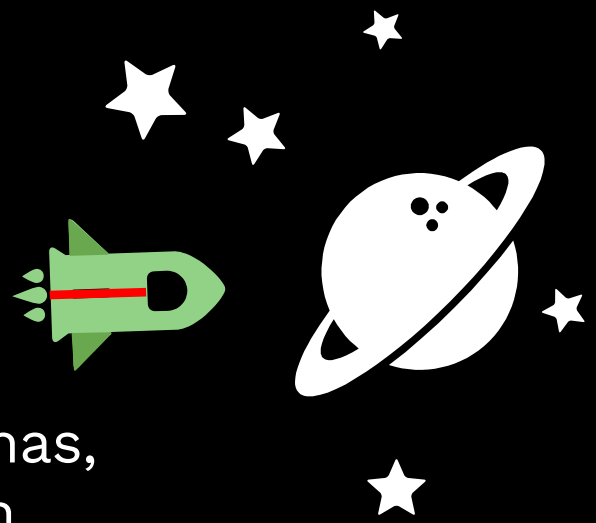
IT IS OUR PRESENTATION.

IST 718 MICHAEL J., DAN C., ALEX R.

BIG CONCEPT

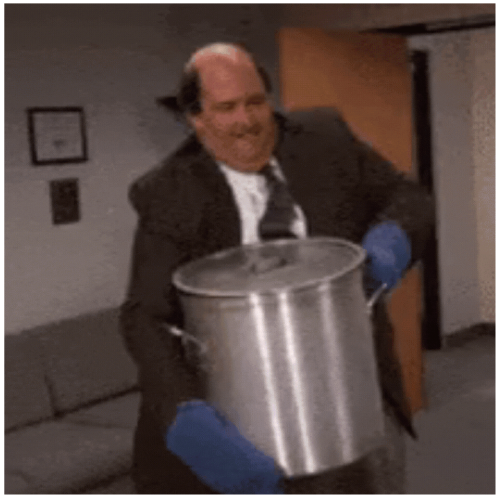


In the world of creating customer personas, many approaches use cookies along with metadata to create a persona. An alternative approach for commerce companies is asking customers which three fictional character they relate to the most.

TABLE OF CONTENTS

01 THE CHARACTERS

Scripts from 4 popular shows from Kaggle:



THE OFFICE(US)

9 seasons

18 characters

SEINFELD

9 seasons

4 characters

RICK & MORTY

3 Seasons

5 Characters

FUTURAMA

6 Seasons

7 Characters

ECOMMERCE BEHAVIOR DATA

42,448,764 rows

That’s a lot of data

5.60 Gigabytes

Still heck of a lot of data

3,022,290 Users

That’s a lot of users

ECOMMERCE BEHAVIOR DATA

381,504 rows

That’s a lot of data

0.11 Gigabytes

Still heck of a lot of data

95,442 Users

That’s a lot of users

02 THE PROCESS

CHARACTER SENTIMENT ANALYSIS

Step 1:
Clean and organize scripts

Step 2:
Check scope of data

Step 3:
NLTK and NRC packages for sentiment analysis

CLEAN & COMBINE CUSTOMER DATA

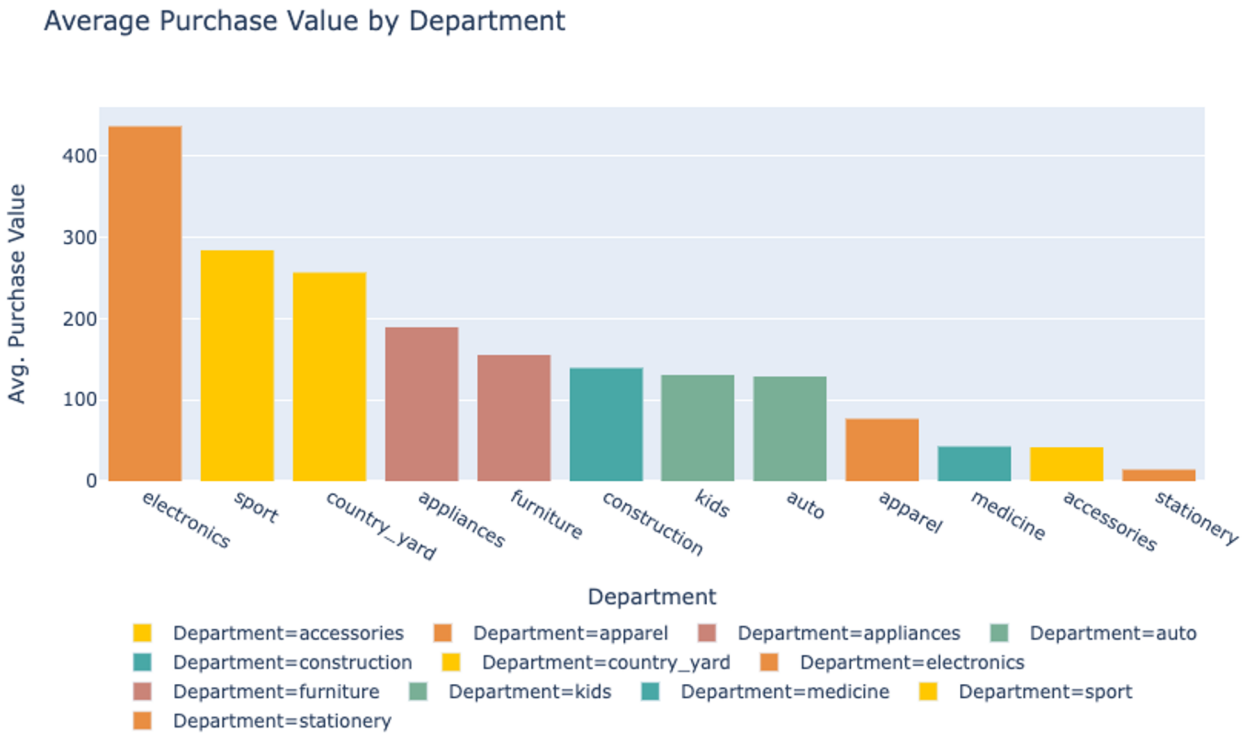
Step 4:
Split customer data into categories

