mjrotter4445 / DA_Analysis_README_1stdraft- Public

☆ 0 stars   0 forks

☆ Star   Unwatch ▾

Code   Issues   Pull requests   Actions   Projects   Wiki   Security

main ▾   ···

mjrotter4445 Update README.md ...   3.

Create new file

Upload files

View code

README.md

Data Analyst Jobs Analysis *1stdraft*

Data Analyst Dataset analysis with Python, Pandas, NumPy and Matplotlib

Project Overview For this project we are analyzing one or two datasets from Kaggle Dataset Repository.

- One dataset is: Data Analyst Jobs - from 1 year ago - approx. 2253 records total.
- Second dataset is:

Attribute Information:

- Job Title

- Rating
- Company Name
- Industry
- Sector
- Salary Minimum - added this column
- Salary Maximum - added this column
- Salary Average - added this column
- ADDED Salary Low Bottom of Range
- ADDED Salary Med Middle of Range
- ADDED Salary High Top of Range
- ADDED Salary Level considerations
- Python Jobs - we could add this
- SQL Jobs - we could add this
- Excel Jobs - we could add this
- Tableau jobs - we could add this

EDA (Data Investigation Data Cleaning & Feature Engineering) EDA (Exploratory Data Analysis) as in “A first look at the data” is used to understand and summarize the content of the dataset, such as initial look at the columns, data types, data quality, data statistics and data relationships. Moreover, data often require a significant amount of work to make it suitable for analysis like cleaning, feature engineering and visualizing. Python packages such as Pandas, NumPy and Matplotlib help work faster and more efficiently when performing EDA.

Data Processing In order to perform sufficient data analysis data needs to be manipulated. In this analysis I used Pandas functions such as groupby, query and binning amongst others. Data processing and data manipulation is essential for any data analysis and with a thoughtful approach we can get useful insights about the given dataset and recommendations beyond this analysis. Research Questions As mentioned above EDA is an important step in data analytics. This critical step can save roughly 15–50% of time on a project because it provides a targeted plan for how to clean, sort, and create smaller datasets that are easier to work with. It is also extremely important to familiarize with the dataset, what various features mean and what values represent. Based on that we can conclude what questions can be answered from the data or do we need to collect more data in order to provide comprehensive analysis.

In this analysis I will be focusing on the following questions:

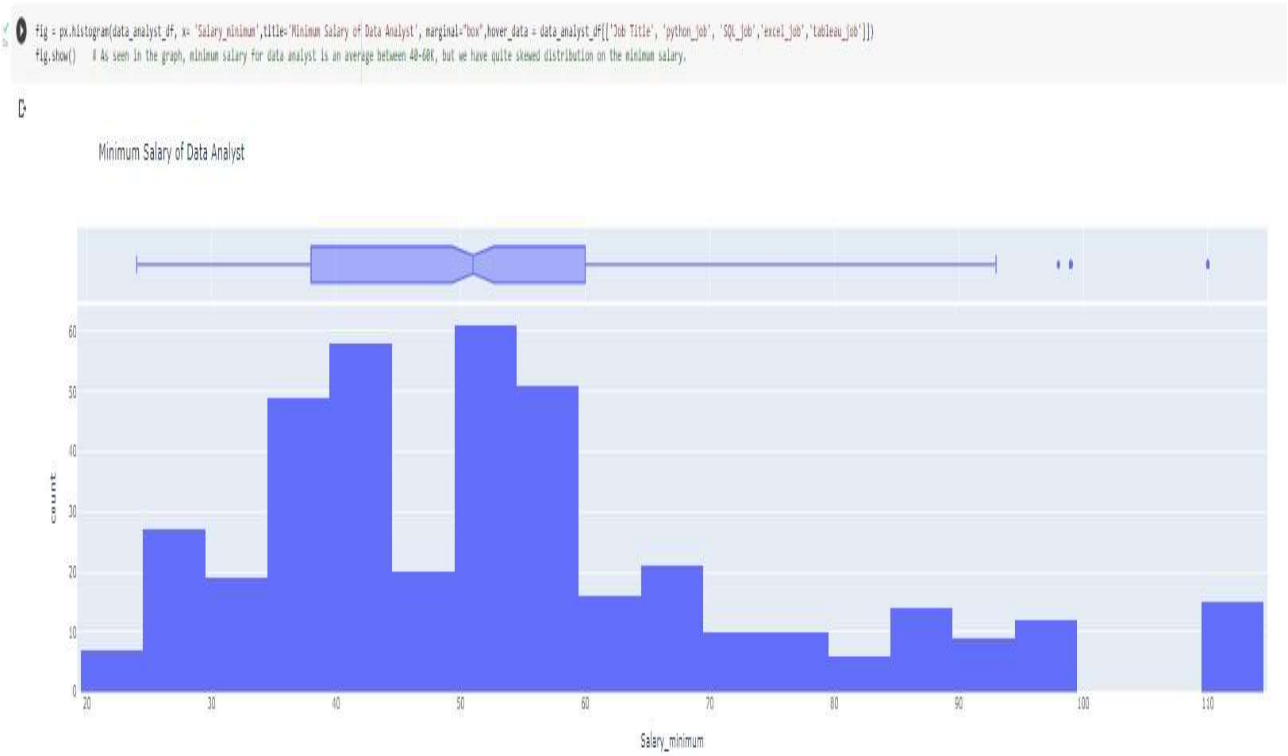
- � What is the Minimum Salary?

