SSec4_FP_DavidBadalamenti_HsuanYunLiau_RunyiZh

Exploring factors influencing the pricing of used cars

Data: Auto Scout Car

The buying ans selling of vehicles form a critical part of economies worldwide, with online marketplaces playing an increasingly important role in connecting buyers and sellers. Auto Scout, the largest pan-European online car marketplace, provides an extensive data set detailing used car listings, including specifications, features, and prices. These details offer a valuable opportunity to investigate a pressing question in the automotive market: *What factors significantly influence the pricing of used cars?* In this report, we will explore how age and mileage of a vehicle affect the price and find which has a more significant impact.

Explain how your data meet the FAIR and/or CARE Principles:

FAIR:

- Find able:
  - data set found through Kaggle, which is a widely recognized and accessible platform for data sharing
  - includes metadata that describes the data set structure and its variables, which helps users locate and understand relevant information
- Accessible:
  - the data set is provided in a CSV format
  - the platform has open access to data set
- Inter operable:

  - the data uses standard and machine readable formats, such as numeric fields for prices and mileage and categorical fields for attributes like fuel type and body type
  - the metadata provides clear descriptions of fields

Reusable:

- the data is accompanied by metadata that ensures the attributes, ensuring clarity for reuse

CARE:

- **Collective benefit**:

The data set enables analyses that can benefit a wide audience, such as consumers, dealers, and researchers

- **Authority to control**:
  - the data contains public information on car listings and does not include personal or community-sensitive data
- **Responsibility:**
  - the data set avoids ethical issues by not including personal identifiers or private information
- **Ethics:**
  - By excluding sensitive information, the data set aligns with ethical principles for data usage.
  - The source (Kaggle) provides transparency about the data set's origin, licensing, and usage conditions

Distribution of car models:

To begin with, we generated a histogram visualizing the frequency distribution of car models in the data set, with the make_model variable on the x-axis and the count of listings on the y-axis. By identifying the models with the highest frequency, we can focus on the top three most common car models for further analysis, which in this case, the Audi A3, Audi A1, and Opel Insignia.

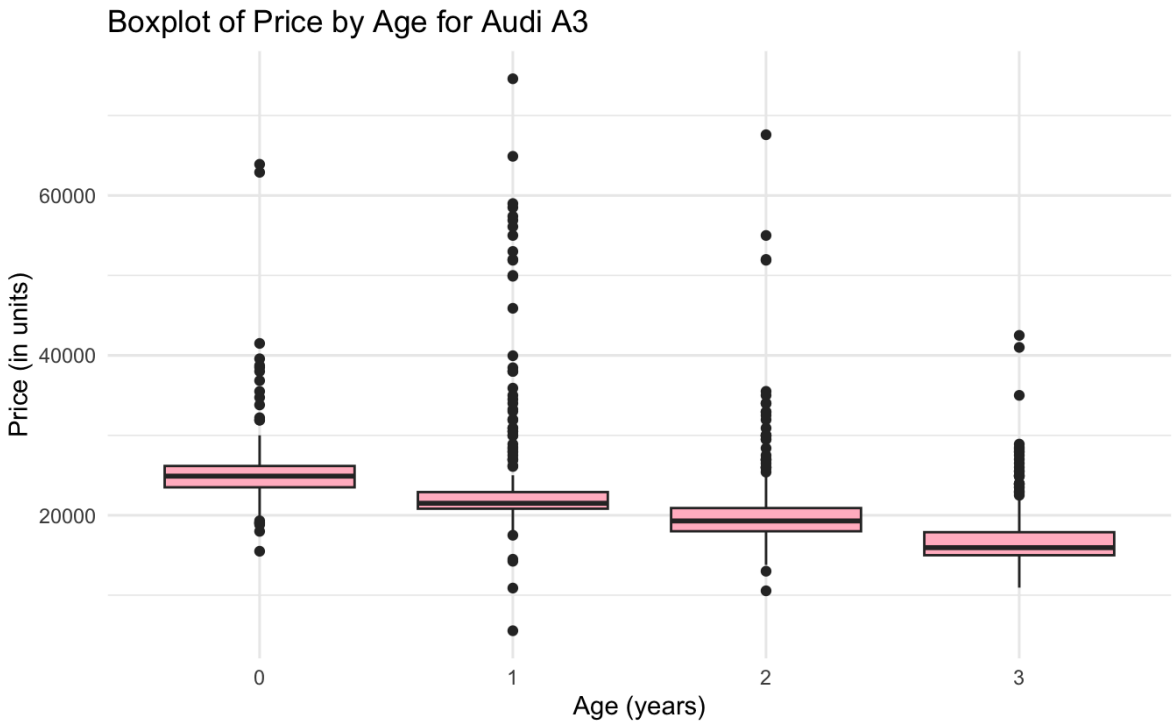
Choosing larger sample size for these models ensures the results are statistically reliable and can help better evaluate how variables like mileage and age impact the prices within and across top models. By finding the regressions coefficient for both age and mileage we can see which has a larger impact on the price of the car.

