STAT 167 - Car Price Prediction Project

2023-06-10

Team Name: The Special Squad

Link to data set: https://www.kaggle.com/datasets/tunguz/used-car-auction-prices (https://www.kaggle.com/datasets/tunguz/used-car-auction-prices)

Link to github repository: https://github.com/RyanSolanki/STAT167_Final_Project_Spring_2023 (https://github.com/RyanSolanki/STAT167_Final_Project_Spring_2023)

Introduction and Project Description:

```
# Load car data
set.seed(167)
car_data <- read_csv("car_prices.csv")
```

```
## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```
## Rows: 558837 Columns: 16
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (11): make, model, trim, body, transmission, vin, state, color, interior...
## dbl  (5): year, condition, odometer, mmr, sellingprice
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Drop NA values in dataset
car_data <- car_data %>% drop_na()

# Remove unnecessary variables
car_data <- subset(car_data, select = -c(vin, seller, saledate))

head(car_data,5)
```

```
## # A tibble: 5 × 13
##    year make  model      trim  body  transmission state condition odometer color
##   <dbl> <chr> <chr>      <chr> <chr> <chr>        <chr>     <dbl>    <dbl> <chr>
## 1  2015 Kia   Sorento    LX    SUV   automatic    ca          5      16639 white
## 2  2015 Kia   Sorento    LX    SUV   automatic    ca          5       9393 white
## 3  2014 BMW   3 Series   328i… Sedan automatic    ca          4.5     1331 gray
## 4  2015 Volvo S60        T5    Sedan automatic    ca          4.1    14282 white
## 5  2014 BMW   6 Series … 650i  Sedan automatic    ca          4.3     2641 gray
## # ℹ 3 more variables: interior <chr>, mmr <dbl>, sellingprice <dbl>
```

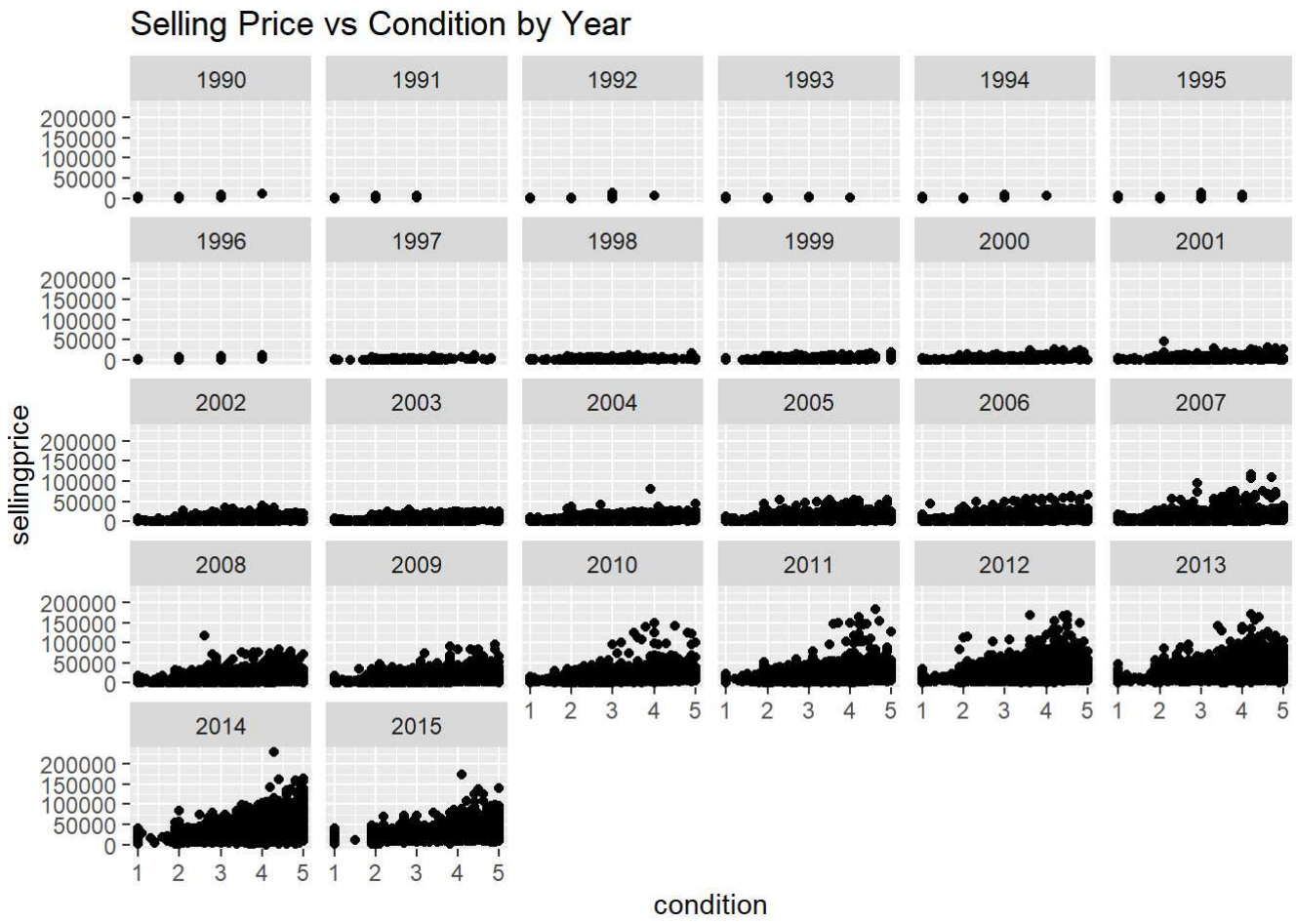
The data set on used car auction prices encompasses a comprehensive range of historical data, spanning from 1982 to 2015. Within this dataset, a remarkable total of 558,837 observations are available, featuring 16 key attributes. These attributes include details such as the year, make, model, trim, body type, transmission, vehicle identification number (VIN), location/state, condition, odometer reading, color, interior features, seller information, Manheim Market Report (MMR), selling price, and sales date. Our project objective and main reason that we chose this data set was to summarize and understand which car attributes had the most impact in selling price within the United States. In addition, we had a couple hypotheses that we wanted to answer:

