Hiring Process Analytics

Project 5

- **Samruddhi Pawar**

Project Description:

The project is about a company's hiring process, which is a very fundamental and important process. The analysis is done over the data set provided. Basically, EDA- Exploratory data analysis is performed where we will try to summarize, visualize, and identify missing values of the data, which will help us answer the various questions related to the data being asked.

As the hiring process is a very crucial process for a company, here, the MNCs get to know about the major underlying trends in the hiring process. Trends such as the number of rejections, number of interviews, types of jobs, vacancies, etc. are important for a company to analyse before hiring freshers or any other individual. Thus, making an opportunity for a Data Analyst job here too!

Following are the tasks that need to be interpreted to respond to specific inquiries.
- ***Hiring:*** Process of taking people into an organization for different kinds of positions. We have to determine the number of males and females who are hired.
- ***Average Salary:*** Adding all the salaries for a select group of employees and then dividing the sum by the number of employees in the group. We have to determine the average salary offered in this company.
- ***Class Intervals:*** The class interval is the difference between the upper-class limit and the lower-class limit. We have to Draw the class intervals for salary in the company.
- ***Charts and Plots:*** This is one of the most important parts of analysis to visualize the data. We have to draw a Pie Chart/Bar Graph of the proportion of people working in different departments.
- ***Charts:*** Use different charts and graphs to perform the task representing the data. We have to represent different post tiers using a chart/graph.

Approach:

We will be performing our analysis on Microsoft Excel using various in-built formulas of Excel and we will go step by step to find out the answers to the questions asked related to the data set.

1. ***Handling Missing Data:*** Check if there are any missing values in the dataset. If there are, decide on the best strategy to handle them.
2. ***Clubbing Columns:*** If there are columns with multiple categories that can be combined, do so to simplify your analysis.
3. ***Outlier Detection:*** Check for outliers in the dataset that may skew your analysis.
4. ***Removing Outliers:*** Decide on the best strategy to handle outliers. This could be removing them, replacing them, or leaving them as is, depending on the situation.
5. ***Data Summary:*** After cleaning and preparing your data, summarize your findings. This could involve calculating averages, medians, or other statistical measures. It could also involve creating visualizations to better understand the data.

The goal of this project is to use the knowledge of Statistics and Excel to draw meaningful conclusions about the company's hiring process. These insights could potentially help the company improve its hiring process and make better hiring decisions in the future.

Tech-Stack Used:

1. ***MS Excel:***
   It is used as a preparation base for data. The software enables users to seamlessly import and organize data from various sources, facilitating a structured foundation for analysis. Data cleaning becomes an intuitive process with Excel's capabilities, allowing users to identify and rectify issues like missing values and duplicates.

2. ***Tableau:***
   It was primarily used to create interactive charts and dashboards. Using the built-in data connections, I was able to work more efficiently, even when collating data from multiple sources and file types. Additionally, files with the same column names can also be combined into one data source, saving time on copying and pasting. Tableau's drag-and-drop interface is intuitive and dynamic, allowing for more flexibility and experimentation.

Dataset Overview:

The dataset contains records of candidates who were interviewed previously with information about hiring status, hiring department, salary, etc.

The Dataset details are:
- Number of Data Points: 7,168
- Number of Features: 6
- Column Details:

1. ***application_id:*** ID of the applicant
2. ***Interview Taken on:*** Date and time of the interview
3. ***Status:*** Hired or rejected
4. ***event_name:*** Gender of the applicant
5. ***Department:*** Name of the department for which the interview was conducted
6. ***Post Name:*** Name of the post offered
7. ***Offered Salary:*** Salary offered for the job

Source of Data:
https://docs.google.com/spreadsheets/d/1GtK--gVfkHXMkrzLCp9fEvWmz0vBsknTe8IOZ_F8BO0/edit?usp=sharing

Data Cleaning:

