Student Name: Joshua Cayetano Student ID: s3722151

# Data Preparation

With Task 1 we needed to combine 11 separate CSV files together into 1 file called “cleaned\_car\_buyers.CSV”. I would then need to later get rid of any potential issues such as white spaces or commas which will be explained later.

* I first started downloading all the associated files and putting them into 1 directory so I could access them later in Jupyter Notebook.
* I would import ‘pandas’ and matplot to read between different data structures and plot graphs from my data later.
* I read all the CSV files and later combined them into one file (“Mainfile”) by creating a new column for each variable and calling upon them in different CSV files. “*Task – Read Files”*
* We do not need to worry about bad lines, as there is no extra value in the CSV value.

*“Task – bad lines”*

Identification of issues

I looked at the contents of each CSV file. I determined potential issues were white spaces, commas, missing values (NaN), impossible values and outliers.

**NOTE:** *I used the sort-function in Excel from Microsoft (2022) as well as functions in Python to identify these values.*

*For each File:*

* Manufacturer: This is a nominal data type. There are 52 car manufacturers. All are valid.
* Model: This is a nominal data type. There is only one white space which we need to remove. ***I’m going to assume that all models are valid for the purpose of the assignment. It would be intensive, time consuming and outside the scope of the course to create another file containing all model types then use it to compare missing values. The reason we would compare our data to the models that are allowed is because there could be a filled in observation which does not match the specifications (E.g. Grease Lightning).***
* Price: This is a discrete data type. There are 20 values that are above the price range. We have 2 values less then 0. We also have 5 blank observations which we should remove or change.
* Transmission: This is a continuous data type. There are 129 observations that are <=0 which we must remove. We can assume this is because they were just bought buy haven’t been used yet. We also have 2 values over 10 which we need to remove. We also have 4 white space values which we must remove or change.
* Power: This is a continuous data type. We have 4 values which are greater than 500 which we need to remove. We have 3 observations which are less than 0 which we need to remove. We also again see 5 observations which are blank and we need to remove.
* Engine CC: This is a continuous data type. There are 17 values which are equal to 0 which violates our conditions which we need to remove.
* Fuel: This is a nominal data type. There are 5 values which are empty which we need to remove.
* Male: This is a discrete data type. There are 4 whitespaces which we need to replace with 0. We also must remove the commas.
* Female: This is a discrete data type. There are 4 whitespaces which we need to replace with 0. We also must remove the commas.
* Unknown: This is a discrete data type. There are 5 whitespaces which we need to replace with 0. We also must remove the commas.
* Total: The data type is discrete. There are 4 whitespaces which we need to remove or replace with 0. We also must remove the commas.

Python ways to check values:

**“**Task - Look at blank excel value.” ***+*** “Task - Python Check”

Removal of issues

## Error 1: Redundant Index

I had originally used the “Unnamed: 0” column to link the tables together. However, it became redundant as there was an automatic index created. So, what I did to solve this is delete it using the drop function. By removing this we also remove one of the major sources of error in Machine Learning – “Leakage”.

Table

Description automatically generated

*“Task – Drop Index”*

## Error 2: Mistakes during data entry

In this case it was being the use of commas in our integer values (Male, Female, Unknown, Total) and us having the need to round the decimals for, Price, Transmission, Power, and Engine. We need to remove commas to sum our values later and do calculations. We round our decimals to make it easier to read the data and standardise the variables.

I aimed to do this by updating each column and using the round and string replace command.

*“Task – Rounding and comma removal”*

## Error 3: Redundant White Space

We need to remove white space as while it is easy to remove it is difficult to spot and can cause errors. I aimed to solve this by using the strip function which would remove any white space for any string characters. This can also be used on discrete values. It can be noted that around 50 observations were removed based on the shape outputted.

“*Task – White Space”*

*“Task – Python Check”*

## Error 4: Impossible values/ outliers

Outliers can affect our data results in the data modelling process later. We also have to remove those that cannot be possible as for instance a car could not be priced over $650. I used drop function and used as references, examples from StackOverflow (2013) as well as examples from Data Science Parichay (2022).

