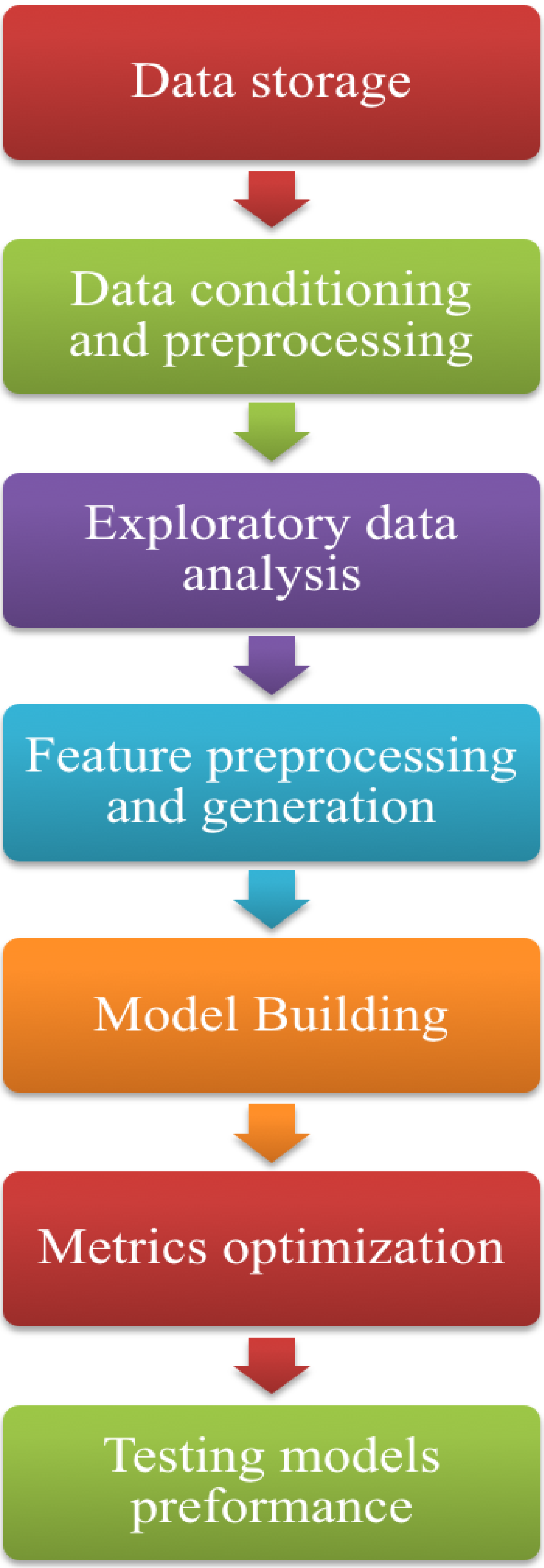
Suprath Technologies Data Analysis Task Report

DAINT1186 - Vipin C
vipincsekar98@gmail.com

Dataset given : Customer review dataset
Rows : 130
Columns : 8 (attributes)

Subject - Subject of the review.
Customer - Name of the Customer.
Location - Location of the customer from where the review is being given.
Rating - Rating of the product which ranges from 1 to 5
Data_time - Date and time of the review
Views - Number of Views for the review given
Total_reviews_by_customer - Total number of reviews given by that customer.
Customer_followers- Number of customers who have followed that customer who have given the review.
Complaint - Complaint entered by that customer.

Workflow :



Data storage

Data conditioning and preprocessing

Exploratory data analysis

Feature preprocessing and generation

Model Building

Metrics optimization

Testing models preformance

Data Storage :

Google Cloud Storage is used.

Data preprocessing performed :

1.Converted “Views” column to numerical data.

2.Converted “Total_reviews_by_customer” column to numerical data.

3.Converted “Customer_follower” column to numerical data.(”NULL” value is converted to 0)

4.There are two rows in the column “Views” which are non-numeric - “READ”. So those two rows have been deleted for further analysis. As the number of erroneous data is only 2, so we have deleted the rows,if the number is high we can replace that rows values to either “Mean”,”Median” or “Mode” for filling those rows.

