Hiring Process Analytics : Statistics

By: Arpan Sharma(Data Analytics Trainee)

Project Description:-

Hiring Process Analytics project is about finding trends and insights about the hiring process of the company. In this project, I have used the hiring statistics dataset provided by trainity and drawn some conclusions. I have provided insights to topics serving the hiring department of the company by answering questions asked by the management team. I have used Google Spreadsheets and Microsoft Excel for data analytics and data visualization.

Approach:-

Firstly, I have used the basics of the data analytics process to clean the raw data and ask questions from cleaned data. Then, I have used data wrangling to make small data frames for relevant insights to answer all the possible questions. Finally, I combined all the results and visuals into this report.

Tech-Stack Used:-

I have used the web based application "**Google Sheets**" which is part of google online docs and "**Microsoft Excel for Mac version 16.70"** for performing various functions on spreadsheets. Both of these software provide ease of work and make data sharing and real time tracking very easy.

Project Insights:-

The database included tables namely: comments, follows, likes, photos, photo_tags, tags, users.

Table Name	No. of Rows	No. of Columns	Name of Columns
Statistics	7168	7	application_id, Interview Taken on, Status, event_name, Department, Post Name, Offered Salary

Table Details:

Column Name	Null Values	Description
application_id	No null values	It is unique identifier given to candidates at time of application for job
Interview Taken on	No null values	It is timestamp at which interview of candidate is conducted
Status	No null values	Status of candidate (If hired or rejected)
event_name	No null values	Gender of the candidate(if male/female/don't want to say)
Department	No null values	Department which is hiring
Post Name	No null values	Internal code for post for which department is hiring
Offered Salary	1	Offered salary to the candidates

Data Cleaning:-

Removing Null values:

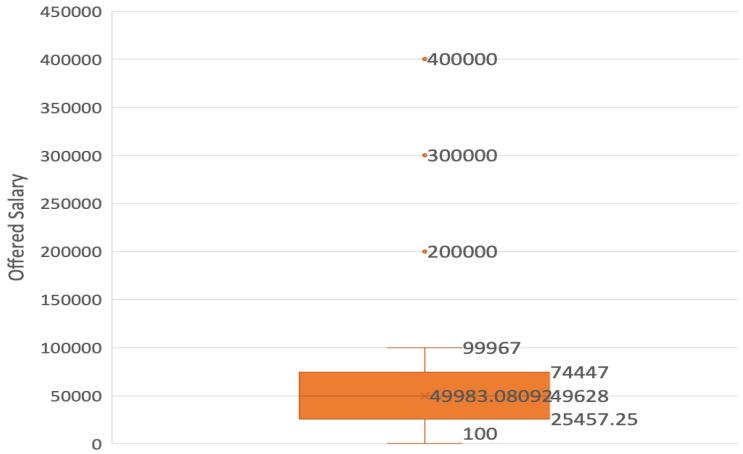
- I find out that there is null value in row 79 in column offered salary:

application_id	Interview Taken on	Status	event_name	Department	Post Name	Offered Salary
114584	5/7/14 8:08	Rejected	Male	Sales Department	i7	

- So, there are two options: either remove the whole row or fill the data with mean as its numerical data. I calculated the mean by selecting the department as sales department, post name as i7 and event_name as male which gave the mean value to be 50355.

Finding outliers in numerical data:

- In offered salary column i have used quartile function to calculate quartile 1 and 3 and their difference to calculate interquartile range using formula:
  - `=QUARTILE(G2:G7169,1)` for Quartile 1
  - `=QUARTILE(G2:G7169,3)` for Quartile 3
  - `(Quartile 3 - Quartile 1)` for Inter Quartile Range
- Then i used formula:-
  - `=IF(G2<$K$9-$K$11*1.5, "Outlier", IF(G2>$K$10+$K$11*1.5, "Outlier", "Normal"))`
  - This formula checks to see if an observation is 1.5 times the interquartile range greater than the third quartile or 1.5 times the interquartile range less than the first quartile. If either is true, the observation is assigned as an outlier.



