Exploratory data analysis:
CAR DATA

TANISHA SINGHANIA
(24057)

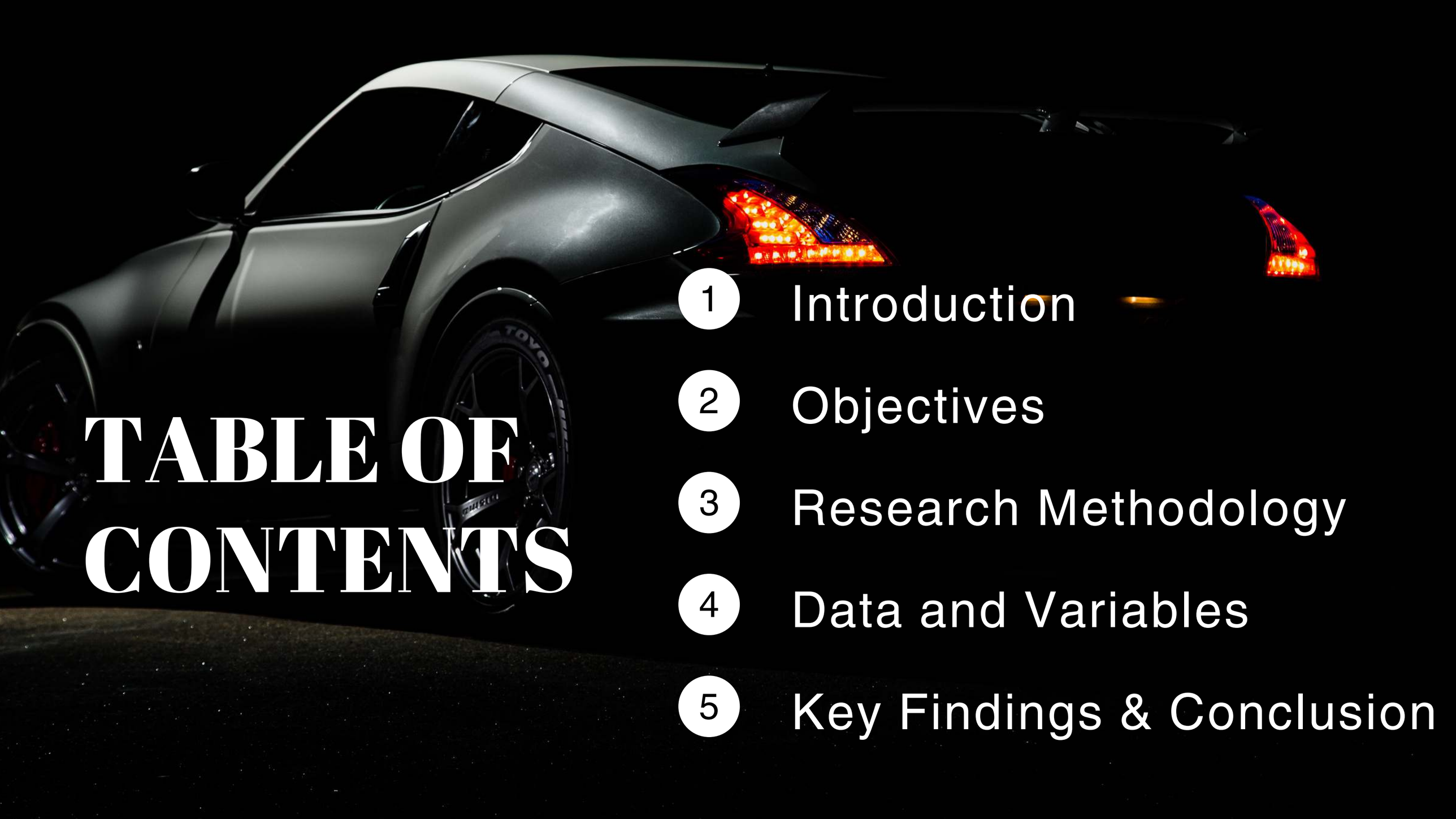


TABLE OF CONTENTS

1. Introduction
2. Objectives
3. Research Methodology
4. Data and Variables
5. Key Findings & Conclusion

INTRODUCTION

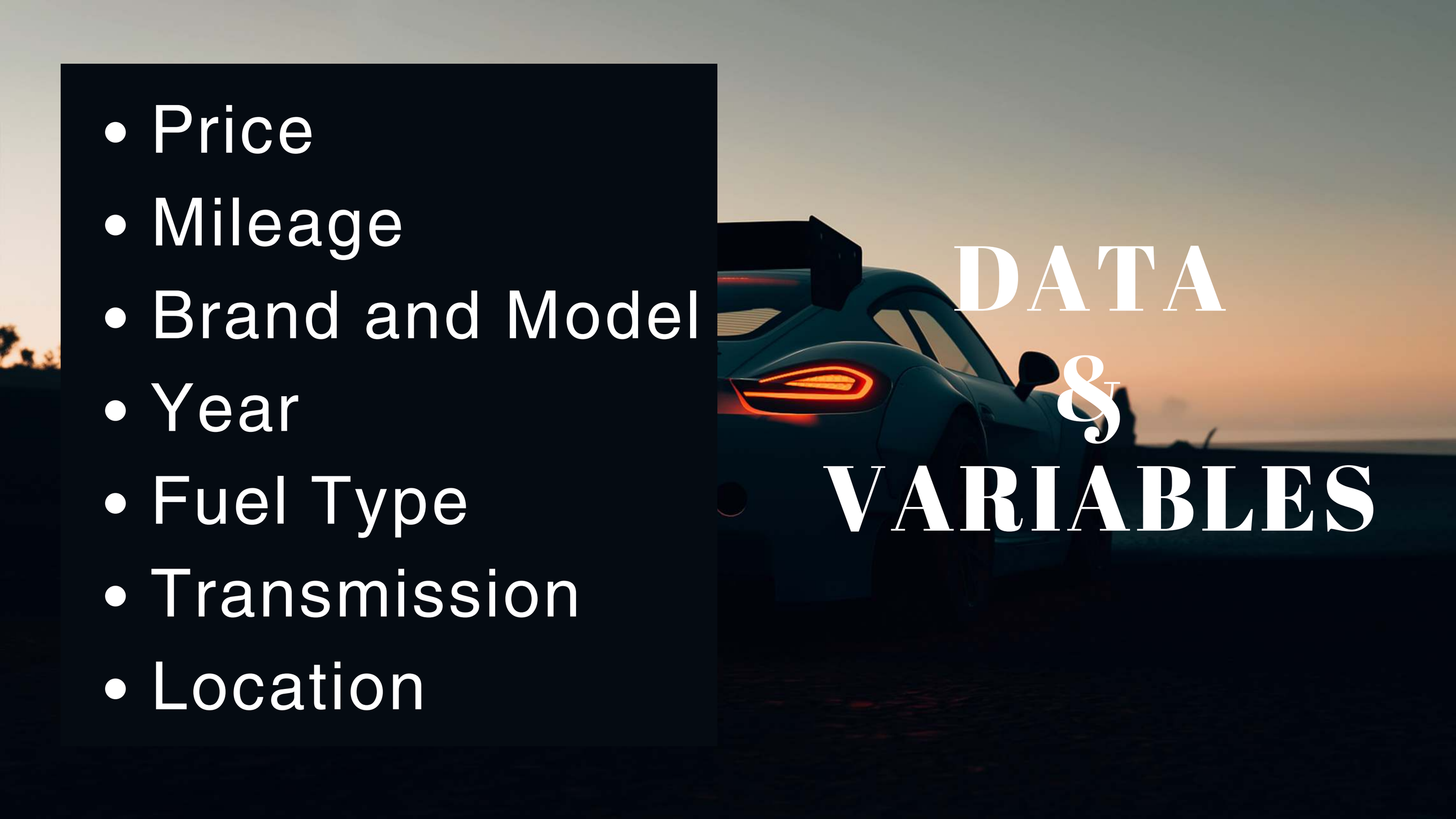
This presentation aims to analyze a dataset on used cars, exploring patterns, trends, and relationships among various attributes such as price, mileage, brand, and fuel type. The analysis provides insights into the used car market, identifying factors that influence prices and highlighting patterns in customer preferences.

OBJECTIVES

1. Analyze the distribution of used cars by brand, model, year, and fuel type.
2. Identify key factors influencing car prices, such as mileage, brand, and year.
3. Discover preferred car models, fuel types, and transmission types.
4. Examine trends in prices, mileage, and listing volume over time.
