**Project – Final Report**

Introduction:-

There are two datasets that include the employed and unemployed forum data that has related scores based on various sets of words described lets us know the conscience of an individual’s thinking process, based on various reaction process throughout their usage of words how, when and where they use it. And simultaneously defining each word with linguistic dimension words, psychological process words and etc., defining the language of the data posted on the forum gives us insights on the basis of score each category of words receive.

Exploratory Data- Analysis:-

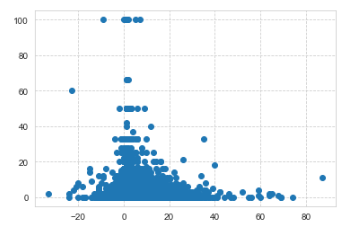
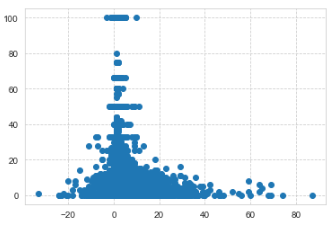
1. Data Exploration –

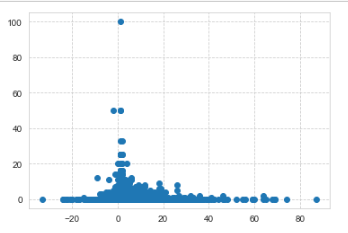
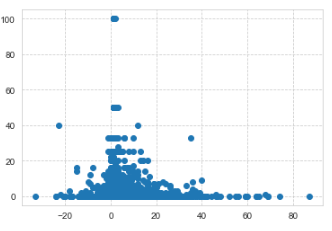
Overview of the dataset by knowing the shape, column data types, names and row features

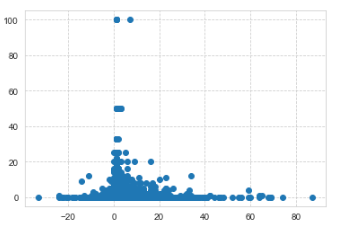
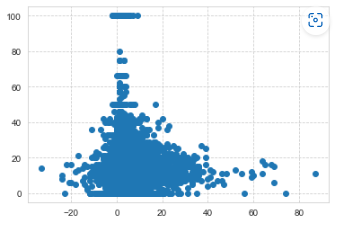
1. Missing value detection
2. Summary Statistics
3. Frequency count of feature columns
4. Data Visualizations –

a) Histogram b) Box Plot c) Scatter Plot

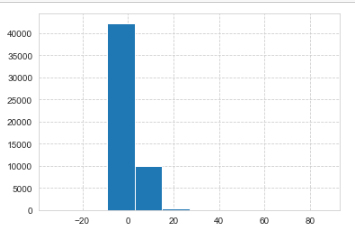
Scatter-Plots shown below are scores associated with the psychological and time orientation related features.

1. Score vs negemo ; b- score vs posemo ; c- score vs anx ; d- anger ; e – sad; f- cogproc
2.  b. 

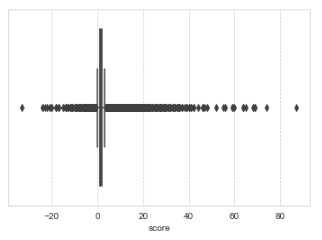
c .  d. 

e .  f . 

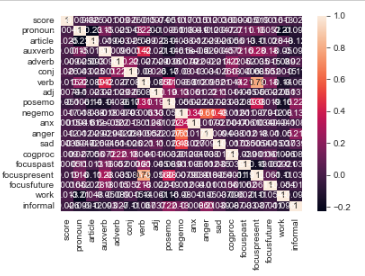
Histogram –



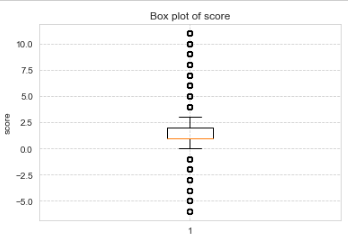
Box-plot –

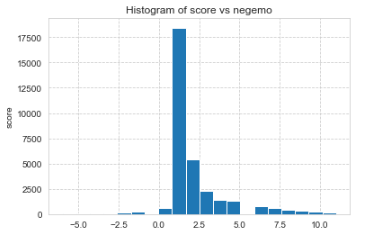
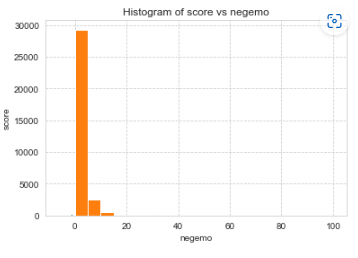
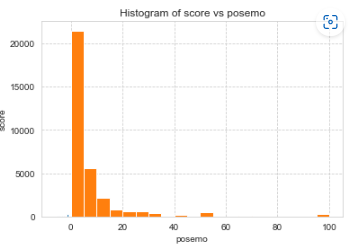
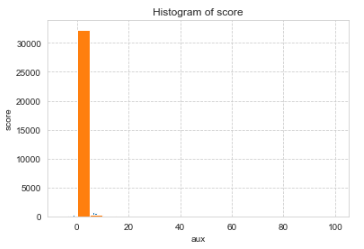