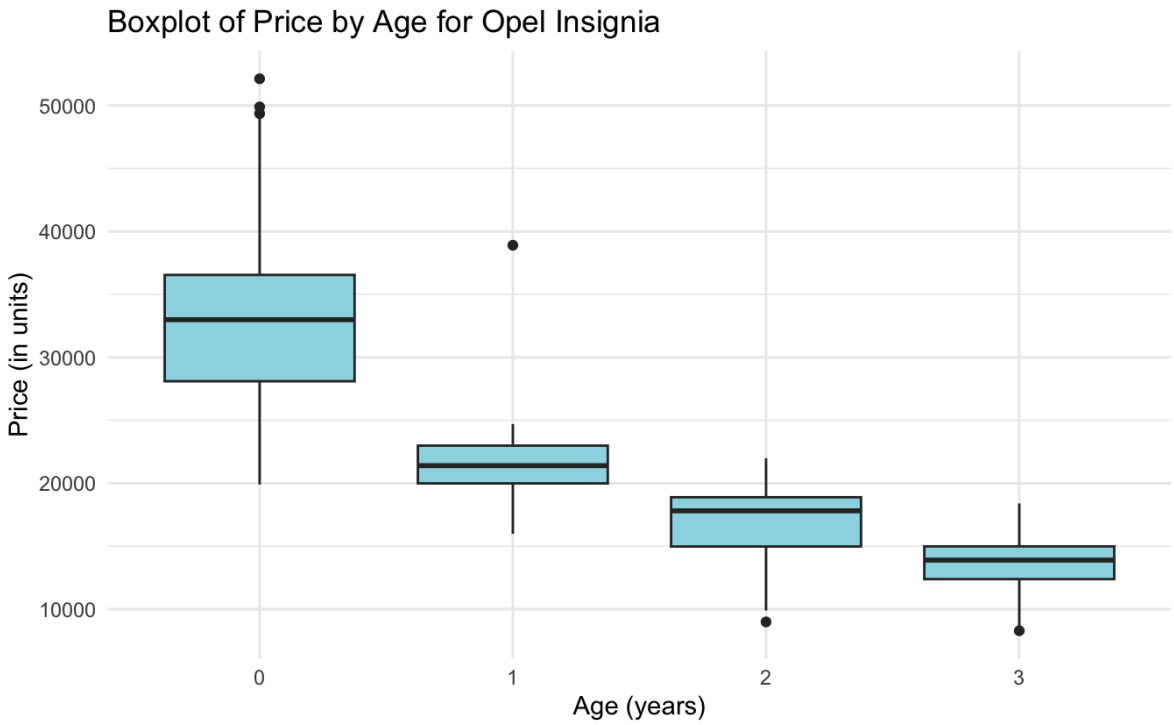
We used a cleaned data set to get the numerical statistics of what we wanted to look at directly age, price and mileage.

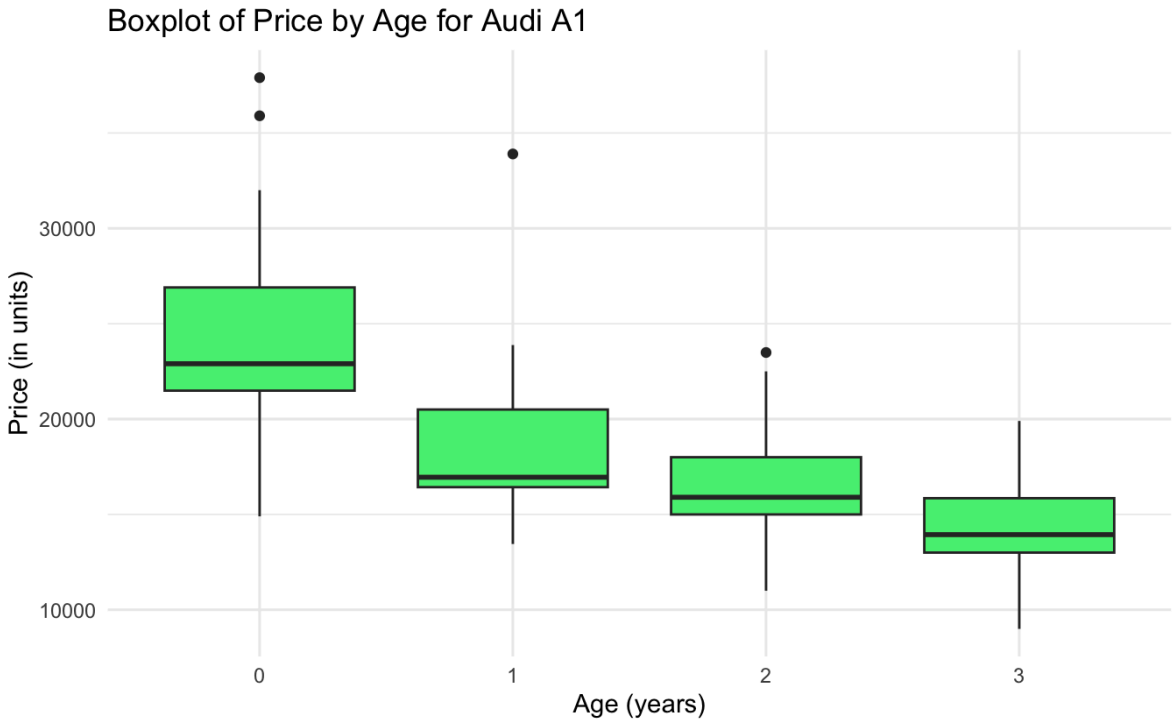
Our initial thoughts were that since the Audi A3 has the highest price and the lowest price the effect of the mileage and age of that model would be the largest

##Age Models

The following are the graphs of the age vs price for the three different car models

