**Project Description:**

Working as a Data Analyst in Google, I have to extract useful insights from the dataset provided. These insights will depict how the applicants are hired in the company based on the number of candidates hired, the type of candidates, positions open, salaries offered, etc. Based on the past data, improvements can be made by the company to hire better employees suited as per their requirement. The approach and the corresponding insights and results will be listed down in detail in the coming pages of this report.

**Tech-Stack Used:**

To implement this project, I have used Microsoft Excel 365 Version 2302. There are two reasons for this:

* MS Excel 365 allows a lot of features that are not available in the previous version such as there are a lot of automated charts which help me to quickly use them on my data without having to perform the calculations manually.
* This subscription is offered to me by the company I currently work in, so I am quite familiar with its ins and outs.

Kindly download the dataset from the link below and view in Excel to see all the charts and tables for the problems (Google sheets doesn’t show all the analytics done).

[Statistics\_DataSet\_Analyzed](https://docs.google.com/spreadsheets/d/1BbW3XQJR3gC3gCbd76A0OqExiprgJyvP/edit?usp=sharing&ouid=104611970421205783778&rtpof=true&sd=true)

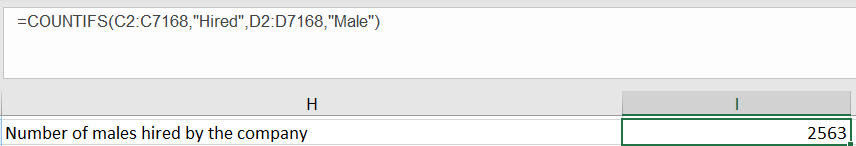
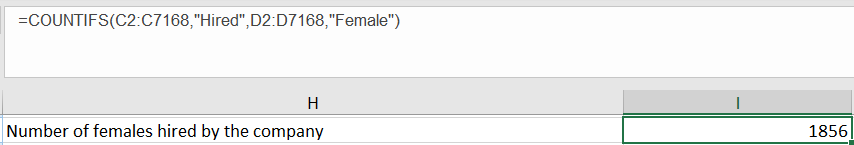
**Handling missing data:**

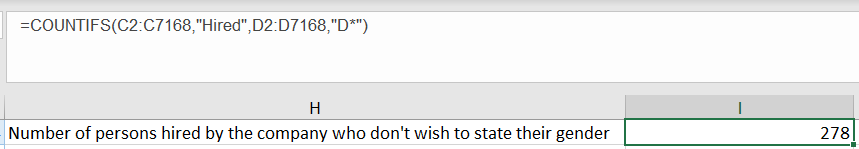
1. We will first create a table out of the data provided and name it “HiredData”. This table will be useful to create shorter pivot tables if needed later.
2. In the dataset provided, in column event\_name, we are given information about the sex of the job candidates. We will rename this column to ‘sex’ for better understanding.
3. Further in this same column we can see that there are three categories – male, female and ‘Don’t want to say’. For some candidates the sex is not specified and is entered with a ‘-‘ symbol.
4. To make our analysis easier we will replace this ‘-‘ symbol with ‘Don’t want to say’ because both of them mean the same thing. We have 15 such records that we will replace.
5. In the offered salary column, we can see that the range is between 1,000 to 1,00,000 but there are 3 records for which the salary is 200000, 300000, and 400000 which seem like outlier values. But we don’t know for sure that these salaries cannot be offered to any employee. The post name is not given in generic terms. There are some posts for which the salary can be considerably much greater than the average salary, for example the director’s salary. So, we will keep these values as they are.
6. We also have a missing cell in the offered salary column. We will replace this missing cell with the average salary offered by the company. Since there are a large number of datapoints, we can fill this missing cell with the average which is 49987.53 ≈ 49988. And even if the offered salary for this post wasn’t equal or near to the average salary it won’t impact our analysis much since the candidate was rejected.
7. Next, we look at the Post Name column and we find that for one record in the Sales Department, the post is missing. To fill this cell, we will look at the Offered Salary column for Sales Department and find the post with the closest salary to the salary with the missing post name. As shown below, we will fill c9 for the missing post since c9 in sales department has the salary which is closest to 85,914.  
   

Problem Description No. 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?

