Project 4 : By Ramana Bansal

Hiring Process Analytics

Project Description : The data is related to 2014 hiring process that took place during the months of May, June, July and August for various positions in a company. Various columns given in the dataset are Application_id, Interview taken on, Status, Event_name, Department, Post_name and Offered Salary. The project focuses on using given data to analyze the hiring statistics, such as gender of applicants hired, various department proportions and salary ranges.

Approach: The data was processed and analyzed using Microsoft Excel.

Tech-Stack Used: Microsoft Excel 2010.

Insights: The project was helpful in getting a general idea about the practical application of data analysis. It helped in understanding how data can be used in almost any process to gain insights and work on improved functioning of an organization.

Result: The project made me feel a little more confident with Microsoft Excel and its use in data cleaning and statistical analysis. It also helped me to understand and apply various statistical formulae. It gave insights regarding how data can be used to gather information regarding hiring processes.

EDA

1. **<u>Understanding data columns and data</u>**

The data is related to 2014 hiring process that took place during the months of May, June, July and August for various positions.

Various columns given in the dataset are:

Application_id : Id of the job applicant

Interview taken on : Date and time of interview

Status : Whether hired or rejected

Event_name : Gender of the applicant

Department: Department of vacancy

Post_name: Name of post being applied for

Offered Salary: Salary offered to the particular applicant

2. **<u>Checking for missing data</u>**

(a) <u>Missing salary data</u>: Only 1 row.
The blank salary can be replaced by average company salary. We use GoTo (Ctrl+G) and special reference to find blank cells. The average salary is computed using average function and selecting Offered Salary column as range. In case of multiple blanks, Ctrl+Enter can be used to fill all blanks with average salary.

2	Missing data						
3	application_id	Interview Taken on	Status	event_name	Department	Post Name	Offered Salary
4	114584	07-05-14 8:08	Rejected	Male	Sales Department	i7	
5							

(b) <u>Missing event_name i.e. gender:</u> 15 such rows.

Events refers to the gender of the interviewee. The blanks can be replaced with 'Don’t want to say' since that seems to affect the data the least.

application_id	Interview Taken on	Status	event_name	Department	Post Name	Offered Salary
195323	09-05-14 12:48	Hired	-	Service Department	i7	81757
742283	02-05-14 8:11	Rejected	-	Service Department	i5	100
227046	27-08-14 18:08	Hired	-	Operations Departmen	b9	76730
711350	16-07-14 13:33	Rejected	-	Operations Departmen	c-10	25785
835053	16-05-14 18:34	Hired	-	Operations Departmen	c5	25583
444043	11-07-14 14:52	Hired	-	Sales Department	c5	80262
352309	20-08-14 10:38	Hired	-	Service Department	i5	4308
204014	09-08-14 16:09	Rejected	-	Purchase Department	c5	96396
901867	18-08-14 9:36	Rejected	-	Service Department	c5	22393
937905	08-08-14 19:29	Hired	-	Marketing Department	c9	94032
564743	28-08-14 10:25	Rejected	-	Production Departmen	c9	4076
245473	14-05-14 18:48	Hired	-	Service Department	c5	66948
411295	22-06-14 14:38	Hired	-	Operations Departmen	i1	98070
487617	30-05-14 16:29	Hired	-	Service Department	c8	12470
827628	30-08-14 15:51	Hired	-	Service Department	i1	3134

(c) Missing post_name : 1 such row.
Doesn't affect our analysis much.

application_id	Interview Taken on	Status	event_name	Department	Post Name	Offered Salary
289907	01-05-14 7:44	Hired	Male	Sales Department	-	85914

3. <u>Clubbing columns with multiple categories :</u>

No such columns in current data.

4. **Checking for outliers**: 3 outliers found based on salary offered.

Q0 :	100	(minimum)
Q1 :	25463.75	(25 percentile)
Q2 :	49628	(Mean)
Q3 :	74429	(75 percentile)
Q4 :	400000	(maximum)

IQR: 48965.25
(IQR = Q3-Q1)

Q : Quartile
IQR: Interquartile range

Lower Limit: -47984.1 (or 0)
Upper Limit: 147876.9

Lower limit= Q1 - (1.5*IQR)
Upper Limit = Q3- (1.5*IQR)

Outliers are either less than lower limit or greater than upper limit.