5. Provide insights to guide buyers, sellers, and dealers in the market.

RESEARCH METHODOLOGY

- **Data Collection:** A dataset of used cars was gathered, covering features like price, mileage, brand etc.
- **Data Preprocessing:** Missing values were addressed, inconsistencies were cleaned.
- **Data Analysis:** Descriptive analysis identified patterns. Visualizations were used to aid interpretation.
- **Validation and Insights**: Statistical tests validated trends, and business implications were drawn from the findings.



DATA & VARIABLES

- Price
- Mileage
- Brand and Model
- Year
- Fuel Type
- Transmission
- Location

KEY FINDINGS

Average car price by location

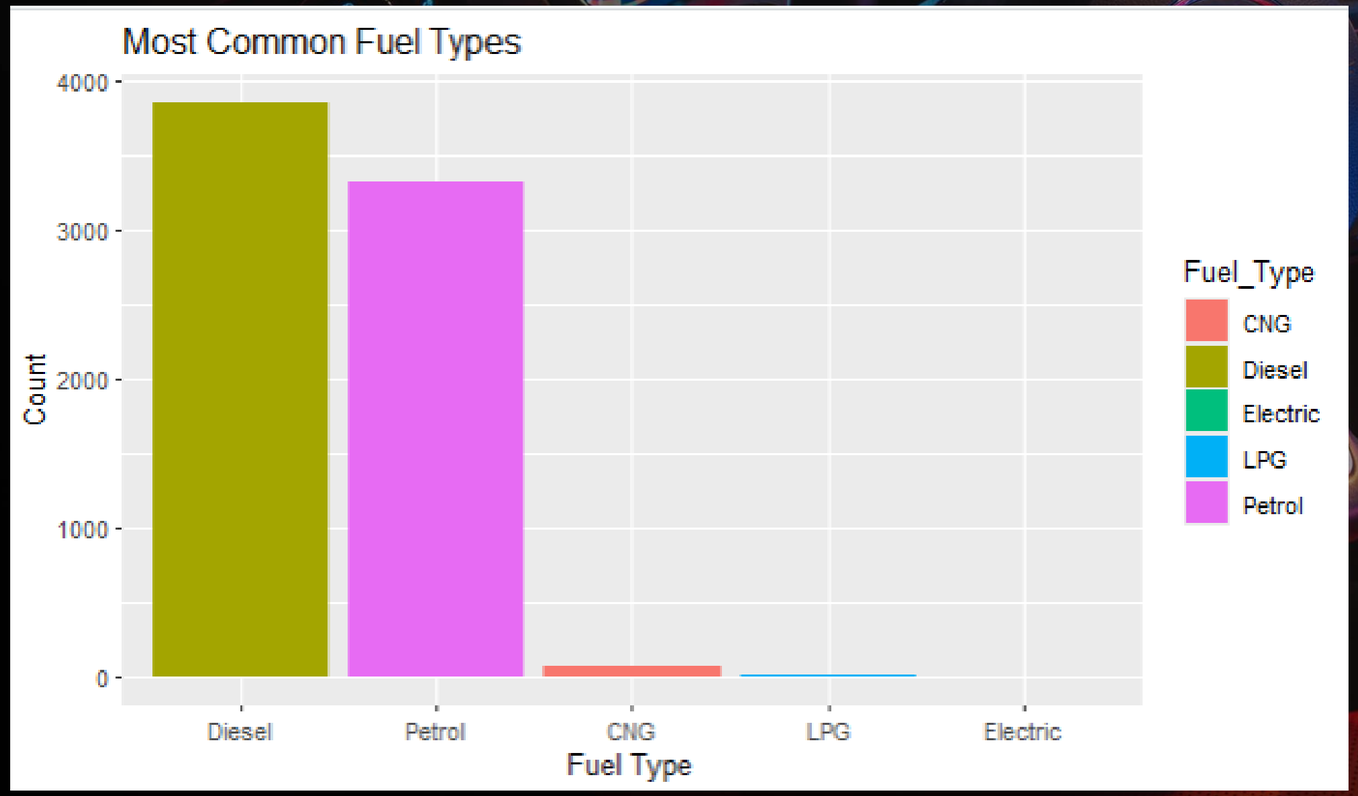
```
avg_price_by_location <- cars_data %>%
  group_by(Location) %>%
  summarize(avg_price = mean(Price, na.rm =
TRUE)) %>%
  arrange(desc(avg_price))
View(avg_price_by_location)
```