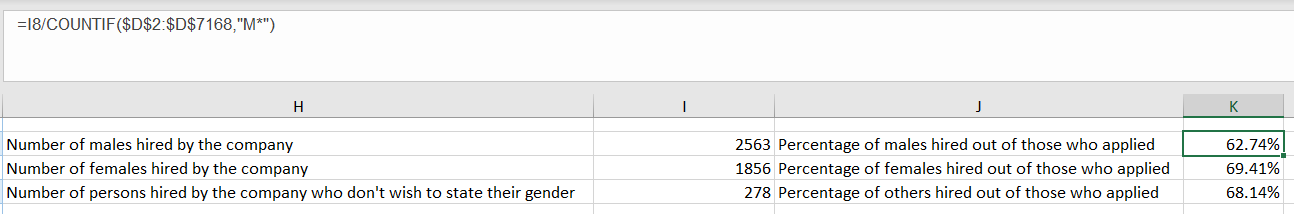
**Approach:**

1. We will look into the columns – status and sex to solve this problem.
2. The excel COUNTIFS function allows us to count the number of records based on multiple criteria.
3. The two criteria are that the status should be hired and the sex should be ‘male’, ‘female’ or ‘Don’t want to say’.
4. Note that in COUNTIFS function **we can’t use an apostrophe sign** when specifying the text criterion in double quotes, therefore, **we will use an asterisk after D** as shown below in the third snapshot.  
    



**Insights:**

From the above count we can see that male candidates are hired more than female candidates. This could be due to a lot of reasons and just finding the count may not give us the exact reason. To find that out we will calculate the percentage of the males and females hired out of the number of males and females who are applying for a job.

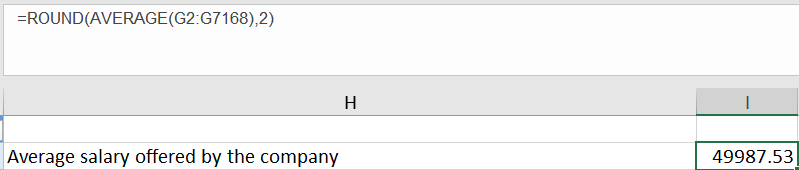
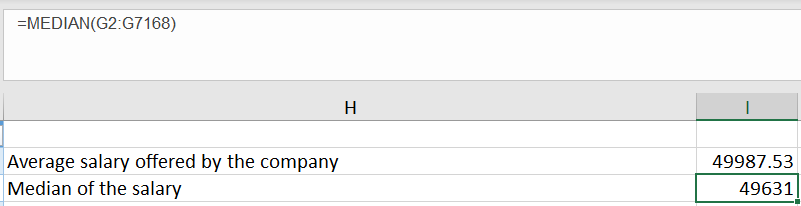
Now if we look at the results, the picture has totally changed. Percentage of females hired is more than that of male candidates.  


This means that even though fewer female candidates have applied for a job, they are hired more. The less number in applications on part of females could be due to gender distribution in the population or various other factors. These numbers can help Google identify those factors and work towards balancing the ratio of men and women working in the organization which can make it a great place to work and a preference for the future job candidates.

Problem Description No. 2

Salary Analysis: What is the average salary offered by this company?

**Approach:**

1. We will use the excel AVERAGE() function over the column **Offered Salary** as shown below: I have also combined it with the ROUND() function to get the result corrected to 2 decimal places.  
   
2. **To check whether the data is symmetrical or not, we will also calculate median** using the excel MEDIAN() function:  
   
3. Now, when we look at the two values, we see that **they are almost equal which means that the data is symmetrical** and that there is no effect of outliers on the dataset.
4. **Note that** this average and median value is calculated after we had replaced the missing salary for one record with the average of the rest of the records. And even if we keep that missing salary as empty, we get the same average and median with the rest of the datapoints.

