Analytics Final Project: Vehicle Sales Analysis

Decision Making with Business Intelligence and Analytics

ISM6358

Prepared by:

Nicholas Sutanto

**Table of Contents**

1. Business and Data Introduction…………….……………………………2
   1. Problem Statement…………………………………………………2
   2. Tools, Data, and Processes……………………………………….2
2. Exploratory Data Analysis……….……………………………………….3
   1. Vehicle Make Distribution………………………………………….3
   2. Vehicle Body Type Distribution…………………………………...4
   3. Vehicle Year Distribution……………………………………….….4
   4. Transmission Distribution………………………………………….5
   5. Vehicle Model Distribution………………………………………....6
   6. Correlation Matrix…………………………………………………...6
   7. Year VS Selling Price……………………………………………....8
   8. Condition VS Selling Price……………………………………..….9
   9. Mileage VS Selling Price…………………………………………10
   10. Average Mileage VS Year…………………………………….….10
   11. Average Mileage VS Condition………………………………….11
3. Machine Learning….…………………………………………………….11
   1. Feature Importance……………………………………………….12
   2. Machine Learning Results……………………………………..…12
4. Recommendations…………………………………………...……..……13
   1. Sell more vehicles from Ford, Chevrolet, and Nissan…………13
   2. Sell more vehicles from the years 2011-2014………………….13
   3. Sell more vehicles in good condition……………………...…….14
   4. Use Random Forests for the selling price predictions……...…14
5. The Three W’s……………………………………………………………14
   1. What Went Well……………………………………………………14
   2. What Did Not Go Well…………………………………………….14
   3. What I Would Do Differently……………………………………...15
6. **Business and Data Introduction**

This section will discuss the problem statement of the project, the tools used, the description of the data, and the processes done.

* 1. **Problem Statement**

A vehicle is a large investment for many consumers. This is one reason why having the most optimal lines of vehicles can help consumers ease the burden of choice. There are many dealerships around the country and through this analysis, we gain a competitive advantage over other dealerships by the results of data-driven decision making. As dealerships, we have to adapt to the changes in consumer behaviors, or else, we may risk losing the business. Through this analysis, we can determine what factors affect the selling price of a vehicle. It is essential for us to implement the important changes found in this analysis to keep up with consumer preferences.

**1.2 Tools, Data, and Processes**

The tools I decided to use in this analysis were Python and PowerBI. There were many other choices I could have gone for but, these were the two tools I wanted to improve at. The dataset chosen was retrieved from Kaggle. The dataset contained only one Excel file, I would have liked to use more than one file (I will explain more in the Three Ws section). The dataset contains information regarding vehicle sales. It contains 558,838 rows and 16 columns. The columns include year, make, model, trim, body, transmission vin, state, condition, odometer, color, interior, seller, mmr, selling price, and sale date.

The data needed to be cleaned as there were missing values that had to be removed/replaced. The preprocessing was done with Python and after it was done, there were 472,325 rows left. An issue I encountered while performing this analysis was processing times. It took a large amount of time as 470,000 rows of data is a lot for my personal desktop computer to process. With this issue, I decided to filter the dataset only using a total of 9 out of 53 vehicle makes. I chose 3 vehicle brands from 3 regions (NA, Asia, EU), the brands are Ford, Chevrolet, Tesla, Nissan, Toyota, Honda, Audi, BMW, and Mercedes-Benz. This resulted in the final row count being 275,642. I also created 2 new variables. These variables are body\_simple, and capacity. Body\_simple uses the data from the original body column to create a more simplified version of that column. For example, if the body type is Double Cab, Crew Cab, or Access Cab, it will return Cab. The capacity column uses the body\_simple column to provide the number of seats available based on the vehicle’s body type. For example, if the body type is sedan, the capacity would be 4. I also used Python to create a correlation matrix and a feature importance chart (discussed in EDA and Machine Learning). Lastly, the charts throughout this analysis were created with Python and PowerBI.

1. **Exploratory Data Analysis**

The exploratory data analysis showed many insights that I was able to gain from this project. Some were more robust than others. In this section, I will talk about my findings throughout the process.