  - Using this we found three outliers in our data that is row number 12,285,6824

application_id	Interview Taken on	Status	event_name	Department	Post Name	Offered Salary	
649039	5/7/14 10:48	Hired	Female	Service Department	b9	200000	Outlier
795330	6/15/14 9:45	Hired	Female	General Management	i4	400000	Outlier
874368	7/21/14 15:39	Hired	Male	General Management	i7	300000	Outlier

Data Analysis :-

A. **Hiring:** How many males and females are Hired ?

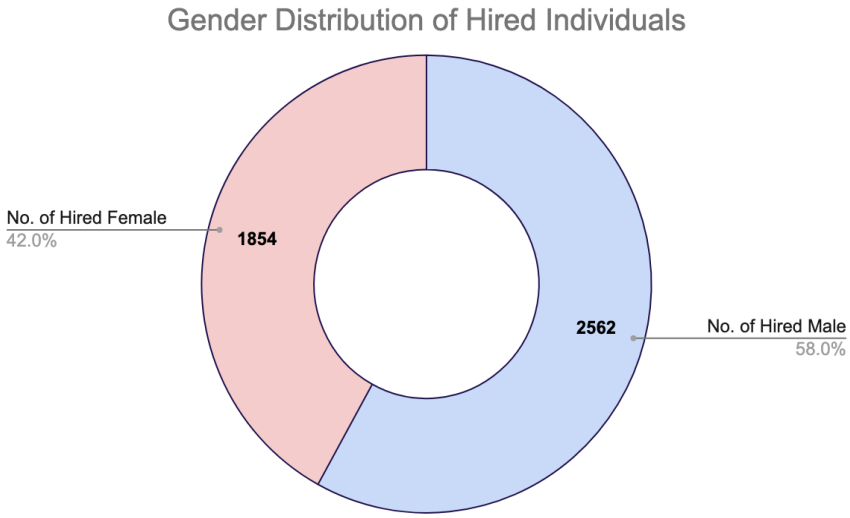
Approach - I used countifs formula to calculate the number of people with condition that their status should be “Hired” and event_name should be “Male” or “Female”. Formula used is-

- `=countifs(D2:D7166,"Male",C2:C7166,"Hired")` “for Male”
- `=countifs(D2:D7166,"Female",C2:C7166,"Hired")` “ for Female”

Result Grid :-

Task 1	How many males and females are Hired
No. of Hired Male	2562
No. of Hired Female	1854

Chart representation :-



Conclusion :- The number of hired male candidates is 2562 and number of hired female candidates is 1854.

