Used Car Prices in New York and Chicago

Shaidatullisa Nadia Saipudin, Xinyi Wang, Jingchun Zhang

2023-11-28

Introduction

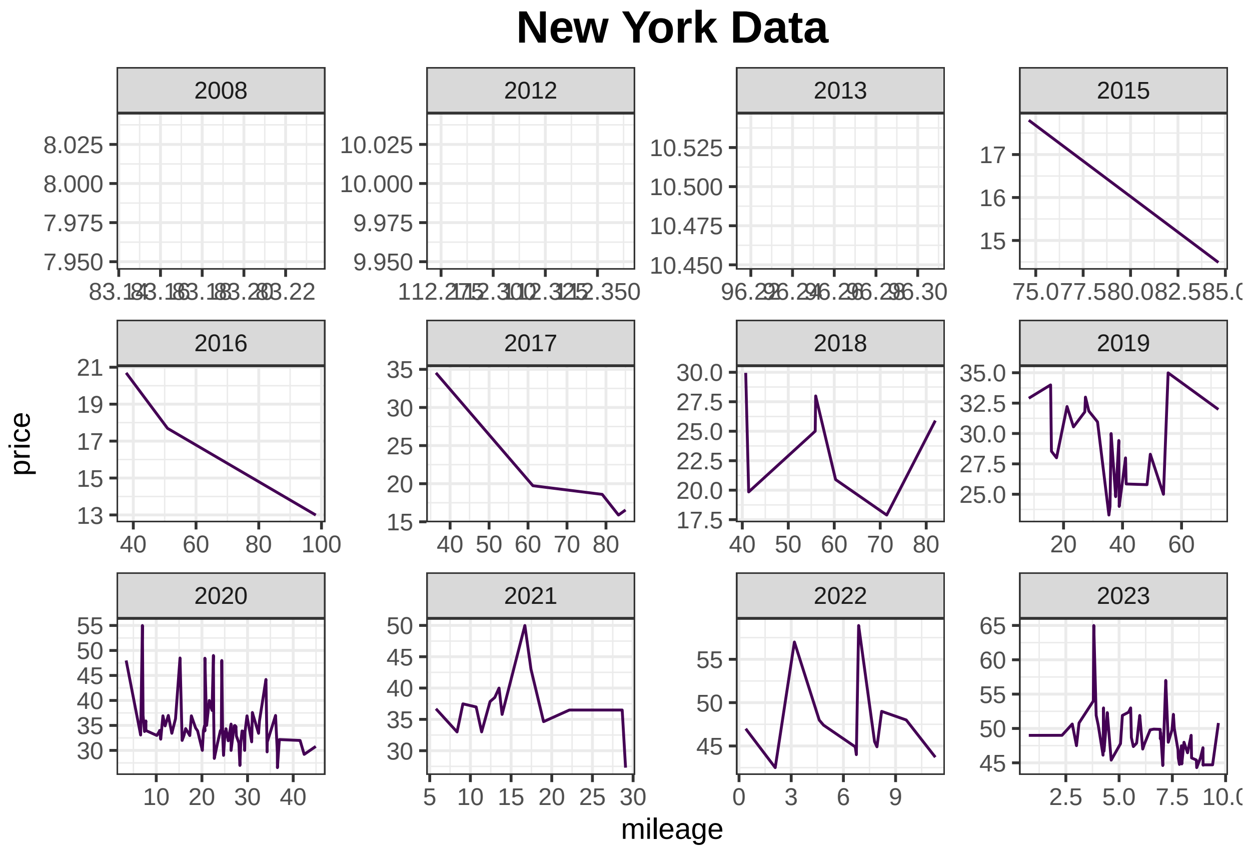
In the United States, the convenience of a car is obvious, whether it's for everyday shopping (at Walmart, for example), moving, or long-distance traveling between cities. To travel from State College to New York, for example, it takes approximately 7 hours by bus, while driving your own car takes only 4 hours. Recognizing the importance of owning a car, we launched this project with the goal of delving into the many factors that influence the pricing of used cars.

