

# SSec4\_FP\_DavidBadalamenti\_HsuanYunLiau\_RunyiZh

## Exploring factors influencing the pricing of used cars

### Data: Auto Scout Car

The buying and 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 dataset 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 dataset, 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.

```
library(readr)
library(dplyr)
```

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

filter, lag

The following objects are masked from 'package:base':

intersect, setdiff, setequal, union

```
data <- read.csv("final_scout_not_dummy.csv")
```

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
```

```
v forcats 1.0.0      v stringr 1.5.1
```

```
v ggplot2  3.5.1      v tibble  3.2.1
```

```
v lubridate 1.9.3     v tidyr   1.3.1
```

```
v purrr     1.0.2
```

```
-- Conflicts ----- tidyverse_conflicts() --
```

```
x dplyr::filter() masks stats::filter()
```

```
x dplyr::lag()     masks stats::lag()
```

```
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
```

```
data <- read.csv("cleaned_scout_data.csv")
```

```
model_counts <- data %>%
```

```
  group_by(make_model) %>%
```

```
  summarise(count = n()) %>%
```

```
  arrange(desc(count))
```

```
top_3_models <- head(model_counts, 3)
```

```
library(ggplot2)
```

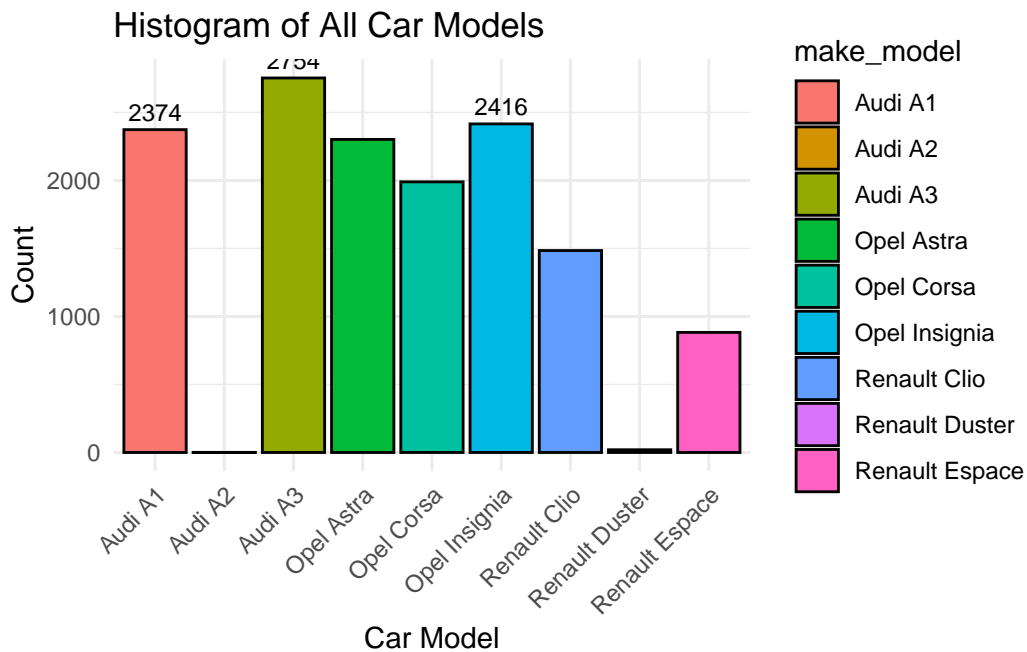
```
ggplot(data, aes(x = make_model, fill = make_model)) +
```

```
  geom_bar(color = "black") +
```

```
  geom_text(data = top_3_models, aes(x = make_model, y = count, label = count),
```

```
    vjust = -0.5, color = "black", size = 3) +
```

```
labs(
  title = "Histogram of All Car Models",
  x = "Car Model",
  y = "Count"
) +
theme_minimal() +
theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



