Name: Đỗ Thị Hiền Nga IRN: 1932300357 GitHub Portfolio Link: https://github.com/DoHana2203		**Comment**
Criteria	**Score**	
Data Cleaning and Preparation		
Data Analysis and Insight		
GitHub Prortfolio Quality		
Total		

1. Data Overview

1.1 Dataset Description

The dataset, called "Used Car Price" - contains information about used cars available for sale, featuring various automobile brands such as Hyundai, KIA, Nissan, Ford, Mercedes, Audi, and more. In this dataset, I use analytical tools, such as Excel to conduct my analysis. The data was most likely collected from the market and has been cleaned and analyzed for academic purposes.

1.2 Dataset Structure

This Raw Data sheet contains 202 rows and 7 columns, including key attributes such as Model, Year, Km, Color, Type, Fuel, and Price.

The dataset consists of 7 columns describing attributes of each car:

- Model: The make and model of the car (e.g., Hyundai Accent, Mercedes GLC).
- Year: The manufacturing year of the vehicle.
- Km/h: The total kilometers driven by the car, indicating usage and wear.
- Color: The exterior color of the car.
- Type: The transmission type, either Automatic or Manual.
- Fuel: The type of fuel used by the vehicle (in this dataset, all cars use Gasoline).
- Price: The listed sale price of the car, expressed in thousand.

2. Data Cleaning

2.1 Missing Data

Identify: The dataset contains some missing values, especially noticeable in the Color with 3 missing column:

  - The row for Hyundai Accent RB, Hyundai Accent and KIA Cerato is missing 3 values in the color column and 1 value in price column.

Model	Year	Km/h	Color	Type	Fuel	Price
Hyundai Accent RB	2011	220.000		Automatic	Gasoline	530.000
Hyundai Accent	2009	280.000		Manual	Gasoline	
KIA Cerato	2015	140.000		Automatic	Gasoline	765.000

- The row for Hyundai Elantra HD, KIA K3, Mitsubishi Xpander is missing 3 value in the Fuel column.

Model	Year	Km/h	Color	Type	Fuel	Price
Hyundai Elantra HD	2009	100.000	Bronze	Automatic		475.000
KIA K3	2017	130.000	Black	Automatic		790.000
Mitsubishi Xpander	2022	200.000	Silver	Automatic		900.000

- The row for Hyundai Accent RB, KIA Picanto, **KIA K3, Mercedes E 200 AMG** is missing 4 values in the Type column.

Model	Year	Km/h	Color	Type	Fuel	Price
Hyundai Accent RB	2012	135.000	Silver		Gasoline	525.000
KIA Picanto	2009	114.000	Silver		Gasoline	415.000
KIA K3	2016	80.000	Petroleum		Gasoline	885.000
Mercedes E 200 AMG	2021	25.000	Black		Gasoline	4.200.000

- That’s 11 missing values in total.

2.2 Decide how to handle them

Handling Missing Data:

- In the Cleaned Data version, the number of rows remains unchanged at 202. While minor adjustments were made, no rows were removed during the cleaning process. Instead, any missing values were addressed using statistical imputation methods: categorical fields such as Color, Type, and Fuel were filled with their most frequently occurring values (mode), while the missing numeric value in the Price column was replaced with the dataset’s median price. This approach ensures data completeness without compromising the dataset’s integrity or volume. All column data types were preserved, with numeric fields stored as integers or floats and categorical fields as text.

Like:

- For the Fuel column, since most cars use Gasoline in this dataset, it is reasonable to:

- Use mode imputation (fill missing values with the most frequent value) by replacing missing Fuel values with "Gasoline".

2.2 Duplicates

Identify duplicates:

- The dataset contains duplicate records, for example:

➔The entry for Hyundai Accent ; 2009 ; 280,000 Km, Manual, Gasoline appears at least twice.

- Duplicate records can distort analysis, such as counts and average prices.