A. Handling Missing Values

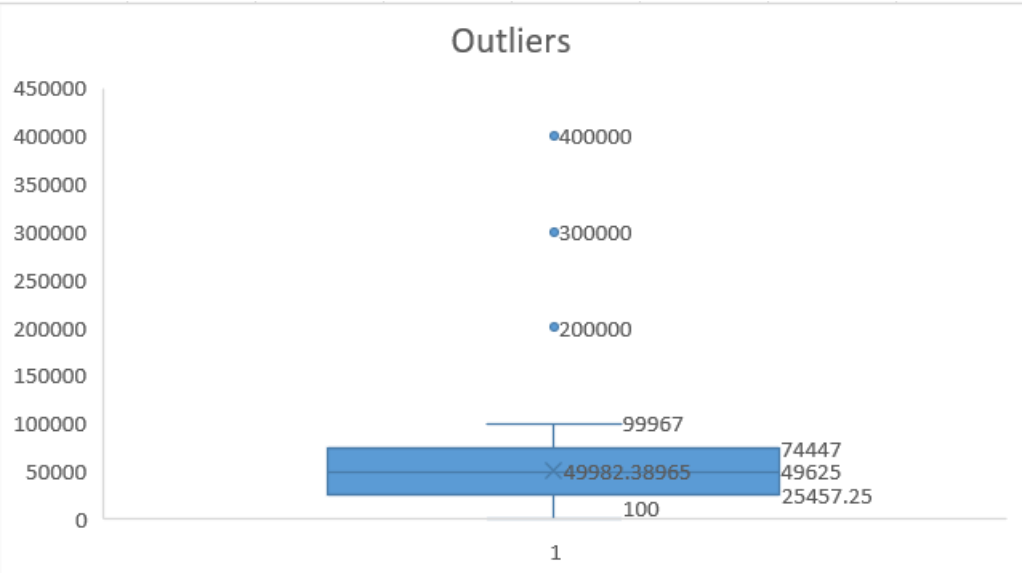
1. Column event_name has 15 rows with "-"as its value. These can be termed Null values. We replaced it with "Don't want to say" as they both imply the same thing in the context of this project i.e. gender of the candidate is not known.
2. Column Offered Salary has 1 row with Null Value. The corresponding value in the Department column is "Sales Department" and the Post Name is "i7". So, we replaced it with a median of Offered Salary for Sales Department and i7 Post Name. The median came out to 45400.
3. Column Post Name has 1 row with "- "as its value. It can be termed a Null value. The corresponding value in the Department column is "Sales Department" and the Offered Salary is "85914". So we replaced it with the majority count of Posts for candidates in the Sales Department whose Offered Salary is between 85,000 and 96,000, which is "c9".

B. Error Rectification

Column Post Name has a category "c-10" which seems to be a typo and the correct category should be "c10" which we rectified.

C. Handling Outliers

From the below Box Plot of Column Offered Salary, we can see that there are three rows whose Column values are outliers and the values are 200000, 300000, 400000. We replaced them with the median value of the Offered Salary for the corresponding Department and Post Name.



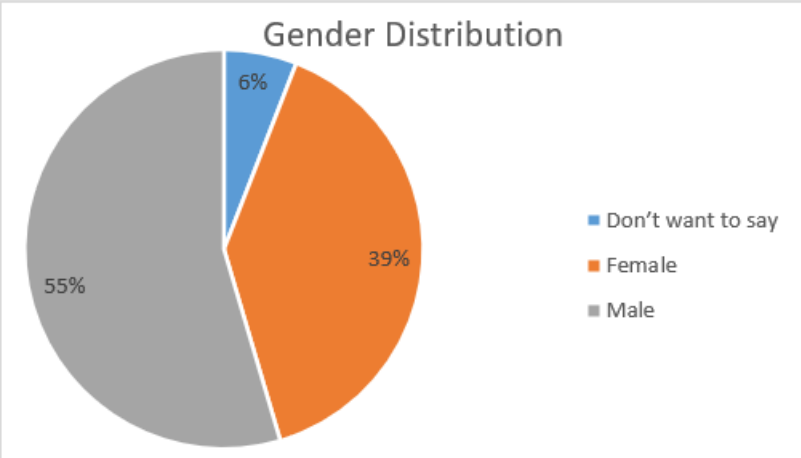
D. **Handling Duplicate Values**
Column application_id has 54 rows with duplicate values. They should either be removed or replaced with the correct value.