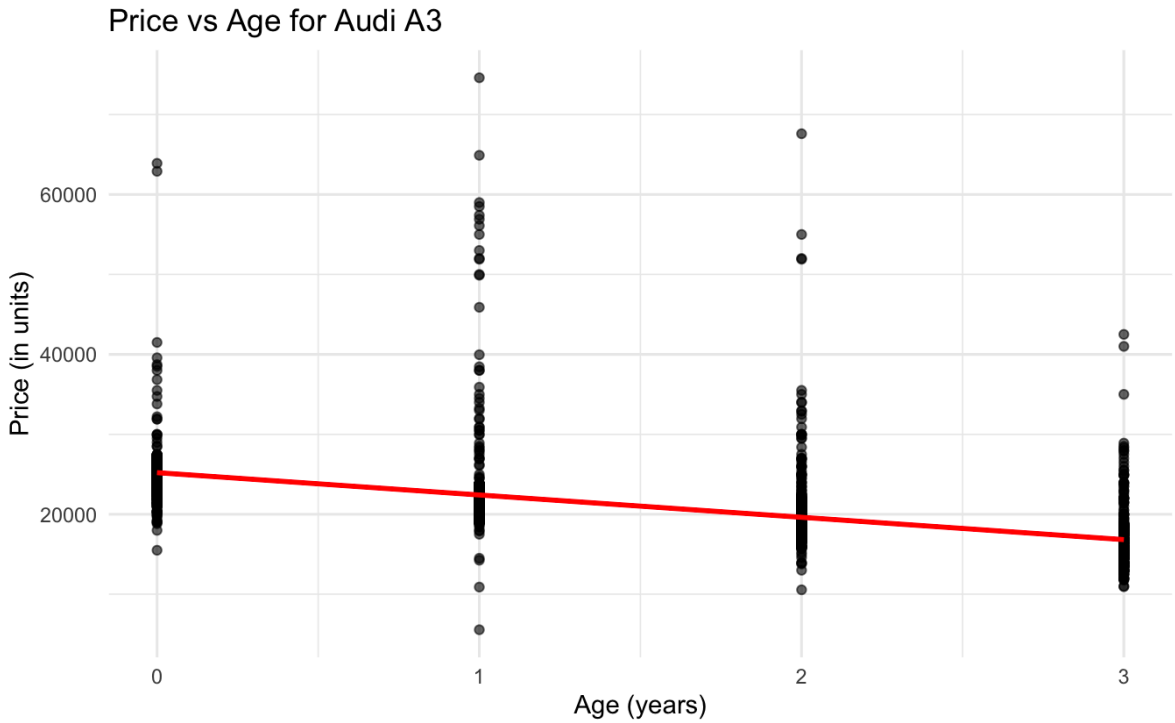


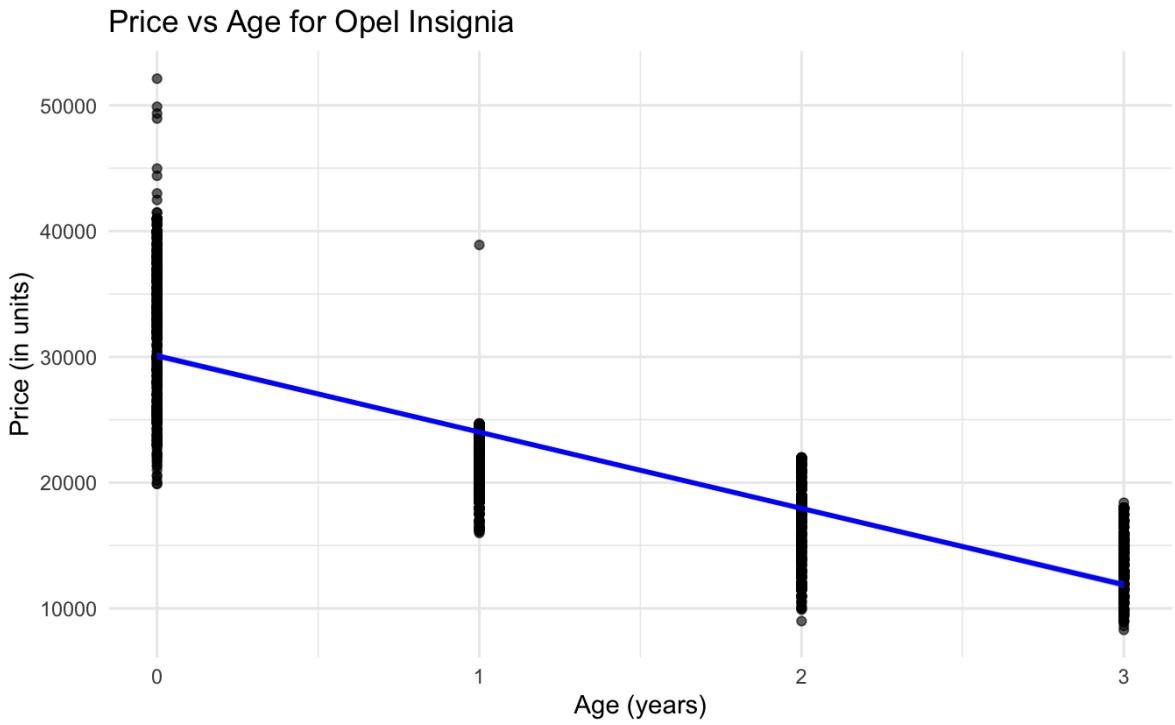


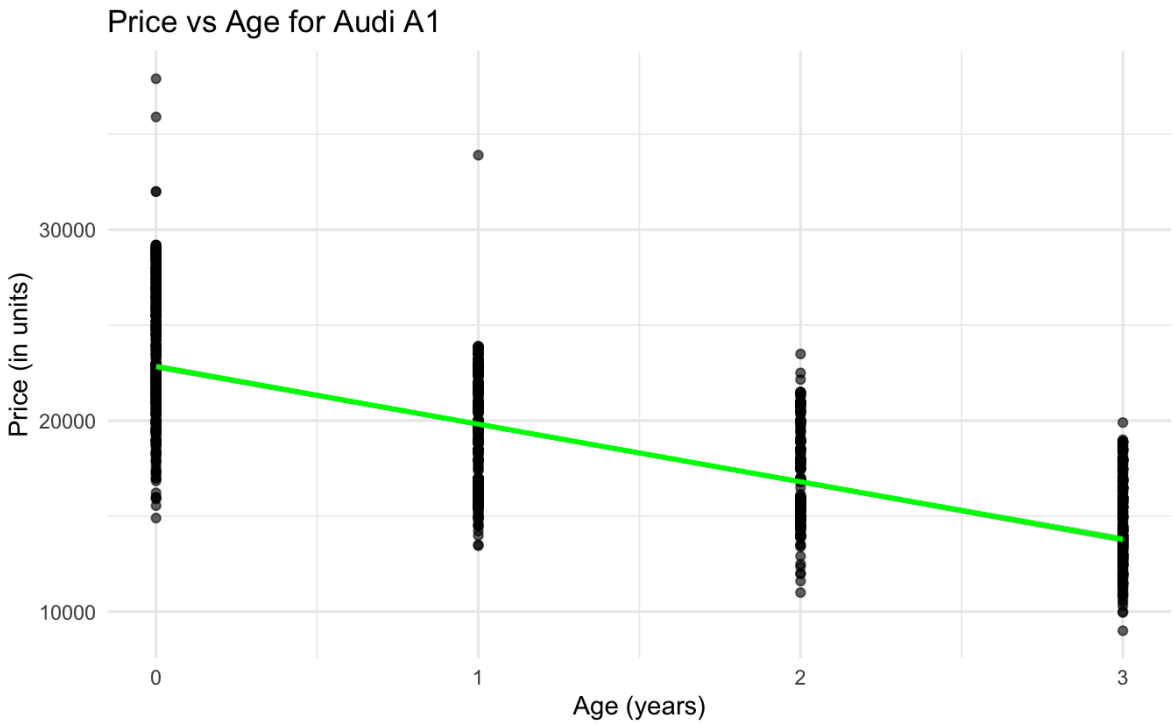


After Visualizing the data with a box plot we create a scatter plot with a regression line to

