HIRING PROCESS ANALYTICS

Project 4 – by Nazireen Sania

METHODOLOGY

The total number of rows in the datasheet is 7168 with 7 columns namely application_id, Interview Taken on, Status, event_name, Department, Post Name, Offered Salary.

1. Handling Missing Data

* In the event_name column, there are blank values, indicated with a hyphen (-) and many cells with the name “Don’t want to say”. There are 393 entries with “Don’t want to say” and 15 entries with a hyphen. By filtering these values using the “Filter” function in Excel, we filter the “-“ and “Don’t want to say” values and delete the rows. There is one missing value under Post Name. Aditionally, we also find a missing value from the ‘find&select’ function. We delete these rows.

Other ways of handling missing data:

- Replacing the missing value with the average value
  - We cannot do that here as gender is a categorical variable.
- Creating a separate category and placing all the missing values in that
  - We choose not to that here because further analysis requires analysing strictly based on the categories and any kind of additional categories will skew the analysis and goes beyond the scope of the analysis.
- Imputing Based on Other Data like inferring from names
  - Since we do not have names, this is not possible. Manually assigning genders to rows from their names for a large dataset is tedious.
- Assigning the most frequent gender in the dataset
  - This will skew the analysis here so we don’t prefer this.

Hence, deleting the rows is preferred.

2. **Clubbing Columns:**

No columns are clubbed.

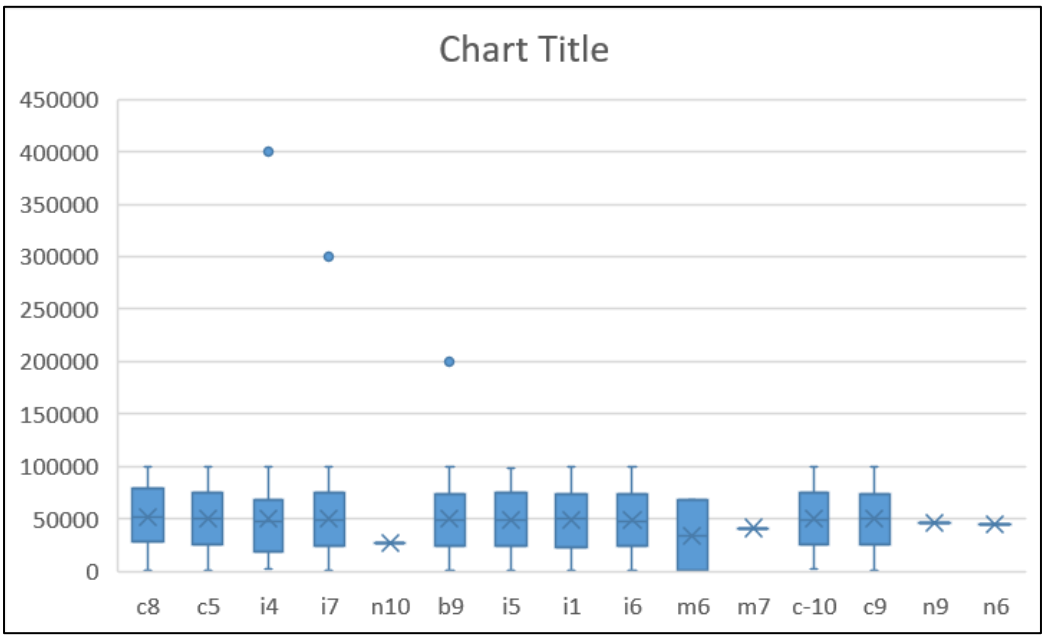
3. Outlier Detection:

There are multiple ways to calculate outliers in Excel:

- Using the Interquartile Range (IQR) Method
- Using Z-Scores (Standard Deviation Method)
- Using Conditional Formatting for Visualization
- Using Box Plot (Graphical Method)

We use the box plot as it is convenient and since we only need to find the outliers for only two columns (Post Name and Offered Salary – Numerical Values), box plot would be best suited.

We select the columns and click on box plot under Insert Charts function.



The X -axis represents the Post Name and the Y-axis represents the Offered Salary. The box plot visually represents salary distributions across different post names, with the **outliers marked as individual dots** above the whiskers. Here's how you can interpret it:

- **Outliers Detected**
  The dots above the boxes indicate **outliers**, meaning these salaries are significantly higher than the typical range.

- To find whether or not a salary value is an outlier, we use excel formulae. For this we find:
  - **Q1 (First Quartile):** =QUARTILE.INC(range, 1)
  - **Q3 (Third Quartile):** =QUARTILE.INC(range, 3)
  - **IQR (Interquartile Range):** =Q3 - Q1
  - **Upper Bound:** =Q3 + (1.5 * IQR)
  - **Lower Bound**: =Q1 - (1.5 * IQR)
  - To identify each entry in the salary column, we give the condition to identify outliers
  - =IF(OR(A2<Lower_Bound, A2>Upper_Bound), "Outlier", "Normal")
    where A2 is the cell.

Oultiers are values that are too small or too large hence if a value is less than lower bound or greater than upper bond, it is an outlier.

Data Analytics Tasks

A. Hiring Analysis

Objective: Determine the gender distribution of hires. How many males and females have been hired by the company?

We have already deleted rows under the event_name to remove values other than male or female.

We use a pivot table to determine the number of males and females in the company. By using rows as event_name and values as the count of event_name, we get the number of males and females in the company.

	Column Labels			
	Female	Male	(blank)	Grand Total
Count of event_name	2675	4083		6758

Besides, we can also calculate the number of males and females in each department in the company by putting Rows as event_name, Columns as Department and Values as the count of event_name.

