Group - Room 6

Compiled by:
smck0583 510678236
kong7456 500459007
jdor8613 500490723
mlim2731 500308222

Topic of Interest

The topic of interest for this analysis is investigating which lifestyle choices have the greatest impact on one’s body mass index (BMI). Data was collected from the National Health Survey from the Australian Bureau of Statistics and was surveyed from the year 2017 to 2018. The data itself was released on April 30, 2019. Recorded data concerns a number of individuals aged 0 to 85 years and over, with random backgrounds. We extracted the participants’ responses to various questions based on 4 lifestyle choices:

1. Physical exercise
2. Smoking use
3. Alcohol consumption
4. Dietary habits

These 4 lifestyle choices were chosen due to their rich data and diverse values. The topic was chosen not only for its relevance to the personal interests of the group members, but also to discover and engage with the data-rich world of healthcare through the lens of data analytics with automated aggregations and charts. Examination of data values in the field of health provides insights and expands the horizon of the vast applications of data analytics within other, more grounded fields of work apart from abstract fields such as finance or economics.

Stakeholder Discussion

The topic in investigation discusses the impacts different health habits have on BMI, and also attempts to reveal a direct link between these behaviours and an increase or decrease in BMI. The use of BMI in determining health risks and categorising people as it does not account for muscle mass, bone density, overall body composition, and racial and sex differences. However, improved understanding of factors that affect BMI could be generally useful to various parties:

- Dieticians, nutritionists and healthcare professionals
- Physical trainers and fitness industry players
- Government bodies who are involved with making health legislation
- Individuals who are interested in health and learning more how these behaviours affect their body

Physical Exercise (510678236)

The data used to analyse physical exercise is taken from a data set with comma separated value format (with two worksheets) entitled “health.xlsx”, it was originally obtained from the National Health Survey conducted by the Australian Bureau of Statistics 2017-18. The data set holds data relating to the health and weekly habits of many respondents and is a sample meant to represent a sample of the Australian population. Within the Data set there are 255 observed variables (here expressed as the columns) of each respondent with 21316 rows, other than the header row, the remaining rows are each respondent's answers/ observations to each of the variables. This original data set was then reduced in size to only include variables to do with BMI measurements and physical exercise habits of the respondents, making up a new data set with 35 columns (indicating the variables) and keeping the original 21316 rows (before data cleaning). The remaining data will be utilized to analyze the attributes of physical exercise against the BMIs of all the respondents, most important of which are listed below:

Variable Name	**Explanation**	**Encoding (if applicable)**
“SEX”	This is the gender of the respondent. This variable was downloaded with a binary encoding system to identify males with the integer 1 and females with the integer 2.	Male - 1 Female - 2
“EXTYPELW”	This variable is a nominal variable that represents what type of exercise the respondent performed within the last 7 days. This variable was downloaded with a label encoding system with 8 values from a range of integer values from (1-8).	No exercise - 1 Walking only - 2 Moderate exercise only - 3 Vigorous exercise only - 4 Walking and moderate - 5 Walking and vigorous - 6 Moderate and vigorous - 7 Walking, moderate and vigorous - 8
“EXNUDTH”	This is a numeric integer variable that represents the number of days in the last week where the respondent performed physical exercise.	
“BMISC”	This is a numeric float variable that measures the BMI of every respondent.	

Part 1: Data Cleaning

In order for the analysis of the data to be correct, the data should first be cleaned such that all missing or incorrect values are omitted. The first cleaning program will remove all of the incorrect data values placed within the variable of interest: column 23 (representing the number of times a respondent has exercised in the past week). Since, all the fields contain data values, any missing values were replaced with the value 97 or 98, therefore to remove the respondents from the data set with missing values, the following program may be used:

```
import csv

file = open("Health2.csv", "r")
new_file = open("cleaned_data.csv", "w")

first_line = True

counter = 0

for row in file:
   variables = row.strip("\n").split(",")
   if first_line:
       new_file.write(row)
       first_line = False
   else:
       exercise = variables[22]
       if exercise == "97" or exercise == "98":
           continue
       else:
           new_file.write(row)
```

Furthermore, since the analysis also looks at values of BMI, any missing values will also need to be removed from the column containing BMI data. Missing values for BMI were inputted into the CSV file as 0s. Therefore to eliminate faulty data the following program may be run:

```
import csv

file = open("cleaned_data.csv", "r")
new_file = open("cleaned_info.csv", "w")

first_line = True

counter = 0

for row in file:
   variables = row.strip("\n").split(",")
   if first_line:
       new_file.write(row)
       first_line = False
   else:
       bmi = variables[-5]
       if bmi == "0":
           continue
       else:
           new_file.write(row)
```

Finally, another variable that is observed is the amount of time exercised in the last 7 days for respondents (column 8 in the data set). Within this column, the missing values are represented by 99998, therefore, to remove the missing values within the data set, the following program may be run:

```
import csv

file = open("cleaned_info.csv", "r")
new_file = open("cleaned_info1.csv", "w")

first_line = True

counter = 0

for row in file:
   variables = row.strip("\n").split(",")
   if first_line:
       new_file.write(row)
       first_line = False
   else:
       exercise = variables[7]
       if exercise == "99998":
           continue
       else:
           new_file.write(row)
```

Part 2: Quantitative Binning Group Aggregate:

The first aggregate summary that may be derived is determining how many respondents within each BMI category perform physical activity every day of the week. The BMI classifications in accordance with The Centre for Disease Control and Prevention with a BMI under 18.5 being underweight, a BMI between 18.5 and 25 being considered healthy weight, a BMI between 25 and 30 being overweight and a BMI of over 30 to be considered obese. Given those conditions the aggregate summary may be produced with the following code:

```
file = open("cleaned_info1.csv", "r")

bmi_dict = {}
max_obese_freq = 0
max_over_freq = 0
max_healthy_freq = 0
max_under_freq = 0
max_overall_freq = 0

bmi_totals_list = [0,0,0,0]
total = 0

first_line = True
for row in file:
   if first_line:
       first_line = False
       continue
   variables = row.strip("\n").split(",")
   weekly_exercise = int(variables[22])
```

```
    gender = variables[0]
    bmi = float(variables[-5])
    if weekly_exercise > 4:
        max_overall_freq += 1
    if bmi > 30 and weekly_exercise > 4:
        max_obese_freq += 1
    elif bmi > 25 and weekly_exercise > 4:
        max_over_freq += 1
    elif bmi > 18.5 and weekly_exercise > 4:
        max_healthy_freq += 1
    elif bmi <= 18.5 and weekly_exercise > 4:
        max_under_freq += 1
    if bmi > 30:
        bmi_totals_list[0] += 1
    elif bmi > 25:
        bmi_totals_list[1] += 1
    elif bmi > 18.5:
        bmi_totals_list[2] += 1
    elif bmi <= 18.5:
        bmi_totals_list[3] += 1
    total += 1

bmi_dict["Obese"] = (max_obese_freq/bmi_totals_list[0])*100
bmi_dict["Overweight"] = (max_over_freq/bmi_totals_list[1])*100
bmi_dict["Healthy_weight"] = (max_healthy_freq/bmi_totals_list[2])*100
bmi_dict["Underweight"] = (max_under_freq/bmi_totals_list[3])*100
bmi_dict["any BMI"] = (max_overall_freq/ total)*100

for key in bmi_dict:
    print("The percentage of respondents that are {} and exercise at least 5 times a
week is: {}".format(key, round(bmi_dict[key])))
```

Table 1A - Percentage of All Respondents Who Performs Physical Activity 5 or More Days in the Last Week in Each BMI Classification:

BMI	Percentage of Respondents Who Perform Physical Activity at least 5 Days of the Last Week (%):
Underweight (BMI<18.5)	29
Healthy Weight (18.5<BMI<25.0)	34
Overweight (25.0<BMI<30.0)	32
Obese (30<BMI)	22
Any BMI	29

Part 3: Nominal Grouped Aggregate:

```
file = open("cleaned_info1.csv", "r")

obese_dict = {}
over_dict = {}
healthy_dict = {}
under_dict = {}

first_line = True
for row in file:
   if first_line:
       first_line = False
       continue
   variables = row.strip("\n").split(",")
   exercise_type = int(variables[-14])
   bmi = float(variables[-5])
   if bmi > 30:
       if exercise_type not in obese_dict and exercise_type > 4:
           obese_dict[exercise_type] = 1
       elif exercise_type in under_dict and exercise_type > 4:
           obese_dict[exercise_type] += 1
   elif bmi > 25:
       if exercise_type not in over_dict and exercise_type > 4:
           over_dict[exercise_type] = 1
       elif exercise_type in under_dict and exercise_type > 4:
           over_dict[exercise_type] += 1
   elif bmi > 18.5:
       if exercise_type not in healthy_dict and exercise_type > 4:
           healthy_dict[exercise_type] = 1
       elif exercise_type in under_dict and exercise_type > 4:
           healthy_dict[exercise_type] += 1
   elif bmi < 18:
       if exercise_type not in under_dict and exercise_type > 4:
           under_dict[exercise_type] = 1
       elif exercise_type in under_dict and exercise_type > 4:
           under_dict[exercise_type] += 1

total = sum(obese_dict.values()) + sum(over_dict.values()) +
sum(healthy_dict.values()) + sum(under_dict.values())

overall_dict = {"Obese": sum(obese_dict.values()), "Overweight":
sum(over_dict.values()), " a Healthy Weight": sum(healthy_dict.values()),
"Underweight": sum(under_dict.values()), "Any BMI": total}

for key in overall_dict:
   print("The total number of respondents that are {} that at least walk and perform
moderate exercise each week is: {}".format(key, overall_dict[key]))
```

Table 1B - Number of Respondents That At Least Walk and Do Moderate Exercise in The Last 7 Days by BMI Classification:

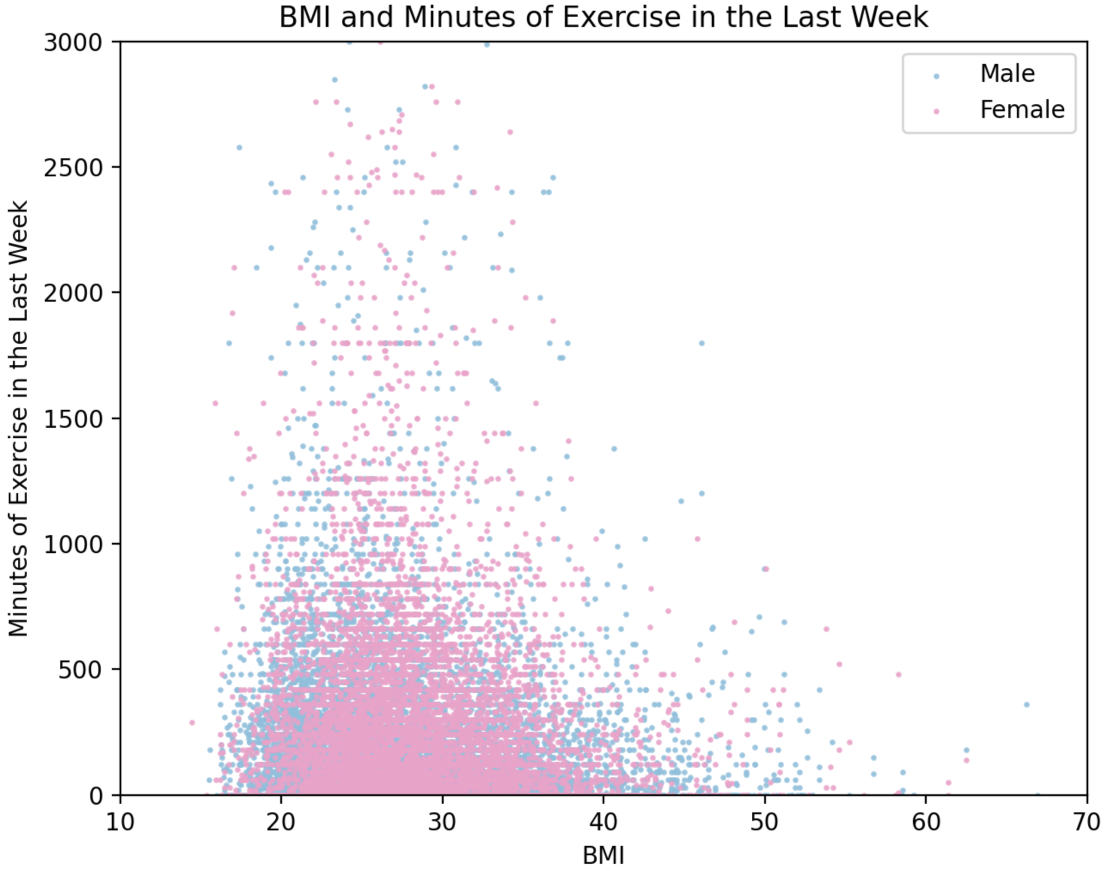
BMI	Number of Respondents That At Least Walked and Perform Moderate Exercise in the Last 7 Days
Underweight (BMI<18.5)	77
Healthy Weight (18.5<BMI<25.0)	2544
Overweight (25.0<BMI<30.0)	2566
Obese (30<BMI)	1793
Any BMI	6980

Part 4: Graph 1A

Graph 1A Code:

```
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv('cleaned_info1.csv')

df_exercise = df["EXLWTIME"]
df_bmi = df["BMISC"]

males = df[df['SEX'] == 1]
females = df[df['SEX'] == 2]

f_bmi = females['BMISC']
f_exercise = females['EXLWTIME']
m_bmi = males['BMISC']
m_exercise = males['EXLWTIME']

plt.scatter(f_bmi, f_exercise, 8, color='#91bfdb', marker='.', label='Male', alpha =
0.8)
plt.scatter(m_bmi, m_exercise, 8, color='#e9a3c9', marker='.', label='Female', alpha =
0.8)

plt.ylim([0,3000])
plt.xlim([10,70])

plt.title('BMI and Number of Days of Exercise Per Week')
plt.xlabel('BMI')
plt.ylabel('Number of Days of Exercise Per Week')
plt.legend()
plt.show()
```

Graph 1A - Relationship between BMI and Number of Minutes Exercised in the Last Week :



Graph 1A Evaluation:

Firstly, the chart type selected for this graph is a scatter plot, this chart type was selected given the two main attributes featured in the graph are two numeric variables. Additionally, given the fact that there is a natural difference in BMI between men and women, the third attribute chosen for the chart is gender of the respondent.

To construct the graph, the data was first separated into male and female data points given the encoding of the gender variable. When the original data set was obtained from the Australian Bureau of Statistics, the gender variable was already encoded using the binary encoding system with the integer value of 1 being associated with males and the integer value 2 being associated with females. Given this encoding system a data frame can be constructed that stratifies the data of BMI and time exercised within a week by each gender. This data is then used to construct the final scatter plot.

