Customer Purchasing Habits Based on Weather Conditions

Group 3: Nirmit Jasapara, Aidan Jones, Phuc Le, Phuoc Le

Project Objective

The objective of this project is to find any correlations between weather/climate data and retail sales data to better predict customer spending trends. Specifically, we aim to predict the quantity sales of certain products based on weather conditions.

Community Contribution

This research could have great implications for various retail businesses. It could help businesses resupply in a more granular manner, ensuring maximum profits while keeping costs down through smarter foreshadowing. Accurate supply prediction can also help to reduce waste by reducing access inventory and preventing unwanted products to be left in storage and eventually be thrown away. This also helps consumers by ensuring the products they need are always in stock and can be acquired through local channels that are faster than online channels.

Data Acquisition: Retail Sales

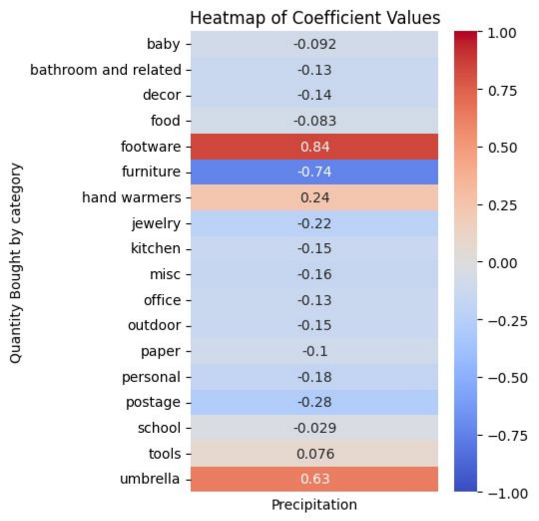
- Customer purchase history dataset: E-Commerce transactions data from retailers from a variety of countries
- Data contains:
    - Customer, unit price, quantity, date, location, and description…
    - Data did not contain column of categories of each product, only a description
- Create new column with category for each entry
    - Python script to assign a category to each entry based on keywords in the product description
    - Allows grouping similar products. Ex) grouping shoes, socks, and boots into footwear

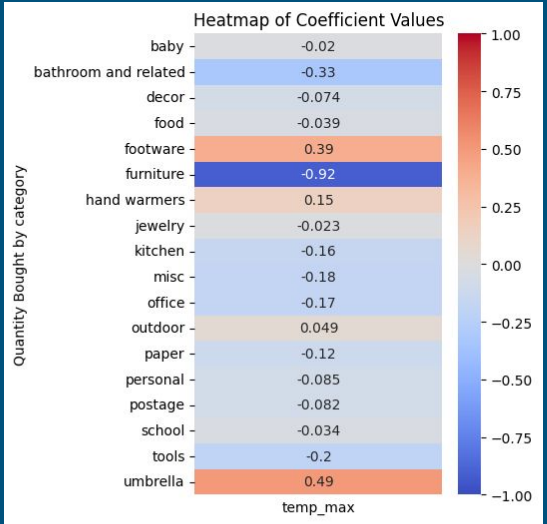
Data Acquisition: Weather

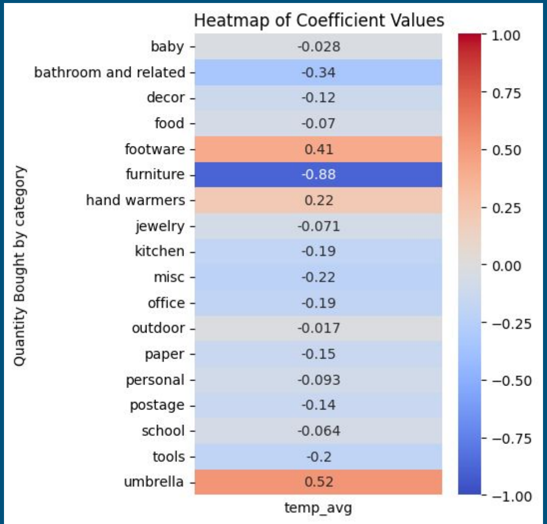
- The weather data retrieved from Meteostat
  - Open API database of historical weather and climate data
- Manually fetched weather and climate data for each row entry in the dataset
  - Weather data fetched based on entry date, time, and location
  - Data includes wind speed, temperature (low, high, average), sunlight, precipitation