Steps performed :

1.Used “colab” by Google for reading the dataset.

2.Date and time variable in the dataset cannot be used directly for any Inferences.So it is converted to a format that can provide more information like the ones given below.

'Year', 'Month', 'Week', 'Day', 'Dayofweek', 'Dayofyear','Is_month_end', 'Is_month_start', 'Is_quarter_end', 'Is_quarter_start', 'Is_year_end', 'Is_year_start']

Tool used for : Pandas

For example, the inference that can be found out using “Is_month_end” can lead to framing a conclusion of the mental state of the customer in the month_end.

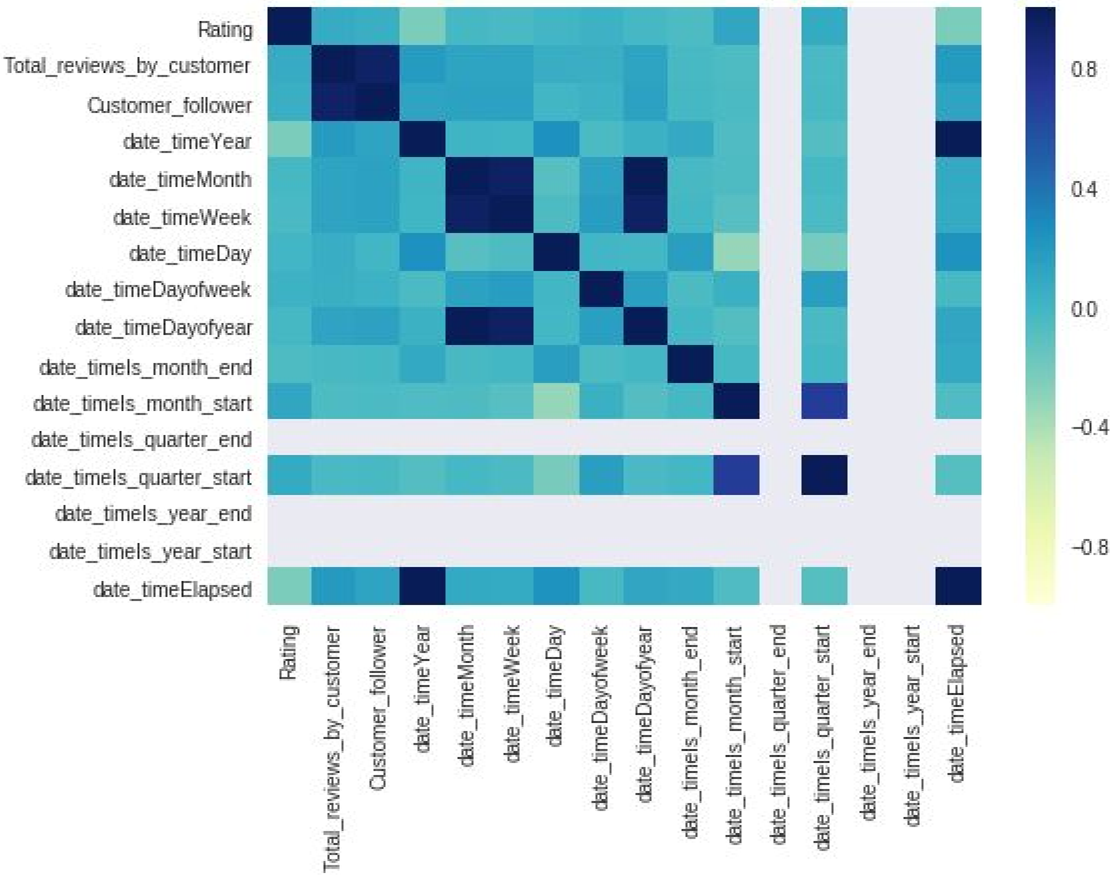
Following are the correlation between the numerical data.

```
Rating                        1.000000
date_timeIs_month_start       0.112877
date_timeIs_quarter_start     0.079190
Total_reviews_by_customer     0.072051
Customer_follower             0.048019
date_timeDayofweek            0.031087
date_timeDay                 -0.002436
date_timeMonth               -0.022219
date_timeDayofyear           -0.022562
date_timeWeek                -0.038689
date_timeIs_month_end        -0.048806
date_timeYear                -0.237371
date_timeElapsed             -0.237778
date_timeIs_quarter_end            NaN
date_timeIs_year_end               NaN
```

```
date_timeIs_year_start           NaN
Name: Rating, dtype: float64
```