*“Task – Impossible Values”*

*“Task - Checking drop function”*

*“Task – Outliers”*

## Error 5: Missing values

I wanted to remove values that were empty as it could affect my descriptive statistics and modelling in the next section by skewing the results. I first attempted to find empty values via “Task – Impossible values” but found there were no more empty areas. I also just in case aimed to fill in empty spaces with the fill.mean function.

*“Task – Impossible values”.*

*“Task – Attempting to fill in missing spaces”*

**After cleaning my data:**

Finally, I wrote the file to a new CSV file with the following code:

Mainfile.to\_csv(**'cleaned\_car\_buyers.csv**', index=False)

# Data Exploration

## Task 2.1

The data columns we are using are Male, Female, Unknown, Total and Model.

* Model: This is a nominal data type. We represent this best with a pie chart, bar chart or stacked column. Our level of measurement is Frequency and Proportions.
* Male, Female, Unknown, Total: These are all Discrete data types. We represent these with bar charts or histograms. This also looks at frequency and proportions for level of measurements.

While I identified that I need to have my data for Model into a Bar Chart, I realised I need to manipulate the data into the top 10 Models. This is because one manufacturer can have many car models, but a model can only belong to one manufacturer. I also realised I needed to leave the values of gender by itself instead of summing it, based on the large quantity of people involved and with the numbers already being provided in the total column.

“Task 2.1 – Model conversion needed”

I then obtained my data as shown through “Task 2.1 – Top 10 Cars by Total”.

NOTE: While the firsts 6 observations are correct in being sorted in descending order, 3 observations ranging at the 9000 total amount, seem to be outliers. This could be because I grouped the data by Model. I did not group by Manufacturer as Manufacturer could appear twice for 2 different car models while also compromising the descending order of Total.

I first mapped the overall data of Top 10 cars into a bar chart. ­**NOTE:** Python cut off the output.

Graphical user interface, application

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

From our Bar Chart data, we can determine that for most car models, majority of them are mostly used by Men (e.g. Leon, 6 and the Accord). People with unknown gender are primarily small stakeholders in the types of car models they get. We can ascertain that Females like the C1 model as Females outnumber the males who buy the car model.

For the bar charts, it would be meaningful to look at only the 1st and 10th observations based on the total of number of people who bought the car model. From the 1st observation, we can ascertain that the people who bought the Leon Model, 65.90% of which were 30.91% of Females and 3.69% of people of Unknown Gender.

Chart, pie chart

Description automatically generated

For the 10th observation, we can observe that of those that bought the Citigo Model, 53.73% were Females, 41.89% were Male and 4.38% were those of an Unknown gender.

Chart, pie chart

Description automatically generated

## Task 2.2

The data columns that we are using is Manufacturer, Model, Price and Power. The reason I added Model is to give context what the variables Price and Power belong to.

Since we are focusing on Price and Power which are both continuous data types, the level of measurement we use can be either be represented as a bar chart or a histogram. It is preferable to use a histogram in this case as we are dealing with numerical data (Price & Power), and we want to see the shape of the distribution.

From our histograms:

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

We can see for both histograms, most of the values are on the left side which means the data is positively skewed. For Power the average power for cars would range between 80 – 120. We also note that the observations are fairly spread out with a possible outlier past 400 Car Power. Meanwhile for Car Price, the most predominant price for Cars is Car Prices being approximately $20.

From the histograms we can identify that as an issue there are outliers which can affect our results which can be removed.

However, another issue which isn’t drawn from our histograms is that there are some mismatches of Prices and Power in relation to Cars. This can be determined from my appendix “Task 2.2 – Sorted Power”. For instance, an Aston-Martin has a higher price of 360.76 and generates 410 for a car. In comparison a Bentleigh Continental has a lower Price of 285.25 that generates more power at 428.20. While some of this could be explained due to the difference of new and older models – this is very ambiguous.

