SHRADDHA ARORA

1910110375

CSD 350

PRESCRIBING DRUGS USING CONSUMER REVIEWS

Introduction

Medication and drug analysis involve credible research and working as it plays the bluff of life and death. In a matter of every few years, there needs to be a revised set of observations on the trends followed by the pharmaceutical industry as to which drugs have proved to be beneficial and which have not.

NLP comes into play here as it helps us to analyze how consumers feel about the product and how they react to it.

Medicine is a very important industry, and its decisions need to be made very carefully, hence, this analysis is extremely important and unavoidable.

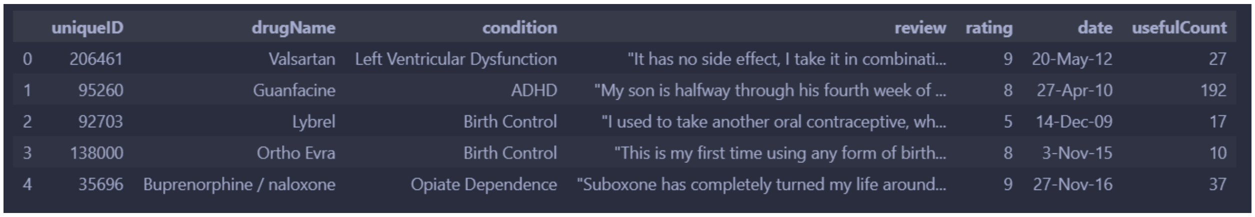
Procedure

IMPORTING THE LIBRARIES:

We start off by importing all necessary libraries including numpy, pandas, string, matplotlib, nltk, ipywidgets etc.

READING THE DATASET

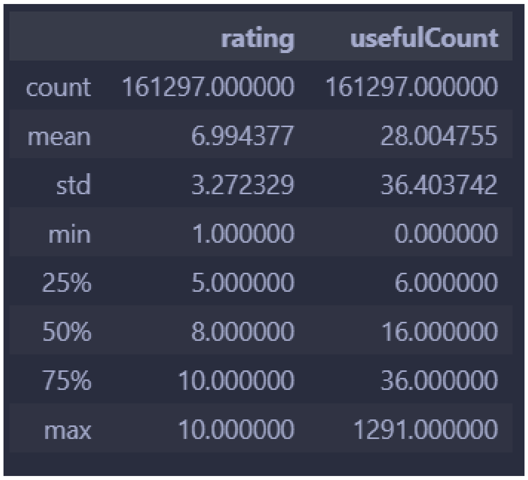
Next, we proceed to reading the dataset and checking the shape of it. Here, we have 161297 entries under 7 categories.



	uniqueID	drugName	condition	review	rating	date	usefulCount
0	206461	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9	20-May-12	27
1	95260	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8	27-Apr-10	192
2	92703	Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5	14-Dec-09	17
3	138000	Ortho Evra	Birth Control	"This is my first time using any form of birth...	8	3-Nov-15	10
4	35696	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has completely turned my life around...	9	27-Nov-16	37

SUMMARISING THE DATASET

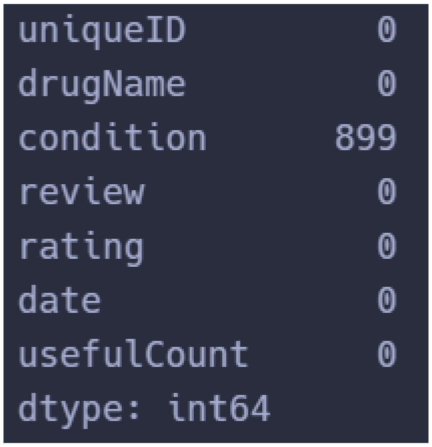
We start by fetching the ratings and useful counts of the drugs.



	rating	usefulCount
count	161297.000000	161297.000000
mean	6.994377	28.004755
std	3.272329	36.403742
min	1.000000	0.000000
25%	5.000000	6.000000
50%	8.000000	16.000000
75%	10.000000	36.000000
max	10.000000	1291.000000

Next, we find out the details of drugs with extreme degrees of use, i.e, useful drugs (with useful count > 1000) and useless drugs (with useful count = 0).

We then check for any null values in the dataset, and find that some reviews have null entries for medical condition records.

