What's This Car Worth? Analysis of Used Cars Data on Ebay

Summary

This project explores a dataset of used car sales on eBay. The primary goal of the project is to use the data available to determine which variables are the best predictors of selling price and how long it takes a used car to sell, and create a model that will allow us to reliably predict those outcomes based on available information. Our analysis identifies associations between predicting and outcome variables, but does not produce a robust predictive model.

Table of Contents

Introduction

Overview and Reason for Study:

Our group has chosen to explore a data set of used car sales on eBay. The members of our group all live in Atlanta and rely on cars to get around the city, and have recently purchased or have considered purchasing a used car. Therefore, we think it would be interesting to understand which factors best predict the price of a used car and how long it might take to sell on a website like eBay.

We will perform our analysis using Kaggle's "Used Car Database," which contains about 370,000 observations of used car advertisements posted on the German eBay site from March 2015 to April 2016. Each observation in our dataset is one instance of a used car that was sold on the site, and includes various attributes such as price, post date, post removal date, whether the car requires repairs, its make, its model, and other details about the car. This set of factors may allow us to analyze whether there is a relationship between these predicting variables and an outcome variable like selling price, and use statistical inference to provide information that could be useful to individuals who are considering buying or selling a used car on a website like eBay. To this end, we used this dataset to explore five questions:

- Question 1: What are the most important predictors of the price of a used car?
- Question 2: Given the same year and model car, how much could having damage significantly impact its selling price?

- Question 3: When posting a used car for sale, how does the post title impact its selling price?
- Question 4: Given a set of factors about a car, can we reliably predict what its selling price will be?
- Question 5: Given a set of factors about a car, can we predict how long it will take to sell on the site?

Expectation:

We expect that we will be able to identify a set of predicting factors that are important when explaining the price of a used car, and that we can therefore predict the selling price of a used car based on a set of known factors about the car. We also expect that we will be able to predict how long a car will be up on the website, based on similar factors.

Raw Data:

Although there are 20 factors included in the raw dataset, we make modifications to the data for our analysis by both adding new columns that impute values based on the data, and by deleting columns that are not relevant for our study. For example, in the column *Seller Type*, more than 99% of the entries are "private" while very few are "professional." For the purposes of this analysis, we want to focus on private sellers; therefore, we delete the observations that are "professional" and then are able to ignore the column, as all its values are now the same. There are also columns in the dataset that are not relevant for our study, such as whether or not the observation was used in internal A/B testing. For our analysis, we omit these irrelevant factors to reduce the size of the dataset.

Likewise, we create a few new columns that we anticipate will be useful for our analysis. This includes calculating the duration time that the advertisement was posted, calculating the age of the car, and measuring some aspects of the advertisement title. These new columns will be explained in more detail in the Questions and Methods portion of the paper. By creating these columns and imputing values based on the other values for the observation, we are better able to explore answers to our five questions.

We also observe that there are many observations with missing values in our data set, and choose to omit these observations entirely. This reduces our data by over 20%, and leaves up with a dataset of almost 250,000 observations. While plenty of data remains with which to perform analysis, we acknowledge that this may increase existing bias in our data. For example, if many sellers choose to omit price for luxury or antique cars, or perhaps chose not to report damage on a low-priced car, this may lead to biased data. However, since we have a lot of data cannot know the extent of this bias without further investigation into the data collection, we find this potential tradeoff acceptable for the purposes of this study.

Questions and Methods

Overview:

We have established five questions that we would like to answer with our data. For each question,we will perform an exploratory data analysis (EDA) to look at general trends, correlations, and relationships in the data. At this point, we will identify any data that we might consider removing, such as outliers or collinear factors. We will also use this phase of the analysis to brainstorm potential models we may use for our analysis, and identify any transformations that may be needed. When determining which

model to use, we will consider whether a model is necessary, or if the question is already sufficiently answered in the EDA portion.

Before we begin the investigation, we will create a few new columns that we think may be relevant for our five questions. First, we create a *duration* column that indicates how long it took a car to sell on the site, which we make by adding a new column that measures the time between when a post was created and when it was deleted. This approach makes the assumption that a post was taken down when the car sold, and that may not always be the case. However, for the purposes of this study, we will work under this assumption.

Second, we estimate the age of the car by subtracting its year of registration from the year the data was scraped. This new column will be *age* of the car in years. In many cases, this may be collinear with other factors, such as mileage or horsepower, and this should be noted in any statistical inference. However, since antique and specialty cars are also listed in this dataset, we do not believe that age of the car should be strictly collinear with any other factors, and it may be a significant predictor in our models.

Additionally, we want to include information about the title of each advertisement to see if it has qualities that are predictive in our models. For example, perhaps advertisements with long, descriptive, emphatic titles are associated with quicker or higher sales. Alternatively, we may find that this is not the case, and in fact the shorter post titles are associated with quicker sales or higher prices. We may also find no relationship, but we are curious to explore this as a potential predicting factor and will create new columns to include in our dataset. To do this, we used regular expression pattern matching in R to create a column with the word count of the post's title included, and an additional binary column that indicates whether or not the title includes one or more exclamation points; these columns are titled *word_count* and *exclamation*, respectively. We note that although only 2% of the observations include one or more exclamation points, we will keep this factor in the dataset to see if there is a relationship between this and any other factors.

Finally, before beginning our EDA, we choose to exclude the factor *model* from the analyses. In the cleaned up data, there are 250 unique car models present in the dataset, which we believe is too many to be useful if included as a categorical variable in our analyses. Instead, we include only the categorical variable *brand*, which we believe will include enough information about the style of the car without the extreme detail of its specific model.

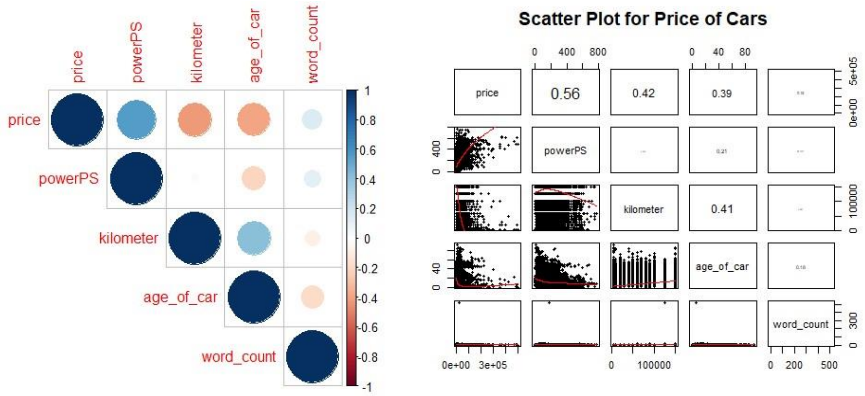
Question 1: What are the most important predictors of the price of a used car?

In order to look at this question, we first considered which factors might be useful to include as potential predictors of price. We decide to exclude the column *duration* from our EDA on this question, as we believe it may be partially dependent on price and may therefore modify our estimated coefficients and statistical inference. We also decide to investigate continuous and categorical variables separately in the EDA.

Once we have selected and grouped the variables we will use in this portion of the EDA, we begin by exploring the correlation between the continuous factors in our dataset and the outcome variable *price*. As seen in Figure 1b, there is a moderate positive correlation of 0.56 between the factor *powerPS* (horsepower) and *price*, and a weak positive correlation between the factor *word_coun*t and *price*. Figures 1a and 1b also show a moderate indirect relationship between the factors *kilometer* and *age* with the outcome variable *price*. Although *word_count* only appears to have a weak correlation with *price*, we still want to include that factor in the model to see if it has a predictive relationship in the presence of other factors.

Although our intention was to explore correlation between the predicting variables and the outcome variable *price*, the visuals in Figures 1a and 1b allow us to see that there may be some factors that are correlated to each other as well. While there may be some evidence of correlation between these predictors such as age of car and kilometers travelled by car, the value is not large enough to for us to be too concerned about collinearity between predictors. To support this hypothesis, analysis of the variance inflation factor is performed, and it reveals no values above 2.2, well below the threshold of 10. Hence it is concluded that collinearity between predictors is not an issue for this study.

