Student Name: Pranav Karnth Mannur **StudentID: s3828461**

Data Preparation:

Step 1:

- Reading all the individual CSV files using read_csv and storing them in a variable.
- Joining all the required columns to a single Dataframe using the DataFrame() method.

Step 2:

- Making sure the Model and Manufacturer in the Dataframe is from the specified list provided in the Assignment1_Template.
- Using loc and isin() method to check for the same.

Step 3:

- In order to make sure Price, Transmission, Power, Engine_CC is in the specified range according to the Assignment1_Template.

Step 4:

- Checking for duplicate values if any in the Dataframe using duplicated and dropping them using drop_duplicates().
- Checking for NaNs values if any in the Dataframe using isna().sum

This is a car buyers dataset and we do not require Duplicated records as it would result in our distribution and give us wrong statistics.

The NaNs values too are not required as they do not provide us clear data.

Error1:

Checking for unique values in Fuel column:

```
'petrol', 'diesel', 'peatrol', 'automatic', 'diasel'
```

Correcting the Spelling mistakes of ‘peatrol’ to ‘petrol’ and ‘diesel’ to ‘diesel’ using .loc[].

Error2:

1. Commas can be spotted in the Values of Male, Female, Gender_Unclassified, Total columns. To correct that, using replace method to remove the commas from the values.

	Manufacturer	Model	Price	Transmission	Power	Engine_CC	Fuel	Male	Female	Gender_Unclassified	Total
3	Renault	Clio	22.100000	5.615385	75.576923	1219.653846	petrol	241287	312556	28,004	581847
4	BMW	320i	47.848370	6.444444	126.111111	1995.777778	petrol	408016	115843	29,125	552984
5	Volkswagen	Polo	18.192500	5.074074	60.962963	1408.055556	petrol	216333	299110	31,701	547144
6	Peugeot	206	20.033750	4.833333	71.333333	1631.500000	petrol	178698	250614	26,135	455447
7	Ford	Mondeo	39.973750	1.750000	130.250000	1998.500000	petrol	357452	69,603	16,550	443605
...	...	...	...	...	...	...	...	...	...	...	...
6097	Land-Rover	Defender	108.747195	7.853659	207.609756	2304.975610	diesel	1,012	150	80	1,242
6098	Toyota	RAV4	43.548516	1.354839	137.774193	2261.193548	petrol	670	482	66	1,218
6099	Alfa-Romeo	Spider	55.200000	6.000000	163.500000	2696.500000	petrol	790	247	81	1,118
6100	Honda	Shuttle	30.081000	4.000000	110.000000	2254.000000	petrol	639	416	49	1,104
6101	Mitsubishi	Space	23.165158	3.947368	82.157895	1817.315789	petrol	721	251	40	1,012

5949 rows × 11 columns

2. Checking datatypes of all columns present in the dataframe and it can be found that the Male, Female, Gender_Unclassified, Total columns are all of the object datatype.

   These have to be changed to int, because it is the total number of owners and it cannot be in any other datatype other than INT, using astype().

Error3:

1. Making sure the total number of owners is equal to the sum of Male, Female and Gender_Unclassified owners.
   df['Total']=df['Male']+df['Female']+df['Gender_Unclassified']

Error4:

For Manufacturer = 'Mitsubishi' it can be spotted that the Price, transmission, Power, Engine_CC, Fuel is exactly the same for the same Model name. Only the number of owners – Male, Female, Gender_unclassified and Total numbers are different.

This can be projected as an error during the Data entry process.

The way this can be solved is to group all the columns and sum the total number of owners in – male, female, Gender_Unclassified and Total.

```
[86]: #It can be seen that the Manufacturer, Model, Price, Transmission, Power, Engine_CC, Fuel is the same but the
      df.loc[df.Manufacturer=='Mitsubishi']
```

Out[86]:

