Assignment 1 — Initial Analysis

Sourabh D. Khot (Student ID 002754952)

College of Professional Studies, Northeastern University

ALY 6070: Communication and Visualization for Data Analytics CRN 81182

Professor Venkata Duvvuri

Jun 9, 2022

Table of Contents

Introduction

The following scenario is created for the purpose of this assignment. My friend is interested in selling his BMW Model 3 Series car directly to a potential buyer but could not decide the price for his listing. I have taken a dataset from Kaggle.com titled 'BMW used car listing' and will try to analyze the pattern and trend to predict car prices from BMW. I will create a dashboard for public use so that any potential BMW car seller can decide the price. They will be able to list their cars with the actual attributes and the price they decide, which will update the dynamic dashboard. Potential buyers can also view the dashboard and get an idea of the car price as per their requirements.

Analysis

What do you see as variables in the data?

There are 9 variables in the data, of which 6 are numerical and 3 are categorical. The variables are described below:

1. model (categorical, nominal) – model of the car from BMW
2. year (numerical, interval, discrete) – the year of registration of the car
3. price (numerical, ratio, discrete) – the price of the car in Euros
4. transmission (categorical, nominal) – the type of gearbox
5. mileage (numerical, ratio, continuous) – total distance traveled by car
6. fuelType (categorical, nominal) – the type of fuel used by the car among petrol or diesel
7. tax (numerical, ratio, discrete) – road tax for the car
8. mpg (numerical, ratio, continuous) – fuel economy of the car in miles per gallon
9. engineSize (numerical, ratio, discrete) – car engine's capacity or displacement in liters

What correlations, patterns and trends do you see?

I have plotted the correlation matrix in figure 1 to understand the relationship between numerical variables. The highest correlation is between mileage and year (between -1 and -0.75), such that when the year increases, mileage decreases, which is logical. Price is the output variable, with the highest correlation between price and mileage/year. The price decreases when year decrease or mileage increases. Price has the next highest correlation with engine size.

Figure 1. Correlation Matrix

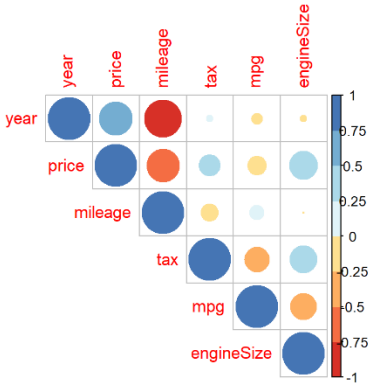
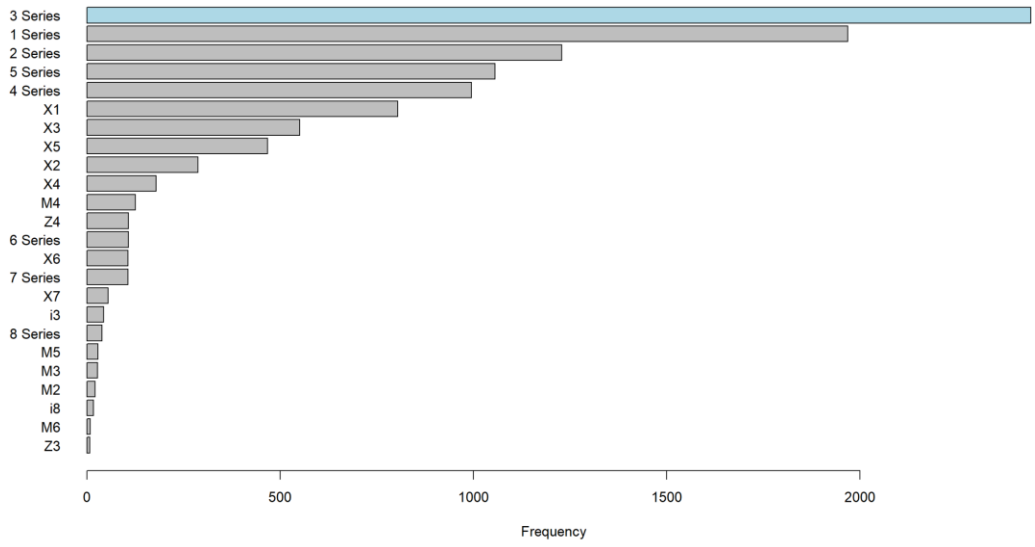
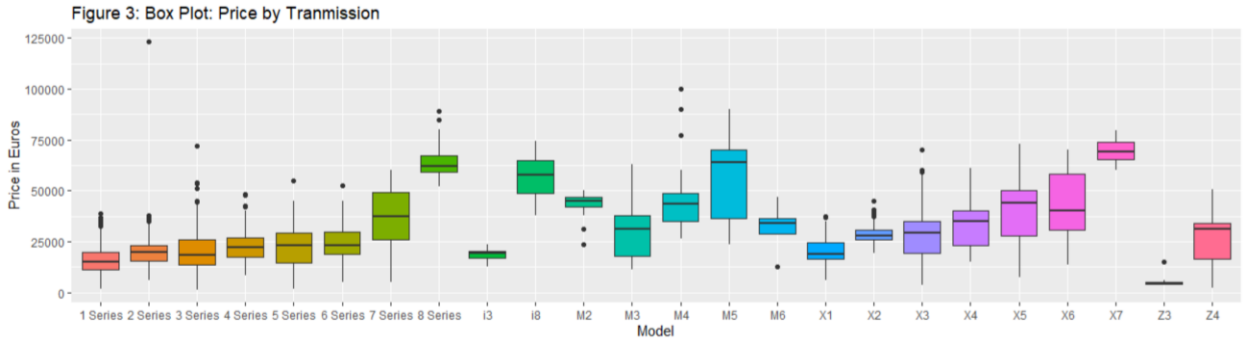