Count of event_name	Column Labels			
Row Labels	Female	Male	(blank)	Grand Total
Finance Department	258	14		272
General Management	152	11		163
Human Resource Department	36	57		93
Marketing Department	102	210		312
Operations Department	960	1639		2599
Production Department	141	220		361
Purchase Department	108	200		308
Sales Department	248	465		713
Service Department	670	1267		1937
(blank)				
Grand Total	**2675**	**4083**		**6758**

B. Salary Analysis:

Objective: What is the average salary offered by this company?

Average salary = (sum of salaries of all employees) / (number of employees)

We give the following formula for average salary where column G consists of salaries offered to the employees

```
=(SUM(G2:G6759)/COUNT(G2:G6759))
```

Average Salary	49990.68

C. Salary Distribution

Objective: Create class intervals for the salaries in the company. This will help you understand the salary distribution.

We can calculate class interval for the salaries using 2 ways:

- Creating class intervals using equal ranges (if we wish to have equal class width)
- Use interquartile Range (IQR) for Class Intervals if you want classes based on salary distribution

(1) CREATING CLASS INTERVALS USING EQUAL RANGES

Class	Salary Range
1	800 - 29,314
2	29,315 - 57,829
3	57,830 - 86,344
4	86,345 - 114,859
5	114,860 - 143,374
6	143,375 - 171,889
7	171,890 - 200,404
8	200,405 - 228,919
9	228,920 - 257,434

- Minimum Salary: 800
- Maximum Salary: 400,000
- Class Width: 28,514.3
- Number of Classes: 14

Class	Salary Range
10	257,435 - 285,949
11	285,950 - 314,464
12	314,465 - 342,979
13	342,980 - 371,494
14	371,495 - 400,000

(2) CREATING CLASS INTERVALS USING INTERQUARTILE RANGE

Q1	25415.75
Q3	74232.25
IQR	48816.5
Upper Bound	147457
Lower Bound	-47809
Q2	49740

Q2 is the second quartile, the median

Defining Class Intervals Based on IQR

- **Below Q1** (Low Salaries)
- **Q1 to Median (Q2)** (Lower-Mid Salaries)
- **Q2 to Q3** (Upper-Mid Salaries)
- **Above Q3** (High Salaries)

Hence, the class intervals are :

- Category	Salary Range
Low Salaries	800 – 25,415
Lower-Mid Salaries	25,416 – 49,970
Upper-Mid Salaries	49,971 – 74,232
High Salaries	74,233 –400,000

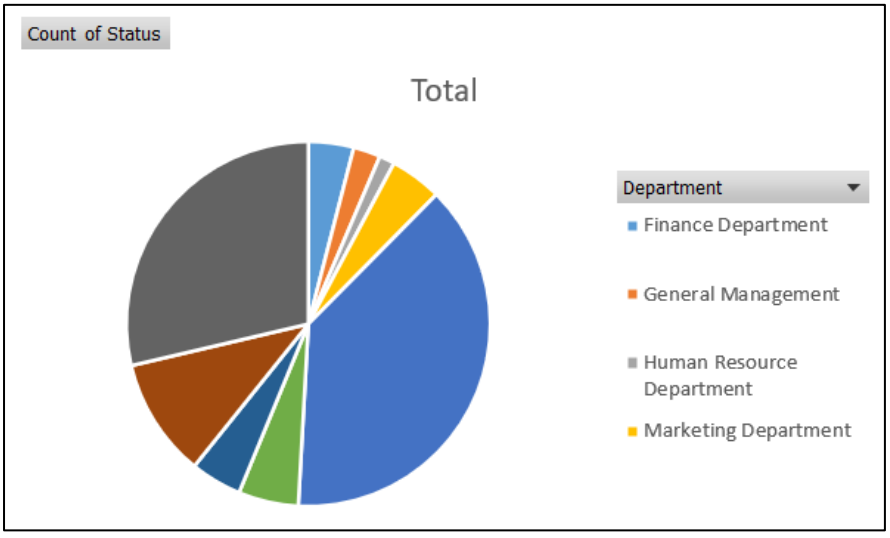
D. Departmental Analysis

Objective: Use a pie chart, bar graph, or any other suitable visualisation to show the proportion of people working in different departments.

Row Labels	Count of application_id
Finance Department	272
General Management	163
Human Resource Department	93
Marketing Department	312
Operations Department	2599
Production Department	361
Purchase Department	308
Sales Department	713
Service Department	1937
(blank)	
Grand Total	**6758**

We create a column using pivot table. Here, the count of application_id represents the number of employees in each department.

After selecting the table values, we create a pie chart using the chart options in Excel.



E. Position Tier Analysis

Objective: Use a chart or graph to represent the different position tiers within the company.

Based on the class intervals that we obtained earlier based on Interquartile Ranges, we can classify employees into different tiers based on their salary since we cannot interpret their position from the labels given, for example, b9, c7, i5, etc. for this, we create a new conditional column that will define the employee's tier.

The IF conditions are:

=IF(G2<=25415,"Low Tier",IF(G2<=49970,"Lower-Mid Tier",IF(G2<=74232,"Upper-Mid Tier","High Tier")))

Where G2 refers to the employee's salary

After labelling each employee into various tiers based on his/her salary, we create a pivot table as follows:

Row Labels	Count of application_id
High Tier	1690
Low Tier	1690
Lower-Mid Tier	1703
Upper-Mid Tier	1675
(blank)	
Grand Total	**6758**

The bar chart for this would be: