Data Analysis & Visualisation

Of Data Analyst Jobs in USA

Nitika

Sept 17, 2020

Ext ID : x178710

COMPSCI 434-007 Programming Python

INTRODUCTION

Who is a Data Analyst ?

A Data Analyst interprets data and turns it into information which can offer ways to improve a business, thus affecting business decisions. Data Analysts gather information from various sources and interpret patterns and trends – so the job of a Data Analyst mainly highlights the analytical nature of the role.

As we know during this pandemic many people lost their jobs, with this dataset it is possible to hone the job search so that more people in need can find employment.

Data Description

The dataset I have used is from kaggle site.

The link : https://www.kaggle.com/andrewmvd/data-analyst-jobs

This dataset contains more than 2000 job listing for data analyst positions (all assumed to be open positions at the time the dataset was published in July 2020), with features such as:

- Unnamed
- Job Title
- Job Description
- Company Name
- Salary Estimate
- Location
- Headquarters
- Founded
- Size
- Rating
- Industry
- Sector
- Revenue
- Easy Apply
- Type of ownership
- Competitors

Objectives

- What kind of jobs get higher salaries? (Job Title)
- What kind of companies pay more? (Rating, Company, Sector etc)
- Does job location matter to salaries?
- Average salary in top companies.
- Number of jobs available in state or city

Data Cleaning

There are some unnecessary columns that need to be cleaned and dropped off.

1) Removed the unnamed column and the Job Description column as it contains the huge text as they aren't needed in this analysis

2) Salary Estimate :

   2.1) Removed the string '(Glassdoor est)' string from the salary estimate column as it is no.

   2.2) Split the column into two column that is Min Salary and Max Salary

3) Location : Location column has city and state information , so will split the location into two columns that are location_city and location_state, so that we can do some analysis over state or city wise.

4) Company Name : In Company Column we can see that data have \n3.8 as a suffix where 3.8 is rating of company and we will eliminate this and only keep the Company name in that column because we already have a separate Rating Column.

5) Removed all the null values , since the csv contains some '-1' , ' -1.0' values as null,

   So we will replaced these values with 'NONE'

6) Removed the ‘employees’ string from Size column

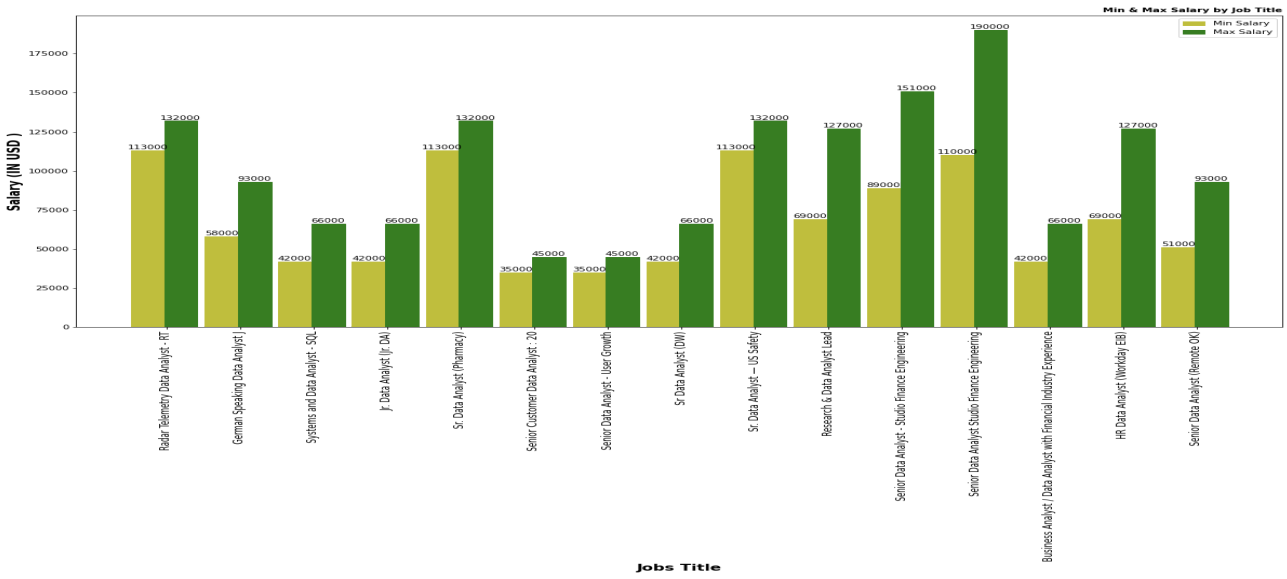
Data Analysis & Visualization

Tier 1 Analysis:

