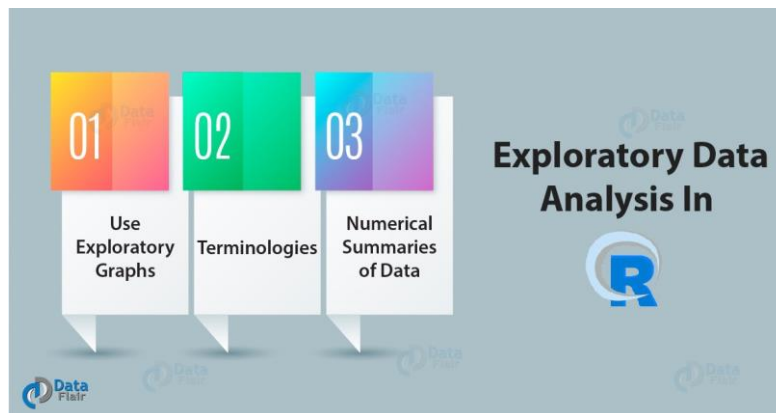


---

## BASIC STATISTICS PRACTICAL EXAM (BD1P2)

# REPORT OF EXPLORATORY DATA ANALYSIS

[RESUME NAME DATASET]



-  
PRATHIBHA.K.S

20BDA15

---

## INTRODUCTION

Employers use resumes to get a deeper understanding of candidate skills, strengths and experience. A resume should reflect achievements, awards, education, experience and any other outstanding accomplishments that align with the career path and goals.

### About the Dataset

#### a) Data description

This data set is a Cross-section data about resume, call-back and employer information for 4,870 fictitious resumes sent in response to employment advertisements in Chicago and Boston in 2001, in a randomized controlled experiment.

#### b) Data preprocessing

- Firstly, the data set was imported using `read.csv`. The original dataset is read as `Datafinal`.
- To make better analysis, the categorical variables “yes and no” were changed to numerical values of 1 and 0 respectively in excel using find and replace. This data is imported and read as `Datafinal1`.
- In order to find if there are missing values, `[which(is.na(Datafinal))]` was used. But it gave the output 0 indicating there aren't any missing values.
- By using `str(Datafinal)`, a compact structure of the dataset was obtained, which makes the further analysis simple based on if the variable is numeric or categorical.
- Other commands like `head(Datafinal)` to obtain the initial rows, `summary(Datafinal)` to produce result summaries of various model fitting functions also were used.

### Questions about the Dataset

The most striking question at the glance of the dataset was if there are missing values? On what basis are the candidates getting a call? How is each variable correlated and dependent on getting a call? Are all the variables responsible? What trends can be seen in the resumes of different candidates and how is the one who got a call superior to the one who has not got a call? In order to find answers to these questions I used the following methodologies, and thought of the possible reasons for the anomalies if any.

