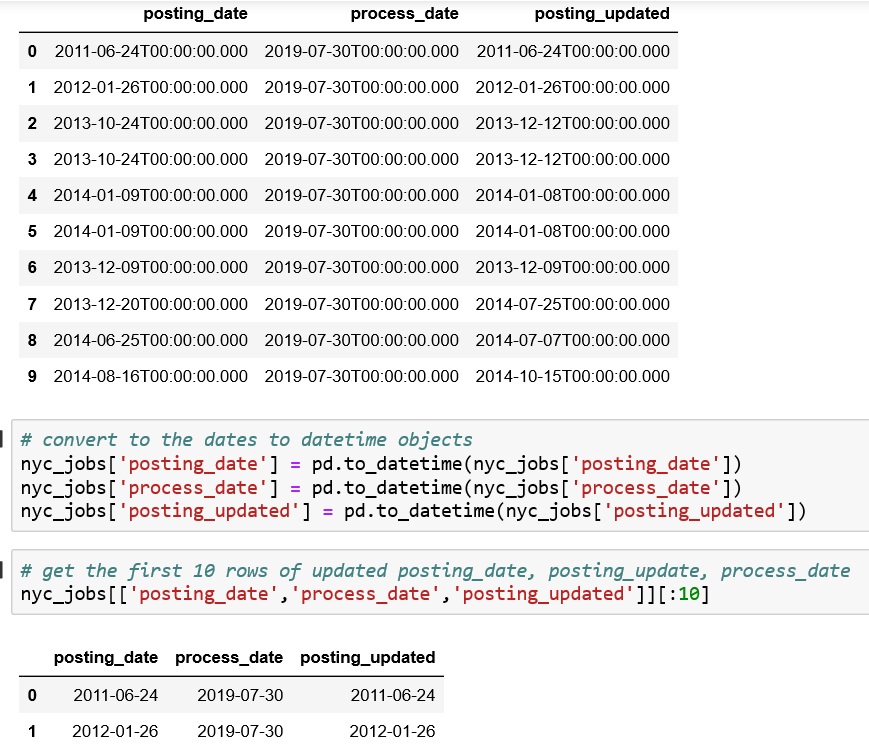
**2.1.3 Data Preprocessing**

Data preprocessing is the first level of data analysis technique where it transforms the raw data into an understandable format. The raw data is often inconsistent, incomplete, contain duplicate data and contain many errors. Using the raw data directly is not a viable idea as it lacks in lots of information and most of the time this data cannot be processed through a model. So, it is very important to preprocess the data for better analysis. It involves multiple steps in data preprocessing. Some of the important steps involved, importing the required libraries or packages, import the dataset, checking for special characters from the column, checking for nominal data, categorical data, checking for missing values, checking for null values, standardize the data and feature scaling.

For this project, as part of the data preprocessing, the following list of important libraries are imported. Pandas are imported for data manipulation and analysis. Numpy is imported for scientific computation. Matplotlib and seaborn are imported for visualization purpose. Wordcloud is imported to generate the word-cloud. Nltk is imported for stopwords. “re” library is imported for the regular expression. Gmap is imported to visualize the map. The dataset was collected in CSV format, so to read the data, pandas ‘read\_csv’ function was used. After loading the dataset, the dataframe was summarized and the first five rows were analyzed. The columns of the dataframe were renamed. The columns were explored further to find out the data types and the columns having nominal data types are dropped. Then it was analyzed to find out if there are null values present, and the columns having null values are dropped. As part of the next preprocessing step, certain values are segregated to get the details of the column. In the next step, the date columns were converted to date time for better analysis.





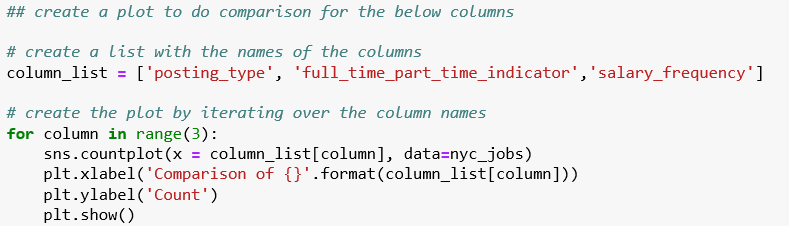
1. **Attributes for Analysis**

Referring to the metadata and the data preprocessing, some of the attributes which need to analyze as part of the data analysis process are:

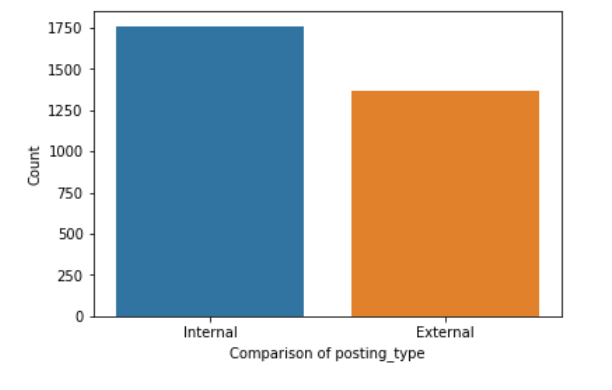
* Comparison of Posting Type, Full Time/Part Time, Salary Frequency
* Which are the Agencies have the highest number of jobs posting?
* Which are the top Civil Service Titles, Division Work Units, Business Titles & Job Categories?
* Which is the frequently used words in Job Description, Preferred Skills, Job Instructions, Additional Information & Minimum Qualification Required?

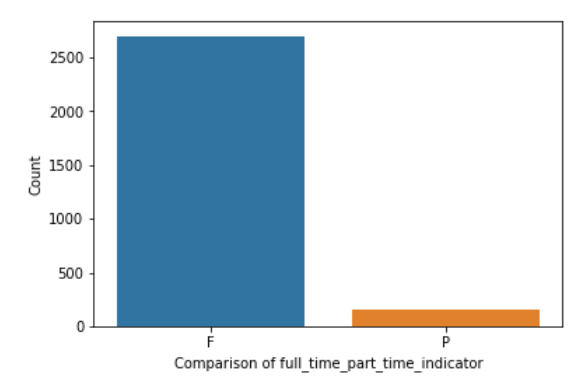
1. **Data Analysis**

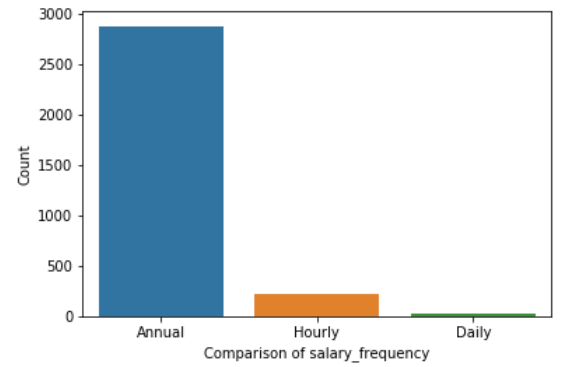
As part of the data analysis process, the primary focus will be to process the data where it transforms the raw data into an understandable format. The next step will be to focus on systematically applying statistical and/or logical techniques on the existing dataset. The process involves multiple steps that include extracting and categorizing the dataset to generate different patterns, relations, connections, and other such valuable insights from it. In the exploratory data analysis, the data is explored to find patterns, spot anomalies, test hypothesis and to check assumptions with the help of summary statistics and graphical representations. The main idea behind representing the data pattern graphically as it helps the user to visualize the pattern and its relation. It helps to reveal the structural insight visually to the users. First, a summary of the dataframe was taken. Then a comparison chart is plotted using some of the attributes present in the dataframe.



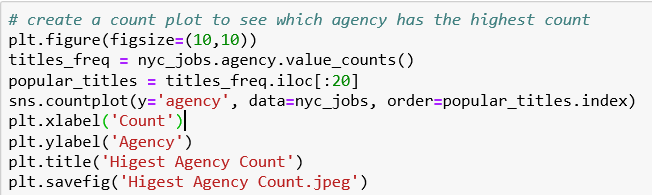
From Figure 3, it can be interpreted that internal posting is more than external posting. The next analysis was done to compare full time and part-time job posting. Figure 4 shows that jobs posted are looking for full-time employees in compare to part-time employees. Then salary frequency column is explored to compare the annual, hourly and daily frequency. Figure 5 shows that Hourly salary frequency is much higher than Hourly. Daily salary has the lowest counts.

Figure 3. Comparison of Posting Types

 Figure 4. Comparison of Full-time and Part-time job posting

Figure 5. Comparison of Salary Frequency

The next analysis was done by visualizing the count of Agencies posted the jobs. First, the unique counts of the first twenty Agency column were generated. Then a count plot is created using the count.



From Figure 6, it can be observed that the Department of Environment Protection has the highest job posting followed by the Department of Health and Mental Hygiene and the Department of Design and Construction. Referring to the same chart, it can be interpreted that Contract Service in Mayor’s Office has the lowest count out of the first 20 unique counts generated. The same process was applied for Civil Service Title, Division Work Unit, and Business Title. Figure 7 shows that for the Civil Service Title column, Community Coordinator got the highest count followed by Agency Attorney and then Civil Engineer. Staff Analyst got the lowest count. After analyzing the Business Title column, it can be observed from Figure 8, that Project Manager got the highest count. Assistant Civil Engineer and Accountable Manager got the same count. Then followed by Engineering Technician 1 and then Director. Civil Engineer 2, Watershed Maintainer, Electrician, Computer Specialist, and Construction Project Manager got the same count, but are in the lowest position of the chart. Administrative Assistant has the lowest count in the plot.