1. Which types of cars had the highest/lowest selling price?
2. How does a car's condition affect the selling price?
3. How does a car's market price compare to its selling price?

Data Exploration and Visualization:

EDA #1:

```
#shows selling price vs condition for each year
ggplot(data = car_data) +
  geom_point(mapping = aes(x = condition, y = sellingprice)) +
  facet_wrap(~ year) +
  labs(title = "Selling Price vs Condition by Year")
```



Between 1990 and 2000, the condition of a car had a negligible impact on its selling price. However, as time progressed, a noticeable positive correlation emerged between the condition of a car and its selling price. Cars were categorized based on a scale from 1 to 5, with 1 representing the poorest condition and 5 indicating the best. Remarkably, vehicles in excellent condition (rated 5) commanded the highest selling prices, while a decline in condition corresponded to a decrease in the selling price as well.

EDA #2:

```
#adds column called diff1 that calculates the differences between selling price and MMR
#selects top 10 differences by make of the car
car_data1 <- car_data
car_data1$diff = car_data1$sellingprice - car_data1$mmr
car_data1 <- car_data1 %>%
  group_by(make) %>%
  arrange(desc(diff)) %>%
  summarise(diff1 = mean(diff)) %>% head(100000)

#plots top 20 differences by make
ggplot(data = car_data1) +
  geom_point(mapping = aes(x = diff1, y = make)) +
  labs(y = "Make",x = "Difference Between Selling Price & MMR", title = "Make vs Diference in Se
lling Price & MMR")
```