**2.1. Vehicle Make Distribution**

Figure 1 shows the most popular vehicle makes in this dataset. As seen in Figure 1, the most popular vehicle make is Ford, with around 81,000 sales. This is followed by Chevrolet coming second with 54,000 sales and closely followed by Nissan with 44,000 sales ending off the top 3. Through this information, we can see that consumers prefer the more affordable American and Japanese cars compared to the more expensive, and luxurious European cars. It also shows that consumers prefer more local vehicle brands such as Ford and Chevrolet (Tesla has under 1000 sales due to the vehicles sold being from the years 1990-2015, Tesla was quite new to the vehicle market and started selling their Model S sedan in 2012).

A graph with blue bars

Description automatically generated

Figure 1. Vehicle Make Distribution Chart

**2.2. Vehicle Body Type Distribution**

Figure 2 shows a chart that displays the count of vehicle body types in the dataset. As shown in Figure 2, there is an overwhelming favorite among all the body types. The most popular vehicle body type is a sedan. A sedan body type has around 129,000 sales followed by SUV with 67,000 and cab with 23,000 sales. This tells us that most consumers prefer a smaller, more practical vehicle for everyday use.

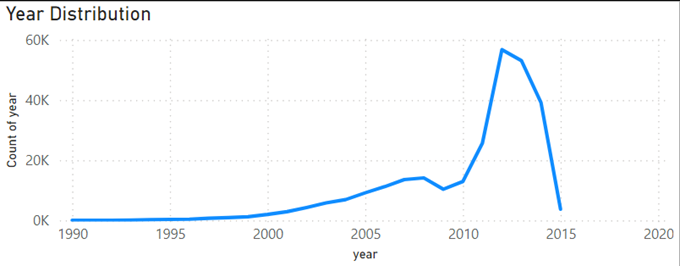
**A graph of body type distribution

Description automatically generated**

Figure 2. Vehicle Body Type Distribution Chart

**2.3. Vehicle Year Distribution**

This next chart shows the years of the vehicles sold in the dataset. Figure 3 shows that consumers tend to prefer purchasing newer vehicles. Starting from the vehicles made in 2011, there is an increase in the number of vehicles sold, peaking in vehicles made in 2012. The most popular years of vehicles sold in this dataset are 2012 with around 56,000 vehicles sold, 2013 with 53,000 vehicles sold, and 2014 with 39,000 vehicles sold. With this information, we can see that consumers prefer newer year vehicles. This could be due to factors such as safety, technology, and value. According to Progressive, vehicles may lose up to 20% of their value in the first year and may lose up to 15% more per year. This may also be a reason why consumers prefer newer, used vehicles (https://www.progressive.com/answers/what-is-car-depreciation/).

**Figure 3. Vehicle Year Distribution Chart**

**2.4. Transmission Distribution.**

The next chart shows the vehicle transmission distribution of the dataset. As shown in Figure 4, automatic transmission is the most popular type of transmission in this dataset by a wide margin. Around 268,000 sales for automatic and 7,400 sales for manual. This tells us that a majority of consumers prefer the ease of use of the automatic transmission compared to a manual transmission.

A blue pie chart with red and blue text

Description automatically generated

Figure 4. Transmission Distribution Chart

**2.5. Vehicle Model Distribution**

This chart shows what vehicle models are the most popular in the dataset. The chart through an image will look messy as I cannot quite show it with the vehicle make filters. As shown in Figure 5, the most popular vehicle in this dataset is the Nissan Altima with around 16,000 sales, followed by the Ford Fusion with 12,000 sales, and the Ford F-150 ending in the top 3 with 11,000 sales. This was quite surprising as Figure 1 shows that Ford has almost double the amount of vehicle sales compared to Nissan but the most popular vehicle sold is from Nissan. The second most sold Nissan is a Nissan Maxima with only around 5,400 sales.

A screenshot of a computer screen

Description automatically generated

Figure 5. Vehicle Model Distribution Chart

**2.6. Correlation Matrix**

This next visualization shows a correlation matrix. A correlation matrix is used to show the relationship between one variable and another. A heatmap is used to visualize the correlation. Figure 6 shows the relationship between the selling price and all the other variables. In Figure 6, we can see that the variables that are most correlated with the selling price are year, condition, and odometer. These are some of the variables that I expected to be correlated with the selling price. I was surprised to see that make is not heavily correlated with with selling price since I believed that the make of a vehicle could influence the selling price. Some other notable correlations include odometer and year, and condition and year.

