Homeworkcheck2

Oliviero Rizzola

Used cars in India price analysis

1a. Where is this data from?

The given dataset contains scraped data from **Cars24**, an online marketplace for buying and selling used cars. More specifically, the dataset used for this analysis is a subset of the original one.

The Five W's and How:

- **Who**: The dataset is composed of individual used cars listed on the CARS24 platform.
- **What**: The dataset includes several variables about the cars as listed in question 1b.
- **When**: The dataset includes listings from cars manufactured in past years, starting from 2011 up until 2023. The dataset is updated annually.
- **Where**: The subset of data under analysis contains cars listed in India.
- **Why**: I assume the data was collected for market analysis.
- **How**: The data was collected from the CARS24 website and shared on Kaggle.

Potential issues with the "Who" and "What":

- Personally I believe that the biggest potential issue is that the dataset may be affected by selection bias. In other words, it only includes cars listed on one platform. Morevoer among the variables listed for example "brand reputation" isn't included, which could offer reasoning to the certain prices.

1b. What are the variable types?

Table 1: Variable types and units

Variable	Type	Units
Model.Name	Categorical	-
Price	Quantitative	Indian Rupees (INR)
Manufacturing_year	Quantitative	Year
Engine.capacity	Quantitative	Cubic centimeters (cm^3)
Spare.key	Categorical	Yes / No
Transmission	Categorical	Manual / Automatic
KM.driven	Quantitative	Kilometers
Ownership	Quantitative	Number of owners
Fuel.type	Categorical	Petrol / Diesel / CNG
Imperfections	Quantitative	Number of imperfections
Repainted.Parts	Quantitative	Number of parts repainted

2a. Investigating Age vs. KM.driven

```r
# Using the mutate() verb as described in the DataCamp lab, make a new variable called Age th
#| echo: true
cars24.mutate <- cars24.data %>%
mutate (Age = 2024 - Manufacturing_year)
```

Think

We are examining the relationship between two **quantitative variables** such as **age**, representing the number of years since the car was manufactured, and **KM.driven**, representing the total distance the car has been driven, in kilometers.

I expect older cars to have been driven more, since they've had more time to accumulate mileage. However, I expect **this not being a perfect relationship**: for example some newer cars might have high mileage if used daily and other older cars might have lower mileage if used as second vehicles.

Since both variables are quantitative, a **scatterplot** is the most appropriate way to display the relationship. I will add a **regression line** to help visualize the trend. I will also compute the **correlation coefficient** (after checking the conditions) to measure the strength and direction of the relationship.

Show

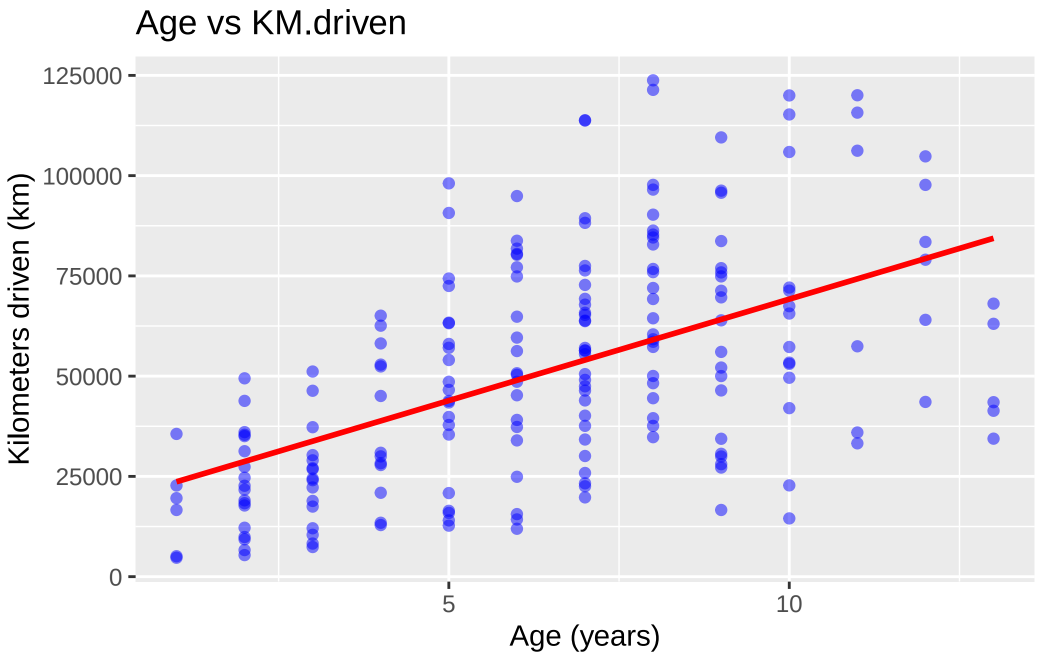

Figure 1: Scatterplot: Age vs. KM.driven

Correlation Coefficient

```
cor(cars24.mutate$Age, cars24.mutate$KM.driven)
```

```
[1] 0.5271851
```