Most Common Fuel Type

```
fuel_counts <- cars_data %>%
 count(Fuel_Type) %>%
 arrange(desc(n))
View(fuel_counts)
```

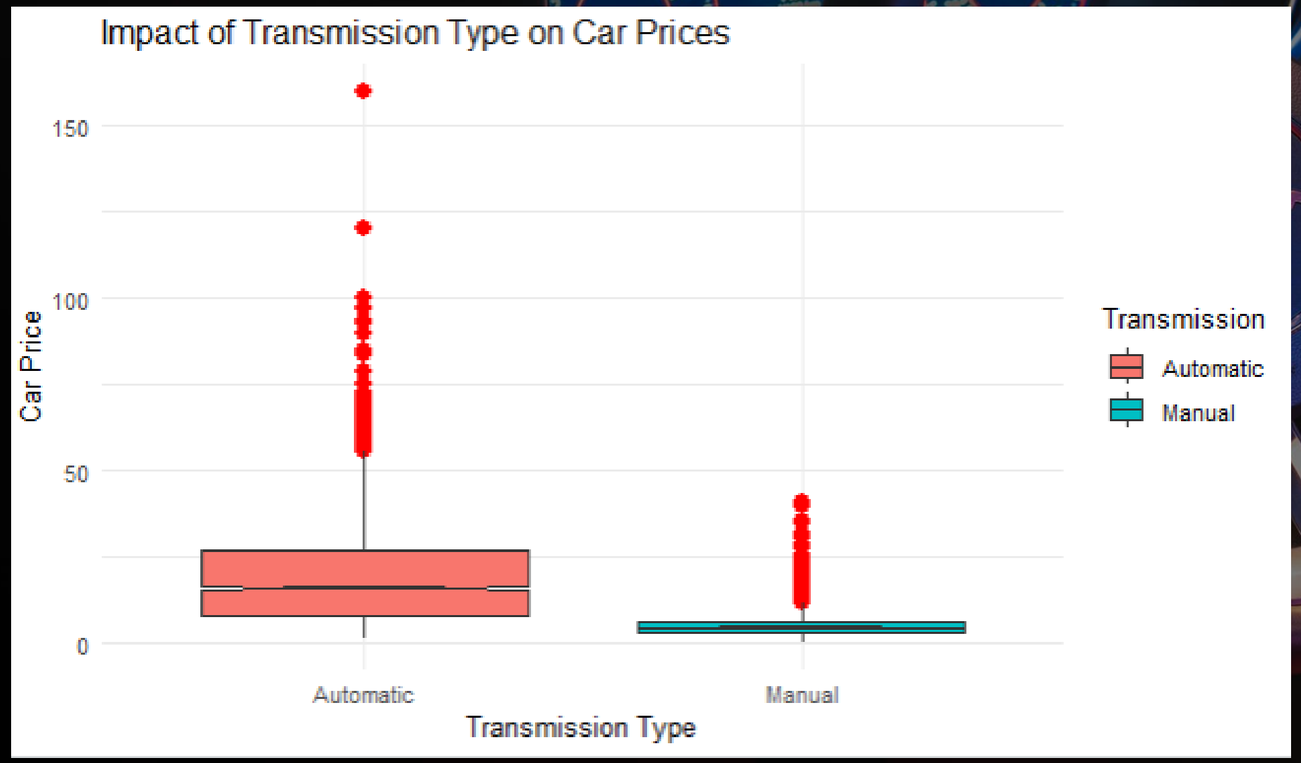
	Fuel_Type	n
1	Diesel	3852
2	Petrol	3325
3	CNG	62
4	LPG	12
5	Electric	2

Bar plot:
Most common fuel type




```
ggplot(fuel_counts, aes(x =
reorder(Fuel_Type, -n), y = n,
fill = Fuel_Type)) +
  geom_bar(stat = "identity") +
  labs(title = "Most Common
Fuel Types", x = "Fuel Type",
y = "Count")
```

Impact of Transmission Type on Car Prices




```
transmission_impact <-
cars_data %>%
  group_by(Transmission) %>%
      summarize(avg_price =
mean(Price, na.rm = TRUE))
View(transmission_impact)
```