We use an OR function to find outliers in data. The value less than lower limit OR greater than upper limit is considered outlier.

OUTLIERS

application_id	Interview Taken on	Status	event_name	Department	Post Name	Offered Salary	Outlier
649039	07-05-14 10:48	Hired	Female	Service Department	b9	200000	TRUE
795330	15-06-14 9:45	Hired	Female	General Management	i4	400000	TRUE
874368	21-07-14 15:39	Hired	Male	General Management	i7	300000	TRUE

5. Removing outliers

Since there are just three outliers in the data set, they can be easily removed by deletion. Thus, the outliers won't skew the analysis.

6. Drawing Data Summary

The outliers in salary range have been removed. Using Descriptive Statistics for Offered Salary from Data Analytics tab:

Offered Salary	
Mean	49878.3318
Standard Error	334.9894768
Median	49614.5
Mode	72843
Standard Deviation	28353.64864
Sample Variance	803929390.9
Kurtosis	-1.179493094
Skewness	0.013177306
Range	99867
Minimum	100
Maximum	99967
Sum	357328369
Count	7164

Count of hired:	4694
Count of rejected:	2471
Total:	7165

The missing values in event_name column, which specifies gender of interviewee, have been filled with third option 'Don’t want to say'.

No. of missing values in event_name column.	15

There are 9 departments involved in hiring process, namely with count:

Finance Department	288
General Management	170
Human Resource Department	97
Marketing Department	325
Operations Department	2771
Production Department	380
Purchase Department	333
Sales Department	747
Service Department	2054
Total	7165

There is one missing value in post name column, but since it won't affect our analysis much, it can be ignored.

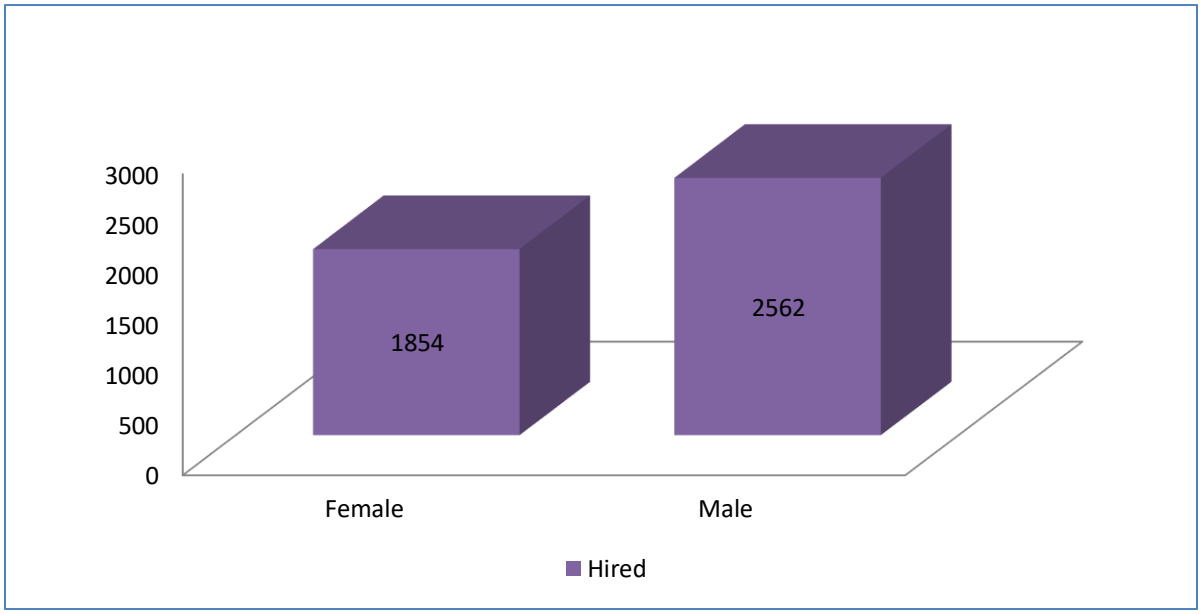
No. of males hired (using countifs)	2562
No. of females hired ((using countifs)	1854
Minimum salary	100
Maximum salary	99574

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

Can be done using pivot table as well as countifs function.

For pivot table, keep event_name in row labels, status in column labels and count of status in values.