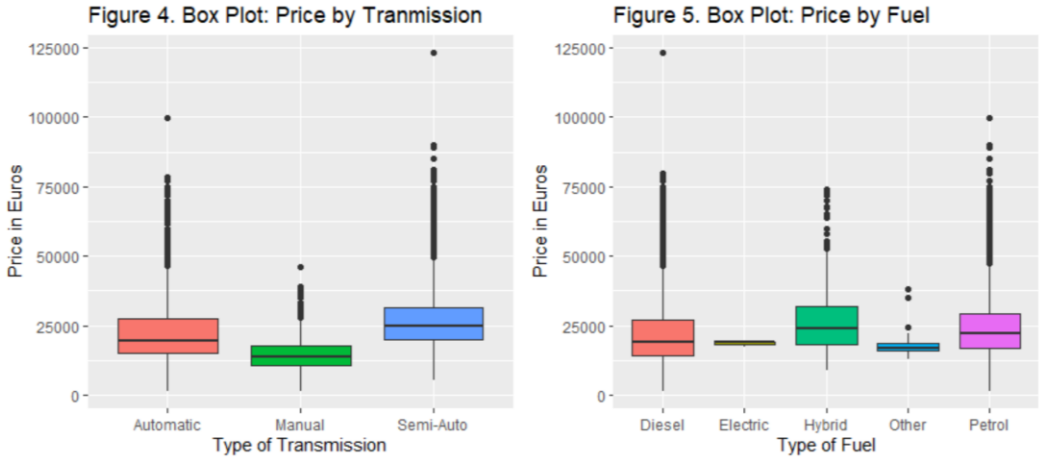
Figure 2. Barplot: Count of Models



Next, I have analyzed the frequency distribution of models in figure 2. I have used the vertical barplot since the categorical variable is of the type nominal scale. The maximum count is of BMW 3 Series, which is highlighted in blue.



A correlation matrix is for numerical variables only. To understand the impact of categorical variables on price, bivariate box plots can be used. The impact of model on price is checked using box plots in figure 3. Visually, we can see that model M5 has the highest price while model Z3 has the lowest price.



Similarly, figures 4 and 5 show the impact of transmission type and fuel type on car prices. Automatic and semi-automatic cars are costlier than manual. Similarly, hybrid vehicles are costlier. Also, the price range of electric vehicles is very narrow compared to other fuel types. In figures 3-5, we see many outliers, possibly due to other factors not considered.