**A screenshot of a graph

Description automatically generated**

Figure 6. Correlation Matrix Heatmap

**2.7. Year vs Selling Price**

This chart visualizes the correlation between the year and the selling price. As seen in Figure 7, the year and the selling price are positively correlated. This means that they will both increase together. As shown in the scatterplot, as the year increases, the selling price also increases.

A graph of a sales growth

Description automatically generated with medium confidence

Figure 7. Year vs Selling Price Chart

**2.8. Condition vs Selling Price**

This next chart shows the correlation between the condition of the vehicle and the selling price. The number that indicates the condition of the vehicle is unclear. There was no documentation stating what the numbers meant. I am unsure if the condition of 1 is good or bad. As shown in Figure 8, the selling price of a vehicle tends to increase when the condition increases. Again, it is unclear if the increase in the number of the condition means that the vehicle is in better or worse condition but, the chart shows that there are higher selling prices in higher conditions which leads me to believe that the higher number of conditions means that the vehicle is in better condition.

**A graph of blue dots

Description automatically generated**

Figure 8. Condition vs Selling Price Chart

**2.9. Average Mileage vs Selling Price**

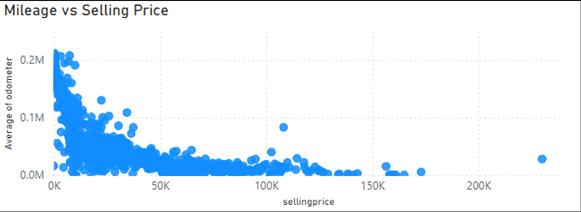
**** This chart visualizes the correlation between the mileage and the selling price. With the mileage, I had to use the average of the mileage for the visualization to show a clearer picture. Figure 9 shows that the mileage and the selling price are negatively correlated. This means that as the mileage lowers, the selling price increases.

Figure 9. Mileage vs Selling Price Chart

**2.10. Average Mileage vs Year**

This next chart shows the correlation between the average mileage and the year of vehicles. Figure 10 shows that there is a negative correlation between the mileage and the year of vehicles. As the year increases (newer vehicles) the vehicles have lower mileage.

**A graph showing the growth of a company

Description automatically generated**Figure 10. Avg Mileage vs Year Chart

**2.11. Average Mileage vs Condition**

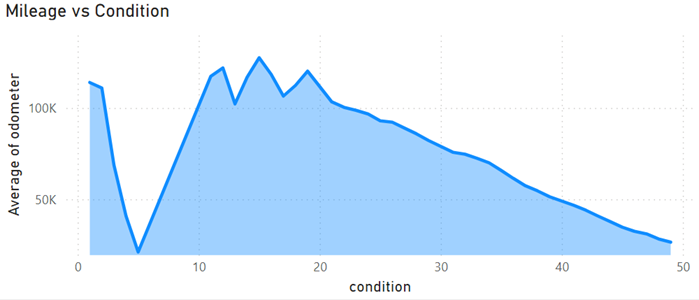
This last chart shows the correlation between the average mileage and the condition of a vehicle. This is another case of negative correlation. Figure 11 shows that the higher the number of condition the vehicle is in, the lower the mileage it has. This is another indicator that shows that the higher the number of condition, the better the vehicle’s condition is.

Figure 11. Avg Mileage vs Condition Chart

1. **Machine Learning**

This section will discuss the results of the machine learning algorithm.

**3.1 Feature Importance**

Feature importance is an important step in the machine learning process. It is used to calculate the score of the variables inputted to show the significance of the importance of those variables. I used the XGBoost algorithm to achieve the scores for this feature importance. I tried using random forests first but, the results were not good. Figure 12 shows the scores of each variable and it looks like make, body\_simple and odometer are important in the decision-making process.

A graph with blue bars

Description automatically generated

Figure 12. Feature Importance Chart

**3.2 Machine Learning Results**

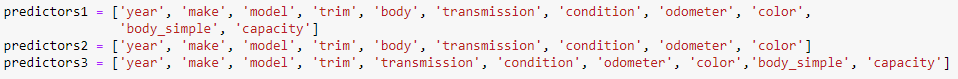
 Figure 13 below shows the variables used as predictors in the machine learning algorithm. This test uses supervised learning methods which indicates that a target variable is required. The target variable in this case is the selling price. I wanted to see the effect of different variables on machine learning algorithms. I tested a total of five algorithms. They are Random Forests, XGBoost, Decision Tree, Gradient Boost, and Logistic Regression.