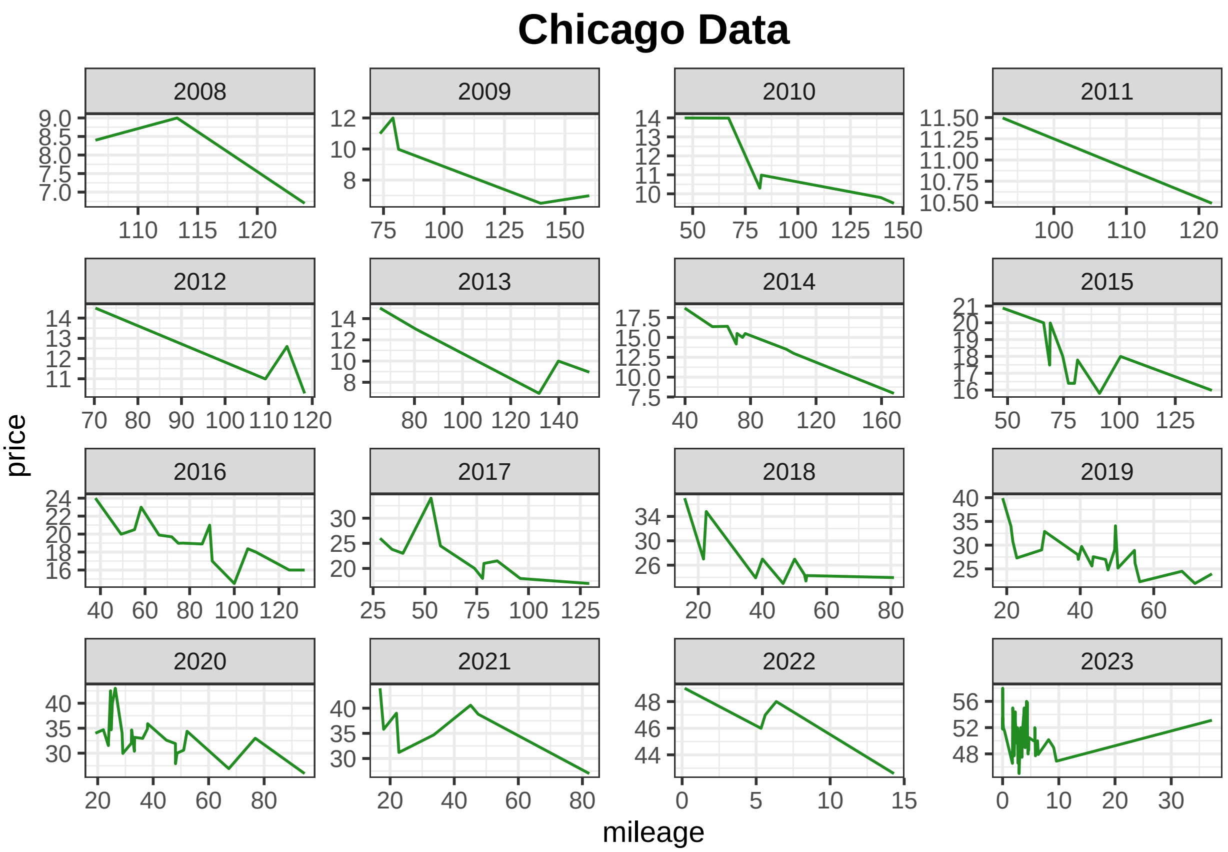
The starting point for our research was to screen and identify data sources that could provide comprehensive and reliable information on automobile pricing. After carefully evaluating numerous sources, we selected a dataset obtained from the website **"myslu.stlawu.edu/~clee/dataset/autotrader/"**, which is a compilation of used car information listed on the autotrader.com website and is available in CSV format. The dataset encompasses a wide range of vehicle models, a wide range of years, and information on different geographic locations, providing a rich data base for our analysis.

For our study, we selected the popular Mercedes-Benz C300 model in the market as the study vehicle, a choice that was made due to its support from official repair stores around State College and the popularity of the model. This enabled us to have a uniform and stable benchmark for comparison when making price comparisons. Considering the potential impact of urban environments on automobile prices, we chose New York and Chicago, two iconic U.S. metropolises, whose level of modernization and size provide us with similar and comparable study conditions, as the subject of comparison.

In this study, we will try to analyze the data to **study how age and mileage affect the pricing of used cars and the difference between Chicago and New York to further understand the market.**

Data preview





After performing an initial visualization and analysis of the collected dataset, we gained a preliminary understanding of the information contained in the dataset. The visual analysis described above revealed to us the relationship between vehicle selling prices and mileage for different years in New York and Chicago. Through careful observation, we found that when attempting to plot an image of the New York dataset, a warning message appeared in the system. We found that there were gaps in the data prior to 2015 in the New York dataset, with some years having no record of data, and others having only a single record of data.