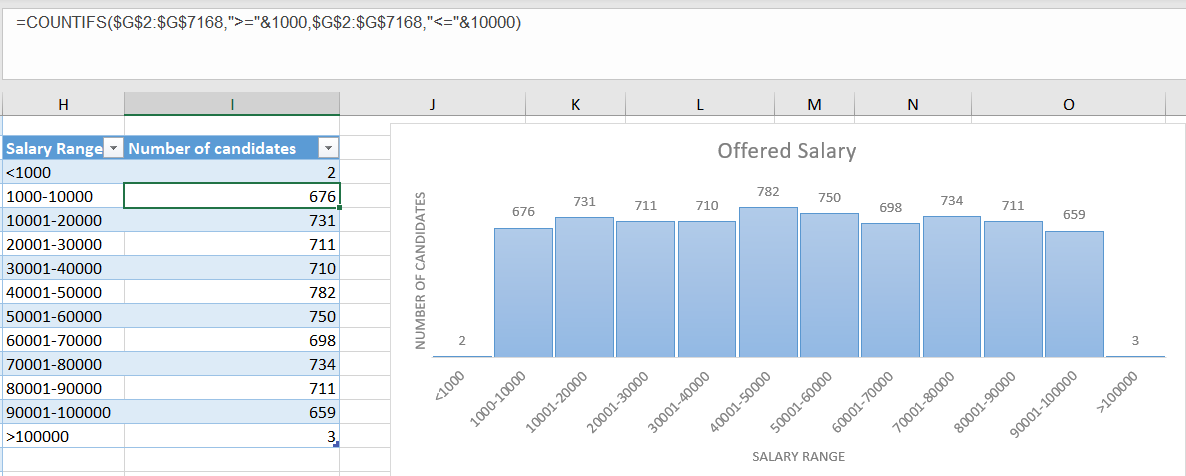
**Insights:**

The above result tells us that the average salary offered by Google is around 50,000 or in other words most open positions are offering 50,000 as salary to the applicants.

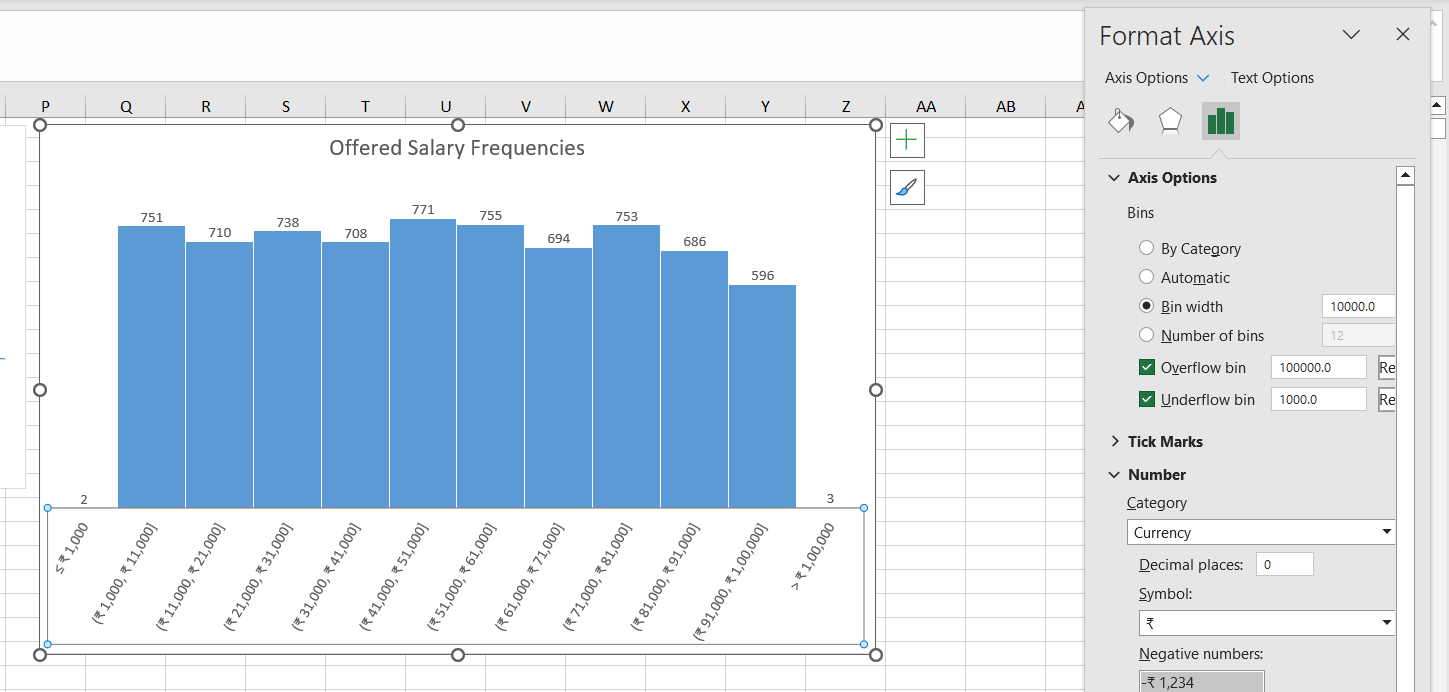
Problem Description No. 3

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.