## METHODOLOGIES USED AND INFERENCES

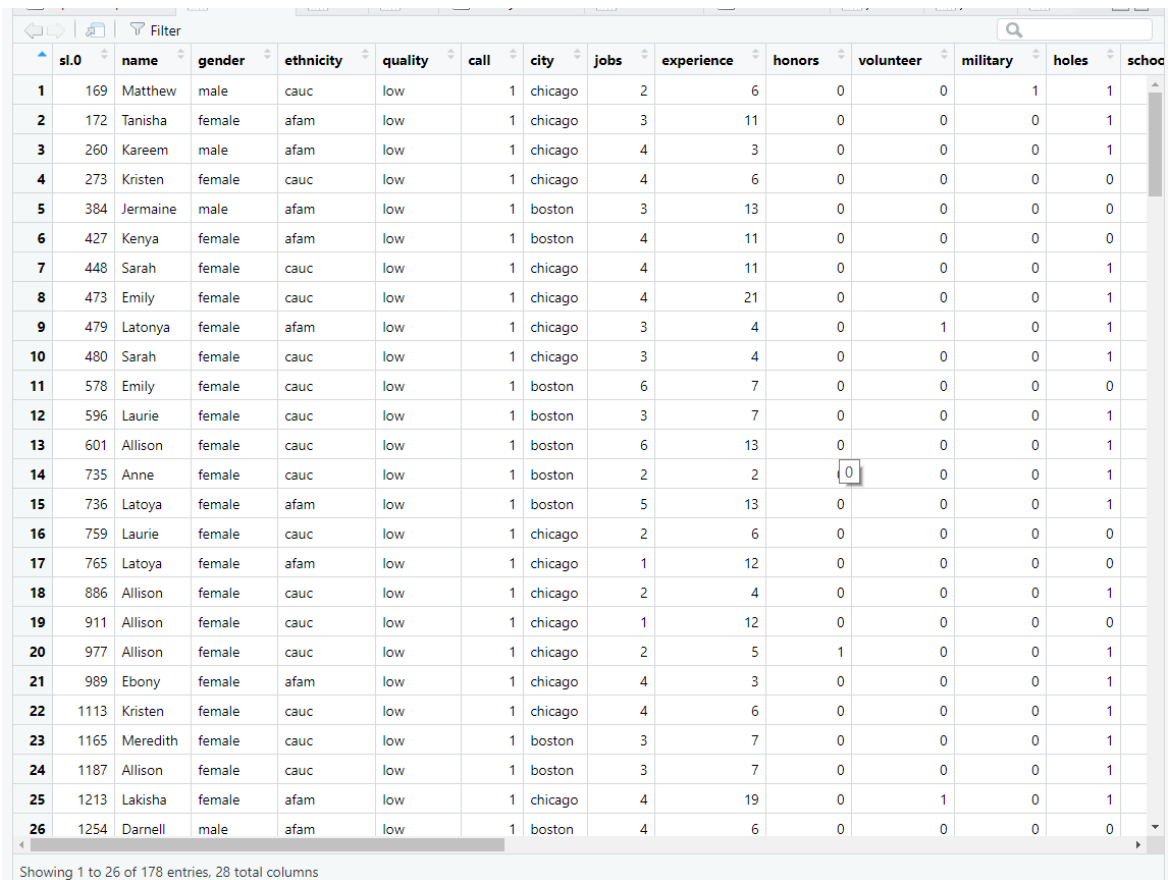
### a) Filtering using of dplyr() library

- filter()

In order to analyze the dependency of getting a call based on high- and low-quality resume, I filtered to see if there are candidates who have got a call even if they have a low-quality resume. Surprisingly, yes. There are 178 candidates with this case.

The reason for this can be found using the “equal” column which gives data about equal opportunity employment. Most of the candidates who have got a call are probably the reserved candidates.

There can be many more factors affecting this anomaly as well.



	sl.0	name	gender	ethnicity	quality	call	city	jobs	experience	honors	volunteer	military	holes	schoc
1	169	Matthew	male	cauc	low	1	chicago	2	6	0	0	1	1	
2	172	Tanisha	female	afam	low	1	chicago	3	11	0	0	0	1	
3	260	Kareem	male	afam	low	1	chicago	4	3	0	0	0	1	
4	273	Kristen	female	cauc	low	1	chicago	4	6	0	0	0	0	
5	384	Jermaine	male	afam	low	1	boston	3	13	0	0	0	0	
6	427	Kenya	female	afam	low	1	boston	4	11	0	0	0	0	
7	448	Sarah	female	cauc	low	1	chicago	4	11	0	0	0	1	
8	473	Emily	female	cauc	low	1	chicago	4	21	0	0	0	1	
9	479	Latonya	female	afam	low	1	chicago	3	4	0	1	0	1	
10	480	Sarah	female	cauc	low	1	chicago	3	4	0	0	0	1	
11	578	Emily	female	cauc	low	1	boston	6	7	0	0	0	0	
12	596	Laurie	female	cauc	low	1	boston	3	7	0	0	0	1	
13	601	Allison	female	cauc	low	1	boston	6	13	0	0	0	1	
14	735	Anne	female	cauc	low	1	boston	2	2	0	0	0	1	
15	736	Latoya	female	afam	low	1	boston	5	13	0	0	0	1	
16	759	Laurie	female	cauc	low	1	chicago	2	6	0	0	0	0	
17	765	Latoya	female	afam	low	1	chicago	1	12	0	0	0	0	
18	886	Allison	female	cauc	low	1	chicago	2	4	0	0	0	1	
19	911	Allison	female	cauc	low	1	chicago	1	12	0	0	0	0	
20	977	Allison	female	cauc	low	1	chicago	2	5	1	0	0	1	
21	989	Ebony	female	afam	low	1	chicago	4	3	0	0	0	1	
22	1113	Kristen	female	cauc	low	1	chicago	4	6	0	0	0	1	
23	1165	Meredith	female	cauc	low	1	boston	3	7	0	0	0	1	
24	1187	Allison	female	cauc	low	1	boston	3	7	0	0	0	1	
25	1213	Lakisha	female	afam	low	1	chicago	4	19	0	1	0	1	
26	1254	Darnell	male	afam	low	1	boston	4	6	0	0	0	0	

