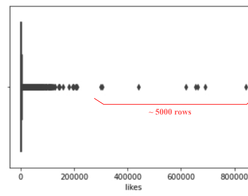
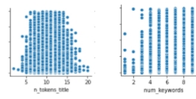
Introduction:
In the previous project, a system was developed to predict customer satisfaction based on both historical opinion of the customer and the characteristics of purchase. It was observed that an important aspect of customer satisfaction is driven by product reviews, specifically by popular reviews.

Here, we aim to build a system that predicts how many likes a comment will get to become one of the popular reviews based on textual features.

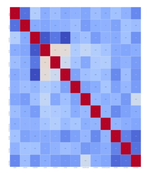
Outliers, Missing , and Duplicate data:
We searched to analyze for any duplicate or missing data and if there were any outliers within our data. In most of the columns, the data was within the same range except in the "likes" column which is the target of our data set. The number of likes are within 20,000 but few hundred "likes" were exceptionally higher. We have tried our models with removing the rows that had "likes" more than 20,000 and that caused us to have a better values for R2 and MAE, but removig data from the dataset specially the target is not the best approach.



Categorical and Numerical data:
In the bike-sharing example, we had the categorical and the numerical data were separated manually and with the select_dtype method. We have used the select_dtype to separate the categorical and the numerical data. Which there was only 2 column for the categorical. But after plotting the pair plot we noticed that we can use some of the columns from the numerical as the categorical. Finally, we have used the dummies to transform the categorical data into numerical data.



Correlated and Uncorrelated data:
We believe this part of the project was one of the main and crucial part of the project which we have noticed some issues with the dataset, which we will get into more details later on in this report. We plotted the heatmap and the pair plot that compared the data with the number of "likes" we removed the data that were correlated with each other and kept the rest. From 37 columns we dropped down to 13 columns.



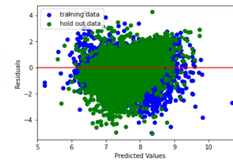
Observation:
After lots of trial and error, doing the logarithmic expression on the dataset, trying with the standard scaler and MinMaxscaler, keeping and adding different columns, balancing numerical and categorical data
we concluded that the dataset given can not accurately predict the desired values.
That is due to the data given is not linearly correlated with the target!

With the given dataset and the models we are using we can not have any "acceptable" model that can be used to predict our value. But for the sake of this project and the models provided, we tried to get the best result possible. We decided not to remove any columns from the dataset, applied the logarithmic expression on the columns that were possible, used the MinMax scaler to scale the data, and decided not to remove any outliers. For the target values since there were some outliers within as mentioned above we applied the logarithmic expression.

After running the dataset with different models, the K neighbors Regressor model and Linear Regression model were our two candidates to choose for our chosen model. but after plotting the error plot we analyzed that the train and the test set were matched better for the Linear Regression model.

K neighbors Regressor

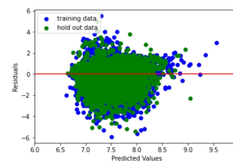
```
MAE  train 0.636 (2413.366440)  test 0.652 (2394.086521)
MSE  train 0.740               test 0.781
RMSE train 0.860               test 0.884
r2   train 0.139               test 0.091
```



We have chosen the Linear Regression model as our chosen model for this dataset. Due to being one of the simplest models and we got one of the best values for Mean Absolute Error(MAE). The R2 value is not the best within the models but it's a trade-off compared to the MAE value. The error plot for our model was not the best but the values for the test set and the trainset were within the same ranges.

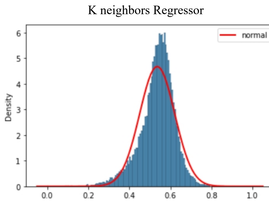
Linear Regressor

```
MAE  train 0.651 (2443.891212)  test 0.654 (2400.811874)
MSE  train 0.763               test 0.778
RMSE train 0.873               test 0.882
r2   train 0.113               test 0.095
```

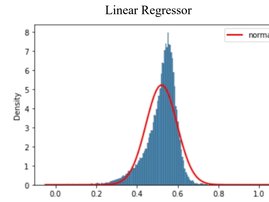


After comparing the error distribution of the K neighbors Regressor and Linear Regression model the distribution of the K neighbors Regressor was more of a normal distribution compared to the Linear Regression model. But still, we have decided to choose the Linear Regression model as our chosen model due to its prediction behavior.

K neighbors Regressor



Linear Regressor



Assignment by: Ali Shadman Yazdi