The analysis of the Chicago dataset showed a more consistent trend of the expected downward trend in vehicle prices as vehicle mileage increased. The absence of warning messages in the Chicago data suggests that enough observations are included in the data for each year to allow us to construct a complete line graph without difficulty.

After having this basic data in place, we began further cleaning and in-depth analysis of the data.

Data combination & cleaning

To ensure accuracy and convenience in our comparative analysis of car prices across different locations, we conducted a series of precise preprocessing steps on the original dataset. Initially, exact geographical tags were appended to each entry, maintaining clear distinction of each vehicle's origin post-merger of the datasets. Subsequently, we consolidated the disparate datasets into a unified data framework, facilitating a centralized analysis.

This integration enables us not only to discern pricing patterns across different geographic locales but also to identify any anomalies, allowing for an in-depth exploration of the pricing dynamics of the Mercedes-Benz C300 in various market conditions. Additionally, we introduced the variable of 'vehicle year' into our dataset, calculating the actual age of each car to add an extra dimension to our analysis. This incorporation allows us to account for the impact of vehicle age on the pricing in distinct city markets such as New York and Chicago.

In response to previous findings of gaps in the data for New York prior to 2015, after integrating the data, we cleaned the data and removed all data before 2015. We will follow up with further analysis focusing on the data from 2015 to the present.

```
##   Location   price    Age mileage
## 1  Chicago 36.5696 2.7468 33.2750
## 2 New York 38.0554 2.2872 21.1223
```

This summary table is a concise representation of the key variables from the combined dataset, showcasing the average values of price, age, and mileage for the Mercedes-Benz C 300 cars in both Chicago and New York. Each row represents one of the two locations, with the columns displaying the mean of the respective variables for the cars sold in that city.

The combined data contains the following relevant information:

Location: This is a character column that indicates the city from which the data was collected, allowing us to compare the two distinct markets.

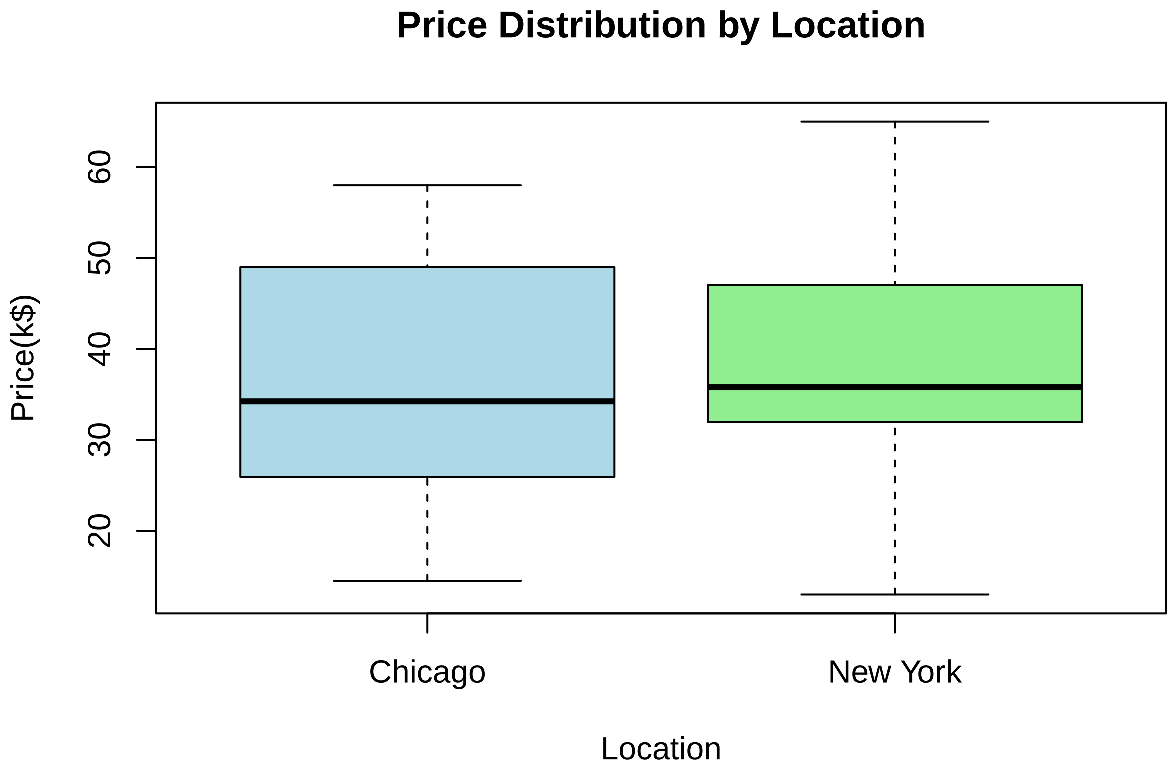
Price: Displayed as a double precision number (floating-point), this column represents the average sale price of the cars in thousands of dollars. The prices are notably higher in New York, with an average of 38.05539 compared to Chicago's 36.56958.