In order to differentiate between the plotted male and female data points, the male points are indicated with a blue dot while the female points are indicated with a pink dot. These visual encoding decisions allow for the viewer to differentiate between men and women respondents to control for any differences between the two samples. Furthermore, given the placement of the majority of the points, a decision was made to increase the transparency of each data point. The alpha variable within matlibplot was lowered from 1 to 0.8, thereby making the overlapping points more visible to the viewers. This decision was particularly important to ensure that the most populated section of the graph may still contain pertaining to both male and female groups. This reason is also why the data point markers were selected to be dots as almost every other data point wouldn't allow for viewers to see all of the data points.

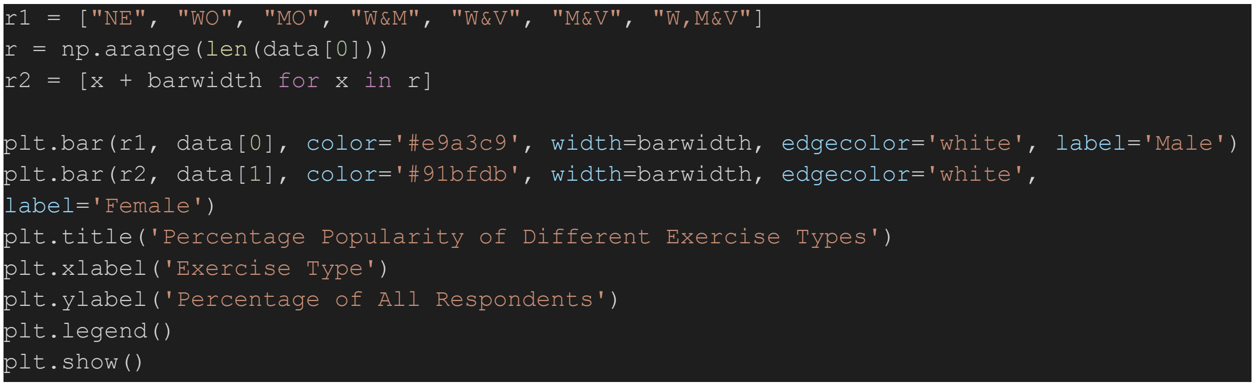
Furthermore, the scale within this graph was edited since the default value included too much white space, leaving it more challenging to locate the smaller dots. Hence, the scale on the x-axis was confined to a 10-70 BMI scale, this was chosen to close in on the most populated location in the chart while also not excluding any points. The y-axis scale was confined to 0-3000 minutes of exercise in the past week. Unlike the x-axis scale, the change in the y-axis scale excluded a few outliers from the produced graph, however, without doing so the true pattern of the data couldn’t be communicated to the audience. Hence, a couple of the outliers were left out to preserve the message of the data from the chart without distorting the meaning of the data.

Lastly, if given more data, the code is written in such a way that is nonspecific to the dataset therefore, the code should work well with more data given. More explicitly, the code converts the csv file within the directory to a dataframe which allows the code to then access each column for the corresponding variable observed. Then it stratifies the file into new data frames for each gender, at which point the values may be plotted. No matter how much more data is given, the code that computes the plottable data will hold well. However, the chart scales and visual encoding was adjusted to suit the presentation of the given data set, with more data the current specifications may not be optimal for communication of the data to audiences.

Part 5: Graph 1B

Graph 1B Code:

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

file = open("cleaned_info1.csv", "r")
male_exercise_type_list = [0,0,0,0,0,0,0]
female_exercise_type_list = [0,0,0,0,0,0,0]

first_line = True
for row in file:
   if first_line:
       first_line = False
       continue
   variables = row.strip("\n").split(",")
   exercise_type = variables[-10]
   gender = variables[0]
   if exercise_type == "1":
       continue
   elif exercise_type == "2":
       if gender == "2":
           female_exercise_type_list[0] += 1
       if gender == "1":
           male_exercise_type_list[0] += 1
   elif exercise_type == "3":
       if gender == "2":
           female_exercise_type_list[1] += 1
       if gender == "1":
           male_exercise_type_list[1] += 1
   elif exercise_type == "4":
       if gender =="2":
           female_exercise_type_list[2] += 1
       if gender == "1":
           male_exercise_type_list[2] += 1
   elif exercise_type == "5":
       if gender == "2":
           female_exercise_type_list[3] += 1
       if gender == "1":
           male_exercise_type_list[3] += 1
   elif exercise_type == "6":
       if gender == "2":
           female_exercise_type_list[4] += 1
       if gender == "1":
           male_exercise_type_list[4] += 1
   elif exercise_type == "7":
       if gender == "2":
           female_exercise_type_list[5] += 1
       if gender == "1":
           male_exercise_type_list[5] += 1
   elif exercise_type == "8":
       if gender == "2":
           female_exercise_type_list[6] += 1
       if gender == "1":
           male_exercise_type_list[6] += 1


perc_male_list = male_exercise_type_list[:]
perc_female_list = female_exercise_type_list[:]

i = 0
while i < len(male_exercise_type_list):
   new_pop = (male_exercise_type_list[i]/sum(male_exercise_type_list))*100
   perc_male_list[i] = new_pop
   i += 1

i = 0
while i < len(female_exercise_type_list):
   new_pop = (female_exercise_type_list[i]/sum(female_exercise_type_list))*100
   perc_female_list[i] = new_pop
   i += 1

data = [
   perc_male_list,
   perc_female_list
]

barwidth = 0.25
```



```
r1 = ["NE", "WO", "MO", "W&M", "W&V", "M&V", "W,M&V"]
r = np.arange(len(data[0]))
r2 = [x + barwidth for x in r]

plt.bar(r1, data[0], color='#e9a3c9', width=barwidth, edgecolor='white', label='Male')
plt.bar(r2, data[1], color='#91bfdb', width=barwidth, edgecolor='white',
label='Female')
plt.title('Percentage Popularity of Different Exercise Types')
plt.xlabel('Exercise Type')
plt.ylabel('Percentage of All Respondents')
plt.legend()
plt.show()
```

Graph 1B - Percentage Popularity of The Different Types of Exercised By Gender:

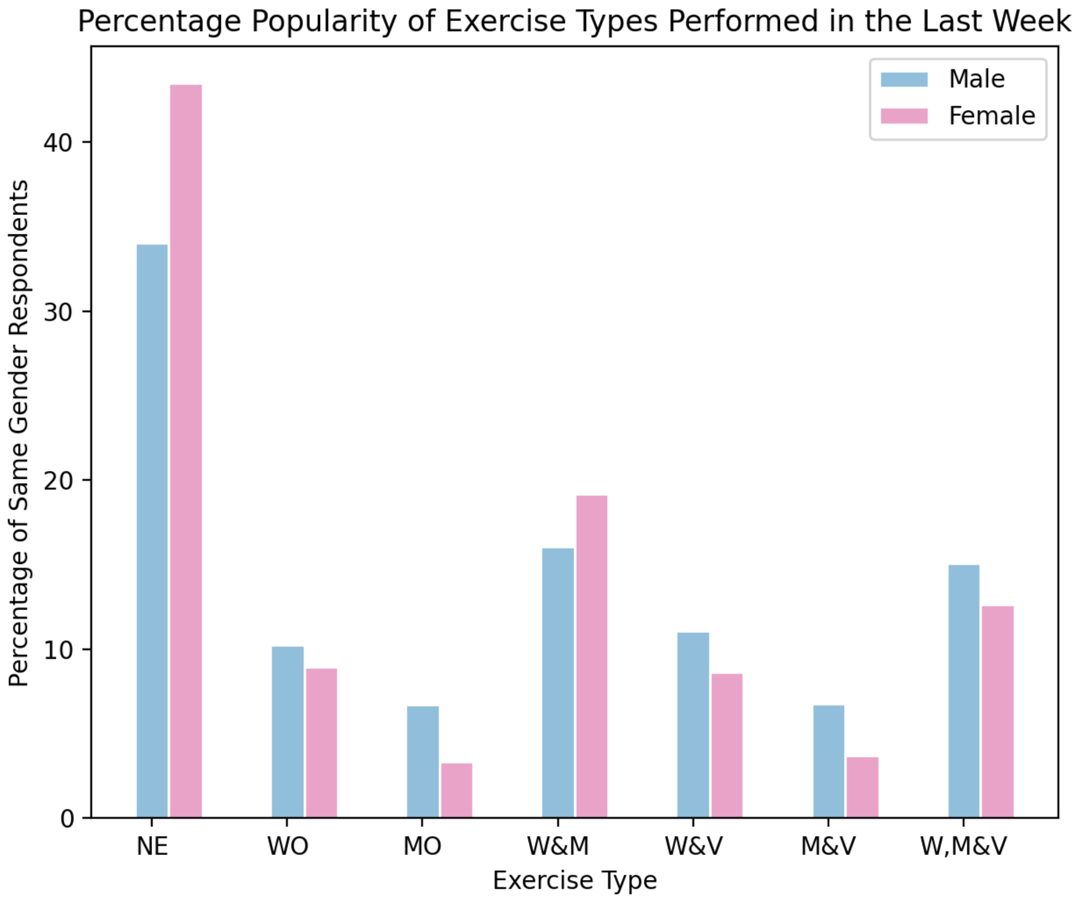


Chart Evaluation:

Firstly, a bar chart was chosen for the output of this graph, given that the main attribute within question (exercise type) is a nominal data type suggesting that a bar chart would be best to capture the trends within the data.

The original data set acquired from The Australian Bureau of Statistics had all of the attributes concerned within graph 1B already encoded. First, since the gender of respondents was recorded as either male or female, the variable uses a binary encoding with 1 to represent male respondents and 2 to represent female respondents. On the other hand, the nominal variable that represents the exercise type of the respondents has 8 different values as mentioned previously. The encoding for this variable is a label encoding system that assigns each exercise type to an integer within the range 1-8 in order from least exercise to most exercise. The code to produce the graph therefore, uses these two encoding systems first to plot the bar charts for each exercise type and then to stratify the male and female samples within each type. By doing so, each of these attributes are communicated to the audience effectively, clearly showing the popular exercise types overall but then also the difference between male and females exercise choices.

In this particular graph, the percentage of each gender group was used as the height for each exercise type bar. It was noticed early on that the total number of respondents for men and women was not equal and hence, comparing the number of each gender within each exercise type may have been misleading. To ensure that the graph is informative and avoid ambiguous presentation, the percentage was used instead, making the comparisons between the exercise groups and between the genders possible. Thereby, this design choice effectively compares each group to one another in a fair way allowing for the possible deductions of the graph conclusive.

In order for viewers to differentiate between the different attributes shown within the charts two visual characteristics were employed. First, the colour of each bar represents the gender that the bar is representing which are chosen in alignment with blue for males and pink for females. Additionally, each exercise type is ordered across the x-axis from least amount of exercise to most amount of exercise in a given week, marked by the short form ticks on the axis. By demonstrating these two elements, viewers can easily compare results not only between genders but also the different amounts of exercise within the same gender. With that being said, the short form notation of each exercise type is not optimal as viewers may not initially understand each shorthand at first, however it was believe that this choice is better for the presentation of the graph such that the audience's attention is directed towards the data and lessens the clutter of the graph.

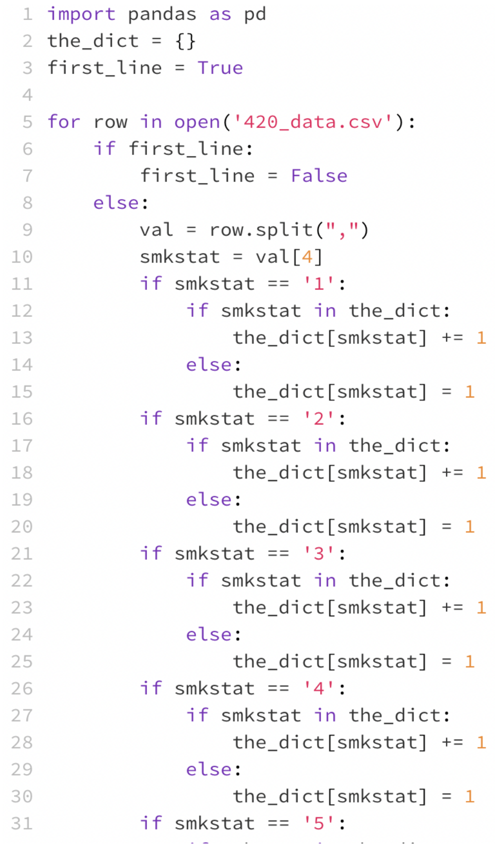
Overall, the graph does an effective job at encapsulating the relationship between the 3 featured attributes and engaging the audience into discovering the patterns within the data and wanting to learn more. Additionally, if given more data, the code would work well to add in the additional data into the graph. Within the code itself, the program utilizes code that is generalized such as within the creation of the lists, transformation of those lists and adding values to the lists that are then shown within the graph. The only exception that comes to mind is the instance within which the nominal variable itself changes to include more types of exercise; however, since the health survey is standard each year this aspect is unlikely to change with more data. Thus, overall the code works effectively to communicate the patterns found within the data and in the instance that more data is provided, the program should correctly incorporate the additional values.

Smoking (500459007)

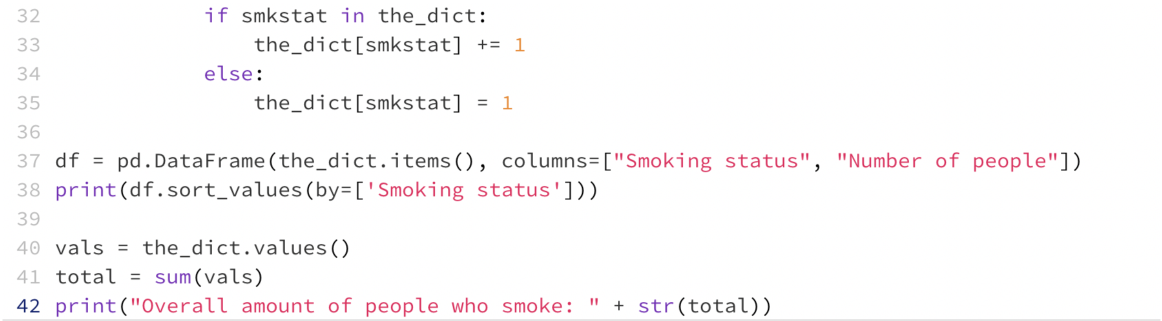
The dataset appropriately named “420_data.csv” was used in analysing if there were any correlations between one’s **smoking habits and BMI**. It is a text file under the comma-separated value (csv) file format. This dataset was originally part of an extremely large dataset from the Australian Bureau of Statistics with over 21,000 values. The specific values regarding individual smoking habits and BMI were extracted and replicated on the new, current spreadsheet on the 28th of September, 2021. It contains 21316 rows with 15 attributes, with values ranging from one’s smoking status to the usual number of cigarettes smoked each week.

Attribute	**Meaning**
BMI	Individual’s body mass index
SMOKEQ1	Individual responses if one currently is a smoker
SMOKEQ5	Individual responses if one has smoked more than 100 cigarettes
SMKDAILY	Individual responses onto one’s *daily* smoking status
SMKSTAT	Individual responses onto one’s smoking status

With the dataset, some analysis were done through Python code in forming grouped aggregate summaries.

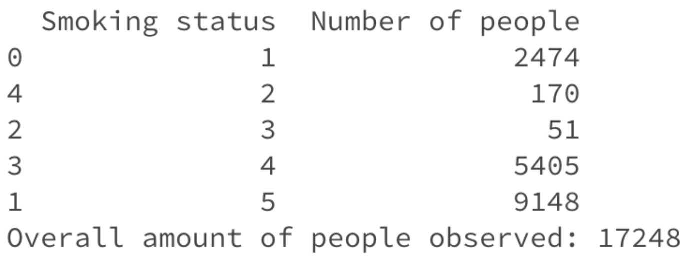