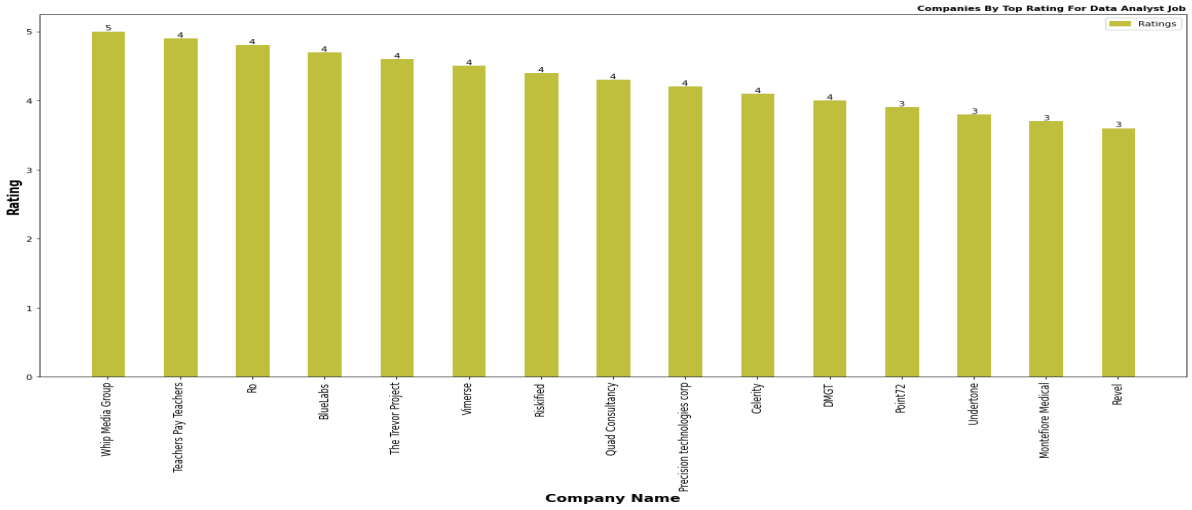
- Number of Data Analyst Jobs In USA : 2251
- Number of company currently open for hiring : 80
- Average Salary for Data Analyst Jobs in USA : 72,106 USD
- Minimum Salary for Data Analyst Jobs in USA : 24,000 USD
- Maximum Salary for Data Analyst Jobs in USA : 1,90,000 USD

Tier 2 Analysis:

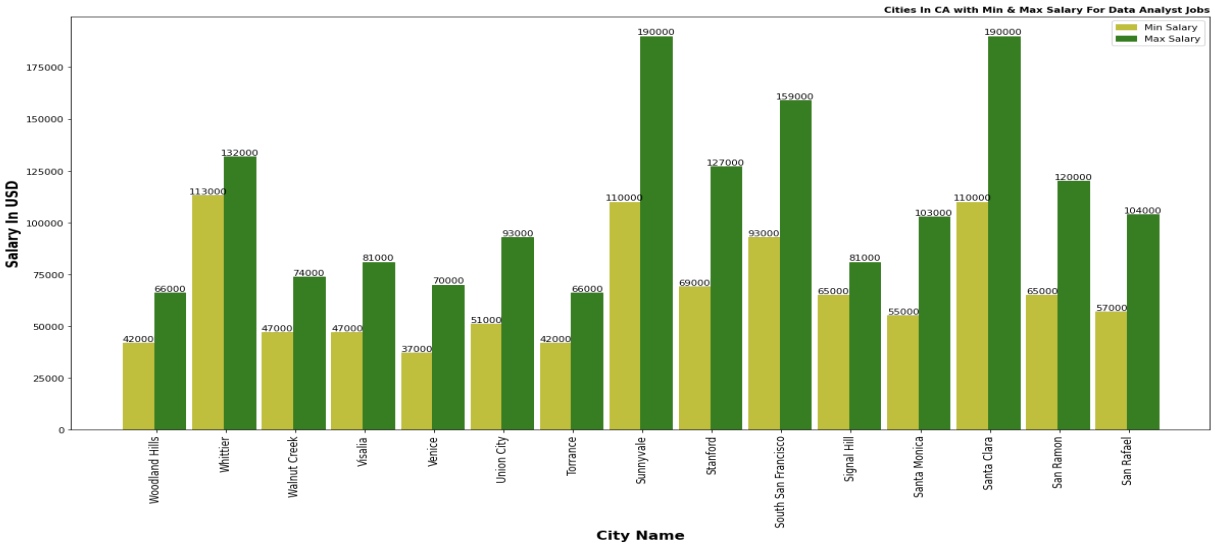
1) This analysis include the min and ,max salary of company by the job title