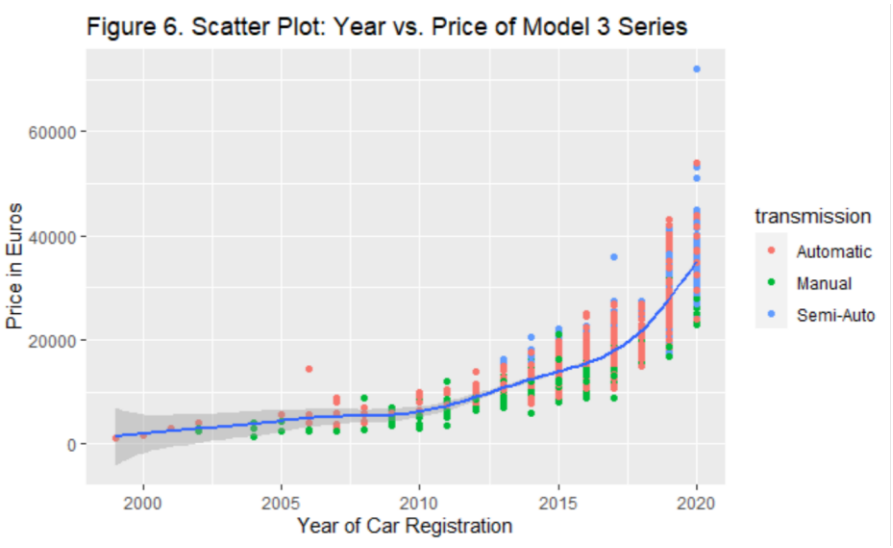


Figure 6 is a scatter plot for year vs. price of only model 3 series cars, additionally indicating the transmission type by color. I am currently interested in model 3 series cars, and I chose the year variable since it is highly correlated with price. We can see a trend that as the year increases, the price also increases. However, it is not linear but a kind of exponential trend with increasing variance as the year increases. We can see that manual transmission cars (green) generally have lower prices than automatic ones (red), while some outlier cars have semi-automatic transmissions (blue).

Who will be the audience of the data? What question would they ask about the data?

BMW owners who want to sell it and potential buyers will be the audience of the data. Sellers would bring attributes of their cars and ask a suitable price range for selling them. Similarly, buyers would select their requirements and budget and try to find a vehicle that fits their criteria and price range. Buyers would be interested to know what other factors impact car prices apart from model and mileage. E.g., those interested in buying automatic cars would ask if cars with the same mileage cost the same for automatic vs. manual, or if automatics are costlier.

What questions would you propose to answer to your audience through the data viz.?

To sellers, I would propose to answer the question of an appropriate price range for their car. I would ask buyers what their rigid requirements are (e.g., price, a specific model), which can be designed as filters in the dashboard. I would also need attributes which they are flexible (like mileage and transmission), which would be part of the x-axis of some of the charts in the dashboard. This would help me design the dashboard and help them analyze trade-offs between attributes by themselves to buy a suitable car.

How could a dashboard be used to show data clearly?

Every seller has a unique car, and every buyer has a different requirement. A dynamic dashboard can clearly show which variables impact the car price and which do not. A buyer can fix their rigid requirements in filters and accordingly decide their preferences to decide the car they want to buy as per their budget. A seller can analyze the price range for similar cars and get a good idea of the price for listing their car.

What types of graphs and charts can be used to clearly explain the data and answer the business question?

As per previous figures, a correlation matrix with colors and shape sizes can visually highlight which factors impact price. Box plots show how the price distribution is different for a particular variable. A horizontal bar plot of model count shows which vehicles are popular in the market. A scatterplot is very useful to understand the price trend with respect to a dependent variable, which is more accurate when filtered for a particular model.

Conclusion

The BMW model 3 series is the most popular series, costing around 20,000 euros. Model M6 and Z3 are very rare, and their prices may not be modeled accurately due to a lack of data.

The price of a BMW car will primarily depend on the car model and the mileage or registration year (mileage and year have multicollinearity). Price also depends upon engine size and type of transmission, and fuel. Tax and mpg have less of an impact on prices. For a particular model, the price increases exponentially as the year of registration increases.