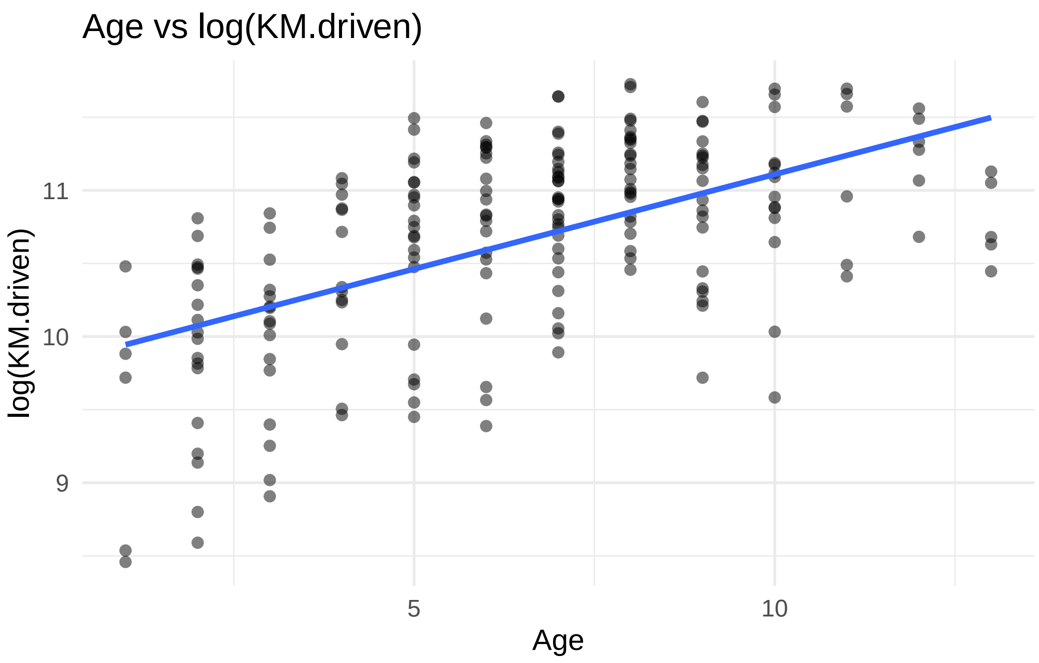

Figure 2: Scatterplot: Age vs. Log_KM.driven

New Correlation Coefficient

```
cor(cars24.mutate.log_KMdriven$Age, cars24.mutate.log_KMdriven$log_KM)
```

```
[1] 0.5589143
```

Tell

The analysis shows a **moderate positive association** between a car's **age** and how many **kilometers it has been driven**. This **aligns with expectations** that older cars generally have been driven more. There are a few visible outliers but they do not appear to dominate the pattern and therefore since the other conditions are matched I calculated the correlation coefficient.

The scatterplot shows **variation at each age level**, which makes sense because not all cars are used equally as anticipated in the expectations.

I decided to apply the **log transformation** to KM.driven in order to make the data more linear and to make it easier to see a consistent pattern.

2b. Investigating Age vs. Imperfections

Think

I will now examine the relationship between a car's **Age** (in years) and the number of **Imperfections** it has. Both variables are **quantitative**.
I expect that older cars will have more imperfections, since they've had more time to accumulate imperfections.A **scatterplot** is appropriate, since both variables are quantitative. Adding a **regression line** helps visualize any trend. If possible I would also compute the **correlation coefficient**.