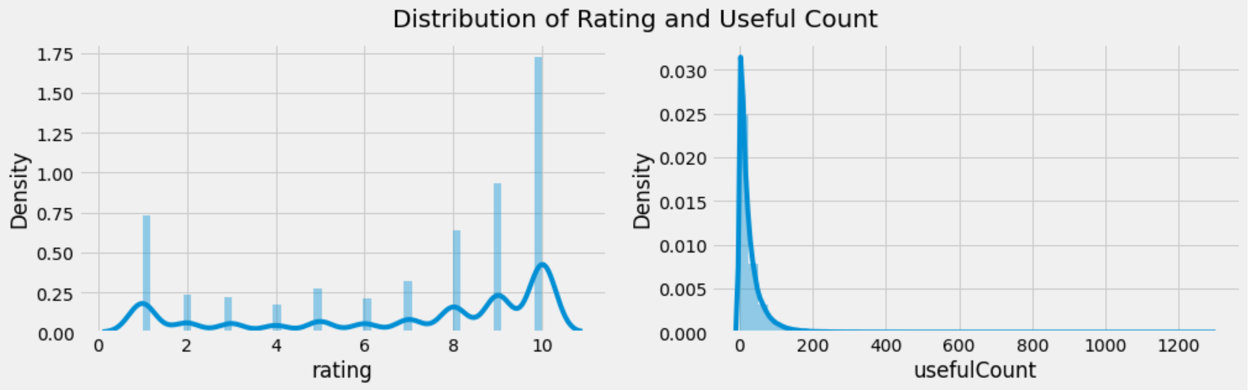


```
uniqueID         0
drugName         0
condition      899
review           0
rating           0
date             0
usefulCount      0
dtype: int64
```

Since knowing about the medical condition is a must for registering any drug review, we remove all the entries which have null values for medical condition.

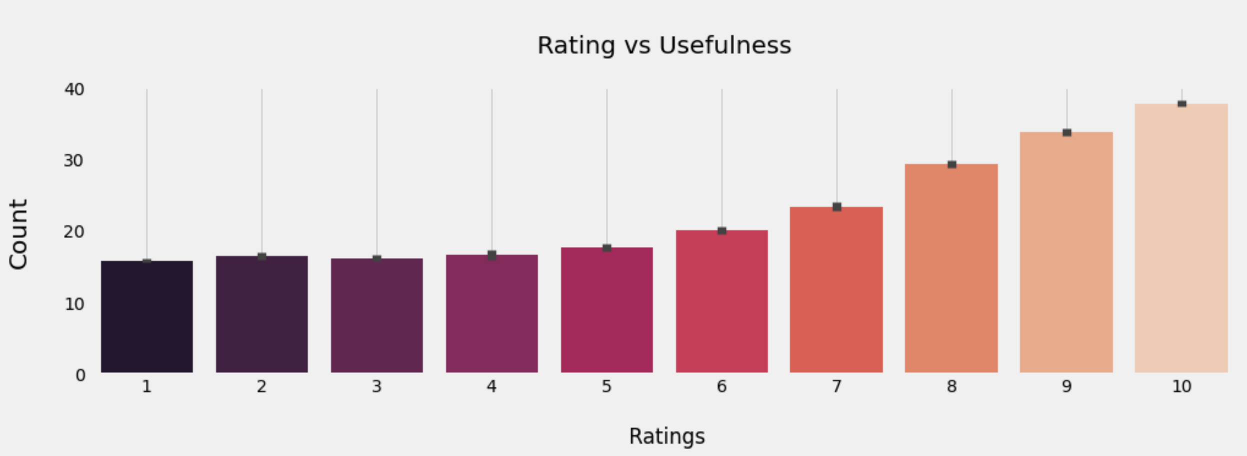
ANALYSING HIDDEN PATTERNS IN THE DATASET

To analyse certain hidden trends in the dataset, we plot the distributions of Rating and Useful Count graphically.