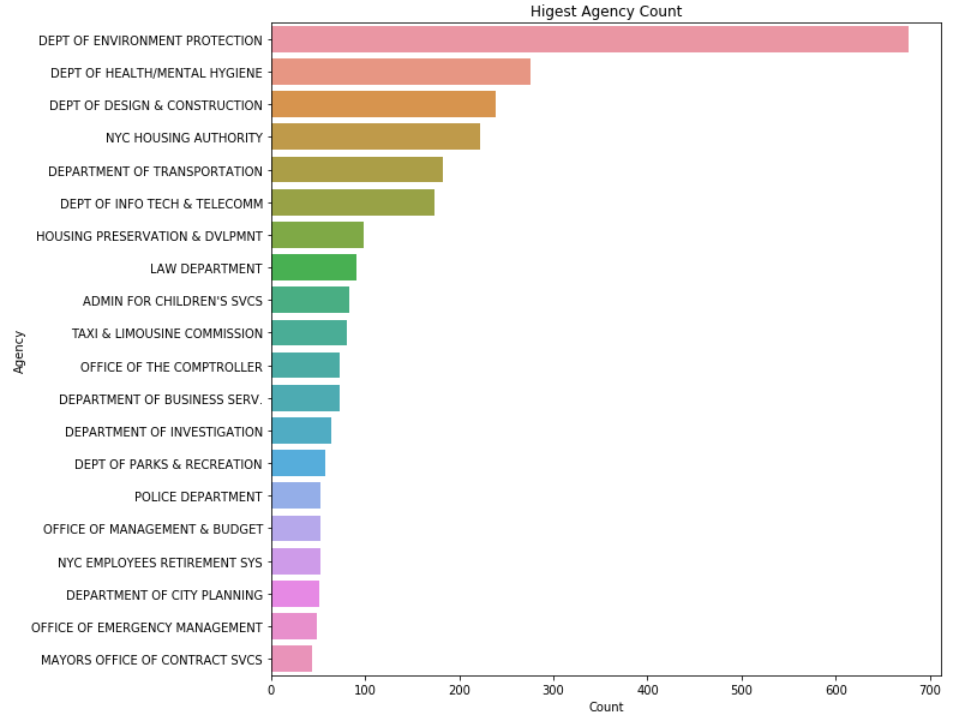


Figure 6. Top 20 Agency Count

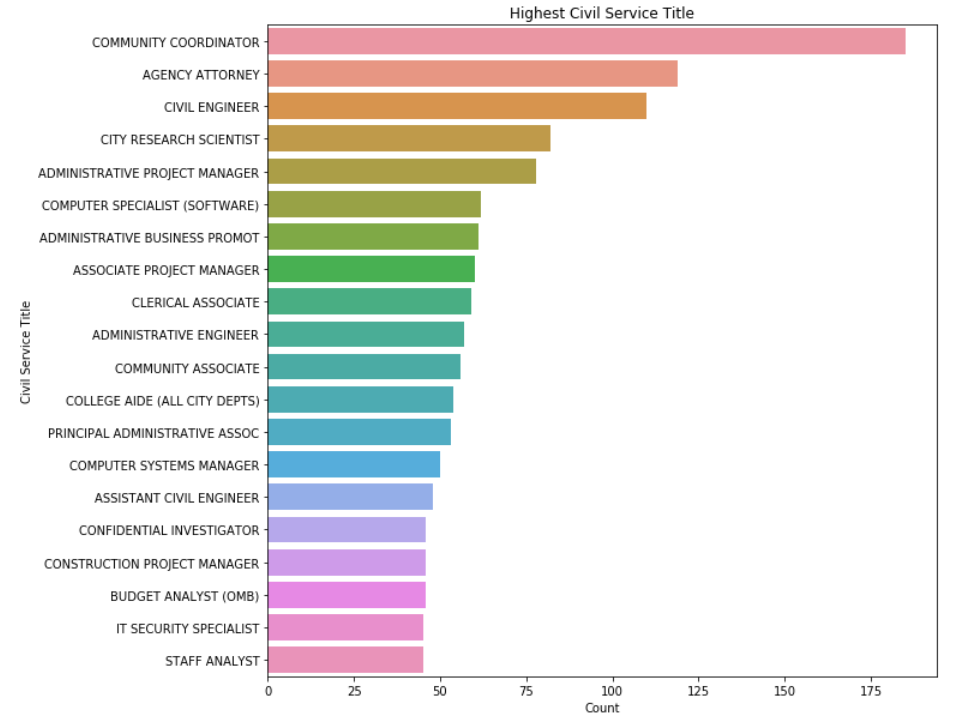


Figure 7. Top 20 Civil Title Count

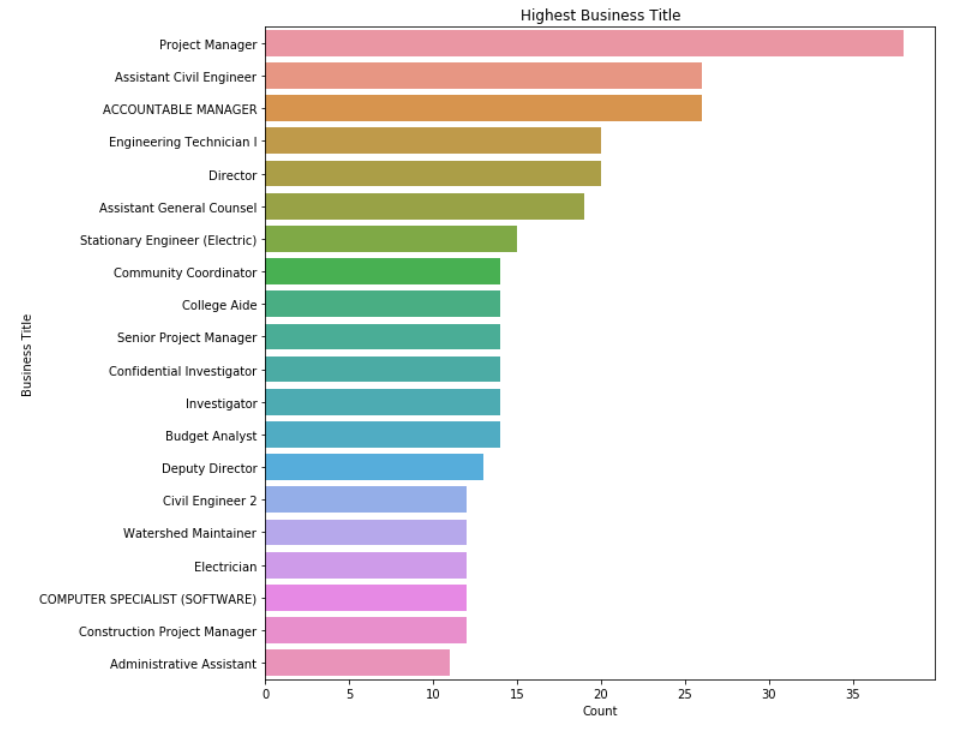
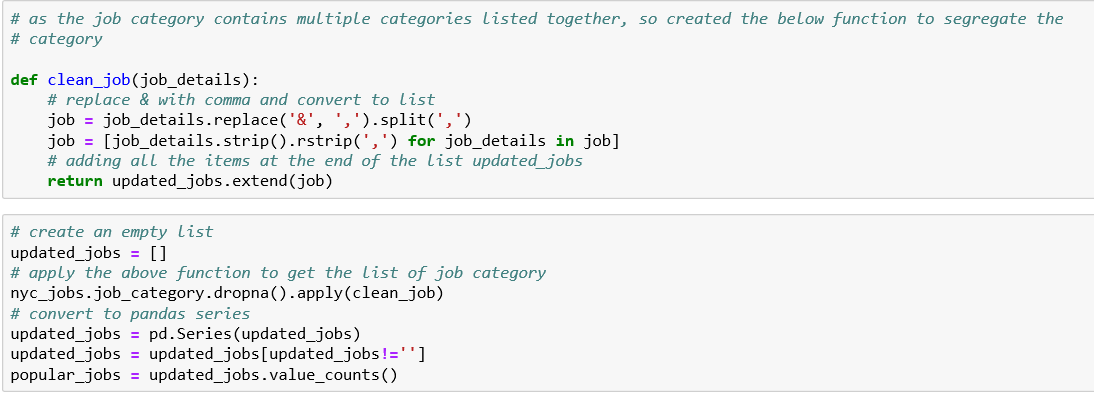


Figure 8. Top 20 Business Title Count

Then, the Job Category column was explored and was found that it contains multiple job categories listed together. To get the count of individual job category, a function was written to split the categories and add the categories on the list.



After the job categories are segregated, the job category is then converted to pandas series and the count of the first twenty job categories are plotted. From Figure 9, it can be observed that Architecture has the highest job posting followed by Engineering, Planning and then Data. Referring to the same chart, it can be interpreted that Building Operation, Health, and Procurement got the lowest count.