B. **Average Salary:** What is the average salary offered in this company.

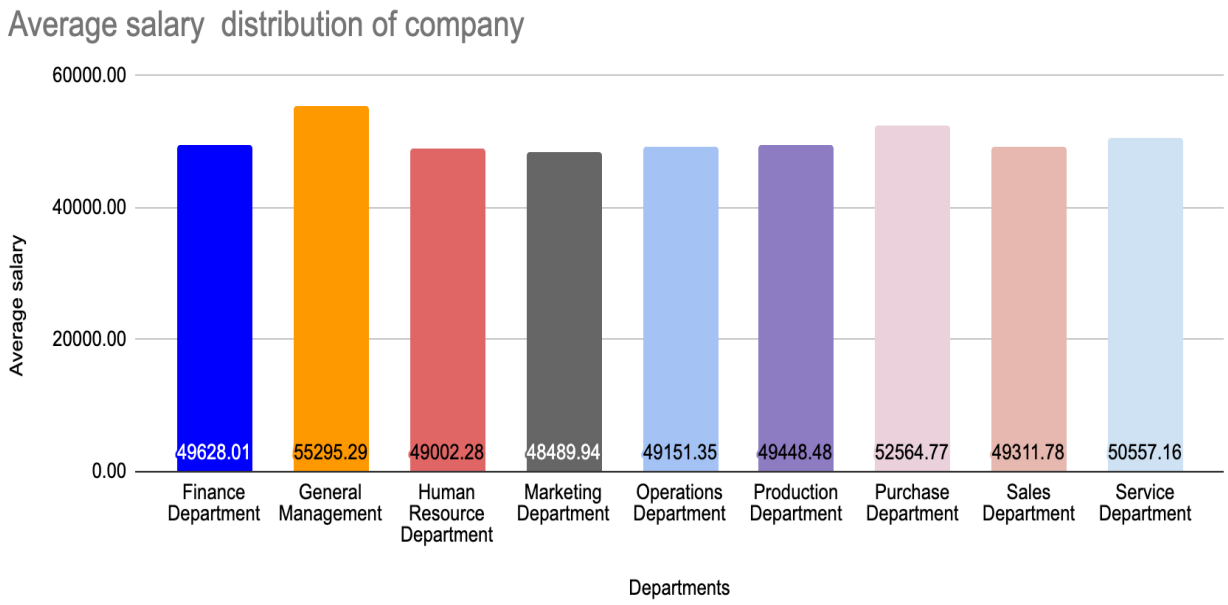
Approach - I have used the Average function to calculate salaries offered by company with and without outliers and Averageif function to calculate department wise average salary. Formula used is-

- `=AVERAGE(G2:G7166)` -after removing outliers
- `=AVERAGE(G2:G7169)` -before removing outliers
- `=AVERAGEif($E$2:$E$7169,"Finance Department",$G$2:$G$7169)` -using names of different departments as a condition to calculate the average of that group.

Result Grid :-

Task 2	Average salary offered in this company
Average Salary(Without Outliers)	49878.39833
Average Salary(With Outliers)	49983.0809
	Average salary offered in each department in this company
Finance Department	49628.00694
General Management	55295.29412
Human Resource Department	49002.27835
Marketing Department	48489.93538
Operations Department	49151.35438
Production Department	49448.48421
Purchase Department	52564.77477
Sales Department	49311.77912
Service Department	50557.16261

Chart representation :-



Conclusion :- The average salary offered by the company is 49878.39833. There is no significant difference in average salary with or without outliers. The average salary offered by each department of the company is presented in the result grid.

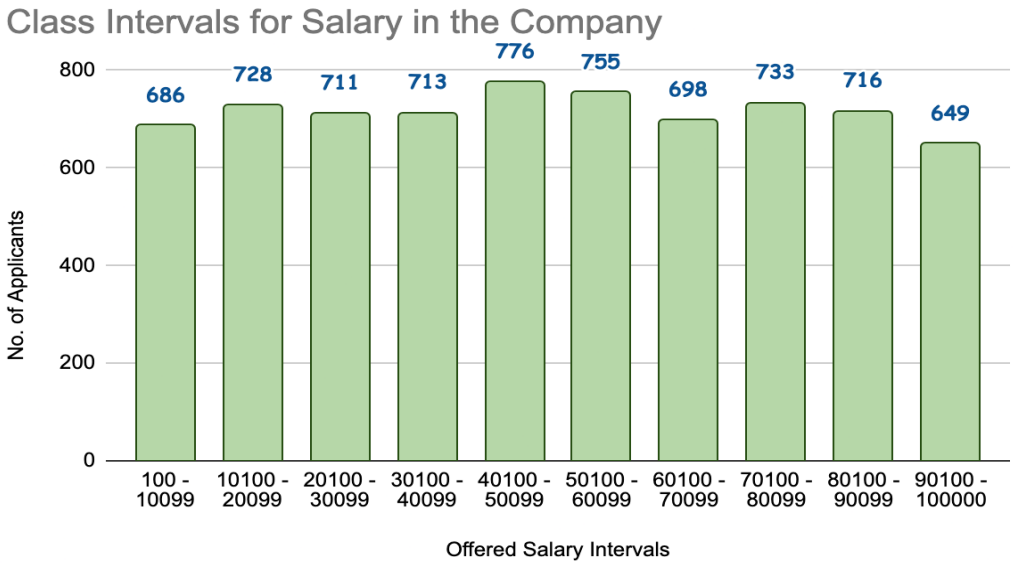
C. **Class Intervals:** Draw the class intervals for salary in the company.