```
import pandas as pd
the_dict = {}
first_line = True

for row in open('420_data.csv'):
    if first_line:
        first_line = False
    else:
        val = row.split(",")
        smkstat = val[4]
        if smkstat == '1':
            if smkstat in the_dict:
                the_dict[smkstat] += 1
            else:
                the_dict[smkstat] = 1
        if smkstat == '2':
            if smkstat in the_dict:
                the_dict[smkstat] += 1
            else:
                the_dict[smkstat] = 1
        if smkstat == '3':
            if smkstat in the_dict:
                the_dict[smkstat] += 1
            else:
                the_dict[smkstat] = 1
        if smkstat == '4':
            if smkstat in the_dict:
                the_dict[smkstat] += 1
            else:
                the_dict[smkstat] = 1
        if smkstat == '5':
```



```
            if smkstat in the_dict:
                the_dict[smkstat] += 1
            else:
                the_dict[smkstat] = 1

df = pd.DataFrame(the_dict.items(), columns=["Smoking status", "Number of people"])
print(df.sort_values(by=['Smoking status']))

vals = the_dict.values()
total = sum(vals)
print("Overall amount of people who smoke: " + str(total))
```

Output:



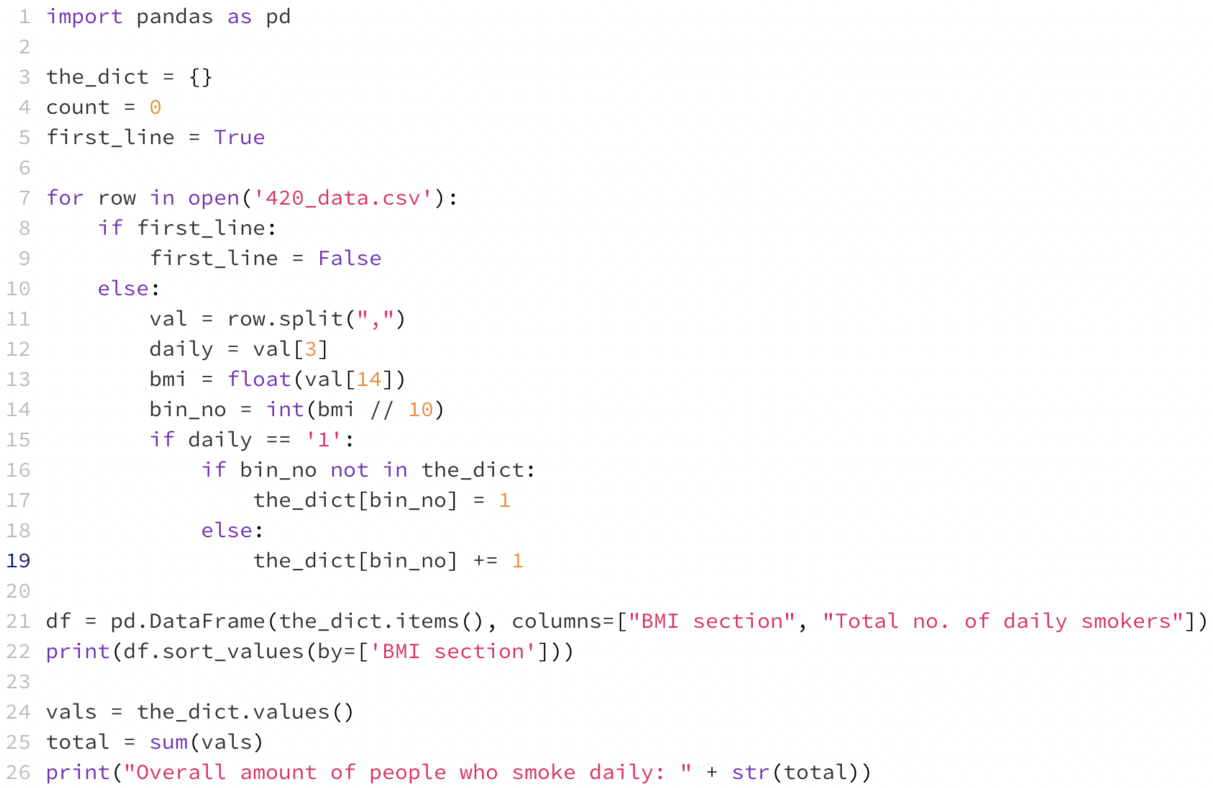
```
  Smoking status  Number of people
0               1              2474
4               2               170
2               3                51
3               4              5405
1               5              9148
Overall amount of_people observed: 17248
```

In the dataset, each person’s smoking status response was categorized into 5 groups with each number representing a certain response:

1. Current daily smoker
2. Current weekly smoker (at least once a week but not daily)
3. Current non-weekly smoker (less than weekly)
4. Ex-smoker (no longer smokes)
5. Non-smoker (does not smoke)

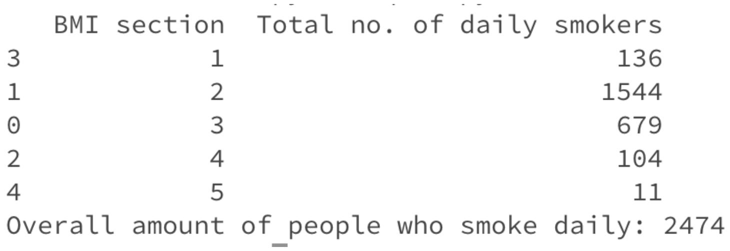
The code above grouped each response with respect to their category and calculated the number of people who responded to each category.

Another section of code (seen below) binned each BMI into sections and calculated the total number of daily smokers in each category.



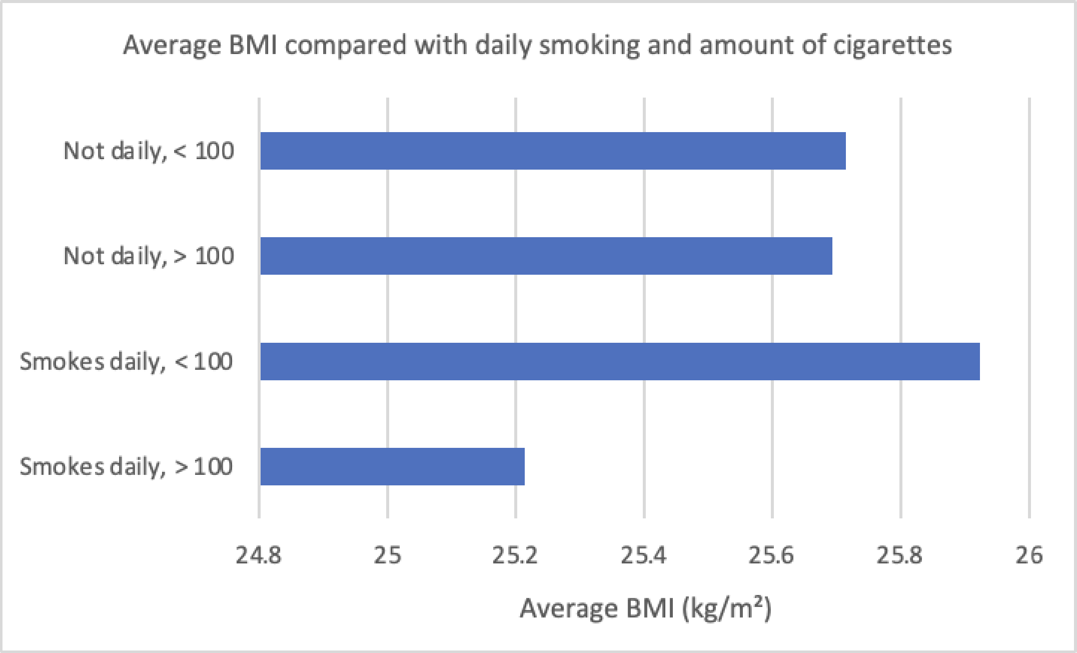
```python
import pandas as pd

the_dict = {}
count = 0
first_line = True

for row in open('420_data.csv'):
    if first_line:
        first_line = False
    else:
        val = row.split(",")
        daily = val[3]
        bmi = float(val[14])
        bin_no = int(bmi // 10)
        if daily == '1':
            if bin_no not in the_dict:
                the_dict[bin_no] = 1
            else:
                the_dict[bin_no] += 1

df = pd.DataFrame(the_dict.items(), columns=["BMI section", "Total no. of daily smokers"])
print(df.sort_values(by=['BMI section']))