- � What is the Maximum Salary?
- � What is the Average Salary?
- � What are the number of Job Openings by Job Title?
- � What are the number of Job Openings by Industry?
- � What are the number of Job Openings by Sector?
- � Correlation between Variables?
- � What attributes are most important in predicting the salary of a job?
- � Is a certain type of ______________associated with _____________?
- � Do companies with higher ratings pay higher salaries?
- � Do companies with lower ratings pay lower salaries?

Answering Questions with Data and Drawing Conclusions

With Python and Matplotlib we can plot a histogram in one line of code, and observe the frequency of distribution for various features. We can read frequency on y-axis and feature values on the x-axis. We can easily plot histograms for the entire dataset, for all features as well as individual features to get a more detailed look. From the histograms below we can see,..................................... The code for histograms and correlation can be found here. >>>

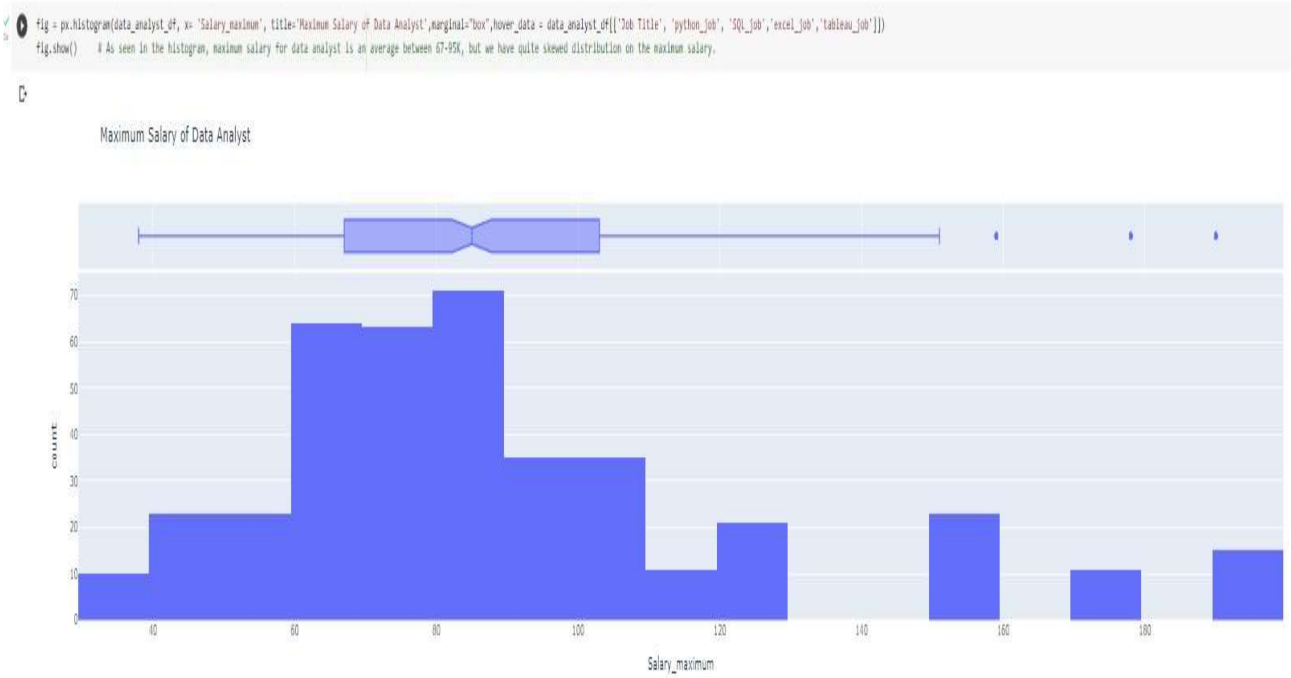
- � What is the Minimum Salary?



```
fig = px.histogram(data_analyst_df, x= 'Salary_minimum',title='Minimum Salary of Data Analyst', marginal="box",hover_data = data_analyst_df[['Job Title', 'python_job', 'SQL_job','excel_job','tableau_job']])
fig.show()    # As seen in the graph, minimum salary for data analyst is an average between 40-60K, but we have quite skewed distribution on the minimum salary.
```