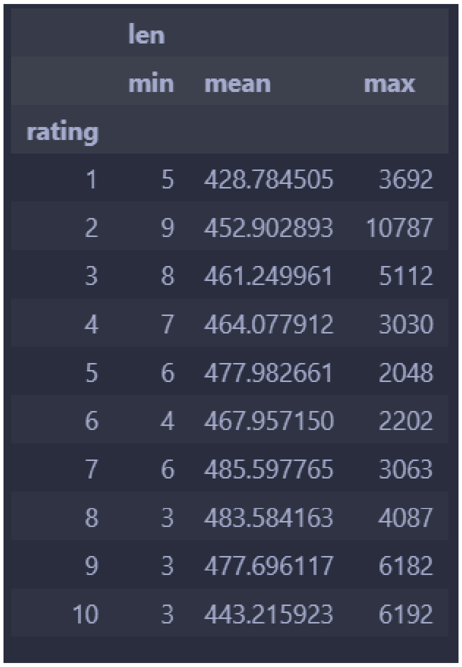


Most of the drugs present in the dataset have useful count < 200.

Next, we check the relationship between useful count and ratings by plotting a bar graph for the same. We observe a positive linear relationship between the two entities.



We also check for any relationship between length of reviews and their ratings but no such relationship is found here. For all ratings, the average length of reviews is similar.



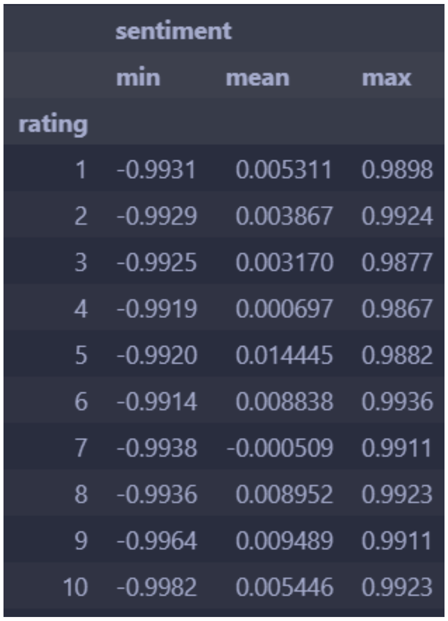
	len		
	min	mean	max
rating			
1	5	428.784505	3692
2	9	452.902893	10787
3	8	461.249961	5112
4	7	464.077912	3030
5	6	477.982661	2048
6	4	467.957150	2202
7	6	485.597765	3063
8	3	483.584163	4087
9	3	477.696117	6182
10	3	443.215923	6192

CLEANING THE REVIEWS

Further, we proceed to clean the data as much as possible, starting with removing punctuations, followed by removing stop words and numbers since they hold no sentimental value.

CALCULATING SENTIMENT FROM THE REVIEWS

We use the Vader Lexicon class in the NLTK Library to carry out the same. After the sentiment scores are calculated, we check their impact on the reviews. We find no clear pattern for the same and hence, conclude that the sentiment scores are meaningless for this analysis.

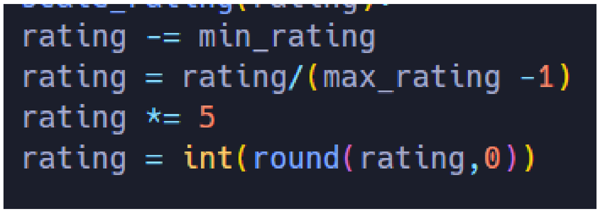


	sentiment		
	min	mean	max
rating			
1	-0.9931	0.005311	0.9898
2	-0.9929	0.003867	0.9924
3	-0.9925	0.003170	0.9877
4	-0.9919	0.000697	0.9867
5	-0.9920	0.014445	0.9882
6	-0.9914	0.008838	0.9936
7	-0.9938	-0.000509	0.9911
8	-0.9936	0.008952	0.9923
9	-0.9964	0.009489	0.9911
10	-0.9982	0.005446	0.9923

CALCULATING EFFECTIVENESS AND USEFULNESS OF DRUGS

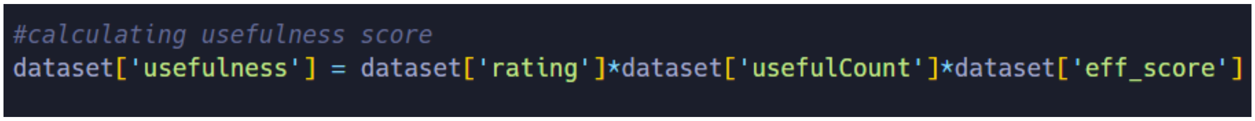
Since sentiment scores proved to be of no use, we calculated the effectiveness and usefulness scores for these drugs in order to prescribe them based on consumer reviews.