Step 5:
Model customer sentiment and purchases

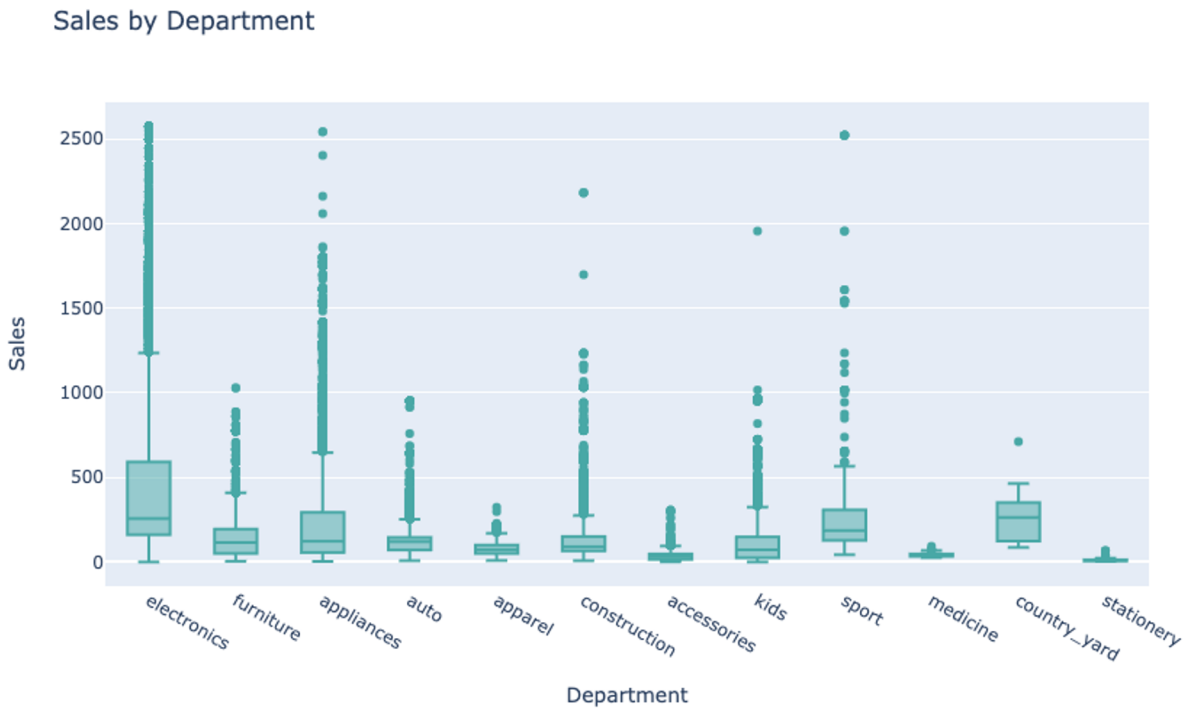
Step 6:
Classify character and customer data