Show

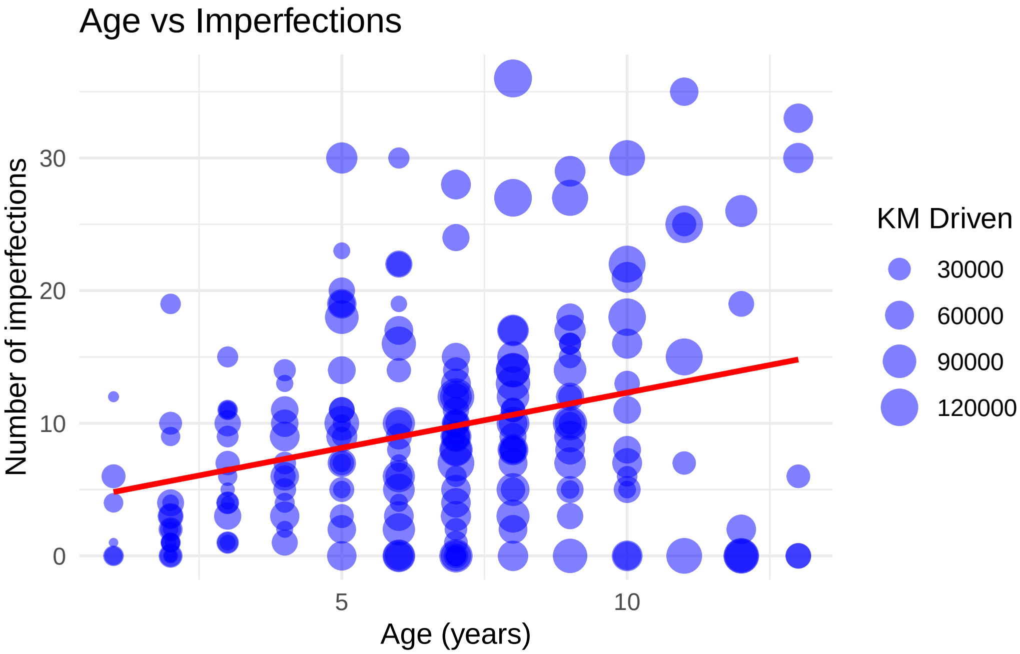

Figure 3: Scatterplot: Age vs. Imperfections

Correlation Coefficient

```
cor(cars24.mutate$Age, cars24.mutate$Imperfections)
```

```
[1] 0.3072834
```

Tell

The analysis shows a **weak positive relationship** between car age and number of imperfections. This means that, on average, older cars tend to have more imperfections, which matches my expectations, but the relationship is not strong.

I decided to calculate the correlation coefficient which provides a helpful summary of the linear association, but may be affected by a few **outliers** as cars with an unusually high number of imperfections. These extreme cases can distort the correlation. I decided to **highlight the kilometers driven to try to explain** this outliers. This result suggests that **age alone does not fully explain a car's condition**. **Other factors**, such as how carefully a car was used, climate exposure, or maintenance play a large role.

2c. Thinking about your results

In question 2a I explored the correlation between **Age and KM.driven which came out to be positive and moderate.** On the other hand, in question 2b I analyzed the relationship between **Age and Imperfections which was weaker, although still positive**, with the scatterplot showing more variability.

These results suggest that the use of cars over time (measured by KM.driven) is fairly predictable, the older the car the more kilometers it has been driven, indicating **steady use over time**. Moreover, we understood that the **number of imperfections is not strongly related to the age of the cars in the Indian market**, suggesting that **other factors** may be more relevant such as maintenance habits, how carefully a car was used the climate exposure and many others.

3a. Investigating Price vs. Age

Think

My expectation is that **as a cars get older**, its **price will decrease** which is a common pattern in the used car market. Both variables are **quantitative**, **age** (years since manufacture) and **price** (in rupees). I believe a **scatterplot** with a **regression line** is the best way to explore this relationship, along with calculating the **correlation coefficient**.