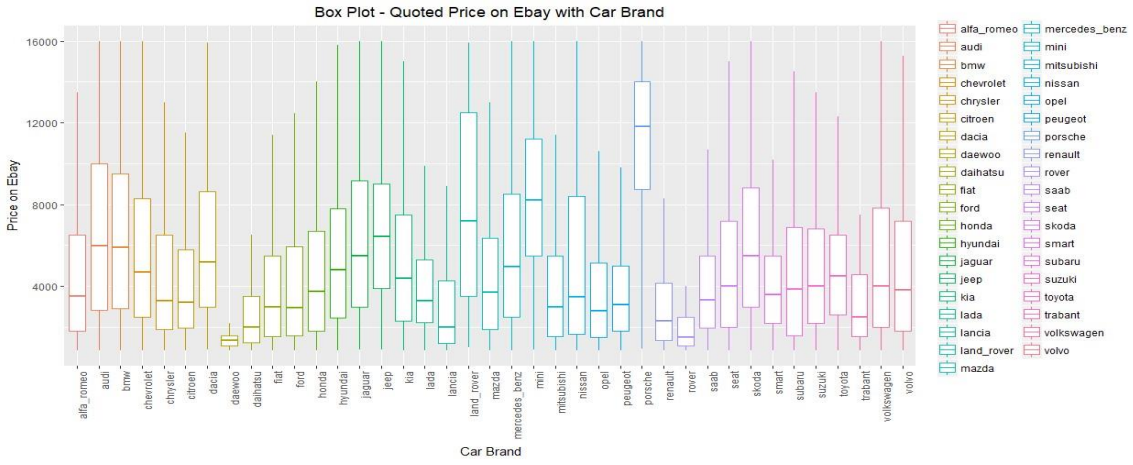
Figure 1a and 1b



Now that we have looked at the relationship between the continuous variables and the outcome variable *price*, we will next choose to explore the relationship between *price* and each of the categorical variables in the dataset. This includes *brand*, *vehicle_type*, *gearbox*, *fuel_type*, *exclamation*, and *not_repaired_damage* (a binary variable indicating if there is the presence of unrepaired damage on the vehicle). Although we will explore all of these factors, we will include details about the two factors that seem most important: the car's brand and the presence of unrepaired damage.

The box plot of *price* and *brand* shown in Figure 2 is highly detailed, as the category has 39 factors; yet, we believe it is important to explore this factor to better understand our dataset. In Figure 2, we see that luxury car brands have higher average prices, as expected; however, we see that the brand Porsche stands out significantly. We may consider removing this brand, or perhaps others, in later analyses.

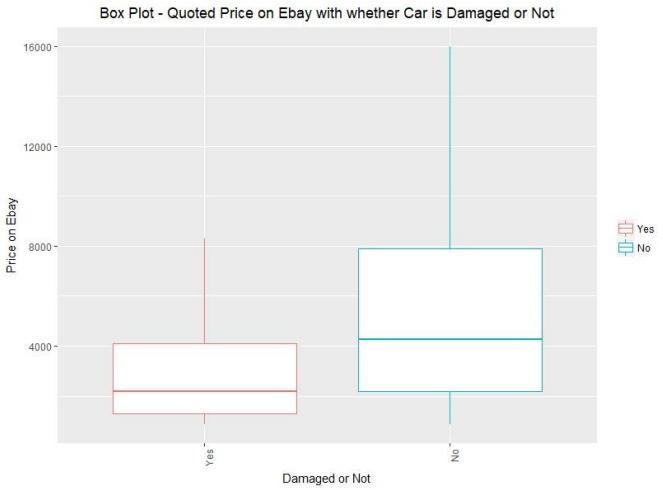
Figure 2



Based on the visual analysis in Figure 2, our group hypothesizes that at least one of the average prices between the different car brands is not equal to the rest. We ran an ANOVA, which confirmed that at least one of the means of these groups is not equal to the rest. We also included an analyses using the Tukey test, which showed more detail between which pairs of brands had the highest difference in means. Based on this, we anticipate that *brand* will have a significant effect in our predictive model.

We are also curious about whether the presence of unrepaired damage on the vehicle has an impact on the average price of cars, as we would expect. The box plot in Figure 3 shows a visible difference between the means of the two groups. Based on this EDA, we determine that we will further explore if this effect is significant in Question 2.

Figure 3



Based on the exploratory analysis performed on the continuous and categorical variables available, we conclude that all of these factors may be important in prediction, and determine that we do not want to eliminate any variables for consideration at this phase. Since we have both continuous and categorical variables, we will choose to include these factors in a Multiple Linear Regression to better determine their predictive power. This will be further discussed in Question 4.

Finally, based on the EDA for this Question, we identify a few opportunities for further analysis. Specifically, we anticipate that we may want to consider reducing the number of brands included, either by grouping the categories or omitting some brands (for example, omitting luxury car brands). We would need to perform two different analyses, one with these brands included and one without, to compare models and see how coefficients may be impacted.

Question 2: Given the controlling factors of age and model, does having damage significantly impact the car's selling price?

In our EDA, we saw that there appears to be a visible difference in average price between cars that have unrepaired damage and those that do not, as shown in Figure 3. Since the difference appears large, we want to perform an ANOVA to see if this effect is truly significant. For this ANOVA, the null hypothesis is that the average price of the two groups is plausibly equal, and the alternative hypothesis is that the two groups do not have equal average prices.

We chose to explore this by testing subsets of coefficients and computing the partial F-test. Our null hypothesis ($H_0 : \alpha_1$ versus $H_A : \alpha_1$ is not zero) is that the coefficients of the factors

notRepairedDamage = 0 and *notRepairedDamage* = 1 are the same, and our alternative hypothesis is that they are not equal. When we perform this test, we see that the F-statistic is 24198 and the p-value is significant at the 99% confidence level, so we conclude that we can reject the null hypothesis that the means of the prices of the cars in the two groups are the same.