CONNECTING THE DATA

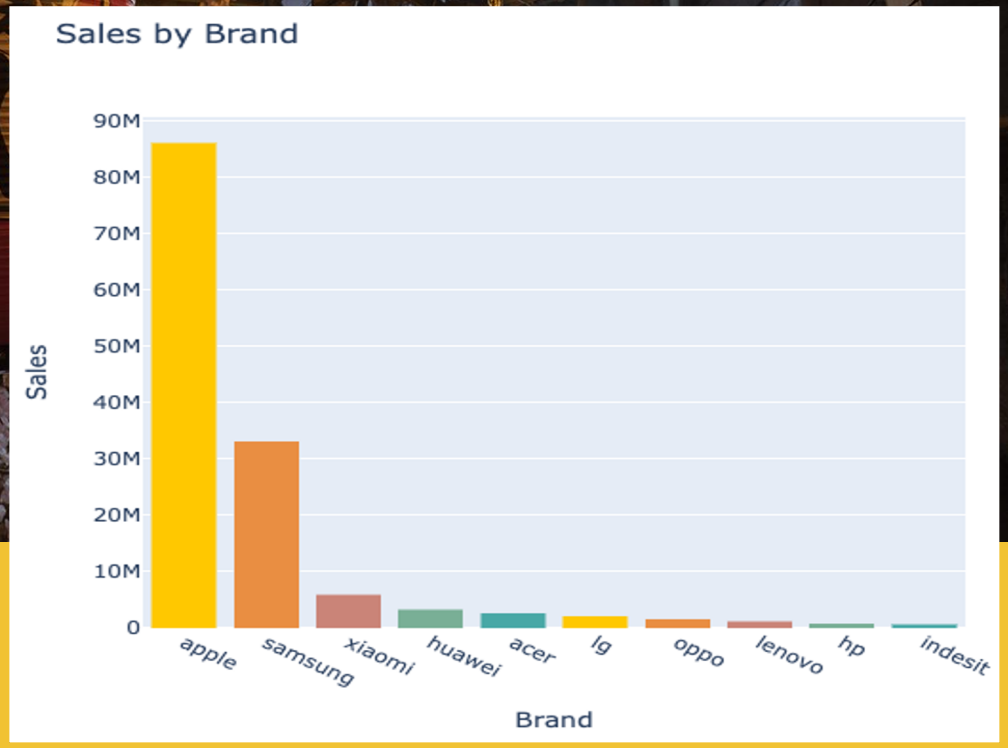
The character sentiment data was connected to the ecommerce data by creating a mapping using total sales. The higher value customer were assigned more positive characters, low value customers were assigned negative characters.

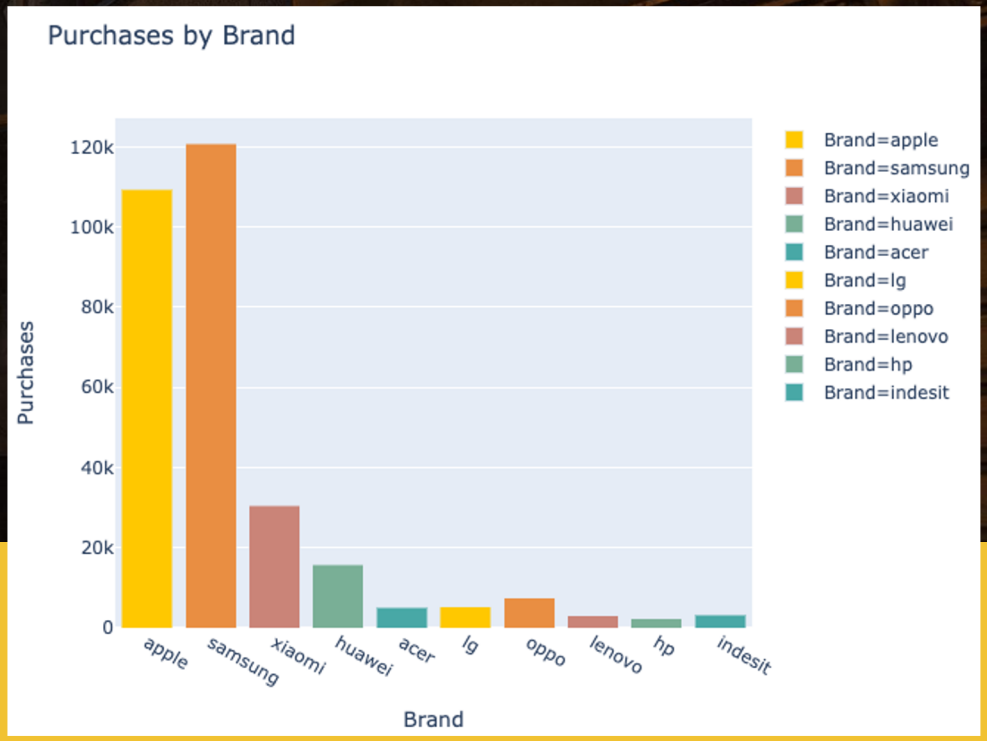


- ➔ Electronics represented roughly 91% of the sales
- ➔ The average purchase value (APV) of electronics was $437 per item
- ➔ With Sports and Country yard around $250