## Task 2.3

Here I have decided to explore the relationship between the number of Male Owners to Model, Transmission and Fuel. I would look at the top 5 results.

For the level of measurement:

* Male is a discrete data type, so I need to use either a histogram or a bar chart.
* Transmission is a continuous data type, so I need to use either a histogram or a bar chart.
* Fuel is a nominal data type, and I would normally represent this with a bar chart or pir chart. However, the results all have the fuel type so it would be redundant to visualise it.
* Model is a nominal data type so I would represent this as a pie chart or bar chart.

From the appendix “Task 2.3 – Males sorted”, we can determine that the number of males in the top 5 observations all use petrol. Furthermore, the highest transmission rate is 6.44 while the lowest transmission rate is 1.75. The highest number of males is 408,016.

From the appendix “Task 2.3 – Redundant values - Fuel”, it can be determined that the L Range has the highest transmission rate from using diesel. However, it must be noted that there are different number of Males for each observation.

# Appendix

## Task 1: Data preparation

“Task 1 - Python Check”

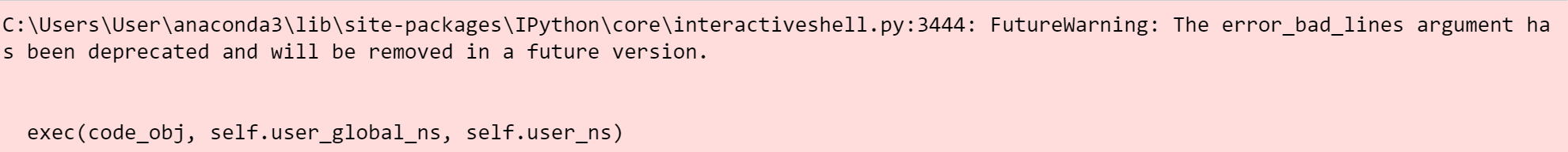
Graphical user interface, text, application, email

Description automatically generated

“Task 1 - Bad Lines”

**A picture containing text

Description automatically generated**



“Task 1 - Drop Index”

Table

Description automatically generated

“Task 1 - Rounding and comma removal”

Text

Description automatically generated

“Task 1 - White Space”

Text, letter

Description automatically generated

“Task 1 - Impossible values”

Text, table

Description automatically generated with medium confidence

““Task 1 - Outliers””

Text

Description automatically generated

“Task 1 - Fill in missing spaces”

Table

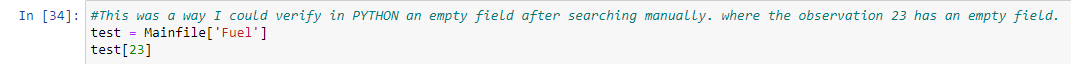
Description automatically generated with medium confidence

“Task 1 - Previous result of filling empty values”

Table

Description automatically generated

“Task 1 - Look at blank excel value.”



“Task 1 - Checking Drop function”

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

“Task 1 - Python Check”

Graphical user interface, text

Description automatically generated

## Task 2: Data Exploration

### 2.1

“Task 2.1 – Model conversion needed”

Text

Description automatically generated

“Task 2.1 – Top 10 Cars by Total”

Table

Description automatically generated

### 2.2

Task 2.2 – Different Prices and Models”

Table

Description automatically generated

“Task 2.2 – Same highest values from different businesses.

Table

Description automatically generated

“Task 2.2 – Sorted Power”

Graphical user interface, text, application

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface, text

Description automatically generated

“Task 2.2 – Sorted Price”

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

### 2.3

“Task 2.3 – Redundant values - Fuel”

Table

Description automatically generated

“Task 2.3 – Males sorted”

Table

Description automatically generated

“Task 2.3 – Transmission sorted”

Table

Description automatically generated

“Task 2.3 – Fuel for Males”

Table

Description automatically generated

“Task 2.3 – Model”

Table

Description automatically generated

# References

**All based on APA 7.**

ASQ. (2022). *WHAT IS A HISTOGRAM?.* ASQ.