Approach: I have used the pivot table to create class intervals of offered salary in the company. I have selected the offered salary as rows and application_id as value to count. Then I used the pivot group rule where I kept the minimum value as 100 and maximum value as 100000 and intervals as 10000.

Result:

Offered Salary Intervals	No. of Applicants
100 - 10099	686
10100 - 20099	728
20100 - 30099	711
30100 - 40099	713
40100 - 50099	776
50100 - 60099	755
60100 - 70099	698
70100 - 80099	733
80100 - 90099	716
90100 - 100000	649
Grand Total	**7165**

Chart representation :-



Conclusion :- The number of applicants in offered salary class intervals can be seen in the result table. Total number of applicants is 7165 after removing 1 null value and 3 outliers. There is very slight variation in the number of individuals getting different classes of salaries. Most number of people that is 776 are offered salaries in the range of 40,100 to 50,099.

D. **Charts and Plots:** Draw Pie Chart / Bar Graph (or any other graph) to show proportion of people working in different departments.

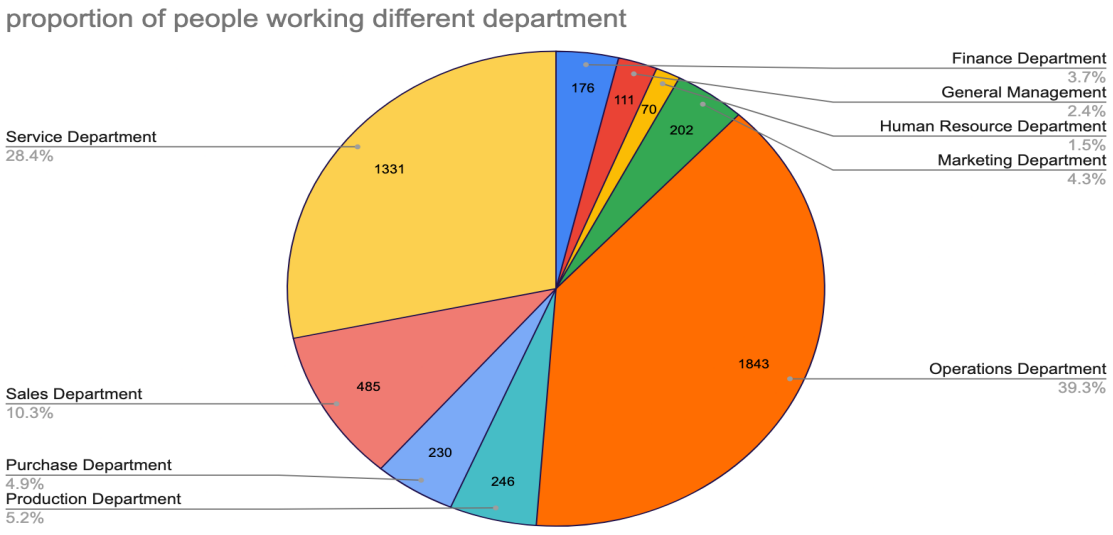
Approach: I used the countifs formula to calculate the number of people with the condition that their "Status" should be "Hired" and their "Department" should be their respective department. I have also used count if to count the total number of individuals working in the company. I have also made an assumption that the "event_name" = "-" is "Don't want to say". Formula used is-

- `=countifs($C$2:$C$7169,"Hired",$E$2:$E$7169,"Finance Department")` for total number of people working in the respective department.
- `=countif($C$2:$C$7169,"Hired")` for total number of people working in company
- `=countifs($C$2:$C$7169,"Hired",$E$2:$E$7169,J21,$D$2:$D$7169,"Female ")` for total number of people of certain gender working in company

