

A thick dark blue vertical bar runs along the left edge of the page. A blue arrow-shaped banner points to the right from this bar, containing the word 'STATISTICS'. In the bottom-left corner, several thin, curved lines in dark blue and light grey sweep upwards and to the right.

STATISTICS

PROJECT 4

Hiring Process Analytics

RAKSHA NAYAK

Project Description: This Project, “Hiring Process Analytics” is about analyzing the given dataset containing records of previous hires and answering certain questions and drawing meaningful insights that can help the company improve its hiring process.

The hiring process is a crucial function of any organization, and understanding trends such as the number of rejections, interviews, job types and vacancies can provide valuable insights for the hiring department.

Approach: The dataset provided is,

<https://docs.google.com/spreadsheets/d/1gAq5sK8L2e7rCP0O0KaNo7gqx6tfnVQk/edit#gid=1029390730>

The dataset after analysis with answers, insights and visualization is,

<https://docs.google.com/spreadsheets/d/1upDvYqcOfBuPII78SxSKgkpLtcYwV6Bm/edit?usp=sharing&ouid=108154584635151678812&rtpof=true&sd=true>

The above dataset contains records of previous hires.

The dataset contains,

Total data points: 7168

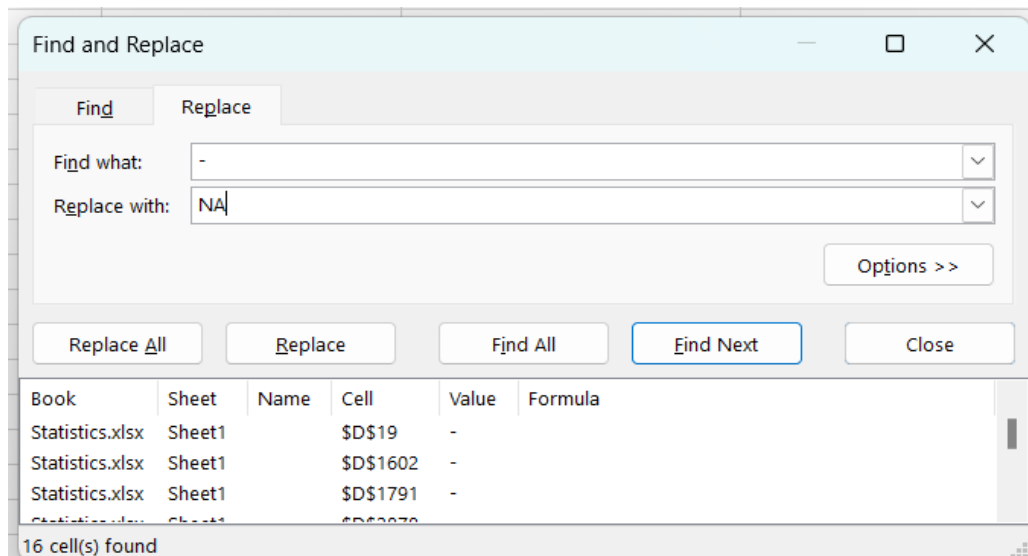
Attributes/Columns: 7

Column names:

1. Application id
2. Interview Taken on
3. Status
4. Event Name
5. Department
6. Post Name
7. Offered Salary

After downloading the data set, we need to check for

1. **Handling Missing Data and errors:** We need to check if there are any missing values in the dataset. If there are, we need to decide on the best strategy to handle them.
 - a) There are 15 values with “-” in event_name. We are replacing it with NA by using “find and replace” option of excel.



- b) Also 1 value in post_name has "-". To find the missing post, we are considering the salary and comparing with rest to check if that salary has appeared in any other post using,
`=INDEX(F:F,MATCH(G7,G:G))`

<code>=INDEX(F:F,MATCH(G7,G:G))</code>					
	C	D	E	F	G
on	Status	event_name	Department	Post Name	Offered Salary
11:40	Hired	Male	Service Department	c8	56553
08:08	Hired	Female	Service Department	c5	22075
08:08	Rejected	Male	Service Department	c5	70069
16:28	Rejected	Female	Operations Department	i4	3207
16:32	Hired	Male	Operations Department	i4	29668
07:44	Hired	Male	Sales Department	c5	85914

- c) There was an empty cell in offered_salary column, so we took the mean of offered_salary based on the respective department and post name and populated that cell with 47731.

Find and Replace

Find Replace

Find what: No Format Set Format...

Within: Sheet ☐ Match case
☒ Match entire cell contents

Search: By Rows

Look in: Values Options <<

Find All Find Next Close

Book	Sheet	Name	Cell	Value	Formula
Statistics.xlsx	Sheet1		\$G\$80		

1 cell(s) found

- d) The column post_name has c-10 as a post name and on observing other post name, seems like there is a typo and therefore we are changing it to c10.

Find and Replace

Find Replace

Find what: c-10 No Format Set Format...

Replace with: c10 No Format Set Format...