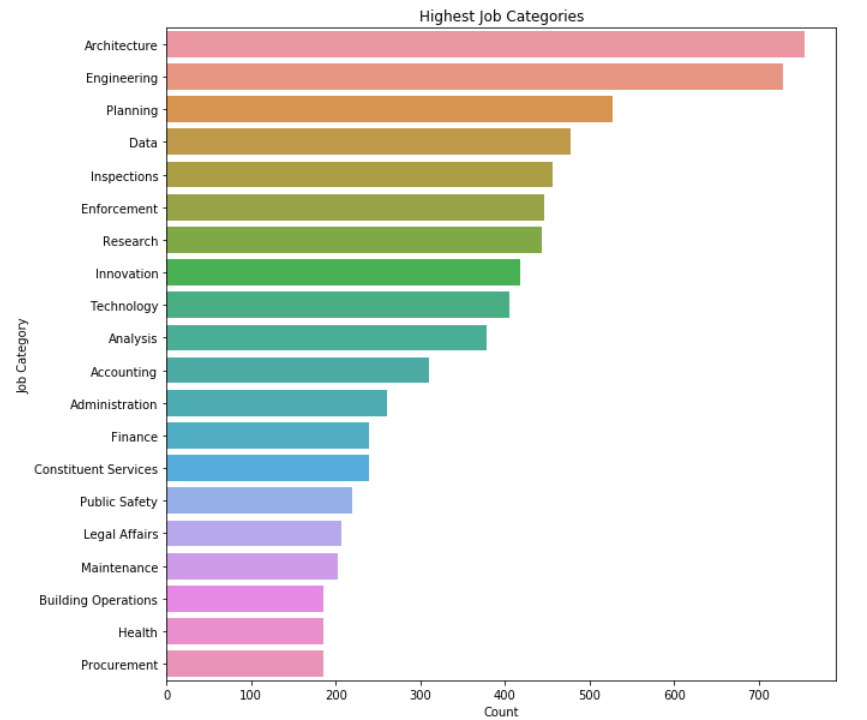


Figure 9. Top 20 Business Title Count

The data was further analyzed using other attributes like Job Description, Preferred Skills, Job Instruction, Additional Information, and Minimum Qualification. As all these attributes contain text data, so the word cloud is formed to visualize the frequency and the importance of each word. To generate the word cloud, the stopwords package is imported from nltk library and the list of stopwords are printed. Then, using the Wordcloud package, the word clouds for the columns were generated.



Figure 10 shows that word cloud is generated for Job Description column, it can be observed that “New York City”, “selected candidate”, “support” are the most frequent word. So, it can be interpreted that the job posting is done for “New York” city; the data is for job posting, so the “selected candidate” is also one of the most frequent words. Whereas, certain words like “wastewater”, “noise”, “hazardous” are the least frequent words, so it can be inferred that jobs related to these kinds of words are less posted.

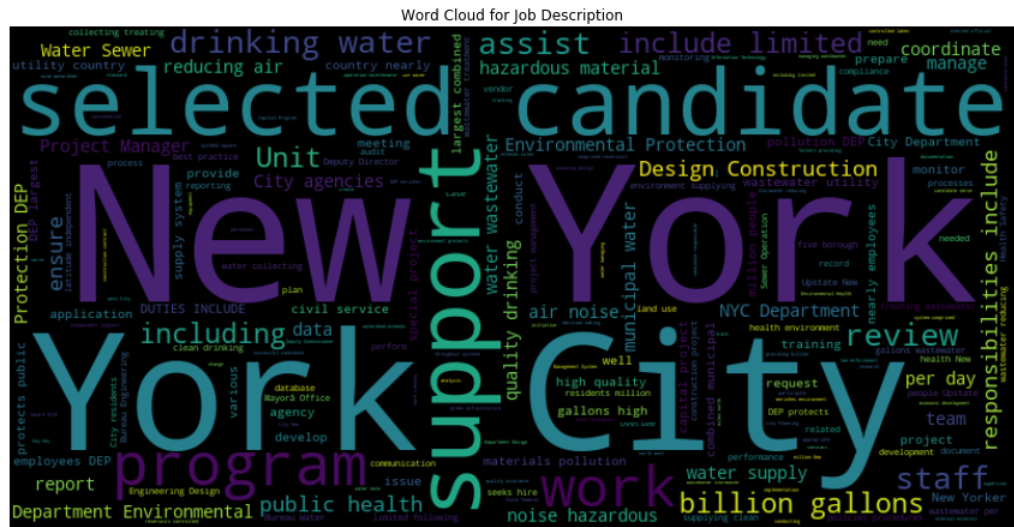


Figure 10. Job Description Wordcloud

Figure 11 shows that word cloud is generated for Preferred Skills column, it can be observed that “Ability”, “work”, “experience”, “communication skills” are the most frequent word. So, it can be interpreted that the job posting is done requires work experience and good communication skills. Whereas, certain words like “handle multiple”, “strong written”, “management system” are the least frequent words, so it can be inferred that jobs related to these kinds of words are less posted.

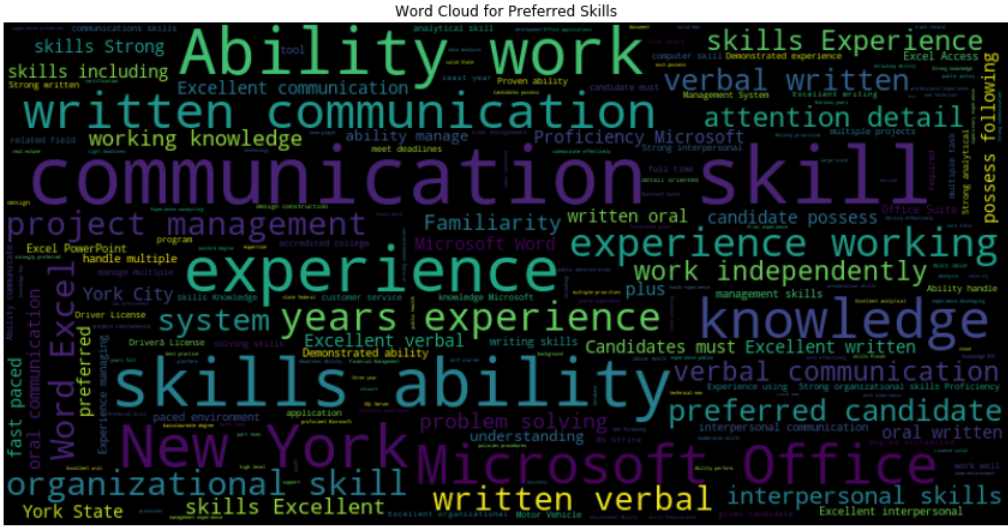


Figure 11. Preferred Skills Wordcloud

Figure 12 shows that the word cloud is generated for the Job Instruction column, it can be observed that “Job”, “ID, “NYC”, “gov” are the most frequent word. So, it can be interpreted that most of the jobs are related to the government and is available in New York City.

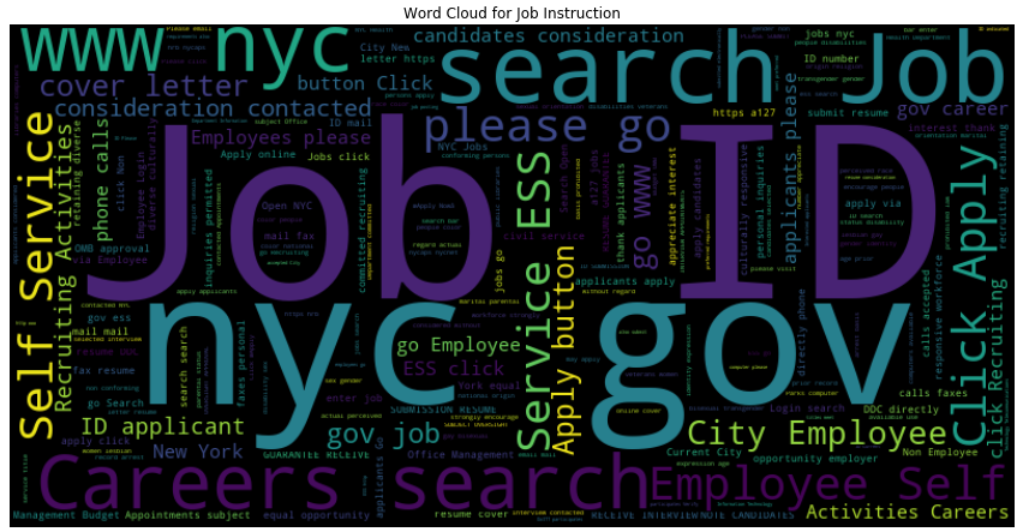
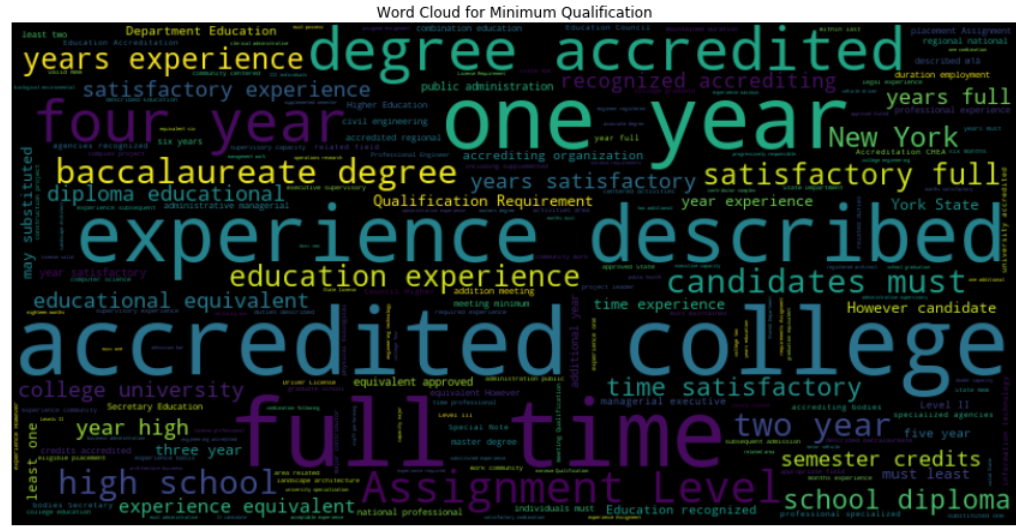