Impact of owner type on avg price

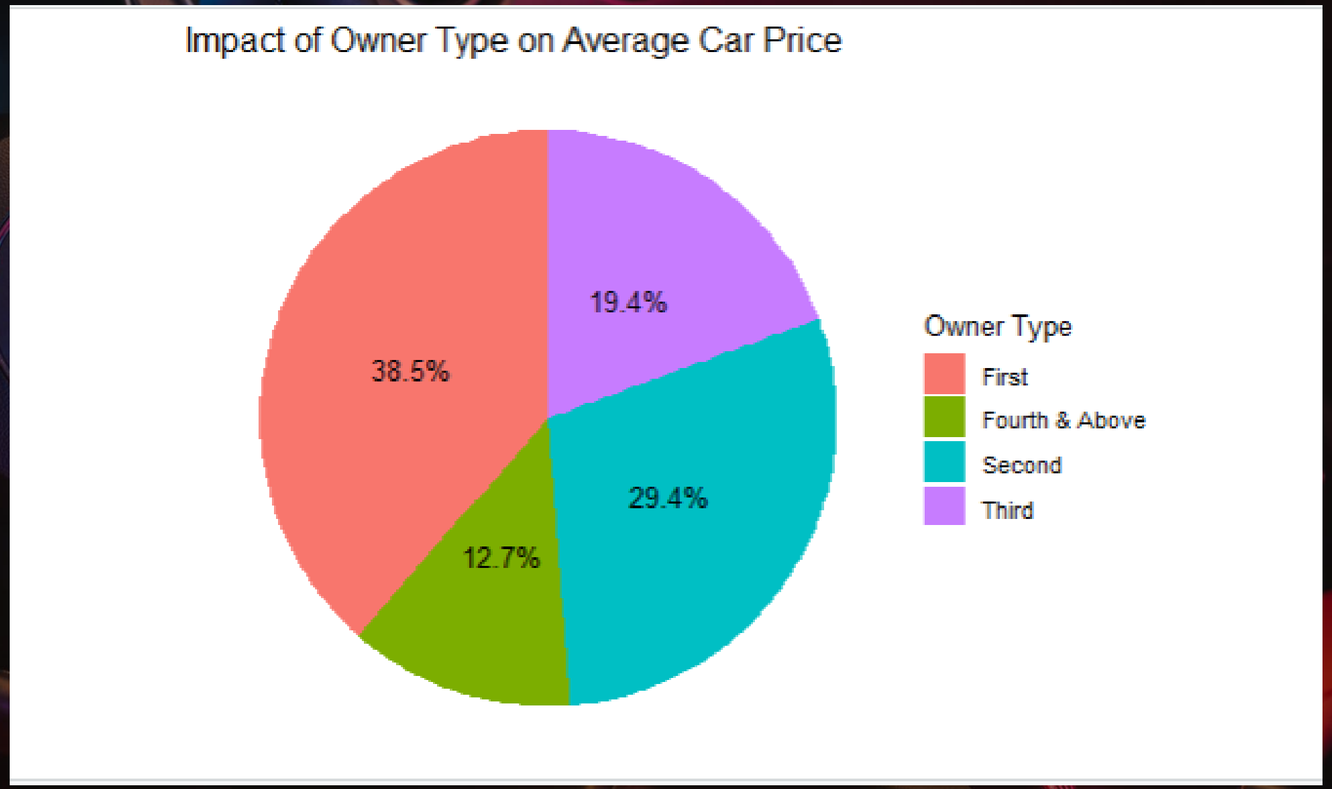
```
owner_price <- cars_data
%>%
group_by(Owner_Type)
%>%
summarize(avg_price =
mean(Price, na.rm = TRUE))
View(owner_price)
```

```
# Add a percentage column
for the pie chart
owner_price <- owner_price
%>%
mutate(percentage =
round((avg_price /
sum(avg_price)) * 100, 1))
```

	Owner_Type	avg_price
1	First	9.962445
2	Fourth & Above	3.280000
3	Second	7.599886
4	Third	5.007257

Pie Chart:
Impact of owner type on avg price

```
ggplot(owner_price, aes(x = "", y = avg_price,
fill = Owner_Type)) +
geom_bar(stat = "identity", width = 1) +
coord_polar("y", start = 0) +
labs(
title = "Impact of Owner Type on Average Car
Price",
fill = "Owner Type"
) +
geom_text(aes(label = paste0(percentage,
"%")),
position = position_stack(vjust = 0.5)) +
theme_void()
```