Showing 1 to 26 of 178 entries, 28 total columns

- **select()**

To understand the above anomaly, further I considered the rows “computer” that tells if the candidate has computer knowledge and “reqcomp” which tells whether the company requires computer knowledge or not. Similarly, the rows “school” and “reqeduc” which tells whether the candidate attended school and if the company needs education as a factor for employment respectively.

This was done using the select() function that helps selecting only the required columns.

From the result, we see that most of the candidates have got a call only if they have satisfied the requirement criteria. However there are contradictions as well.

[Here, 1 represents “yes” and 0 represents “no”]

	computer	reqcomp	school	reqeduc
1	1	1	0	0
2	1	1	1	0
3	1	1	1	0
4	1	1	0	0
5	1	1	1	0
6	0	0	0	0
7	1	0	1	0
8	1	1	0	0
9	1	0	0	0
10	0	0	1	0
11	1	1	0	0
12	1	1	1	0
13	1	1	0	0
14	1	1	1	0
15	0	0	0	0
16	1	0	0	0
17	0	0	1	0
18	0	0	0	0
19	0	0	1	0
20	0	0	0	0
21	1	0	1	0
22	1	0	0	0
23	0	0	0	1
24	1	1	0	1
25	1	0	0	1
26	1	1	1	1

Showing 1 to 27 of 4,870 entries, 4 total columns

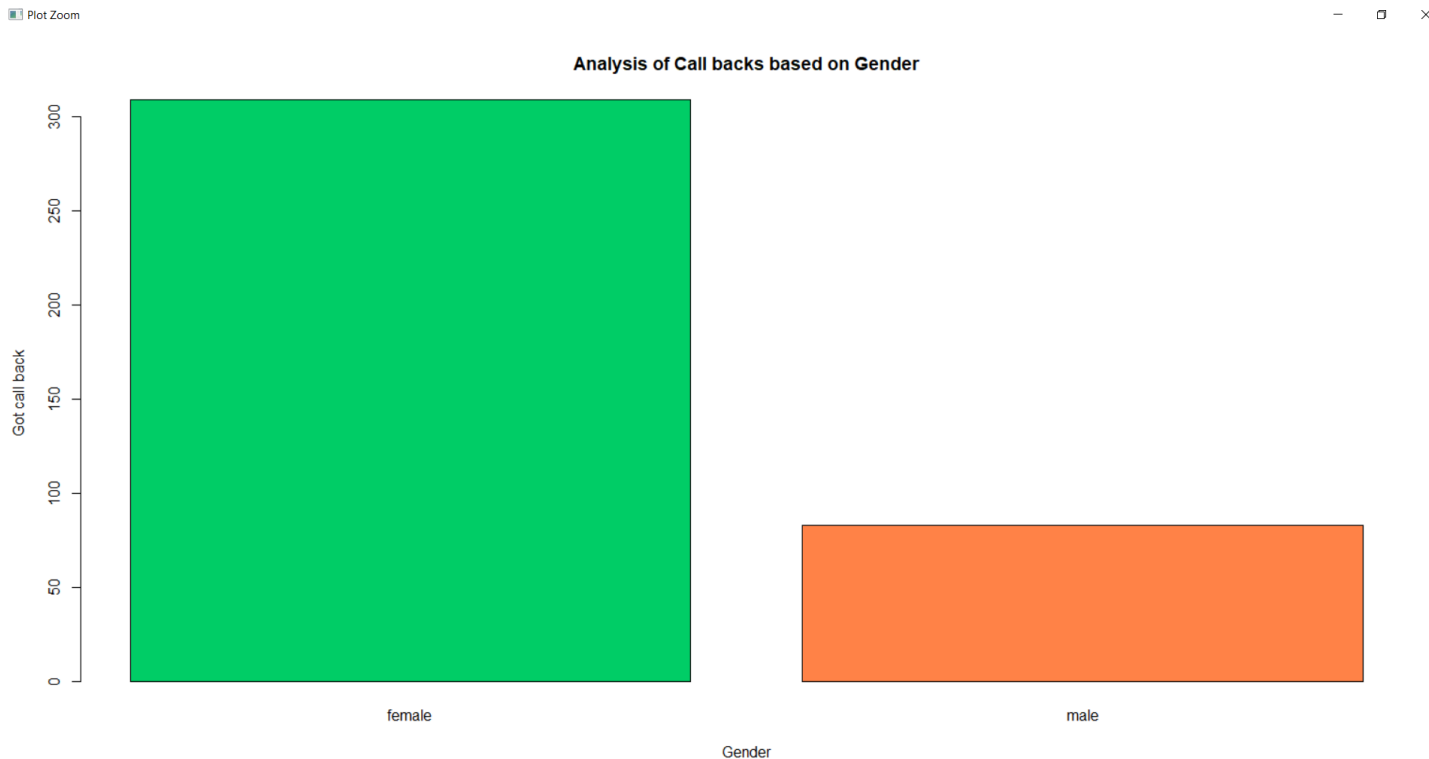
## b) Bar Graphs

A bar chart can be drawn for discrete or categorical data. For each data item, we simply draw a 'bar' showing its frequency.

- **Analysis of Call backs based on Gender**

In order to analyze if there is a bias of getting calls based on gender, a bar graph of gender and call is plotted.

One can find that the female candidates have got exclusively higher calls than male candidates. This might be for encouraging women employment, the posts can be more applicable for women, women are known for their better decision-making.