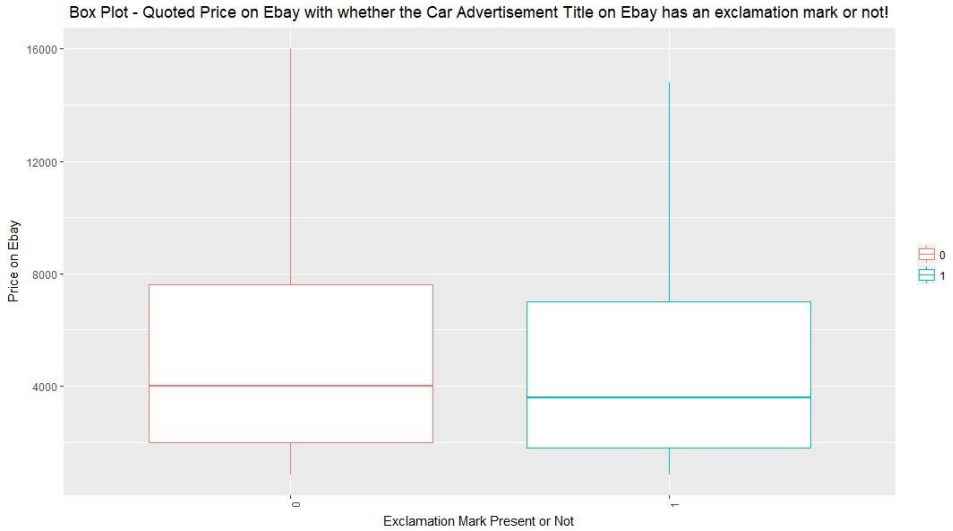
There may be reasonable opportunity to perform further analysis for this question, as it is possible that for some types of cars, damage may have a smaller or larger effect on the sale price. For example, cars that are sold as antiques may not exhibit this effect, and likewise, if cars are being sold for parts, one might see a different relationship between *price* and the factor *notRepairedDamage*. For particular cases, it may be worth analyzing this relationships further.

Question 3: When posting a used car for sale, how does the post title impact its selling price?

Since the title of the advertising post is one of the first things a potential buyer will see about a used car for sale, it may be important for sellers to make a good first impression with the title of their post. To this end, we want to explore our data from this angle to see if we can identify any patterns between the length of the post title and the selling price of the car. Likewise, we want to consider whether the presence of one or more exclamation points in the title may impact the selling price. In order to perform this analysis, we used the columns *word_count* and *exclamation*, which we created using regular expressions to indicate these characteristics in the post title.

When investigating the factor *exclamation*, the initial EDA suggests that there is possibly a significant difference in the mean price between cars that include an exclamation point in the ad title and those that do not, with those including an exclamation point selling for less on average, as shown in Figure 4. However, this box plot is not enough to draw statistical inference, so we will continue by performing an ANOVA. The ANOVA confirms that there is indeed a significant difference between the mean prices of the group of cars whose post titles contain one or more exclamation points and the group of cars whose titles do not. However, we also note that only 2% of the ~250,000 observations in our dataset include an exclamation point, so further analysis may be needed to see if this is indeed a factor that relates to the selling price of the car, or is perhaps an exaggerated effect due to some other bias.

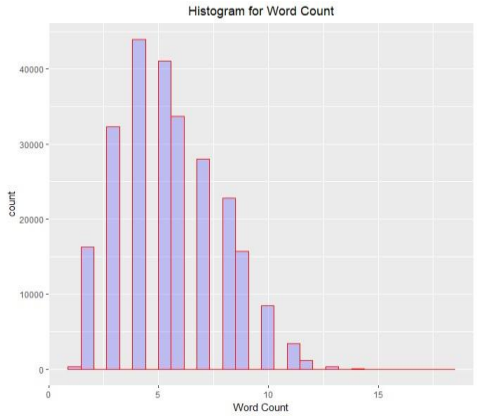
Figure 4



Although we included the factor *word_count* in the initial correlation portion of the EDA for Question 1, we will explore further for this Question. In summarizing the factor, we see that the average number of words per post title in our dataset is only 5.5. However, the count of words per title ranges