Inference from the above table : There is high correlation between “Rating” and “month_start” which says that there as more customers who rate during the start of the month.

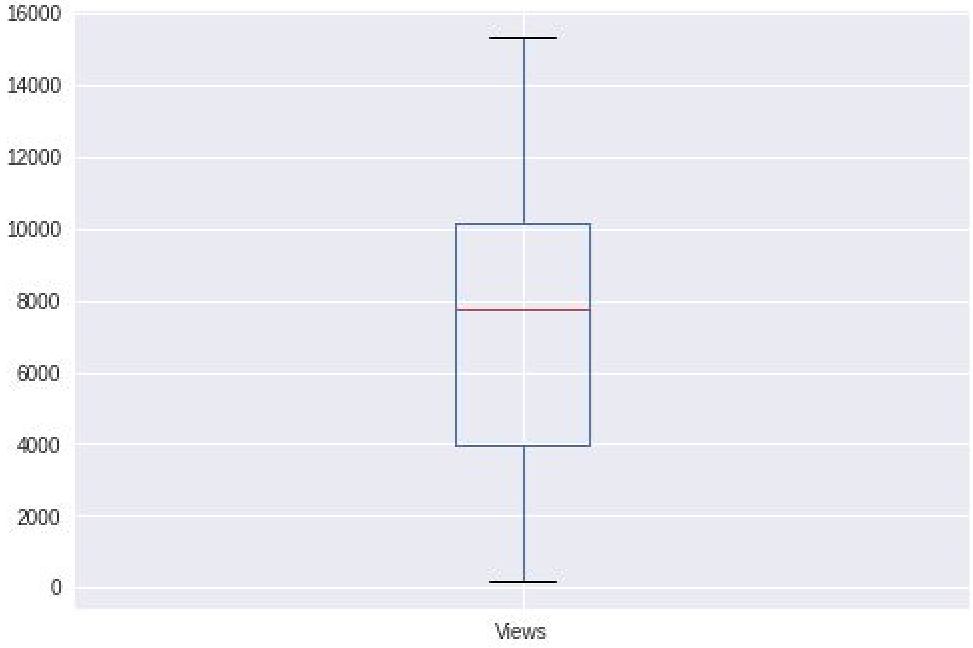
4.Heatmap is drawn for the processed data(All attributes)