Figure 12. Job Instruction Wordcloud

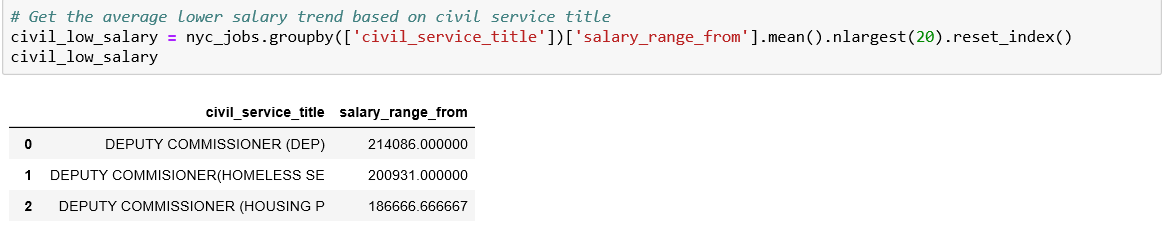
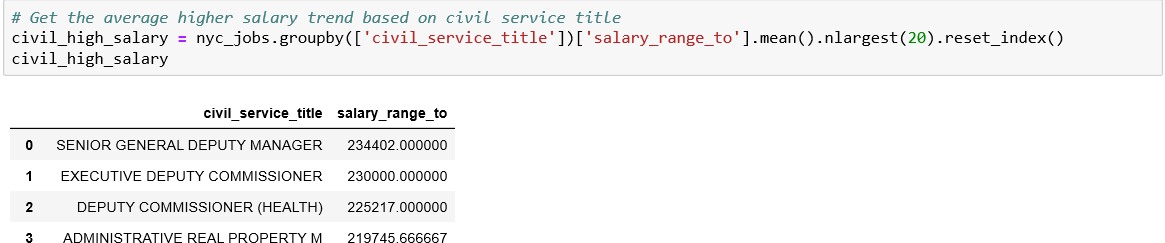
Figure 13 shows that the word cloud is generated for Additional Information column, it can be observed that “public”, “service”, “civil”, “Loan” is the most frequent word. So, it can be interpreted that the jobs are related to the public service industry.

Figure 13. Additional Information Wordcloud

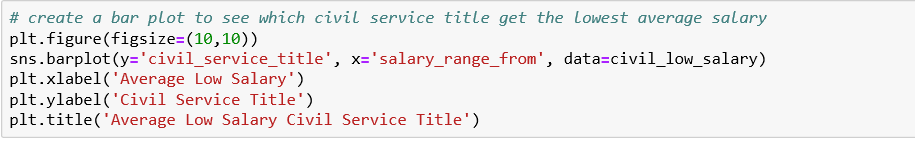
Figure 14 shows that the word cloud is generated for Minimum Qualification column, it can be observed that “one”, “year”, “college” are some of the most frequent words. So, it can be interpreted that the job posting requires a college degree and working experience.

Figure 14. Minimum Qualification Wordcloud

The data was further analyzed using some of the other attributes like a salary from, salary to, civil service title and business title. Based on the salary columns, the average higher and lower salary was generated using pandas group\_by function based on civil service title.

Using the data generated by using pandas group\_by function, bar charts are plotted to get the average low salary and average high salary for civil service title. Figure 15 shows that Deputy Commissioner (DEP), Deputy Commissioner (Homeless Service), Director of Research and Policy are some of the titles which are part of the low salary range. Though, Administrative Real Property and Director of Investigation are in the lower range of the same chart.



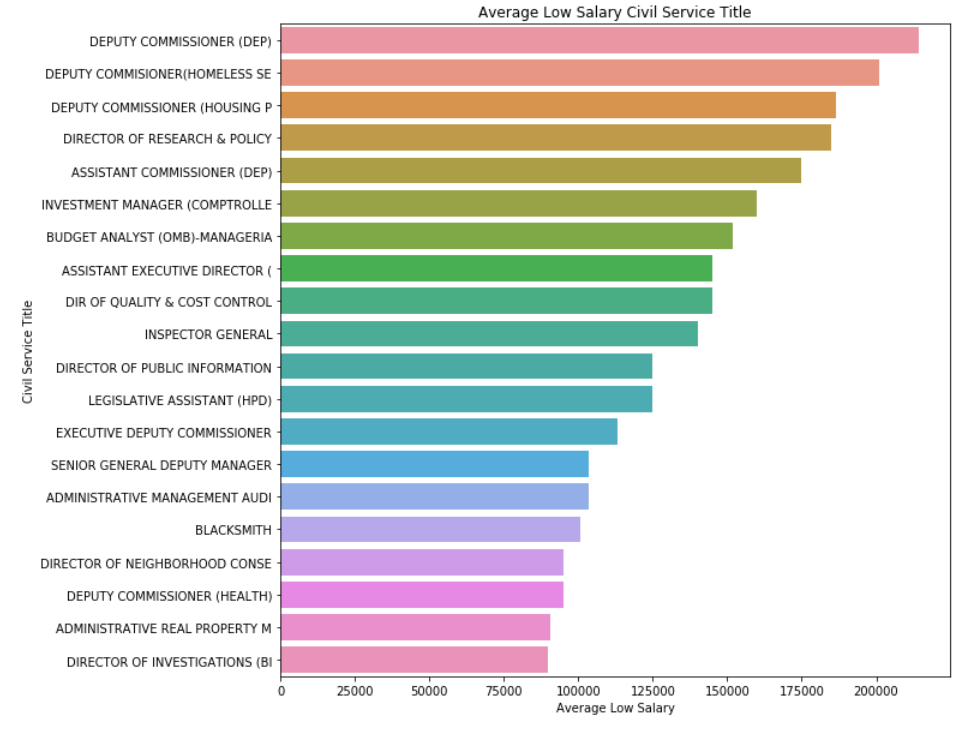
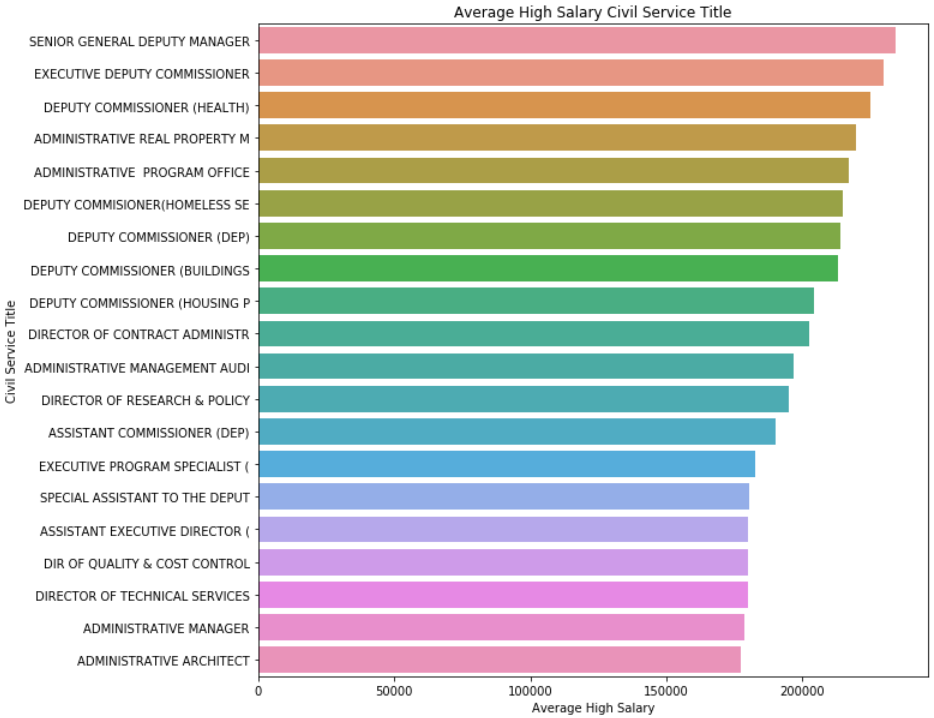


Figure 15. Average Low Salary Civil Service Title

Figure 16 shows that Senior General Deputy Manager, Executive Deputy Commissioner are some of the titles which are part of the high salary range. Administrative Manager and Administrative Architect are in the lower range of the same chart.

 Figure 16. Average High Salary Civil Service Title

The data is sliced to get the average low and high salary based on business title using the salary frequency to find out the business title get the highest annual, hourly and daily salary.