Decide how to handle duplicates:

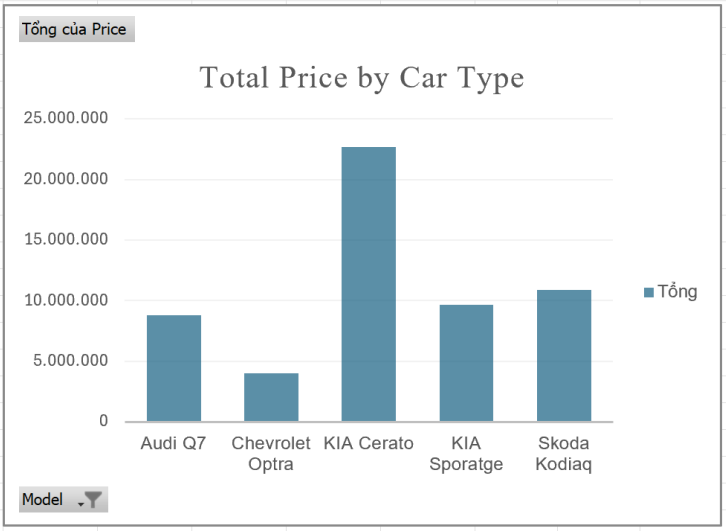
- It is important to detect and and was modified based on the most frequently occurring value to ensure data accuracy.

- Keep only one instance of each duplicate record.

3. Key Insights

Key Insight 1: The distribution of Kia Cerato in tha dataset compared with other brands, beased on Price

Nhãn Hàng	Tổng của Price
Audi Q7	8.800.000
Chevrolet Optra	4.005.000
KIA Cerato	22.665.000
KIA Sporatge	9.630.000
Skoda Kodiaq	10.850.000
Tổng Cuối	**55.950.000**



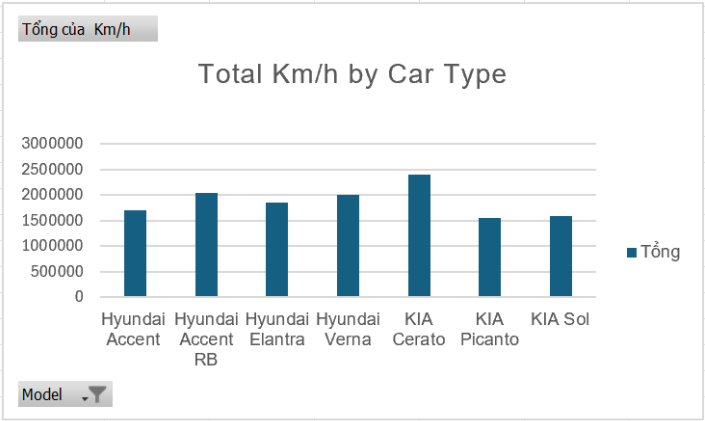
To explore the total car price distribution in the dataset of 202 entries, five representative models were selected: Audi Q7, Chevrolet Optra, KIA Cerato, KIA Sportage, and Skoda Kodiaq.

The KIA Cerato stands out with the highest total price of 22,665,000. In contrast, Chevrolet Optra recorded the lowest total price, only 4,005,000. Both KIA Sportage and Skoda Kodiaq show moderate total values of 9,630,000 and 10,850,000, respectively.

The Audi Q7, a luxury vehicle, has a total price of 8,800,000, which is surprisingly lower than some mid-range models. This comparison suggests that model popularity and market availability may influence total price contribution more than segment or brand positioning alone.

Key Insight 2: The distribution top cars have high numbers in the dataset compared with other brands, beased on Km/h

Nhãn Hàng	Tổng của Km/h
Hyundai Accent	1713000
Hyundai Accent RB	2039100
Hyundai Elantra	1864451
Hyundai Verna	2000034
KIA Cerato	2403500
KIA Picanto	1563000
KIA Sol	1584000
Tổng Cuối	**13167085**



To analyze the total kilometers driven by different car types, data from seven models were reviewed: Hyundai Accent, Hyundai Accent RB, Hyundai Elantra, Hyundai Verna, KIA Cerato, KIA Picanto, and KIA Sol.

Among them, the KIA Cerato leads with the highest total of 2,403,500 km. Hyundai Accent RB and Hyundai Verna follow closely with totals of 2,039,100 km and 2,000,034 km, respectively. Hyundai Elantra and Hyundai Accent recorded moderate values of 1,864,451 km and 1,713,000 km. Meanwhile, KIA Picanto and KIA Sol show the lowest figures, at 1,563,000 km and 1,584,000 km.

This comparison highlights the strong presence of KIA Cerato in terms of usage, while suggesting that other models may have more limited mileage due to factors like usage purpose or market reach.