**Approach:**

1. I have created intervals of 10,000 for each class except the salaries <1000, 1000-10000 and >100000.
2. In the offered salary column, we will use COUNTIFS() function as shown below and create a table to store the frequency, i.e., the number of salaries offered within a particular interval.  
   
3. I have written the formula once with the help of absolute rows and columns using $ sign (or press F4 after writing the range) and copied it down the other intervals changing only the interval limits in the formula (the same as that given in column H)
4. From the table, I have created a 2-D column graph to depict the frequencies.

**Alternative Automated Approach:**

1. We can simply select the **Offered Salary** column from the dataset given and go to Insert tab and select Histogram Chart.
2. I have customized the chart to remove the vertical axis since I am using data labels on the bars.
3. I have the underflow and overflow bins to state the outlier salaries.
4. I have set the bin size to 10000.
5. Chart Title I have renamed to “Offered Salary Frequencies”.
6. And finally, the category of the horizontal axis I have selected as “Currency”. Symbol I have selected as Rupee.  
   
7. The key difference between this automated approach and manual approach here is that in the automated chart we can see that the intervals are inclusive of the boundaries whereas in the manually created chart they are exclusive. In other words, the employee who has salary as 21000 will be counted in **both** the bins 11000-21000 and 21000-31000 in the automated chart.
8. In the manual chart we can provide this exclusion along with stating the exact interval size for each bin specifically which is not possible in automated method.