vals = the_dict.values()
total = sum(vals)
print("Overall amount of people who smoke daily: " + str(total))
```

Output:



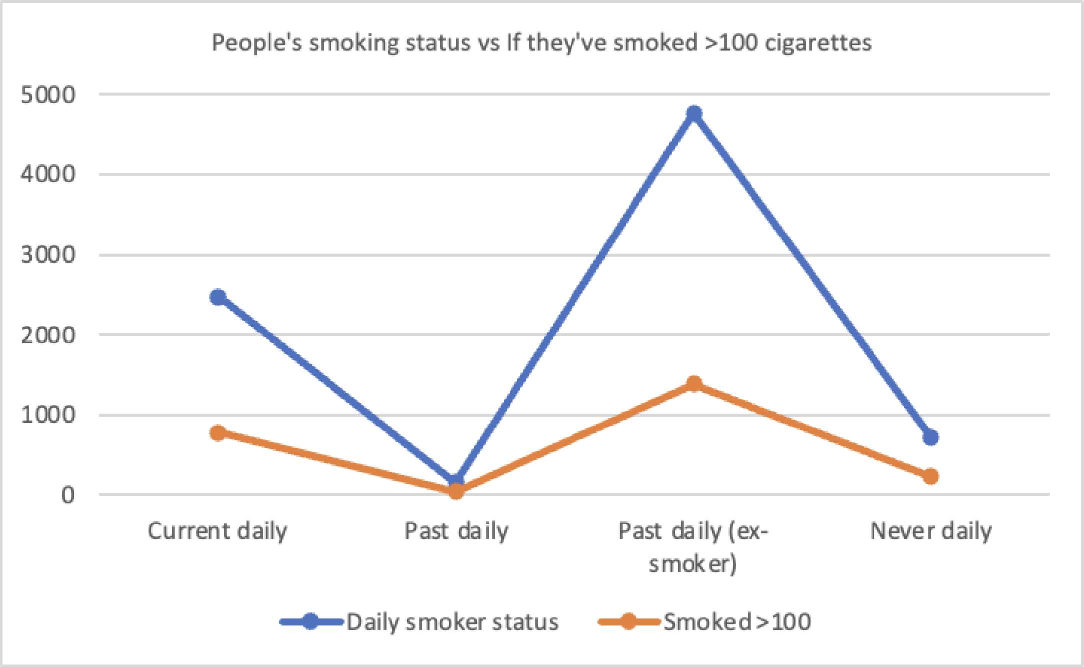
```
   BMI section  Total no. of daily smokers
3            1                         136
1            2                        1544
0            3                         679
2            4                         104
4            5                          11
Overall amount of people who smoke daily: 2474
```

Along with coding, some charts were formed in presenting some variables within the dataset. The following horizontal bar chart was produced via Microsoft Excel through utilising the available responses with whether individuals **smoked daily** and whether individuals have **smoked greater than 100 cigarettes** in their lifetime and comparing those with the average **BMI** for each category in those variables. The chart was made through filtering the responses in each specific category through the *filter function* on Excel and calculating the mean BMI for the all the individuals in each filtered category using the $= AVERAGE$ function and selecting the BMI column of data. The resulting data was then separated and used to form the chart through the *insert chart function* on Excel.



With the chart above, the average BMI for each category is encoded by the x-position (horizontal) across the graph with the length of each bar. The specific chart style of the horizontal bar chart was used as it provides viewers with an effective and efficient visual representation of the average BMI values. Although comparing the length of objects is not at the forefront for ease of human perception, the chart depicts each bar's length clearly and precisely, with one category having a larger value compared to the remaining categories providing viewers with a clear answer of which category has the highest average BMI value. If another chart style were to be used, the visualisation of recorded values would not be as effective or 'easy to understand' due to the differences between each value being simply just 1 to 1.5 off. From an abstract perspective the values may be clear as to which is larger. However from a visual perspective with other charts such as a line chart or a pie chart, the difference between each value is too small to be clearly observed by viewers, making the bar chart the best, most effective option. If there were more data obtained and included in the chart, the graph would hold and maintain its effectiveness. However, it could be improved with adding separate colours with each bar in visualising each variable independently, as noticing different colours heavily aligns with the ease of human perception.

Another chart made through Microsoft Excel was the following marked line chart which depicts the total number of people in each **smoking status** and compares that data with the amount of people in each category who **smoked greater than 100 cigarettes** in their lifetime. This chart was made primarily using the $= COUNTIF$ function on Excel which totalled the number of specific responses in each column of data, resulting in the total number of individuals for each smoking status. The same steps were repeated for the second variable of whether individuals smoked greater than 100 cigarettes in their lifetime and the results of both were then extracted and used to form the marked line chart once again through the *insert chart function* on Excel.



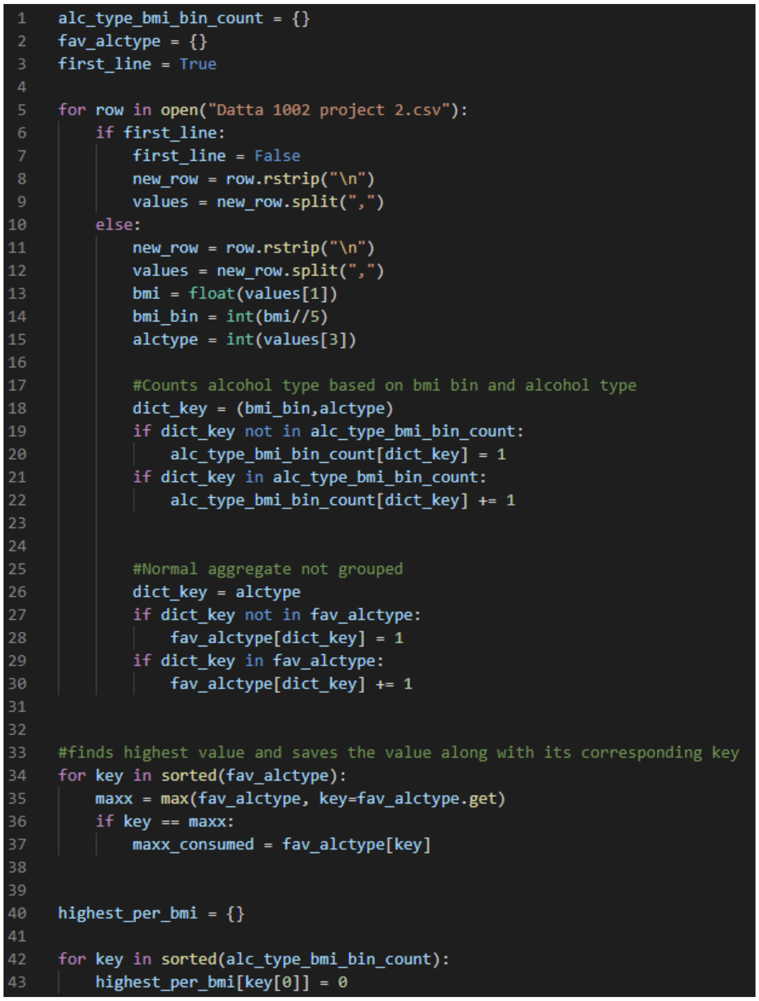
With the chart, the number of people in each smoking category as well as the number of those specific people who have smoked greater than 100 cigarettes in their lifetime are both encoded via the y-axis position (vertical) on the graph. The blue line is used to encode and distinguish the number of people in each daily smoking status and the green line is used to distinguish the number of those people who smoked more than 100 cigarettes. The specific chart style of the marked line chart was used in comparing the two quantitative values as it effectively visualises not only the values of each variable but also the trends and patterns from each value compared to other variables. This encoding and chart style appeals to the ease of perception and visualisation with positioning noted as the preferred encoding style for most quantitative variables. The different colours used for each variable also aids towards the effectiveness of the chart's visual communication amongst viewers as distinguishing, contrasting colours for variables further eases human perception of different variables. If more data were to be obtained and included on this chart, it would not hold its effectiveness as the chart would appear more cluttered and 'crowded' with each value being squeezed onto the chart. This would decrease the overall visual effectiveness of the chart as viewers may be left confused or trends in the data (if there are any) may not be visible or are difficult to identify.

500490723 - Alcohol

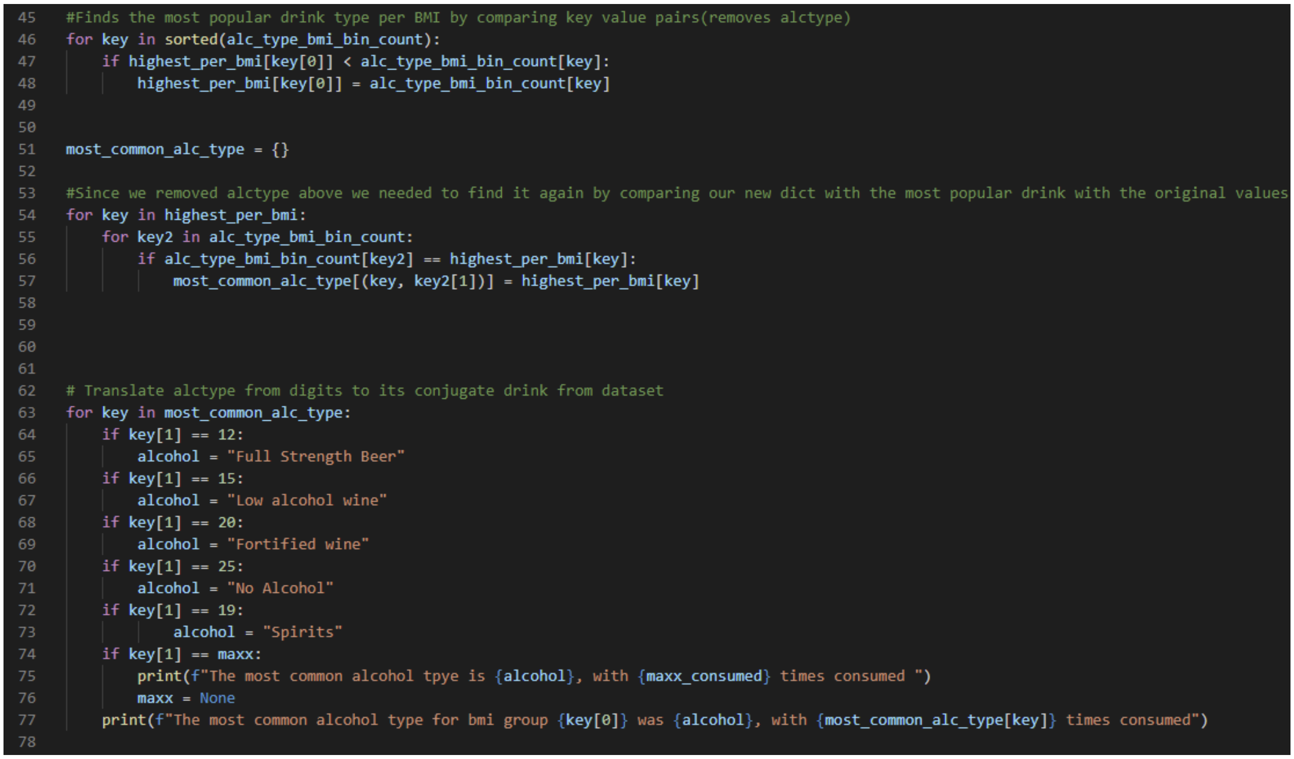
The data for my dataset was originally obtained from the National Health Survey conducted by the Australian Bureau of Statistic from 2017-2018. Specific data points were picked out according to its relevance to my research topic and put into a new dataset. The new dataset contains approximately 20,500 rows and 10 attributes. Description of data:

BMISC	Body mass index (BMI) - score measured
ALCDAY	Order of consumption - most recent 3 days in last week
DAYFLG	Day of consumption(Monday, Tuesday, etc.)
TOTPA3	Total quantity of pure alcohol consumed by day - in mls
TOTPAL	Total quantity of pure alcohol consumed(week) - in mls
STDDRKDY	Number of standard drinks by day
ALCSDCDT	Number of standard drinks consumed by type of drink consumed(week)
ALCTYPE	Type of drink consumed

Most Favoured alcohol type per each BMI group(BMI Bin):



```python
alc_type_bmi_bin_count = {}
fav_alctype = {}
first_line = True

for row in open("Datta 1002 project 2.csv"):
    if first_line:
        first_line = False
        new_row = row.rstrip("\n")
        values = new_row.split(",")
    else:
        new_row = row.rstrip("\n")
        values = new_row.split(",")
        bmi = float(values[1])
        bmi_bin = int(bmi//5)
        alctype = int(values[3])

        #Counts alcohol type based on bmi bin and alcohol type
        dict_key = (bmi_bin,alctype)
        if dict_key not in alc_type_bmi_bin_count:
            alc_type_bmi_bin_count[dict_key] = 1
        if dict_key in alc_type_bmi_bin_count:
            alc_type_bmi_bin_count[dict_key] += 1


        #Normal aggregate not grouped
        dict_key = alctype
        if dict_key not in fav_alctype:
            fav_alctype[dict_key] = 1
        if dict_key in fav_alctype:
            fav_alctype[dict_key] += 1


#finds highest value and saves the value along with its corresponding key
for key in sorted(fav_alctype):
    maxx = max(fav_alctype, key=fav_alctype.get)
    if key == maxx:
        maxx_consumed = fav_alctype[key]


highest_per_bmi = {}

for key in sorted(alc_type_bmi_bin_count):
    highest_per_bmi[key[0]] = 0
```



```
#Finds the most popular drink type per BMI by comparing key value pairs(removes alctype)
for key in sorted(alc_type_bmi_bin_count):
    if highest_per_bmi[key[0]] < alc_type_bmi_bin_count[key]:
        highest_per_bmi[key[0]] = alc_type_bmi_bin_count[key]


most_common_alc_type = {}

#Since we removed alctype above we needed to find it again by comparing our new dict with the most popular drink with the original values
for key in highest_per_bmi:
    for key2 in alc_type_bmi_bin_count:
        if alc_type_bmi_bin_count[key2] == highest_per_bmi[key]:
            most_common_alc_type[(key, key2[1])] = highest_per_bmi[key]




# Translate alctype from digits to its conjugate drink from dataset
for key in most_common_alc_type:
    if key[1] == 12:
        alcohol = "Full Strength Beer"
    if key[1] == 15:
        alcohol = "Low alcohol wine"
    if key[1] == 20:
        alcohol = "Fortified wine"
    if key[1] == 25:
        alcohol = "No Alcohol"
    if key[1] == 19:
            alcohol = "Spirits"
    if key[1] == maxx:
        print(f"The most common alcohol tpye is {alcohol}, with {maxx_consumed} times consumed ")
        maxx = None
    print(f"The most common alcohol type for bmi group {key[0]} was {alcohol}, with {most_common_alc_type[key]} times consumed")

```

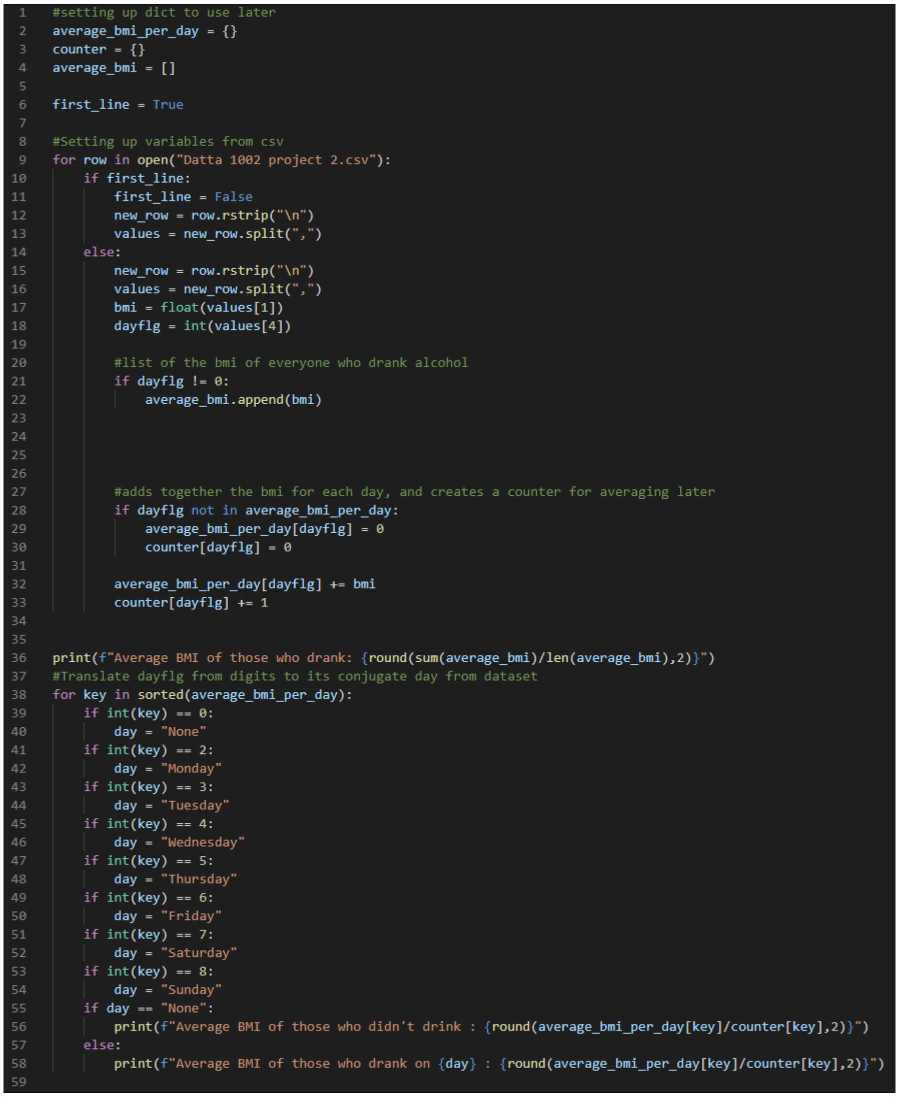
Output: Table 3a

BMI Group (BMI // 5)	Favoured Alcohol Type	Times consumed
2	Full Strength beer	94
3	Full Strength beer	568
4	Full Strength beer	1086
5	Full Strength beer	1267
6	No Alcohol	673
7	Full Strength beer	284
8	Full Strength beer	92
9	No Alcohol	41
10	Fortified Wine	15
10	No alcohol	15
10	Low alcohol wine	15
11	Low alcohol wine	6
11	Fortified wine	6
11	No Alcohol	6
All(no grouped aggregate)	Full Strength beer	4048

Note:

As there were less and less data points for the later BMI groups(BMI>50), there were bmi groups with a 7 way tie for favoured alcohol with times consumed being very small (one of them was only 2). While these are technically correct results, each group only represents 0.009% of the total population and it is impossible to deduce anything for these values. Therefore I stopped showing the last few bmi groups as it still gives the same general idea.

Average BMI per day of week which alcohol is consumed:



```python
#setting up dict to use later
average_bmi_per_day = {}
counter = {}
average_bmi = []

first_line = True

#Setting up variables from csv
for row in open("Datta 1002 project 2.csv"):
    if first_line:
        first_line = False
        new_row = row.rstrip("\n")
        values = new_row.split(",")
    else:
        new_row = row.rstrip("\n")
        values = new_row.split(",")
        bmi = float(values[1])
        dayflg = int(values[4])

        #list of the bmi of everyone who drank alcohol
        if dayflg != 0:
            average_bmi.append(bmi)