Within: Sheet ☐ Match case
☒ Match entire cell contents

Search: By Columns

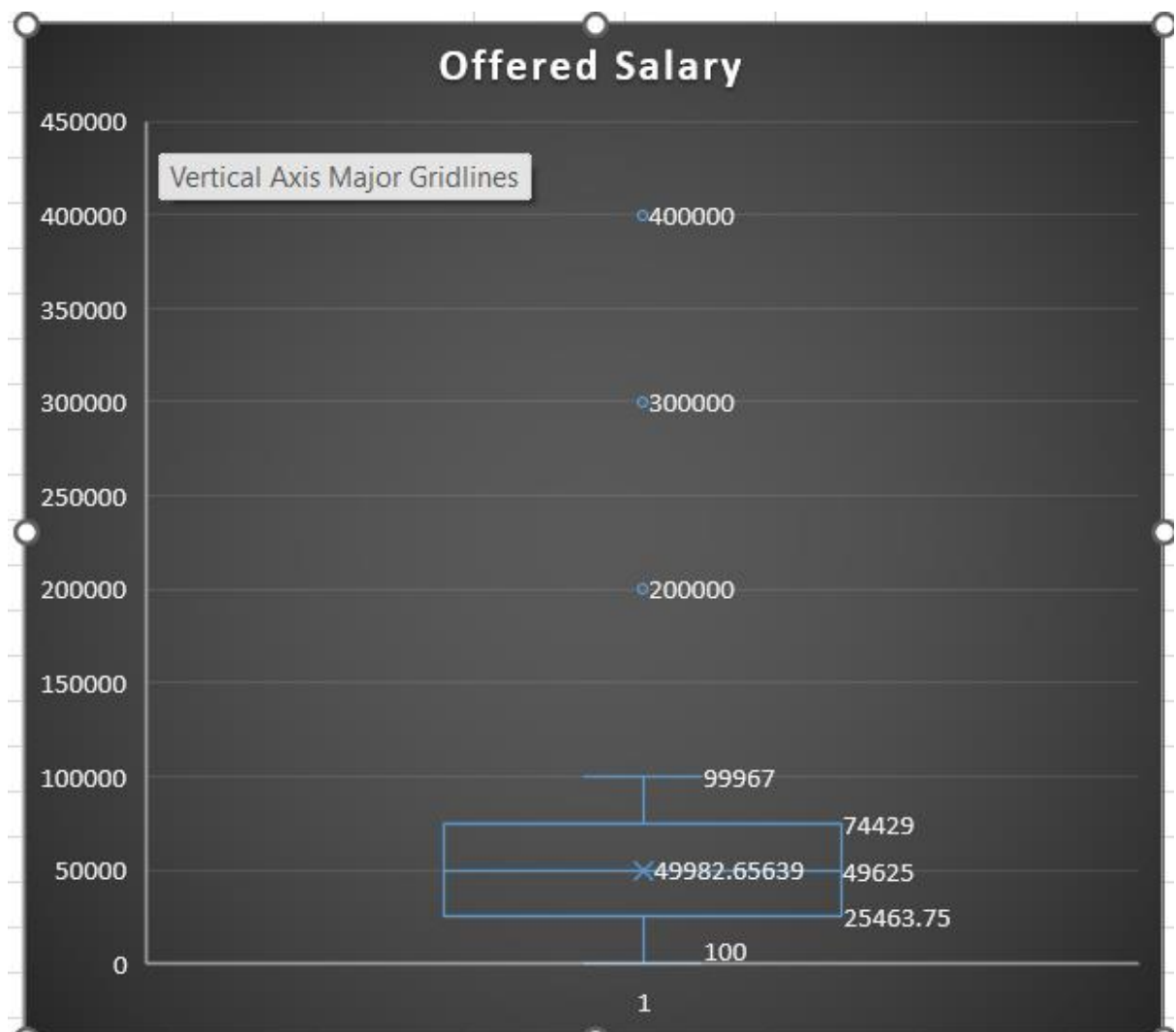
Look in: Formulas Options <<

Replace All Replace Find All Find Next Close

Book	Sheet	Name	Cell	Value	Formula
Statistics.xlsx	Sheet1		\$F\$142	c-10	
Statistics.xlsx	Sheet1		\$F\$143	c-10	
Statistics.xlsx	Sheet1		\$F\$144	c-10	
Statistics.xlsx	Sheet1		\$F\$145	c-10	

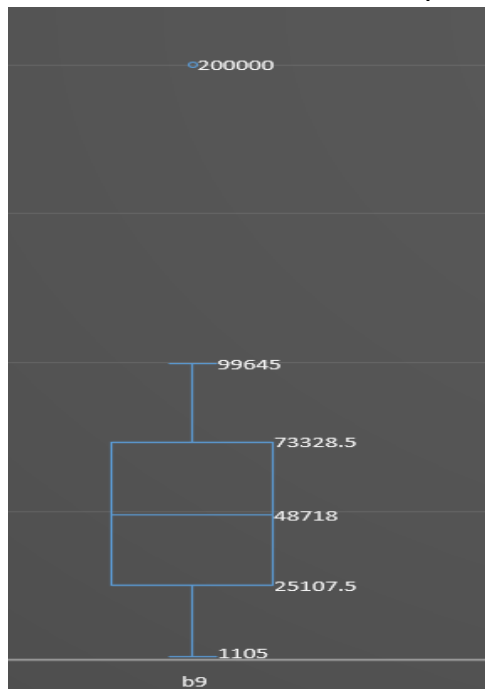
232 cell(s) found

2. **Clubbing Columns:** If there are columns with multiple categories that can be combined, do so to simplify your analysis – There are no columns that needs to be combined.
3. **Outlier Detection:** From the below box plot, we get to know there are 3 outliers of 200000, 300000 and 400000 for offered_salary and that may skew our analysis.

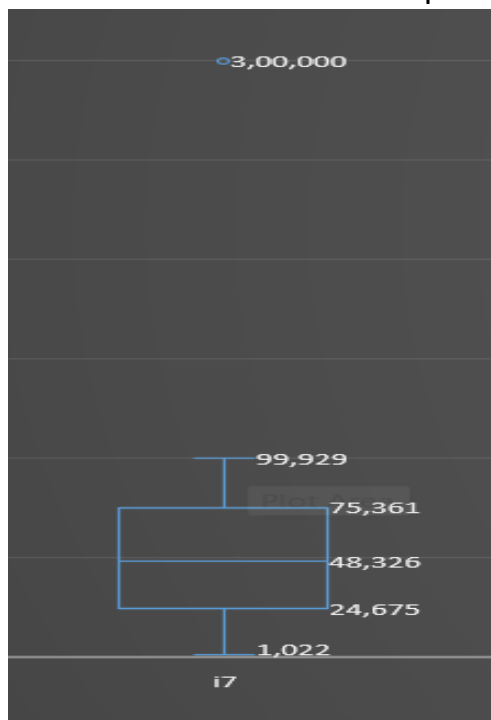


4. **Removing Outliers:** We are going to replace outlier values with median of the respective department and post name. As we know that when having outliers, it is best to *use median* as a measure of central tendency.