see the change in price by age





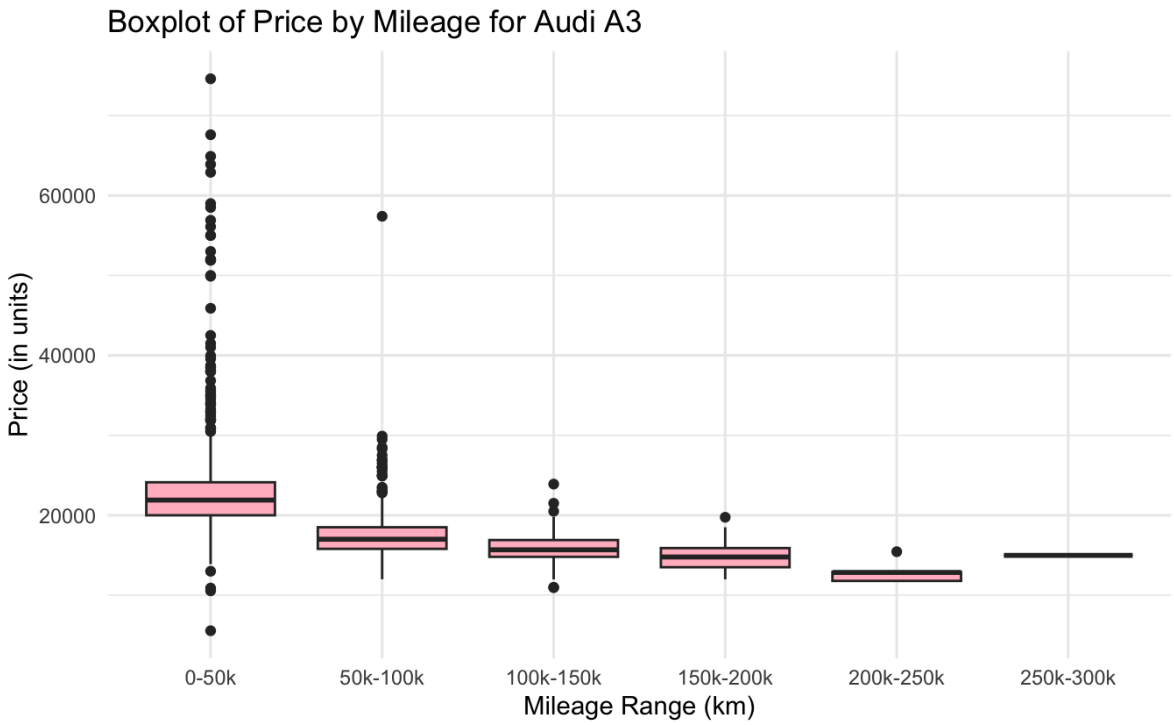


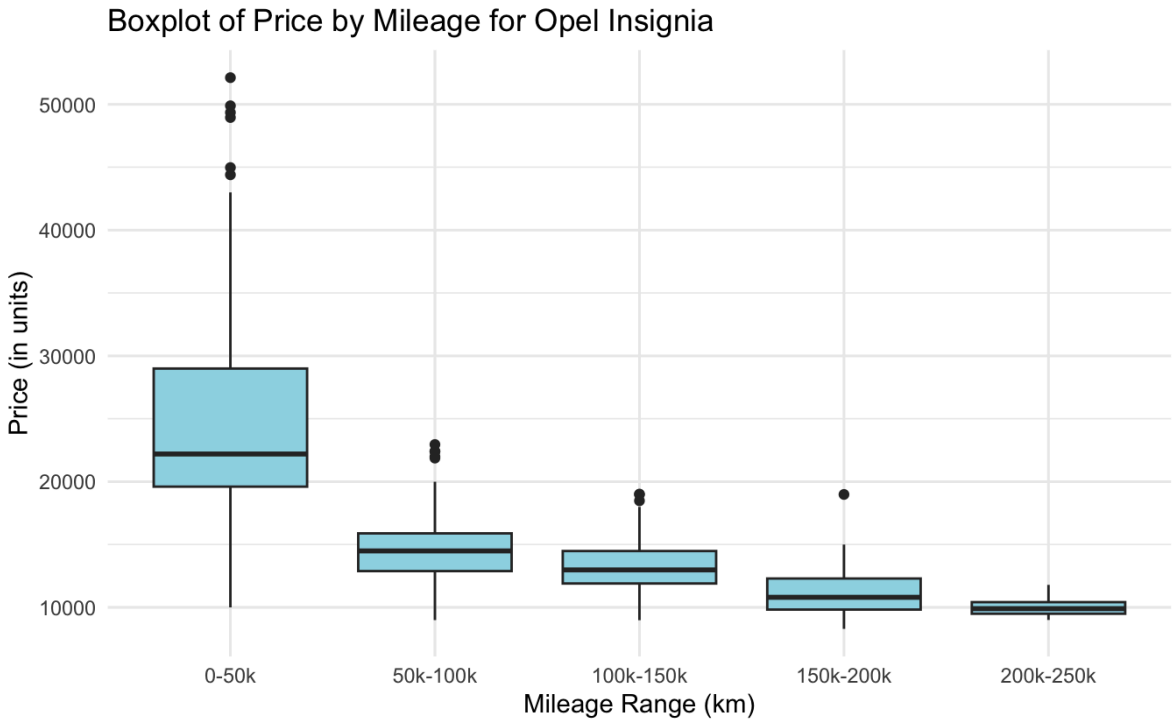
As expected the age has a large affect on the price of the car decreasing over time we then take the regression line and find the regression coefficient to determine the amount the price changes per year per model of car ##Age Coefficients For the Audi A3 the coefficient was -2792.16 meaning that per year the price of car decreases by almost 2,800 Euro For the Audi A1 the coefficient was -3016.65 decreasing at