**Insights:**

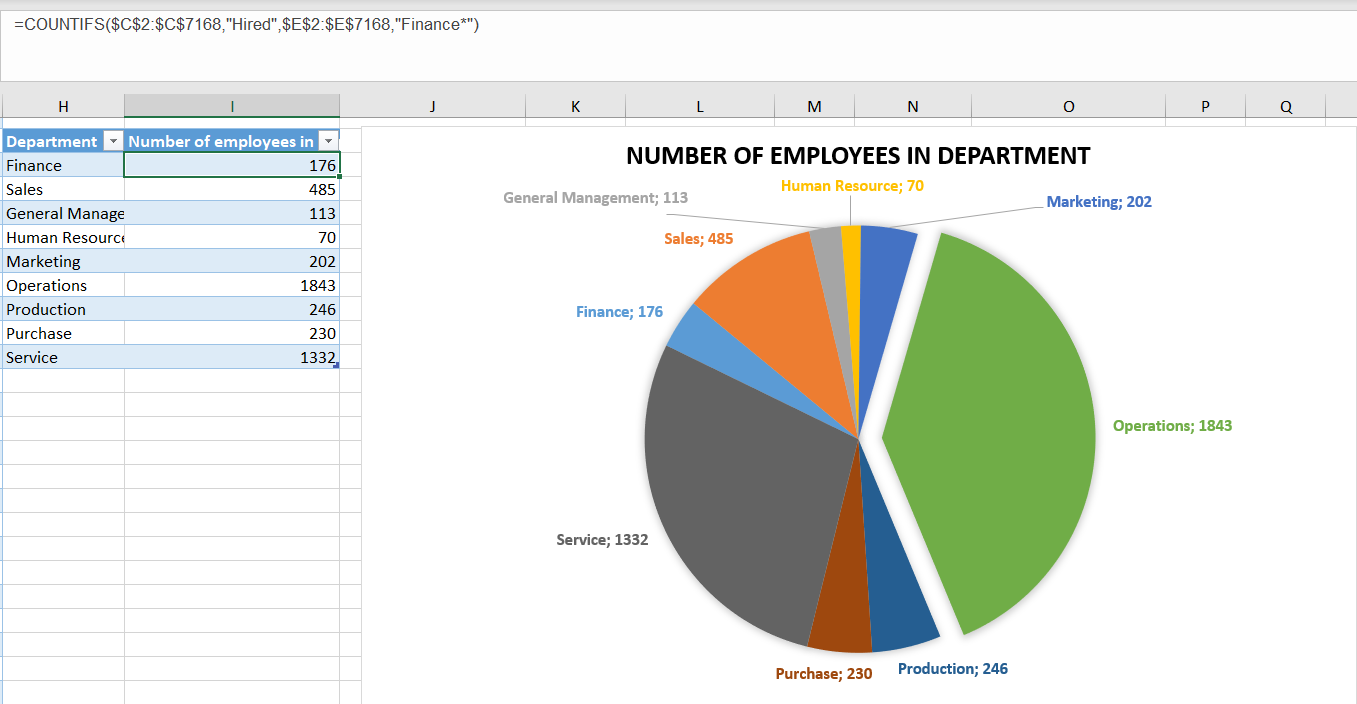
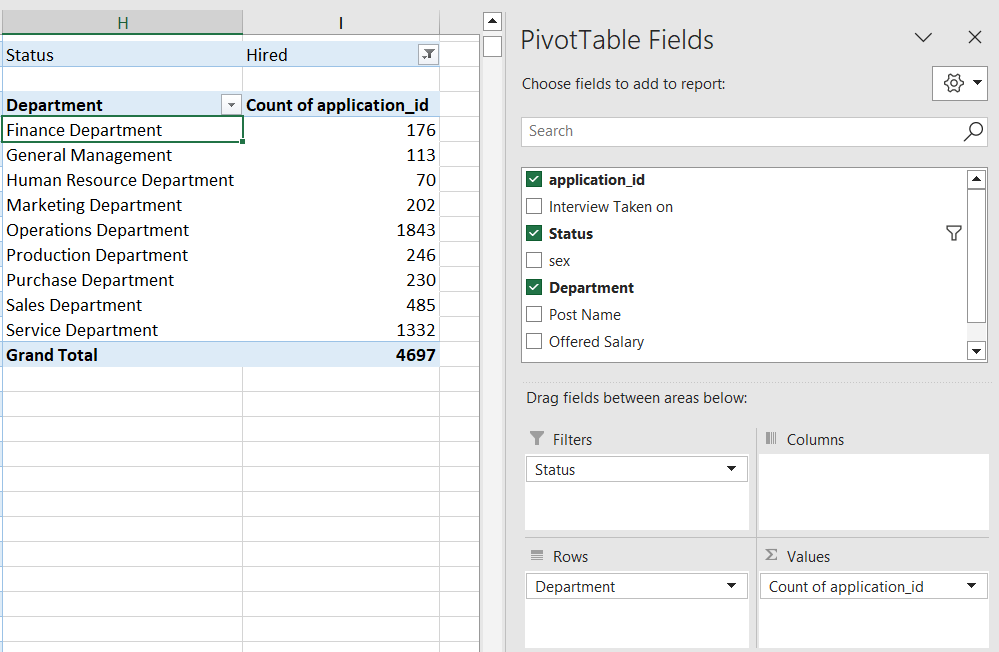
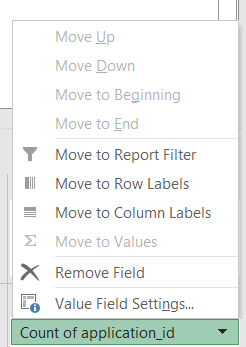
When we look at the chart above, we can see that the most number of openings – 782, are for the salary class 40001-50000 which also matches the average salary which is being offered by Google.

Further, we can see there are 2 salaries which are below 1000 and 3 salaries above 100000.

Problem Description No. 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.

**Approach:**

1. To calculate the people **working** in different departments, we need to first filter out only those people who are hired by the company.
2. To do this, we will use the excel COUNTIFS() function over the two columns – Status (C) and Department (E) and store the count of the number of employees in each department in a separate table as shown below:  
   
3. In the above formula, I have used absolute rows and columns with the help of F4 Key ($) since my range is the same and wildcards with the help of \* for efficiency to reduce typing out the formula again and again.
4. The overall table and the pie chart which depicts the distribution looks like this:  
   
5. Alternatively, we can also make use of Pivot Tables as shown below and generate a pie chart from this table:  
   
6. In the Filters, I have selected Status column to choose only hired candidates. In Rows, I have selected the Department column to depict the category and in Values I have selected the Count of application\_id to display the share of each department.
7. By the default, in the Values section, we get the sum but that can be changed to count by selecting the Value Field Settings which can be accessed through the drop-down icon.  
   