Show

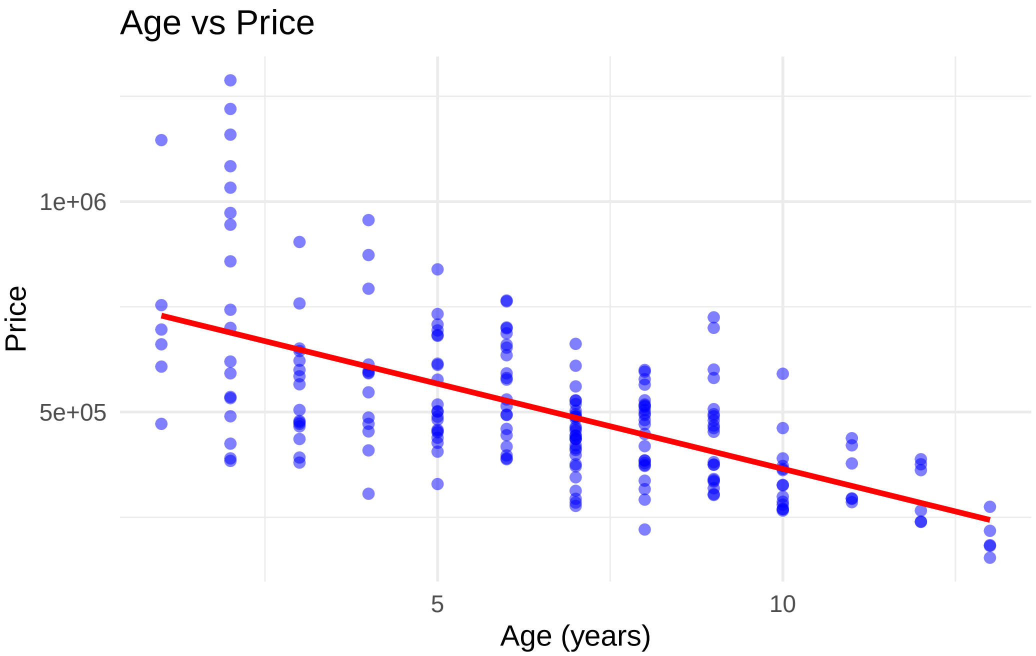

Figure 4: Scatterplot: Age vs. Price

Correlation Coefficient

```r
cor(cars24.mutate$Age, cars24.mutate$Price)
```

```
[1] -0.623393
```

Tell

This analysis confirms that **car price generally decreases with age** in the Indian used car market. The moderately strong negative correlation (r -0.62) means that age is an important factor in determining value, but not the only one.

There are still cars of the same age that sell for very different prices, likely due to other variables such as: kilometers driven, model, engine capacity, transmission type, condition.

In conclusion, **age is a strong predictor of price**, but it's not the full story. The spread in the data suggests that other features , especially usage and condition, play a role in how much a used car sells for.

3b. Checking model fit

Think again

In question 3a the analysis suggested that as a car's age increases, its price decreases. The correlation I found, more specifically, was negative and moderately strong. Now I am going to check whether the **linear model** (Price - Age) is a **good fit** using several tools. For instance I am going to analyze:

- R^2
- Se
- Residuals
- Comparison with the 'braindead' model

```
model_age_price <- lm(Price ~ Age, data = cars24.mutate)
summary(model_age_price)
```

```
Call:
lm(formula = Price ~ Age, data = cars24.mutate)

Residuals:
    Min      1Q  Median      3Q     Max 
-304699  -88798  -26637   74452  599301 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   769591      25712   29.93   <2e-16 ***
Age           -40446       3605  -11.22   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 150700 on 198 degrees of freedom
Multiple R-squared:  0.3886,	Adjusted R-squared:  0.3855 
F-statistic: 125.9 on 1 and 198 DF,  p-value: < 2.2e-16
```