Age: Also a double precision number, this indicates the average age of the cars at the time of sale. On average, cars in New York are younger at 2.29 years compared to Chicago's 2.75 years.

Mileage: This double precision number represents the average mileage (in thousands) on the cars when they were sold. Cars in New York have significantly lower mileage, with an average of 21.1222, suggesting they may be newer on average when sold or used less intensively before the sale compared to those in Chicago, which have an average mileage of 33.275.

Price Distribution by Location

In order to meticulously analyze the price distribution of cars in both New York and Chicago, we used box plots for visualization.



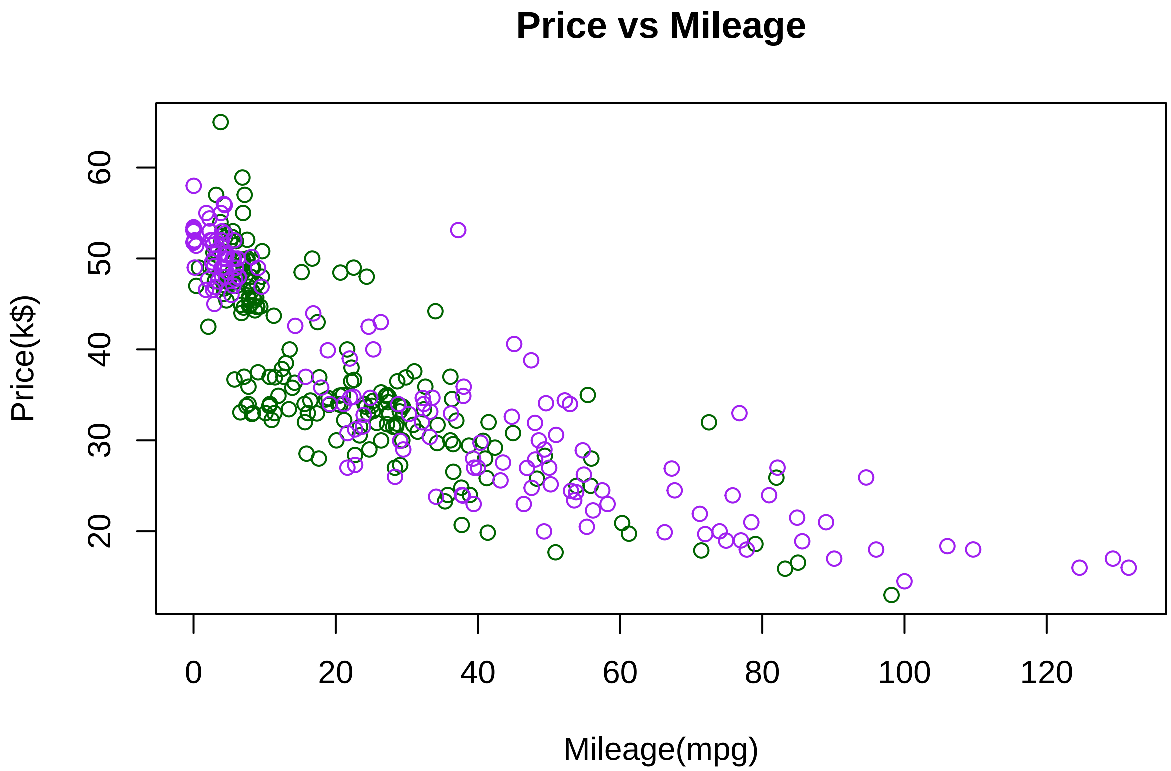
As observed from the box plots, the median price of the Mercedes-Benz C300 in the New York market is significantly higher than that in Chicago, and the quartile spacing of its price distribution is wider, which indicates that the price volatility range is larger in New York.

More strikingly, the box plot for the New York market shows multiple high outliers, implying that a subset of Mercedes-Benz C300s are selling for prices well above the region's regular price levels. These outliers may be due to the fact that some of these vehicles are more luxuriously equipped.