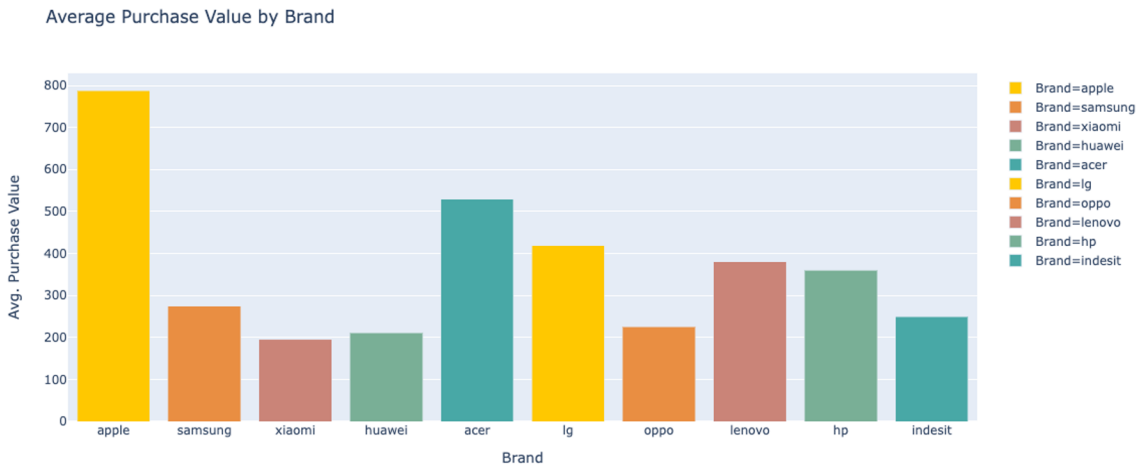


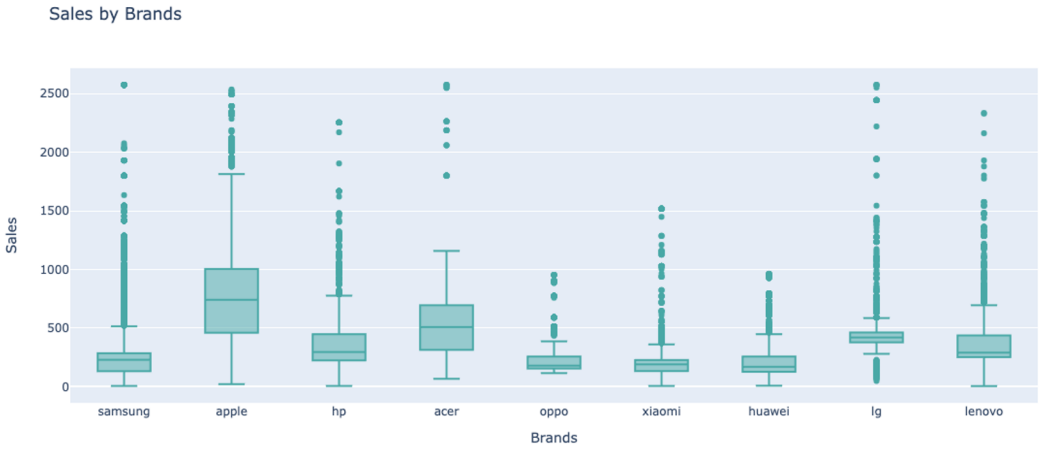
- Although the APV was $437 for electronics, the median purchase value was $256
- Electronics has a lot of outliers which makes sense with products ranging from headphones to TV's.