1			
2	**1. Hiring**		
3	**Count of Statu**	**Column Labels**	
4	**Row Labels**	**Hired**	**Grand Total**
5	Female	1854	1854
6	Male	2562	2562
7	**Grand Total**	**4416**	**4416**
8			



Atleast 2562 males and atleast 1854 females were hired. A total of 4694 interviewees were hired, however some of them either left the gender blank or didn't want to mention it.

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

The outliers have been removed from the data before calculating average salary. Can be done either using average function or pivot tables.
For pivot tables, put offered salary in values, and select average in value field settings.

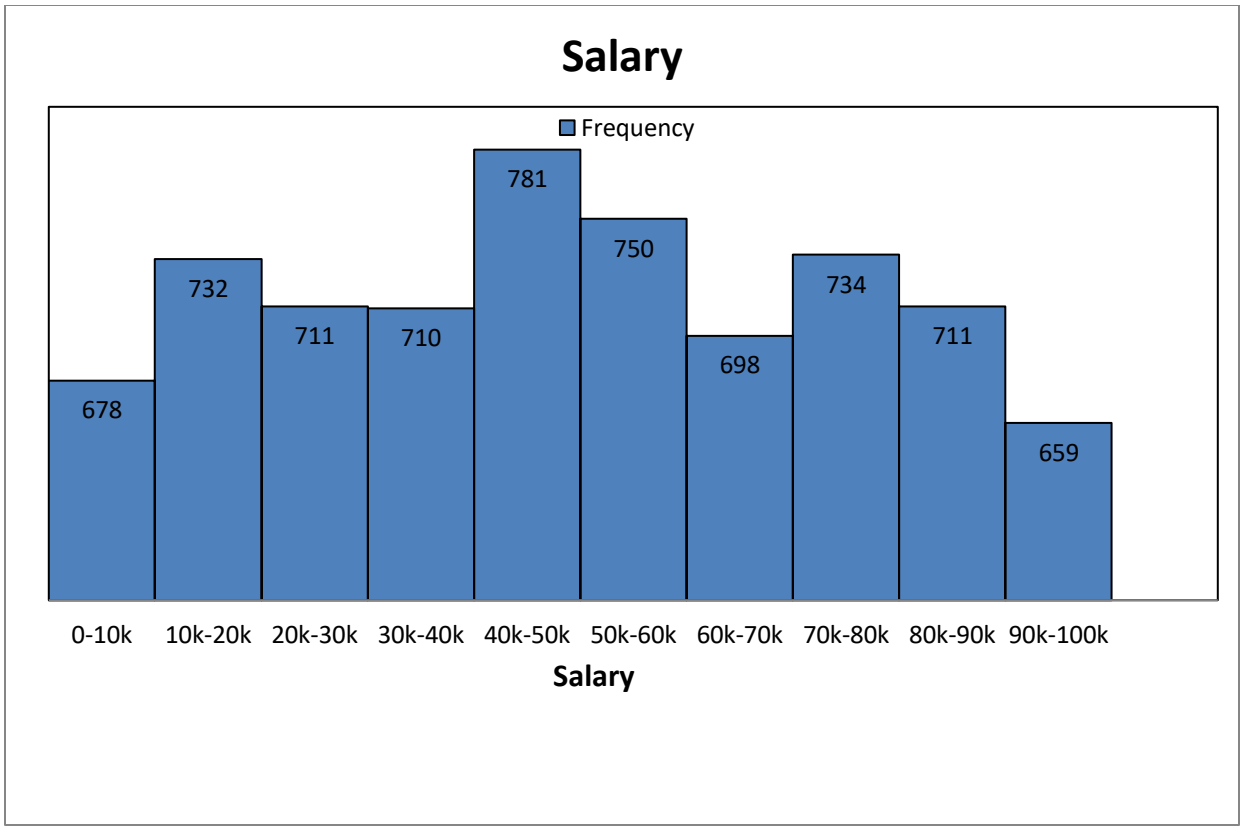
C. **Class Intervals:** Draw the class intervals for salary in the company?

Minimum salary	100
Maximum salary	99574

(Three Outliers have been removed.)