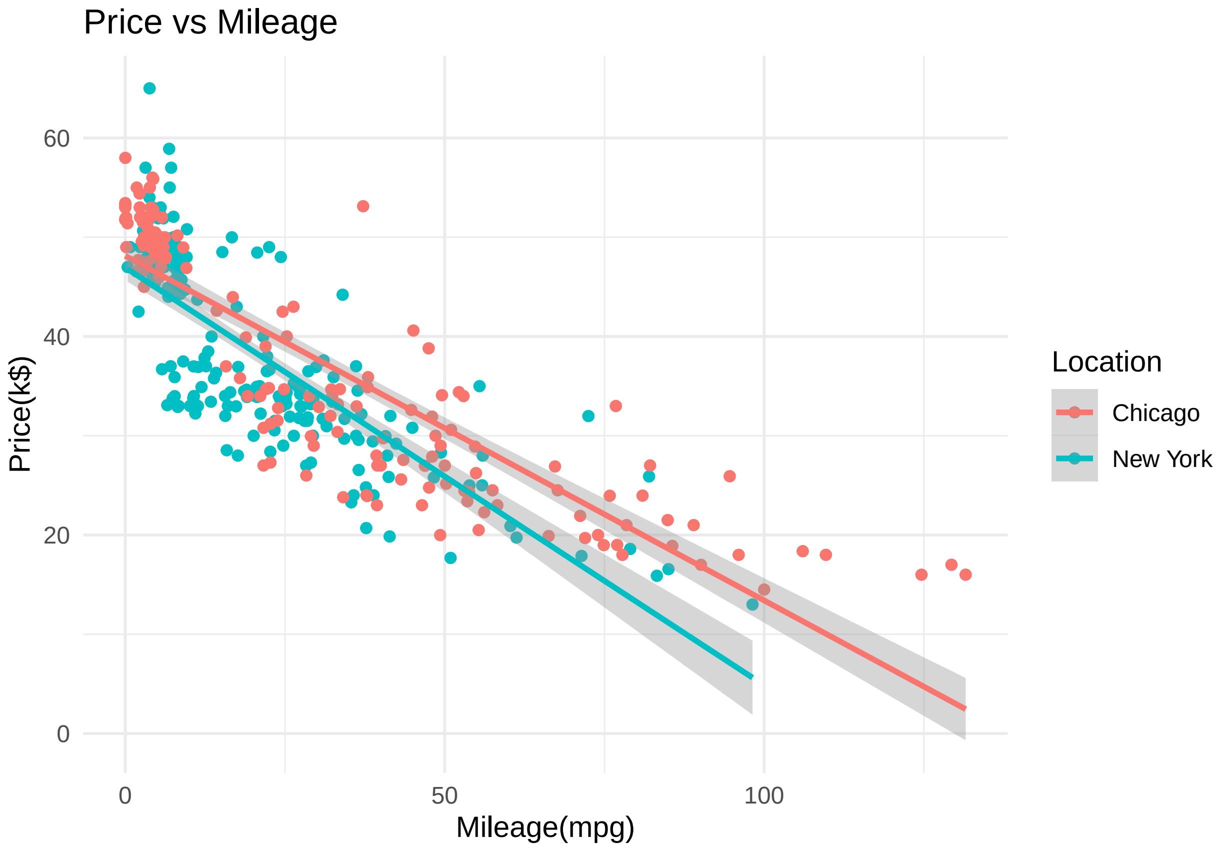
The box plot for the Chicago market is relatively compact, suggesting that Mercedes-Benz C300 prices in the region are relatively concentrated and less volatile. This may reflect the fact that the Chicago market is more competitive in terms of vehicle prices or that consumers are more price sensitive.

Price vs Mileage

After analyzing the distribution of vehicle prices in the two regions and making initial comparisons, our investigation proceeded to examine the impact of various factors on vehicle pricing. We began to evaluate the impact of mileage on vehicle costs. The purpose of this analysis was to determine if there is a correlation between a vehicle's accumulated mileage and its market value, thereby providing insight into how depreciation affects vehicle pricing in these metropolitan areas.



```
## `geom_smooth()` using formula = 'y ~ x'
```