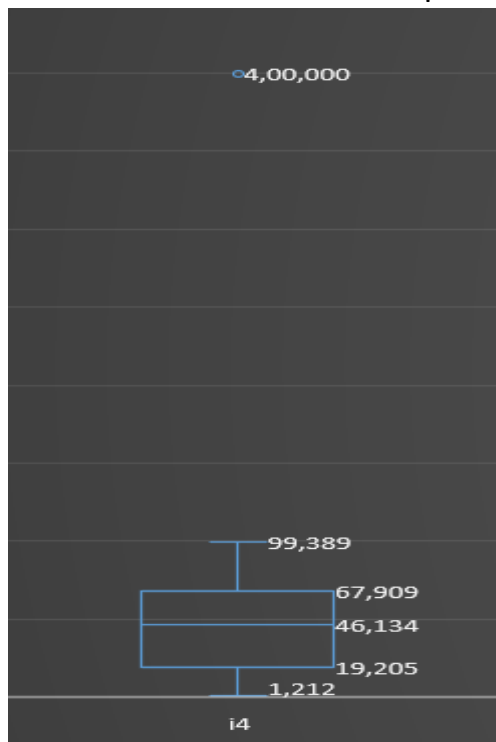
The below box plot helps us understand, b9 has the outlier of 200000 and its median is used to replace 200000 with 48718.



The below box plot helps us understand, i7 has the outlier of 300000 and its median is used to replace 300000 with 48326.



The below box plot helps us understand, i4 has the outlier of 400000 and its median is used to replace 400000 with 46134.



5. **Data Summary:** We have cleaned the data by managing missing values, errors and replacing the outliers with median values. Summary of the cleaned data is as shown below,

Summary	
Column1	
Mean	49877.07296
Standard Error	334.8032356
Median	49586
Mode	72843
Standard Deviation	28345.79518
Sample Variance	803484104.7
Kurtosis	-1.178488934
Skewness	0.013313708
Range	99867
Minimum	100
Maximum	99967
Sum	357518859
Count	7168

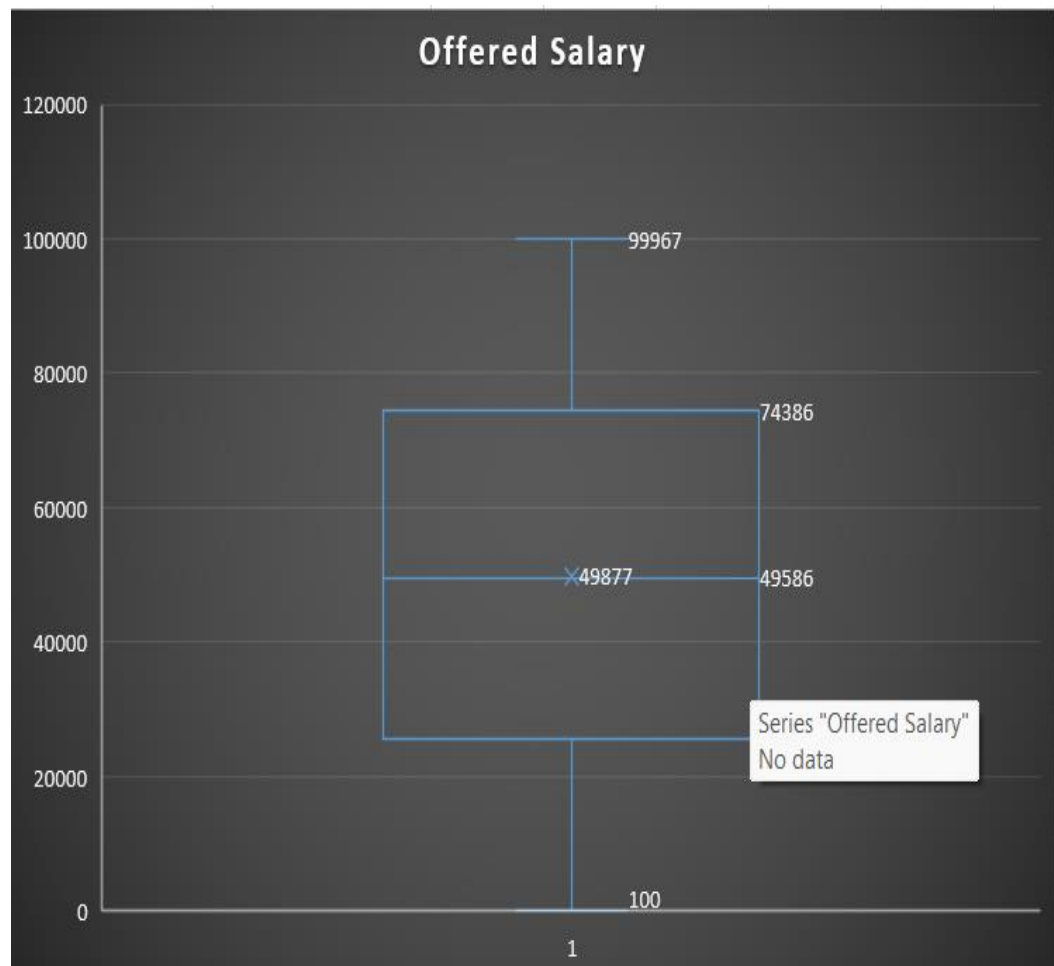
Above summary can be visualized through box plot,

Max Salary = 99967

Min Salary = 100

Mean = 49877

Median = 49586



Tech stack used: Microsoft Excel Version 2407 – Excel is a spreadsheet editor developed by Microsoft. It features calculation or computation capabilities, graphing tools, pivot tables etc.