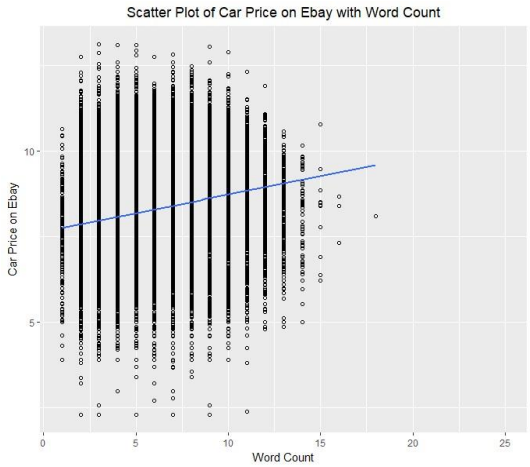
from 1 to 519. Upon further exploration, we see that the observation with a word count of 519 is an extreme outlier, and we choose to remove that particular point from our dataset. With this point removed, we looked at both the distribution of *word_count* and the scatterplot of its relationship with *price*. As seen in Figure 5, the distribution is skewed to the right, showing the majority of posts have between 1 and 10 words, with a long tail to the right.

Figure 5



In Figure 6, we see that the scatterplot of *word_count* and *price* aligns with our earlier observation of a weak positive correlation between these two factors of .15. Since this further exploration does not show that *word_count* has a strong linear relationship with *price*, we do not think it will be a very significant predictor of *price*.

Figure 6



We also investigated these factors to see if they have a relationship with the factor *duration*, but did not find a significant relationship between either *word_count* or *exclamation* and the length of time a post stayed online. Therefore we will not pursue a predictive model with these predicting variables and *duration* as the outcome variable.

We believe that further exploration and inference on this topic could potentially be informative to help sellers maximize their selling price or minimize how long it takes a car to sell; however, the analysis would need to be performed correctly, perhaps even with experimental design, to infer causation. It would be interesting to consider whether titles including certain keywords or car details received more attention and subsequently sold faster or for a higher price; however, we cannot infer this from observation alone.

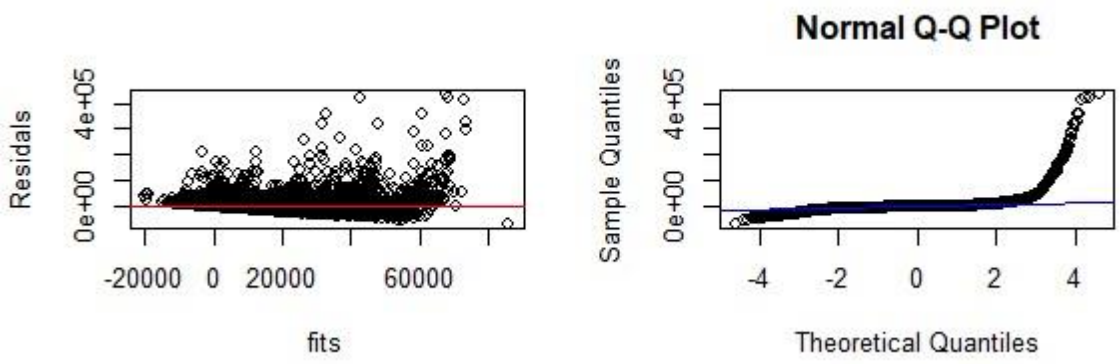
Question 4: Given a set of factors about a car, can we reliably predict what its selling price will be?

Based on our EDA, we would like to build a multiple linear regression model comparing the outcome variable *price* with all of the predicting variables that we have included in EDA so far. This includes the continuous variables *powerPS, Kilometer, exclamation, word_count, age* and the categorical variables *gearbox, brand, notRepairedDamage.*

We anticipate that we may need to perform model selection and consider various transformations of the predicting or response variables in order to create a model that has high predictive power while also being a good fit for the dataset. We will outline our approach below.

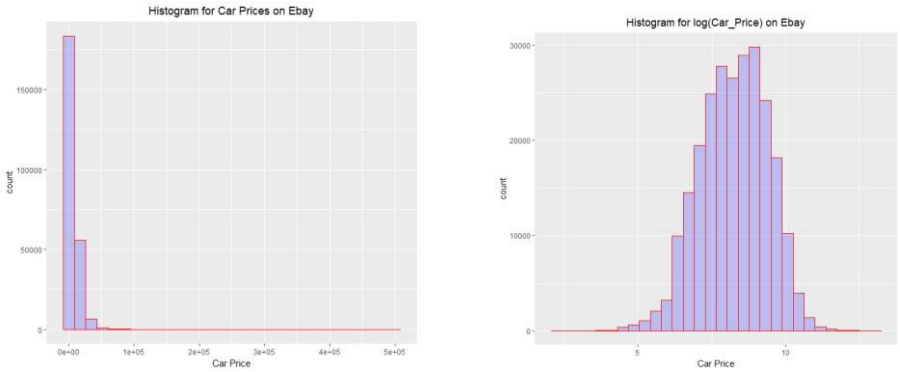
We initially perform the multiple linear regression with the factor *price* set against all the other variables, and achieve an R^2 of 56% with almost all predicting factors indicated as statistically significant. Our next step is to remove any outliers with a Cook's distance greater than 1; upon completing this step, we re-ran our model, and this did not improve our model's R^2. Even though the R^2 is not as high as we would like, we are still interested to see if the model is a good fit for the data by checking for violations of goodness of fit assumptions; this may give us more information about how to construct a good model. We perform the goodness of fit tests in Figure 7, and conclude that this model is not a good fit for our data because there appears to be violation of the constant variance and normality assumptions.