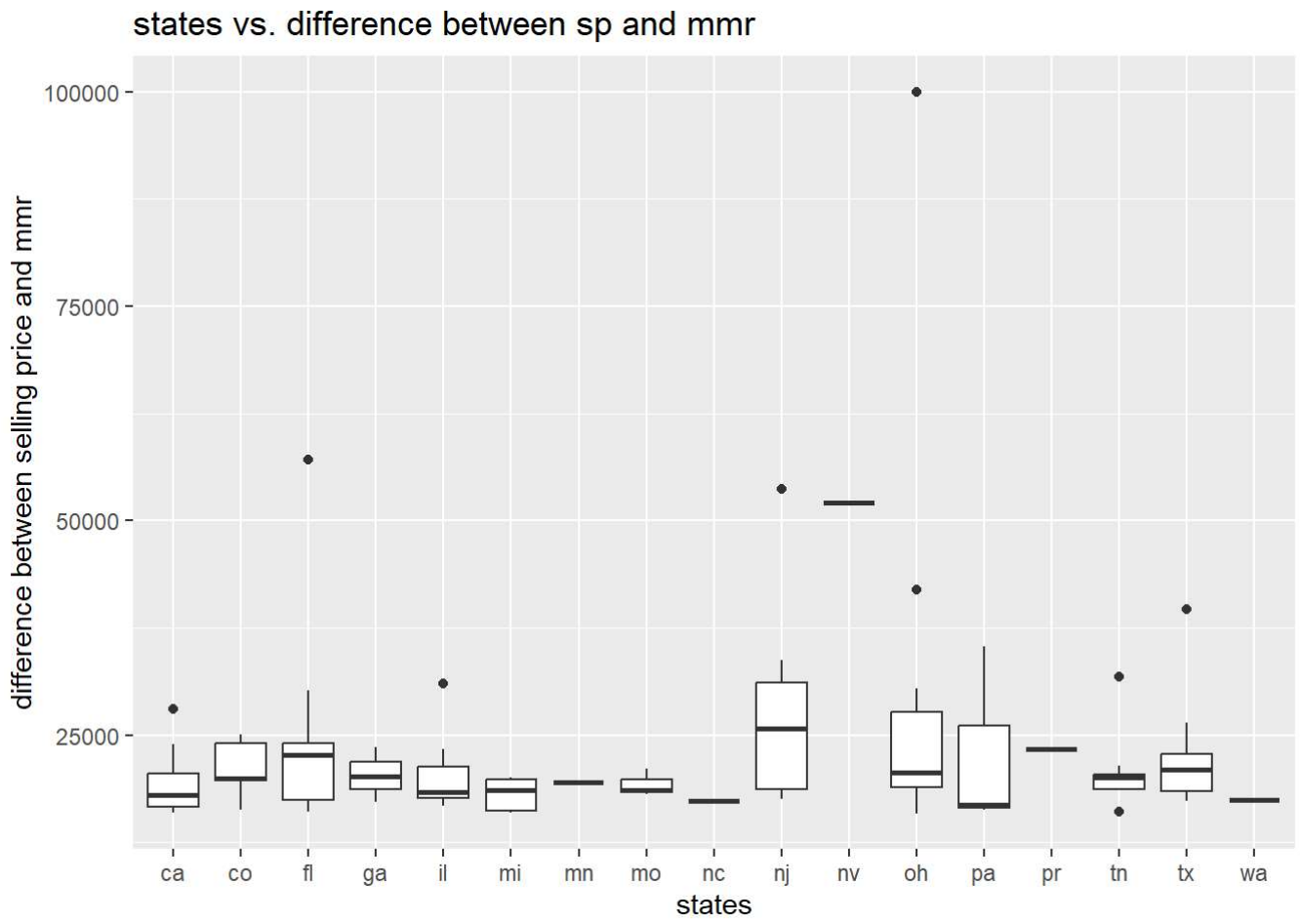
We have introduced a new attribute to capture the disparities between the selling price and MMR. Through our analysis, we discovered that luxury car brands like Fisker, Ferrari, Bentley, and Aston Martin exhibited notable distinctions. Aston Martin emerged as the car brand with the highest positive difference between the selling price and MMR, indicating a higher selling price than estimated. Conversely, Fisker, the lowest-ranked car brand in this regard, displayed the most significant negative difference, suggesting a selling price lower than the estimated value.

EDA #3:

```
car_data <- mutate(car_data, diff = sellingprice - mmr)
difference_sp_mmr <- car_data %>%
  group_by(make) %>%
  arrange(desc(diff)) %>%
  head(100)

ggplot(data = difference_sp_mmr) +
  geom_boxplot(mapping = aes(x = state, y = diff)) + labs(x = "states", y = "difference between
selling price and mmr", title = "states vs. difference between sp and mmr") + ylim(low = NA, hig
h = 100000)
```

```
## Warning: Removed 1 rows containing non-finite values (`stat_boxplot()`).
```