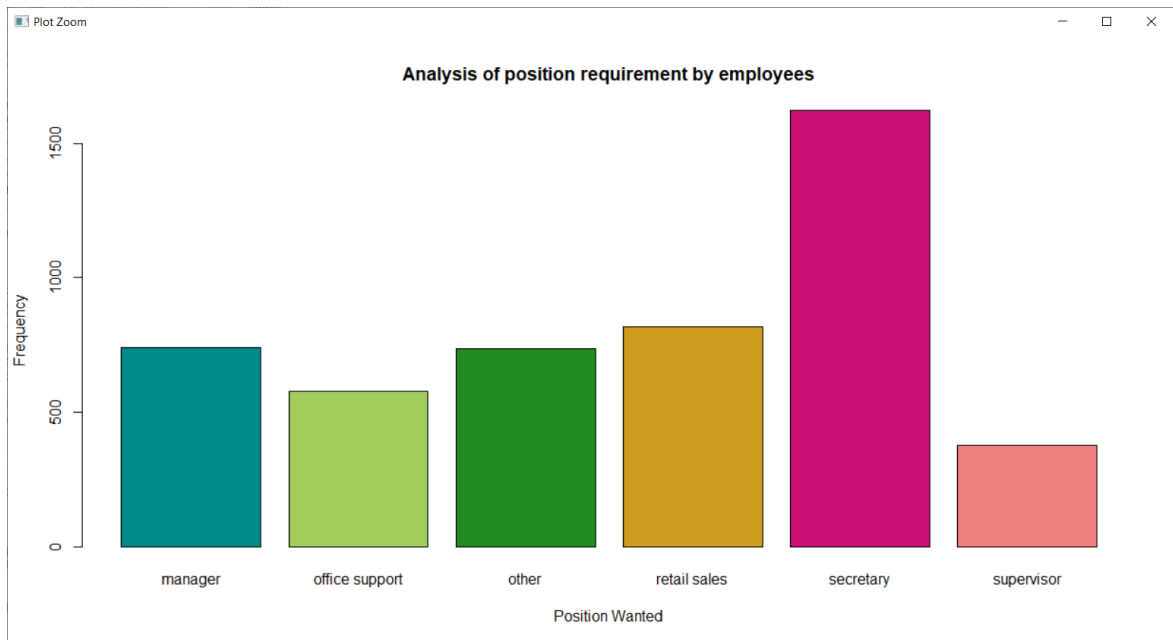


---

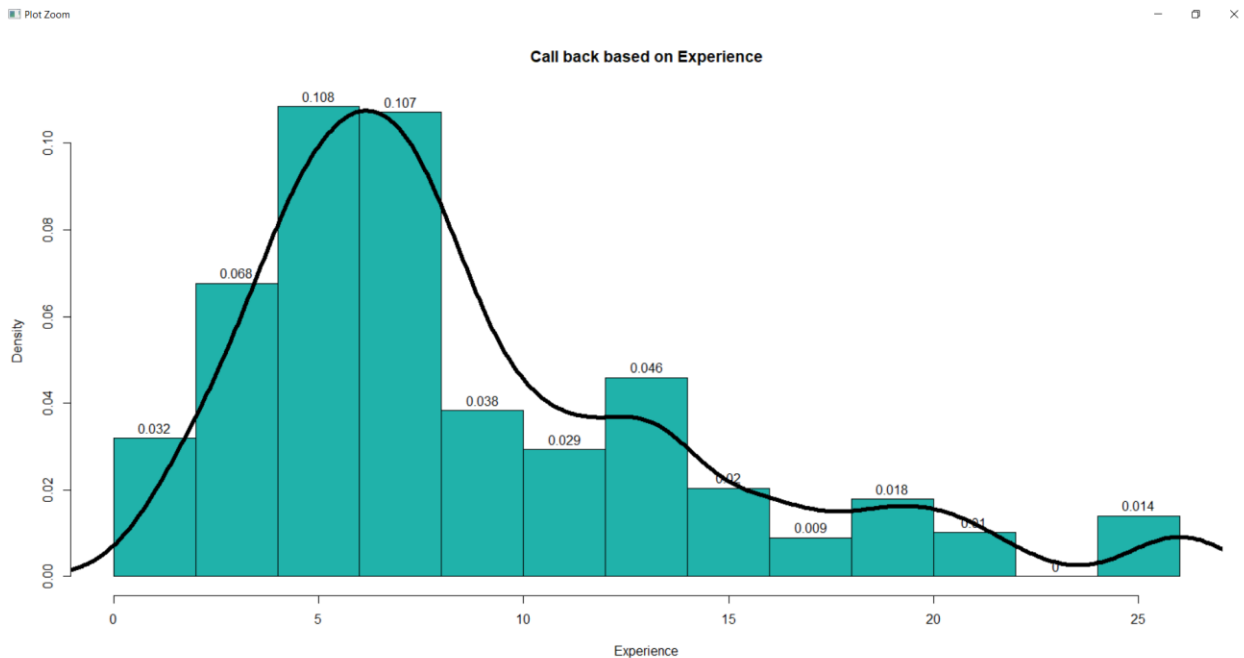
- **Analysis of position requirement by employees**

In order to analyze which post has the highest demand a bar graph of wanted and frequency is plotted.

From the bar graph of analysis of required positions, we can find out the highest applied posts by the candidates is the secretary post with a frequency higher than 1500. This may be because it is the highest paid job with lesser demands.



### c) Histogram (Analysis of Call back based on Experience)



In order to find which age group people have applied the most for the job, a histogram of experience is plotted.

The histogram is positively skewed/ right skewed as most of the data are clustered on the right side. We can see here the peak is higher in the range of 4 to 8. This shows that most of the people who have applied for a the job have an experience range of 4 to 8.

From this observation we can predict the age category of the candidates who have applied for the job. Since, most candidates have experience of 4-8 years they are probably in their 20s or early 30s.

From the previous two graphs it is seen that most of youngsters have applied for secretary post the most.

#### d) Box plot

A boxplot (also called a box and whisker plot) is another way of showing data, from which here I have found the median job and a few outliers.

The rectangle (box) in the middle represents the middle 50% of the data (between the values that are a  $\frac{1}{4}$  and  $\frac{3}{4}$  of the way through the data). The lines (whiskers) extend from the box to the smallest and largest values. The plot also shows the middle value (called the median).

Here, the median is approximately around 4.2, the smallest and the largest values being approximately 1.8 and 5.2 respectively. However, there are also blue diamond markers that represent the outliers above and below the end of the whisker with jobs of 6,7 and 1.





