CARS RENTAL COMPANY

INTRODUCTION

Car trading is a profitable and exciting business, due to the importance of the car as a means of transportation. The diversity of car brands has made this field competitive and in rapid development of cars. Each brand is distinguished from the other by the prices and technologies in the car, and this makes a group of people attracted to changing the car continuously, by selling the used car and buying another.
This type of trade can be used to open a car rental exhibition by purchasing used cars and then renting them.

• Problem understanding:

Through a fake scenario. One of the companies seeks to open an exhibition of used cars with strong income and wants to know several things, including: the most popular brands, knowledge of prices, and several other things, for example: the number of miles the car has traveled. Therefore, we will ask a number of questions:
1- What is the brand with the highest price and in which cities it is?
2- Does the model year affect the price of the car?
3- What states have clean cars, and what are the most popular brands in each state?
4- What are the most common car colors in each state?

• Motivation

One of the most important motives that we seek to achieve is to analyze the data by answering some questions:

1. What is the brand with the highest price and in which state it is?
2. Does the model year affect the price of the car?
3. What states have clean cars, and what are the most popular brands in each state?
4. What are the most common car colors in each state?

• Data Description

With 13 columns, 12 features for each car, and 2,500 rows, a dataset has three types of data, as shown in Table 1, which includes lists the features, data type, and description for each column. The dataset has a size of 284 KB that called US Cars' data.

Table 1: Describe columns by clarifying features and data type.

Price	Integer	The sale price of the vehicle in the ad
Years	Integer	The vehicle registration year
Brand	String	The brand of car
Model	String	model of the vehicle
Color	String	Color of the vehicle
State/City	String	The location in which the car is being available for purchase
Mileage	Float	miles traveled by vehicle
Vin	String	The vehicle identification number is a collection of 17 characters (digits and capital letters)
Title Status	String	This feature included binary classification, which are clean title vehicles and salvage insurance
Lot	Integer	A lot of numbers are identification the number assigned for a particular quantity or a lot of materials from a single manufacturer.For cars, a lot

		numbers are combined with a serial number to form the Vehicle Identification Number.
Condition	String	Time

Tools

Terminology: jupter Notebook.

Libraries: Numpy , Matplotlib.pyplot, Pandas, Seaborn, Stats, Math, Random, Plotly.express.

METHODOLOGY

1. Data Per_processing

Load data & validation

Several libraries have been imported, the most important of which is matplotlib.pyplot, pandas, numpy, seaborn, stats, math, random, plotly.express.
The data set was read with a extension file .csv, and then the first and last rows were displayed. One of the most important features of this step is knowing the number of columns and rows, viewing data, and knowing the type of data for each column of the dataset. Also to displays summary statistic for each numerical column, and knowing the relationship between columns.

Data Cleaning

Data cleaning is one of the important stages that help make the data ready for analysis.

Duplicate or unnecessary data:

This step is to search for repeated rows in the dataset by filtering data down as appropriate. This is done through the use of pandas.duplicated() where the results are that there are no duplicate values and rows.

Inconsistent text and typos:

Each data column was checked by the minimum and maximum of numerical values and unique values of categoricals. The results showed that the data set does not contain inconsistent text and typos.

Missing values:

In this step, pandas.info() is used to check the missing values for each column of the dataset, and then use pandas.isna() which shows the boolean True/False for each element values. The results showed that the dataset does not contain missing values. But by using pandas.value_counts() , where the price column contained a value of 0 and these values in the price column are missing values. The number of zero values is 43 rows, all of which have been replaced by mean or median values by using pandas.fillna() to fill missing values with a mean value.

Outliers:

Through box plots to visualize numeric data for check from outlires value . The results showed that the data set does not contain outliers, at least in the columns that were used in the analysis.

Managing Columns of Data:

In this step we did first: Combine state & country columns into a single column called state. And combine brand & model columns into a single column called brand.
Second: Delete some unimportant columns such as: Unnamed: 0, vin, country, model, lot by using pandas.drop() .

2. Exploratory Data Analysis (EDA)

Price column: prices were categorized into three classes: Low, Medium, and High. By using binning the values of max, mean, and min were calculated and then stored in a new column called price_binncd. Then the average of these values (max, mean, and min) was represented and displayed in the count plot by creating price_cat function(). The results showed that the average prices are medium, higher than the average of high prices and the average of low prices. As the average price ranges between (26-19090).
Brand column: by using pandas.unique(), the results showed that the brand column contains 180 brands of cars available for rent. Results for the top 10 and last 10 of the most popular car brands are shown by countplot().
Year column: the dataset contains a variety of model years from 2020 to 1973. We note that in 2019 it contains the largest number of cars estimated at 892.
It came in second place 2018 with an estimated number of 395 cars, then 2017 with an estimated number of 377 cars. The results were presented through countplot().
State column: using countplot() displays the contents of the state column which shows the number of each state which is estimated at 44 states. Whereas Pennsylvania ranks first in terms of the number of cars, and Montana is the lowest state with the number of cars.
Color column: Through plot() the colors of the cars were shown, where the white color was higher than other colors, and the number of cars with white color was estimated at 707.

3. Correlation Between Features:

By using pandas.corr() which shows the relationships between the numerical column in the dataset. We note that there is a strong correlation relationship between the columns. And through a heat map() that plots the correlation relationship between the numerical column . Where the top of the side bar of dark color represents the strength of the correlation relationship between the columns, and the correlation relationship at the bottom of the bar decreases when the color is light. For this we note that price & year is a strong correlation. And price & mileage is a weak correlation.

4. Questions:

This stage is one of the stages that describes data analysis by answering the questions that were listed in the Problem understanding.
1- Since white is the most common color in cars, we want to know what cities white colors are in, and how many white cars are in each brand?
Through the code line (df.loc[df.color == 'white', ['state', 'brand']].value_counts()), we were able to answer the question.

2- What are the most common car colors in each state?
By using groupby() to specify state and color , in this step, will show the colors of the cars in each state. And through px.scatter() the results will be displayed.
We noticed that the number of white color in cars was the highest, but in the low price rate.
3- For this reason we want to know the effect of the price by color?
Through groupby() and barplot() , we find that kona blue metallic ranks first, then ruby red, then royal crimson metallic tinted clearcoat.
4- What is the brand with the highest price and in which cities it is?
Two methods were used to answer this question. The first method is to display all prices, from high prices to low prices.
The second method is to specify high prices only through the use of filtering for rows. By using px.scatter() which displays only six states with high price cars, it contains 9 different brands, and the high price ranges from 56700 to 84900. The mercedes-benz-sl-class in Florida has the highest price.
5- Does the model year affect the price of the car?
Through regplot() where the y axis represents prices, and the x axis represents years. We note, yes, the year of production of the car affects the price, as the year of production was more than 7 years, the price will decrease.
6- What states have clean cars, and what are the most popular brands in each state?
By using px.scatter(), which details the title status for each brand whether it is a clean vehicle or salvage insurance, and also indicates in which state.

CONCLUSIONS

Through this project, which aimed to data analysis through several stages: problem understanding, identification of the tools used, description of the dataset, where the methodology is summarized in several stages: data per-processing, which includes (loading data & validation, data cleaning), exploratory data analysis , correlation between features, questions.
We noticed that both the brand and year of production of the car affect the price, also when the car is a clean vehicle or salvage insurance, has an effect on the price. We also noticed that the color of the car is special when it is white, the car is not expensive.