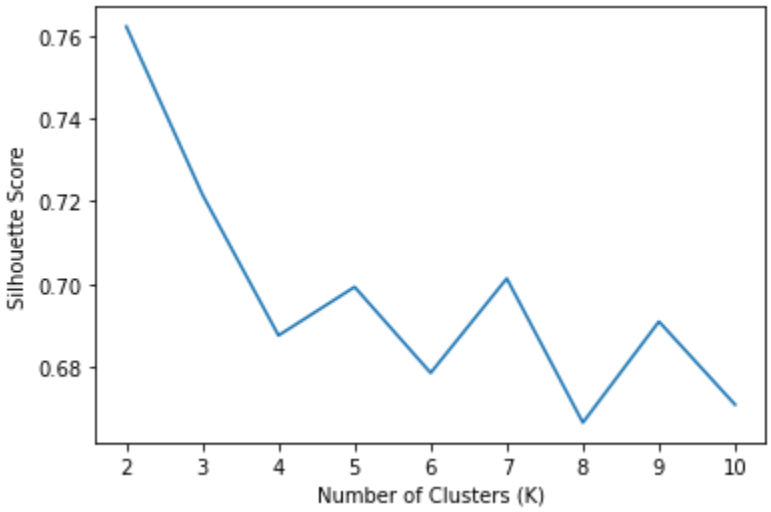
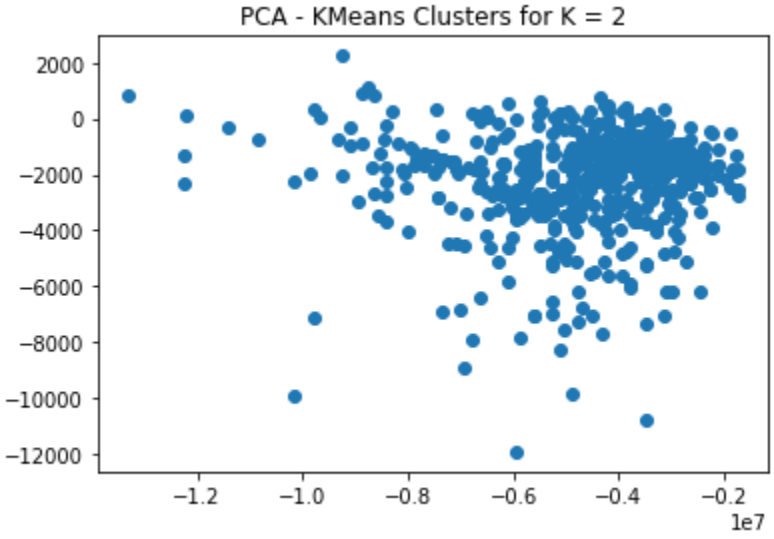
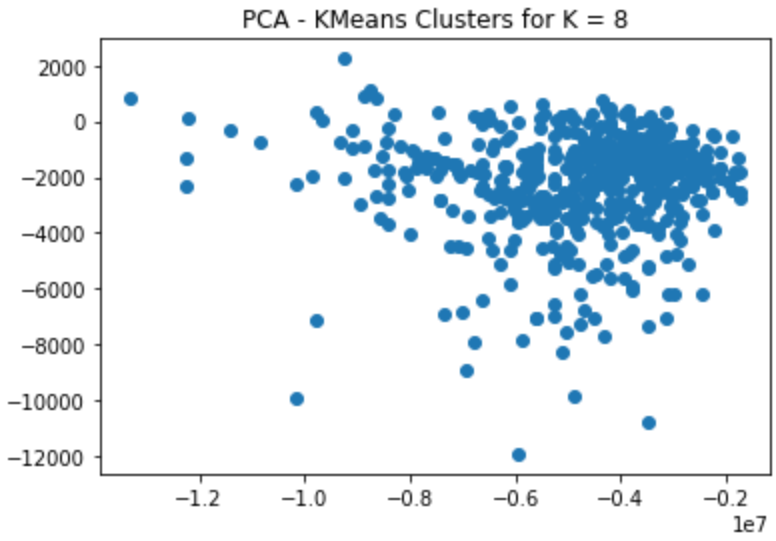
K-Means Clustering