**Insights:**

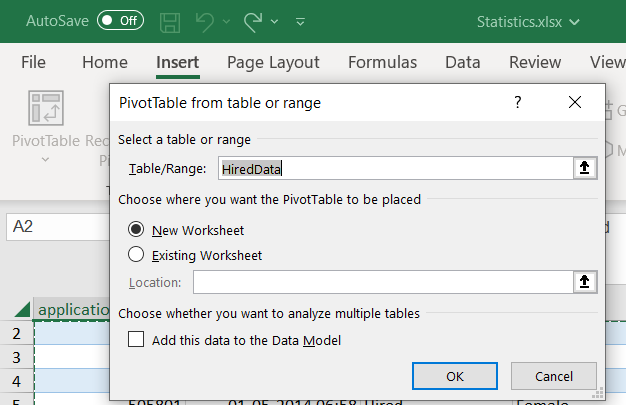
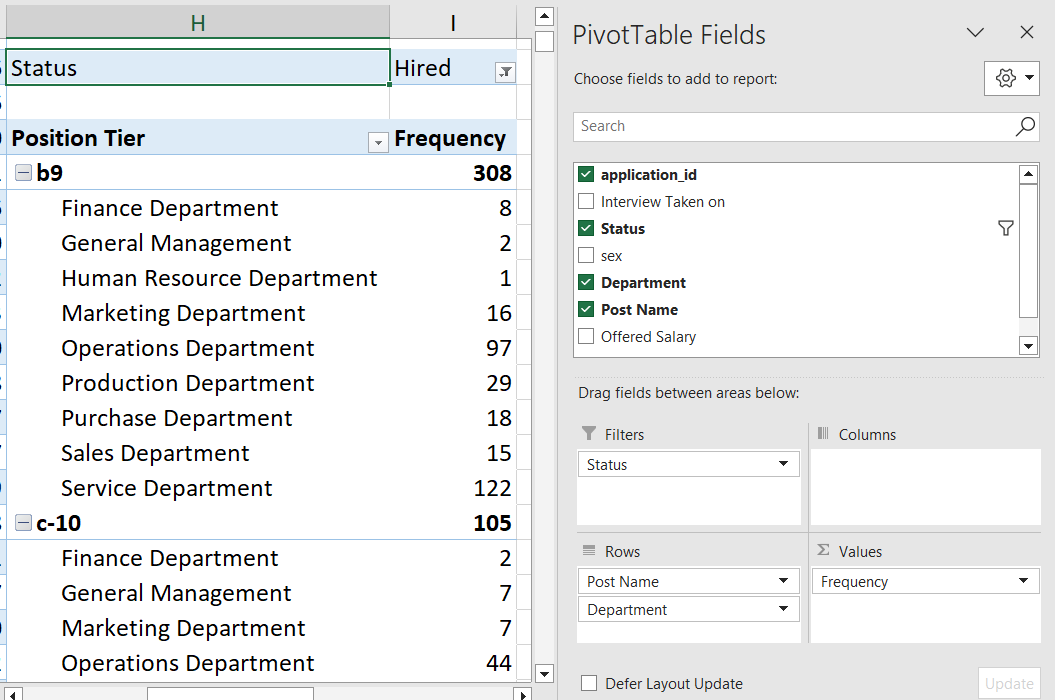
From the first look itself at the pie chart, it is evident that maximum number of employees = 1843 are working in the **Operations Department**, followed by the **service department** with 1332 employees. And we can see that the **Human Resource Department has the least number** equal to 70 employees. This will help Google analyze which department is the most vulnerable department considering the overall company strength through which they can focus for which roles they should create more openings in the company based on their resource requirement.