Result:-

Task 4	Male	Female	Don't want to say	Total
Company	2562	1854	278	4694
Finance Department	10	154	12	176
General Management	9	94	8	111
Human Resource Department	43	26	1	70
Marketing Department	127	66	9	202
Operations Department	1033	695	115	1843
Production Department	128	104	14	246
Purchase Department	133	76	21	230
Sales Department	294	171	20	485
Service Department	785	468	78	1331

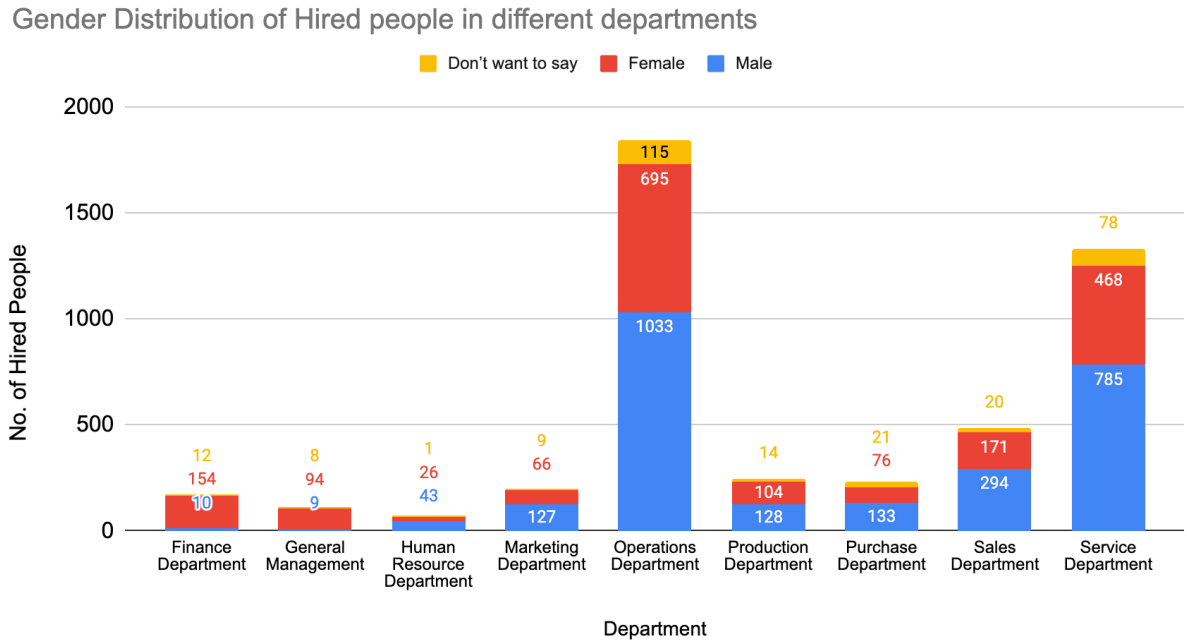
Chart representation 1 :- Proportion of people working different department



Conclusion: From the pie chart it is clearly visible that the majority of the workforce works in operations and service department accounting for 39.3 and 28.4 % respectively. Sales department accounts for around 10% of the workforce. The people working in Purchase and

Production account for around 5% each whereas the same in Finance and Marketing department account for around 4% each. Only 2.4% and 1.5% of total people work for the General management and HR department respectively.

Chart representation 2 :- Gender Distribution of Hired people in different departments



Conclusion: The stacked column chart representing gender distribution of hired people in different departments represents the ratio of different gender working in each department. It is clearly visible that male gender is dominant in operations, service, sales, purchase, production, marketing and HR department. On the other hand, Women are hired in finance and general management in higher numbers than any other gender. 278 people who did not mention their gender, are working for every department and majority of them are working in operations and service departments.

E. **Charts:** Represent different post tiers using chart/graphs.

Approach : I have made 3 chart representations to present different post tiers. I have also made two assumptions that the "post name" = "-" must be a null value and the "event_name" = "-" is "Don't want to say".