Data Analytics Tasks:

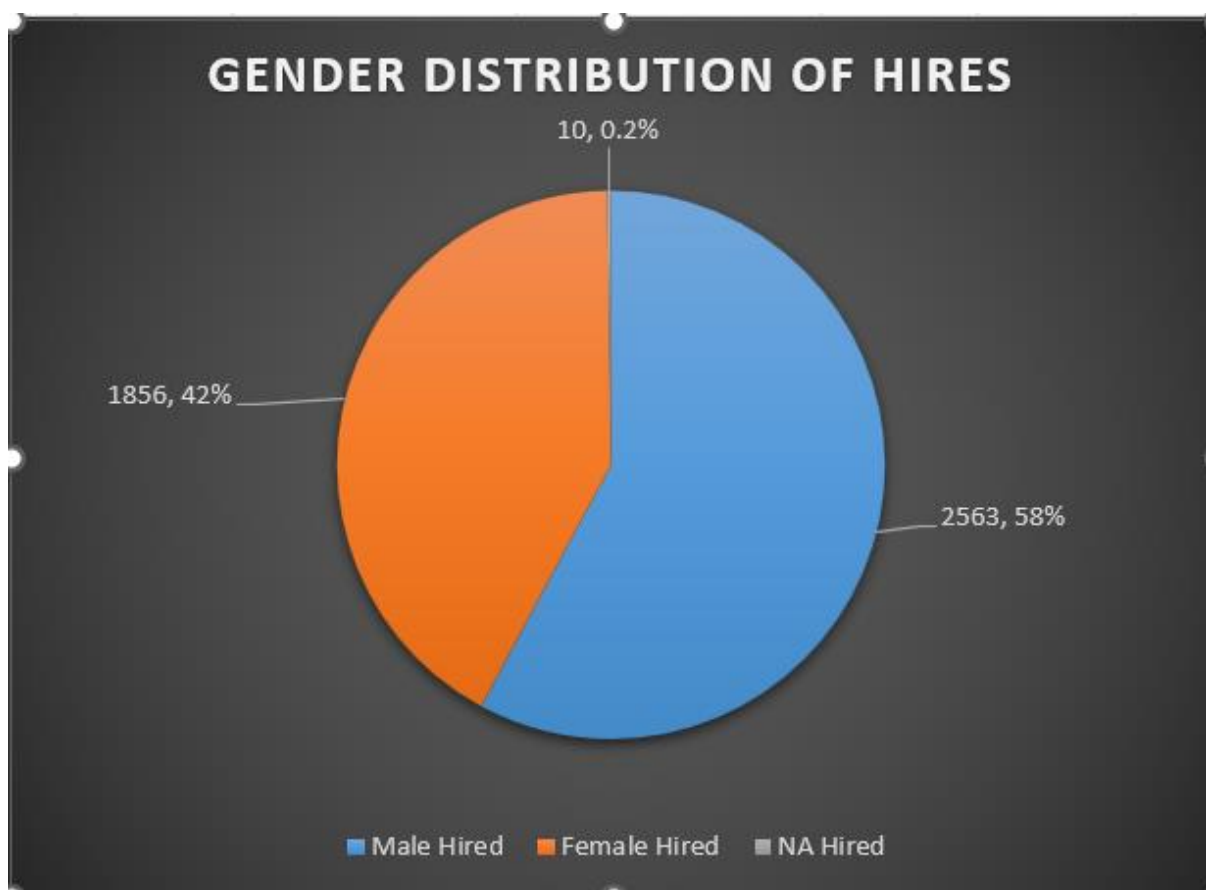
A. Hiring Analysis: The hiring process involves bringing new individuals into the organization for various roles.

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

Output:

event_name	Status	Count
Male	Hired	2563
Female	Hired	1856
NA	Hired	10
TOTAL		4429

Visual Representation:



Insight: 58% (More than half) of the hired candidates are **Male**, 42% **Female** and 0.2% NA(those who haven't disclosed their gender). Therefore, company should focus on

decreasing the gender ratio by bringing it close to 1 as it may negatively impact the image of the company in the public domain.

Also, the Data Quality Team should work on providing complete and relevant data **without any missing or error values** to ensure the smooth analysis process.

B. 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.

Output:

=AVERAGEIF(C:C,C2,G:G)							
	E	F	G	H	I	J	K
	Department	Post Name	Offered Salary			Average Salary of Hired Candidates	49591.7673
=AVERAGE(G:G)							
	Department	Post Name	Offered Salary			Average Salary of Hired Candidates	49591.7673
	Service Department	c8	56553			Average Offered Salary	49877

Insight: We have 2 outputs above,

1st Output calculates average salary of hired candidates, which is 49591.7673.

2nd Output calculates average offered salary irrespective of whether the offer was accepted or rejected, which is 49877.

Observation: There is no much difference between average salary of hired candidates and average offered salary. This shows that the hiring team is following the pre-determined salary ranges of the company.

C. 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 to understand the salary distribution.

Output:

MAX SALARY	99967	
MIN SALARY	100	
Count	7168	
CLASS INTERVAL	FREQUENCY OF SALARY OFFERED	FREQUENCY OF HIRED CANDIDATE SALARY
0	0	0
20000	1410	928
40000	1421	943
60000	1535	1024
80000	1432	929
100000	1370	873
Count	7168	4697
To Create Frequency Distribution Table		
CLASS INTERVAL	FREQUENCY OF SALARY OFFERED	FREQUENCY OF HIRED CANDIDATE SALARY
0-20000	1410	928
20000-40000	1421	943
40000-60000	1535	1024
60000-80000	1432	929
80000-100000	1370	873

Formula's Used:

{=FREQUENCY(G:G,\$K\$5:\$K\$10)}

G	H	I	J	K	L	M
MAX SALARY				99967		
MIN SALARY				100		
Count				7168		
CLASS INTERVAL				FREQUENCY OF SALARY OFFERED	FREQUENCY OF HIRED CANDIDATE SALARY	
0				0	0	0