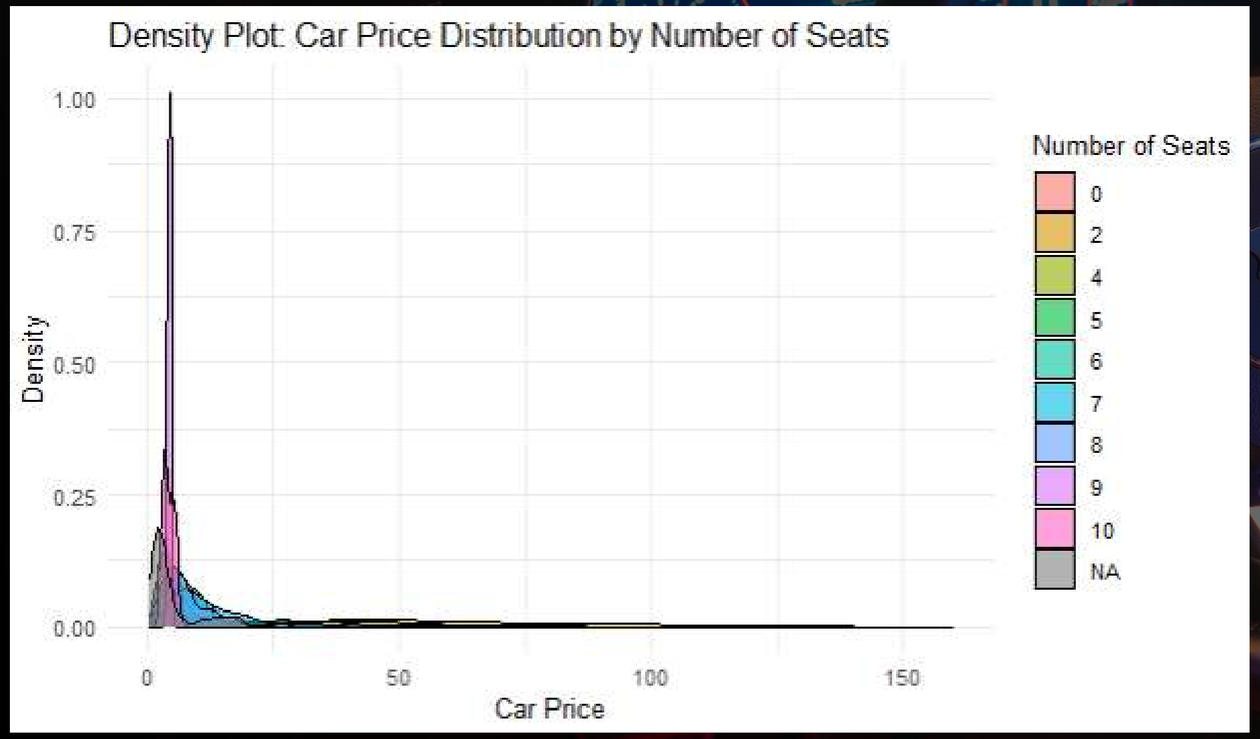


Group data by Seats and calculate the average price

```
seats_avg_price <-cars_data %>%
  group_by(Seats) %>%
  summarize(avg_price = mean(Price, na.rm =
TRUE)) %>%
  arrange(desc(avg_price))
View(seats_avg_price)
```

	Seats	avg_price
1	2	55.211875
2	4	20.752525
3	0	18.000000
4	7	14.837463
5	6	9.511290
6	5	8.478791
7	8	7.458881
8	NA	6.162619
9	9	4.450000
10	10	4.280000

Density plot:
car price distribution by no. of seats




```
ggplot(cars_data, aes(x = Price, fill =
factor(Seats))) +
 geom_density(alpha = 0.6) +
 labs(
  title = "Density Plot: Car Price Distribution by
Number of Seats",
  x = "Car Price",
  y = "Density",
  fill = "Number of Seats"
 ) +
 theme_minimal()
```

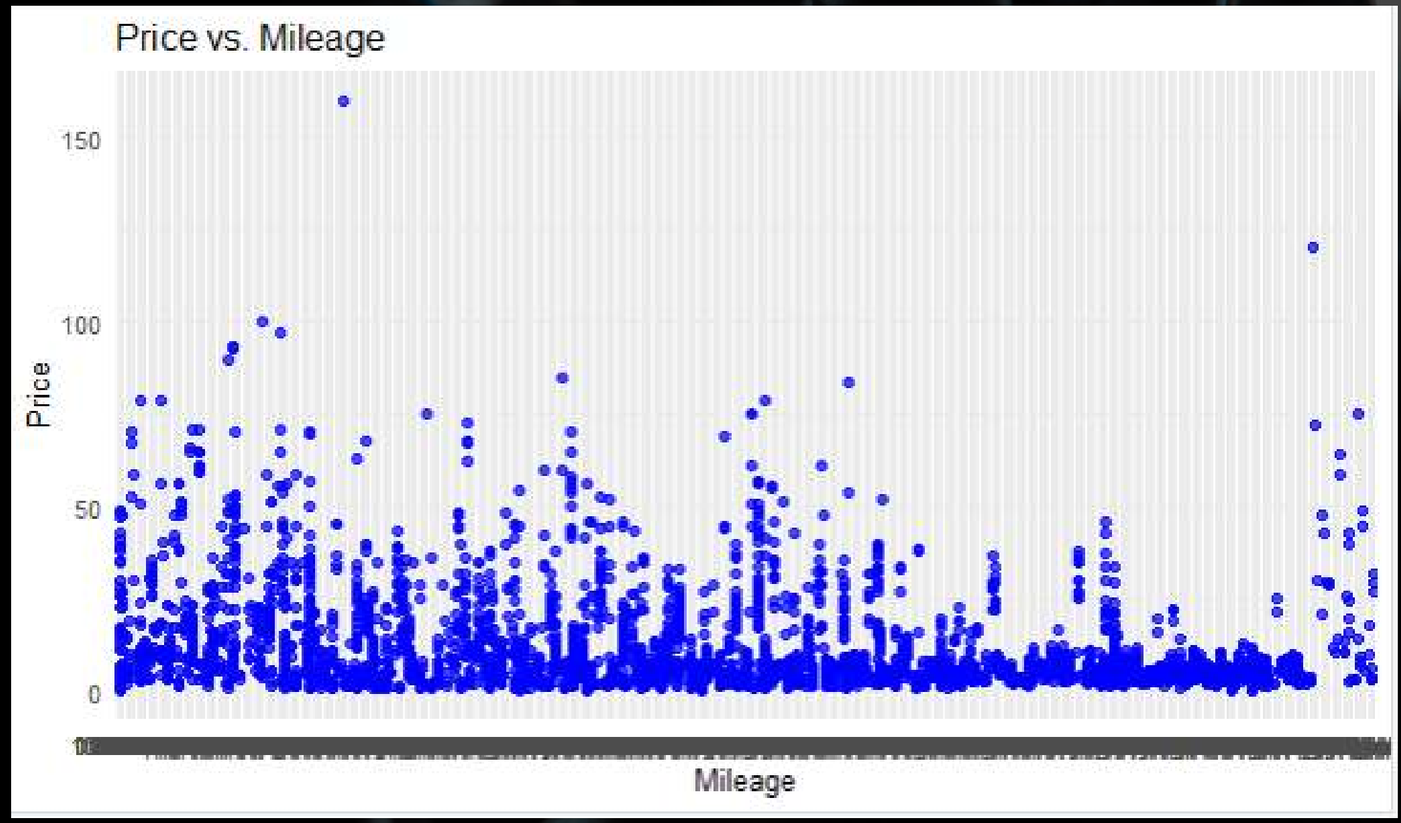
Price vs Mileage

```
mileage_analysis <- cars_data %>%
  mutate(
    mileage_category = case_when(
      Mileage < 15 ~ "Low Mileage",
      Mileage >= 15 & Mileage < 25 ~
"Medium Mileage",
      Mileage >= 25 ~ "High Mileage"
    )
  ) %>%
  group_by(mileage_category) %>%
  summarize(
    avg_price = mean(Price, na.rm =
TRUE),
    count = n()
  )
View(mileage_analysis)
```

	mileage_category	avg_price	count
1	High Mileage	7.154021	568
2	Low Mileage	15.758162	1583
3	Medium Mileage	7.786037	5102