        #adds together the bmi for each day, and creates a counter for averaging later
        if dayflg not in average_bmi_per_day:
            average_bmi_per_day[dayflg] = 0
            counter[dayflg] = 0

        average_bmi_per_day[dayflg] += bmi
        counter[dayflg] += 1


print(f"Average BMI of those who drank: {round(sum(average_bmi)/len(average_bmi),2)}")
#Translate dayflg from digits to its conjugate day from dataset
for key in sorted(average_bmi_per_day):
    if int(key) == 0:
        day = "None"
    if int(key) == 2:
        day = "Monday"
    if int(key) == 3:
        day = "Tuesday"
    if int(key) == 4:
        day = "Wednesday"
    if int(key) == 5:
        day = "Thursday"
    if int(key) == 6:
        day = "Friday"
    if int(key) == 7:
        day = "Saturday"
    if int(key) == 8:
        day = "Sunday"
    if day == "None":
        print(f"Average BMI of those who didn't drink : {round(average_bmi_per_day[key]/counter[key],2)}")
    else:
        print(f"Average BMI of those who drank on {day} : {round(average_bmi_per_day[key]/counter[key],2)}")
```

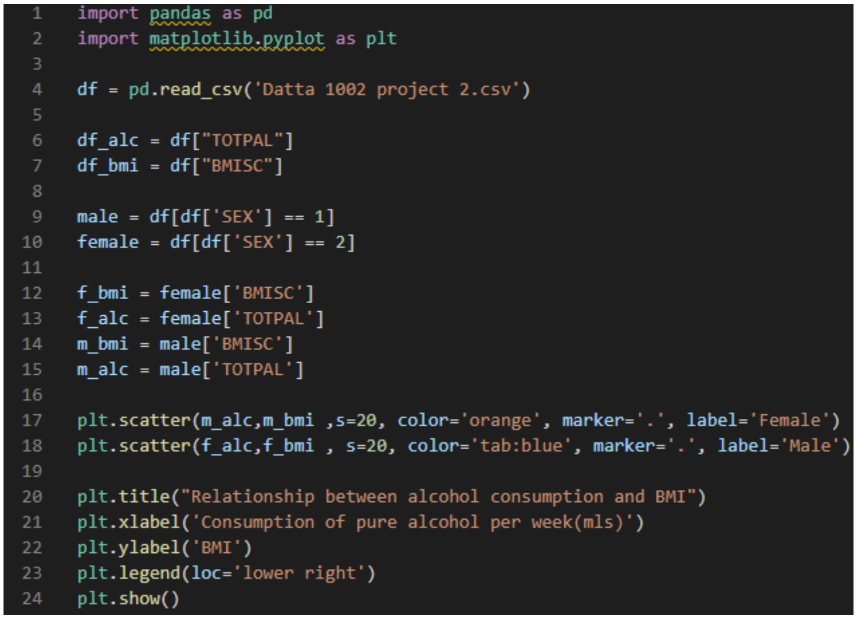
Output: Table 3b

Day of consumption	Average BMI
All(no grouped aggregate)	26.43
None	26.4
Monday	26.52
Tuesday	26.64
Wednesday	26.27

Thursday	26.29
Friday	26.45
Saturday	26.42
Sunday	26.42

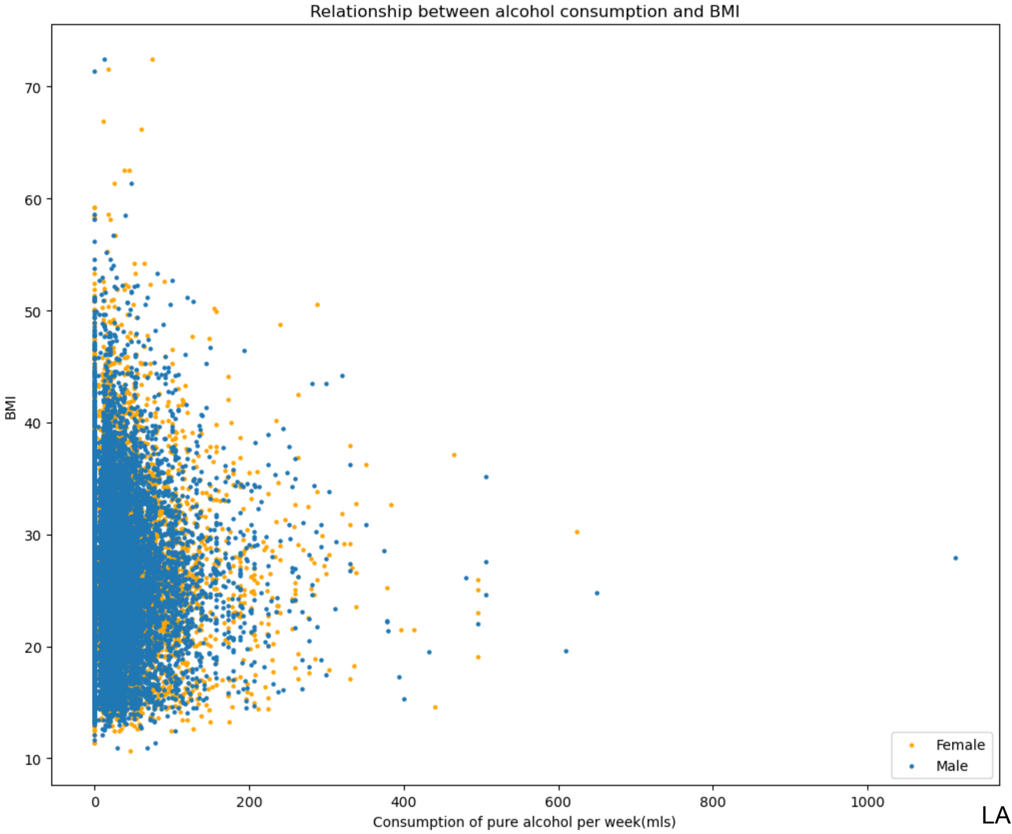
Charting:

The first chart was created through the following python code:



```python
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv('Datta 1002 project 2.csv')

df_alc = df["TOTPAL"]
df_bmi = df["BMISC"]

male = df[df['SEX'] == 1]
female = df[df['SEX'] == 2]

f_bmi = female['BMISC']
f_alc = female['TOTPAL']
m_bmi = male['BMISC']
m_alc = male['TOTPAL']

plt.scatter(m_alc,m_bmi ,s=20, color='orange', marker='.', label='Female')
plt.scatter(f_alc,f_bmi , s=20, color='tab:blue', marker='.', label='Male')

plt.title("Relationship between alcohol consumption and BMI")
plt.xlabel('Consumption of pure alcohol per week(mls)')
plt.ylabel('BMI')
plt.legend(loc='lower right')
plt.show()
```

The resulting graph: Chart 3a

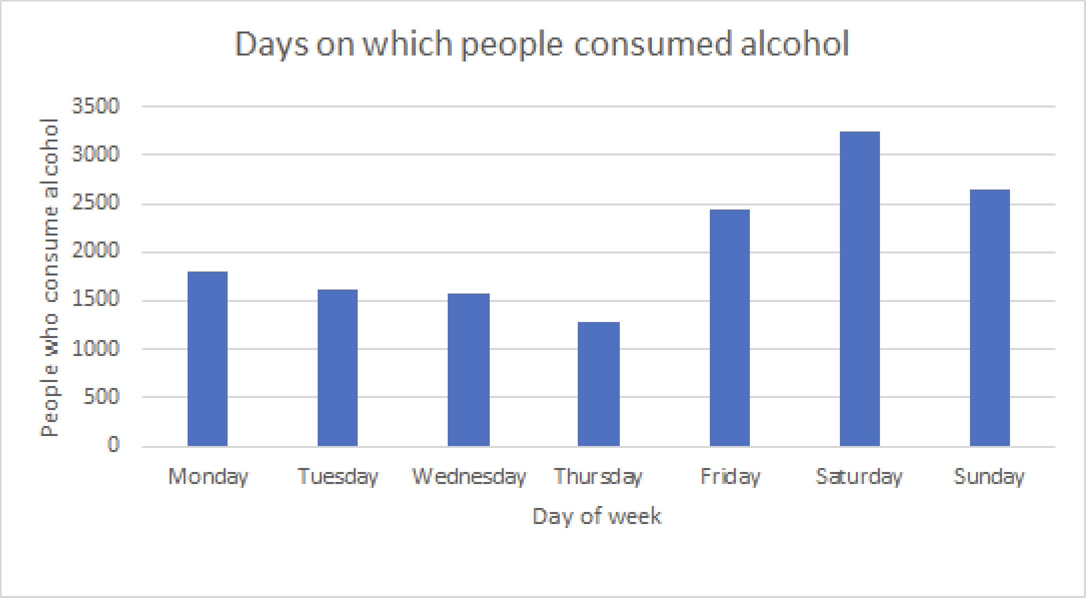


For this graph, the x axis was encoded with the variable relating to alcohol consumption, and the y axis was encoded with BMI.Both BMI and alcohol consumption was easily turned into a dataframe from the original csv through pandas as they both already had the desired format of an integer. To encode this data with the last variable (gender) I needed to convert the strings “male” and “female” to 1 and 2 respectively. This data could now be used to colour code the other data points resulting in the above graph.

The style of graph was chosen to be a scatterplot as it is one of the more efficient ways to produce a chart between 2 quantitative data types with a lot of data points (20,000+) which have not been summarised. Small dots were chosen to visually display the data points as since there was so much data, it was the best chance to display as much data as possible. Despite this, solid sections of blue appear, completely covering up the underlying orange data points. Changing the order in which the male and female data points are added to the graph swaps the layering of the colours but the same issue occurs; sections of solid orange coving up the underlying blue data. This definitely affects the effectiveness of this graph as it is harder to extract information at first because the majority of data is covered up by other data. If the assumption is made that the data underneath looks similar to what's on top the effect lessens.In the end it still did a reasonable job displaying the relationship between the data.

The chart worked reasonably well displaying the data I wanted, but it already started to struggle with the sheer amount of data that needed to be portrayed. If even more data was added the best way for this chart to work is simply displaying a sample of the data, or summarising the data somehow to reduce the amount of data points on the graph.

Chart 3b:



This second chart was not made using python code but rather through excel. I used the ‘=countif’ function on the column of data pertaining to the day of consumption (DAYFLG variable) to create a total for each day. These totals were then placed in the same column under the heading ‘People who consumed alcohol’. A new column was created(with a heading of ‘day of week’) adjacent to the aforementioned column which contained the days of the week (monday, tuesday,etc.), lining up with their respective values. These totals are then displayed by each individual bar of the bar graph.

A bar graph was then chosen as it effectively portrays the relationship with adjacent days as well as separated ones. The several grid lines also makes it even easier to discern the difference between each day. The simplicity of the chart, having only 2 variables, makes it very easy and effective to portray the information as there is only 1 visual aspect which needs to be comprehended; the length of the bar, there is no secondary bar of split bars with colour coding which could make it less easily readable.

This chart is fully adaptable to almost any size of data as it already does a great job portraying data with 20,000+ individual data points. The only potential issue could be the y axis numbers being too big e.g. 200000000, which could simply be solved by using a different scale.

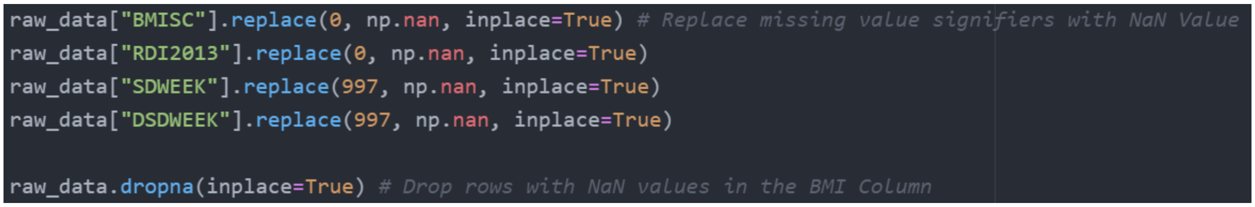
500308222 - Dietary Habits

The dataset used is “NHS17SPB.csv” obtained from MicrodataDownload run by the Australian Bureau of Statistics.

Data Variables:

- BMISC: Body mass index (BMI) – score measure
    - Continuous variable
    - 0: Not applicable; Single unit values <0.01…96.99>
- RDI2013: Whether vegetable and fruit consumption met recommended guidelines (2013 NHRMC guidelines)
    - Categorical variable
    - 0: Not Applicable; 1. Met both fruit and vegetable guidelines; 2. Met vegetable guideline only; 3. Met fruit guideline only; 4. Did not meet either fruit or vegetable guideline
- SEX: Sex of person
    - Categorical variable
    - 1: Male; 2: Female
- SDWEEK: Number of metric cups of selected sugar sweetened drinks usually consumed per week
    - Continuous variable
    - 0: Do not usually consume selected sugar sweetened drinks; Number of cups in single unit values <1…700>; 997. Not applicable
- DSDWEEK: Number of metric cups of selected diet drinks usually consumed per week
    - Continuous variable
    - 0: Do not usually consume selected diet drinks; Number of cups in single unit values <1…700>; 997. Not applicable

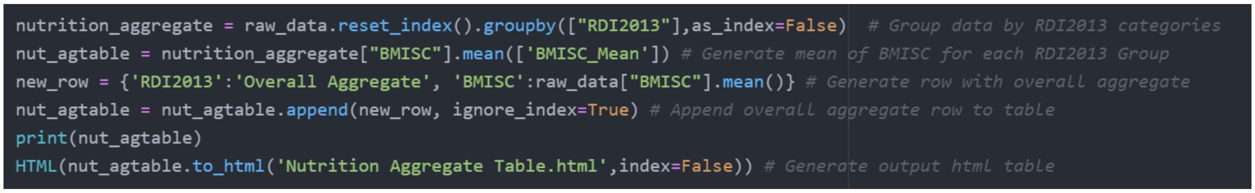
Missing value signifiers like 0: Not applicable, or 997: Not applicable were converted into NaN values and rows with NaN values were removed



```
raw_data["BMISC"].replace(0, np.nan, inplace=True) # Replace missing value signifiers with NaN Value
raw_data["RDI2013"].replace(0, np.nan, inplace=True)
raw_data["SDWEEK"].replace(997, np.nan, inplace=True)
raw_data["DSDWEEK"].replace(997, np.nan, inplace=True)

raw_data.dropna(inplace=True) # Drop rows with NaN values in the BMI Column
```

Grouped-Aggregate Summaries:

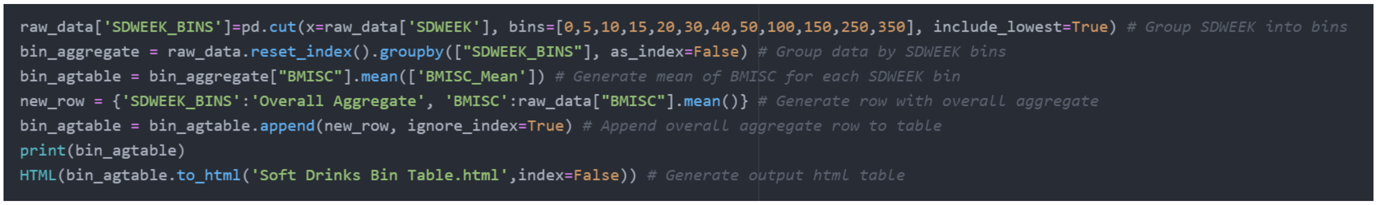
1. Summary of BMISC Means grouped by RDI2013