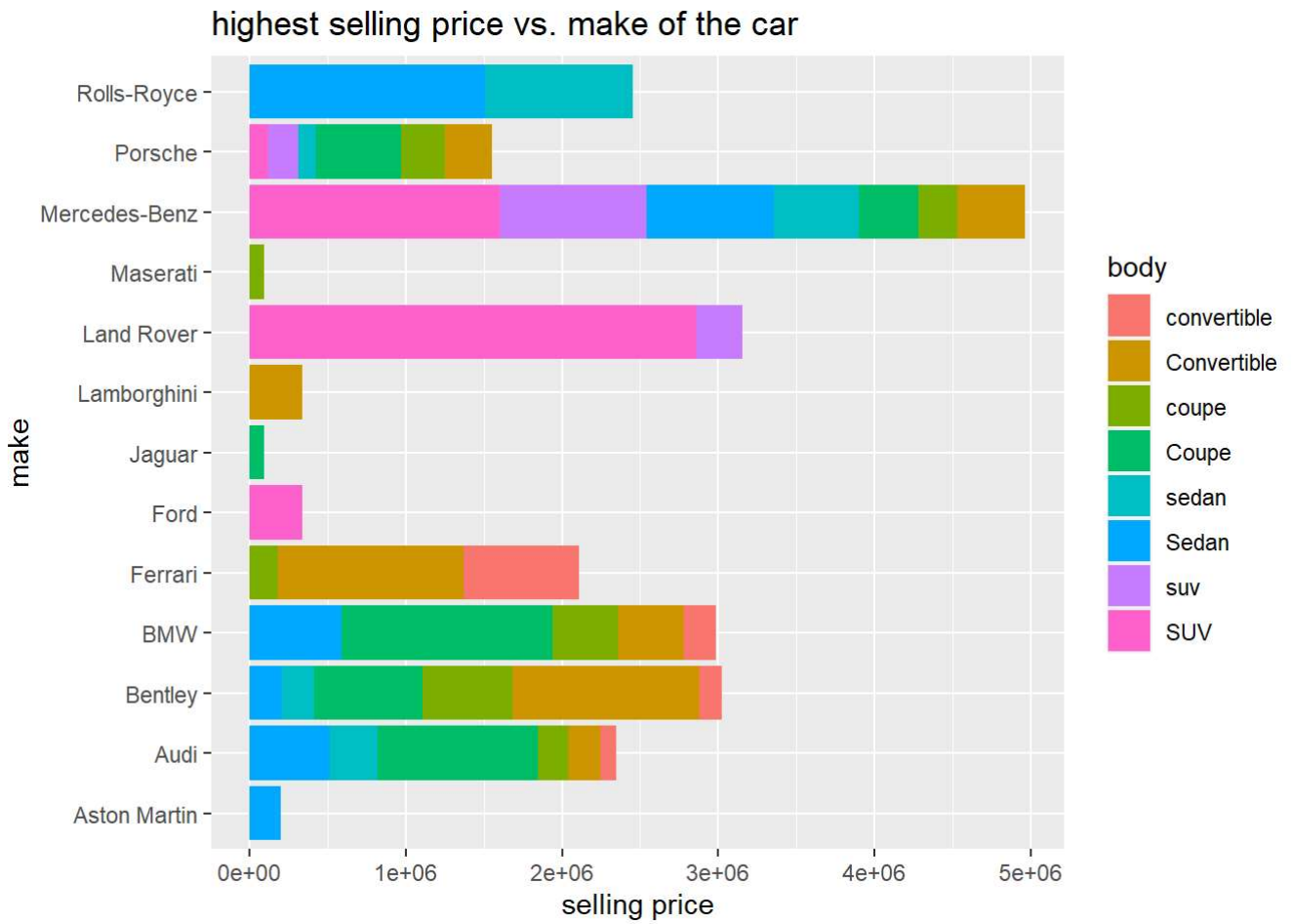


Since our main variable to measure is car prices, the difference/disparities between the selling price and MMR might be a key figure in telling us more about our key variable. When comparing the difference between selling price and MMR, there came the question of whether this difference was equal for all states that were measured in the United States. Using a box plot, it was found that this was not true. The states of Nevada, Ohio, Pennsylvania and New Jersey had the highest disparity among all the states in the data set. In the special case of Missouri, the median difference was quite low but its maximum value was among the highest in the set.

