**PROJECT 3 REPORT**

**Glassdoor Jobs (2017-18)**

**Introduction**

What is the best location in the country for job opportunities?

This study leads interested candidates to answers in the form of data analysis and assessment, based on the job opportunities posted on Glassdoor.com.

**Data**

Glassdoor data was obtained from Kaggle.com:

<https://www.kaggle.com/datasets/thedevastator/jobs-dataset-from-glassdoor>

The file name: eda\_data.csv

In Jupyter Notebook, a Python file ‘glassdoor\_Data’ was created to ‘clean’ the dataset.

The cleaned data has 742 records (job postings), and 14 features, viz.,

1. 'job\_state',
2. 'job\_region',
3. 'Sector',
4. 'Company\_Age',
5. 'Avg\_Company\_Size',
6. 'Job\_Title',
7. 'Avg\_Salary',
8. 'Max\_Salary',
9. 'Usr\_Rating',
10. 'Python\_y\_n',
11. 'R\_y\_n',
12. 'Spark\_y\_n',
13. 'AWS\_y\_n',
14. 'Excel\_y\_n'

A SQLite database is created and a table Glassdoor\_clean is created, in the database, with 14 features. The SQLite file is accessed in Jupyter notebook for analysis using Python.

Re-formatted JSON files are also created for analysis using JavaScript.

**Data Analysis**

An overview of job prospects around the country (USA) in graphs:

Figure 1 shows box-and whiskers plots of all continuous numeric features in the dataset. Each box includes 50% of the data within a feature.

Observations:

* Most companies are between the ages of 10 years and 60 years. Ages of some newly founded companies are not listed.
* Most average annual salaries range between $75k and $130k.
* Most maximum average salaries range between $90k to $160k.
* Most employment sizes of companies range between 500 to 7500.
* Most Glassdoor user ratings of a company, defined in the interval of 0 to 5, is between 3.3 and 4.

These ranges when viewed, give a general understanding of the overall data, without other categorical specifications, such as ‘state’ and ‘region’ where the jobs are located.

Figure 1

A diagram of a company

Description automatically generated

Figure 2 is a violin plot displaying the types of jobs by title and corresponding salary range. All titles except those of ‘Manager’ and ‘Director’ have outliers toward higher ‘Average Salary’, indicating the availability of jobs with salaries higher than the norm.

The title ‘Research Scientist’ displays salaries, both lower and higher than the normal range.

Figure 2

A graph showing different colored rhombuses

Description automatically generated

Figure 3 is a bar chart of average annual salary by sectors. 24 sectors are listed in the dataset. The bars are in descending order of salaries. The 4 most paying jobs are in ‘Media’, ‘Accounting & Legal’, Information Technology and ‘Biotech & Pharmaceuticals’, all of which have more than $100k as average salaries. ‘Construction, Repair & Maintenance’ is the lowest paying of all the listed sectors, with the average annual salary of less than $30k.

Figure 3

A graph of a number of people

Description automatically generated with medium confidence



Figure 4 follows Figure 3 in showing the number of jobs available in each sector.

‘Information Technology’ tops the list followed by ‘Biotech & Pharmaceuticals’. Both of these sectors are in the top 4, both in terms of job numbers and salary. ‘Agriculture & Forestry’ has the least number of jobs.

Figure 4

A graph with different colored bars

Description automatically generated

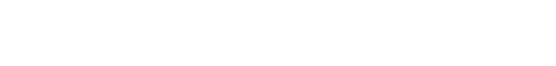


Figure 5 combines Figure 3 and Figure 4, to create a bubble chart to assess the number of jobs and the average salary for a sector at a glance. The height of the bubble from the x-axis shows the salary and the size of the bubble represents the number of jobs posted for a particular sector.

The chart clearly shows that ‘Information Technology’ (IT) and ‘Biotech & Pharmaceuticals’ (Biotech) are two of the highest and largest bubbles.