about 3,000 Euro per year For the Opel Insignia the coefficient was -6068.48 decreasing at a rate of about 6,000 Euro per year

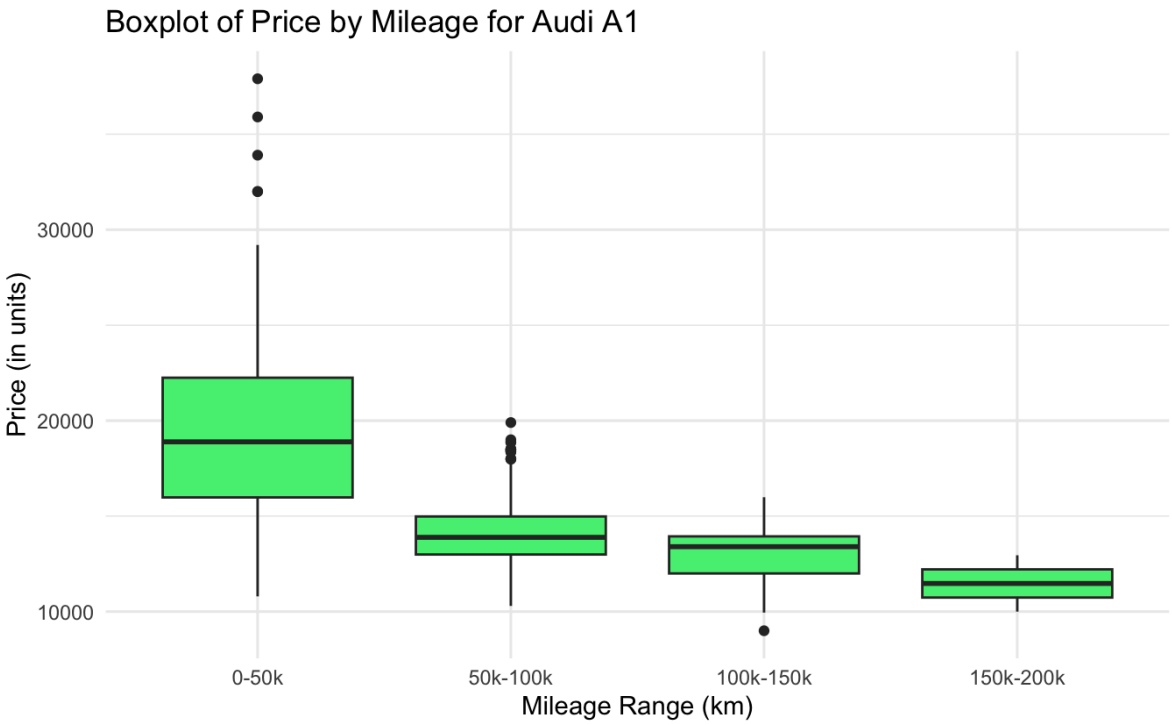
Considering our initial guess that the Audi A3 would have the larger coefficient because of the large difference in the max and min prices seeing the Opel's be more than double was very surprising. When looking at the regression lines and the box plots however it does make a little more sense as the max and min are outliers compared to the majority of the data.