Figure 17 shows that there is not much difference between the business titles which are part of the annual high salary range. The few top ones are Executive Vice President for Operation, Executive Deputy Commissioner, Division Mental Hygiene, Budget Analyst. Executive Deputy Commissioner for Operations, Deputy Commissioner Water and Sewer Operations are in the lower range of the same chart.

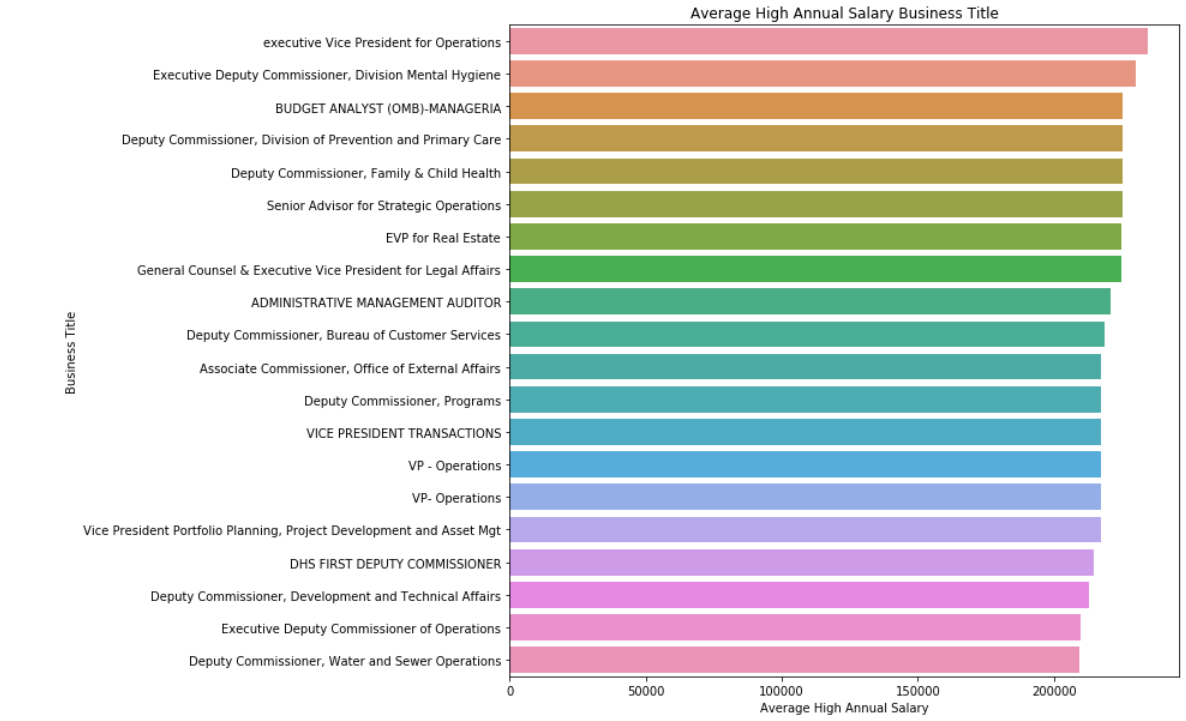
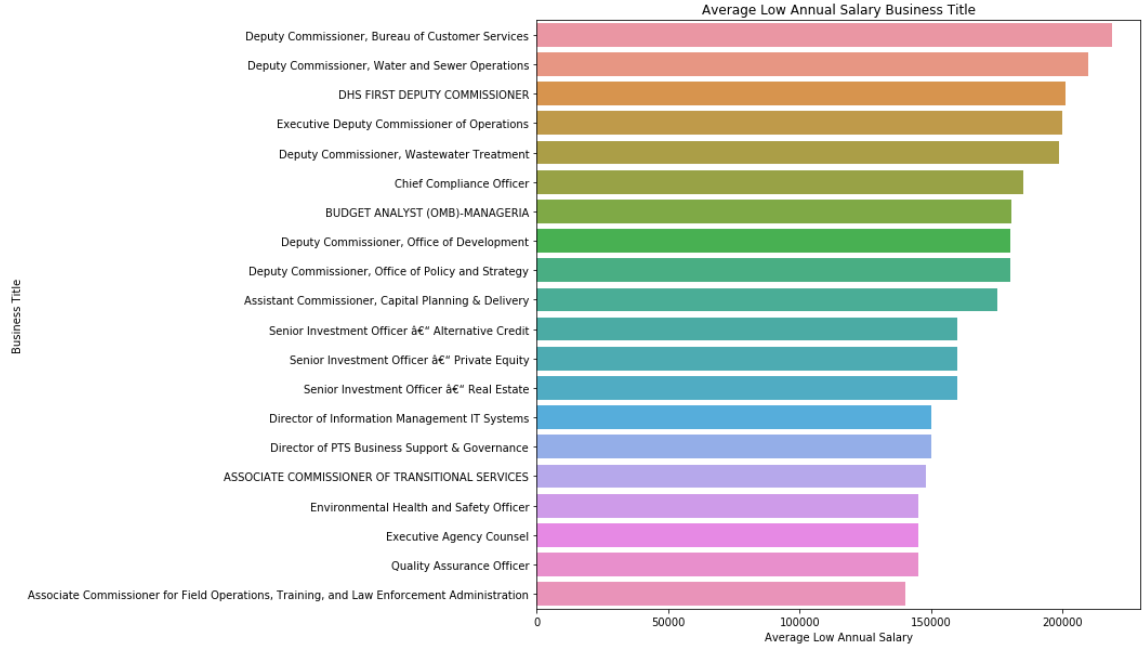
 Figure 17. Average High Annual Salary Business Title

Figure 18 shows that Deputy Commissioner – Bureau of Customer Services, Deputy Commissioner Water and Sewer Operations are at the top of the same chart. Whereas Quality Assurance Officer, Associate Commissioner for Field Operations are at the lower range of the chart.

Figure 18. Average Low Annual Salary Business Title

While plotting the hourly salary for the business titles, it can be observed that there is not much difference between the average lower salary range and average higher salary range. Figure 19 shows that Inter-agency data manager, City medical specialist are some of the business titles present at the top of the chart and School nurse and public health nurse are at the lower range of the chart. Figure 20 shows that Inter-agency data manager, City medical specialist are some of the business titles present at the top of the chart and public health nurse is at the lower range of the chart. Figure 21 and Figure 22 are the bar plots for average high daily salary and average low daily salary respectively for business titles, both the plots don’t show much difference. So, it can be concluded that business titles fall under hourly and daily salary ranges don’t have many differences in comparison to an annual salary range. But referring the salary frequencies and business titles, the job posting can be grouped by high-level, mid-level and low-level application.