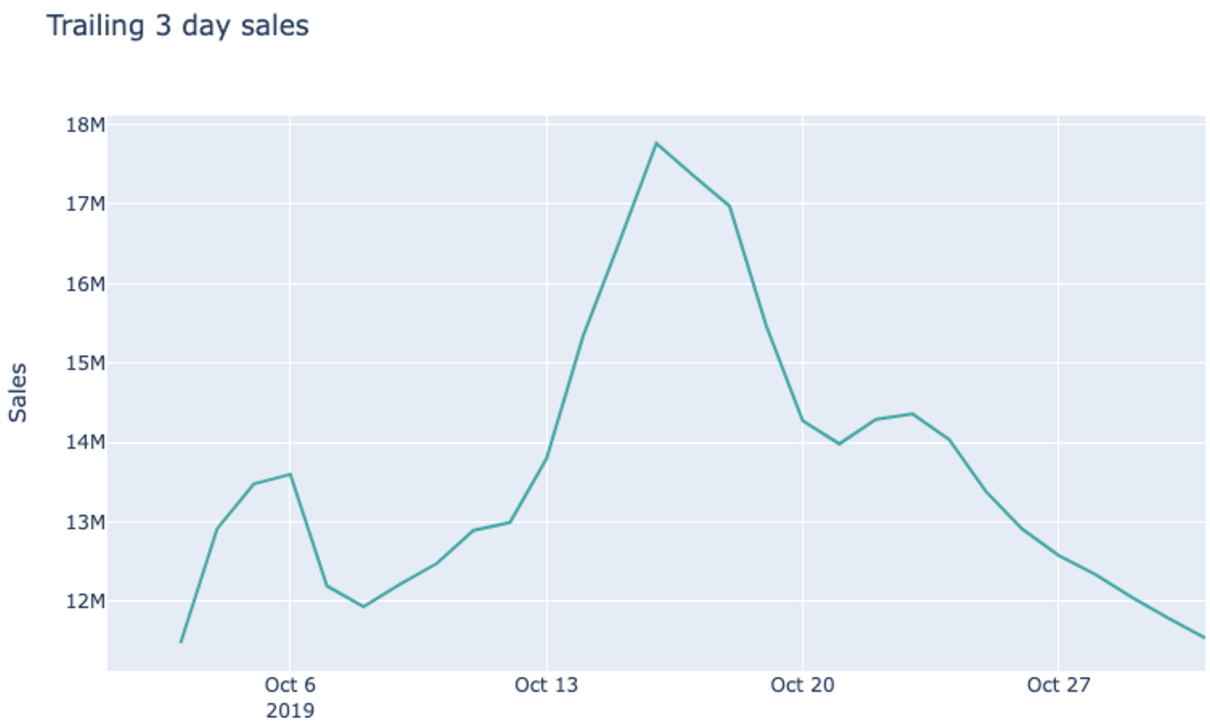


Apple represents a large amount of the sales but Samsung takes the win in selling the most products.

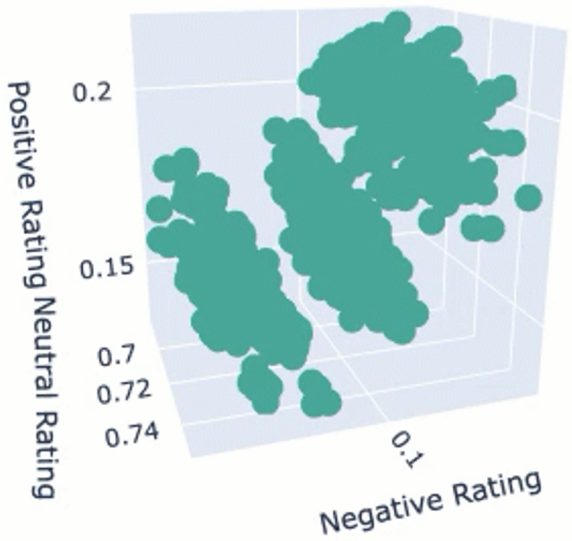




- ➔ Apple has the highest average purchase value.
- ➔ The highest median purchase value.
- ➔ Samsung has nearly the lowest with large outlier.
- ➔ This makes sense with the vast span of products they have.



- The beginning and end of the month sales are much lower than the middle of the month.
- Sales peaked on October 15th around $17.5 million than declined down.



- Users sentiment score plotted on a 3D plot.
- There was a distinct difference between 3 groups.
- We will explore this in the classification side.