Figure 13. List of Predictors

A screenshot of a computer

Description automatically generatedOnce I have input the predictors into the machine learning algorithms, the outcome is as follows:

Figure 14. Prediction Accuracy

As shown in Figure 14, each machine learning algorithm produced different accuracies depending on the predictors. Across the board, the ranking of these machine learning algorithms stayed the same. Overall, the Random Forest algorithm had the highest accuracy using all the variables (Predictors 1) in the dataset with an accuracy of 95.3%. It also has the highest accuracy for Predictors 2 and 3 with 94.8% and 95.2% respectively. Through this testing, we can conclude that all variables including the new variables are important and the Random Forest algorithm is the best to use to predict the selling price of vehicles.

1. **Recommendations**

This section will discuss the recommendations gathered from the analysis. The first three recommendations may be classified as one recommendation but, that would leave out the option of choice. If we were to only sell vehicles from Ford, Chevrolet, and Nissan from the years 2011-2014 and in good condition, that would ignore any variety of choices for our consumers and may hurt our sales. There is an ethical concern we should look at and it has to do with upcharge. We should be careful with the amount of upcharge we have on newer, used vehicles as it can tarnish relationships with frequent customers and jeopardize business.

**4.1.** **Sell more vehicles from Ford, Chevrolet, and Nissan**

As we can see from Figure 1, Ford, Chevrolet, and Nissan are the top 3 most popular manufacturers that were sold. On average, these manufacturers cost less than the European vehicles that were also sold. This tells us that consumers prefer more affordable low to mid-end vehicles from American and Japanese brands compared to more luxurious and high-end European vehicles. Not to say that we should not sell these European vehicles, a variety of choices would still benefit us but, more resources should be put into purchasing and selling American and Japanese manufactured vehicles.

**4.2. Sell more vehicles from the years 2011-2014**

Figure 3 shows us that the most popular years of vehicles sold range from the years 2011 to 2014. This tells us that consumers are interested in relatively newer, used vehicles compared to older, used vehicles. These newer used vehicles will also most likely have lower mileage according to Figure 10. Newer, used vehicles can also benefit us as they can be sold for a higher price which increases our return.

**4.3. Sell more vehicles in good condition.**

A vehicle’s condition is also an important factor in the consumer’s decision to purchase a vehicle. As seen in Figures 6 and 8, a vehicle’s condition is correlated with its selling price. The better the condition, the higher the selling price. With better condition vehicles, consumers may also feel like the money spent on the vehicle would be more worthwhile than an old damaged vehicle.

**4.4. Use Random Forests for the selling price predictions.**

The last recommendation I have is to use the Random Forests algorithm for the most accurate selling price predictions. A 95.3% accuracy score is quite high for a machine-learning model. It is also important to remember that all variables are important to use in the machine-learning model.

1. **The Three Ws**

In this section, I will talk about my final thoughts after finishing the project.

**5.1. What went well:**

A few things went well while doing the project. I continued to learn how to use Python and PowerBI. I think it is always good to keep practicing your skills. Another aspect that went well was that the project went smoothly. I did not run into any major issues that could have slowed down my progress. The dataset was easy to work with but required some quick preprocessing. The last thing that went well was the overall results of the analysis. I was able to get some insightful information from the visualizations and the machine-learning algorithms provided excellent accuracy.

**5.2. What did not go well**:

There were a few things also did not go well. The main problem I had was not being able to use multiple datasets. I did try to find a dataset with multiple files but I was unable to find a place to find it. This is due to the combination of not knowing how to search for it and where to find it. I had an idea of using files from a database with multiple datasets but again, I was not sure where or what to look for. In the end, I had to use a single dataset. Looking for a topic was also difficult. I was not sure what topics were good for analysis. Similar to finding a dataset, I was not sure what to look for in a topic. I went with the safe route and used a sales dataset. Processing times were also another issue. The original dataset contained over 550,000 rows. My desktop computer took way too long to process all that data.

**5.3. What I would do differently:**

There are a few processes that I would do differently. One of them is to choose a more complex topic. I think being able to analyze a topic that I am not familiar with can help me improve my analytical skills. Another thing I would do differently is to use multiple datasets. I would like to practice using multiple datasets for analysis. I am interested in how different an analysis can be with multiple datasets. I would also have liked to use a different visualization tool like Tableau. I was not able to use Tableau this time because my student license has expired and I did not want to purchase a license just to use it once.