- For first chart representation, I have used the countif formula to calculate total number of people working on the respective post and number of people of certain gender working at that post.
  - `=countifs($C$2:$C$7169,"Hired",$F$2:$F$7169,"b9")` for total people.
  - `=countifs($C$2:$C$7169,"Hired",$F$2:$F$7169,"b9",$D$2:$D$7169,"Male")` for specific gender
- For second chart representation, I have used conditional to find the number of people working at the respective post in different departments.

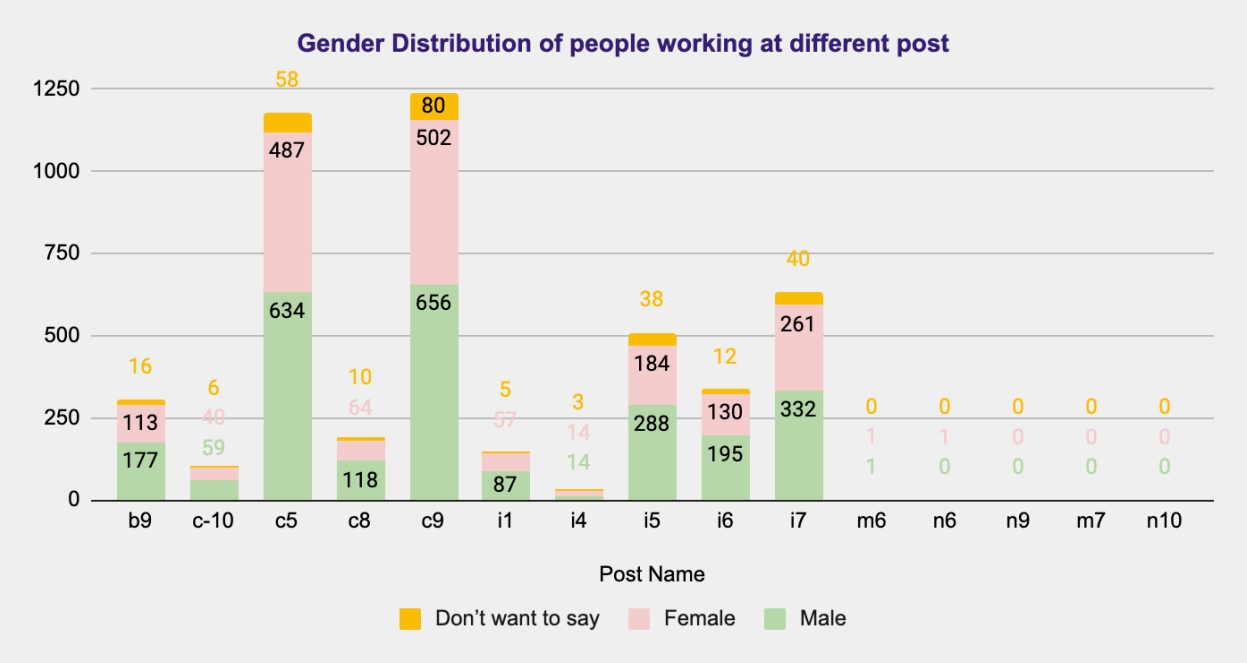
- =countifs($C$2:$C$7169,"Hired",$F$2:$F$7169,"b9",$E$2:$E$7169,"Finance Department")

- For third chart representation, I have used the Averageif function to calculate the average salary offered to hired individuals at a given post.
  - =AVERAGEiFS($G$2:$G$7169,$F$2:$F$7169,"b9",$C$2:$C$7169,"Hired")

Result 1:-

Post Name	Total	Male	Female	Don't want to say
b9	307	177	113	16
c-10	105	59	40	6
c5	1182	634	487	58
c8	193	118	64	10
c9	1239	656	502	80
i1	151	87	57	5
i4	31	14	14	3
i5	511	288	184	38
i6	337	195	130	12
i7	634	332	261	40
m6	2	1	1	0
n6	1	0	1	0
n9	0	0	0	0
m7	0	0	0	0
n10	0	0	0	0