Formula used for calculating Effective Rating:



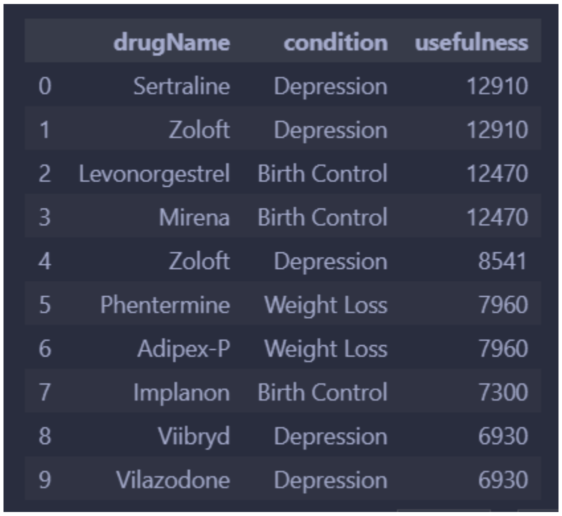
```
rating -= min_rating
rating = rating/(max_rating -1)
rating *= 5
rating = int(round(rating,0))
```

Formula used for calculating Usefulness Score:



```
#calculating usefulness score
dataset['usefulness'] = dataset['rating']*dataset['usefulCount']*dataset['eff_score']
```

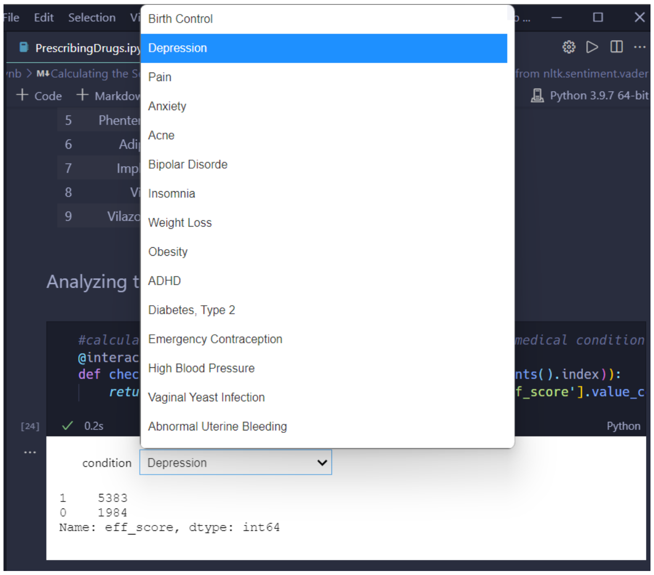
Based on this Usefulness score calculated, we get a list of the Top 10 drugs, with the medical condition they are associated with.



	drugName	condition	usefulness
0	Sertraline	Depression	12910
1	Zoloft	Depression	12910
2	Levonorgestrel	Birth Control	12470
3	Mirena	Birth Control	12470
4	Zoloft	Depression	8541
5	Phentermine	Weight Loss	7960
6	Adipex-P	Weight Loss	7960
7	Implanon	Birth Control	7300
8	Viibryd	Depression	6930
9	Vilazodone	Depression	6930

ANALYSING THE MEDICAL CONDITIONS

Based on the effective scores calculated, we get the number of useful and useless drugs for each medical condition (which can be selected from the drop down list provided).



```
5    Phente
6       Adi
7       Imp
8         V
9     Vilazo
```