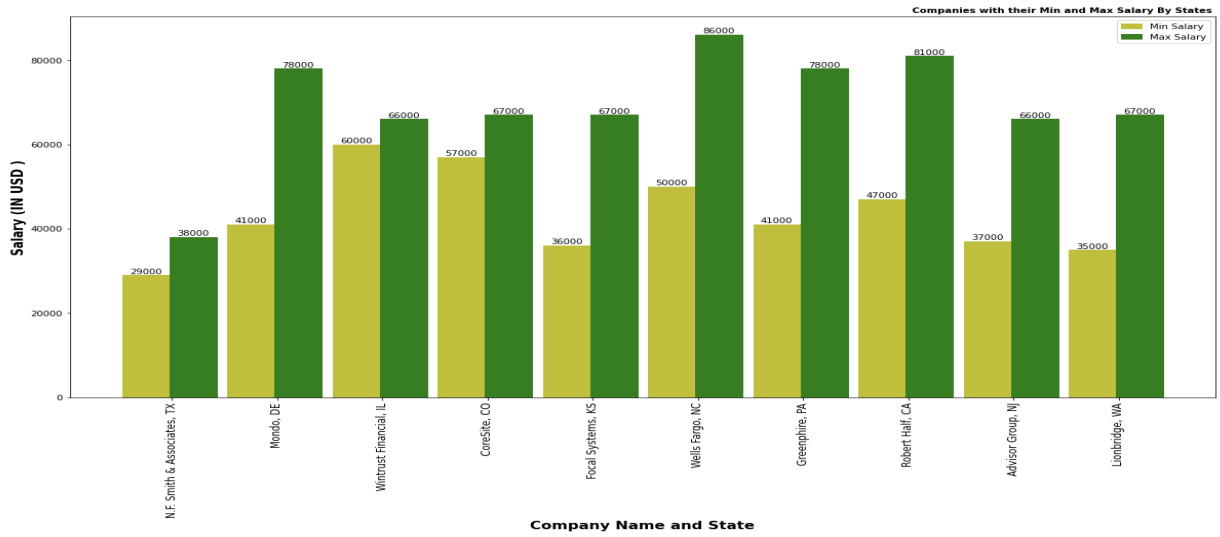
2) This analysis include the top 10 companies having top ratings in USA