Figure 1 -Min Salary is between 40-60K

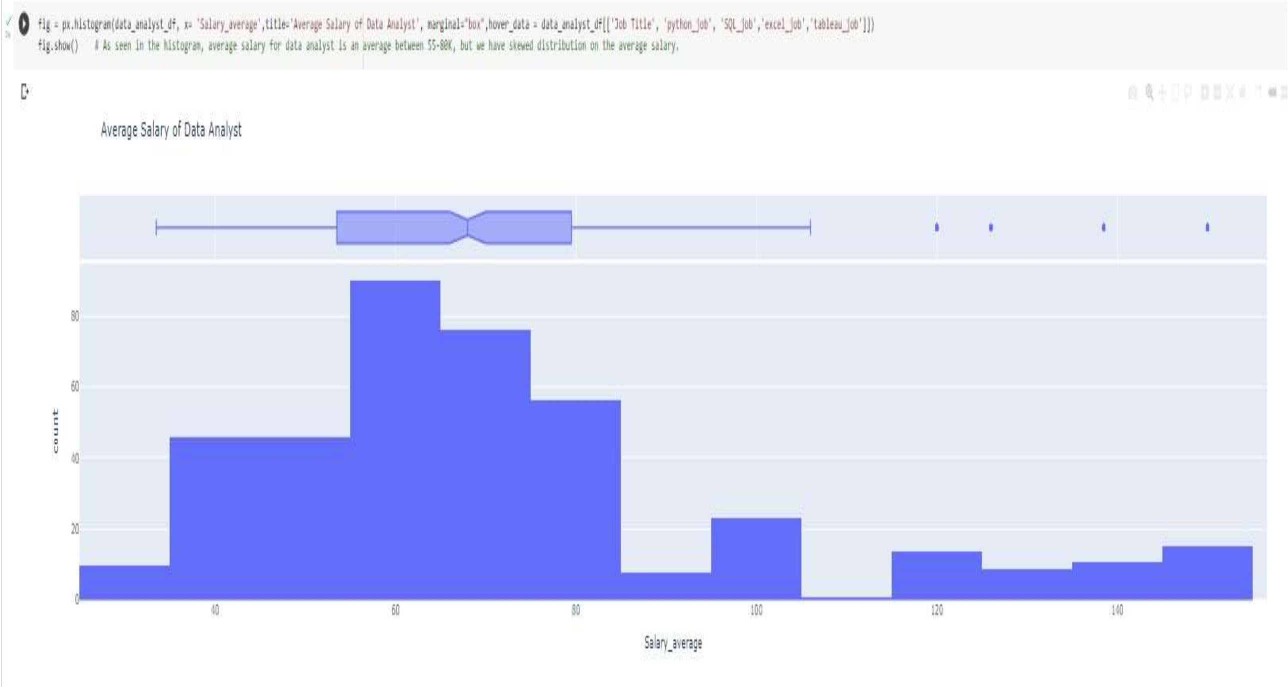
- � What is the Maximum Salary?



```
fig = px.histogram(data_analyst_df, x= 'Salary_maximum', title='Maximum Salary of Data Analyst',marginal="box",hover_data = data_analyst_df[['Job Title', 'python_job', 'SQL_job','excel_job','tableau_job']])
fig.show()    # As seen in the histogram, maximum salary for data analyst is an average between 67-95K, but we have quite skewed distribution on the maximum salary.
```

