Project Milestone 1

Gunadheep Sakthivel

Introduction:

Data mining and preparing the dataset for analytics was a critical part of our project initial milestone. We collected our dataset from cars.exportauction.com, a site which attains reputation by selling clean and serviced cars and taken it from Kaggle https://www.kaggle.com/datasets/doaaalsenani/usa-cers-dataset/data. Varies different car characteristics are estimated by the dataset such as price, lot number, brand, model, color, mileage, VIN, and title status and the vehicle condition. Every horizon of those 32 columns of characteristics is a different type of vehicle.

Data Cleaning and Preprocessing

Data Loading

We loaded the dataset into a Pandas DataFrame using the **read_csv** function. We inspected the first few rows of the dataset to get a glimpse of its structure.

```
car_data=pd.read_csv("USA_cars_datasets.csv")
car_data.head()
```

	Unnamed: 0	price	brand	model	year	title_status	mileage	color	vin	lot	state	country	condition
0	0	6300	toyota	cruiser	2008	clean vehicle	274117.0	black	jtezu11f88k007763	159348797	new jersey	usa	10 days left
1	1	2899	ford	se	2011	clean vehicle	190552.0	silver	2fmdk3gc4bbb02217	166951262	tennessee	usa	6 days left
2	2	5350	dodge	mpv	2018	clean vehicle	39590.0	silver	3c4pdcgg5jt346413	167655728	georgia	usa	2 days left
3	3	25000	ford	door	2014	clean vehicle	64146.0	blue	1ftfw1et4efc23745	167753855	virginia	usa	22 hours left
4	4	27700	chevrolet	1500	2018	clean vehicle	6654.0	red	3gcpcrec2jg473991	167763266	florida	usa	22 hours left

Data Cleaning

Removing Unnecessary Columns

We removed the 'Unnamed: 0' column as it seemed to be an index column and was not required for our analysis.

```
car_data = car_data.drop(columns = ['Unnamed: 0'])
```

```
car_data.head()
```

	price	brand	model	year	title_status	mileage	color	vin	lot	state	country	condition
0	6300	toyota	cruiser	2008	clean vehicle	274117.0	black	jtezu11f88k007763	159348797	new jersey	usa	10 days left
1	2899	ford	se	2011	clean vehicle	190552.0	silver	2fmdk3gc4bbb02217	166951262	tennessee	usa	6 days left
2	5350	dodge	mpv	2018	clean vehicle	39590.0	silver	3c4pdcgg5jt346413	167655728	georgia	usa	2 days left
3	25000	ford	door	2014	clean vehicle	64146.0	blue	1ftfw1et4efc23745	167753855	virginia	usa	22 hours left
4	27700	chevrolet	1500	2018	clean vehicle	6654.0	red	3gcpcrec2jg473991	167763266	florida	usa	22 hours left

Handling Missing Values

We checked for missing values in the dataset and found that there were no missing values.

```
In [5]: #missing data
        car_data.isnull().sum().sort_values(ascending=False)
```

```
Out[5]: price          0
        brand          0
        model          0
        year           0
        title_status   0
        mileage        0
        color          0
        vin            0
        lot            0
        state          0
        country        0
        condition      0
        dtype: int64
```

Handling Zero Values in Price

We replaced zero values in the 'price' column with the median price of the dataset to ensure data integrity.

```
In [6]: median_price = car_data['price'].median()
        car_data['price'] = car_data['price'].astype(int)
        car_data['price'].replace(0,median_price ,inplace=True)
```

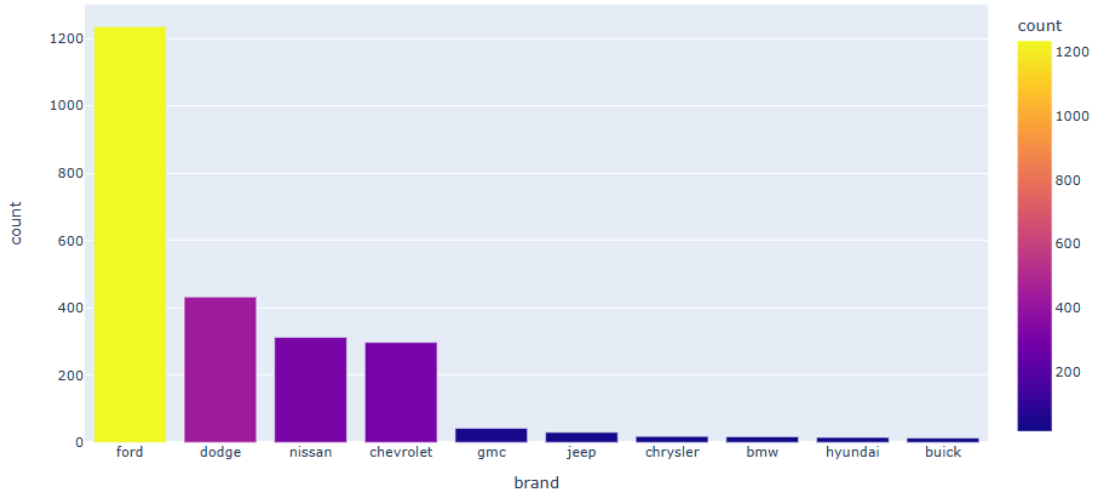
Initial Analysis

Brand-wise Distribution of Cars

We grouped the data by brand and counted the number of models for each brand. We then visualized the top 10 brands with the highest number of models using a bar chart.

```
In [7]: brand_of_car = car_data.groupby('brand')['model'].count().reset_index().sort_values('model',ascending = False).head(10)
        brand_of_car = brand_of_car.rename(columns = {'model':'count'})
        fig = px.bar(brand_of_car, x='brand', y='count', color='count')
        fig.show()

        #You can reach a lot of information about car brand and their count
```



Self-Evaluation:

Areas of Success

Data Loading: read_csv could be used to load this dataset into the Pandas DataFrame and it worked perfectly.

Data cleaning: The median price was used as a stand-in for NA values in the "price" column and any of the excess columns were omitted.

First Data Analysis: A bar chart was used in the first step in the analysis in order to gauge the number of cars recorded against the various brands in the data set.

Areas for Improvement

Data exploration: Although at a glance we have reviewed the initial rows of the data, a strategic EDA would have unveiled other details concerning distribution, features and possible extreme values in the statistics set.

Data Visualization: At the same time, we used the bar chart to show brand-wise distribution of automobiles. Other data visualization techniques such as pie charts, line charts, or histograms will, in turn, provide a spooky-dooker story and complete picture of the same.

Beyond Expectations

Managing Missing Values: Ease of preparation procedure was hardened due to the existence of no missing elements in the data set.

Data Loading and Cleaning: The task was carried out with no difficulties stumbling along smoothly and there were no errors found.

Unmet Expectations

Extensive Data research: The original research limitations were solved by brand-wise distribution visualization single. A more detailed study which investigated the correlations between the variables and looked deeper into the qualities not covered in the dataset could have given deeper insight.

Data Quality Assessment: While we did the work with detective and treatment of the zero values in the 'price' column, we did not complete a comprehensive evaluation of the data that comprised looking for duplicates, outliers, and discrepancies in other columns as well.

Conclusion:

In this milestone release, we achieved the goal of gathering the dataset, the preprocessing and cleaning part of the data as well as the analysis to determine the brand-wise distribution of the cars in our dataset. The subsequent stage will be the creation of some interactive visualisation tools that will be used to deepen the understanding of the data set.