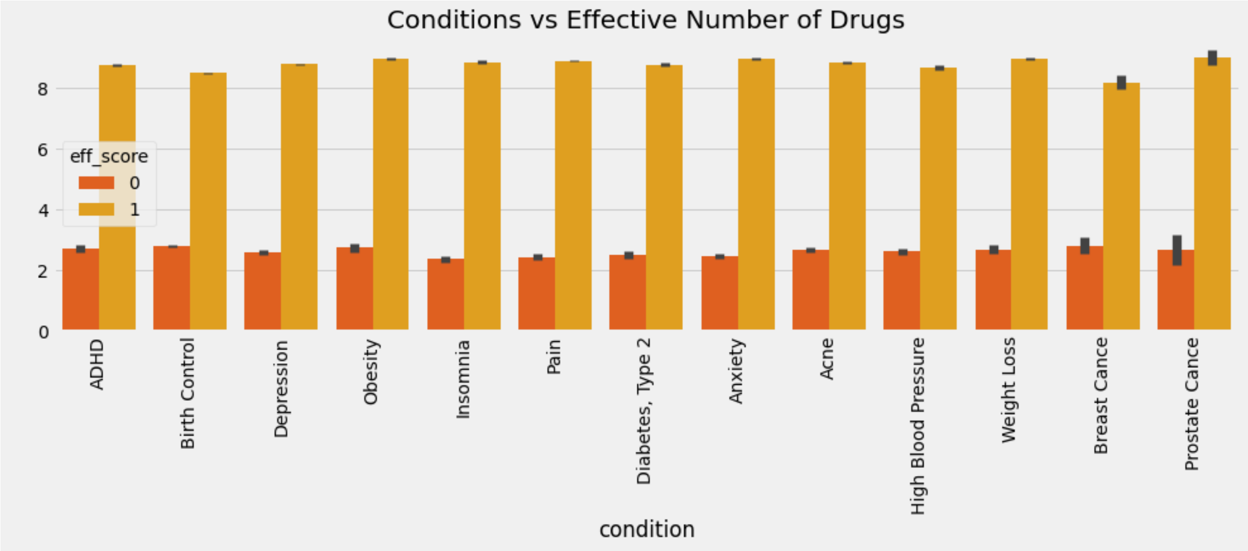
```
#calcula                                    medical condition
@interac
def chec                                    nts().index)):
    retu                                    f_score'].value_c
```

condition: Depression

```
1    5383
0    1984
Name: eff_score, dtype: int64
```

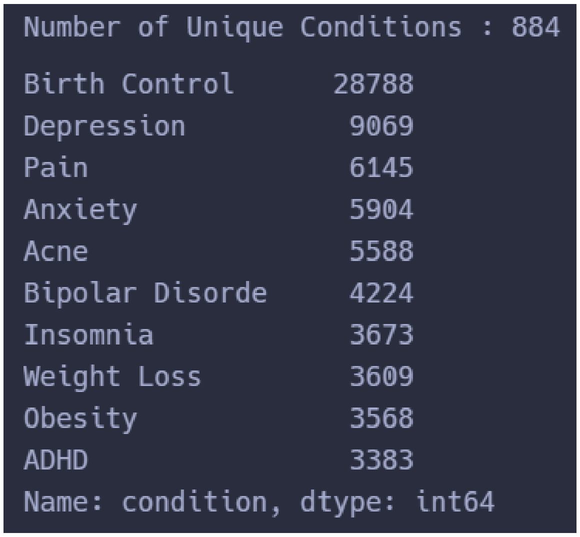
We plot the same on a graph for some common medical conditions.

Note: We could not have plotted this graph for all conditions since there are 800+ medical conditions mentioned in the dataset.



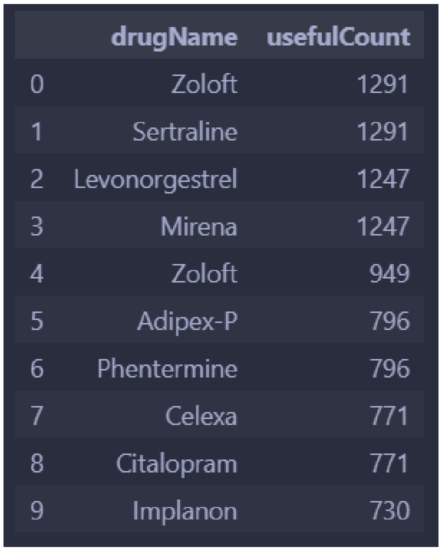
Looking at the pattern in the graph, it is evident that **about 30% drugs are useless for most of the medical conditions.**

Next, we find out the list of most common medical conditions amongst patients:



```
Number of Unique Conditions : 884
Birth Control      28788
Depression          9069
Pain                6145
Anxiety             5904
Acne                5588
Bipolar Disorde     4224
Insomnia            3673
Weight Loss         3609
Obesity             3568
ADHD                3383
Name: condition, dtype: int64
```