##Mileage Models

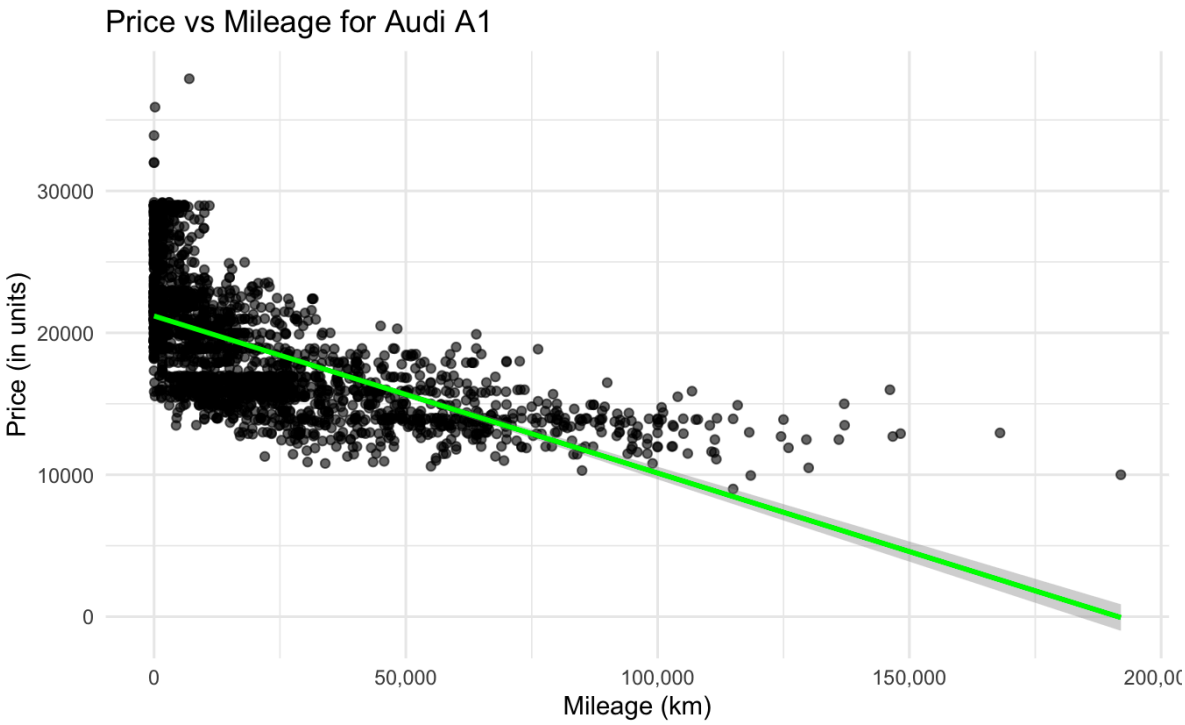
Following the pattern from the age models we first created box plots of the price vs the mileage for each of the cars, since unlike age the mileage is a lot more distributed we grouped them by every 50,000 Kilometers

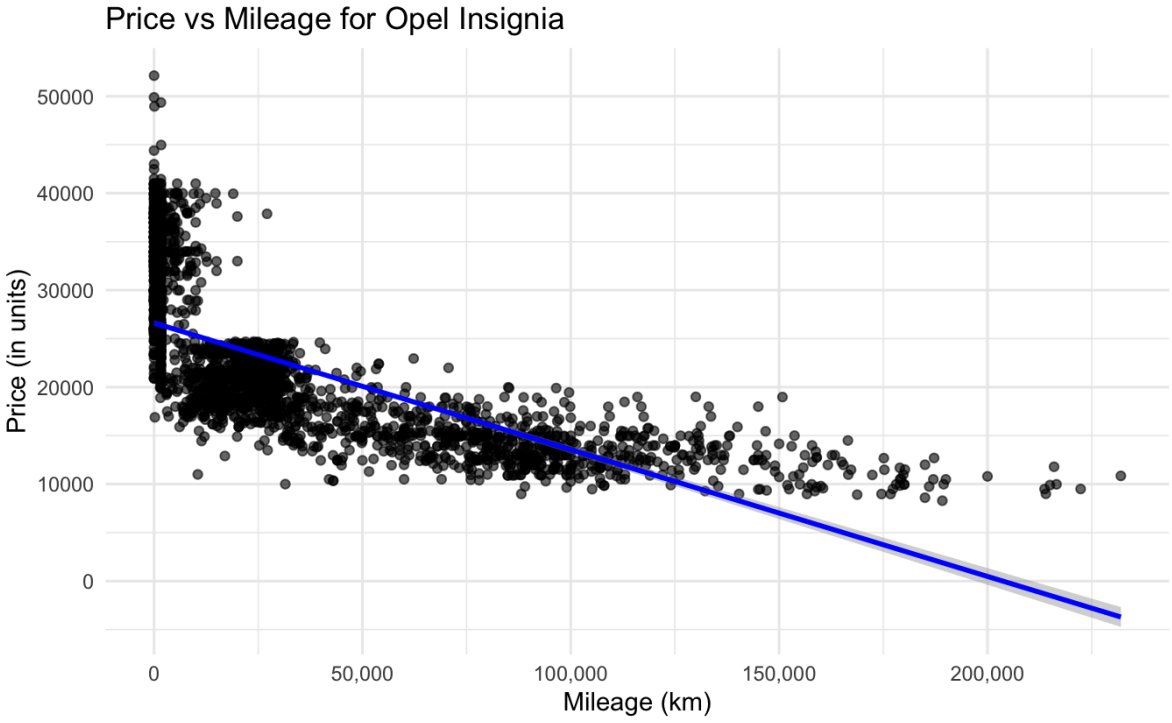

Boxplot of Price by Mileage for Audi A3
Price (in units)
60000
40000
20000
0-50k
50k-100k
100k-150k
150k-200k
200k-250k
250k-300k
Mileage Range (km)