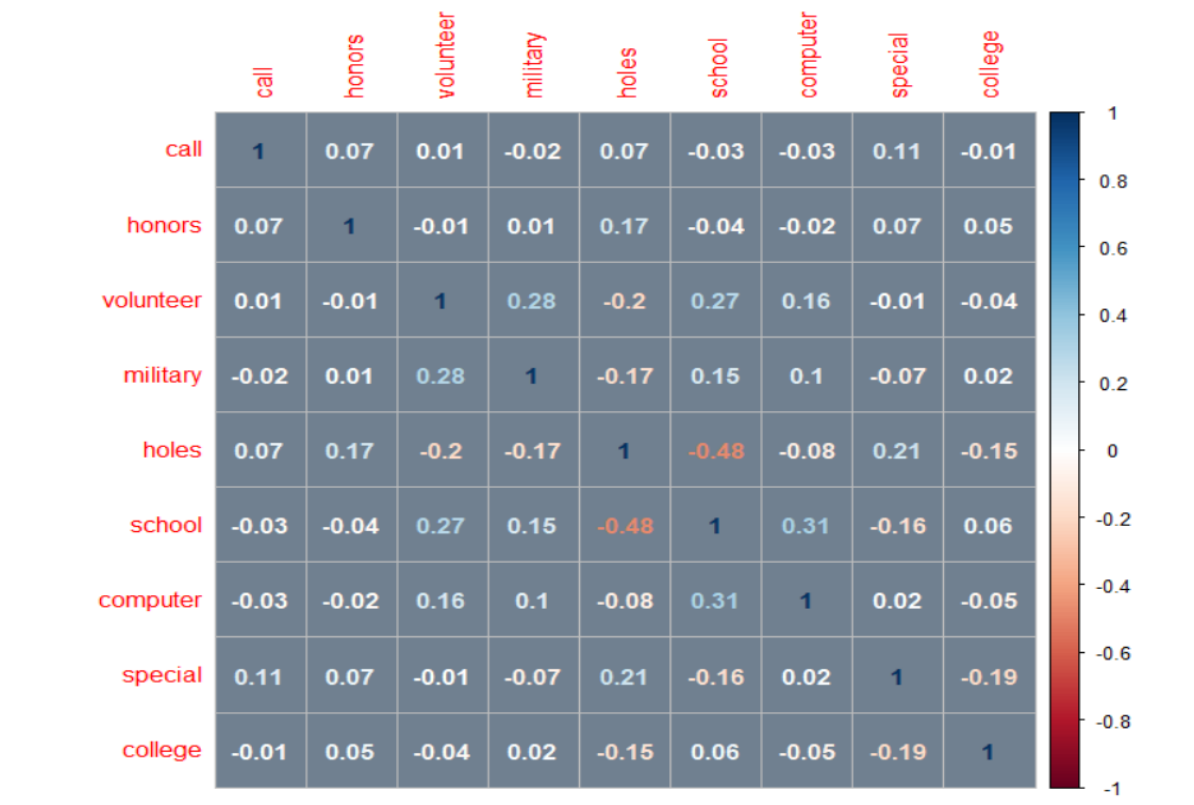
### e) Correlation matrix

In order to analyze how different variables are correlated to each other and with getting a call, a correlation matrix was plotted. Correlation is a statistical measure that expresses the extent to which two variables change together at a constant rate.

- **Pearson's correlation**

Correlation coefficients are used to determine how strong a relationship is between two variables, and a table indicating the same is called the correlation matrix.

The correlation coefficient can take values from -1 to +1. The closer the value is towards 1, the more correlated it is.



From the above result, one can see that if a candidate has any special skill, he is more likely to get a call, which is followed by honors and holes. Hence, by just looking at the correlation matrix, we can see which variables are responsible the most for whether a candidate will get a call or not.

---

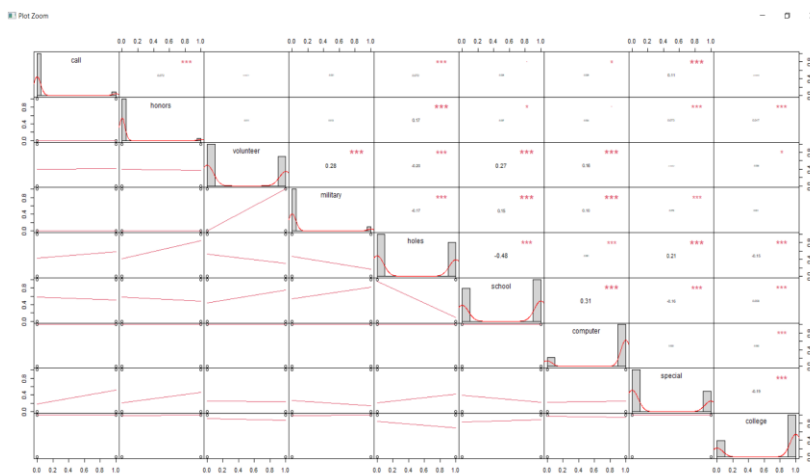
- **Correlation scatterplots**

- The distribution of each variable is shown on the diagonal.
- On the bottom of the diagonal: the bivariate scatter plots with a fitted line are displayed
- On the top of the diagonal: the value of the correlation is displayed.