<https://asq.org/quality-resources/histogram#:~:text=When%20to%20Use%20a%20Histogram,can%20meet%20the%20customer's%20requirements>

Data Science Parichay. (2022). *Pandas – Delete rows based on column values*. Data Science Parochay.

<https://datascienceparichay.com/article/pandas-delete-rows-based-on-column-values/>

# Bibliography

Adam Smith. (2022). *How to sum rows of a pandas DataFrame in Python.* Adam Smith.

<https://www.adamsmith.haus/python/answers/how-to-sum-rows-of-a-pandas-dataframe-in-python>

GeeksforGeeks. (2022). *Bar Plot in Matplotlib.* GeeksforGeeks. <https://www.geeksforgeeks.org/bar-plot-in-matplotlib/>

Lynn, S. (2022). *Use Pandas Groupby to Group and Summarise DataFrames.* Shanelynn.

<https://www.shanelynn.ie/summarising-aggregation-and-grouping-data-in-python-pandas/>

Nightingale, N. (2022). *How to remove comma from integer in python.* Grepper. <https://www.codegrepper.com/code-examples/python/how+to+remove+comma+from+integer+in+python>

Nic, Dr.(2022). *Types of Data: Nominal, Ordinal, Interval/Ratio - Statistics Help*. [Dr Nic's Maths and Stats](https://www.youtube.com/channel/UCG32MfGLit1pcqCRXyy9cAg). <https://www.youtube.com/watch?v=hZxnzfnt5v8#t=03m15s>

Microsoft. (2022). *Split text into different columns with the Convert Text to Columns Wizard.* Microsoft. <https://support.microsoft.com/en-us/office/split-text-into-different-columns-with-the-convert-text-to-columns-wizard-30b14928-5550-41f5-97ca-7a3e9c363ed7>

pandas. (2022). *Panda.DataFrame.dropna.* Pandas. <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.dropna.html>

Python Guides. (2022). *Python Pandas Drop Rows Example*. Python Guides.

<https://pythonguides.com/python-pandas-drop-rows-example/#:~:text=Python%20pandas%20drop%20rows%20by%20index,-In%20this%20section&text=To%20remove%20the%20rows%20by,the%20index%20number%20or%20name>.

Ren,Y. (2022, April 9). *Practical Data Science: Data Curation* [ Lecture Material ]*.* Canvas@RMIT University. <https://rmit.instructure.com/courses/90503/files/23226842?module_item_id=3941935>

Ren,Y. (2022, April 9). *Practical Data Science: Data Summarisation: Descriptive Statistics and Visualisation*.[Lecture Material]*.* Canvas@RMIT University.

RMIT University. (2022). *Easy Cite referencing tool -RMIT University.* RMIT University. <https://www.lib.rmit.edu.au/easy-cite/>

Stackoverflow. (2013). [*How to delete rows from a pandas DataFrame based on a conditional expression [duplicate]*](https://stackoverflow.com/questions/13851535/how-to-delete-rows-from-a-pandas-dataframe-based-on-a-conditional-expression)*,* Stackoverflow.

<https://stackoverflow.com/questions/13851535/how-to-delete-rows-from-a-pandas-dataframe-based-on-a-conditional-expression>

Statology. (2022). *How to Change the Position of a Legend in Matplotlib*. Statology.

<https://www.statology.org/matplotlib-legend-position/#:~:text=To%20change%20the%20position%20of%20a%20legend%20in%20Matplotlib%2C%20you,legend()%20function.&text=The%20default%20location%20is%20%E2%80%9Cbest,avoids%20covering%20any%20data%20points>.

TutorialsPoint. (2022). *Matplotlib – Bar Plot.* Tutorialspoint. <https://www.tutorialspoint.com/matplotlib/matplotlib_bar_plot.htm>

3schools. (2022). *Matplotlib Pie Charts.* 3schools. <https://www.w3schools.com/python/matplotlib_pie_charts.asp>