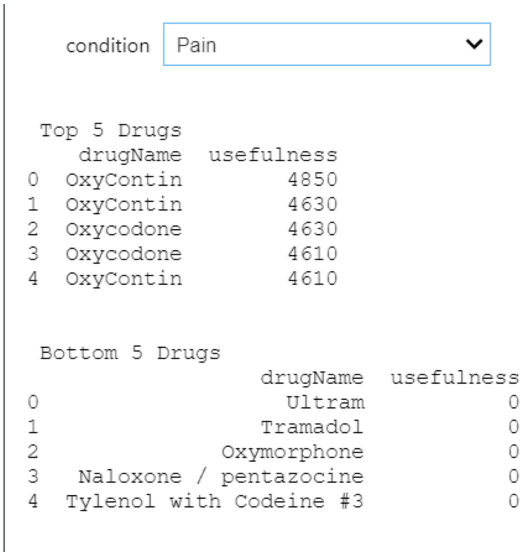
Furthermore, we fetch the list of Top 10 drugs which were helpful to the highest number of people:



	drugName	usefulCount
0	Zoloft	1291
1	Sertraline	1291
2	Levonorgestrel	1247
3	Mirena	1247
4	Zoloft	949
5	Adipex-P	796
6	Phentermine	796
7	Celexa	771
8	Citalopram	771
9	Implanon	730

FINDING MOST USEFUL AND USELESS DRUGS ASSOCIATED TO EACH MEDICAL CONDITION

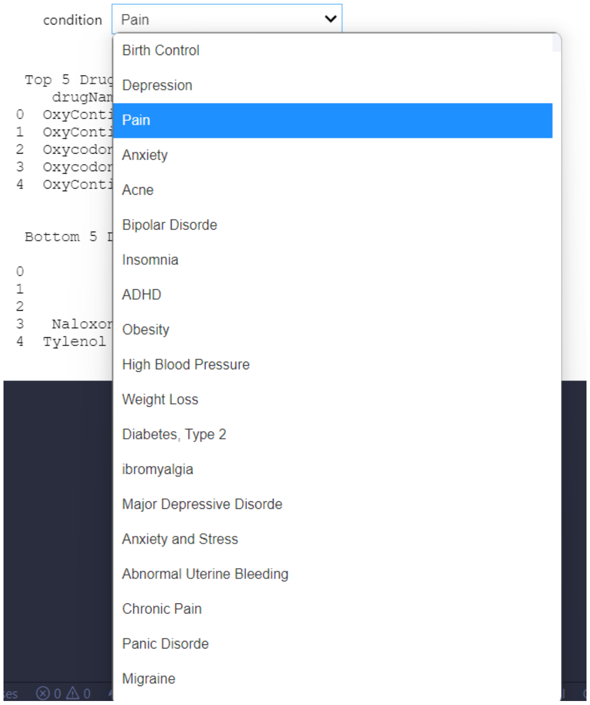
We conclude the project by finally fetching the highest and lowest rated drugs (Top 5) for each medical condition (which can be opted for from the dropdown list).



condition: Pain

```
 Top 5 Drugs
    drugName  usefulness
0  OxyContin        4850
1  OxyContin        4630
2  Oxycodone        4630
3  Oxycodone        4610
4  OxyContin        4610


 Bottom 5 Drugs
                   drugName  usefulness
0                    Ultram           0
1                  Tramadol           0
2                Oxymorphone          0
3   Naloxone / pentazocine            0
4  Tylenol with Codeine #3            0
```



condition: Pain

```
 Top 5 Drug
   drugNam
0  OxyConti
1  OxyConti
2  Oxycodor
3  Oxycodor
4  OxyConti

 Bottom 5 I

0
1
2
3   Naloxor
4  Tylenol
```

Conclusion

This project was solely based on how the pharmaceutical industry reviews drugs based on consumer reviews using NLP and how it can change the course of medication. Further, I intend on trying to convert this project into a predictive model by training a Machine Learning model to predict the usefulness of a drug by looking at the consumer’s review.

Acknowledgement

I would like to thank Professor Ketan Bajaj, for extending us this opportunity to work on such projects, which help us widen our horizons of perspective and knowledge. Further, his guidance and motivation have provided me with the diligence to work on this project with utmost sincerity.