Salary	*No. of employees*
0-10k	678
10k-20k	732
20k-30k	711
30k-40k	710
40k-50k	781
50k-60k	750
60k-70k	698
70k-80k	734
80k-90k	711
90k-100k	659

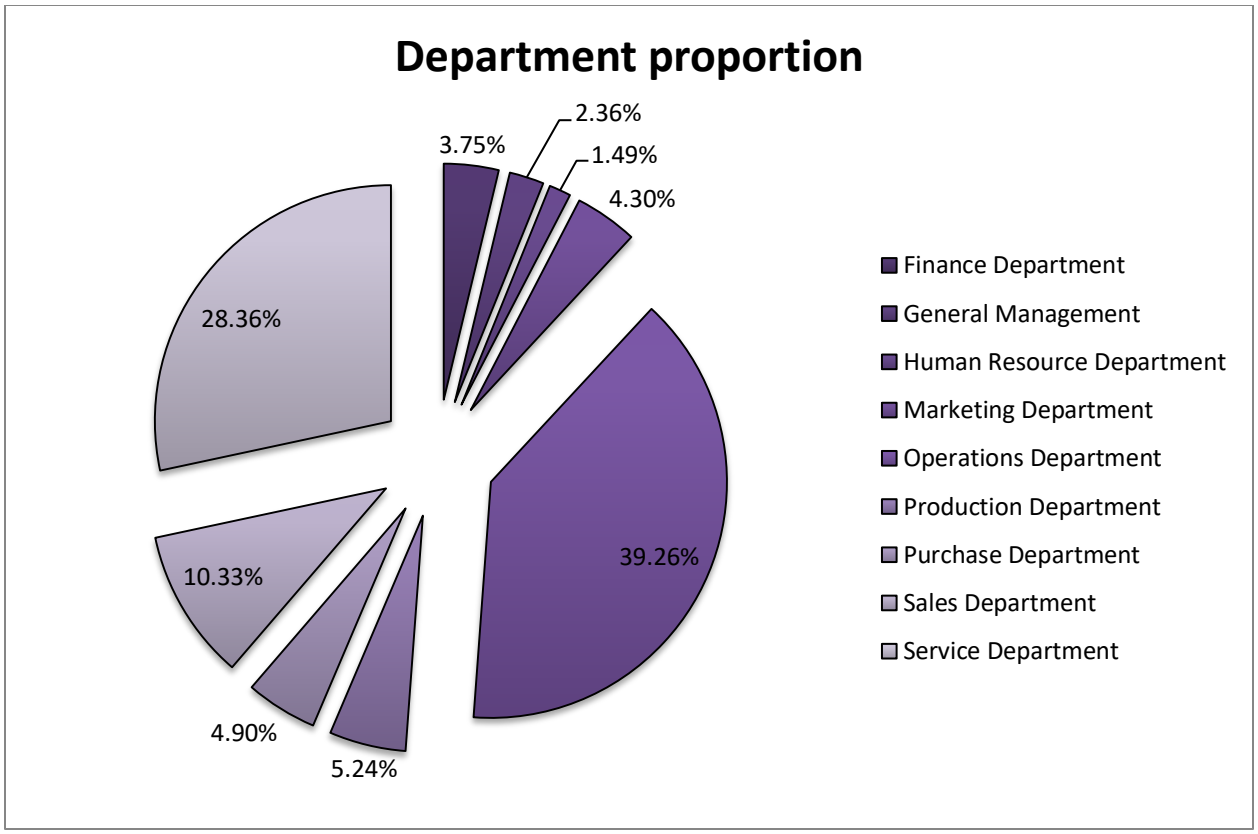
The minimum and maximum salaries were calculated using min and max functions. The minimum salary was 100 and maximum was 99574. Three Outliers were removed before calculating these. Therefore, range from 0 to 100000 was used to derive class intervals with a gap of 10000 each.



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

Using pivot table, Department in row labels, Status (hired) in column labels and count of application_id and count of application_id as percentage in values.