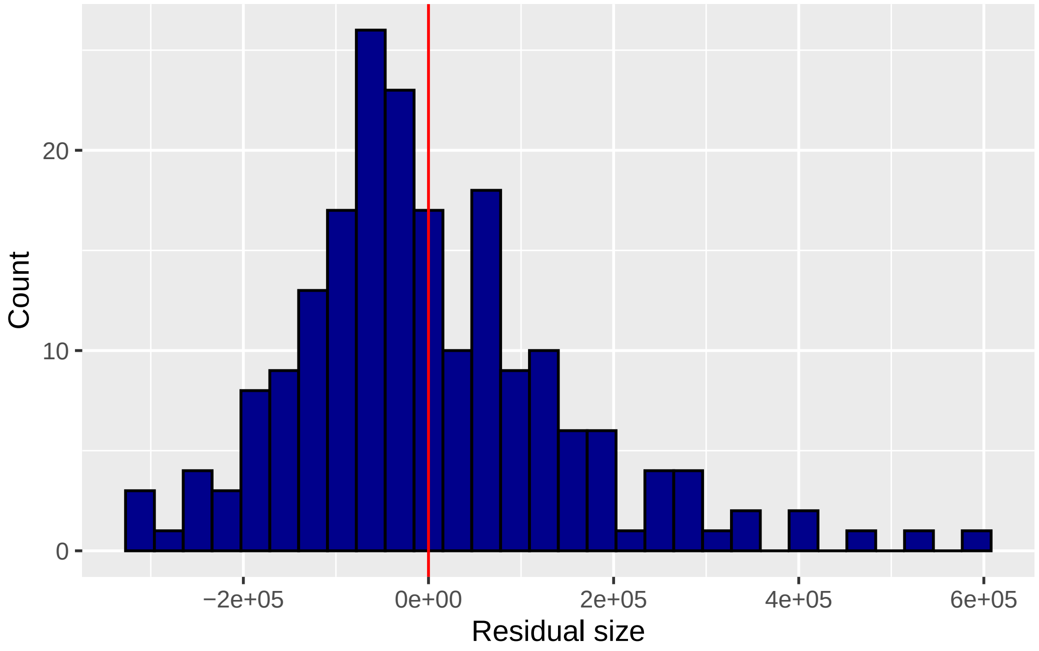

Figure 5: Histogram of residuals from the linear model Price ~ Age

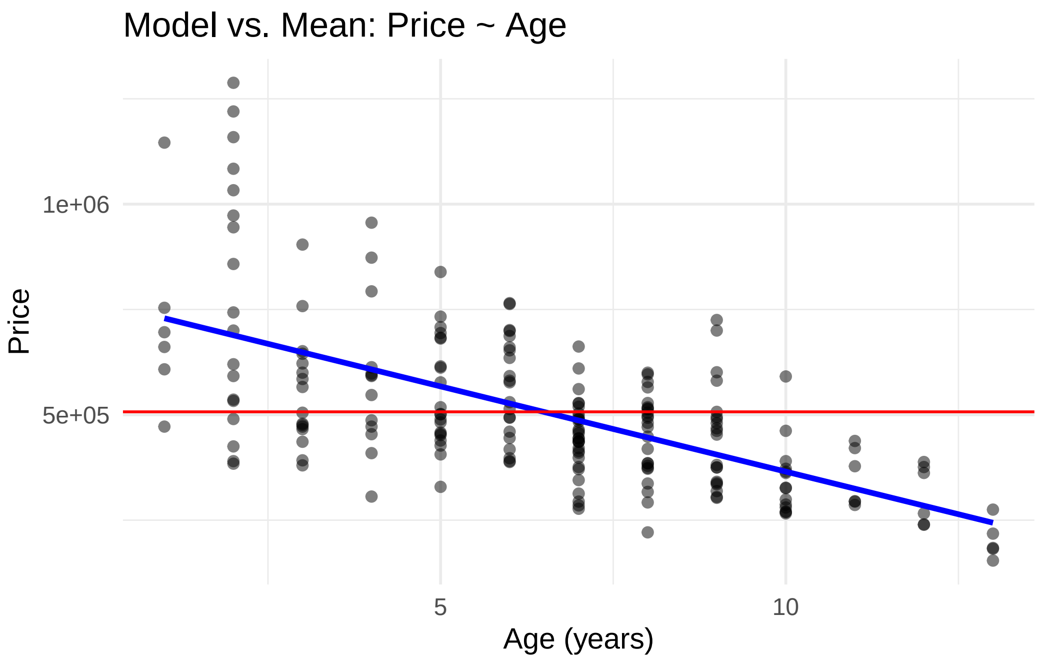

Figure 6: R²- comparison between two models

Revising conclusions

From the regression results we can appreciate an **R-squared** of 0.3855 that suggests that this is model is a moderate fit where age clearly matters, but other variables may also influence price. Moreover the **p-value** for the slope is < 2e-16, indicating a statistically significant relationship between the variables in the linear regression model. The **residual standard error** (Se), which gives an idea of how far off predictions typically are, is around 150000 which again suggests that while age is informative, it's not enough on its own. From the regression summury output one can read as follows:

- **Min:** –304,699
- **1st Quartile (Q1):** –88,798
- **Median:** –26,637
- **3rd Quartile (Q3):** 74,452
- **Max:** 599,301

This shows that the **residuals are centered near 0**, which is positive, but that there is some skew. The histogram confirms this being mostly symmetric, but **slightly right-skewed.**

Finally, the **comparison between the two models** shows that the predictions (blue line) from my model are much closer to the actual points than the red line, brainded model, is, meaning the chosen model is suitable.