```
nutrition_aggregate = raw_data.reset_index().groupby(["RDI2013"],as_index=False)  # Group data by RDI2013 categories
nut_agtable = nutrition_aggregate["BMISC"].mean(['BMISC_Mean']) # Generate mean of BMISC for each RDI2013 Group
new_row = {'RDI2013':'Overall Aggregate', 'BMISC':raw_data["BMISC"].mean()} # Generate row with overall aggregate
nut_agtable = nut_agtable.append(new_row, ignore_index=True) # Append overall aggregate row to table
print(nut_agtable)
HTML(nut_agtable.to_html('Nutrition Aggregate Table.html',index=False)) # Generate output html table
```

RDI2013	BMISC Mean
1	25.942592
2	27.776511
3	25.635973
4	27.353401
Overall Aggregate	26.424420

2. Summary of BMISC Means grouped by binned SDWEEK Variable



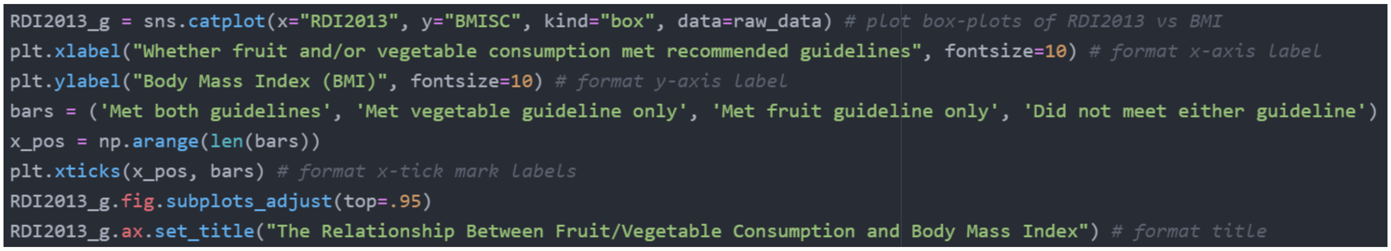
```
raw_data['SDWEEK_BINS']=pd.cut(x=raw_data['SDWEEK'], bins=[0,5,10,15,20,30,40,50,100,150,250,350], include_lowest=True) # Group SDWEEK into bins
bin_aggregate = raw_data.reset_index().groupby(["SDWEEK_BINS"], as_index=False) # Group data by SDWEEK bins
bin_agtable = bin_aggregate["BMISC"].mean(['BMISC_Mean']) # Generate mean of BMISC for each SDWEEK bin
new_row = {'SDWEEK_BINS':'Overall Aggregate', 'BMISC':raw_data["BMISC"].mean()} # Generate row with overall aggregate
bin_agtable = bin_agtable.append(new_row, ignore_index=True) # Append overall aggregate row to table
print(bin_agtable)
HTML(bin_agtable.to_html('Soft Drinks Bin Table.html',index=False)) # Generate output html table
```

SDWEEK_BINS	BMISC
(-0.001, 5.0]	26.275029
(5.0, 10.0]	26.780041
(10.0, 15.0]	27.370037
(15.0, 20.0]	26.967385
(20.0, 30.0]	28.340479

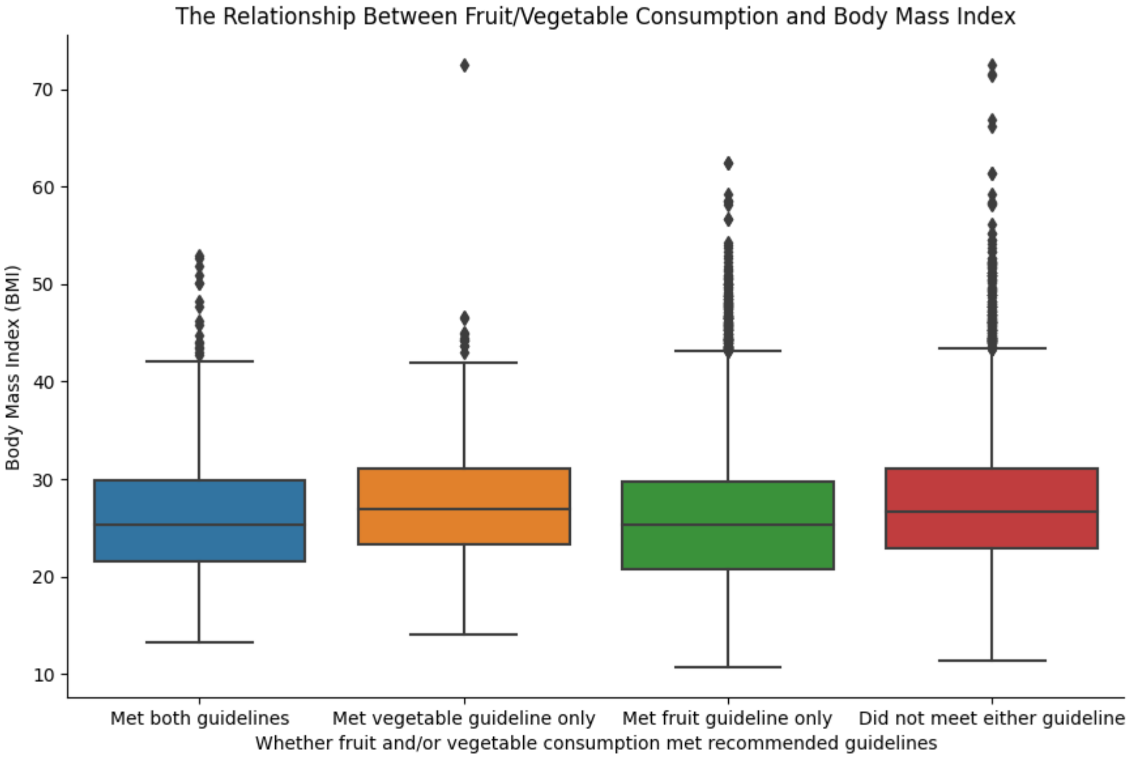
(30.0, 40.0]	28.188611
(40.0, 50.0]	29.521310
(50.0, 100.0]	28.517113
(100.0, 150.0]	23.595000
(150.0, 250.0]	30.275000
(250.0, 350.0]	32.800000
Overall Aggregate	26.424420

Chart Production:

A. RDI2013 vs BMI Chart



```
RDI2013_g = sns.catplot(x="RDI2013", y="BMISC", kind="box", data=raw_data) # plot box-plots of RDI2013 vs BMI
plt.xlabel("Whether fruit and/or vegetable consumption met recommended guidelines", fontsize=10) # format x-axis label
plt.ylabel("Body Mass Index (BMI)", fontsize=10) # format y-axis label
bars = ('Met both guidelines', 'Met vegetable guideline only', 'Met fruit guideline only', 'Did not meet either guideline')
x_pos = np.arange(len(bars))
plt.xticks(x_pos, bars) # format x-tick mark labels
RDI2013_g.fig.subplots_adjust(top=.95)
RDI2013_g.ax.set_title("The Relationship Between Fruit/Vegetable Consumption and Body Mass Index") # format title
```



- The BMI of each participant in a category is encoded in the y-position of the graph, and the levels of RDI2013 are encoded across the x-axes

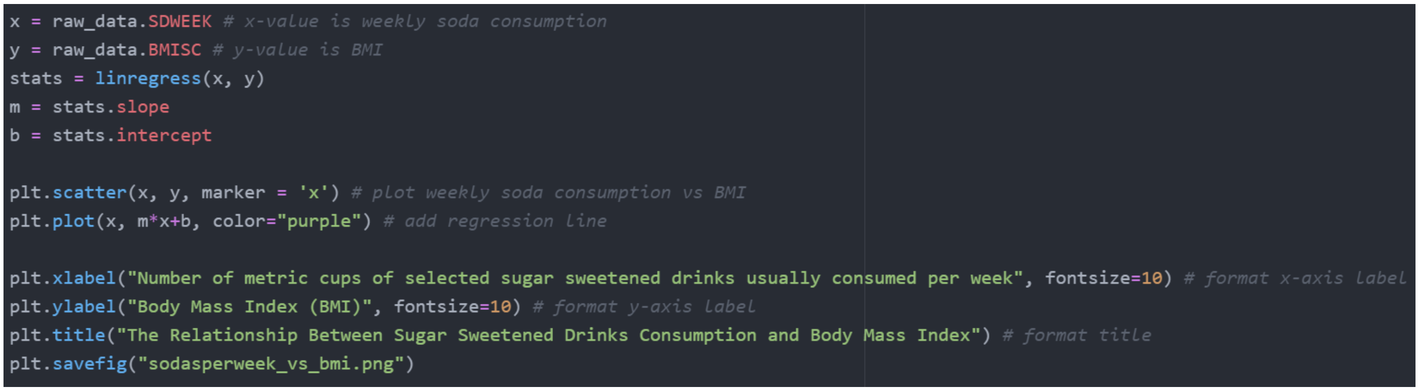
Style of chart: box plots corresponding to each level of the categorical variable RDI2013

- Chose box plots to represent the relationship between a quantitative variable and a categorical variable because they are visually simple and allow the viewer to quickly see the relationship between the levels and their corresponding BMI values (mean, interquartile ranges, overall range)
- The colours are distinct, and the clear space separating each boxplot leads to less confusion when comparing the variable levels
- It allows appreciation of the BMI values of each participant in a way that a bar chart or a line graph could not, which in turn better displays the complex relationship between BMI and nutritional intake
- The graph is quite wide to accommodate the long variable level names (‘met both guidelines’, ‘met vegetable guideline only’, etc.)

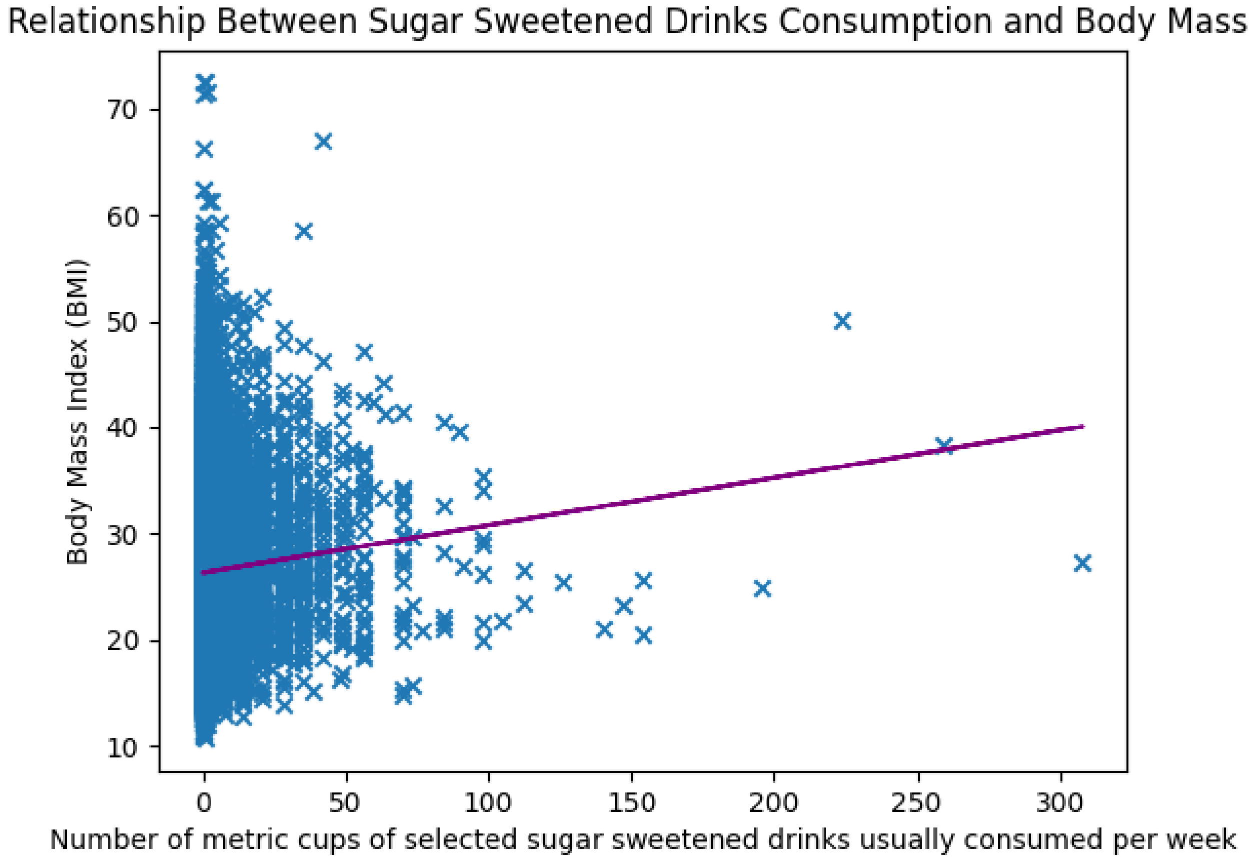
More data points obtained:

- As the data contains a large amount of data points as is (>20,000) and because the boxplots encompass all data points available from the data, it is likely that if more data points were to be added, the chart would remain effective. There may be an increase in outliers, or shifting in the position of the boxplots

B. SDWEEK vs BMI Chart