Figure 2 -Max Salary is between 67-95K

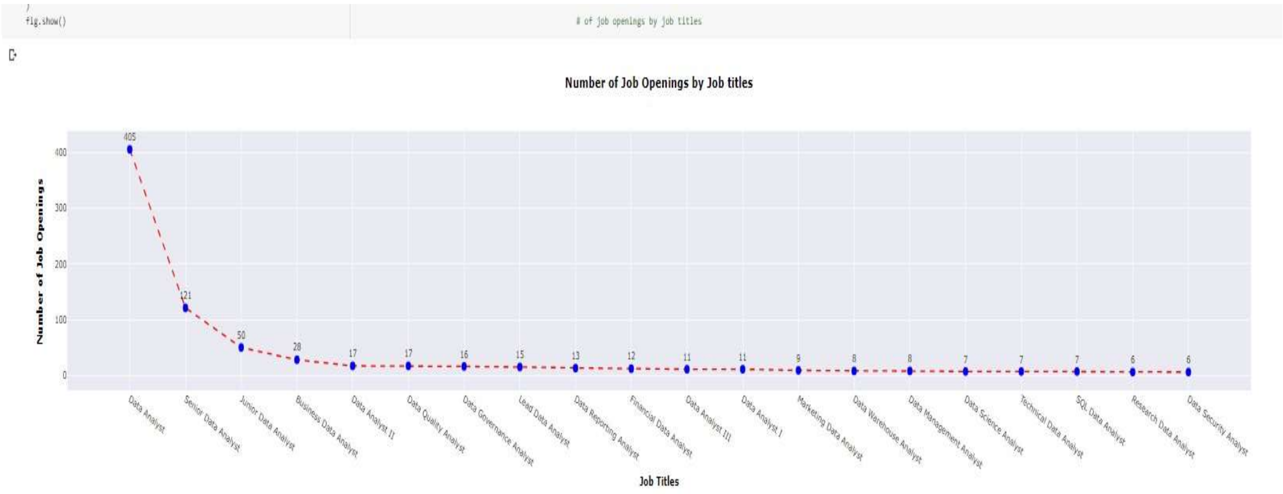
- � What is the Average Salary?



```
fig = px.histogram(data_analyst_df, x= 'Salary_average',title='Average Salary of Data Analyst', marginal="box",hover_data = data_analyst_df[['Job Title', 'python_job', 'SQL_job','excel_job','tableau_job']])
fig.show()    # As seen in the histogram, average salary for data analyst is an average between 55-80K, but we have skewed distribution on the average salary.
```

Figure 3 -Avg Salary is between 55-80K

- � What are the number of Job Openings by Job Title?



```
)
fig.show()                                      # of job openings by job titles
```

Figure 4 -Most common used titles in job advertisements are Data, Senior,Junior, Business

- � What are the number of Job Openings by Industry?