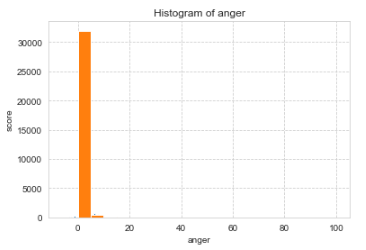
1. Conversion of Float columns to integer columns
2. Cor-relation Matrix:- It helped detect which features I could take into consideration for my analysis on the getting to know the indigenous thought process of a person based on which how many times he/she uses those words.



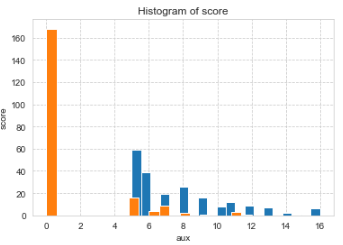
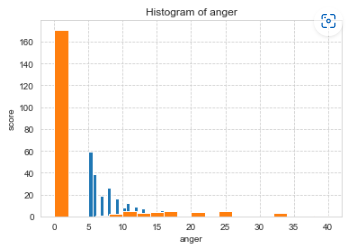
1. Outlier Detection Visualizations :-

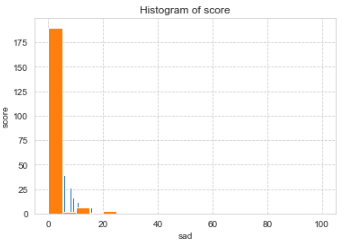
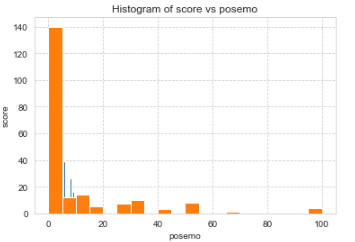




* With Outliers:-

**Goal- 1: Prediction**

**Features used:-**

1. Anx
2. Sad
3. Anger
4. Cogproc
5. Poseme
6. negemo
7. Focuspresent
8. Focuspast
9. Focusfuture
10. score

**Models Used and Scores Obtained:-**

* Decision Tree: - Comparatively decision tree gives us the modules of the features that are contributing to the better score. Here it is cognitive process.

Accuracy – score: - 0.769123

Tree modular has high score for

MSE – 0.3451

RMSE – 0.43127

R2 value – 0.62147823

* Linear Regression:-

Scores for Linear Regression are farther away and gives least amount of insights and through it we can decide on dropping the irrelevant columns.

R2 value- 0.38761

* Polynomial Regression:-

R2 Value- 0.5987124

* Ridge Regression:-

R2 Value- 0.234567

* Elastic-Net Regression:-

R2 - -9.567124

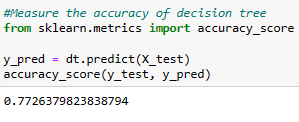
* XGBoost:-

R2 – 0.6793451

Process and Prediction Mechanism:-

Main Idea – 1. The goal focuses on predicting the sensitivity and emotions of the getEmployed dataset where the frequency of words that are used tells us the thinking capacity of the person.

2. According to the feature columns selected which is mostly based on the time orientation and psychological words that a person uses at a particular timeframe can help us detect the frequency of emotions and accordingly predict us the score of those words at different intervals of timeframe.

* 

**Goal – 2: Pattern Detection**

Clustering Analysis used - KMeans

* Analyzing the pattern between the psychological words and cognitive process words taken the main idea is to see what emotions shown by a person gives us what kinds of groups those clustered together and have the cause related to usage of those words
* By using KMeans –