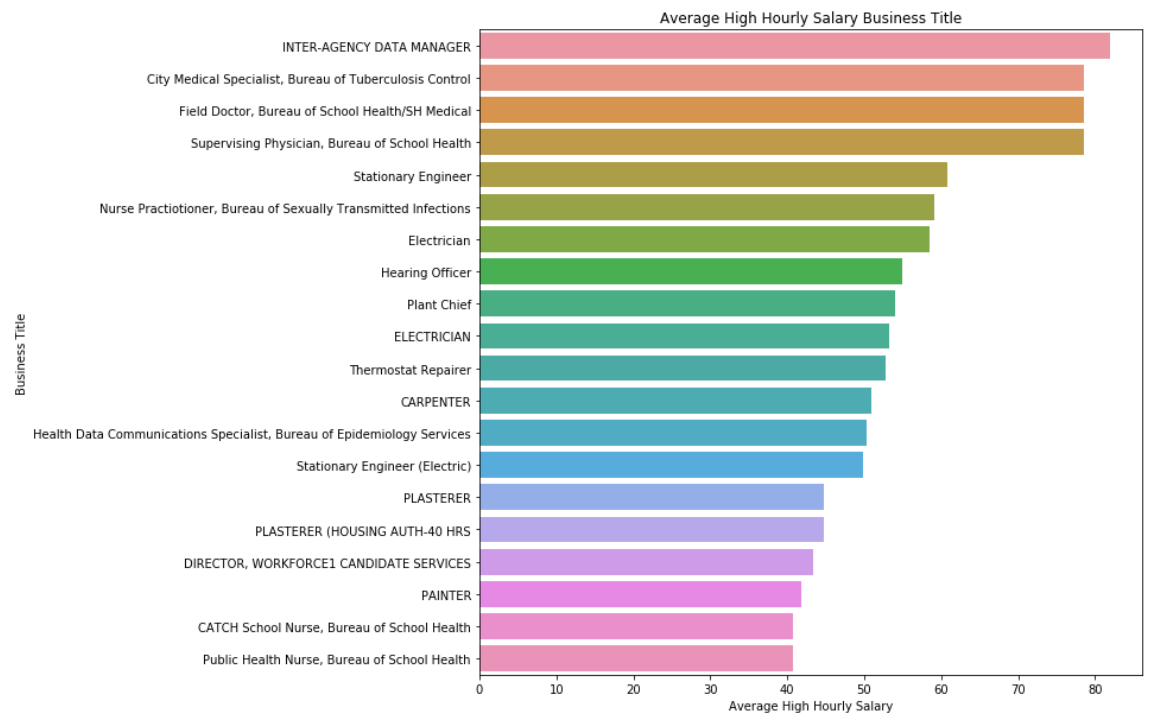


Figure 19. Average High Hourly Salary Business Title

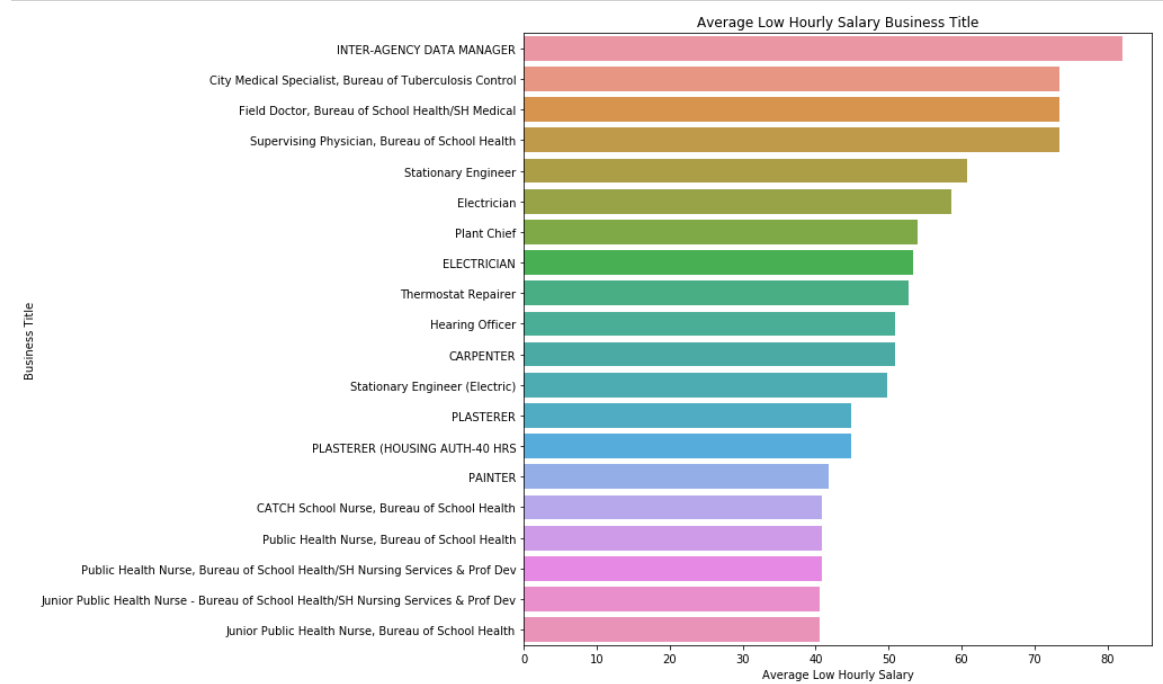
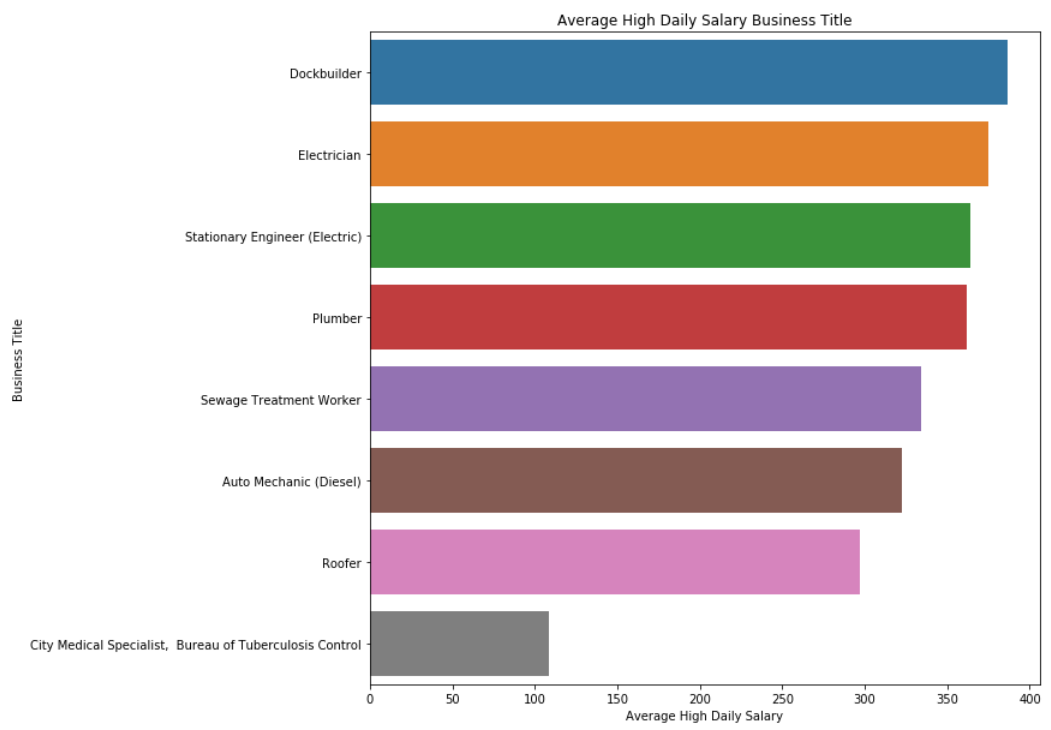


Figure 20. Average Low Hourly Salary Business Title

 Figure 21. Average High Daily Salary Business Title

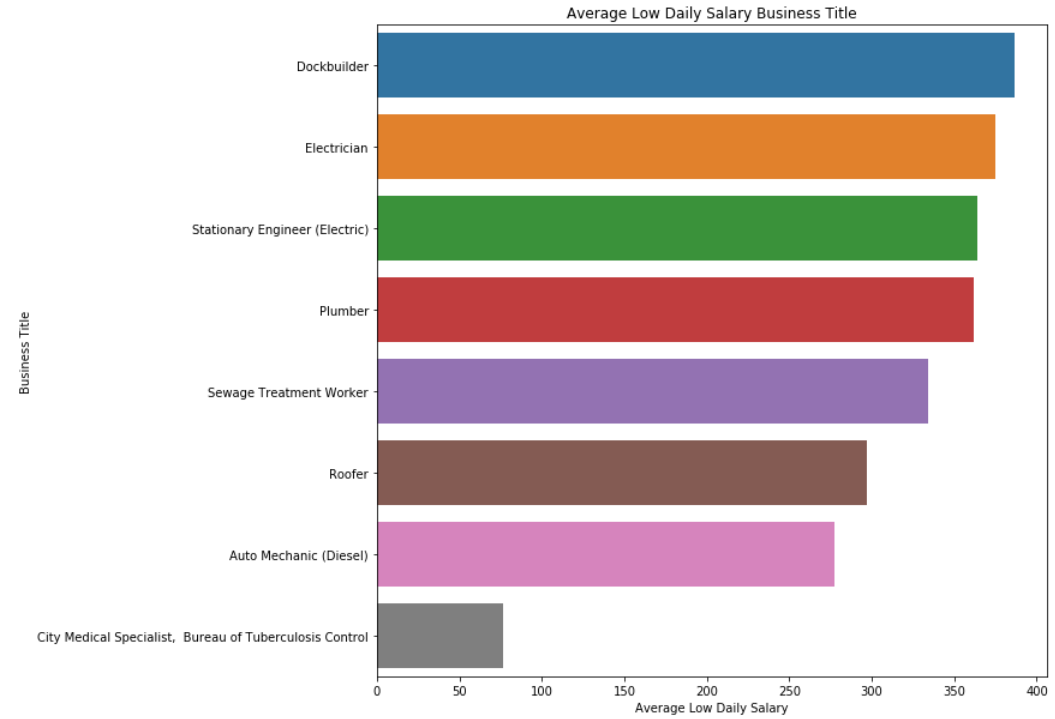
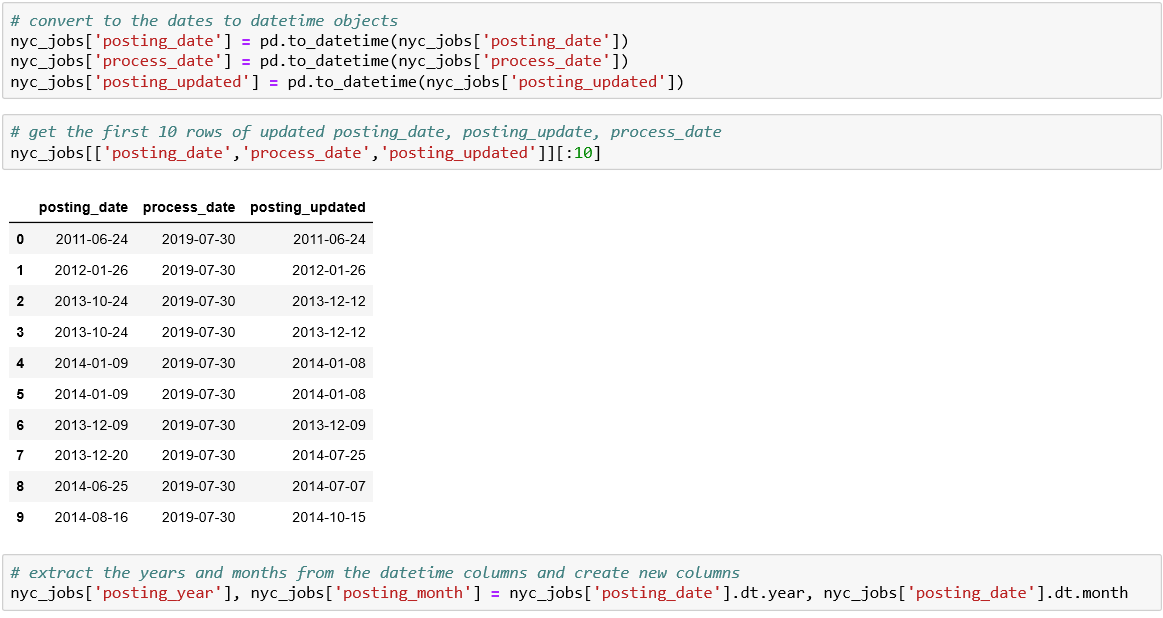


Figure 22. Average Low Daily Salary Business Title

To further analyze the data, the date columns – posted date, process date, and posting date were converted into date-time objects. From the posted date column, the year and the months are extracted and added as two separate columns.