EDA #4:

```
selling_prices_high <- car_data %>%
  group_by(make) %>%
  arrange(desc(sellingprice)) %>%
  head(200)
selling_prices_high <- selling_prices_high %>% drop_na()

ggplot(data = selling_prices_high) +
  geom_col(mapping = aes(x = make, y = sellingprice, fill = body)) +
  labs(x = "make", y = "selling price", title = "highest selling price vs. make of the car") + c
oord_flip()
```



Furthermore, comparing the selling price of the car with the make of the car reveals an expected but also telling idea about how specific brands will be related to high selling prices while others are related to low selling prices. After arranging the data set by the selling prices and choosing the top 200 results, it revealed the makes of cars such as Mercedes, Bentley, BMW and Rolls-Royce were among the highest prices. While on the other hand, after arranging the data set in the opposite direction, we found the makes of cars such as Toyota, Chevrolet and Ford to be the most consistent.

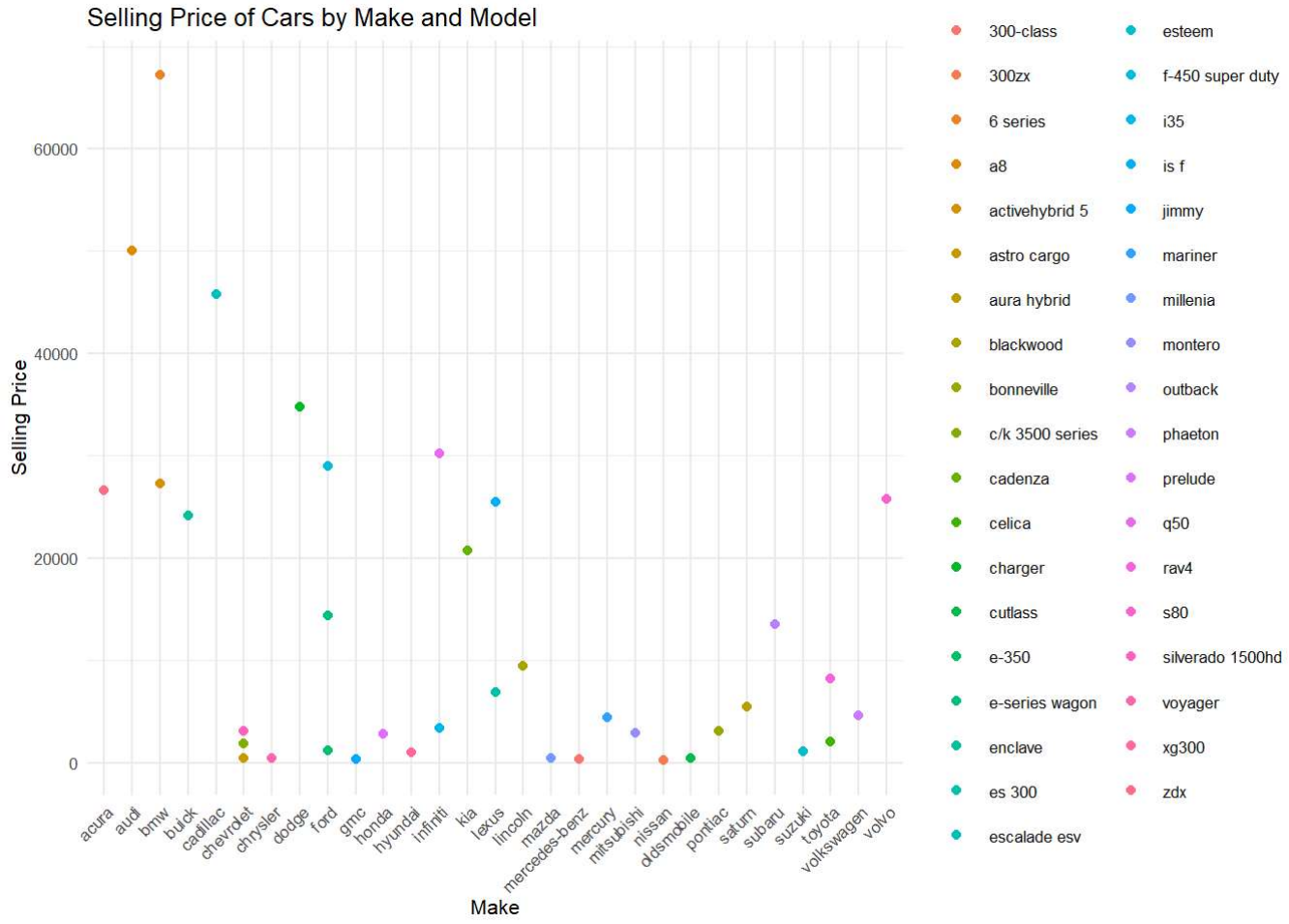
EDA #5:

```
# Code for selling price vs model of car bar plot
# Dot Plot Showcasing Selling Price Disparity Between Various Car Makes & Their Models
dataset_no_duplicates <- car_data
dataset_no_duplicates$make <- tolower(dataset_no_duplicates$make)
dataset_no_duplicates$model <- tolower(dataset_no_duplicates$model)

#remove duplicate rows to only contain singular makes
dataset_no_duplicates <- dataset_no_duplicates[!duplicated(dataset_no_duplicates$model), ]