Figure 7



Because this model violates the goodness of fit assumptions and is therefore not appropriate for this data as is, we will next proceed with a transformation of the outcome variable, *price*. First, we will perform exploration to see whether a transformation of *price* may improve the fit of our model. Figure 8, the histogram of *price*, shows that it has a very long tail on the right, so we consider building a model incorporating the log(*price*) instead of the raw data. When we view the histogram of log(*price*) in Figure 9, it is approximately normally distributed, so we prefer this metric for our analysis.

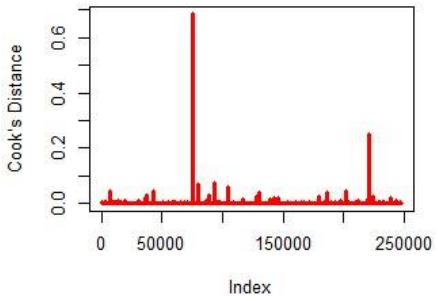
Figure 8 Figure 9



Replacing *price* with log(*price*) in our model increases the R2 to .6967, indicating that the predictive variables in the model explain about 70% of the variation of *price* in the dataset.

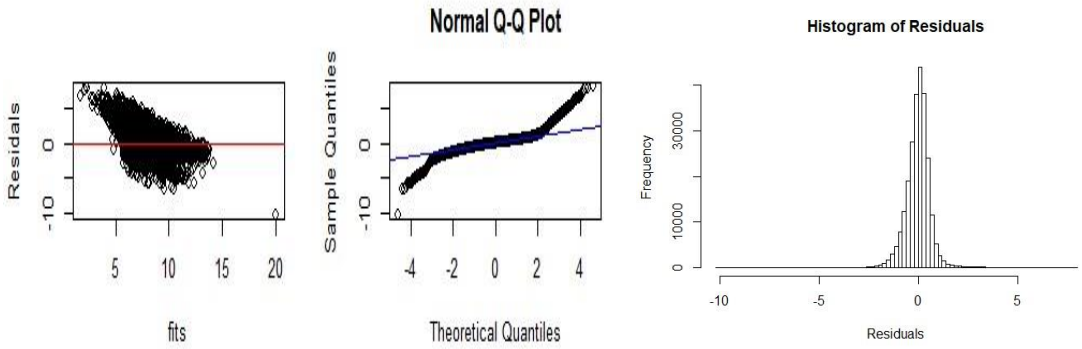
For our new model, we now need to check for outliers using Cook's distance again. We find that there is only one observation that has a Cook's distance greater than 1, so we will remove this point from the dataset and rerun our model to see if it is an influential point that may impact the results of the model. After rerunning our model without this point, we find that there is no difference between the R^2 values of the two models.

Figure 10



Moving forward, we will check the assumptions of the model of log(*price*) with outliers removed to determine its goodness of fit. Figure 11 shows the results of the various graphical tests for violations of goodness of fit.