	Manufacturer	Model	Price	Transmission	Power	Engine_CC	Fuel	Male	Female	Gender_Unclassified	Total
105	Mitsubishi	Colt	17.038767	5.302326	73.488372	1374.046512	petrol	20915	25109	2781	48805
180	Mitsubishi	Lancer	23.016615	4.730769	90.884615	1709.538462	petrol	12773	3703	607	17083
204	Mitsubishi	Outlander	37.446667	2.611111	120.333333	2085.055556	petrol	8910	3970	620	13500
213	Mitsubishi	Space	23.165158	3.947368	82.157895	1817.315789	petrol	7314	4690	716	12720
214	Mitsubishi	Carisma	22.690000	4.800000	82.200000	1773.700000	petrol	9003	3084	630	12717
...	...	...	...	...	...	...	...	...	...	...	...
5989	Mitsubishi	Colt	17.038767	5.302326	73.488372	1374.046512	petrol	10362	8619	707	19688
5998	Mitsubishi	Galant	25.082667	4.533333	93.600000	2127.333333	petrol	14098	2477	488	17063
6016	Mitsubishi	Space	23.165158	3.947368	82.157895	1817.315789	petrol	6860	2921	732	10513
6017	Mitsubishi	Lancer	23.016615	4.730769	90.884615	1709.538462	petrol	7048	2641	303	9992
6101	Mitsubishi	Space	23.165158	3.947368	82.157895	1817.315789	petrol	721	251	40	1012

Data Exploration:

Task 2.1:

A bar graph is used when we have to plot a categorical variable. A categorical variable represents types of data which are most likely to be divided into groups. Example: Sex (Male, Female) , Race, Age group.

Therefore, plotting a bar graph to analyse the composition of the total number of vehicle owners by gender for the top ten vehicles with the most owners.

Task 2.2:

Task 2.2.1 - Price Column:

For the Price column, for the initial dataset (before cleaning), it contains NaNs values which do not provide us data. The data collector makes an entry error when collecting data.

	Manufacturer	Model	Price	Transmission	Power	Engine_CC	Fuel	Male	Female	Gender_Unclassified	Total
23	Ford	Focus	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
32	Seat	Ibiza	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
54	Toyota	Aygo	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
65	Ford	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
75	NaN	106	NaN	4.647059	NaN	1366.294118	NaN	41,148	NaN	NaN	NaN

After cleaning the dataset according to the requirements in the template, there are no NaNs in the Price column.

As shown in the Boxplot there are outliers present in the dataset before and after cleaning. These are not be removed from the dataset as it is not acceptable to remove values just because it is an outlier. So there is no error in this perspective.

Task 2.2.1 – Power Column:

For the Power column, for the initial dataset (before cleaning), it contains NaN values which do not provide us data. The data collector makes an entry error when collecting data.

	Manufacturer	Model	Price	Transmission	Power	Engine_CC	Fuel	Male	Female	Gender_Unclassified	Total
23	Ford	Focus	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
32	Seat	Ibiza	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
54	Toyota	Aygo	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
65	Ford	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
75	NaN	106	NaN	4.647059	NaN	1366.294118	NaN	41,148	NaN	NaN	NaN

After cleaning the dataset according to the requirements in the template, there are no NaNs in the Power column.

As shown in the Boxplot there are outliers present in the dataset before and after cleaning. These are not to be removed from the dataset as it is not acceptable to remove values just because it is an outlier. So there is no error in this perspective.

Task 2.3:

Task 2.3.1:

A bar graph is used when we have to plot a categorical variable. A categorical variable represents types of data which are most likely to be divided into groups. Example: Sex (Male, Female) , Fuel type (Petrol, diesel), Age group.

Therefore, plotting a bar graph to analyse the composition of the total number of vehicle owners by Male and Female for the top ten vehicle Manufacturers with the most owners.

Task 2.3.2:

A pie chart can be used on a categorical variable, each slice can be used to represent a specific category. A categorical variable represents types of data which are most likely to be divided into groups. Example: Sex (Male, Female) , Fuel type (Petrol, diesel), Age group.

Therefore, plotting a pie chart to project the percentage of total number of vehicle owners by Fuel type for the top 100 owners.

Task 2.3.3:

A bar graph is used when we have to plot a categorical variable. A categorical variable represents types of data which are most likely to be divided into groups. Example: Sex (Male, Female) , Fuel type (Petrol, diesel), Age group.

Therefore, plotting a bar graph to analyse the composition of the total number of vehicle owners by Price of car models for the top ten most expensive cars models.

References:

1. GeeksforGeeks. 2022. *Matplotlib Tutorial - GeeksforGeeks*. [online] Available at: < https://www.geeksforgeeks.org/matplotlib-tutorial/ > [Accessed 10 March 2022].
2. GeeksforGeeks. 2022. *Pandas Tutorial - GeeksforGeeks*. [online] Available at: < https://www.geeksforgeeks.org/pandas-tutorial/ > [Accessed 10 March 2022].
3. GeeksforGeeks. 2022. *Matplotlib Tutorial - GeeksforGeeks*. [online] Available at: < https://www.geeksforgeeks.org/matplotlib-tutorial/ > [Accessed 5 April 2022].