Using the pandas value\_counts function, the unique counts for the years are generated and count plot was created. Based on the count plot in Figure 23, it can be observed that year 2019 has the highest number of counts which says that most of the job posting is done in 2019, whereas 2013 and 2014 are having the lowest job posted.

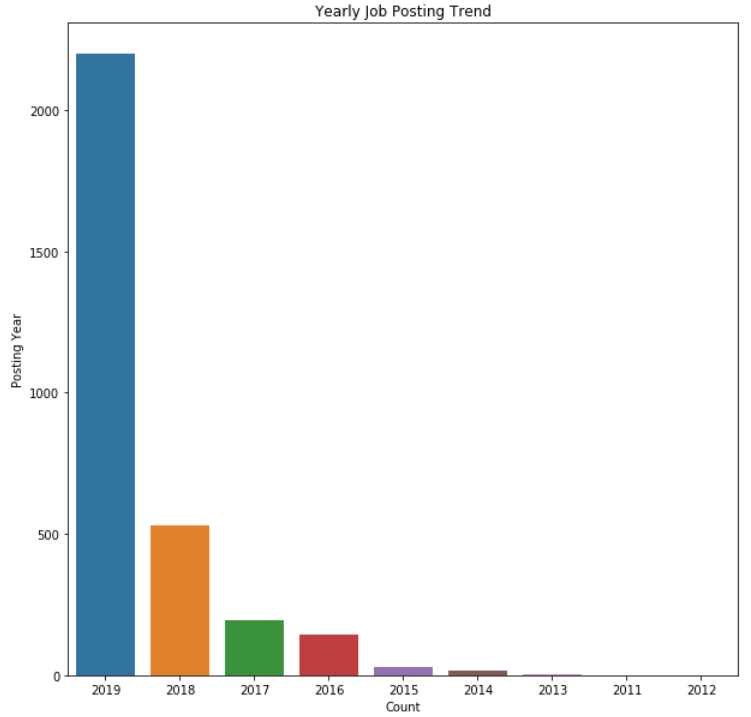
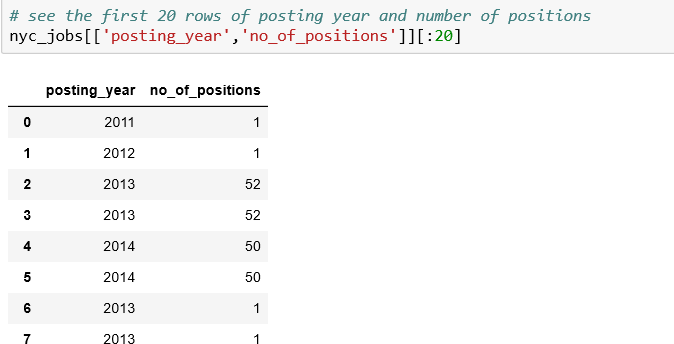


Figure 23. Yearly Job Posting Trend

To further analyze, how many positions opened in each year, the data is sliced using the number of positions and year columns.



From Figure 24, it can be observed that even though most of the job was posted in 2019, the number of positions was high in the year 2013 in comparison to other years.

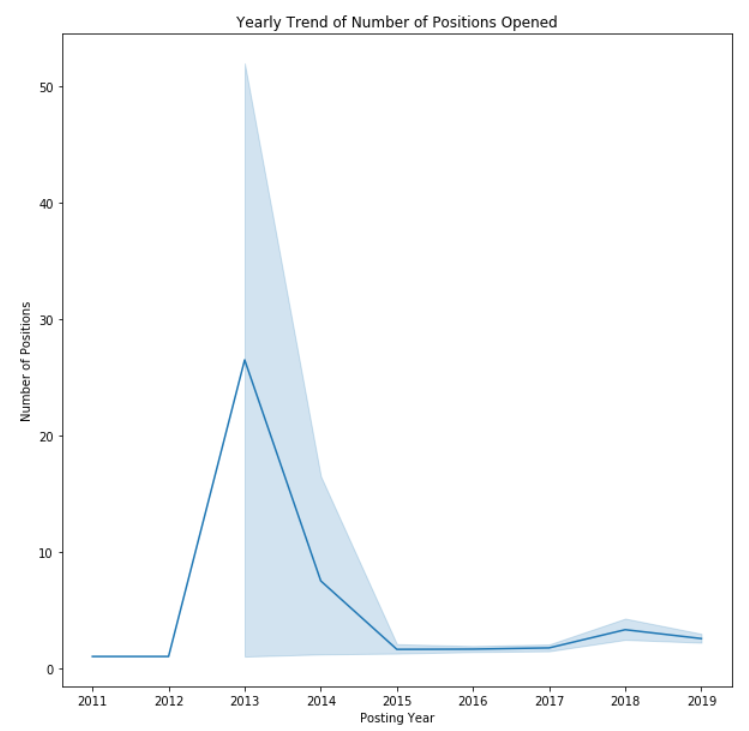


Figure 24. Yearly Trend of Number of Positions Opened

The data were further analyzed to see the job posting trend for every month. From Figure 25, shows that most of the job posting was done on the month of July, followed by June and May. September has the lowest job posting. To do more analysis, the number of positions opened are checked for each month. Figure 26, shows that January and December have the most positions opened, and September has the lowest number of open positions.

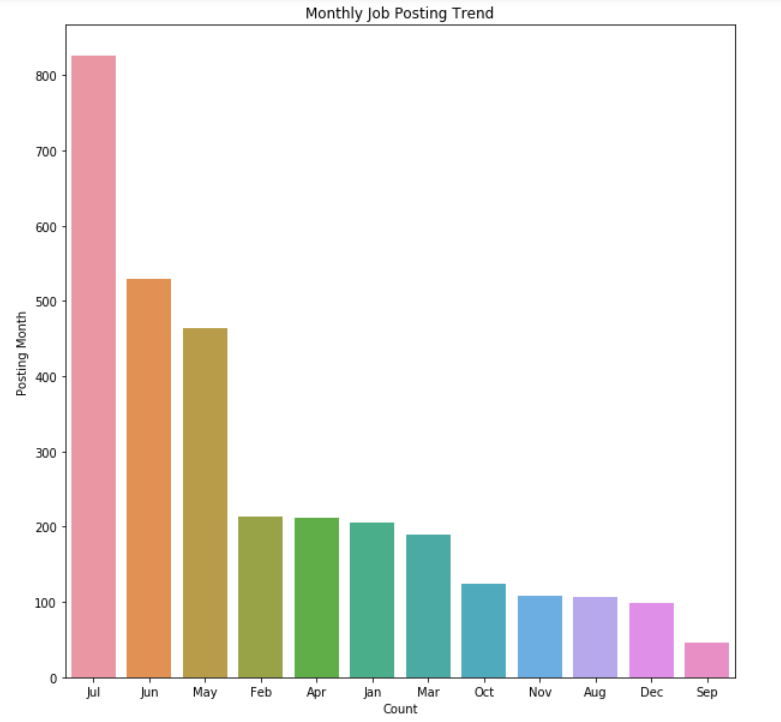
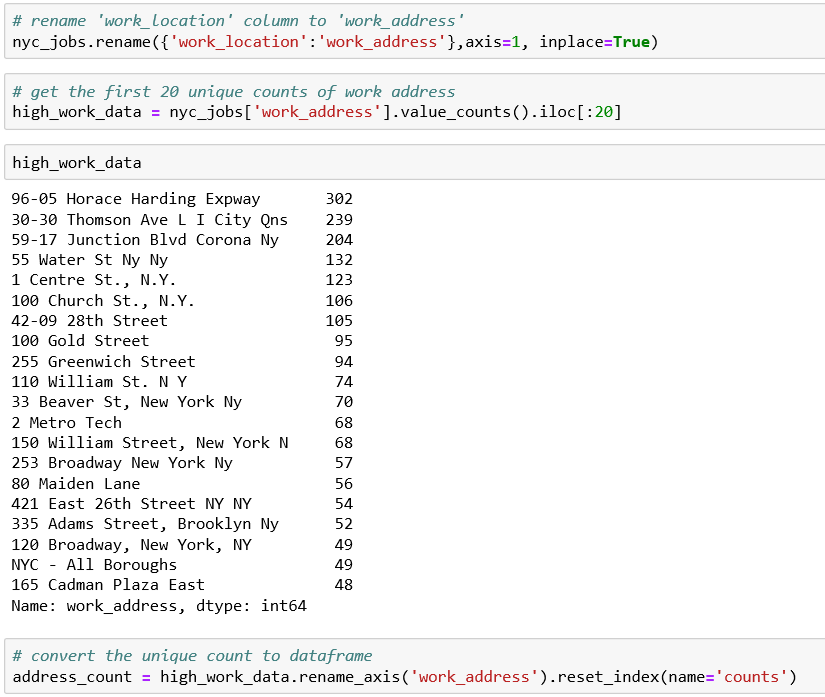