Data Analysis

- Looked at quantity sold by categories for certain weather conditions
  - Group each category with the weather condition and summing them by quantity
  - Create heat map of the relationships
- Moderate and high correlation of
  - Footwear, furniture, umbrella







Predictive Models

- For different categories of products, given the weather information, we can predict the quantity of the item that will be purchased
- Two separate ensemble classifiers were created
  - Decision tree
  - Support Vector Machines
  - K-nearest neighbors (only used in one of the classifiers)
- Classifier with KNN used for categories where there was limited data for the category
  - Footwear
  - Furniture

Model Performance

- Accuracies vary significantly from category to category, and at times from model to model.
- Some categories of products have low accuracy for all models, indicating weather does not play a large role in purchase habits for these items
    - Outdoor items
    - Tools
    - Decor
- Other categories performed poorly due to sparse data
    - Furniture, footwear

Model Accuracy

```
hand warmers                Decision Tree
,postage                     Decision Tree
,kitchen                               NaN
,misc                                  NaN
,decor                                 NaN
,jewelry                     Decision Tree
,school                      Decision Tree
,paper                       Decision Tree
,office                                KNN
,bathroom and related                  NaN
,personal                    Decision Tree
,tools                                 NaN
,food                                  NaN
,umbrella                    Decision Tree
,outdoor                               NaN
,footware                              NaN
,baby                                  KNN
,furniture                   Decision Tree
,dtype: object
```

	Decision Tree	SVM	KNN	Ensemble
hand warmers	0.666667	0.0	0.666667	0.666667
postage	0.9	0.9	0.9	0.9
kitchen	0.550265	0.449735	0.550265	0.534392
misc	0.529545	0.518182	0.509091	0.518182
decor	0.538835	0.490291	0.490291	0.509709
jewelry	0.636364	0.636364	0.454545	0.636364
school	0.722222	0.5	0.694444	0.694444
paper	0.75	0.5	0.636364	0.659091
office	0.72	0.56	0.76	0.76
bathroom and related	0.590909	0.318182	0.454545	0.454545
personal	0.6	0.6	0.4	0.4
tools	0.5	0.4	0.5	0.4
food	0.409091	0.590909	0.181818	0.363636
umbrella	0.666667	0.666667	0.666667	0.666667
outdoor	0.333333	0.166667	0.166667	0.166667
footware	0.0	0.0	NaN	0.0
baby	0.444444	0.666667	0.777778	0.777778
furniture	1.0	1.0	NaN	1.0

Model Prediction

	Description	Quantity	category	temp_avg	temp_max	precipitation	Season_Spring	Season_Summer	Season_Winter	Quantity_Category	predicted
0	HAND WARMER BIRD DESIGN	96	hand warmers	-2.2	-0.1	43.0	0	0	1	high	high
1	HAND WARMER BABUSHKA DESIGN	48	hand warmers	-2.2	-0.1	43.0	0	0	1	medium	high
2	CANDY SPOT EGG WARMER HARE	12	hand warmers	6.0	11.1	24.0	1	0	0	medium	medium
3	HAND WARMER RED LOVE HEART	96	hand warmers	15.6	19.9	89.0	0	0	0	high	high
4	HAND WARMER BIRD DESIGN	96	hand warmers	15.6	19.9	89.0	0	0	0	high	high

```
accuracy_score(predicted_data['Quantity_Category'], predicted_data['predicted'])
```

```
0.8219013237063778
```

K-means

- To improve accuracy of model:
  - Split dataset, remove all category with low accuracy for all models
  - Use K-means to do unsupervised learning to cluster and created new categories
  - Model with the new categories from K-means

```
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=10, random_state=123)
kmeans.fit(unsupervised_data[['Quantity', 'temp_avg', 'temp_max', 'precipitation']])
unsupervised_data['category'] = kmeans.labels_
unsupervised_data['category'] = unsupervised_data['category'].map(lambda x: f'Category {x}')
unsupervised_data.head()
```

K-means Cluster Models

- Using K-means - created 9 clusters or new categories

	Decision Tree	SVM	KNN	Ensemble
Category 8	0.681481	0.677778	0.637037	0.692593
Category 0	1	NaN	NaN	NaN
Category 2	0.590909	0.587413	0.552448	0.597902
Category 6	1	NaN	NaN	NaN
Category 9	1	NaN	NaN	NaN
Category 7	1	NaN	NaN	NaN
Category 3	1	NaN	NaN	NaN
Category 1	0.615385	0.615385	0.538462	0.615385
Category 5	0.591304	0.530435	0.582609	0.573913
Category 4	1	NaN	NaN	NaN