Insights:

1. **Hiring Analysis:** The hiring process involves bringing new individuals into the organization for various roles.
    - ***Task:*** Determine the gender distribution of hires. How many males and females have been hired by the company?

    - ***Results:***

Status	Hired
Gender	**Count**
Don't want to say	278
Female	1856
Male	2563
Grand Total	**4697**



- ***Insight:***

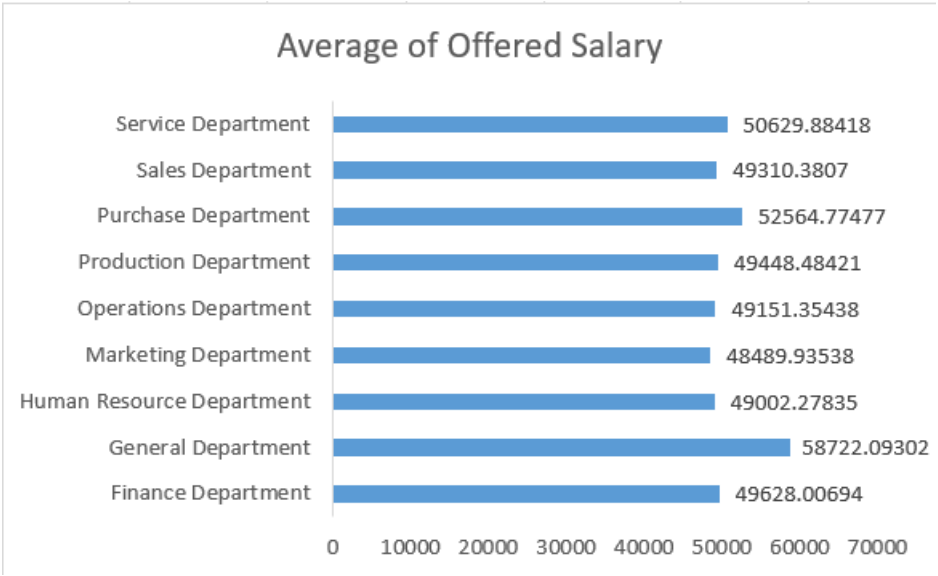
- More than half of the hired candidates are Male and only 39.51% of them are Female. The rest haven't disclosed their Gender. A high Gender Ratio (Ratio of Male to Female) may negatively impact the Organization's image in the public domain. The Organization should therefore focus on decreasing the Gender Ratio bringing it close to 1.
- The Data Quality Team should ensure that the Data is complete and relevant as incomplete or irrelevant data may hamper the analysis process.

2. **Salary Analysis:** The average salary is calculated by adding up the salaries of a group of employees and then dividing the total by the number of employees.
    - ***Task:*** What is the average salary offered by this company? Use Excel functions to calculate this.

    - ***Result:***

Department	Average
Finance Departmer	49628
General Departme	58722
Human Resource D	49002
Marketing Departn	48490
Operations Depart	49151
Production Depart	49448
Purchase Departm	52565
Sales Department	49310
Service Departmer	50630
Total Average	49983

