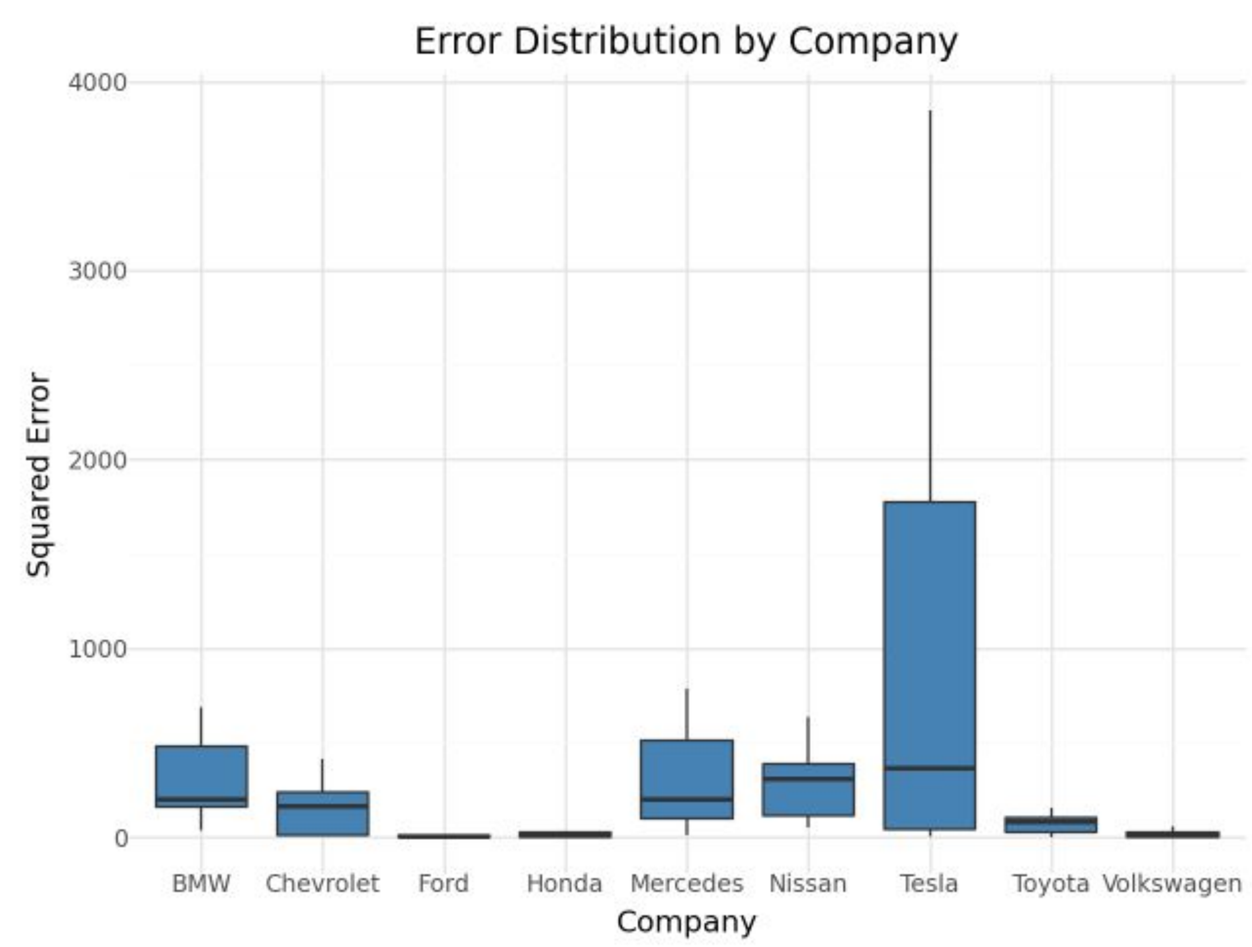
Automotive Stock Analysis

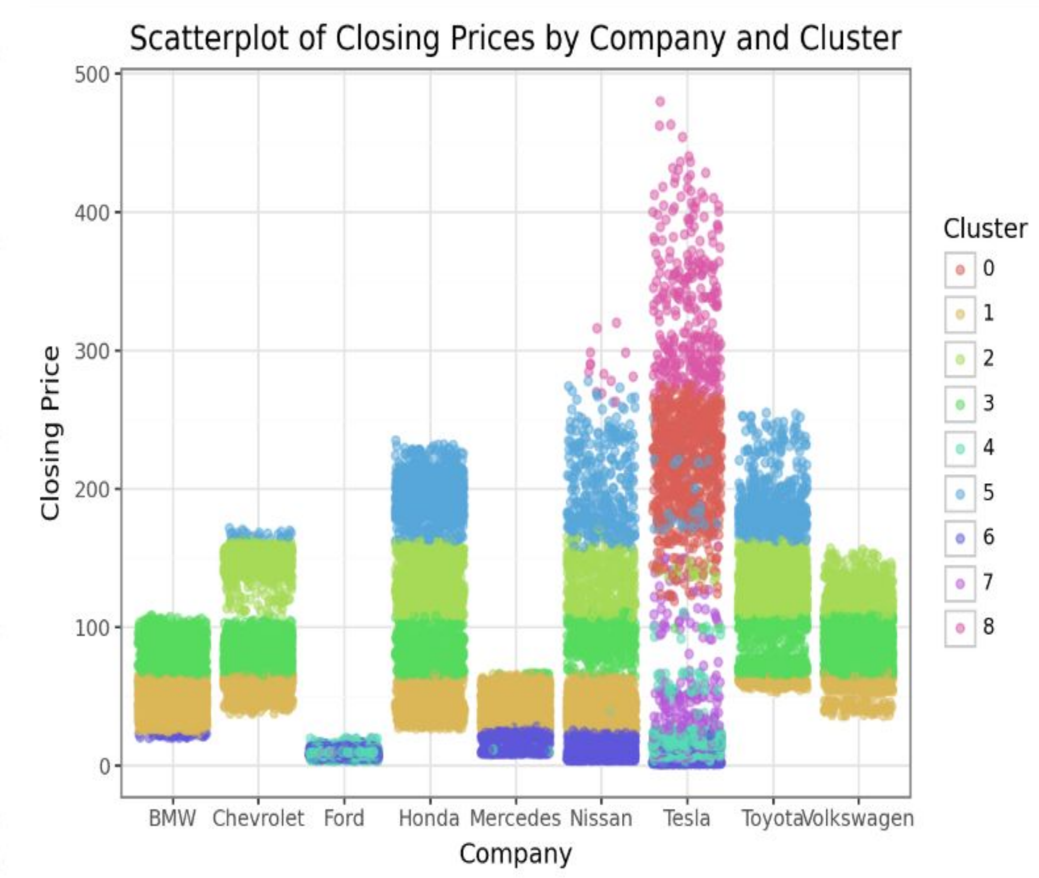
Owen Loughery, Dilraj Dhaliwal, Ian Hodgson
Data 301 - Final Project

Introduction

- Automotive companies are a large part of the global economy. The stock prices in the company are heavily influenced by politics and global events. One thing to note is Covid-19 and electric vehicles. Covid-19 made many factories shut down and new policies have made electric vehicles more popular.
- In the dataset we are working with BMW, Chevrolet, Nissan, Ford, Honda, Mercedes, Tesla, Toyota, and Volkswagen. We have the information for 2010 - 2025. The features we worked with were close price, open price, volume, low price, high price and season.
- **Primary Research Question**: Do certain automotive stock prices follows seasonal patterns which can make the price predictable?
- **Secondary Research Question**: Can we predict the company by the stock prices?

Visualization





Methods

Modifications to original data:
- Each company had their own dataset where we added a company and region column to indicate the origin of the data
- We then combined all the datasets using the concat function
- Next, because Tesla was the dataset in which we had the least observations, we dropped all dates that did not line up with the entire Tesla dataset.
- Finally, we converted the date variable into a pandas datetime variable type, and used that date to create a function to sort each observation into a season

Features Considered:
- We created a new dataset that grouped the combined dataset by season, year, and company
- within these groups, we found the mean open, high, low, and closing prices along with the mean volume

Models:
- K-Nearest Neighbors regression model
  - We chose to answer our primary research question using KNN because It relies on historical data to make predictions, which is useful when attempting to determine if stocks follow seasonal trends.
  - If a stock follows a seasonal pattern, KNN should perform well in predicting unknown seasonal stock prices.
  - If no seasonal pattern exists, KNN should struggle to make accurate predictions, resulting in higher prediction errors (mean squared error), allowing us to identify which companies, if any, follow seasonal trends.
- K Means clustering model
  - We chose to answer our secondary research question using K means because it is an unsupervised model, meaning we do not have to specify the companies to make the clusters
  - If we can use stock prices to predict the company, then there will be one cluster in which most of a company's observations will fall into.