```
x = raw_data.SDWEEK # x-value is weekly soda consumption
y = raw_data.BMISC # y-value is BMI
stats = linregress(x, y)
m = stats.slope
b = stats.intercept

plt.scatter(x, y, marker = 'x') # plot weekly soda consumption vs BMI
plt.plot(x, m*x+b, color="purple") # add regression line

plt.xlabel("Number of metric cups of selected sugar sweetened drinks usually consumed per week", fontsize=10) # format x-axis label
plt.ylabel("Body Mass Index (BMI)", fontsize=10) # format y-axis label
plt.title("The Relationship Between Sugar Sweetened Drinks Consumption and Body Mass Index") # format title
plt.savefig("sodasperweek_vs_bmi.png")
```



- The BMI of each participant is encoded in the y-position of the graph, and the number of metric cups of sugary drinks consumed by the participant is encoded in the x-position of the graph, each participant is represented by a blue cross
- The line in purple is a regression line, which shows the overall trend of the data

Style of chart: Scatter plot

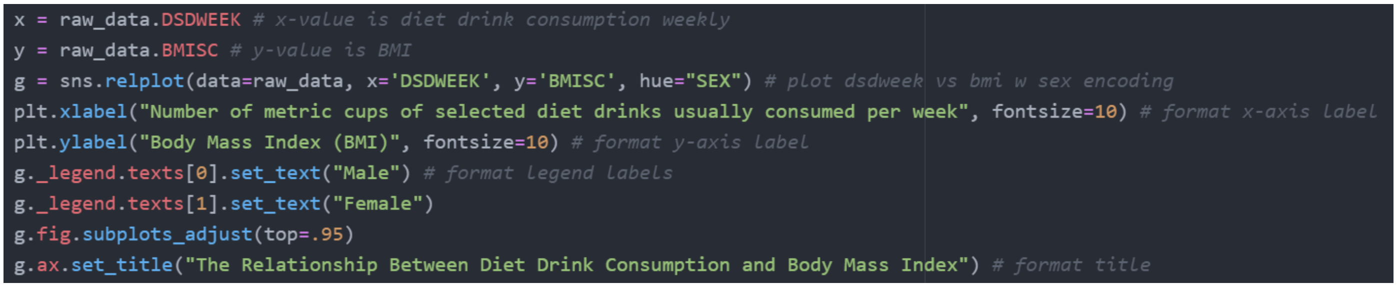
- Scatter plots are effective at showing the relationship between two quantitative variables in a way that bar graphs, line graphs are not because scatter graphs can display all the variables, and also show patterns when the data are taken as a whole

- The two colours chosen are blue and purple to avoid miscomprehension due to red/green colour-blindness. The box-plots above were physically distinct and so, colour-blindness was not an issue that was considered as the colour was not strictly necessary for analysis/did not affect the comprehension of the graph overtly. However, since in the scatterplot, the regression line and the data points physically overlap, they need to be in colours that everyone can see to avoid any confusion.
- Crosses were chosen over circles/dots due to the concentration of data points on the left; upon comparison, the circles formed a blob-like shape while the crosses, still not very distinct, looked slightly easier to distinguish

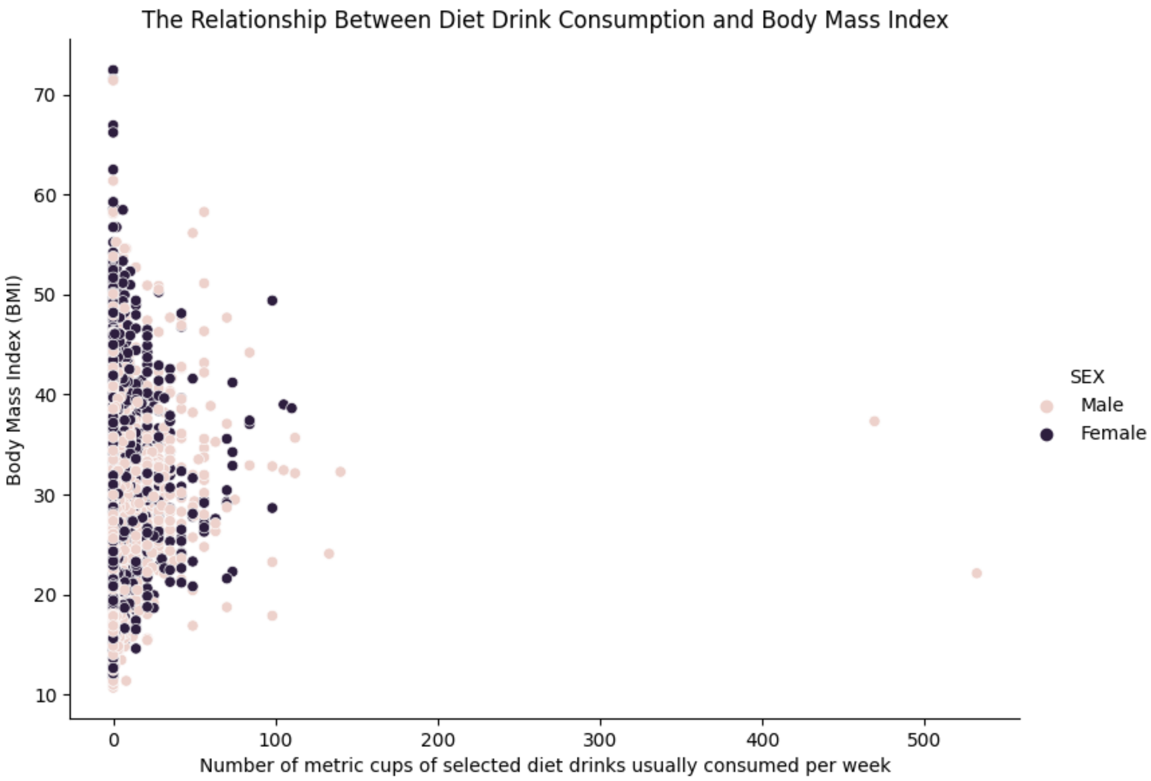
More data points obtained:

- As the data contains a large amount of data points as is (>20,000) and due to the widely applicable nature of scatter plots as they show all data points, it is likely that if more data points were to be added, the chart would remain effective. There would most likely be an increase in concentration at the already concentrated bottom left corner of the graph

C. DSDWEEK vs BMI Chart with SEX encoding



```
x = raw_data.DSDWEEK # x-value is diet drink consumption weekly
y = raw_data.BMISC # y-value is BMI
g = sns.relplot(data=raw_data, x='DSDWEEK', y='BMISC', hue="SEX") # plot dsdweek vs bmi w sex encoding
plt.xlabel("Number of metric cups of selected diet drinks usually consumed per week", fontsize=10) # format x-axis label
plt.ylabel("Body Mass Index (BMI)", fontsize=10) # format y-axis label
g._legend.texts[0].set_text("Male") # format legend labels
g._legend.texts[1].set_text("Female")
g.fig.subplots_adjust(top=.95)
g.ax.set_title("The Relationship Between Diet Drink Consumption and Body Mass Index") # format title
```



- The BMI of each participant is encoded in the y-position of the graph, and the number of metric cups of diet drinks consumed by the participant is encoded in the x-position of the graph, each participant is represented by a pink and purple circle/dot - a pink dot represents a Male participant and a purple dot represents a Female participant

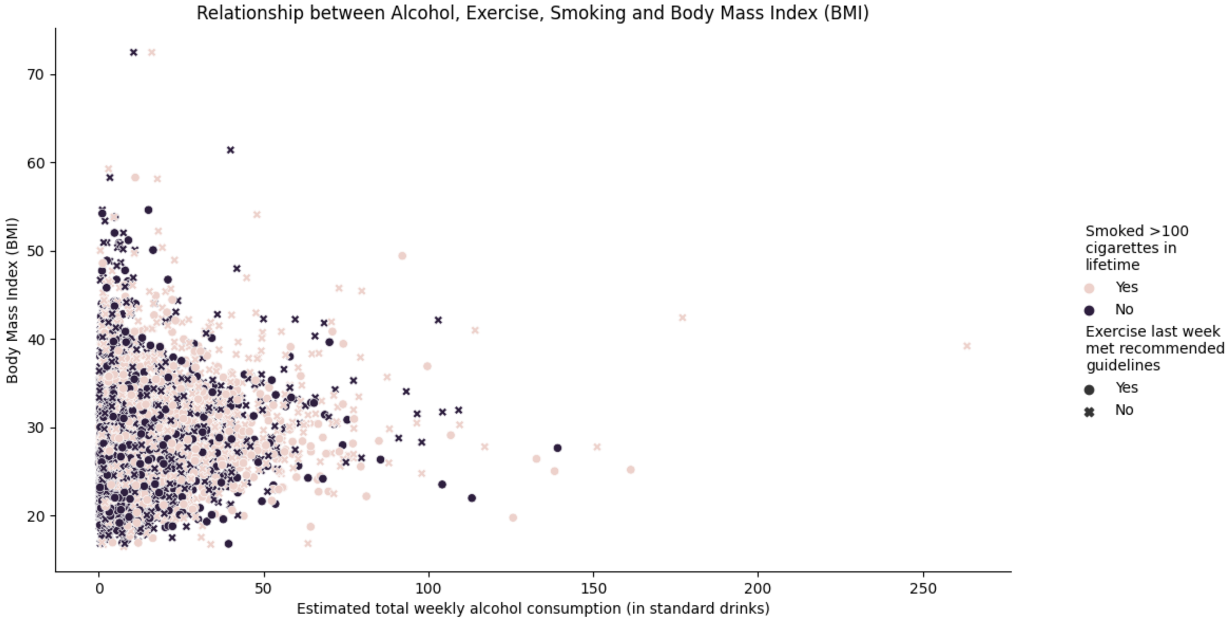
Style of chart: Scatter plot

- A scatter plot was chosen once again, as they are effective at showing the relationship between two quantitative variables in a way that bar graphs, line graphs are not because scatter graphs can display all the variables, and also show patterns when the data are taken as a whole
- The two colours chosen are pink and purple to avoid miscomprehension due to red/green colour-blindness
- Circles were chosen due to the two colours; upon comparison with crosses, it was easier to distinguish the two colours with circular data points

More data points obtained:

- As the data contains a large amount of data points as is (>20,000) and due to the widely applicable nature of scatter plots as they show all data points, it is likely that if more data points were to be added, the chart would remain effective. Once again, there would most likely be an increase in concentration at the already concentrated bottom right corner of the graph.

4-Variable Graph:



The featured graph above features 4 different attributes. On the x-axis is the estimated total weekly alcohol consumption (with a scale of 0-250), on the y axis is the BMI index (with a scale of 0-70), given that both these variables are numerical variables a scatter plot was selected to represent their data points to effectively communicate the relationship between the two variables. Additionally, there are two more variables that are encoded onto the graph visually through markers and colour. First, a binary nominal variable is encoded onto each data point with the pink colour signifying if the respondent has smoked more than 100 cigarettes in their lifetime. Second, is another nominal binary variable that records if the respondent has reached the recommended exercise guidelines in the last week, this variable is encoded onto the graph through the use of data point markers with the dot signifying that the respondent has met the requirement whereas the cross means that they have not. Through the plotting of both the numeric and binary nominal variables, the chart features four unique attributes that describe the relationship between different lifestyle choices on each other but also those variables on BMI.

The chart already starts to struggle in portraying the data effectively with the current amount of data due to the high density of data, with many data points overlapping and covering up others. A higher amount of data would only make this worse and make it even harder to discern a relationship between BMI and the 2 binary variables whereas a trend between alcohol consumption would still work but also end up being less effective.

Conclusion

Throughout this study we sought to find a correlation between BMI and several different lifestyle choices, as well as attempt to find which out of the lifestyle choice had the greatest impact. Each member investigated their own topic and explored their findings below.

Physical activity conclusion

From the data collected within the tables and charts on the impact of physical activity on BMI, one may deduce that there is a weak negative relationship with frequency of physical activity and BMI. Firstly, table 1A would indicate that as BMI increases from the healthy weight to obese classification, the percentage of respondents within that group that exercise at least 5 days a week decreases (seen below):

BMI	Percentage of Respondents Who Perform Physical Activity at least 5 Days of the Last Week (%):
Underweight (BMI<18.5)	29
Healthy Weight (18.5<BMI<25.0)	34
Overweight (25.0<BMI<30.0)	32
Obese (30<BMI)	22
Any BMI	29

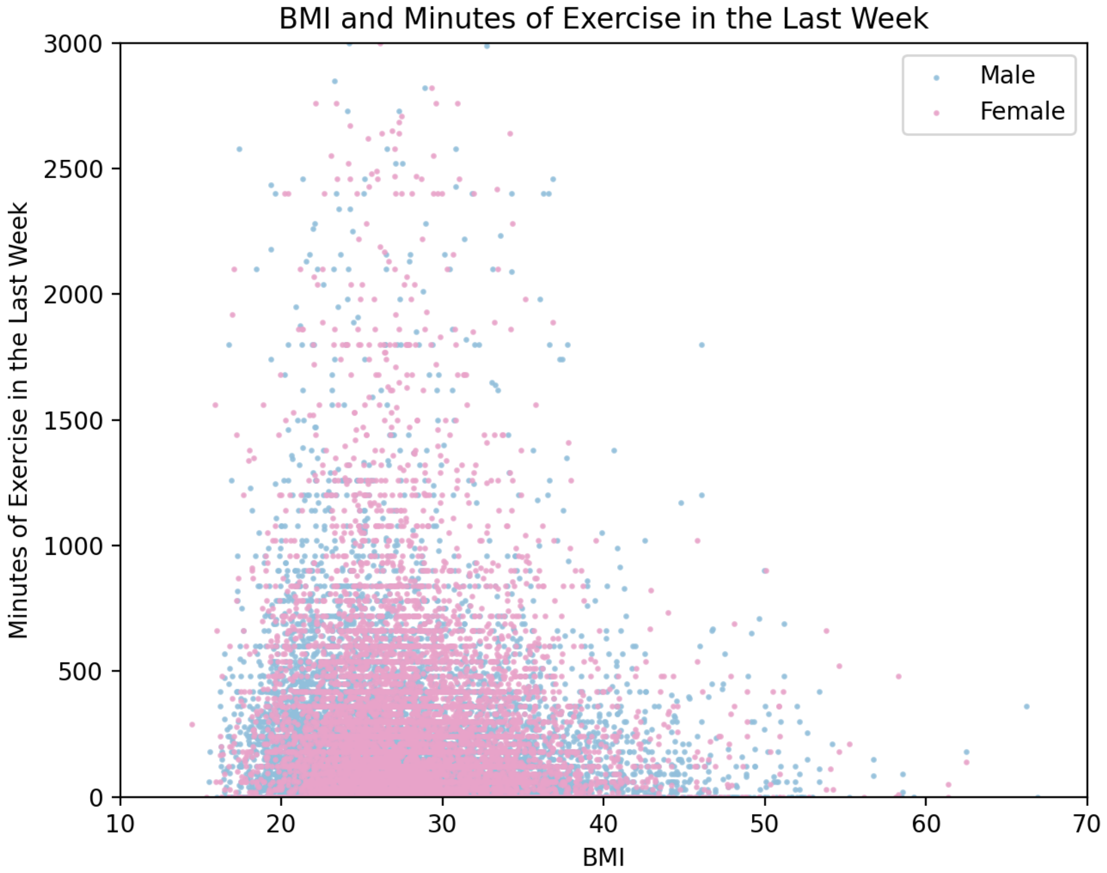
The table seems to indicate the about one third of people classified as a healthy weight exercise at least five times a week whereas, the corresponding obese population have just above one fifth of the sample that exercise over five times a week. Additionally, it seems that out of the entire respondent population 29% exercises more than 5 times a week, therefore, the obese population is far below that proportion while the healthy weight individuals are above that proportion. Thus, this data seems to support the proposition of a negative relationship between BMI and frequency of exercise. Despite this, individuals classified as overweight show similar results to the healthy weight group in table 1A, suggesting that the relationship is more apparent within the extremes of the BMI groups.

Furthermore, the data presented within Table 1B seems to allude to the same conclusion as Table 1A, wherein there are significantly more healthy weight classified individuals that exercise often than obese individuals (shown below).

BMI	Number of Respondents That At Least

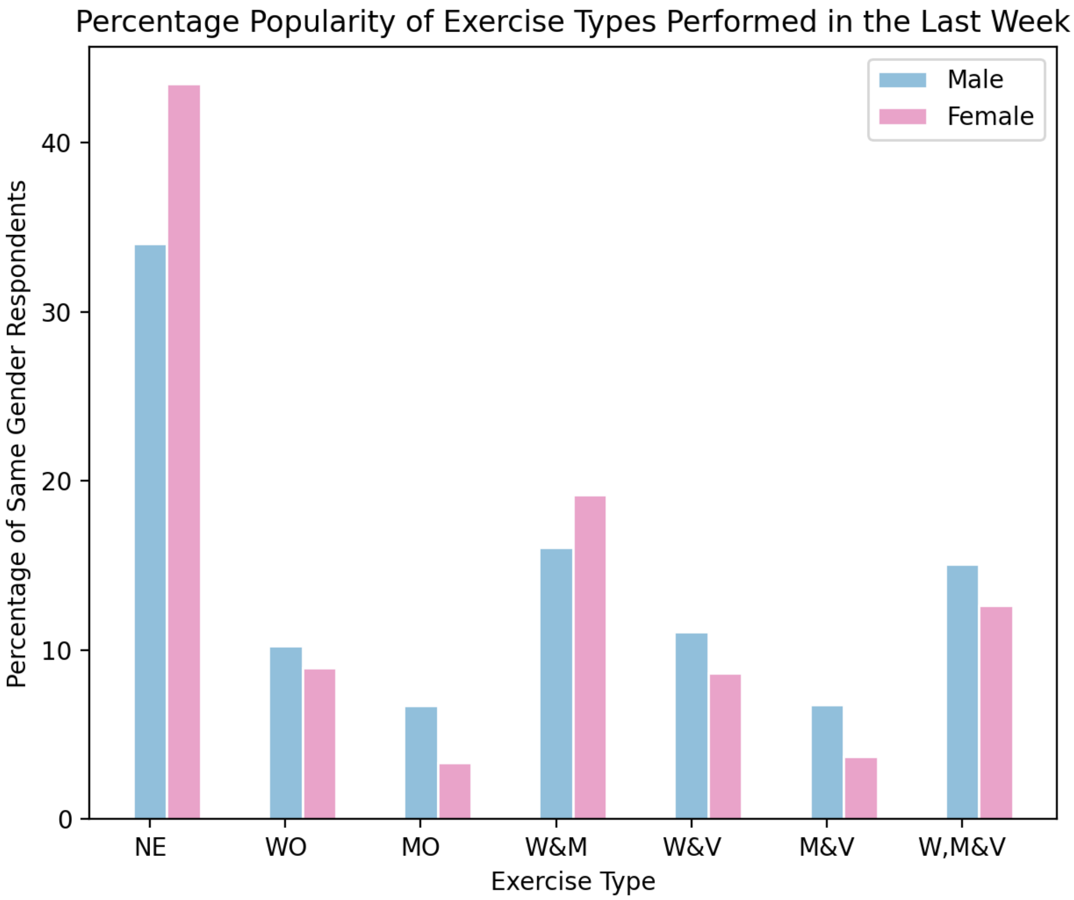
	Walked and Perform Moderate Exercise in the Last 7 Days
Underweight (BMI<18.5)	77
Healthy Weight (18.5<BMI<25.0)	2544
Overweight (25.0<BMI<30.0)	2566
Obese (30<BMI)	1793
Any BMI	6980

On the other hand, both Graphs 1A and 1B seem to be less conclusive than the tables. Within Chart 1A the majority of data points are centered around a BMI from 20-40 and with a large portion of respondents documenting fewer than 1000 minutes of exercise in the last week.



Unlike the tables, Graph 1A doesn't seem to suggest a significant difference between respondents with a BMI from 15-25 vs respondents with a BMI within the 30-40 range on the number of minutes exercised (other than a slight decrease in female respondents). It seems that the similar amount of physical exercise between healthy weight and overweight individuals is also present within the graphs, as the average respondent behaves very similarly while there are a larger number of outliers within the healthy population that exercise far more than the average.

Additionally, this consensus is support with Graph 1B:



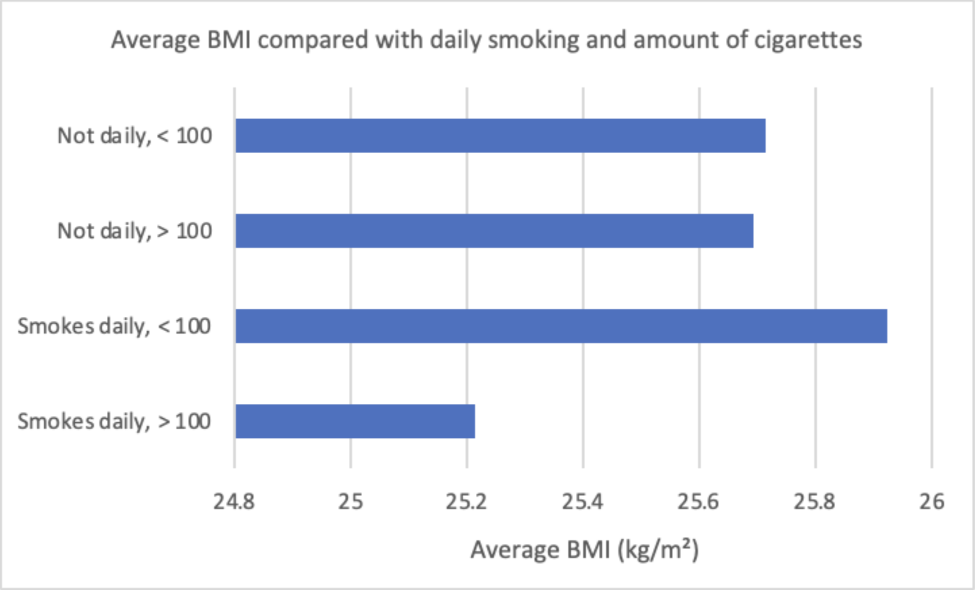
This graph shows that the most common type of exercise for both men and women is reported as “No exercise”.

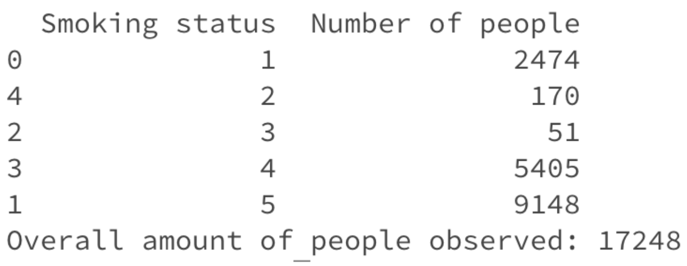
Additionally, there are some significant limitations and uncertainties associated with the data that limit its ability to be fully conclusive about the relationship between BMI and exercise frequency. Firstly, the dataset only included variables that reported respondents' physical activity in the past week, with such a limited range of time, the much more insightful relationship between habitual exercise (over a year for instance) and BMI couldnt be revealed. Not to mention the potential for bias within the respondents' answers. Furthermore, BMI is a health proxy that is dependent on many factors so, isolating the impact or relationship no less between physical exercise and BMI is unlikely to be conclusive unless more stratification of the data is produced to control for potentially confounding variables. More specifically, the data may feature Simpson’s Paradox which hinders the ability of a researcher to draw conclusions and may in fact lead research to conclude incorrectly. Given this, it perhaps would have been better to add another attribute to Graph 1A to control for age, hereditary disease, and/or mental health to discern the relationship between the attributes data. Moreover, within the cleaning process the data values were reduced from 21316 respondents to only 17161. Due to such a large proportion of the data being omitted, the final results may have been subject bias that produce inaccurate conclusions. And, lastly the image of Graph 1A despite an attempt to increase the separation between data points, still increases the difficulty of viewers to discern the presence of blue data points. This, although small detail, limits the quality of the graph.

All in all, despite the limitations of the data and graphs, the relationship between BMI and frequency of physical exercise seems to be somewhat revealed from the summaries and graphs produced in this document. Given the importance and significance of the issue of physical health that the several stakeholders are concerned with, in particular the Australian government, it would be advised that stakeholders continue research that involves collecting high quality data with more variables to be able to better distinguish the explicit relationship between physical exercise and BMI. Though from the limited conclusivity of the results, stakeholders may be intrigued to know that the majority of Australians within the sample still do not exercise at a rigorous level and at the recommended frequency of 5 days a week. And so, this may very well be contributing to the large presence of obese individuals within the sample.

Smoking conclusion

As for the smoking habits of individuals, there were no positive correlations between an individual's BMI and the amount of cigarettes smoked. Surveyed individuals with their daily smoking status and whether they used greater 100 cigarettes were compared alongside the average BMI in each category. As a result, all categories (smoked daily/non-daily and greater/less than 100 cigarettes) resulted with similar average BMI values whose differences were not significant enough to conclude any correlation. As seen in the chart below, the category with the highest average BMI were individuals who smoked daily and had less than 100 cigarettes. However, any potential conclusions with BMI and amount of cigarettes smoked was not possible as the category with the lowest average BMI were found to be individuals who smoked daily but had more than 100 cigarettes.



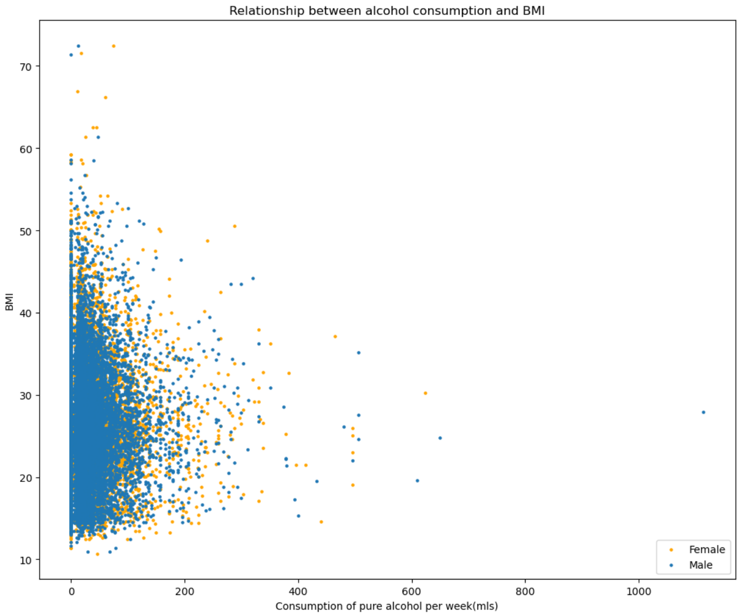