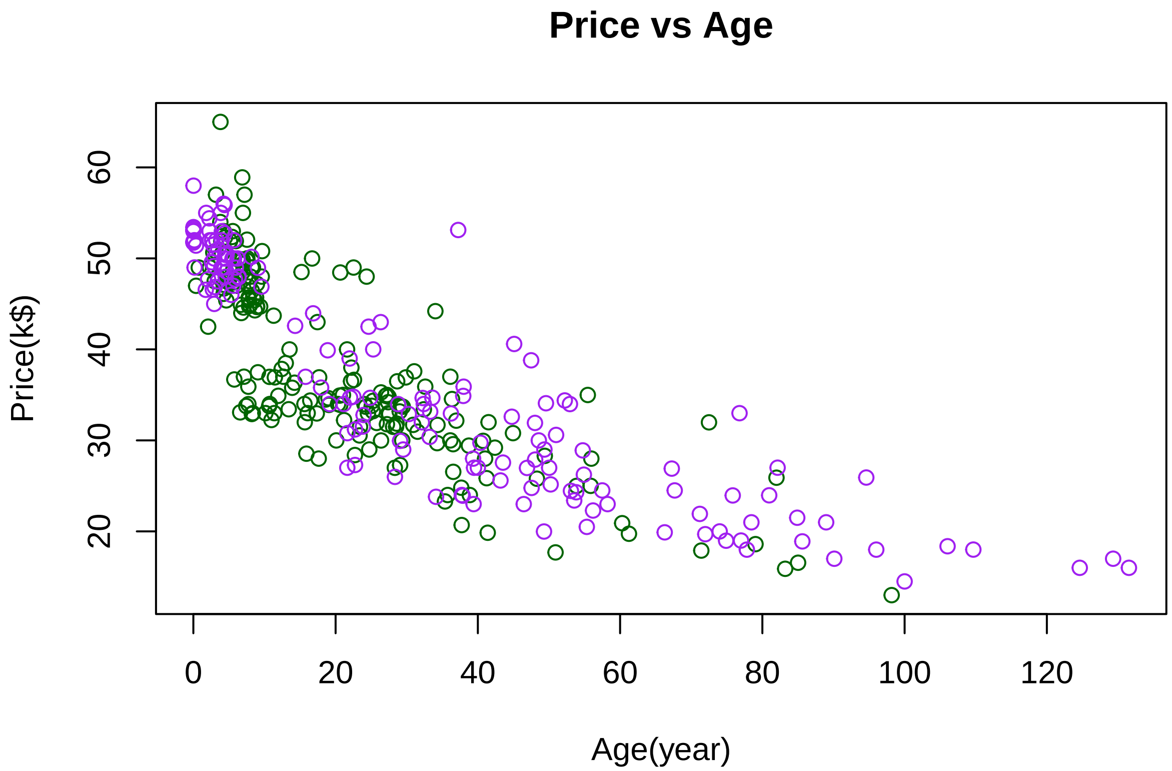


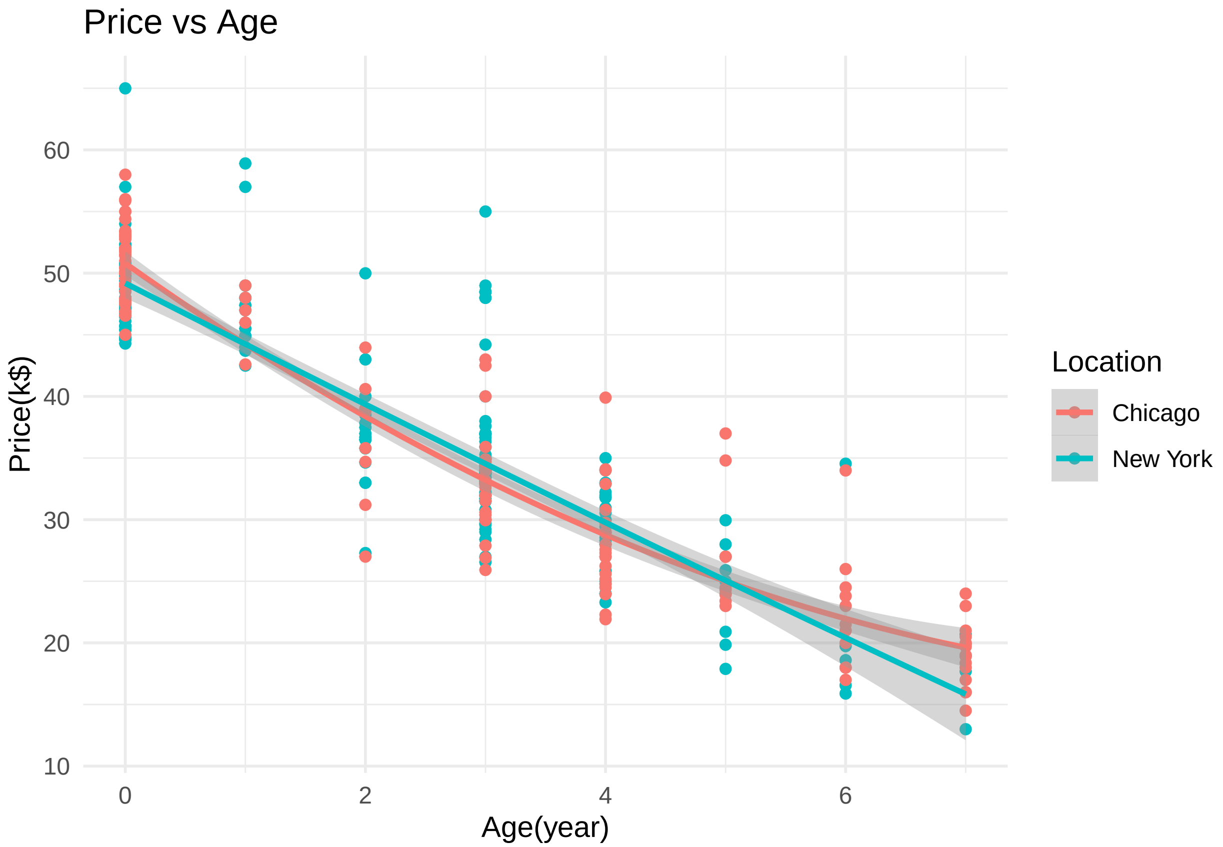
The scatter plots and the fitted linear model curves demonstrate the relationship between mileage and price for cars sold in both New York and Chicago. In both cities, there is a clear negative relationship: as mileage increases, the price decreases. This trend is consistent with the strong negative correlation found in the statistical analysis. The fitted linear model lines on the scatter plot also show a clear downward trend, reinforcing the idea that higher mileage is associated with lower car prices. This suggests that mileage is a significant predictor of the price of a Mercedes-Benz C 300, regardless of the city.

Price vs Age

Our third analysis dives into a key factor affecting how much a used Mercedes-Benz C300 costs—how old the car is. We're looking at this to help folks thinking about buying or selling this car figure out the best time to do it based on their money situation. The goal is to give people smart insights so they can make good choices and even predict the car's price depending on its age.

This kind of analysis is useful for both buyers and sellers. It helps everyone understand the connection between a car's age and how much it's worth in the market. By clarifying this link, people can navigate the tricky world of selling and buying used cars, choosing the right time to get involved based on what's happening in the market and their own financial picture.


Price vs Age
Price(k$)
20
30
40
50
60
Age(year)
0
20
40
60
80
100
120



The scatter plots and fitted linear model above shows the relationship between the price and age of the used Mercedes-Benz C300 in New York and Chicago. As shown in the graph, there is a clear negative relationship: as age increases, the price decreases in both cities. This trend is consistent with the strong negative correlation found in the statistical analysis. The