Figure 5

A graph showing different colored circles

Description automatically generated

Figure 6 is a bar chart that shows the number of available jobs by types and sectors. This shows that the number of job types like ‘Data Scientist’, ‘Data Analyst’ and ‘Data Engineer’ are available in most sectors, the greatest numbers being in ‘IT’ (gray). ‘Research Scientist’ is most needed in ‘Biotech’ (violet) sector. Title ‘Machine Learning Engineer’ (MLE) is available in sectors, ‘Education’, ‘IT’, ‘Aerospace & Defense’ and ‘Finance’.

Figure 6

A graph of different colored lines

Description automatically generated

Figure 7 Shows the Glassdoor user ratings data versus the age of the company. The pyramid shape suggests a normal curve for the data. The mean /median is as user-rating of 3.7. Well established companies older than 100 years received user rating close to the median value of 3.7. This could be due to a large number of people employed by a company over the long period of existence of the company resulting in a large number of reviews and ratings leading to the average rating close to 3.7.

Younger companies less than 100 years old may not have a large set of review/rating data, hence the wide range of ratings ranging from 1.9 to 5, for companies less than 25 years old.

The size of the bubbles is proportionate to average salaries offered by companies. A general trend of lower ratings for lower salaries and higher ratings for higher salaries can be deduced.

Figure 7

A graph of a company age

Description automatically generated

Figure 8 is a pie chart showing 5 geographical regions of the USA (courtesy National Geographic) represented by the slices. The size of a slice is equivalent to the percentage of average salary for the region out of the total amount around the country.

The following is a list of regions and corresponding state codes of states that belong to each region:

* 'West': WA, MT, OR, ID, WY, CA, NV, UT, CO, AK, HI,
* 'Southwest': AZ, NM, TX, OK,
* 'Midwest': ND, SD, NE, KS, MN, IA, MO, WI, IL, MI, IN, OH,
* 'Southeast': KY, WV, VA, AR, TN, NC, SC, GA, AL, MS, LA, FL,
* 'Northeast': PA, NY, VT, ME, NH, MA, RI, CT, NJ, DE, MD, DC

The chart shows that the earning capacity in ‘West’ exceeds that of the rest of the regions. Each region, however, may include states with very high paying jobs along with states with low paying jobs.

Figure 8

A pie chart with different colored circles

Description automatically generated

Figure 9 is a bar plot showing the average salary vs. company size, i.e., the number of employees. It can be observed from the chart that there is no correlation between the average salary and the number of employees in a company.

The average salary for the companies with number of employees ‘not listed’, offer the highest pay. Some such companies are relatively new and are ≲ 10 years old. Some such companies are: Persivia, Kronos Bio, ALIN, Monte Rosa and Muso.

Figure 9

A graph of pink bars

Description automatically generated

From all the above data visualization charts it can be inferred that the highest paying jobs as well as the largest number of jobs are in the sector: **Information Technology**. All ‘IT’ jobs require the knowledge of one or more of the following Programming Languages or Application Packages listed in the Glassdoor dataset:

* Python,
* R,
* Spark,
* Amazon Web Services (AWS)
* Excel

Therefore, in the following plots, data is explored to analyze the earning potential with programming skills.

In Figure 10, each plot is a histogram of job-count versus salary for each programming skill. All plots except one for ‘R’, resemble a ‘normal’ curve skewed right. This implies that the average value of salary moves to the right and salary range is extended to the right.

The jobs requiring such skills offer much more than the median value of salary listed here:

* Python: $107k
* R: $70k
* Spark: $108k
* AWS: $107k
* Excel: $92k

These median values are also the mode values with the largest number of jobs.

Figure 10

A group of graphs showing different colored lines

Description automatically generated with medium confidence

Figure 11 illustrates the mean values of salaries for jobs requiring one or more of the 5 programming skills. Knowledge of Python, Spark and AWS are among the highest paying skills.