K MEANS CLUSTERING

➔ **Clustering goal:**

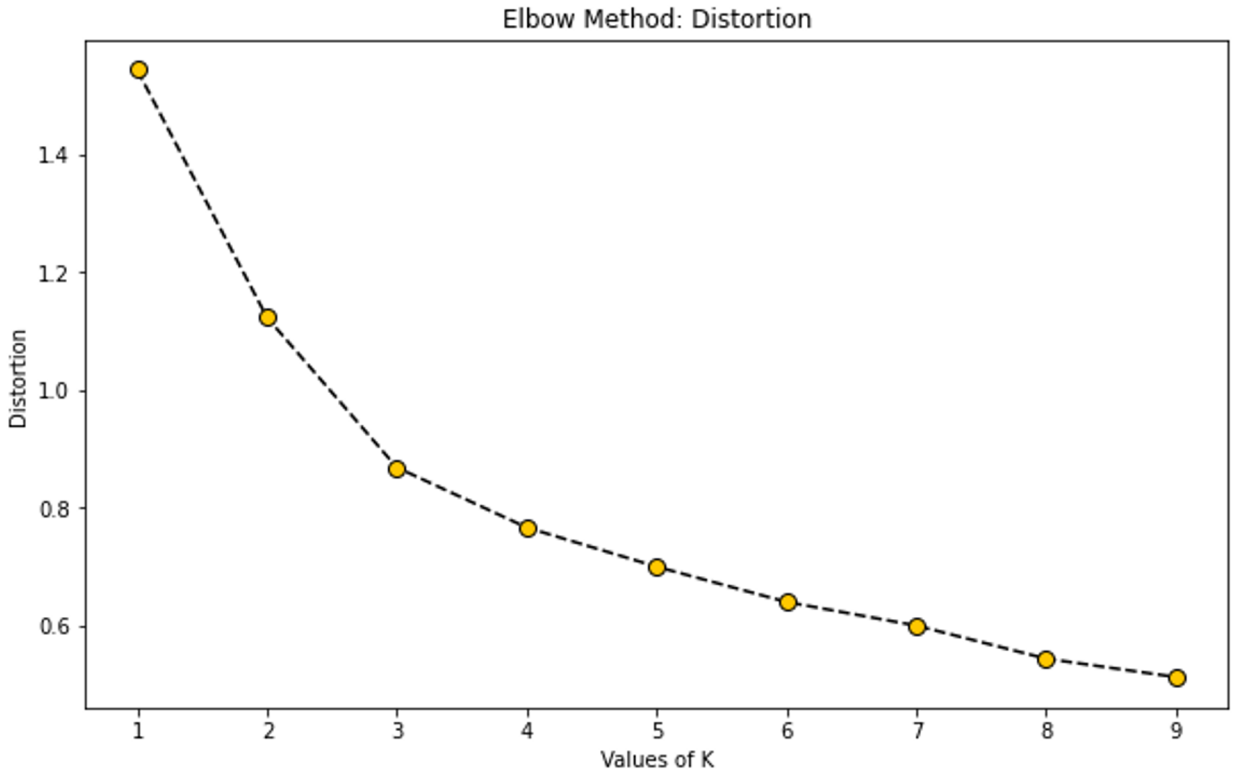
identify a way to group customers based on sentiment analysis from character choices.

➔ Identified the best K values based on distortion (Sum of Squared Errors) and Inertia (distance between points in a cluster).

DISTORTION BY K VALUE

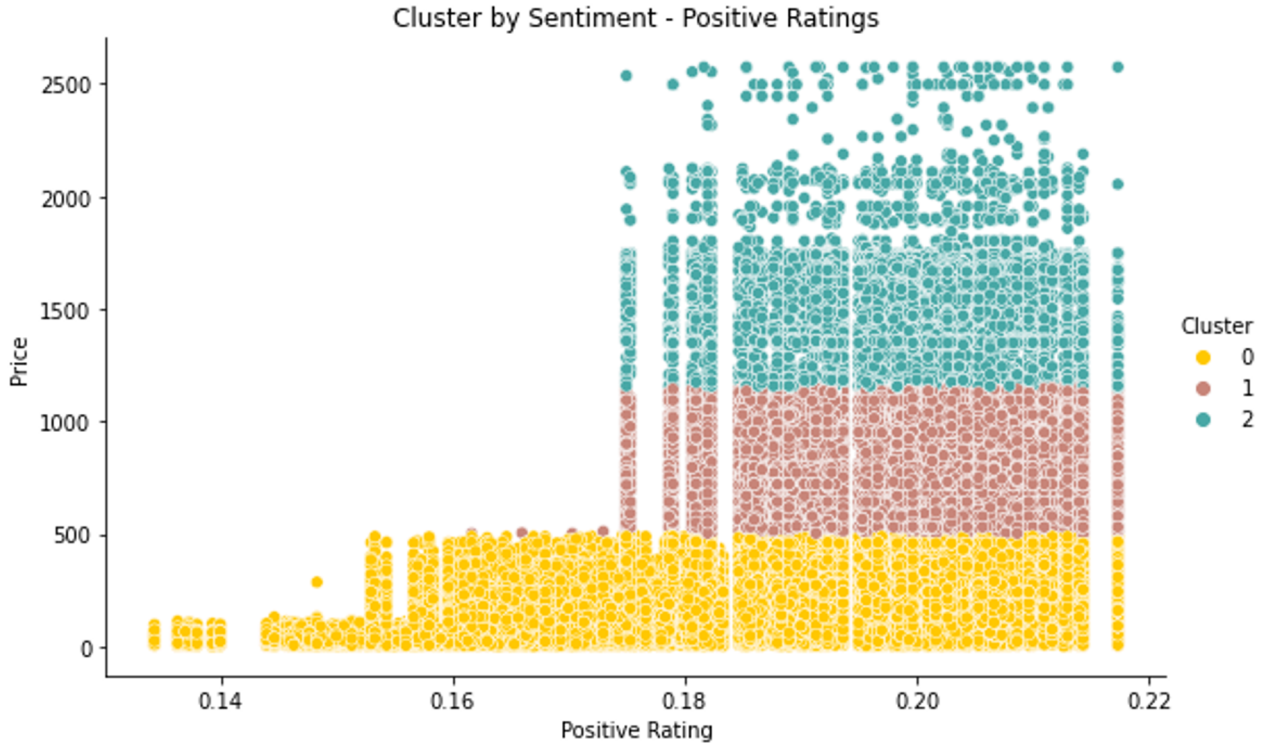
The K values chosen were based on the the sum of the squared distances between each observation vector and its centroid.