#Stratified sampling due to large dataset size
smaller_dataset <- stratified(dataset_no_duplicates, "make", size = 0.05)

#Relationship between selling price and a car's make and model
 ggplot(smaller_dataset, aes(x = make, y = sellingprice, color = model)) +
  geom_point() +
  labs(x = "Make", y = "Selling Price", title = "Selling Price of Cars by Make and Model") +
  theme_minimal(base_size = 8) + # change the base_size to a smaller value
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

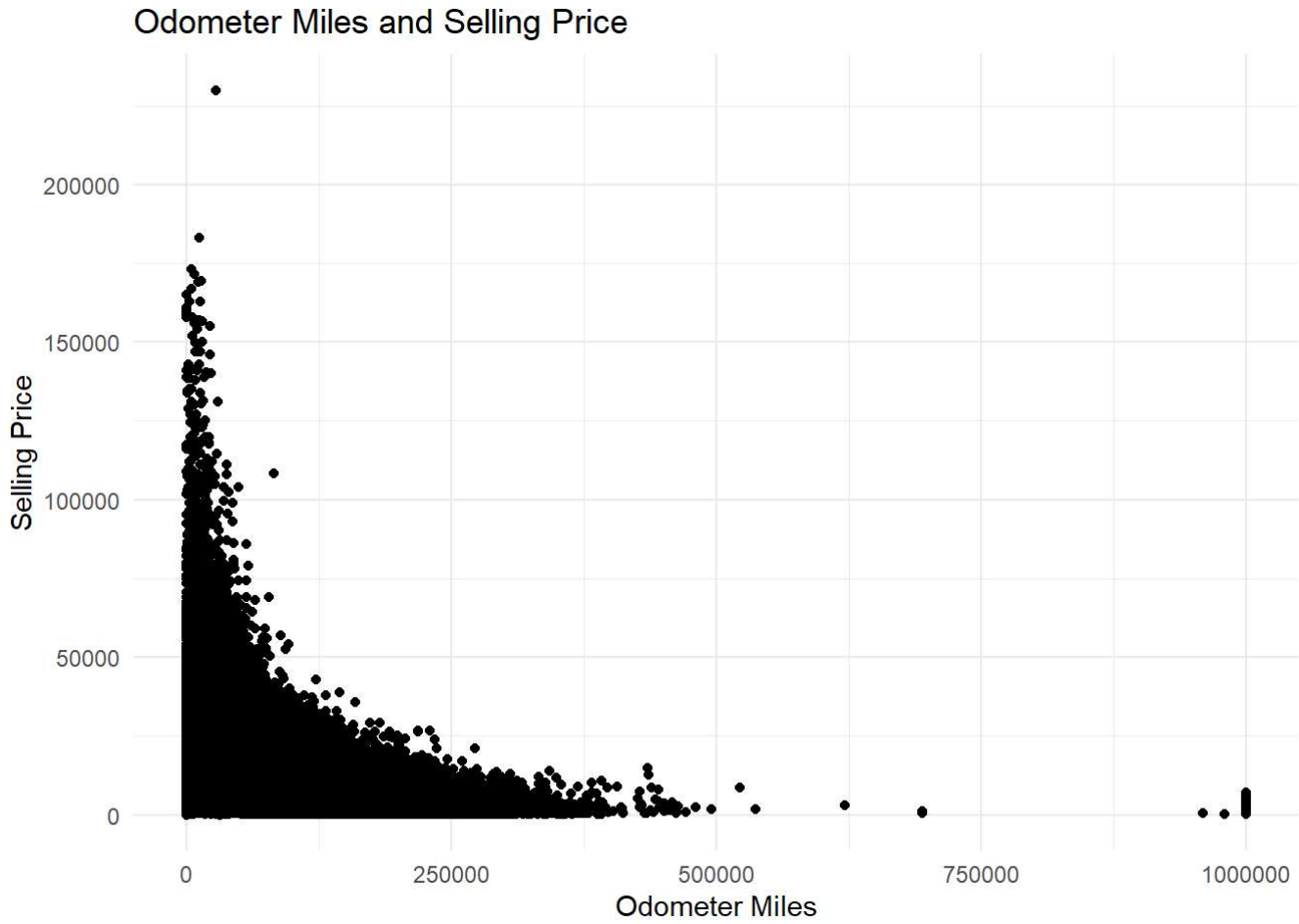


Following suit with EDA #4, we wanted to better visualize and showcase the disparity between certain car brands among others in accordance to their prices. We all know the notion of price disparity between brands but just how glaring is this difference? Furthermore, while the existence of price difference is clear among car brands, the same cannot be said for the plethora of models within these brands. Which model is cheaper, which is more expensive, the knowledge of this concept is often unknown to the average consumer. Through the integration of a color coded

dot plot, we are now able to visually see the disparity between these relations. Akin to common knowledge, brands such as Acura and Lexus are more expensive than their cheaper alternatives Honda and Toyota and sportier models fetch a far higher price within each brand in comparison to their sedan counterparts.

EDA #6:

```
#Relationship between odometer miles and selling price
ggplot(car_data, aes(x = odometer, y = sellingprice)) +
  geom_point() +
  labs(x = "Odometer Miles", y = "Selling Price", title = "Odometer Miles and Selling Price") +
  theme_minimal()
```



Another piece of common assumption we wanted to visualize was the relation between odometer mileage and the cars final sale price. We all know the notion of wear and tear, with a car's price decreasing in correlation to its usage's increase. To showcase this relation, we chose to create a scatter-plot to better present this trend. As the graph shows, the car starts at an extremely aggressive dip before slowly taming out at the 125,000 mile mark. From this plot we can assume that a car's initial odometer mileage is far more important than its added miles later on, hinting at an importance of the newness of a car.

Data Analysis, Modeling, Predictions:

Utilizing multiple linear regression of the scaled data, we employed a forward step-wise selection approach in our analysis to predict the selling price of used cars at auction. This method involved starting with an empty model and gradually incorporating variables one by one. In each forward step, we selected the variable that offered the most significant enhancement to our model. By incorporating all fifteen attributes, our multiple linear regression model

achieved outstanding results, boasting an impressive R-squared value of 0.9733. Furthermore, the adjusted R-squared value also stood at 0.9733, demonstrating the robustness and reliability of our model. Notably, this model showcased the lowest mean squared error (MSE), indicating its superior predictive accuracy.

Model Evaluation and Validations:

To ensure robust and reliable model evaluation, we employed K-Fold Cross-Validation, a technique that partitions the data set into ten approximately equal-sized groups. In each iteration, one fold is designated as the holdout set, while the model is trained on the remaining k-1 folds. This process is repeated ten times, with a different set acting as the holdout in each iteration. By calculating the test Mean Squared Error (MSE) on the observations within each fold, we obtained a comprehensive assessment of our model's performance. Overall, the average of the ten test MSE values, which was approximately 0.0267, served as a reliable indicator of our model's predictive accuracy and generalizability.

Conclusion:

After completing our project, we can draw some conclusions about the hypothesis we originally had. We can now say that the market value of a car has the highest correlation to it's selling price (small difference between market value and selling price for most cars). We also found that the newer a car is, the more its condition affects the selling price. From our exploratory data analysis, we saw that sports and luxury cars sell for the highest prices, and vans, wagons, and SUVs sell for the lowest prices.

As a group, we were also able to take away a couple of learning lessons from this project. One of the things we learned was that your first data set or model might not work very well, but that is okay. This happened to us, and we were able to explore other options until we found a new data set that worked for us. Another important thing we learned is that it is important to normalize your data before evaluating your model. We did not do this at first and were puzzled because we had such a large MSE value for our k-fold cross validation. After normalizing the data, this issue was resolved.

Team Contributions:

Ryan - Contributed to helping find the dataset, data cleaning, multiple linear regression model, regression graph.

Jake - Helped with data cleaning, multiple linear regression, and data validation using k-fold cross validation

Harrison - Contributed to creating EDA's, adding stepwise regression, and performing cross validation.

Girum - contributed to the EDA's

Shuqiao - contributed to the EDA's