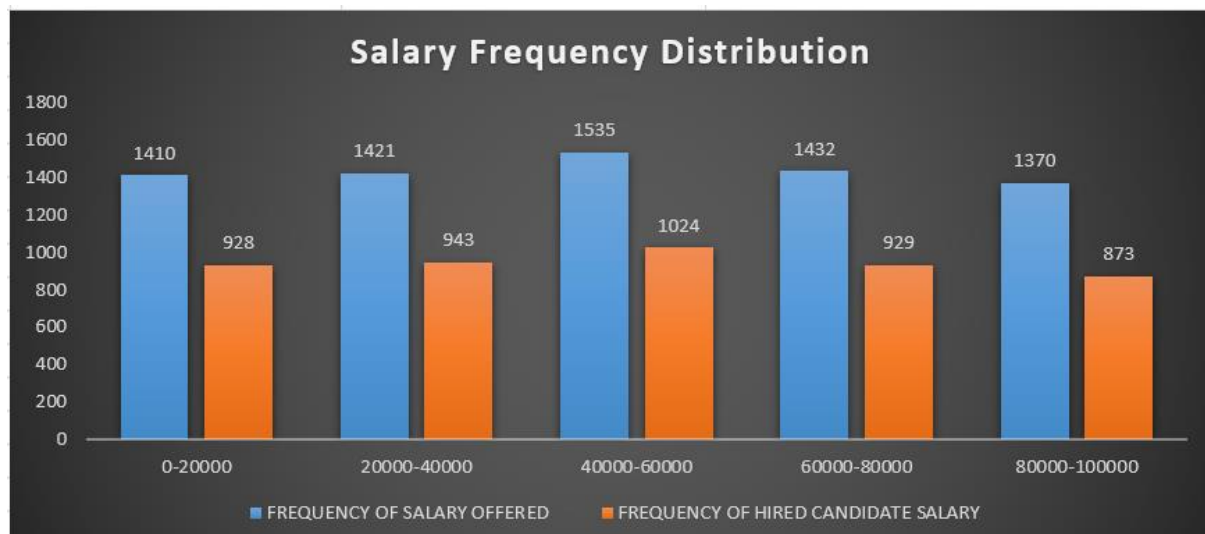
{=FREQUENCY(IF(C:C=\$C\$2,G:G),\$K\$5:\$K\$10)}

G	H	I	J	K	L	M
MAX SALARY				99967		
MIN SALARY				100		
Count				7168		
CLASS INTERVAL				FREQUENCY OF SALARY OFFERED	FREQUENCY OF HIRED CANDIDATE SALARY	
0				0	0	0

To create the intervals or bins, lets first find out maximum, minimum and total count of salaries.

To make it simple, each bin has an interval of 20,000 each. I have created 2 frequency distribution table. One for offered salary and the other for hired candidates' salary. This will help us in understanding and studying which range has maximum salary offers or accepted offers.

Visual Representation:



Insight: Maximum salary offered is in the range of 40000-60000 and minimum offered salary is in the range of 80000-100000.

Observation: This shows the maximum hiring was for the mid-level roles and minimum hiring for the Senior roles.

Maximum salary accepted by the hired candidates falls in the range of 40000-60000 and minimum salary accepted by the hired candidate falls in the range of 80000-100000.

Observation: Most of the positions filled in an organization is by mid-level roles and least by senior level roles.

D. 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.

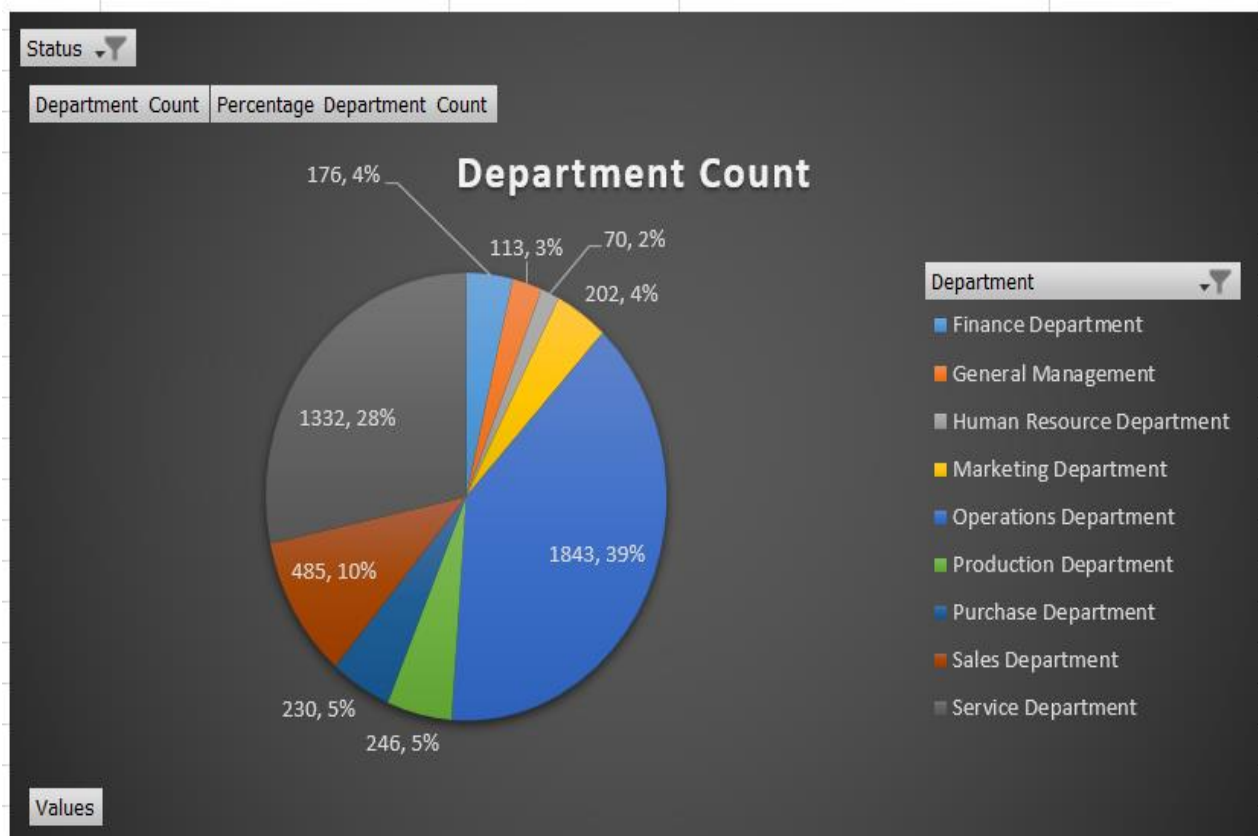
Approach: Have created pivot table to summarize various departments and number of people working for each department. To visualize and understand better, let's take percentage of total number of people working for each department.