3) This analysis includes cities in california with min and max salary

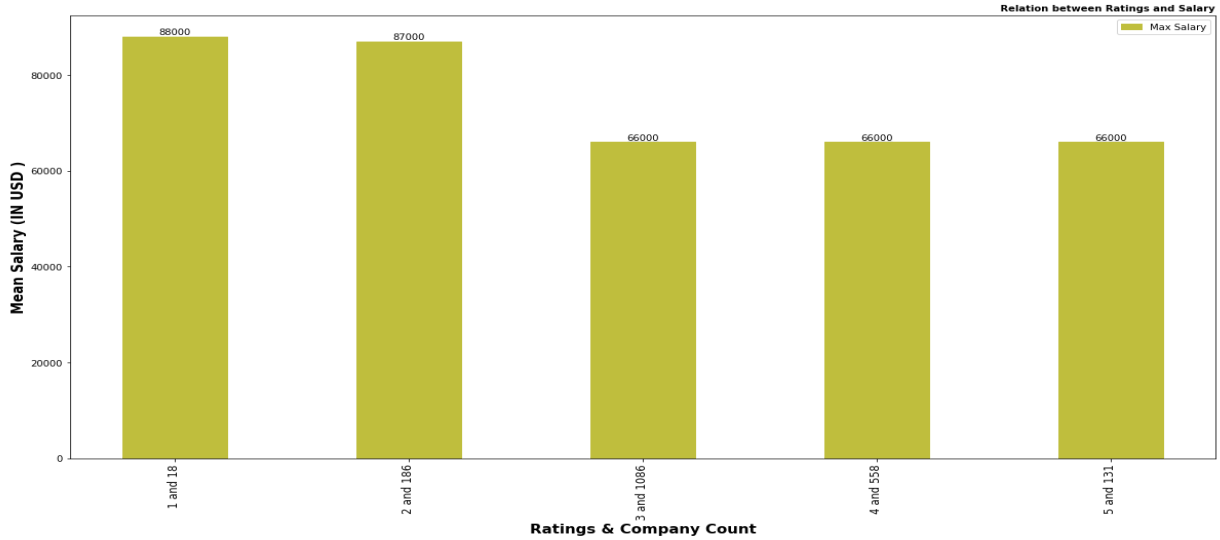


4) This analysis includes the state wise companies with their min and max salary



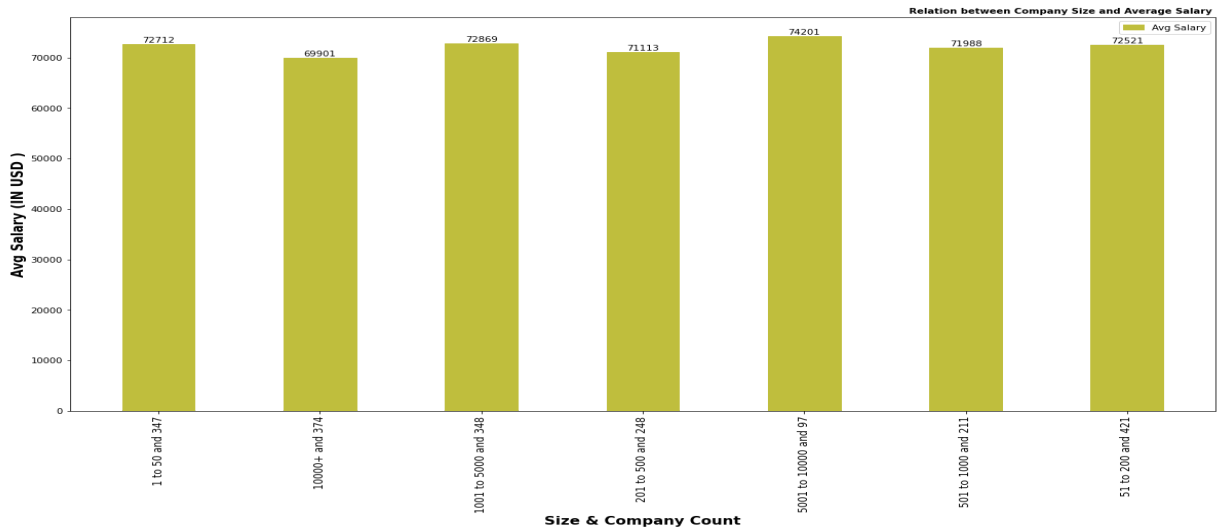
Tier 3 Analysis:

1) This analysis show if there is any relationship between the ratings of the company and the salary given to the data analyst :



We can see that the salary of the data analyst is not directly proportional to rating.

2) This analysis see if size of the company matter while giving the salary:



We can see that the size of the company does not decide what salary is to be given.

Limitation and Assumption

- The results only reflect the outcome at the time the dataset was published, which is presumed to be July 2020. Seasonal variation is disregarded (not a time-series data).
- The salary estimates come from Glassdoor, which may not reflect the actual salaries.
- The dataset is assumed to reflect the traits of the actual job market.

Future Enhancements:

- Job Description Columns have a huge amount of text data which is not included in this analysis But we can use this for future analysis.
- For Revenue Column , we can split it into maximum and minimum revenue for further analysis, this can also be included in future analysis.