The graph indicates difficulty using this method as there is no clear point of change, known as the “elbow”.



CLUSTERS COMPARISON

The clusters picked up change points at random price intervals but not in relation to the sentiment rating.



THE MODELS 03

Main findings...

LINEAR REGRESSION

- **Linear Regression goal:**

  Identify the best accuracy value for predicting price based on sentiment analysis. Both SKLearn and Statsmodels were used.
- **SKLearn Accuracy**: 0.0974
- **Statsmodel p-values:**

  Positive < 0.05

  Neutral < 0.05

  Negative < 0.05
- **Statsmodel R Squared:** 0.121

RANDOM FOREST - REGRESSION

- **Random Forest goal:**

  Identify the best number of estimators based on accuracy for classifying customer price spending.

- **Estimators tried:**

  10, 100, 500, 1000

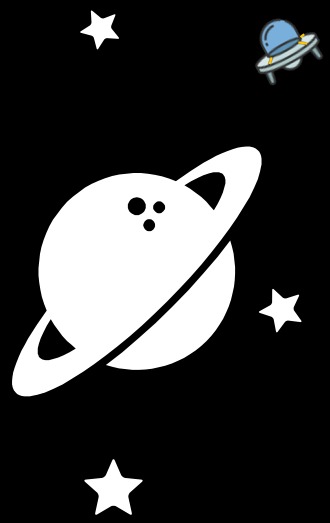
- **Best Values:**

  Estimators: 500
  Accuracy: 0.1544

REGRESSION ACCURACY

Model	Accuracy
Linear Regression	**9.74%**
Random Forest	**15.44%**

CLASSIFICATION PIVOT



Low regression accuracies required moving to classification based on spending habits.

Customers grouped by spending:

- ➔ Low: under $135.01 spent
- ➔ Moderate: over $135.01 and under $527.65 spent
- ➔ High: over 527.65 spent

Classification goal:

Successfully group customer spending based on sentiment analysis from characters.

K-NEAREST NEIGHBORS

- **KNN goal:**

  identify the best K value with the lowest error rate for classifying customer spending based on sentiment analysis.

- **K Values tried:** 50
- **Best Values:**

  K: 48

  Accuracy: 0.5320

RANDOM FOREST - CLASSIFICATION

- **Random Forest goal:**

  identify the best number of estimators based on accuracy for classifying customer spending based on sentiment analysis.

- **Estimators tried:** 250, 500, 1000
- **Best Values**:

  Estimators: 250

  Accuracy: 0.5382

BAYES CLASSIFICATION

- **Bayes goal:**

  accurately classify customer spending based on sentiment analysis.

- **Best Accuracy:** 0.4725
- **Confusion Matrix** results (High, Moderate, Low):