Output:

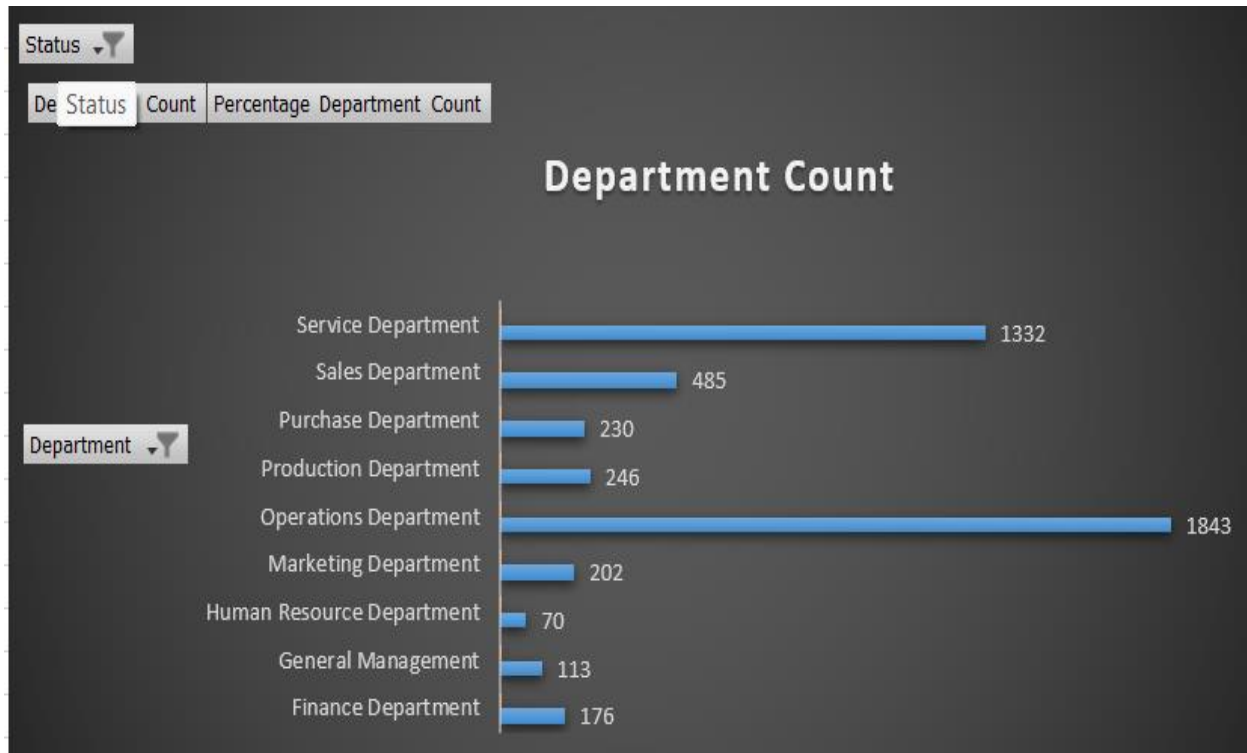
Status	Hired	
Row Labels	Department Count	Percentage Department Count
Finance Department	176	3.7%
General Management	113	2.4%
Human Resource Department	70	1.5%
Marketing Department	202	4.3%
Operations Department	1843	39.2%
Production Department	246	5.2%
Purchase Department	230	4.9%
Sales Department	485	10.3%
Service Department	1332	28.4%
Grand Total	4697	100.0%

Visual Representation:

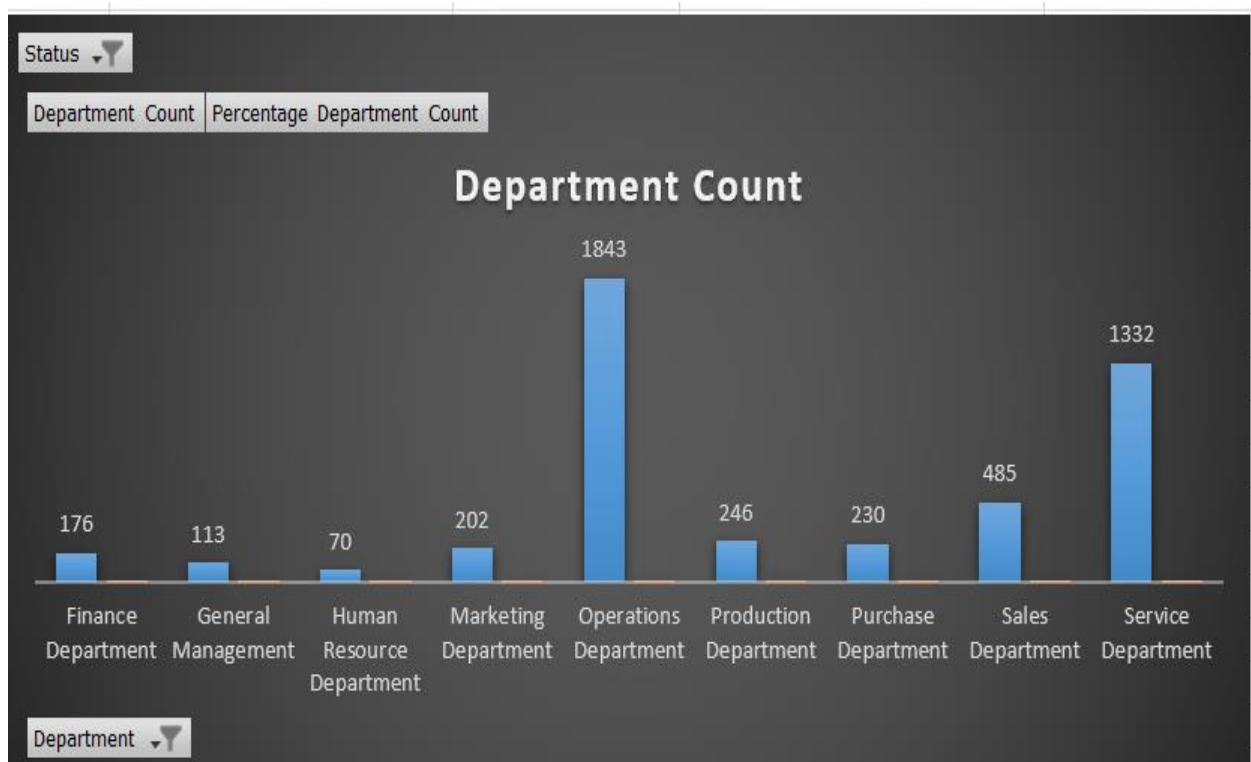
Pie Chart



Bar Graph



Column Graph



Insight: The above graph helps to understand organization structure with size of each department and their importances in the organization.