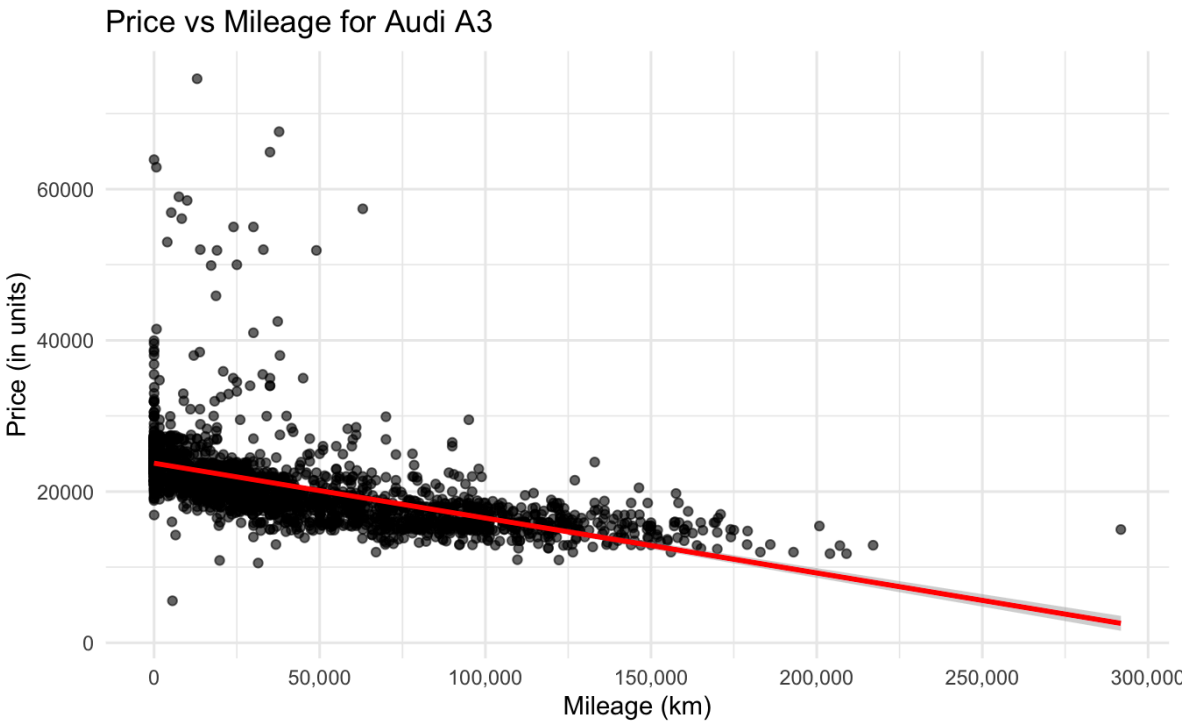

Boxplot of Price by Mileage for Opel Insignia
Price (in units)
50000
40000
30000
20000
10000
0-50k
50k-100k
100k-150k
150k-200k
200k-250k
Mileage Range (km)



The higher ends of the mileages tend to have less cars, so the higher the mileage the less likely someone is to sell? We next got modeled a scatter plot just as we did with age to get a regression model and find the regression coefficient for each car ##Scatter plots


Price vs Mileage for Audi A1
Price (in units)
30000
20000
10000
0
0
50,000
100,000
150,000
Mileage (km)


Price vs Mileage for Opel Insignia
Price (in units)
50000
40000
30000
20000
10000
0
0
50,000
100,000
150,000
200,000
Mileage (km)



Visually you can see the slope of the Regression line on the Audi A3 mileage was lower than the other two models so it would logically follow that the mileage of the audi has a less affect than the mileage of the other models

##Mileage Coefficients

For the mileage the coefficient is given as the change per mile but that number was in the cents and didn't make sense for the groups we made so we multiplied the coefficient by 50,000 to get the change per 50,000 km

For the Audi A3 the coefficient was -3629.50 meaning that per year the price of car decreases by about 3600 Euro For the Audi A1 the coefficient was -5532.67 decreasing at about 5,500 Euro per year For the Opel Insignia the coefficient was -6535.30 decreasing at a rate of about 6,500 Euro per year

##Conclusion

For the three cars the Opel Insignia was affected the most by age and mileage having the largest decrease per year and per mile than either Audi Model. The Audi A3 was affected less than the A1 model in both age and mileage but the A1 had a much larger difference from age to mileage than the A3 increasing from -3016.65 to -5532.67 while the A3 model only increased from -2792.16 to -3629.50 this shows us that the mileage of the A1 had a larger impact on the price that the mileage on the A3 model. This could come from the type of car itself if one is a more sporty car the buyer might feel that the mileage has a larger impact on what they would

pay for the vehicle. Some of the key points we found were that the Opel Insignia is affected by both age and mileage the most, we also discovered that while the A3 has the largest gap from minimum car price to maximum car price it was affected the least by age and mileage.

The cars lose euro cents every kilometer so representing the change in price by mileage as groups of 50,000 km made more sense especially when trying to compare the price vs age and mileage. It also helped to visually see if the change in price would decrease after a certain point or if the change per group stayed relatively similar These changes in the price we are looking at are also a little skewed as you cant inspect the car visually or account for things like damages aswell as if the car had been changed by the owner in anyway. Directly comparing the age to the mileage is hard as they have different subgroups but the comparing the affect of the age or mileage to the different types of cars was much easier. If we were to do further investigations we might try and find what makes some of those cars outliers like what attributes lead to an extreme price or if it was maybe the seller changing the price of the car beyond what it would actually sell for.