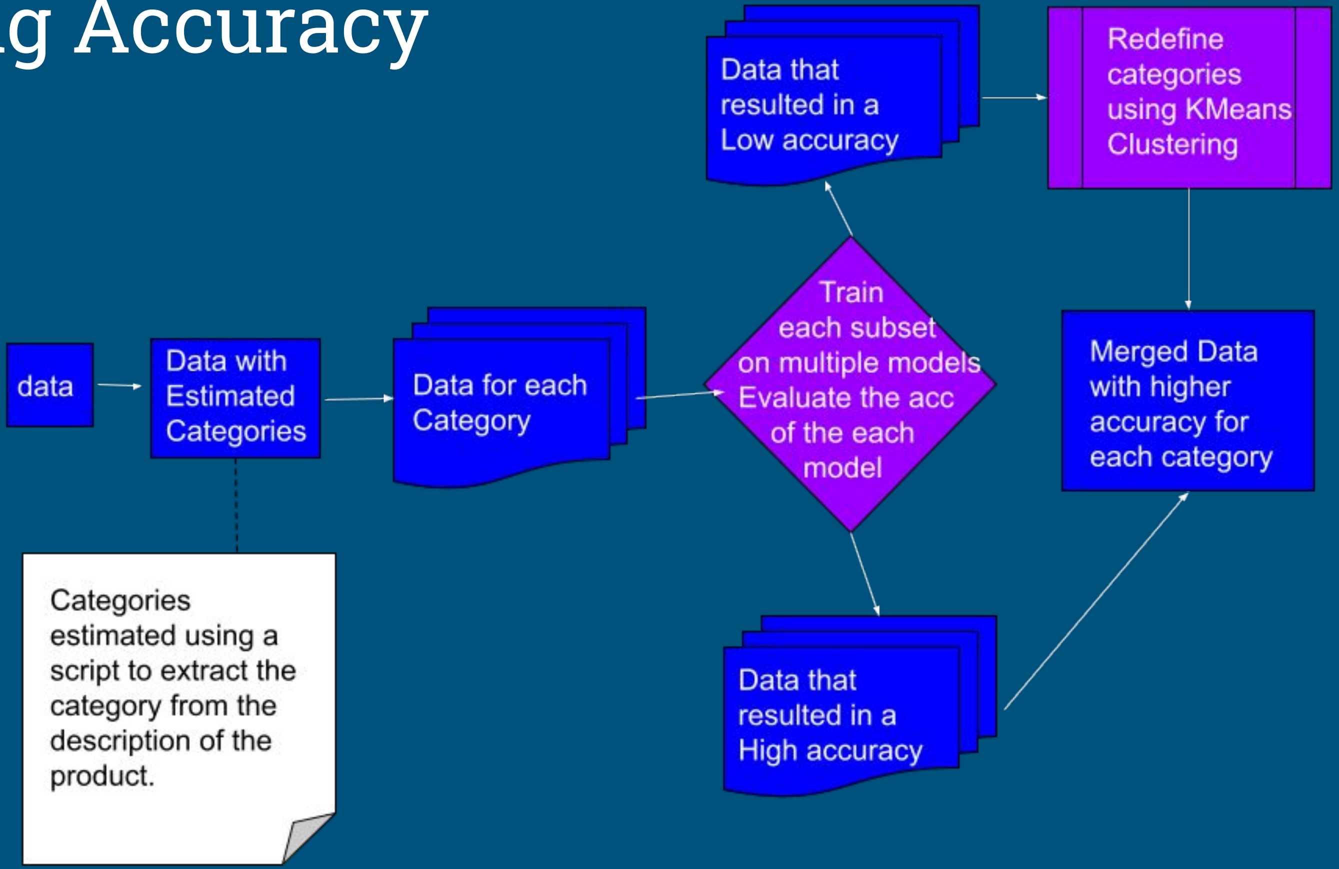
Final Model

- Combine back all data into one data set to get all categories
- Overall accuracy of model is improved

```
best_models

hand warmers    Decision Tree
,postage        Decision Tree
,jewelry        Decision Tree
,school         Decision Tree
,paper          Decision Tree
,office                   KNN
,personal       Decision Tree
,umbrella       Decision Tree
,baby                     KNN
,furniture      Decision Tree
,Category 8          Ensemble
,Category 0     Decision Tree
,Category 2               NaN
,Category 6     Decision Tree
,Category 9     Decision Tree
,Category 7     Decision Tree
,Category 3     Decision Tree
,Category 1     Decision Tree
,Category 5               NaN
,Category 4               NaN
,dtype: object
```