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 dataset to get the numerical statistics of what we wanted to look at directly age, price and mileage.

```
library(readr)
library(dplyr)
data <- read.csv("final_scout_not_dummy.csv")

library(tidyverse)

data <- read.csv("cleaned_scout_data.csv")
```

```

model_counts <- data %>%
  group_by(make_model) %>%
  summarise(count = n()) %>%
  arrange(desc(count))

top_3_models <- head(model_counts, 3)

library(ggplot2)
ggplot(data, aes(x = make_model, fill = make_model)) +
  geom_bar(color = "black") +
  geom_text(data = top_3_models, aes(x = make_model, y = count, label = count),
            vjust = -0.5, color = "black", size = 3) +
  labs(
    title = "Histogram of All Car Models",
    x = "Car Model",
    y = "Count"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

#Provides key 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)

```

```

audi_a3_data <- data %>% filter(make_model == "Audi A3")

# 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 for price by age for Audi A1
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()

audi_a3_data <- data %>% filter(make_model == "Audi A3")

# 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 for price by age for Audi A1
# 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()
#
# #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, 250000),
labels = c("0-50k", "50k-100k", "100k-150k", "150k-200k", "200k-250k"),
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, 200000, 250000),
labels = c("0-50k", "50k-100k", "100k-150k", "150k-200k", "200k-250k"),
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
# This plot shows the distribution of price for each mileage group
ggplot(audi_a1_data, aes(x = cut(mileage_km, breaks = c(0, 50000, 100000, 150000, 200000, 250000),
labels = c("0-50k", "50k-100k", "100k-150k", "150k-200k", "200k-250k"),
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()

#Scatter plot
ggplot(audi_a1_data, aes(x = mileage_km, y = price)) +
  geom_point(alpha = 0.6) +

```

```

geom_smooth(method = "lm", se = TRUE, color = "green") +
scale_x_continuous(labels = scales::comma, breaks = seq(0, 300000, by = 50000)) +
labs(
  title = "Price vs Mileage for Audi A1",
  x = "Mileage (km)",
  y = "Price (in units)"
) +
theme_minimal()
#Scatter plot
ggplot(opel_insignia_data, aes(x = mileage_km, y = price)) +
geom_point(alpha = 0.6) +
geom_smooth(method = "lm", se = TRUE, color = "blue") +
scale_x_continuous(labels = scales::comma, breaks = seq(0, 300000, by = 50000)) +
labs(
  title = "Price vs Mileage for Opel Insignia",
  x = "Mileage (km)",
  y = "Price (in units)"
) +
theme_minimal()

#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()

```

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

```

library(readr)
library(dplyr)
data <- read_csv("final_scout_not_dummy.csv")

```

```

library(tidyverse)

data <- read.csv("cleaned_scout_data.csv")

model_counts <- data %>%
  group_by(make_model) %>%
  summarise(count = n()) %>%
  arrange(desc(count))

top_3_models <- head(model_counts, 3)

library(ggplot2)
ggplot(data, aes(x = make_model, fill = make_model)) +
  geom_bar(color = "black") +
  geom_text(data = top_3_models, aes(x = make_model, y = count, label = count),
            vjust = -0.5, color = "black", size = 3) +
  labs(
    title = "Histogram of All Car Models",
    x = "Car Model",
    y = "Count"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

#Provides key 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)

audi_a3_data <- data %>% filter(make_model == "Audi A3")

# 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 for price by age for Audi A1
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()

audi_a3_data <- data %>% filter(make_model == "Audi A3")

# 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 for price by age for Audi A1
# 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()
#
# #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, 250000),
                              labels = c("0-50k", "50k-100k", "100k-150k", "150k-200k", "200k-250k"),
                              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, 200000, 250000),
                              labels = c("0-50k", "50k-100k", "100k-150k", "150k-200k", "200k-250k"),
                              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
# This plot shows the distribution of price for each mileage group
ggplot(audi_a1_data, aes(x = cut(mileage_km, breaks = c(0, 50000, 100000, 150000, 200000, 250000),
                              labels = c("0-50k", "50k-100k", "100k-150k", "150k-200k", "200k-250k"),
                              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()