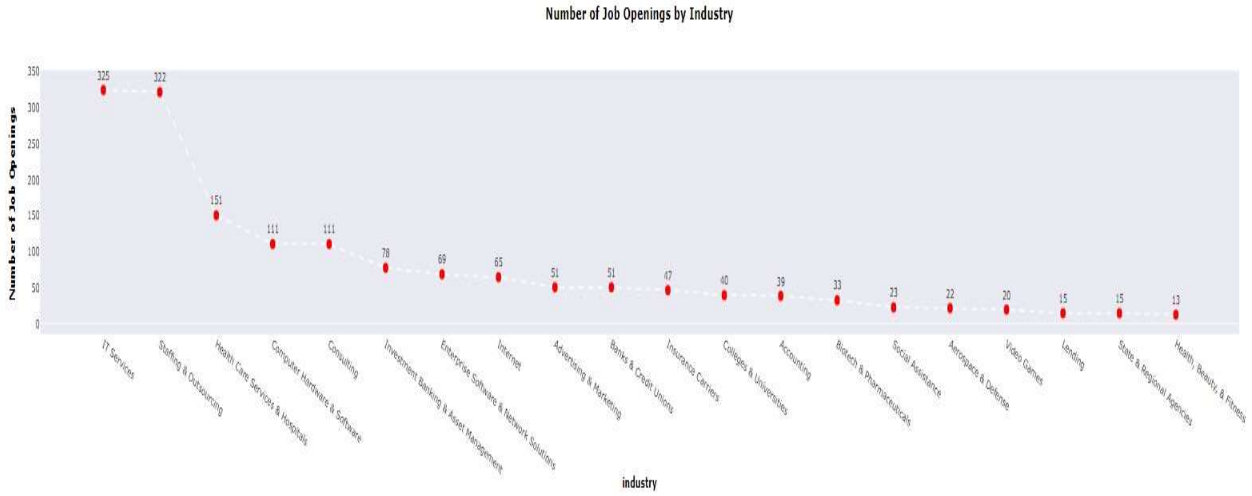


Figure 5 -Most Common

- � What are the number of Job Openings by Sector?

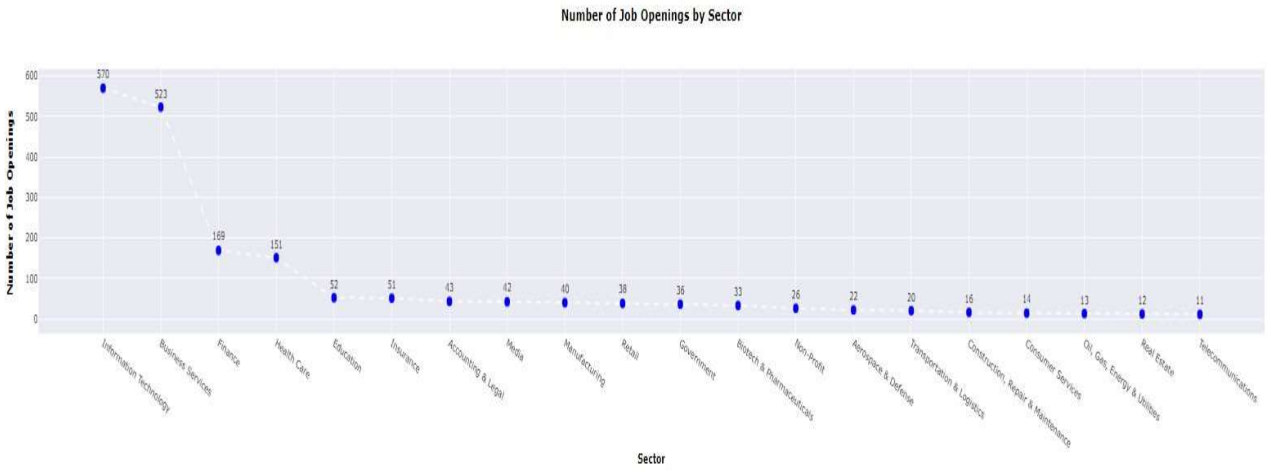


Figure 6 -Most Common

DISTRIBUTIONS & CORRELATIONS

- � What are the distributions based on Minimum Salary, by Company?