3c. Investigating Price vs. KM.driven

Think

I expect that as the number of kilometers driven increases, the price of the car should decrease. More driving usually means more wear and tear, worse conditions of the vehicle due to usage, so buyers would likely pay less for high-mileage cars. Both variables, **KM.driven and Price are quantitative**. Since both are quantitative, a **scatterplot** is the right way to show the relationship, along with a **regression line** to highlight the trend and a **correlation coefficient** to quantify it.

Show

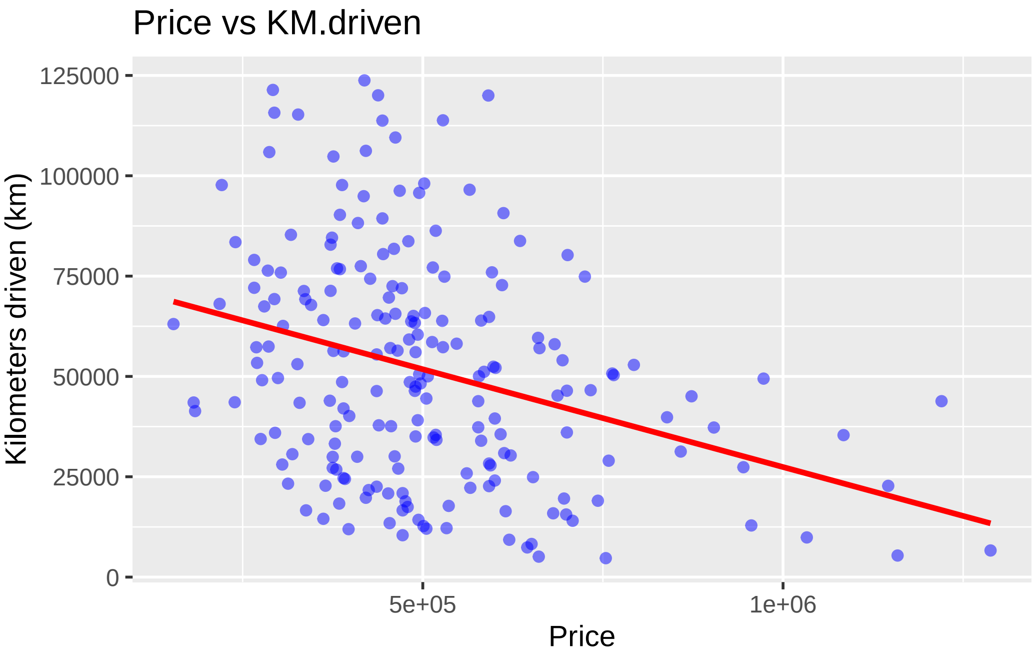

Figure 7: Scatterplot: Price vs. KM.driven

```
cor(cars24.mutate$Price, cars24.mutate$KM.driven)
```

```
[1] -0.329403
```

Tell

This analysis shows that there is a **negative relationship** between how much a car has been driven and its price. In other words, cars with higher kilometers driven tend to sell for less. The **correlation coefficient of -0.329** supports this, indicating a **moderately weak negative** association.

However, the relationship is not very strong, meaning that KM.driven alone does not fully explain price. **Other factors**, as also mentioned in previous questions, like age, car model, condition, or ownership history likely **also play important roles in determining price**.

In conclusion, while more use typically means lower value, the variability in price at similar KM levels suggests that a more complete model should include additional variables beyond just KM.driven.

3d. Thinking about your results

In Questions **3a–3c**, I explored how **Age** and **KM.driven** relate individually to **Price**. Both variables showed the expected trends, but with different strengths of association.

- **Age vs. Price** showed a **stronger negative relationship** (correlation -0.62, R^2 0.39) highlighting how generally older cars tend to sell for less, which is expected due to depreciation over time.
- **KM.driven vs. Price** showed a **weaker negative relationship** (correlation -0.33). Cars with more kilometers generally have lower prices, but there's more variation

As mentioned in the "Tell" section of the different questions, there are likely several other factors influencing price that weren't accounted for yet. This possible **lurking variables** might be car model, car conditions, ownership history, fuel type and transmission.

Therefore, Age and KM.driven are both important in determining car prices in India, but **neither fully determines it alone**. Age has a stronger impact, but to truly understand price, I believe one **might have to include more variables to capture the full picture**.