```

```

#Scatter plot
ggplot(audi_a1_data, aes(x = mileage_km, y = price)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, color = "green") +
  scale_x_continuous(labels = scales::comma, breaks = seq(0, 300000, by = 50000)) +
  labs(
    title = "Price vs Mileage for Audi A1",
    x = "Mileage (km)",
    y = "Price (in units)"
  ) +
  theme_minimal()
#Scatter plot
ggplot(opel_insignia_data, aes(x = mileage_km, y = price)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, color = "blue") +
  scale_x_continuous(labels = scales::comma, breaks = seq(0, 300000, by = 50000)) +
  labs(
    title = "Price vs Mileage for Opel Insignia",
    x = "Mileage (km)",
    y = "Price (in units)"
  ) +
  theme_minimal()

#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()

```

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

```

library(readr)
library(dplyr)
data <- read.csv("final_scout_not_dummy.csv")

```

```

library(tidyverse)

data <- read.csv("cleaned_scout_data.csv")

model_counts <- data %>%
  group_by(make_model) %>%
  summarise(count = n()) %>%
  arrange(desc(count))

top_3_models <- head(model_counts, 3)

library(ggplot2)
ggplot(data, aes(x = make_model, fill = make_model)) +
  geom_bar(color = "black") +
  geom_text(data = top_3_models, aes(x = make_model, y = count, label = count),
            vjust = -0.5, color = "black", size = 3) +
  labs(
    title = "Histogram of All Car Models",
    x = "Car Model",
    y = "Count"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

#Provides key 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)

audi_a3_data <- data %>% filter(make_model == "Audi A3")

# 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 for price by age for Audi A1
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()

audi_a3_data <- data %>% filter(make_model == "Audi A3")

# 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 for price by age for Audi A1
# 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()
#
# #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, 250000),
                             labels = c("0-50k", "50k-100k", "100k-150k", "150k-200k", "200k-250k")),
                        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, 200000, 250000),
                             labels = c("0-50k", "50k-100k", "100k-150k", "150k-200k", "200k-250k")),
                        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
# This plot shows the distribution of price for each mileage group
ggplot(audi_a1_data, aes(x = cut(mileage_km, breaks = c(0, 50000, 100000, 150000, 200000, 250000),
                             labels = c("0-50k", "50k-100k", "100k-150k", "150k-200k", "200k-250k")),
                        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()

```



```

#Scatter plot
ggplot(audi_a1_data, aes(x = mileage_km, y = price)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, color = "green") +
  scale_x_continuous(labels = scales::comma, breaks = seq(0, 300000, by = 50000)) +
  labs(
    title = "Price vs Mileage for Audi A1",
    x = "Mileage (km)",
    y = "Price (in units)"
  ) +
  theme_minimal()
#Scatter plot
ggplot(opel_insignia_data, aes(x = mileage_km, y = price)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, color = "blue") +
  scale_x_continuous(labels = scales::comma, breaks = seq(0, 300000, by = 50000)) +
  labs(
    title = "Price vs Mileage for Opel Insignia",
    x = "Mileage (km)",
    y = "Price (in units)"
  ) +
  theme_minimal()

#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()

```

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 Lastly 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

```
library(readr)
library(dplyr)
data <- read_csv("final_scout_not_dummy.csv")

library(tidyverse)

data <- read_csv("cleaned_scout_data.csv")

model_counts <- data %>%
  group_by(make_model) %>%
  summarise(count = n()) %>%
  arrange(desc(count))

top_3_models <- head(model_counts, 3)

library(ggplot2)
ggplot(data, aes(x = make_model, fill = make_model)) +
  geom_bar(color = "black") +
  geom_text(data = top_3_models, aes(x = make_model, y = count, label = count),
            vjust = -0.5, color = "black", size = 3) +
  labs(
    title = "Histogram of All Car Models",
    x = "Car Model",
    y = "Count"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

#Provides key 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)

audi_a3_data <- data %>% filter(make_model == "Audi A3")

# 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 for price by age for Audi A1