Average	49983
Median	49625
Minimum	100
Maximum	400000



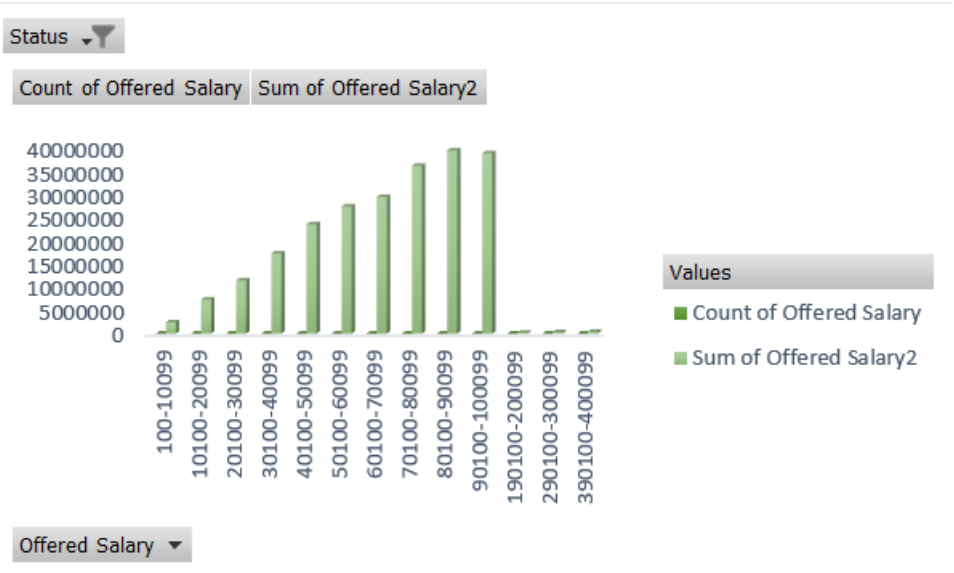
- ***Insights:***

- The Average Offered Salary is Rs. 49983 by the company
- If you calculate department-wise, the General department has the highest average salary of Rs. 58722.09

3. **Salary Distribution:** Class intervals represent ranges of values, in this case, salary ranges. The class interval is the difference between the upper and lower limits of a class.
    - ***Task:*** Create class intervals for the salaries in the company. This will help you understand the salary distribution.

    - ***Result:***

Status	Hired	
Row Labels	**Count of Offered Salary**	**Sum of Offered Salary2**
100-10099	444	2436140
10100-20099	487	7339573
20100-30099	457	11445148
30100-40099	488	17223090
40100-50099	523	23494867
50100-60099	496	27333948
60100-70099	450	29350614
70100-80099	479	36054597
80100-90099	462	39362436
90100-100099	408	38748940
190100-200099	1	200000
290100-300099	1	300000
390100-400099	1	400000
Grand Total	**4697**	**233689353**



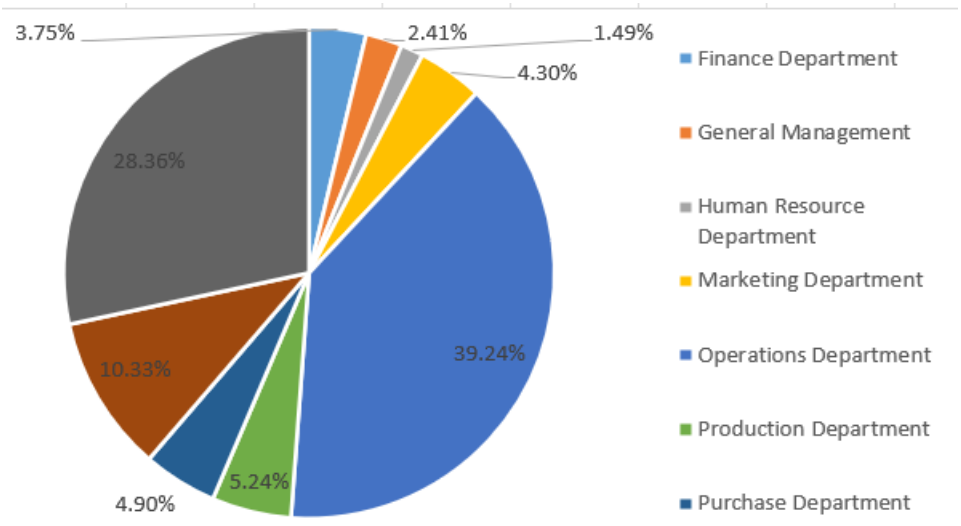
- ***Insights:***

- We can observe that the highest count of status for the offered salary is in the interval of Rs.40100–50099. That is most of the job requirements were for middle experience posts and least for senior posts and for freshers (assuming salary is directly proportional to work experience).
- We can also observe that the distribution of Salaries of hired candidates also follows the same pattern i.e. most candidates are hired for middle experience posts and least for senior posts and for freshers (assuming salary is directly proportional to work experience).