fitted linear model lines on the scatter plot also show a clear downward trend, reinforcing the idea that the higher age of car is associated with lower car prices. In addition, the rate for the price to decrease is also similar for both New York and Chicago which indicate that the pattern is uniform throughout the years. This suggests that age is a significant predictor of the price of a Mercedes-Benz C 300, regardless of the city and there is not much difference in the price decreasing rate.Therefore, the buyer can choose the right time for them to buy the car according to the age of the car that may influence other factors according to their own preferences and financial situation.

Mileage vs Age vs Price

Within this segment, we will explore the interconnections among mileage, age, and price to discern the relative strengths of correlation between these variables. This examination aims to provide valuable insights for conducting a comprehensive analysis of the relationships among these three factors.

```
##          mileage     Age   price
## mileage  1.0000  0.8144 -0.8250
## Age      0.8144  1.0000 -0.9138
## price   -0.8250 -0.9138  1.0000
```

The presented correlation matrix delineates the interrelationships among the variables ‘mileage’, ‘Age’, and ‘price’. This matrix reveals the correlation coefficients, which quantify the strength and direction of linear associations between pairs of variables.

The correlation coefficient between **‘mileage’ and ‘Age’** is 0.8144, indicating a strong positive correlation. This implies that as the mileage of the vehicles increases, there is a substantial tendency for the age of the cars to also rise.

The correlation coefficient between **‘mileage’ and ‘price’** is -0.8250, reflecting a robust negative correlation. This signifies that as the mileage of the vehicles increases, there is a noteworthy inclination for the price of the cars to decrease.