Using my analysis, a BMW car seller can decide the price to list his car, and a buyer can get an idea of the price for their requirement.

References

An Introduction to corrplot Package: R-project.org. (2021). Retrieved from https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html

APA Style Table: APA.org. (n.d.). Retrieved from https://apastyle.apa.org/style-grammar-guidelines/tables-figures/tables

BMW used car listing: Kaggle.com. (n.d.). Retrieved from https://www.kaggle.com/datasets/mysarahmadbhat/bmw-used-car-listing

Canvas ALY6070: Assignment 1 Initial Analysis. (n.d.). Retrieved from https://northeastern.instructure.com/courses/110061/assignments/1351343

Canvas ALY6070: Lesson 3-8 The Basics of Baseplot and GGplot. (n.d.). Retrieved from https://northeastern.instructure.com/courses/110061/pages/lesson-3-8-the-basics-of-baseplot-and-ggplot

Knaflic, C. N. (2015). *Storytelling with Data - A Data Visualization Guide for Business Professionals.* New Jersey: Wiley.

Appendix: Code

```
# 1. Importing ####

library(dplyr) # for data manipulation
library(corrplot) # for correlation
library(RColorBrewer) # for color in correlation
library(gridExtra)    # for visualization: grid.arrange()

df <- read.csv('Week2/bmw.csv')

# 2. Cleaning ####

head(df)
View(df)

# show unique from categorical rows
lapply(df[,c(1,4,6,9)], unique)

df$model <- factor(df$model)
df$transmission <- factor(df$transmission)
df$fuelType <- factor(df$fuelType)

# 3. EDA ####

cors = cor(df[,c(2,3,5,7,8,9)], use='pairwise')
corrplot(cors, type='upper', col=brewer.pal(n=8, name="RdYlBu"),
        main = "Figure 1. Correlation Matrix", mar=c(0,0,2,0),)

## a. Univariate ####

model_counts <- df %>% group_by(model) %>% count(model) %>%
  rename(counts = n) %>% data.frame () %>% arrange(counts)

barplot(model_counts$counts,
       names.arg = model_counts$model,
       xlab = 'Frequency',
       col =
c('grey','grey','grey','grey','grey','grey','grey','grey','grey','grey','grey','grey','grey','grey','grey','grey',
'grey','grey','grey','grey','grey','grey','grey','light blue'),
```

```
      horiz = TRUE,
      las = 1,
      main = 'Figure 2. Barplot: Count of Models')

## b. Bivariate ####

qplot(x=model, y=price, data=df, fill=model, geom='boxplot', main='Figure 3: Box Plot: Price
by Tranmission') + guides(fill=FALSE)+
  xlab("Model")+
  ylab("Price in Euros")

x <- qplot(x=transmission, y=price, data=df, fill=transmission, geom='boxplot', main='Figure 4.
Box Plot: Price by Tranmission') + guides(fill=FALSE)+
  xlab("Type of Transmission")+
  ylab("Price in Euros")
y <- qplot(x=fuelType, y=price, data=df, fill=fuelType, geom='boxplot', main='Figure 5. Box
Plot: Price by Fuel') + guides(fill=FALSE)+
  xlab("Type of Fuel")+
  ylab("Price in Euros")
grid.arrange(x, y, nrow=1)
rm(x, y)

## c. Multivariate ####

# ggplot(data=df)+
#  geom_point(mapping = aes(x=year,y=price,color=model) )+
#  xlab("Year of Car Registration")+
#  ylab("Price in Euros")+
#  geom_smooth(mapping = aes(x=year,y=price) )+
#  ggtitle("Figure 7. Scatter Plot: Year vs. Price of all Models")

ggplot(data=df[df$model==" 3 Series",])+
    geom_point(mapping = aes(x=year,y=price,color=transmission) )+
    xlab("Year of Car Registration")+
    ylab("Price in Euros")+
    geom_smooth(mapping = aes(x=year,y=price) )+
    ggtitle("Figure 6. Scatter Plot: Year vs. Price of Model 3 Series")
```

<End of Report>