Figure 11

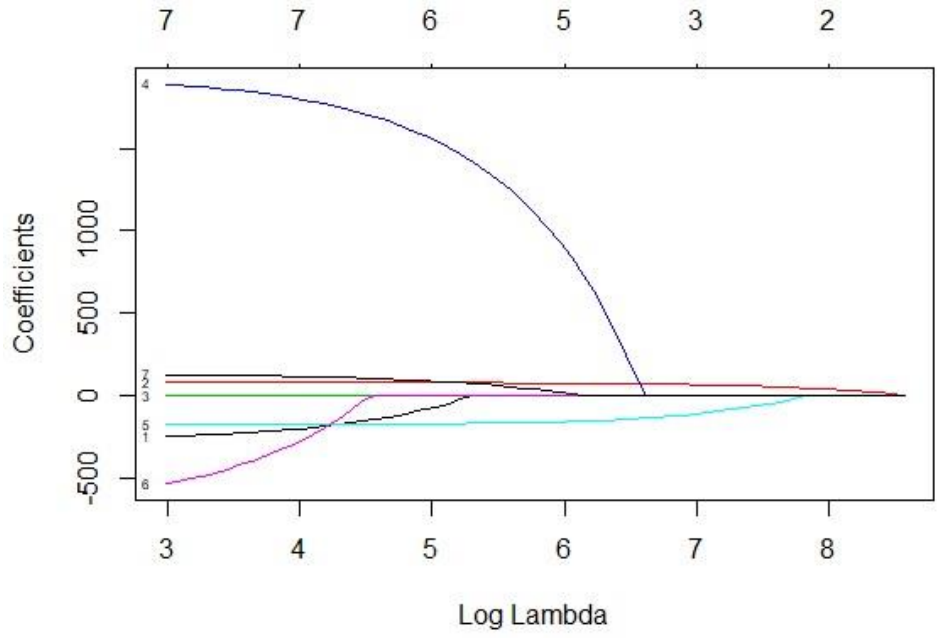


Upon reviewing the plots in Figure 11, we can see that the model using log(*price*) demonstrates a clear improvement in goodness of fit compared with the initial model that only used *price*, seen earlier in Figure 10. However, there are still some potential issues with constant variance and normality, so we will explore various transformations of y and variable subsets using variable selection to see if we can improve the fit of our model.

First, we implement the stepwise variable selection tool using AIC criteria to see if we can achieve model improvement by dropping any of our variables; however, this approach does not omit any of our predicting variables. Next, we used the *drop1* function in R to drop the variable that would reduce the AIC the most; this model dropped *gearbox*. Therefore, we may consider omitting this column from our model in future iterations.

At this point it would be interesting to see how important each of our variables are in determining the price of the car. We attempt a lasso model, to determine which variable enters the linear regression model at which stage. This can be seen in Figure 12.

Figure 12



The variables used in the above graph are 1) *gearbox*, 2) *powerPS*, 3) *kilometer*, 4) *notRepairedDamage*, 5) *age_of_car*, 6) *exclamation*, 7) *word_count*. From the above plot we can see that variable 2) *PowerPs* is the last to enter the model, this means that the power of the vehicle given all other predicting variables contributes the least in explain the variability in the linear regression model. The most important factors according to the lasso graph above are *gearbox* and number of kilometers driven by the car. Additionally, the presence of an exclamation mark in the title is one of the first variables to enter the model.

We will now attempt to transform our outcome variable *price* by using the Box-Cox transformation to see if this improves the fit of our model. Investigation with the dataset shows that the optimal lambda for the Box-Cox transformations is 0.1818. We explored this new model and found that while the conformity to goodness of fit assumptions are slightly better, the transformation brought our R^2 down to only 7.3% - a huge decrease in explanatory power which makes our model useless. Based on this outcome, we will not proceed with using this model to make predictions.

Our next approach involved scaling a few of the predicting variables before producing the model. We scaled numerical variables *kilometer, powerPS, duration,* and *age*. We then reran the model regressing our scaled predictors and our other categorical variables again log(*price*) and there was no improvement in our model. We also performed additional transformations of x, including log transformations of quantitative variables, alternative scaling, polynomial transformations, and various interaction terms.

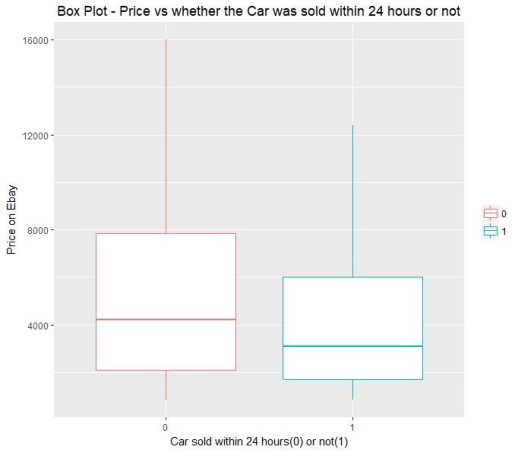
Finally, we would like to note that in our exploration of this question, we were unable to build a model that has both a good R^2 value but also does not violate the assumptions. We could plug in values to produce a "prediction," however we believe that they may not be reliable estimates of *price*. We would recommend further research, analysis and exploration of more variables to find a model with a better fit that best explains the variability in price of used cars on Ebay

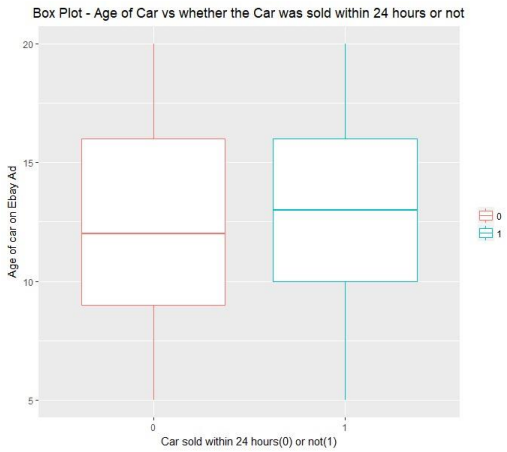
Question 5: What factors are associated with cars that sell in one day, versus cars that take longer to sell?