Chart representation 1 :-

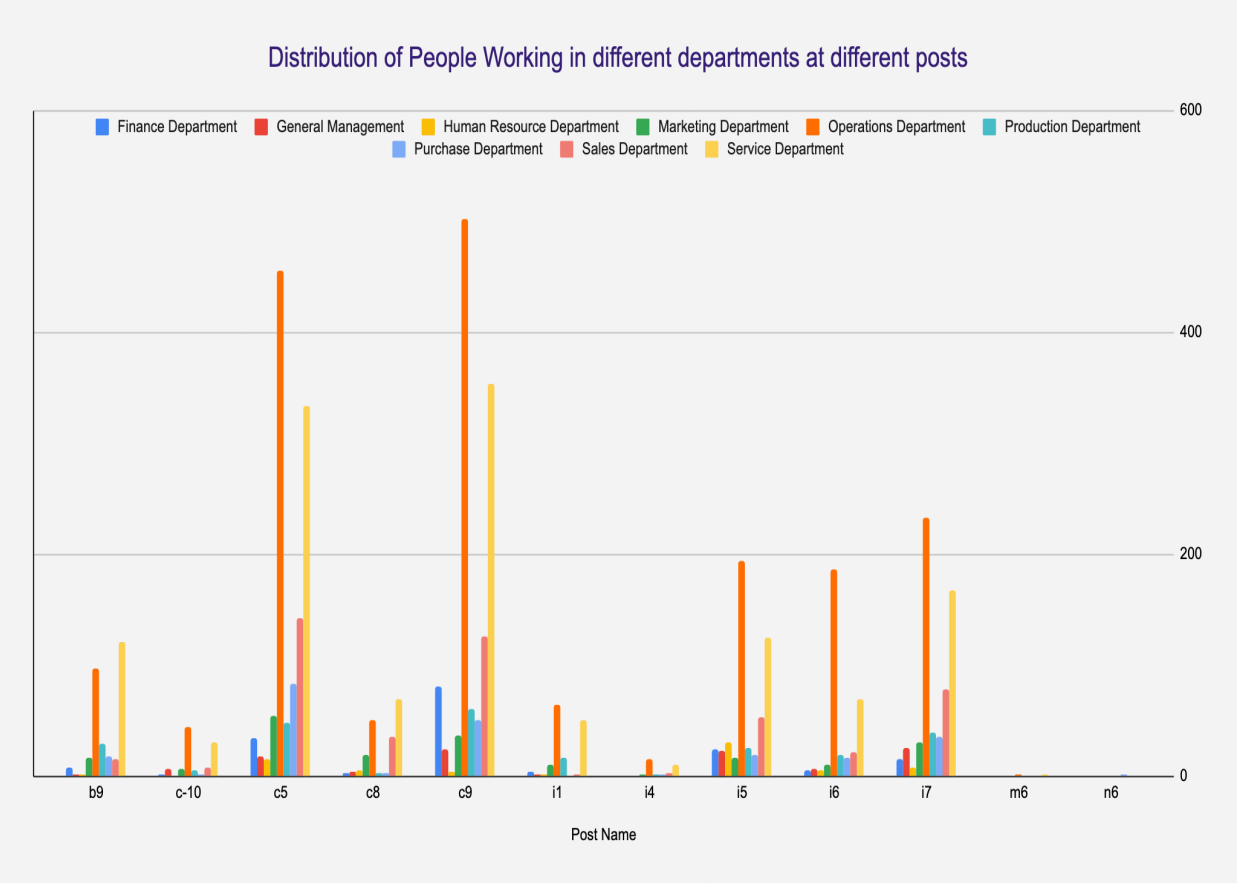


Conclusion:- No one is working at post name m7, n9 and n10. Majority of people are working at post "c9" and "c5". Except post "n6","i4" and "m6", Male are working in majority at every post.