Figure 25. Yearly Trend of Number of Positions Opened

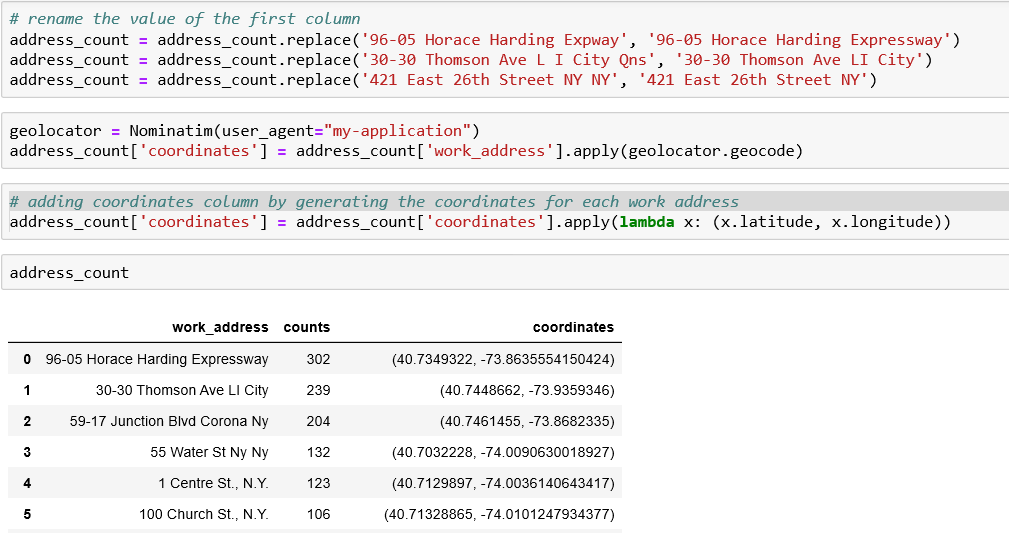


Figure 26. Monthly Trend of Number of Positions Opened

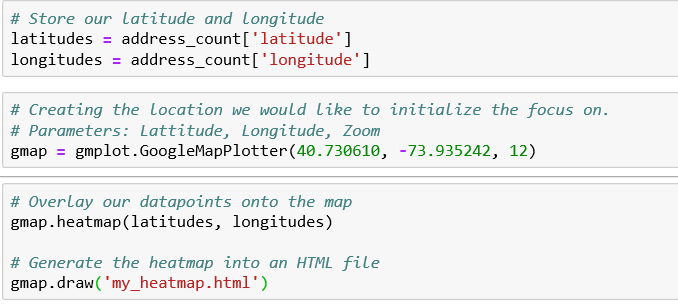
To do the map-based spatial analysis, the work location column is renamed to work address and the value of the first 20 address are collected.



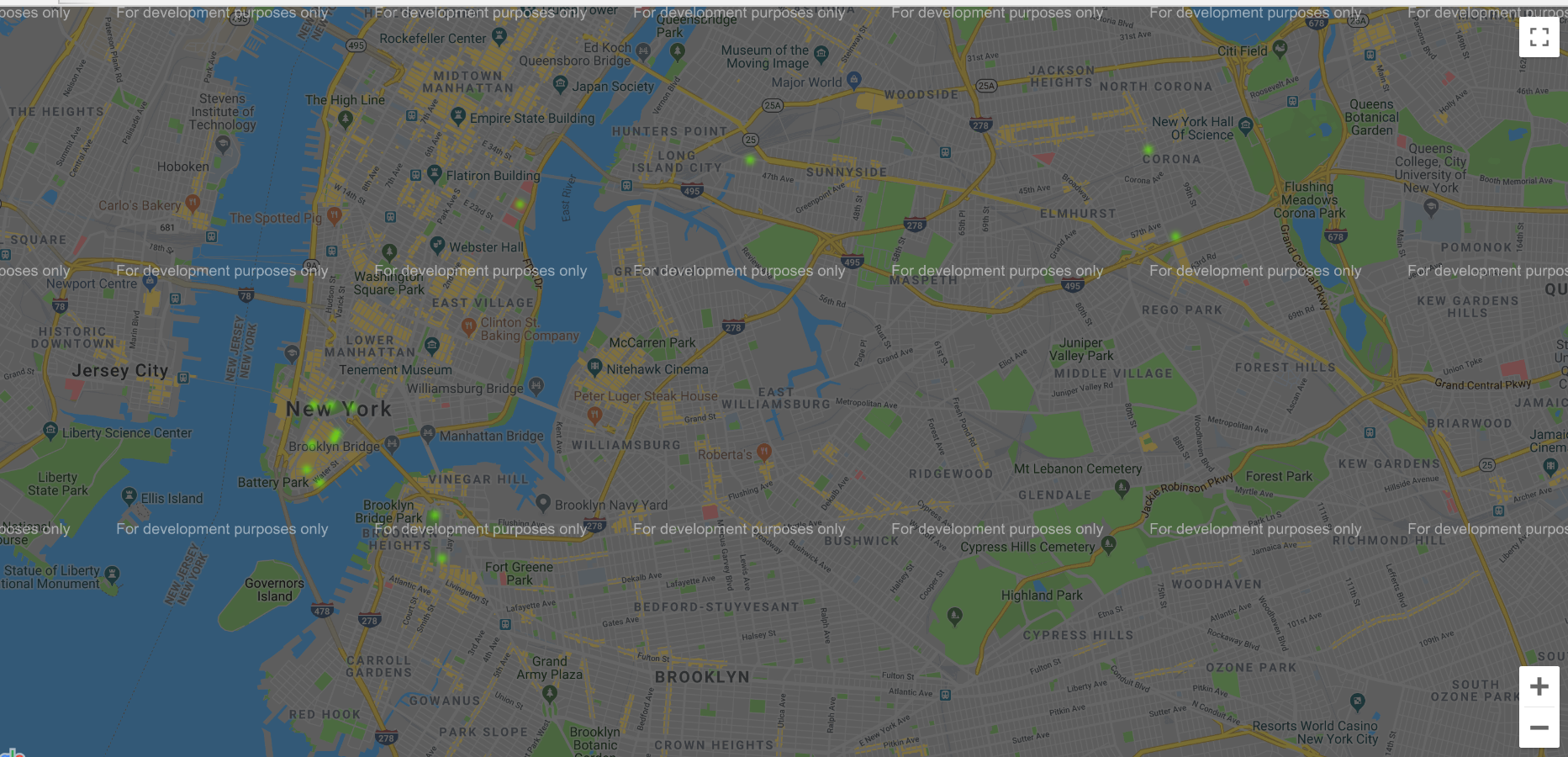
The work address and the counts were then converted to dataframe. Using the geopy library, the coordinates of the address were generated. The coordinates column was then broken into two more columns – latitude and longitude.



Then, using gmplot library, the heatmap is generated using the work address’s latitude and longitude



The heatmap was then generated in Figure 27 and was highlighted on the New York map. As Google API was not working, the map generated was for “development purpose only”

 Figure 27. Heatmap for Job Posting in New York City

**4. Conclusion**

Overall this work aims to provide a detailed analysis of the job posting in New York City. The analysis is multifaceted, covering the unstructured data like job categories or preferred skills as well as structured data like business titles, number of positions. To do the analysis with the data set, most of the time the first twenty rows is selected to analyze the trend. According to the descriptive analysis, the massive amount of data reflects several valuable information. The internal job posting is much more in comparison to external posting, similarly, most of the jobs are targeted for the full-time applicant in compare to part-time applicants. Also, most of the jobs listed are for annual salary range which shows that mostly high-level job market is targeted. Without doubt, Architecture and Engineering are the two most preferred job categories. Also based on the word clouds, it can be easily deduced that most of the job posting requires a college education and work experience, and the job posting is primarily based on New York City location.

It is concluded that most of the job posting targets high-level annual salary jobs which are related to Architecture and Engineering. Most of the companies try to absorb internal employees to the vacant positions, before publishing for external applicants. So, it can be concluded that the companies try to utilize the existing employees, to avoid the additional cost in terms of interviewing and evaluating new candidates and doing background checks. Department of Environment Protection is the top agent and post most of the jobs. When referred to the yearly job market, it seems most of the posting was done on 2019, but when number of positions is compared yearly, it shows that 2013 has the highest number of vacancies, so it can be concluded that even though jobs openings are more, but it was never filled up due to some reason which cannot be extracted from this dataset and may need additional data. Referring to the monthly job posting trend, it can be found that July has the greatest number of jobs posted, but when the number of positions is analyzed using month column, it shows that the positions are mostly opened during December and January. So, it can be concluded that even though most of the company fiscal end-year happened in December, but they still look for some positions to fill up during the winter season. This process may get extended till January which is when the fiscal year starts and company try to absorb more employees for the open positions during that time. The data need to be further analyzed to find out some additional descriptive analysis like top skills and qualifications required for each job category. Also, predictive analysis can be done using predictive models to find out if certain job posting can be predicted based on the dataset.