The correlation coefficient between **‘Age’ and ‘price’** is -0.9138, denoting a notably strong negative correlation. This underscores the discernible trend that as the age of the cars increases, there is a substantial proclivity for the prices to diminish.

The overall pattern in the correlation matrix indicates a clear and logically consistent set of relationships between the considered variables. The negative correlations observed between ‘mileage’ and both ‘Age’ and ‘price’ align with expectations in the context of used cars. Furthermore, the pronounced negative correlation between ‘Age’ and ‘price’ underscores the impact of aging on the pricing dynamics of the examined vehicles.

Conclusion

In conclusion, this report provides a comprehensive analysis of used Mercedes-Benz C300 car prices in New York and Chicago. It highlights how factors like age and mileage significantly affect car prices, with a notable negative correlation between mileage and price. The analysis reveals that cars in New York are generally priced higher and exhibit greater price variability than in Chicago, potentially due to differences in market demand and condition of the cars. Additionally, the cars in New York tend to be newer and have lower mileage. These insights are crucial for buyers and sellers in the used car market, offering a deeper understanding of how various factors influence car pricing in different urban markets.

Code Appendix

```r
knitr::opts_chunk$set(echo = FALSE)
# Libraries and datasets already loaded
NYdata <- read.csv("C:/Users/User/Downloads/C300NY.csv")
ILdata <- read.csv("C:/Users/User/Downloads/C300IL.csv")
library(ggplot2)

# New York data preview
ggplot(NYdata) +
  aes(x = mileage, y = price, group = year) +  # Added group aesthetic
  geom_line(colour = "#440154") +
  labs(title = "New York Data") +
  theme_bw() +
  theme(
    plot.title = element_text(size = 17L,
                              face = "bold",
                              hjust = 0.5)
  ) +
  facet_wrap(vars(year), scales = "free")

#Chicago data preview
ggplot(ILdata) +
  aes(x = mileage, y = price) +
  geom_line(colour = "#228B22") +
  labs(title = "Chicago Data") +
  theme_bw() +
  theme(
    plot.title = element_text(size = 16L,
    face = "bold",
    hjust = 0.5)
  ) +
  facet_wrap(vars(year), scales = "free")
# Data preprocessing
# Adding location labels
NYdata$Location <- 'New York'
ILdata$Location <- 'Chicago'

# Merging the datasets
combined_data <- rbind(NYdata, ILdata)

# Calculating the age of the car
combined_data$Age <- 2023 - combined_data$year

# Data cleaning
combined_data <- combined_data[combined_data$year > 2015, ]

# Creating a summary table
library(dplyr)
summary_table <- aggregate(cbind(price, Age, mileage) ~ Location, data = combined_data, mean) %>%
  mutate_at(vars(Age, mileage, price), ~round(., 4))

print(summary_table)
```

```r
# Create a boxplot - Price distribution by location
boxplot(price ~ Location, data = combined_data,
        main = "Price Distribution by Location",
        xlab = "Location", ylab = "Price(k$)",
        col = c('lightblue', 'lightgreen'))
# Plotting price vs mileage scatter plot
plot(combined_data$mileage, combined_data$price,
     main = "Price vs Mileage",
     xlab = "Mileage(mpg)", ylab = "Price(k$)",
     col = ifelse(combined_data$Location == "New York", 'darkgreen', 'purple'))

# Adding a smooth curve
library(ggplot2)
ggplot(combined_data, aes(x = mileage, y = price, color = Location)) +
  geom_point() +
  geom_smooth(method = "lm") +
  theme_minimal() +
  labs(title = "Price vs Mileage", x = "Mileage(mpg)", y = "Price(k$)")
# Plotting price vs age scatter plot
plot(combined_data$mileage, combined_data$price,
     main = "Price vs Age",
     xlab = "Age(year)", ylab = "Price(k$)",
     col = ifelse(combined_data$Location == "New York", 'darkgreen', 'purple'))

# Adding a smooth curve
library(ggplot2)
ggplot(combined_data, aes(x = Age, y = price, color = Location)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ poly(x, 2)) +
  theme_minimal() +
  labs(title = "Price vs Age", x = "Age(year)", y = "Price(k$)")



# Create a correlation table
cor_table <- cor(combined_data[, c("mileage", "Age", "price")])

# Round the correlation matrix to 4 decimal places
rounded_cor_table <- round(cor_table, 4)

# Print the rounded correlation matrix
print(rounded_cor_table)
```