Model

Primary:
Model Implementation & Feature Selection
- The dataset was split into training (data from years ≤ 2021) and testing (data from years > 2021) to ensure that predictions were based on past observations.
- The features selected for the model include:
  - Quantitative Variables: Mean High, Mean Low, Mean Volume, and Year (means calculated as the average for that season)
  - Categorical Variables: Company, Season
- We used a column transformer to apply:
  - Standard scaling to the quantitative variables to ensure distance calculations in KNN are meaningful.
  - One-hot encoding to categorical variables to ensure company and season information were properly incorporated into the model

Model Training & Cross-Validation
- We performed 5-fold cross-validation to estimate the test error and select the best hyperparameters for KNN.
- Hyperparameter tuning was conducted using GridSearchCV to optimize:
  - Number of neighbors: Tested values from 1 to 30.
  - Distance metric: Compared Euclidean distance and Manhattan distance.
- The best model parameters were selected based on root mean squared error (RMSE), ensuring that the chosen model minimizes prediction error

Secondary:
Model Implementation
- We used the five quantitative variables relating to stock prices and a KMeans model with 9 clusters inside of a pipeline, and fit the combined dataset
- We created a new data frame displaying the 9 clusters as columns and the 9 companies as rows
- We added the cluster number of each observation back to the combined a dataset, and made a visualization of these clusters

Results / Implications

Primary:
The model was evaluated on the test set, and we compared the predicted stock prices vs. actual stock prices.
- Mean squared error (MSE) was used as a key metric:
  - A low MSE suggests that the stock price follows a consistent seasonal pattern, making it easier to predict.
  - A high MSE suggests that the stock price does not follow seasonal trends, meaning other external factors might influence price fluctuations.
- By analyzing MSE across different companies, we were able to identify which automotive companies exhibit seasonal trends in their stock prices

From the graph representing the error distribution for each company we can see a clear difference in the mean squared error gathered from the model.
- Ford, Honda, Toyota, and Volkswagen
  - Smaller IQR and lower medians indicate that KNN makes more accurate predictions for these companies.
  - This suggests that these stocks might follow seasonal trends better, making them easier to predict
- BMW, Chevrolet, Mercedes, and Nissan
  - Errors are higher than companies like Ford and Toyota but lower than Tesla.
  - This could indicate some level of seasonality, but possibly other factors influencing stock price fluctuations.
- Tesla
  - The largest IQR and highest median indicate that the KNN model struggles the most with Tesla's stock prices.
  - The long whiskers and high outliers suggest very unpredictable price movements, meaning Tesla's stock likely does not follow seasonal trends well.

What we gain from these results:
- Investors and analysts could leverage the results of which companies follow seasonal patterns because the predictability can help make more informed trading decisions. For example companies with low error scores could benefit from seasonal forecasting models for stock price prediction.
- Companies with high error scores might require a more complex approach, such as incorporating additional features (e.g., economic indicators, market trends, news sentiment analysis). Also it is good to know that these companies follow a more volatile stock price pattern as this is important to take into consideration when making trades.

Secondary:
Results:
- After graphing Closing price by company using clusters as the color, we found that the clusters followed the stock prices rather than the company

What we gain from these results:
- the quantitative features of the stock price behavior alone are not sufficient to distinguish between companies in a meaningful way.

Ethics / Limitations

One key important limitation from this model is that it does not actually predict future prices and should not be used as an investment tool for that measure. Instead the model finds stock prices that follow seasonal trends by using KNN.

Another factor to pay attention to is that KNN does not account for external market events. KNN only uses the given dataset, meaning it cannot react to real-world financial events like:
- Economic recessions
- Oil price changes (which affect car manufacturers)
- Government regulations on emissions
- Trade restrictions on car exports/imports

These could have impacted the results from the model unknowingly so they should be taken into consideration before using the results from the model to make any financial decisions.