##Citations

```
#Style Guide: Tidyverse Style Guide (https://style.tidyverse.org/)
{r keyStats, ref.label=knitr::all_labels(), echo=TRUE, eval=FALSE}
#Provides statistics for price, mileage, and age
summary_table <- data %>%
  filter(make_model %in% c("Audi A3", "Opel Insignia", "Audi A1")) %>%
  group_by(make_model) %>%
  summarise(
    Avg_Price = mean(price, na.rm = TRUE),
    Median_Price = median(price, na.rm = TRUE),
    Min_Price = min(price, na.rm = TRUE),
    Max_Price = max(price, na.rm = TRUE),
    Avg_Mileage = mean(mileage_km, na.rm = TRUE),
    Median_Mileage = median(mileage_km, na.rm = TRUE),
    Min_Mileage = min(mileage_km, na.rm = TRUE),
    Max_Mileage = max(mileage_km, na.rm = TRUE),
    Avg_Age = mean(age, na.rm = TRUE),
    Median_Age = median(age, na.rm = TRUE),
    Min_Age = min(age, na.rm = TRUE),
    Max_Age = max(age, na.rm = TRUE)
  )

print(summary_table)

{r ageModels, ref.label=knitr::all_labels(), echo=TRUE, eval=FALSE}
audi_a3_data <- data %>% filter(make_model == "Audi A3")
```

```r
# Scatter plot with regression line
ggplot(audi_a3_data, aes(x = age, y = price)) +
geom_point(alpha = 0.6) +
geom_smooth(method = "lm", se = TRUE, color = "red") +
labs(
title = "Price vs Age for Audi A3",
x = "Age (years)",
y = "Price (in units)"
) +
theme_minimal()
audi_age_model <- lm(price ~ age, data = audi_a3_data)
summary(audi_age_model)


# Boxplot
ggplot(audi_a3_data, aes(x = factor(age), y = price)) +
  geom_boxplot(fill = "pink") +
  labs(
    title = "Boxplot of Price by Age for Audi A3",
    x = "Age (years)",
    y = "Price (in units)"
  ) +
  theme_minimal()

opel_insignia_data <- data %>% filter(make_model == "Opel Insignia")

# Scatter plot with regression line
ggplot(opel_insignia_data, aes(x = age, y = price)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, color = "blue") +
  labs(
    title = "Price vs Age for Opel Insignia",
    x = "Age (years)",
    y = "Price (in units)"
  ) +
  theme_minimal()

opel_age_model <- lm(price ~ age, data = opel_insignia_data)
summary(opel_age_model)

# Boxplot
ggplot(opel_insignia_data, aes(x = factor(age), y = price)) +
```

```
  geom_boxplot(fill = "lightblue") +
  labs(
    title = "Boxplot of Price by Age for Opel Insignia",
    x = "Age (years)",
    y = "Price (in units)"
  ) +
  theme_minimal()
audi_a1_data <- data %>% filter(make_model == "Audi A1")

# Scatter plot with regression line
ggplot(audi_a1_data, aes(x = age, y = price)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, color = "green") +
  labs(
    title = "Price vs Age for Audi A1",
    x = "Age (years)",
    y = "Price (in units)"
  ) +
  theme_minimal()

audi1_age_model <- lm(price ~ age, data = audi_a1_data)
summary(audi1_age_model)

# Boxplot
ggplot(audi_a1_data, aes(x = factor(age), y = price)) +
  geom_boxplot(fill = "lightgreen") +
  labs(
    title = "Boxplot of Price by Age for Audi A1",
    x = "Age (years)",
    y = "Price (in units)"
  ) +
  theme_minimal()
{r mileModels, ref.label=knitr::all_labels(), echo=TRUE, eval=FALSE}
#scatter plot
ggplot(audi_a3_data, aes(x = mileage_km, y = price)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  scale_x_continuous(labels = scales::comma, breaks = seq(0, 300000, by = 50000)) +
  labs(
    title = "Price vs Mileage for Audi A3",
    x = "Mileage (km)",
    y = "Price (in units)"
```

```
  ) +
  theme_minimal()

audi_mile_model <- lm(price ~ mileage_km, data = audi_a3_data)
summary(audi_mile_model)

# Boxplot
ggplot(audi_a3_data, aes(x = cut(mileage_km, breaks = c(0, 50000, 100000, 150000, 200000, 250
                                 labels = c("0-50k", "50k-100k", "100k-150k", "150k-200k", "2
                         y = price)) +
  geom_boxplot(fill = "pink") +
  labs(
    title = "Boxplot of Price by Mileage for Audi A3",
    x = "Mileage Range (km)",
    y = "Price (in units)"
  ) +
  theme_minimal()

# Boxplot
ggplot(opel_insignia_data, aes(x = cut(mileage_km, breaks = c(0, 50000, 100000, 150000, 20000
                                       labels = c("0-50k", "50k-100k", "100k-150k", "150k-200
                               y = price)) +
  geom_boxplot(fill = "lightblue") +
  labs(
    title = "Boxplot of Price by Mileage for Opel Insignia",
    x = "Mileage Range (km)",
    y = "Price (in units)"
  ) +
  theme_minimal()

# Boxplot

ggplot(audi_a1_data, aes(x = cut(mileage_km, breaks = c(0, 50000, 100000, 150000, 200000, 250
                                 labels = c("0-50k", "50k-100k", "100k-150k", "150k-200k", "2
                         y = price)) +
  geom_boxplot(fill = "lightgreen") +
  labs(
    title = "Boxplot of Price by Mileage for Audi A1",
    x = "Mileage Range (km)",
    y = "Price (in units)"
  ) +
  theme_minimal()
```