Scatter plot:
Price vs. Mileage

```
ggplot(cars_data,
aes(x = Mileage, y =
Price)) +
 geom_point(color =
"blue",alpha=0.7) +
 labs(title="Price vs.
Mileage", x="Mileage",
y="Price")+
 theme_minimal()
```



Most Popular Car Models

	Name	n
1	Mahindra XUV500 W8 2WD	55
2	Maruti Swift VDI	49
3	Maruti Swift Dzire VDI	42
4	Honda City 1.5 S MT	39
5	Maruti Swift VDI BSIV	37
6	Maruti Ritz VDi	35
7	Toyota Fortuner 3.0 Diesel	35
8	Honda Amaze S i-Dtech	32
9	Honda Brio S MT	32
10	Honda City 1.5 V MT	32
11	Hyundai Grand i10 Sportz	32

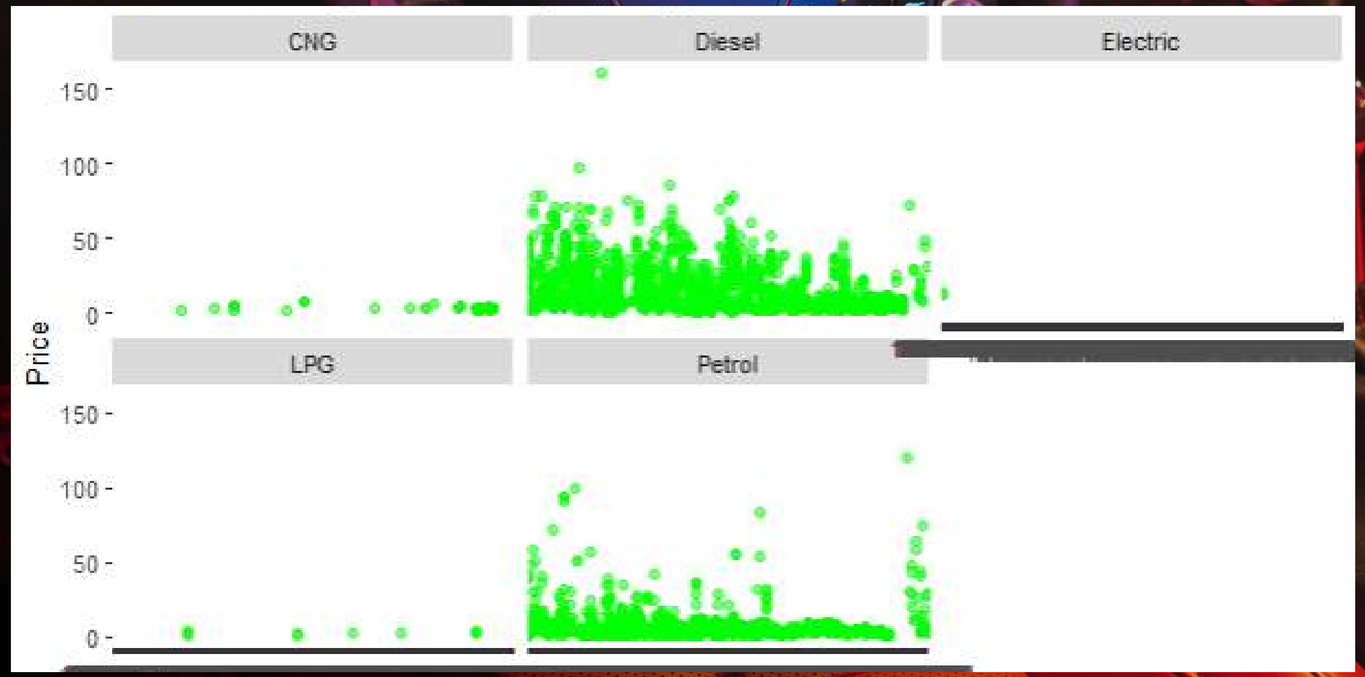
```
top_models <- cars_data %>%
count(Name, sort = TRUE) %>%
top_n(10)
View(top_models)
```

Key Findings

Certain models dominate in terms of listings and popularity.