```

```

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()

audi_a3_data <- data %>% filter(make_model == "Audi A3")

# 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 for price by age for Audi A1
# 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()
#
# #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, 250000),
                              labels = c("0-50k", "50k-100k", "100k-150k", "150k-200k", "200k-250k")),
                        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, 200000),
                              labels = c("0-50k", "50k-100k", "100k-150k", "150k-200k", "200k-250k")),
                        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
# This plot shows the distribution of price for each mileage group
ggplot(audi_a1_data, aes(x = cut(mileage_km, breaks = c(0, 50000, 100000, 150000, 200000, 250000),
labels = c("0-50k", "50k-100k", "100k-150k", "150k-200k", "200k-250k"),
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()

#Scatter plot
ggplot(audi_a1_data, aes(x = mileage_km, y = price)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, color = "green") +
  scale_x_continuous(labels = scales::comma, breaks = seq(0, 300000, by = 50000)) +
  labs(
    title = "Price vs Mileage for Audi A1",
    x = "Mileage (km)",
    y = "Price (in units)"
  ) +
  theme_minimal()

#Scatter plot
ggplot(opel_insignia_data, aes(x = mileage_km, y = price)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, color = "blue") +
  scale_x_continuous(labels = scales::comma, breaks = seq(0, 300000, by = 50000)) +
  labs(
    title = "Price vs Mileage for Opel Insignia",
    x = "Mileage (km)",
    y = "Price (in units)"
  ) +
  theme_minimal()

#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()

```

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**

```

library(readr)
library(dplyr)
data <- read_csv("final_scout_not_dummy.csv")

library(tidyverse)

data <- read_csv("cleaned_scout_data.csv")

model_counts <- data %>%
  group_by(make_model) %>%
  summarise(count = n()) %>%
  arrange(desc(count))

top_3_models <- head(model_counts, 3)

library(ggplot2)
ggplot(data, aes(x = make_model, fill = make_model)) +
  geom_bar(color = "black") +
  geom_text(data = top_3_models, aes(x = make_model, y = count, label = count),
            vjust = -0.5, color = "black", size = 3) +
  labs(
    title = "Histogram of All Car Models",
    x = "Car Model",
    y = "Count"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

#Provides key 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)

audi_a3_data <- data %>% filter(make_model == "Audi A3")

# 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 for price by age for Audi A1
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()

audi_a3_data <- data %>% filter(make_model == "Audi A3")

# 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 for price by age for Audi A1
# 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()
#
# #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, 250000),
                           labels = c("0-50k", "50k-100k", "100k-150k", "150k-200k", "200k-250k")),
                        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, 200000),
                                       labels = c("0-50k", "50k-100k", "100k-150k", "150k-200k", "200k-250k")),
                             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
# This plot shows the distribution of price for each mileage group
ggplot(audi_a1_data, aes(x = cut(mileage_km, breaks = c(0, 50000, 100000, 150000, 200000, 250000),
labels = c("0-50k", "50k-100k", "100k-150k", "150k-200k", "200k-250k")),
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()

#Scatter plot
ggplot(audi_a1_data, aes(x = mileage_km, y = price)) +
geom_point(alpha = 0.6) +
geom_smooth(method = "lm", se = TRUE, color = "green") +
scale_x_continuous(labels = scales::comma, breaks = seq(0, 300000, by = 50000)) +
labs(
  title = "Price vs Mileage for Audi A1",
  x = "Mileage (km)",
  y = "Price (in units)"
) +
theme_minimal()

#Scatter plot
ggplot(opel_insignia_data, aes(x = mileage_km, y = price)) +
geom_point(alpha = 0.6) +
geom_smooth(method = "lm", se = TRUE, color = "blue") +
scale_x_continuous(labels = scales::comma, breaks = seq(0, 300000, by = 50000)) +
labs(
  title = "Price vs Mileage for Opel Insignia",
  x = "Mileage (km)",
  y = "Price (in units)"
) +
theme_minimal()

#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()
```

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 Lastly for the Opel Insignia the coefficient was -6535.30 decreasing at a rate of about 6,500 Euro per year