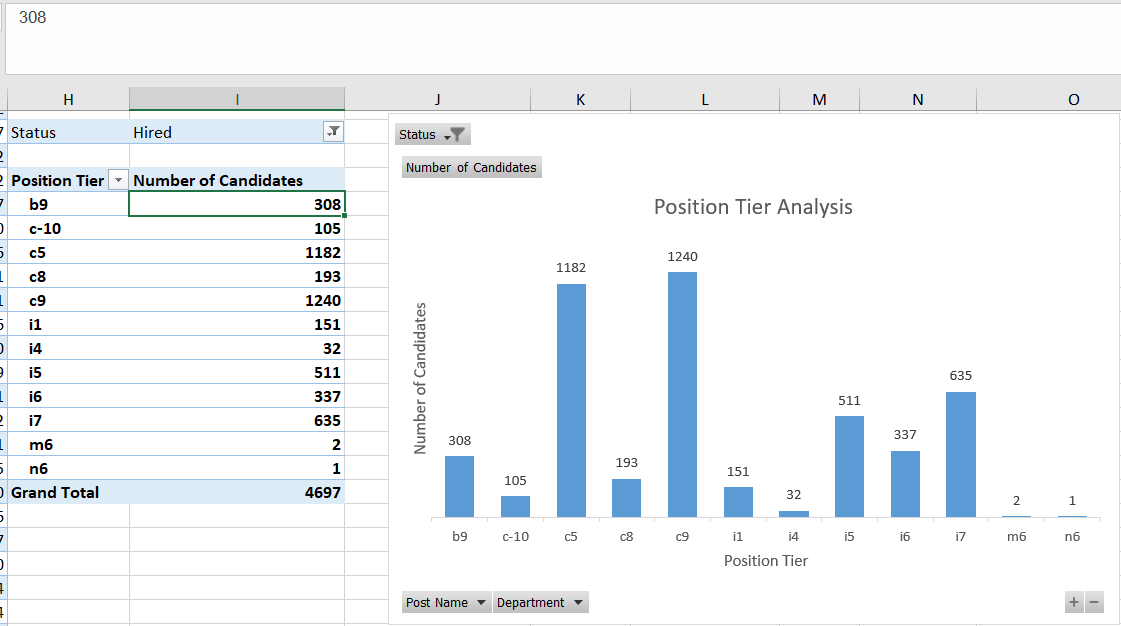
Problem Description No. 5

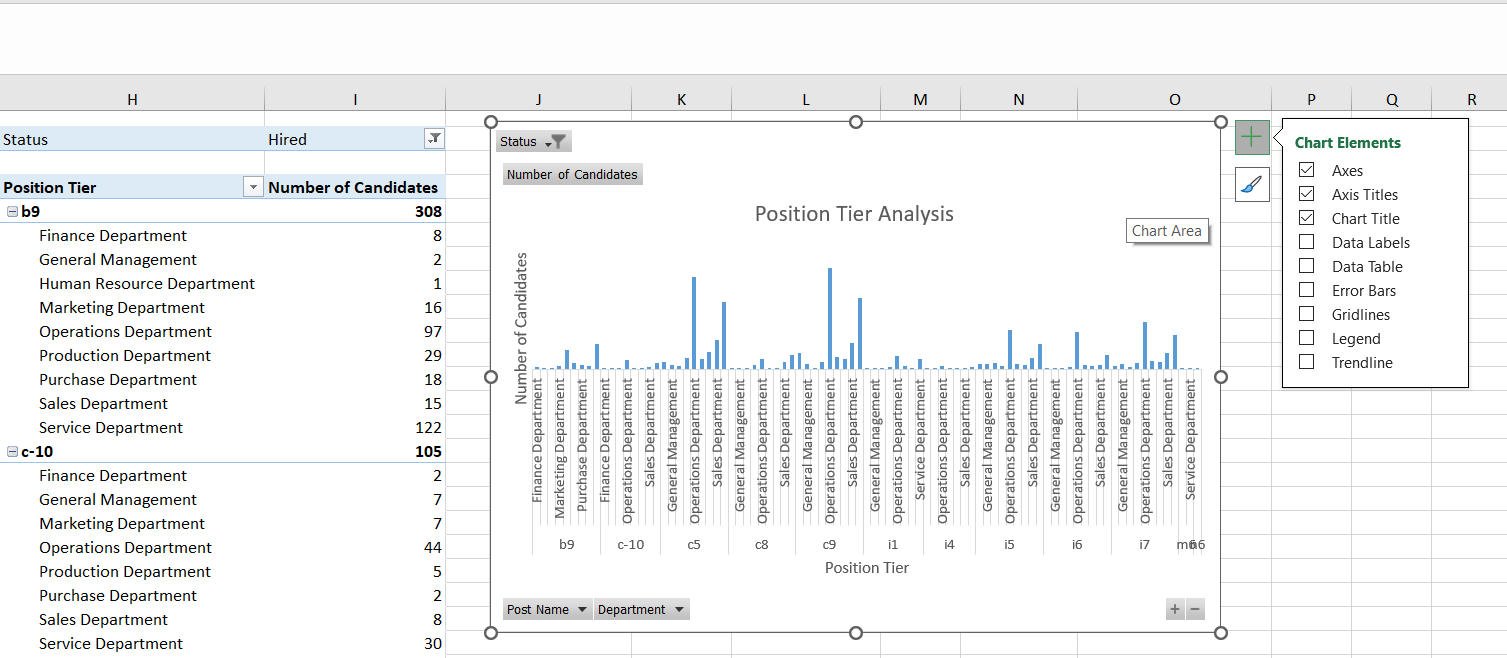
Different positions within a company often have different tiers or levels.  
Task: Use a chart or graph to represent the different position tiers within the company. This will help to understand the distribution of positions across different tiers.