Figure 11

A graph of blue bars

Description automatically generated with medium confidence

Figure 12 illustrates a combination of employee satisfaction and average salary for jobs with each skill. The blue dots in each scatter plot represent jobs with a particular skill for the plot. In the plots for Python, Spark and AWS, the number of blue dots increases for salaries higher than the average value as well as the median ‘usr\_rating’.

Figure 12



A group of blue and orange dots

Description automatically generated

Choosing **one** particular programming language, the number of jobs and salary can be observed in the following charts. **Python** is chosen as an example.

In Figure 13, the number of jobs that require Python are compared with the number of jobs that do not, for each sector. Knowledge of Python could be attributed to obtaining a high paying job, as can be noted from the following charts.

The number of jobs available that uses Python is abundant compared to those that do not, in most jobs using technology. Information technology tops the list.

Figure 13

A graph with different colored bars

Description automatically generated

Using Python as an example, Figure 14 illustrates the salaries in sectors for jobs that require knowledge of Python versus those that do not. With a few exceptions, most sectors offer a higher pay for jobs that require knowledge of Python compared to others that do not.

Figure 14

A graph of green and blue bars

Description automatically generated

**Data Presentation**

**Part I: Flask App**

From the database glassdoor\_jobs in SQLite, table glassdoor\_clean is extracted to create a flask app.

An API with various query options is created. Each option when chosen displays a chart pertaining to job data.

A screenshot of a computer

Description automatically generated

Using JavaScript, a program module is created with a ‘click’ button. The options listed above are included in the module to be accessed using the ‘click’ option. The click button is activated using an html file, which uses a CSS file to ‘style’ the webpage.

The following illustrates one instance of the page using the ‘click’ button:

A group of graphs showing different colored lines

Description automatically generated with medium confidence

**Part II: Map App**

Cleaned Glassdoor data was reformatted for information regarding jobs for each state as follows:

A table of information with text

Description automatically generated with medium confidence

This data was saved as a JSON file to be used in a ‘map’ application.

A map module is created using JavaScript. A base tile layer was added using Leaflet street map:

<https://www.openstreetmap.org/copyright>

Each region is given a unique color. The states within are demarcated with borders. The display is created using HTML file and CSS styling the page.

A map of the united states

Description automatically generated

A ‘mouse-over’ function increases the opacity to 90% when the mouse is over a state.

Clicking on a state, Oregon, in the following example, enlarges the state to fit the screen, and the state job data appears in a pop-up box.

A pink map with black text

Description automatically generated

**Part III: Interactive Dashboard**

For a complete overview of state-by-state job opportunities, an interactive dashboard is created.

Data from glassdoor.json is re-formatted include job information for each region pertaining to each state. All data for each state is isolated so as to be exclusively viewed, one state at a time.

Using JavaScript, functions are created in a program file, which are accessed in a HTML file to create the dashboard. The following is an instance of the dashboard for the state of New York:

A screenshot of a computer

Description automatically generated

In the dashboard, a dropdown option allows a user to choose a state code, such as NY for New York, and graphs and job-data pertaining to the state appears. The gauge on the upper right shows information regarding the average salary in the region, the chosen state belongs to.

**Conclusion**

Glassdoor data of job postings for 2017-2018 is analyzed to assess the following:

* Identifying the dominating factors affecting salaries,
* Determining the locations (state, region) that invite higher paying jobs,
* Number of jobs available in a category, such as job types (titles) and sectors.

From the visual exploration and analysis of data, it can be deduced that:

* Tech jobs related to ‘data processing’ are among the highest paying jobs,
* Such jobs are mostly concentrated in companies less than 50 years old,
* Company size (number of employees) is not a factor in high paying jobs,
* Jobs that require programming skills pay the highest, and
* West and Northeast regions have the greatest abundance of tech jobs related to data processing.

The Glassdoor dataset considered here could be significantly useful for candidates seeking careers in fields related to ‘data science’ and ‘data processing’.