For our last question, we intend to explore if there are factors that are associated with cars that sell very quickly, so a seller may be able to predict how long it will take their car to sell. However, EDA showed that performing a linear regression between any set of factors and the outcome variable *duration* would not be promising, and initial models had very low R^2 and violated goodness of fit assumptions. However, we are still interested in the question as we think it may be interesting for both sellers and buyers, so we developed another approach.

We decided to explore *duration* as a categorical variable with two factors, rather than as a continuous variable, to see if this approach would exhibit a more robust relationship with other factors in the model. To do this, we split the observations into two groups: cars that sell in 1 day compared to cars that sell in more than 1 day. There are significant number of cars in our dataset that sell in 1 day or less, and we thought this would be an interesting point on which to split the data. We then created box plots comparing the factors of *age* and *price* between these two groups, seen in Figures 13 and 14 below, and further explored the significance of these relationships using ANOVA. Using these two tools, we see that there is a statistically significant difference in both the means of *price* and *age* of the two groups (those that sell in 1 day and those that sell in more than 1 day).

Figure 13 Figure 14





After performing additional EDA, we attempt to produce a linear regression model using any subset of predicting variables and the outcome variable *duration*. Although we created many models with different numbers of variables and transformations of both x and y, every model we tried had an R^2 between 0% - 10%, so we did not pursue any of those models further. We conclude that although there appears to be some relationship between *duration* and the factors *price* and *age* in the EDA, this relationship does not appear to hold when analyzed with regression. We would recommend further research to understand what factors might be more closely related to how quickly the car sells online.

Conclusion

Findings and Implications:

Our primary goal, to predict price of a used car based on various factors of the vehicle, is explored in the models created in Question 4. Our most promising model predicted log(*price*) against the other predicting variables and had an R2 of ~70%. However, this model violates the constant variance assumption, so we acknowledge that the estimated coefficients and any predictions made with this model may not be very reliable.

All of the variables included in this model are significant, with the exception of a few *brand* categories (listed in the Appendix). Of note, the multiple linear regression model estimates that the coefficient for *not_Repaired_Damage*, a binary variable indicating whether or not a car had damage, is .072. In context, this means that for the presence of damage on the car (binary variable 1), *price* is predicted to change by $e^{-0.72}$, or decrease by ~52%. This makes sense given what we know about the used car market, and we believe this would be useful information for a seller to know if he or she is considering selling a damaged used car.

We next move into prediction using this model, and we are able to use a relevant case to test our prediction. One of our group members recently purchased a new car, so we tested our model to see what it would predict as the price of that specific vehicle. The model predicted a price of $12,224, with a 95% confidence interval between $3,553 and $41,820. The true price of the car is included in the confidence interval and was ~$2,000 away from the predicted price. Although this is only one test, it confirms our suspicion that the predictions from this model may not be very reliable.

The other questions we explored yielded interesting observations and associations that we believe could be useful to someone in the market to buy or sell a used car, if interpreted correctly. For example, we find using an ANOVA that advertisements that include one or more exclamation points in their titles sell for less than those without an exclamation point. We certainly cannot say this is causal, but the

association may exist based on some other factor; for example, perhaps individuals using exclamation points are trying to sell cars quickly and are willing to settle for a lower price, or are perhaps trying to differentiate their car among a brand or model with many options.

Finally, running an ANOVA on not_Repaired_Damage also indicates that there is a statistically significant difference in the mean prices of damaged and non-damaged cars, which aligns with our results in Question 4.

Further Study:

Throughout this paper, our group has identified many opportunities for further research that could improve statistical inference and reliability of prediction. We would recommend that other groups perform further research to understand how to make a model for prediction of *price* that does not violate the assumptions necessary for goodness of fit. We also see opportunity for further exploration and subsetting around the categorical variable *brand*, as we suspect one might find more reliable predictive results by looking at a smaller subset of similar vehicles.

Furthermore, since the dataset is so large, we would recommend that future groups train a promising model with a training set, and reserve a testing set for assessing predictive power. We did not include this step, as we did not anticipate robust predictions from our model. Finally, we would recommend more robust variable selection, and that future groups consider subsetting the data to work with a much smaller dataset, as this may provide more relevant results.