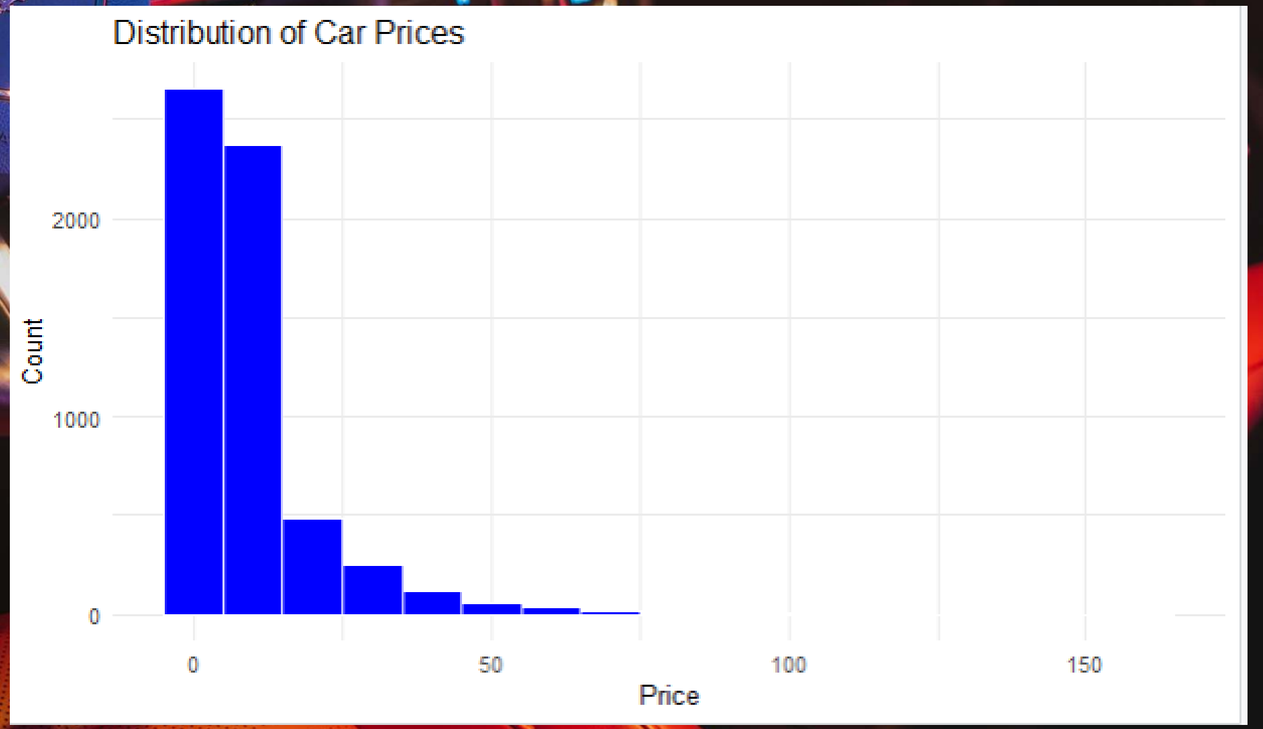
Faceted Plot: Price vs. Mileage by Fuel Type

```
ggplot(cars_data, aes(x = Mileage, y = Price)) +
  geom_point(alpha = 0.5, color = "green") +
  facet_wrap(~ Fuel_Type) +
  labs(title = "Price vs. Mileage Faceted by Fuel Type", x = "Mileage", y = "Price")
```



Histogram: car prices

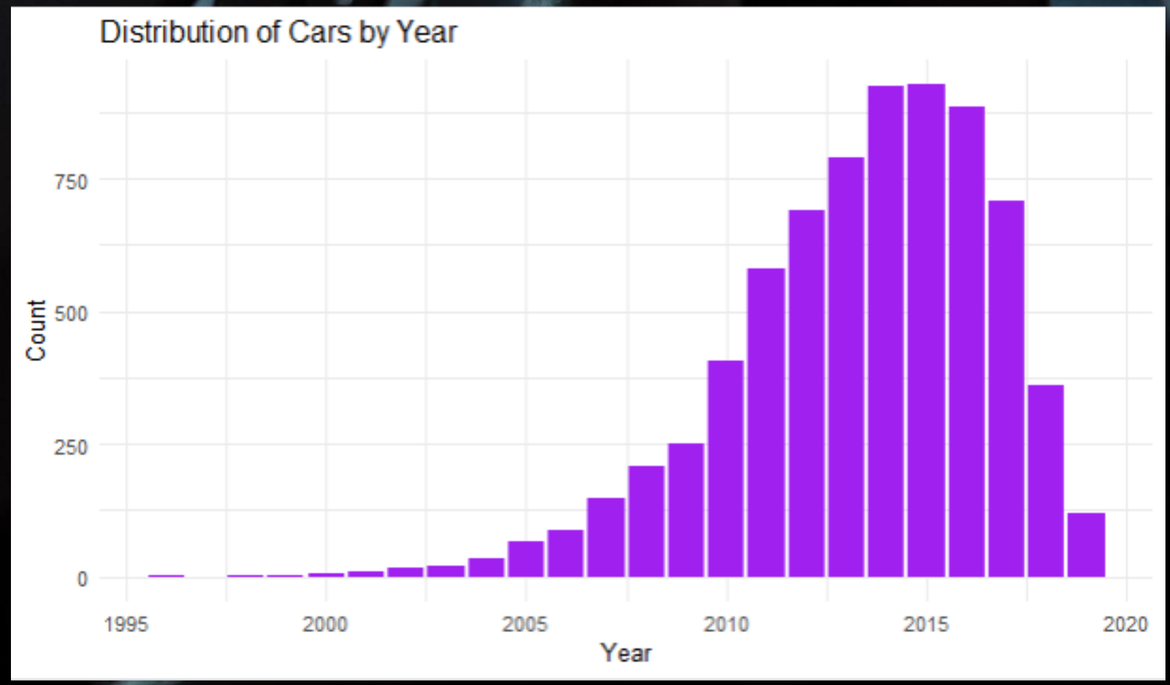
```
ggplot(cars_data, aes(x = Price)) +
geom_histogram(binwidth = 10, fill = "blue", color = "white") +
theme_minimal() +
labs(title = "Distribution of Car Prices", x = "Price", y = "Count")
```



Car prices exhibit a skewed distribution, with most cars concentrated in lower price ranges.

Distribution of cars by year of manufacture

```
ggplot(cars_data, aes(x = Year)) +
  geom_bar(fill = "purple") +
  theme_minimal() +
labs(title = "Distribution of Cars by Year", x = "Year", y = "Count")
```



Trend in average price by year

```
avg_price_by_year <- cars_data %>%
group_by(Year) %>%
summarise(Average_Price = mean(Price, na.rm = TRUE))
View(avg_price_by_year)
```

The average prices fluctuate over time, showing historical and market-related patterns.