Observation: Operations Department has highest number of employees working with 39% and Human Resource Department has least number of employees with 2%.

E. Position Tier Analysis: Different positions within a company often have different tiers or levels.

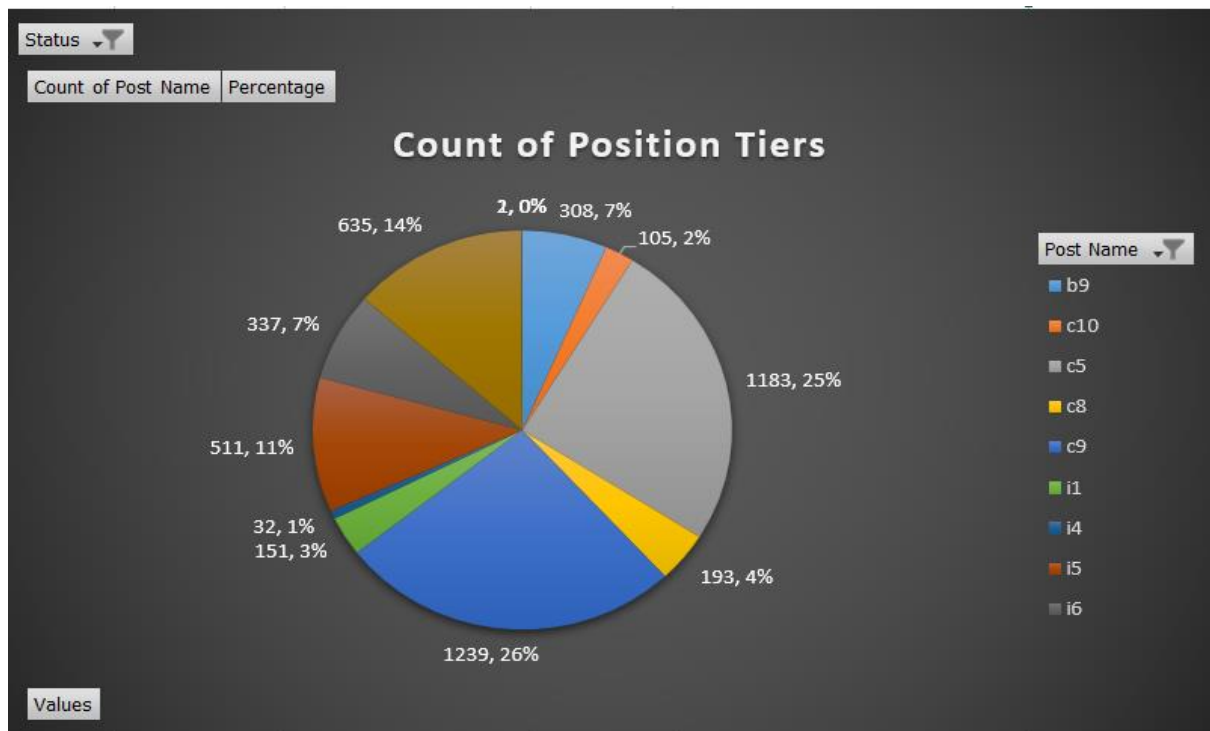
Task: Use a chart or graph to represent the different position tiers within the company. This will help you understand the distribution of positions across different tiers.

Output:

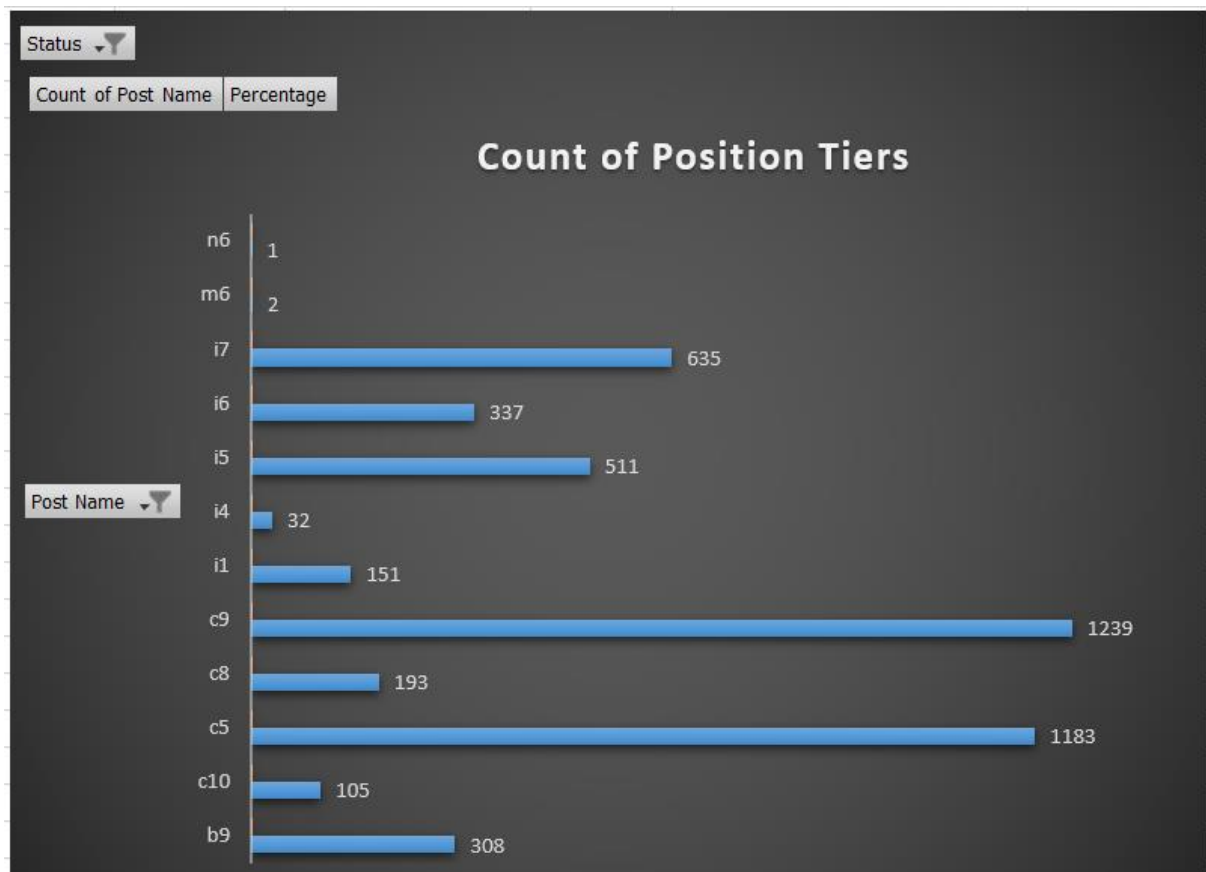
Status	Hired	
Row Labels	Count of Post Name	Percentage
b9	308	6.56%
c10	105	2.24%
c5	1183	25.19%
c8	193	4.11%
c9	1239	26.38%
i1	151	3.21%
i4	32	0.68%
i5	511	10.88%
i6	337	7.17%
i7	635	13.52%
m6	2	0.04%
n6	1	0.02%
Grand Total	4697	100.00%