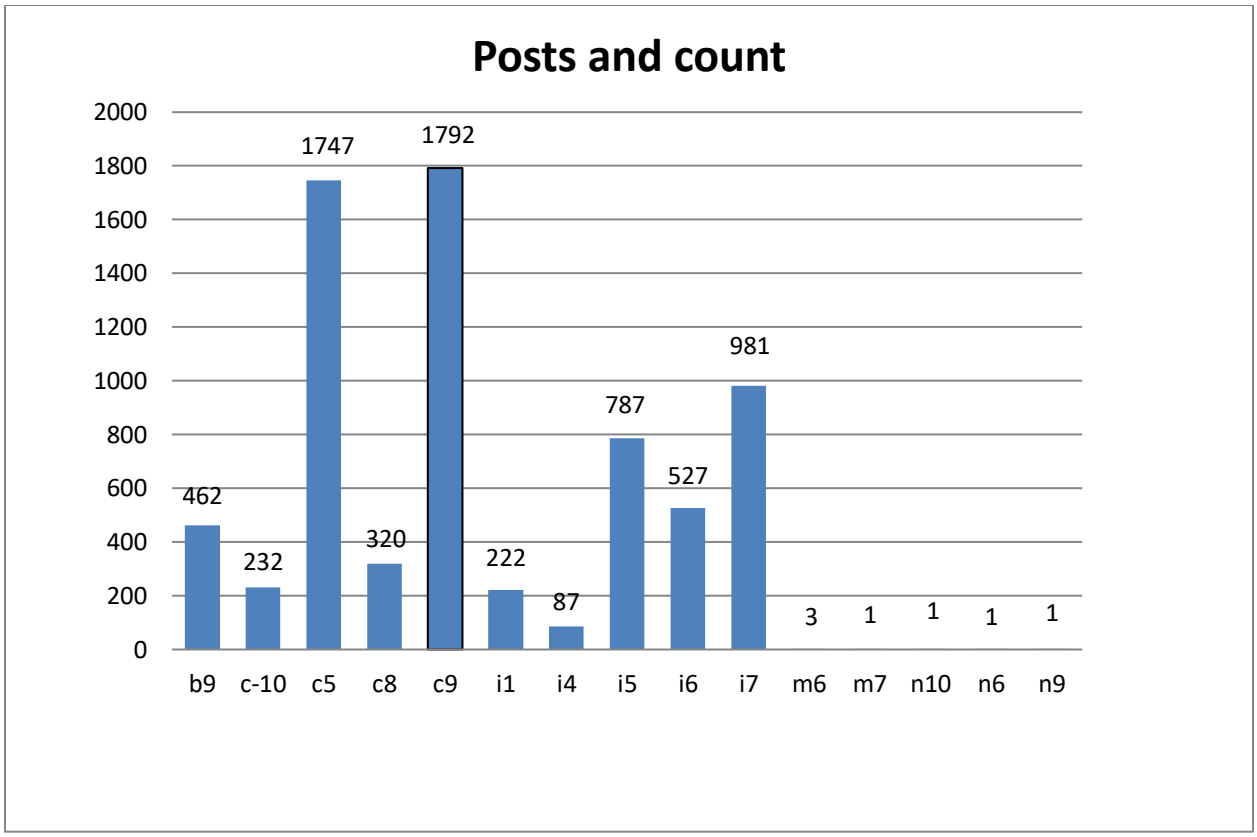
	A	B	C	D	E	F
106						
107		Column Labels				
108		Hired		Total Count of application_id	Total Count of application_id2	
109	Row Labels	Count of application_id	Count of application_id2			
110	Finance Department	176	3.75%	176	3.75%	
111	General Manageme	111	2.36%	111	2.36%	
112	Human Resource De	70	1.49%	70	1.49%	
113	Marketing Departme	202	4.30%	202	4.30%	
114	Operations Departm	1843	39.26%	1843	39.26%	
115	Production Departm	246	5.24%	246	5.24%	
116	Purchase Departme	230	4.90%	230	4.90%	
117	Sales Department	485	10.33%	485	10.33%	
118	Service Department	1331	28.36%	1331	28.36%	
119	**Grand Total**	**4694**	**100.00%**	**4694**	**100.00%**	
120						



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

This was done using pivot table, with post name in row label and count of applicant_id in values.

	Row Labels	Count of application_id2
149		
150		
151	**Row Labels**	**Count of application_id2**
152	b9	462
153	c-10	232
154	c5	1747
155	c8	320
156	c9	1792
157	i1	222
158	i4	87
159	i5	787
160	i6	527
161	i7	981
162	m6	3
163	m7	1
164	n10	1
165	n6	1
166	n9	1
167	**Grand Total**	**7164**
168		
169		


Posts and count
2000
1800
1600
1400
1200
1000
800
600
400
200
0
462
232
1747
320
1792
222
87
787
527
981
3
1
1
1
1
b9
c-10
c5
c8
c9
i1
i4
i5
i6
i7
m6
m7
n10
n6
n9