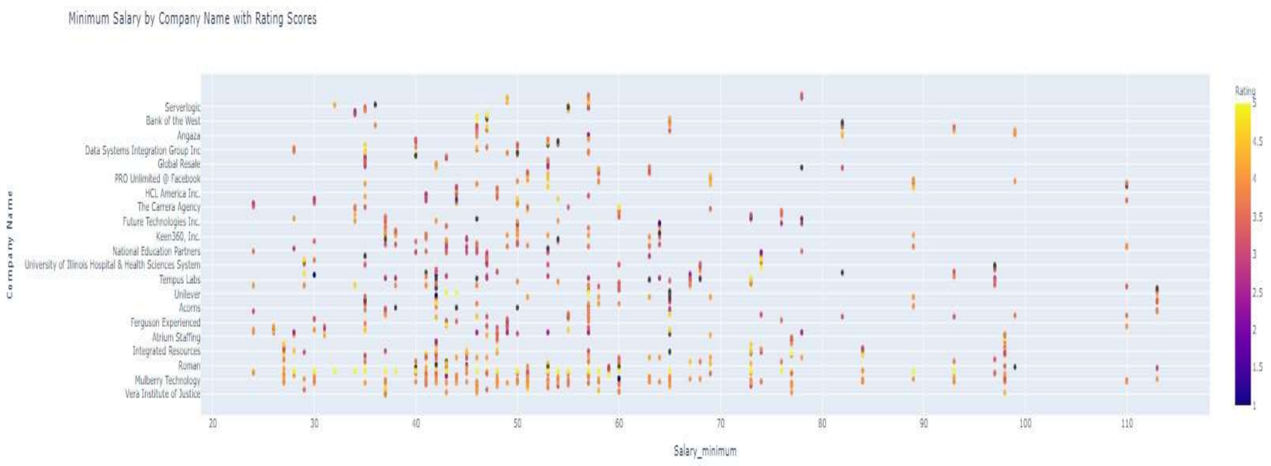


Figure 7 -best at min

- � What are the distributions based on Maximum Salary, by Company?

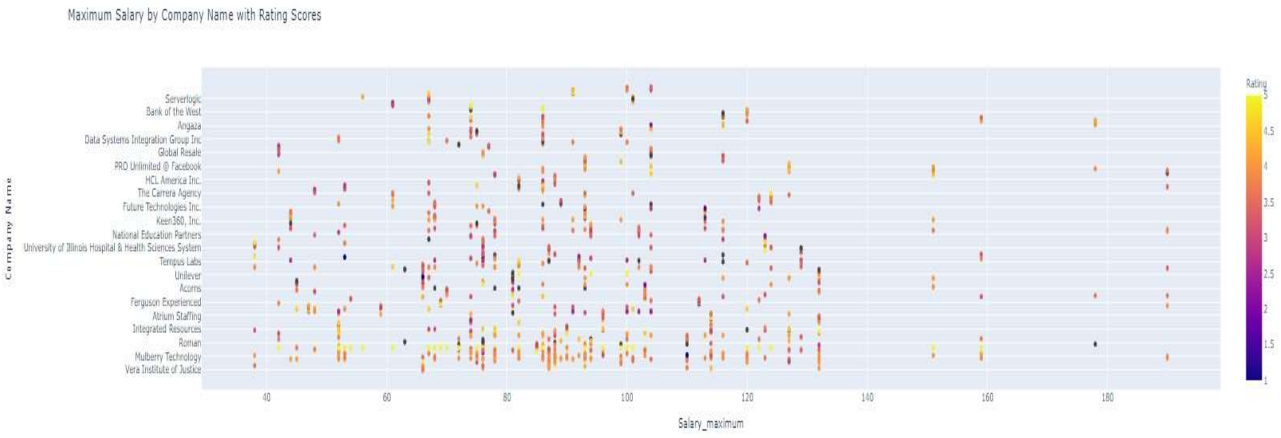


Figure 8 -best at max

HEAT MAP IT
A great way to explore, to get familiar with the data, finding patterns and building intuitions is to calculate, visualize and uncover complex and unknown relationships between variables. With the *corr function* we can plot the entire dataset and explore correlations between variables. For example correlation between _________and _____ is _____. Low values indicates weak positive correlation, while higher values indicates almost no correlation.

Figure 4 could be a Correlation between variables - HEAT MAP type

What attributes are most important in predicting the salary of a job?

Releases

No releases published

Create a new release

Packages

No packages published

Publish your first package