Visual Representation:

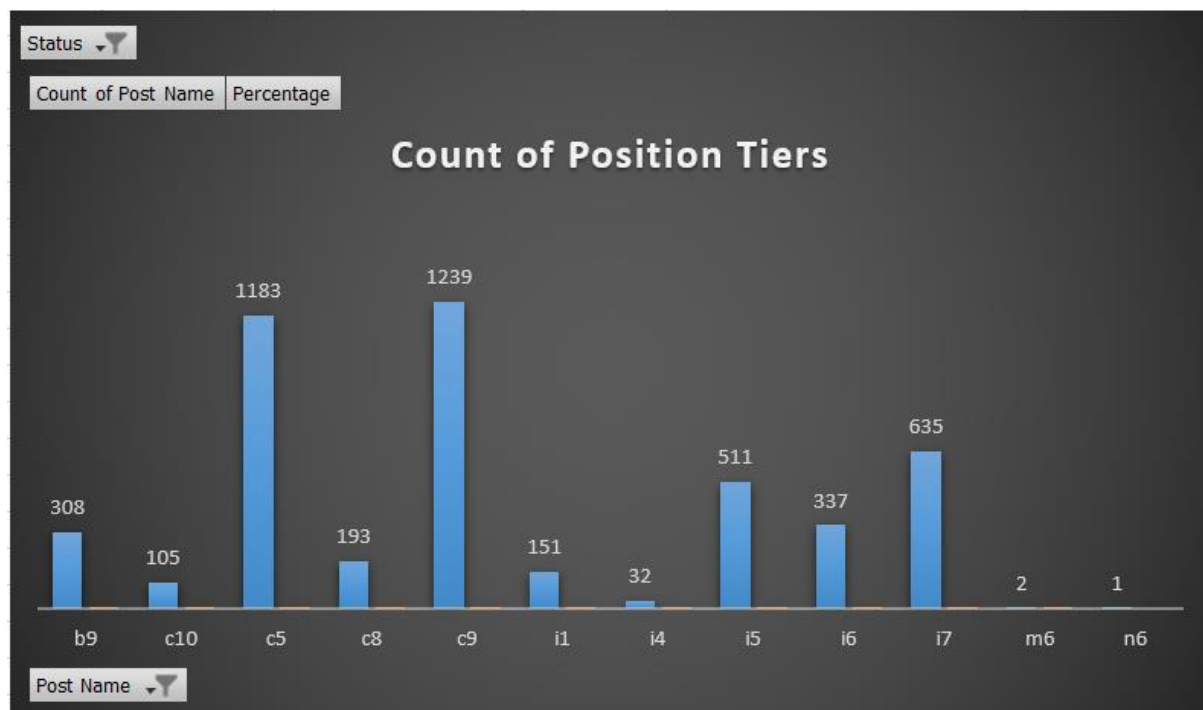
Pie Chart



Bar Graph



Column Graph



Insight: The above graph helps to understand various positions/posts in an organization along with number of each positions/post available.



Observation: Highest number of employees are working for post, c9 with 26%, followed by c5 and least number of employees are working for post, n6 with 0.02% followed by m6.

Result: This project helped me to understand

- How to clean the data by removing duplicate values, handling missing values and errors.
- Helped me understand how to find the outlier and use best strategy to remove or replace them.
- Has made me think and apply logic and use the best out of them with proper explanation so that the audience can relate with it.
- How to retrieve/extract the answers and what technique or formulas to use to achieve it.
- Has made me think like an analyst, considering what makes the entire thing perfect with relevant and valuable insights.
- Along with approach and output, visual representation plays a very important role for faster understanding of analysis and this project has made me more confident and aware about various options that are available.

- Have tried providing visual representations where ever it seems fit and easy to understand. Thus, gave me a chance to use my excel knowledge of graphs and made the entire process more engaging.
- Has helped me understand how important data analytics is for the Hiring Process of any organization as it provides valuable insights to hiring department to improve and take right decisions.

Links: My excel worksheet link with different sheets for each task,

Note: In the first sheet of Data Cleaning, I have highlighted the changes and formula cells in  colour and references in  colour

<https://docs.google.com/spreadsheets/d/1upDvYqcOfBuPII78SxSKgkpLtcYwV6Bm/edit?usp=sharing&oid=108154584635151678812&rtpof=true&sd=true>