Result 2:-

Post Name	Finance Department	General Management	Human Resource Department	Marketing Department	Operations Department	Production Department	Purchase Department	Sales Department	Service Department
b9	8	2	1	16	97	29	18	15	121
c-10	2	7	0	7	44	5	2	8	30
c5	34	18	15	54	455	48	83	142	333
c8	3	4	5	19	51	3	3	36	69
c9	81	24	4	37	502	60	51	126	354
i1	4	1	2	10	65	16	0	2	51
i4	0	0	0	1	15	1	1	3	10
i5	24	23	30	17	194	26	19	53	125
i6	5	6	5	10	186	19	16	21	69
i7	15	26	8	31	233	39	36	78	168
m6	0	0	0	0	1	0	0	0	1
n6	0	0	0	0	0	0	1	0	0

Chart representation 2 :-

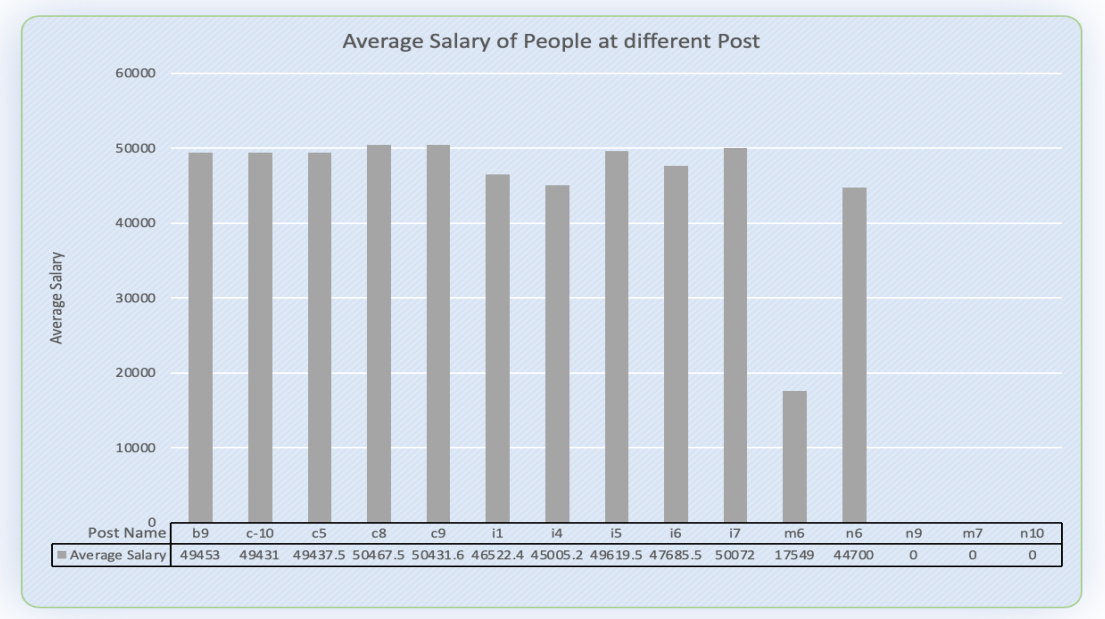


Conclusion:- Except post “n6”, personnel of the operations department are working at every post in majority. Except post “i5”, personnel of HRD are working at every post level in minority.

Result 3:-

Post Name	Average Salary
b9	49452.98697
c-10	49430.9619
c5	49437.51861
c8	50467.53368
c9	50431.55044
i1	46522.38411
i4	45005.22581
i5	49619.48728
i6	47685.54006
i7	50071.96215
m6	17549
n6	44700

Chart representation 3 :-



Conclusion: Average salary of people working at different posts in all the departments does not vary much except that of post “m6” which is approximately less than half than all other departments.

Result:-

I have answered all the questions asked by the company in this project and explained the result grid and conclusion under the project insights part. While doing the project I applied my learning of statistics and understanding of different functions, pivot tables, conditionals used in spreadsheets.