**Approach:**

1. We will create a pivot table from the HiredData table provided to us as shown below:  
   In the location field we can select the cell where we want to place our table.

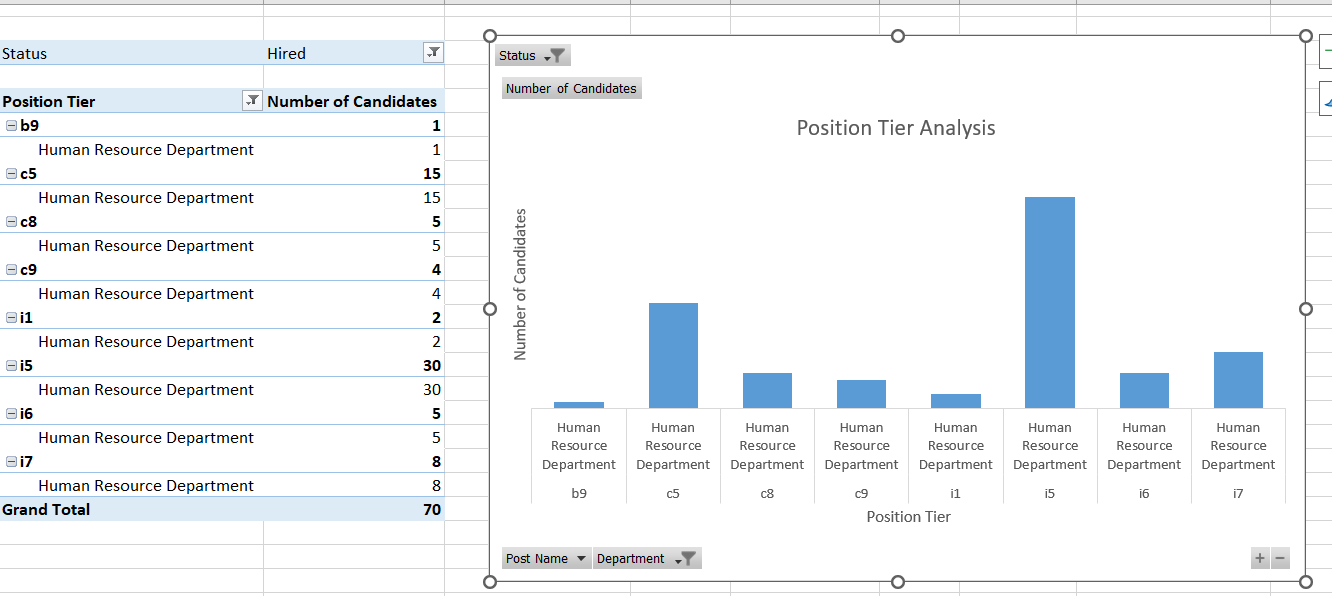
  


1. As in problem description no. 4 I have selected all the attributes in their respective sections and additionally selected the Post Name in the Rows section.
2. With this table, we can create a 2D column graph and either keep it expanded to show the departments for a particular post or minimize it to show the distribution of just the different posts or position tiers.  
   



**In the expanded graph, I have unchecked the Data Labels option** because I want to focus on the bigger picture here, which says that for both the positions c5 and c9, Operations Department has the highest number of openings.

1. We can also filter out either the post name or the department to visualize different results with the help of the Pivot Table that we have created. This would not have been possible if we had used the COUNTIFS() function and then created a table manually.

For example, for the HR department, we have highest number of openings for the i5 position.  


**Insights:**

From the first graph, it is evident that highest number of openings are for the position c9 closely followed by c5. And if we want to find out for exactly which department the posts are for, we can view the expanded chart and see that it is the Operations Department.

**Results:**

Working on this project has allowed to work around the functions that excel offers and apply them to real world problems. Along with this, data visualization through charts and their formatting and customization with respect to the audience has increased my understanding of what wonders Excel can do for any business. It is, I realize a very important tool for a Data Analyst as it has a lot of features which can be used to derive insights from a raw data set. Another way that this project has helped me is to apply my learnings of the excel features and find out that there are multiple approaches to answer the same problem. Even though the approaches are different, they provide the same exact analytical results.