11,467	17,181	0
5,492	8,395	14,593
19,208	28,038	10,078

SVM CLASSIFICATION

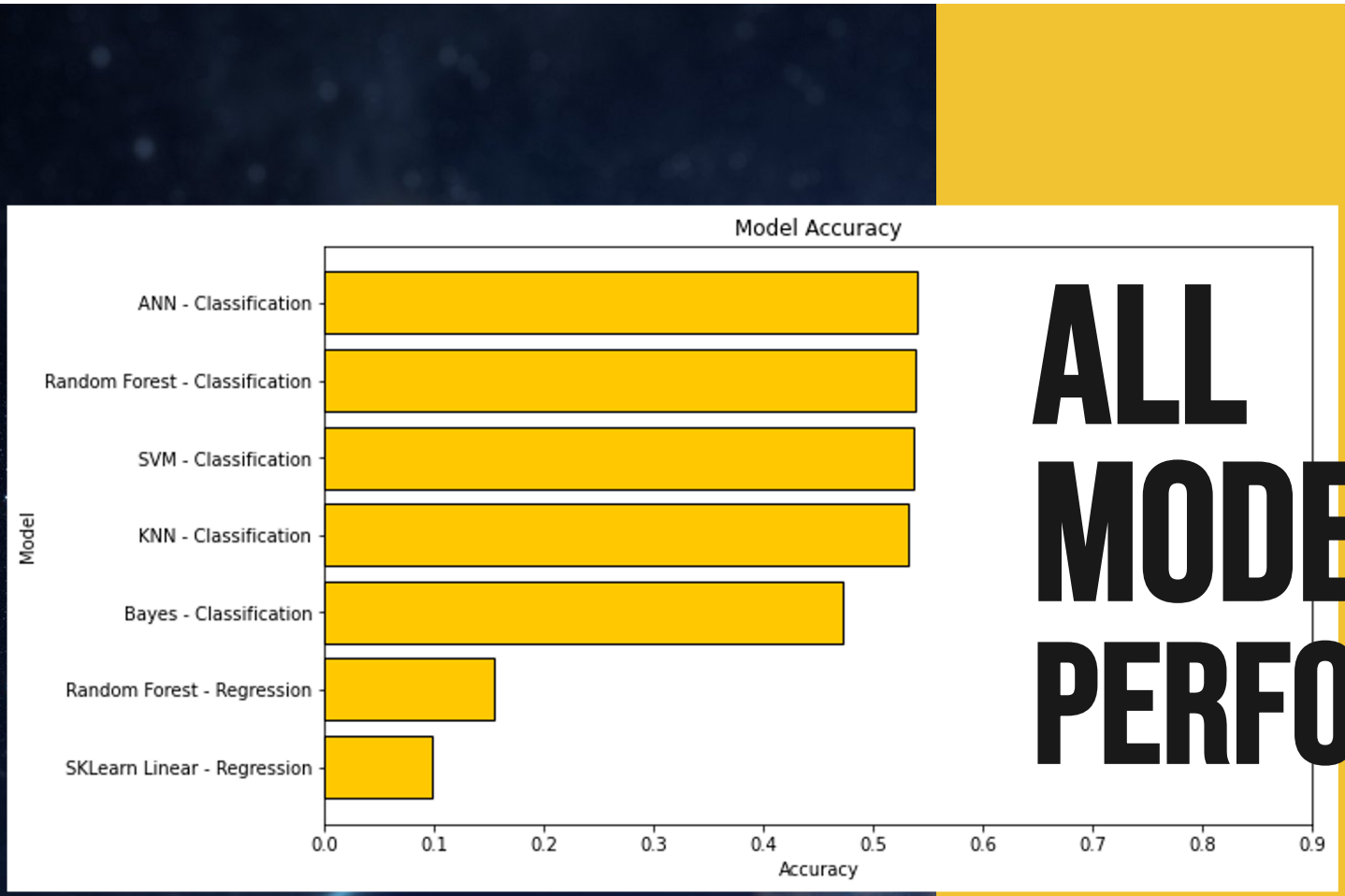
- ➔ SVM goal: find the best cost and gamma parameters to accurately classify customer spending based on sentiment analysis.
- ➔ Best Values:
  - ◆ C: 10
  - ◆ Gamma: 0.1
  - ◆ Kernel: RBF
  - ◆ Accuracy: 0.5368
- ➔ Note: very high computational burden.

ARTIFICIAL NEURAL NETWORK

- ANN goal: find the best number of layers, neurons, and learning rate to accurately classify customer spending based on sentiment analysis.
- Best Values:
  - Layers: 4
  - Learning Rate: 0.01
  - Accuracy: 0.5401
- Note: high computational burden.

CLASSIFICATION ACCURACY

Model	Accuracy
K-Nearest Neighbors	53.23%
Random Forest	53.82%
Bayes	47.25%
Support Vector Machine	53.68%
Artificial Neural Network	54.01%



ALL MODELS PERFORMANCES

04

THE CONCLUSION

LET'S REVIEW SOME CONCEPTS

Character Data

Character sentiment analysis conducted with NLP assigns ratings for positive, negative, and neutral based on analysis of their scripts.

Ecommerce Customer Data

The Ecommerce data contains transactions for customers in October 2019. Customers are assigned sentiment values based on character selections.

Clustering Attempts

Clustering the customers did not provide a great solution for understanding their spending habits based on the associated sentiment analysis.

Regression Difficulties

Regression modeling could not use customer sentiment analysis to predict how much a customer would spend and resulted in low accuracy scores.

Classification Improvements

Classification of customers by spending habits resulted in higher accuracy scores but still did not result in confident outputs based on sentiment.

Overall Model Accuracy

The models tested did not result in actionable results for the company. The best accuracy score using Random Forest only resulted in 54%.

FINAL RECOMMENDATIONS

- ➔ **Develop additional character choices to include a wider variety of characters in the analysis.**
- ➔ **Include additional sentiment ratings based on additional factors outside positive, negative, and neutral.**
- ➔ **Collect real customer character choices to better align purchasing habits with character sentiments.**
- ➔ **Invest more resources into examining product recommendations with sentiment using Association Rule Mining or Collaborative Filtering.**


Questions?