The primary function of a scatter plot is to visualize the strength of correlation between the two plotted variables.

- Two variables with a strong correlation will appear as a number of points occurring in a clear and recognizable linear pattern. The line does not need to be straight, but it should be consistent and not exactly horizontal or vertical.
- Two variables with a weak correlation will appear as a much more scattered field of points, with only a little indication of points falling into a line of any sort.

Here, the more linear the scatterplots lines are, the highest is the correlation between the variables and hence can find which variable is responsible the most for getting a call.



One can see that if a candidate has any special skill, he is more likely to get a call, as the plot of special skill is the most linear .

## f) Chi-square

From the correlation matrix we found how correlated/dependent the numerical variables(yes=1,no=0) were. In order to find whether the categorical variables are dependent on getting a call back, chi-squared test is performed.

Chi-squared test in R can be used to test if two categorical variables are dependent, by means of a contingency table.

```
> table(Datafinal1$call,Datafinal1$ethnicity)
      afam cauc
no    2278 2200
yes   157  235
> ftable(Datafinal1[c("call","experience","jobs")])
```

		jobs							
		1	2	3	4	5	6	7	
call	no	experience							
	1	4	29	9	0	0	0	0	
	2	0	40	53	237	0	0	0	
	3	5	40	109	25	6	0	0	
	4	17	86	237	90	41	22	0	
	5	28	76	199	113	38	3	14	
	6	4	159	143	447	14	1	0	
	7	1	35	325	71	33	34	0	
	8	0	14	73	149	54	246	0	
	9	8	13	32	40	26	25	0	
	10	1	41	35	29	7	2	0	
	11	16	10	21	98	9	1	0	
	12	8	23	16	10	7	0	0	
	13	2	7	26	17	75	1	0	
	14	0	29	8	69	33	0	0	
	15	0	15	3	12	3	0	0	
	16	0	0	21	29	2	27	0	
	17	0	0	1	2	0	0	0	
	18	0	0	1	2	2	62	3	
	19	0	12	3	3	21	0	0	
	20	0	0	2	21	1	4	0	
yes	1	1	2	0	0	0	0	0	
	2	0	8	3	11	0	0	0	
	3	0	2	5	2	0	0	0	
	4	0	8	24	6	4	2	0	
	5	8	9	7	6	3	1	2	
	6	1	22	5	18	3	0	0	
	7	0	3	27	4	2	6	0	
	8	1	4	9	18	3	7	0	
	9	1	0	1	2	5	6	0	

---

There are two ways to tell if they are independent:

- **By looking at the p-Value:** If the p-Value is less than 0.05, we fail to reject the null hypothesis that the x and y are independent. So for the example output above, (p-Value=2.954e-07), we reject the null hypothesis and conclude that x and y are not independent.
- **From Chi.sq value:** For 2 x 2 contingency tables with 2 degrees of freedom (d.o.f), if the Chi-Squared calculated is greater than 3.841 (critical value), we reject the null hypothesis that the variables are independent.

```
> chisq.test(table(Datafinal1$call,Datafinal1$ethnicity), correct = FALSE)

Pearson's Chi-squared test

data:  table(Datafinal1$call, Datafinal1$ethnicity)
X-squared = 16.879, df = 1, p-value = 3.984e-05

> summary(table(Datafinal1$call,Datafinal1$ethnicity))
Number of cases in table: 4870
Number of factors: 2
Test for independence of all factors:
  Chisq = 16.879, df = 1, p-value = 3.984e-05
>
> #chi-square test for city and call
> chisq.test(table(Datafinal1$call,Datafinal1$city), correct = FALSE)

Pearson's Chi-squared test

data:  table(Datafinal1$call, Datafinal1$city)
X-squared = 14.28, df = 1, p-value = 0.0001575

> summary(table(Datafinal1$call,Datafinal1$city))
Number of cases in table: 4870
Number of factors: 2
Test for independence of all factors:
  Chisq = 14.28, df = 1, p-value = 0.0001575
> |
```

By looking into the results of the chi-square test, we can find that both ethnicity and city are dependent on whether a candidate gets a call or not as they diverge from the critical values.

---

## CONCLUSION

Performing “Exploratory Data analysis using R” on the resume dataset using libraries like `dplyr()`, `tidyr()`, `ggplot2()`, `corrplot()`, `Hmisc()`, `PerformanceAnalytic()` , various statistical and visualization methods, we have drawn numerous conclusions and inferences.