4. **Departmental Analysis:** Visualizing data through charts and plots is a crucial part of data analysis.
   - ***Task:*** Use a pie chart, bar graph, or any other suitable visualization to show the proportion of people working in different departments.

   - ***Result:***

Status	Hired
Department	**Count**
Finance Department	3.75%
General Management	2.41%
Human Resource Department	1.49%
Marketing Department	4.30%
Operations Department	39.24%
Production Department	5.24%
Purchase Department	4.90%
Sales Department	10.33%
Service Department	28.36%
Grand Total	**100.00%**



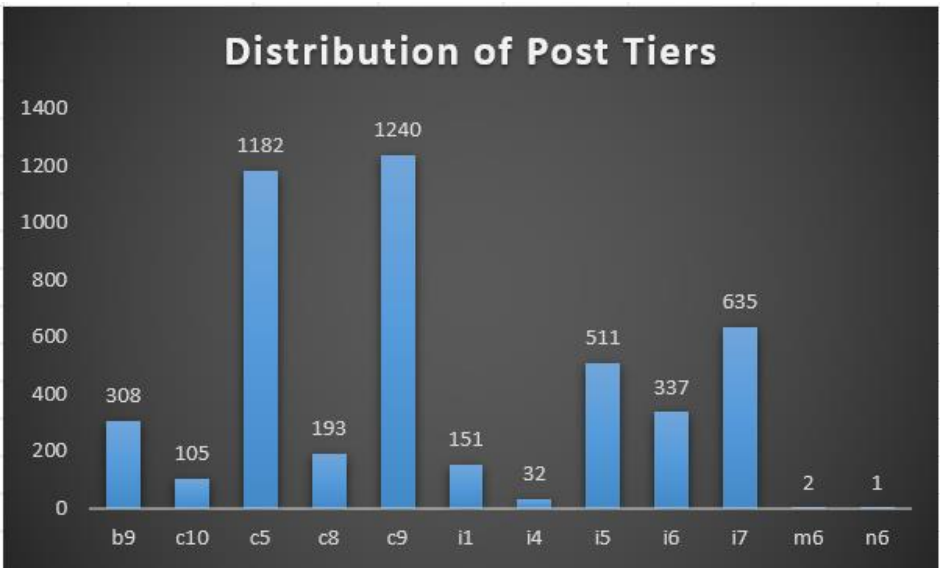
- ***Insights:***

- From the above pie chart, we can observe that most candidates are hired in the Operations Department followed by the Services Department and Sales Department and the least candidates are hired in the Human Resources Department.
- These numbers may indicate the size of teams and the importance of the departments in the organization.

5. **Position Tier Analysis:** Different positions within a company often have different tiers or levels.
    - ***Task:*** Use a chart or graph to represent the different position tiers within the company. This will help you understand the distribution of positions across different tiers.

    - ***Result:***

Status	Hired
Post Name	**Count of Post Name**
b9	308
c10	105
c5	1182
c8	193
c9	1240
i1	151
i4	32
i5	511
i6	337
i7	635
m6	2
n6	1
Grand Total	**4697**



- ***Insights:***

- Here, we can observe that the organization has hired most candidates for post-tier c9 followed by c5 and then i7 at a distant third.

Summary:

- Males are 55% of all the employees, 39% are females and 6% don't want to say.
- The General Department has the highest average salary of Rs. 58722
- We can observe that the highest count of status for the offered salary is in the interval of Rs.40100–50099.
- Most candidates are hired in the Operations Department and the least candidates are hired in the Human Resources Department.
- Most of the employees are in the post-tier of c9 with a total of 1240 employees.

Conclusion:

This project has helped in get hands-on experience with real-life data sets and how we clean, manipulate, visualize, and draw insights from the data The questions asked in the data set have been answered to the best knowledge, and tried to plot the required graphs and chat as per requirement and my understanding.

The Exploratory Data Analysis part has helped me to understand that before moving towards making further analytical treatment of data and making it fit for making models. We have to do the EDA to make the data error-free and bias-free so that the inference drawn from the data is a good a fit to the further statistical and analytical treatment.