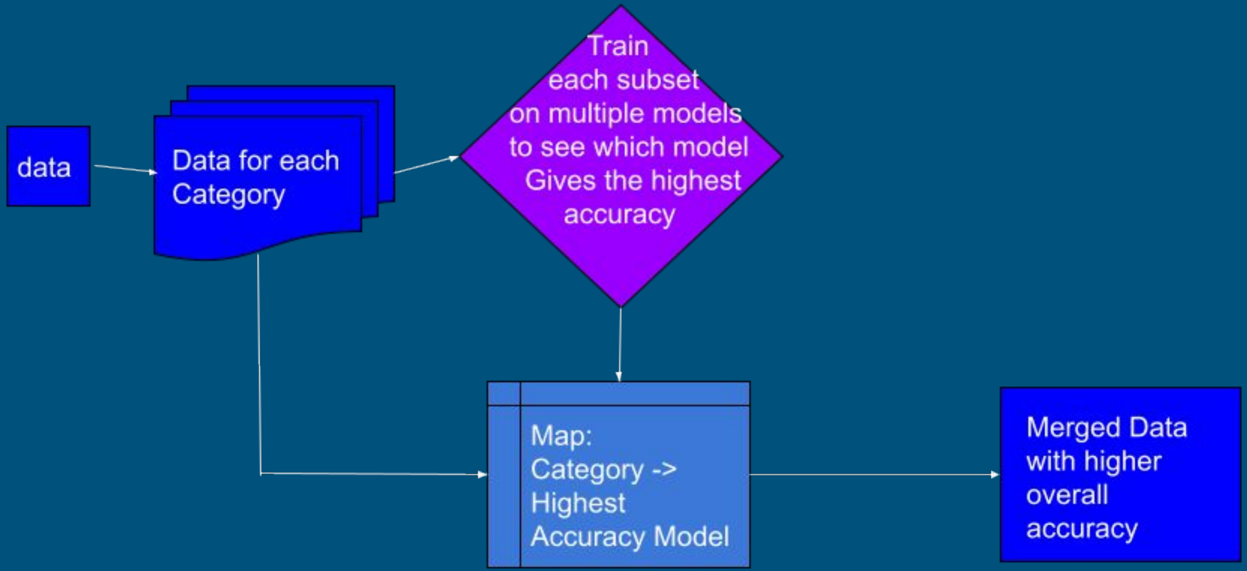
```
accuracy_score(predicted_data['Quantity_Category'], predicted_data['predicted'])

0.9166666666666666
```

Boosting Accuracy



Boosting Accuracy (Continued)

	Decision Tree	SVM	KNN	Ensemble
hand warmers	0.666667	0.0	0.666667	0.666667
postage	0.9	0.9	0.9	0.9
jewelry	0.636364	0.636364	0.454545	0.636364
school	0.722222	0.5	0.694444	0.694444
paper	0.75	0.5	0.636364	0.659091
office	0.72	0.56	0.76	0.76
personal	0.6	0.6	0.4	0.4
umbrella	0.666667	0.666667	0.666667	0.666667
baby	0.444444	0.666667	0.777778	0.777778
furniture	1.0	1.0	NaN	1.0
Category 8	0.681481	0.677778	0.637037	0.692593
Category 0	1	NaN	NaN	NaN
Category 2	0.590909	0.587413	0.552448	0.597902
Category 6	1	NaN	NaN	NaN
Category 9	1	NaN	NaN	NaN
Category 7	1	NaN	NaN	NaN
Category 3	1	NaN	NaN	NaN
Category 1	0.615385	0.615385	0.538462	0.615385
Category 5	0.591304	0.530435	0.582609	0.573913
Category 4	1	NaN	NaN	NaN



Problems Encountered and Future Work

- Some categories of products were too sparse to create effective model
  - One possible solution we did not have time for is to generate synthetic data with SMOTE
- Many categories are largely unaffected by weather conditions
- A larger and more diverse dataset will lead to better models
  - Larger, more diverse, and/or additional datasets
- Need more comprehensive sources for weather data and features of weather

Thank You