The baseline of all the analysis techniques is whether a candidate who has applied for a particular job gets a call back or not. Visualization techniques like bar plots, histogram, box plot, scatterplot and statistical tools like chi-square test and correlation were used, which helped in finding which variables are highly responsible for getting a call back for the interview.

## FUTURE PROSPECTS:

The future prospect of this analysis can be to design a predictive model that helps predict if a candidate has to be called for the interview or not by considering all inferences drawn from Exploratory data analysis of the Resume names dataset.

---

## APPENDIX

```
library(dplyr)
library(tidyr)
library(tidyverse)
library(ggplot2)
library(corrplot)
library(Hmisc)
library(PerformanceAnalytics)

#-----import dataset-----

Datafinal=read.csv("C:/Users/prathibha k s/OneDrive/Desktop/report 1p2/ResumeNames.csv")
Datafinal1=read.csv("C:/Users/prathibha k s/OneDrive/Desktop/report
1p2/ResumeNames1.csv")

#-----pre-requisite analysis-----

#to find if there are missing values
which(is.na(Datafinal))

#finding the summary of the dataset
summary(Datafinal)

#to visualize the 1st 5 rows of the dataset
head(Datafinal)

#to find the datatypes of different variables
str(Datafinal)

# -----Analysis-----
#-----Data wrangling with dplyr()-----

# checking if any person with low quality got call
callback <- Datafinal %>% filter(call==1,quality=='high')
callback1 <- Datafinal %>% filter(call==1,quality=='low')

# checking the number of males and female who got a call
call<-Datafinal %>% filter(call==1)
female<-Datafinal %>% filter(call==1,gender=='female')
male<-callback %>% filter(call==1,gender=='male')

# analysis of resume and requirements
resreq<- call %>% select(computer,reqcomp,school,reqeduc,)
resreq
```

---

#-----Bar Graphs-----

#plotting a bar graph to visualize the number of males and female who got a call

```
bar <- table(call$gender)
barplot(bar,
        xlab= "Gender",
        ylab= "Got call back",
        main= "Analysis of Call backs based on Gender",
        col= c("female" = "springgreen3", "male" = "sienna1"),
        border = "black")
```

#bar plot to analyze position requirement by employees

```
bar <- table(Datafinal$wanted)
barplot(bar,
        xlab= "Position Wanted",
        ylab= "Frequency",
        main= "Analysis of position requirement by employees",
        col= c("darkcyan", "darkolivegreen3", "forestgreen", "goldenrod3", "deeppink3", "lightcoral"),
        border = "black")
```

#-----Histogram-----

#histogram to visualize the number of call

```
hist(call$experience,
     xlab="Experience",
     main="Call back based on Experience",
     breaks = 12,
     col="lightseagreen",
     border="black",
     labels=T,
     prob=T)
lines(density(call$experience),col="black",lwd=5)
```

#-----Box Plot-----

#box plot

```
boxplot(Datafinal$jobs,data=Datafinal,
        xlab="Jobs",
        main="Boxplot of jobs",
        cex=1.5, #magnification
        pch=18, #Try like 3 different from 1 to 20 numbers for this argument.
        col="lightpink2",
        border="darkblue")
```

---

```
#-----Correlation matrix-----
```

```
#pearson's R
```

```
mydata <- Datafinal %>%  
select(call,honors,volunteer,military,holes,school,computer,special,college)  
mydata.rcorr = rcorr(as.matrix(mydata))  
mydata.coeff = mydata.rcorr$r  
corrplot(mydata.coeff,method = "number", bg="slategray")
```

```
#correlation scatterplot
```

```
chart.Correlation(mydata, histogram=TRUE, pch=19)
```

```
#-----Chi square test-----
```

```
#chi-square test
```

```
table(Datafinal1$call,Datafinal1$ethnicity)  
ftable(Datafinal1[c("call","experience","jobs")])
```

```
#chi-square test for ethnicity and call
```

```
chisq.test(table(Datafinal1$call,Datafinal1$ethnicity), correct = FALSE)  
summary(table(Datafinal1$call,Datafinal1$ethnicity))
```

```
#chi-square test for city and call
```

```
chisq.test(table(Datafinal1$call,Datafinal1$city), correct = FALSE)  
summary(table(Datafinal1$call,Datafinal1$city))
```