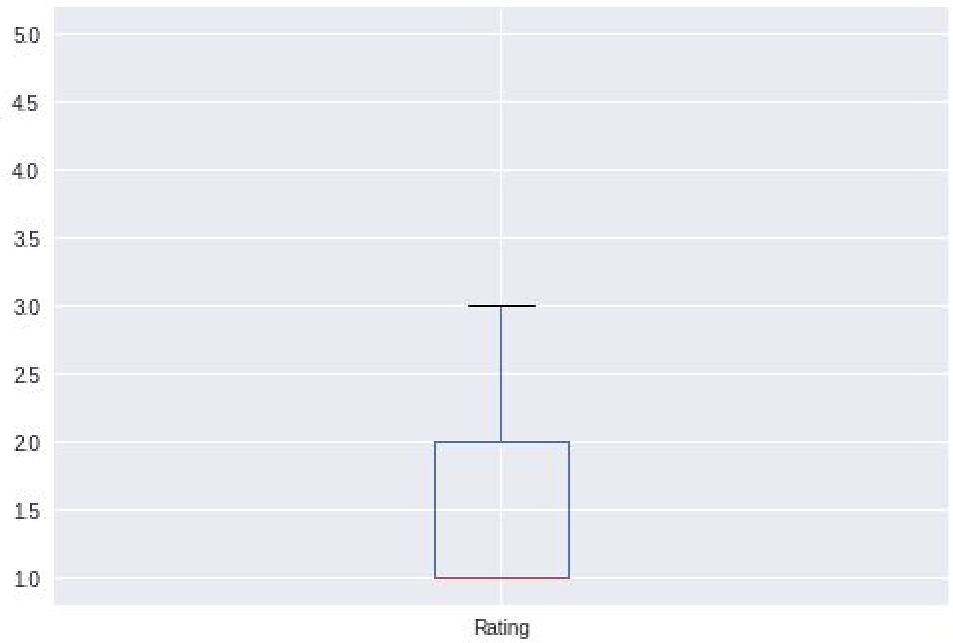


Sample Inferences from the above Heatmap

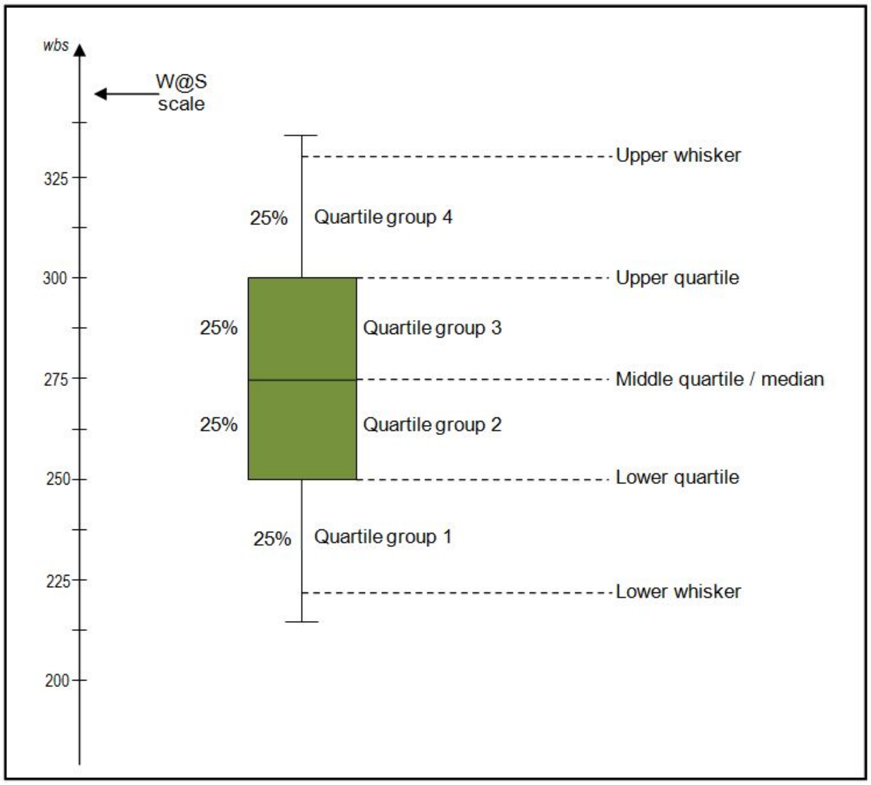
- date_timeYear and date_timeElapsed have positive correlation so it is darkly coloured.
- date_timeElapsed and Rating have negative correlation so it is lightly coloured.

5.Boxplot for various attributes are shown below.


16000
14000
12000
10000
8000
6000
4000
2000
0
Views



5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Rating



wbs
W@S
scale
325
300
275
250
225
200
Upper whisker
25%
Quartile group 4
Upper quartile
25%
Quartile group 3
Middle quartile / median
25%
Quartile group 2
Lower quartile
25%
Quartile group 1
Lower whisker