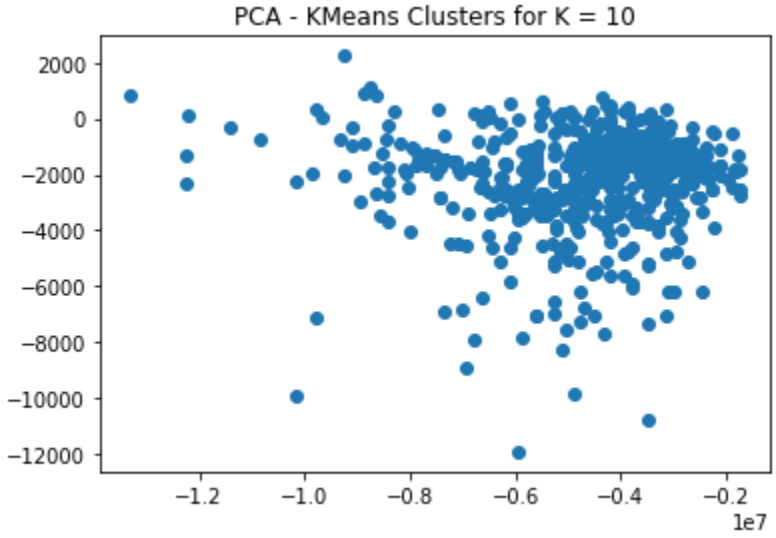


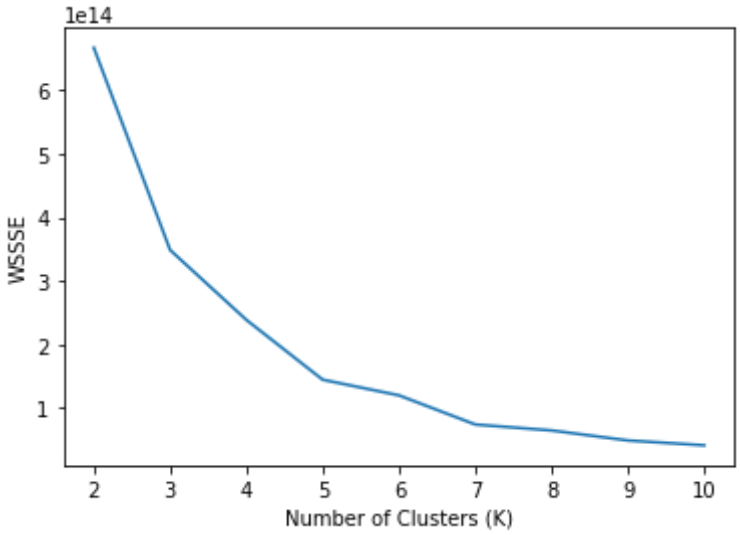
a.

b. I will only show some of the plots to save space but each plot is very similar unlike our plot from the silhouette plot which shows clear differences between each number of clusters. Our optimal value of K remains the same




PCA - KMeans Clusters for K = 8
2000
0
−2000
−4000
−6000
−8000
−10000
−12000
−1.2
−1.0
−0.8
−0.6
−0.4
−0.2
1e7


PCA - KMeans Clusters for K = 10
2000
0
−2000
−4000
−6000
−8000
−10000
−12000
−1.2
−1.0
−0.8
−0.6
−0.4
−0.2
1e7



c. We can see a steady decline in WSSE as K increases, this falls in line with our reasoning

2. Recommendation Systems

```
Mean Squared Error (MSE) = 0.738850228591604
```

a.

   i. We got a pretty high MSE which is not ideal, to mitigate this we should look into hyper-parameter tuning and perhaps preprocess the data better to avoid such a high MSE

3. Text Classification and Statistics

```
▸ train_df: pyspark.sql.dataframe.DataFrame = [text: string, label: string]
▸ test_df: pyspark.sql.dataframe.DataFrame = [text: string, label: string]
```

```
+-----+-----+
|label|count|
+-----+-----+
|    0|20019|
|    1|19981|
+-----+-----+
```

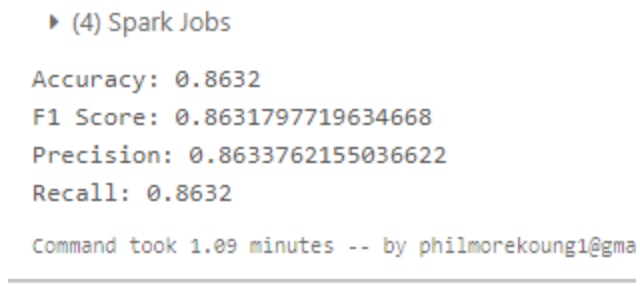
```
+-----+-----+
|label|count|
+-----+-----+
|    0| 2495|
|    1| 2505|
+-----+-----+
```

a.

```
Command took 3.67 seconds -- by philmorekoung1@gmail.com at 12/4/2023, 5:48:59 PM
```

70-30 train/test split

b.



i. Our accuracy of 0.8632 indicates that our model is correct 86.32% of the time at classifying both types of sentiments (0 and 1)
ii. F1 Score of 0.8631, or the harmonic mean of Precision and Recall, takes both false positives and false negatives into account. Given our high F1-Score, our model is very accurate in these cases
iii. Precision of 0.8634 means that we are able to identify 86.34% of the positives as positives
iv. Recall of 0.8632 indicates that we were able to capture 86.32% of the actual positive cases we tested our model against