1. Subset 1:-

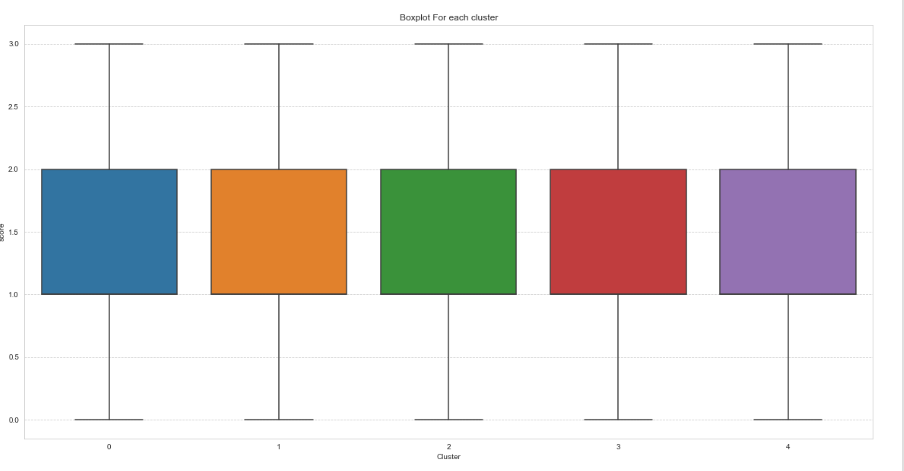
5 clusters array ranges taken for the psychological used words and after finding the centroids – a pattern that can be observed is with each data array there is no distinct significance or difference in 5 different array clusters. All the words have similar significance.

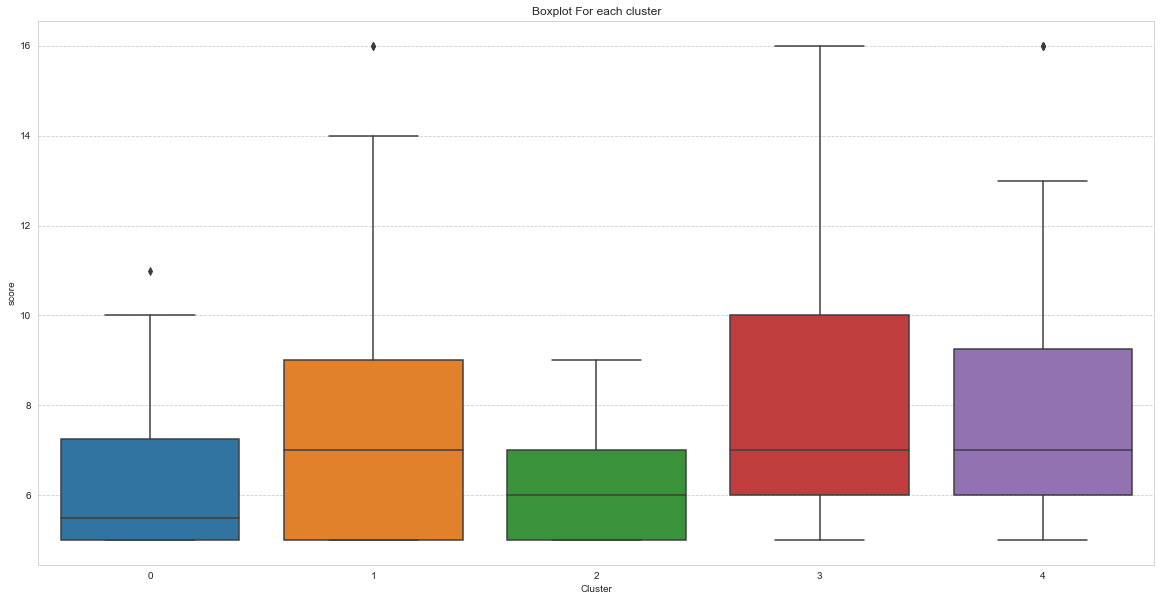
1. Subset 2:-

When feature columns such as time orientation and cognitive process features are added and the array observations are noticed the words show a pattern of getting more inclined towards and used more often at focuspresent feature words and the score rate is 1.

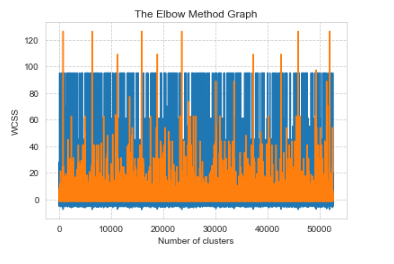
* The prediction will give us how well does any type of word when tallied with emotional words shows the emotion quotient of every group.

**Visualizations-**

****

****

* Kmeans Inertia procedure with Elbow Method Graph:-



**Conclusion-**

1. From the model approach the we can conclude that the predicting group of words taken Decision tree and XGBoost have better scores and evaluation metric
2. Kmeans algorithm clusters help detect the group which gives nearly accurate result on detecting which time frame psychological words are used by people more and what is there thinking process.