	Year	Average_Price
1	1996	NaN
2	1998	1.432500
3	1999	0.835000
4	2000	1.175000
5	2001	1.543750
6	2002	1.294000
7	2003	2.440000
8	2004	1.941290
9	2005	2.026842
10	2006	3.355897
11	2007	3.204000
12	2008	3.917759
13	2009	5.177727

Count the number of listings by year

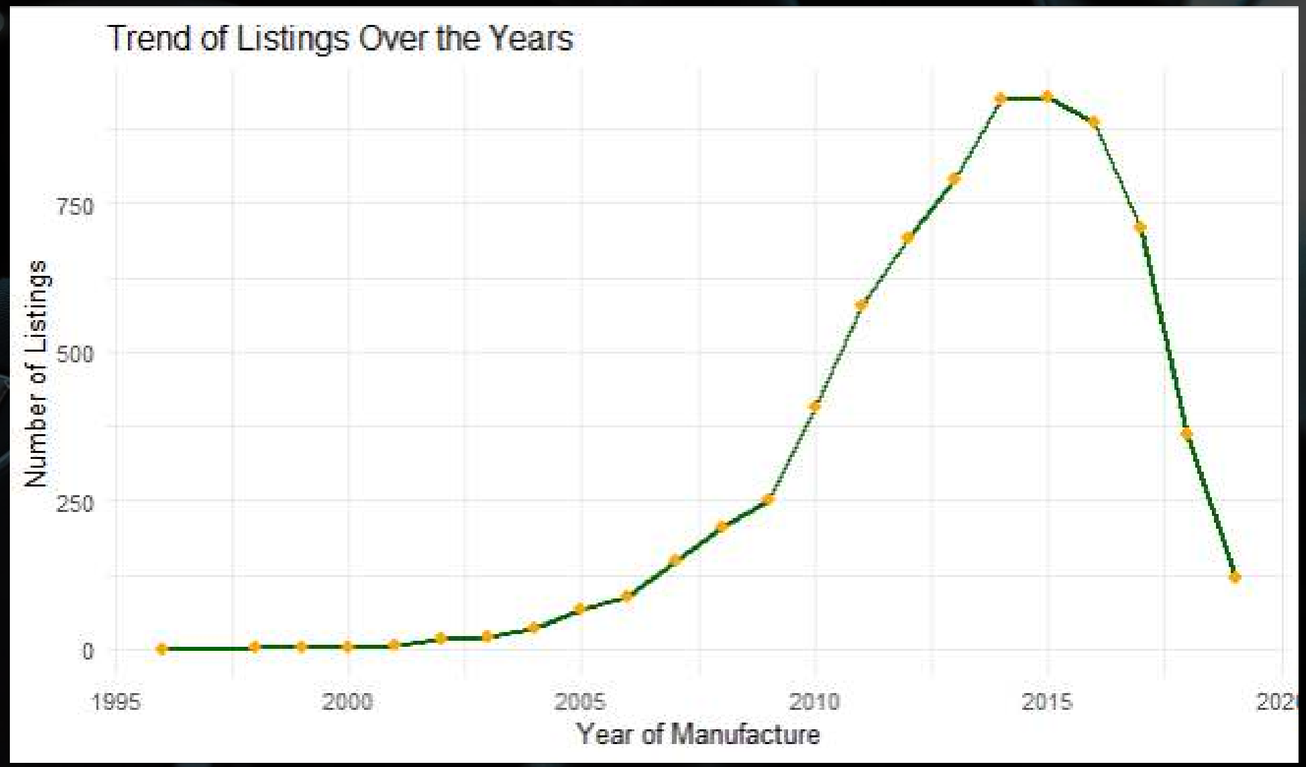
	Year	Count
1	1996	1
2	1998	4
3	1999	2
4	2000	5
5	2001	8
6	2002	18
7	2003	20
8	2004	35
9	2005	68
10	2006	89
11	2007	148
12	2008	207
13	2009	252

```
listings_by_year <- cars_data %>%
  group_by(Year) %>%
  summarise(Count = n(), .groups = "drop")
View(listings_by_year)
```

The number of listings fluctuate over time, showing historical and market-related patterns.

Line chart for the trend in listings

```
ggplot(listings_by_year, aes(x = Year, y = Count)) +
  geom_line(color = "darkgreen", size = 1) +
  geom_point(color = "orange", size = 2) +
  theme_minimal() +
  labs(
  title = "Trend of Listings Over the Years",
    x = "Year of Manufacture",
    y = "Number of Listings"
  )
```



The line chart shows fluctuations in car listings over the years, with peaks indicating high market activity and declines sales.

Find the most common fuel type by year

```
common_fuel_by_year <- cars_data %>%
  group_by(Year, Fuel_Type) %>%
  summarise(Count = n(), .groups = "drop") %>%
  arrange(Year, desc(Count)) %>%
  filter(row_number() == 1)
  View(common_fuel_by_year)
```

	Year	Fuel_Type	Count
1	1996	Diesel	1

The most common fuel type changes with manufacturing year, reflecting technological and market shifts.

KEY FINDINGS

Average Price by Location:
Significant variation in average car prices exists across locations.
Visualization highlights regional trends.

Fuel Type Distribution:
Some fuel types are more prevalent in the dataset.
Diesel and petrol cars are typically dominant.

Transmission Impact:
Transmission type influences car prices.
Automatic cars generally have higher average prices.

KEY FINDINGS

Owner Type Influence:
First-owner cars fetch higher average prices compared to others.
Pie chart illustrates the percentage contribution to total prices by owner type.

Seating Capacity:
Cars with more seats often have higher average prices, but trends may vary.
Density plots reveal price distribution differences based on seating capacity.

Mileage Analysis:
Higher mileage tends to correspond to lower prices, although exceptions exist.
Mileage categories (low, medium, high) correlate with different price brackets.

CONCLUSION

This analysis provides insights into the used car market, highlighting the importance of factors like location, fuel type, transmission, and mileage in pricing and trends. Stakeholders, such as buyers and sellers, can use this data to make informed decisions, tailoring their strategies to market realities.


THANK YOU