```
  Smoking status  Number of people
0              1              2474
4              2               170
2              3                51
3              4              5405
1              5              9148
Overall amount of_people observed: 17248
```

With the table above, it is made evident that among the 4 smoking categories not including those who are non-smokers (daily, weekly, non-weekly and ex-smokers) most surveyed individuals were ex-smokers, indicating that there is a higher ratio of those who quit smoking to those who still currently smoke. This could be indicative of people opting not to smoke as compared to picking up the habit in hopes of maintaining a healthy lifestyle.

Alcohol conclusion

With my current data it is very difficult to make a conclusion about any correlation between BMI and alcohol consumption. The overarching comparison between BMI and alcohol consumption can be seen in the following graph:



While initially there is an increase in BMI when alcohol consumption increases, it flattens out and starts to decrease as BMI increases above 30. There are several variables that could be responsible for this such as the distribution of the data. As the majority of people are going to be between a BMI of 20 to around 30,with the average BMI of my dataset being around 26.4(Table 3b), we naturally have a greater deviation in alcohol consumption as there is simply more data for that range.

Another possibility is that it is very difficult to maintain a BMI of >30 by high alcohol consumption as that would require the person to drink very high amounts of alcohol and spend a significant amount of their day drunk which is impossible for your everyday person. This also explains why people with a very high BMI don't drink much alcohol, as their main source of calories has to come from food.

This is also backup up by the following sections of chart 3a:

BMI Group (BMI // 5)	Favoured Alcohol Type	Times consumed
2	Full Strength beer	94
3	Full Strength beer	568
4	Full Strength beer	1086
5	Full Strength beer	1267
6	No Alcohol	673
7	Full Strength beer	284
8	Full Strength beer	92
All(no grouped aggregate)	Full Strength beer	4048

With most BMI groups preferring Full strength beer and the overall favorite also being Full strength beer, it would become increasingly difficult to maintain a high weight through high beer consumption due to cost, liver damage, intoxication, and a high volume of liquid consumption. Interestingly, those with a BMI of 30-34 preferred to drink no alcohol but as shown by the graph shown earlier, there are still a lot of people who did drink in that bmi group, so it is quite hard to conclude anything from this anomaly.

Lower end of chart 3a:

9	No Alcohol	41
10	Fortified Wine	15
10	No alcohol	15
10	Low alcohol wine	15
11	Low alcohol wine	6
11	Fortified wine	6
11	No Alcohol	6

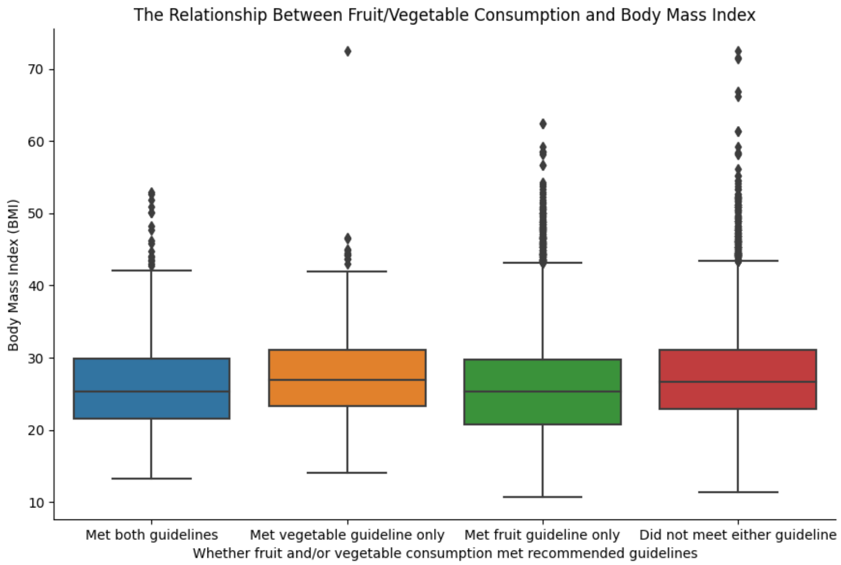
Interestly for the lower end of the chart pertaining to BMIs of 45-55, lower alcohol wine and no alcohol were the most common out of all the groups. As mentioned before, this could be the result of the fact that it is much harder to maintain a high BMI with alcohol rather than solid food, and the more alcohol you drink, the less food you can eat.

Overall there does seem to be some positive correlation between alcohol consumption and an increase in BMI, but it does not seem to be a leading factor of high BMI, with everyday factors and bodily limits restricting alcohol being one of the main causes of a constantly increasing BMI. Despite there being some possible correlation further investigation would have to be performed to confirm it, as there were no immediate trends found in any table or chart and many possible variables that could affect BMI were not investigated. Therefore no solid correlation can be confirmed between alcohol consumption and BMI.

Diet conclusion

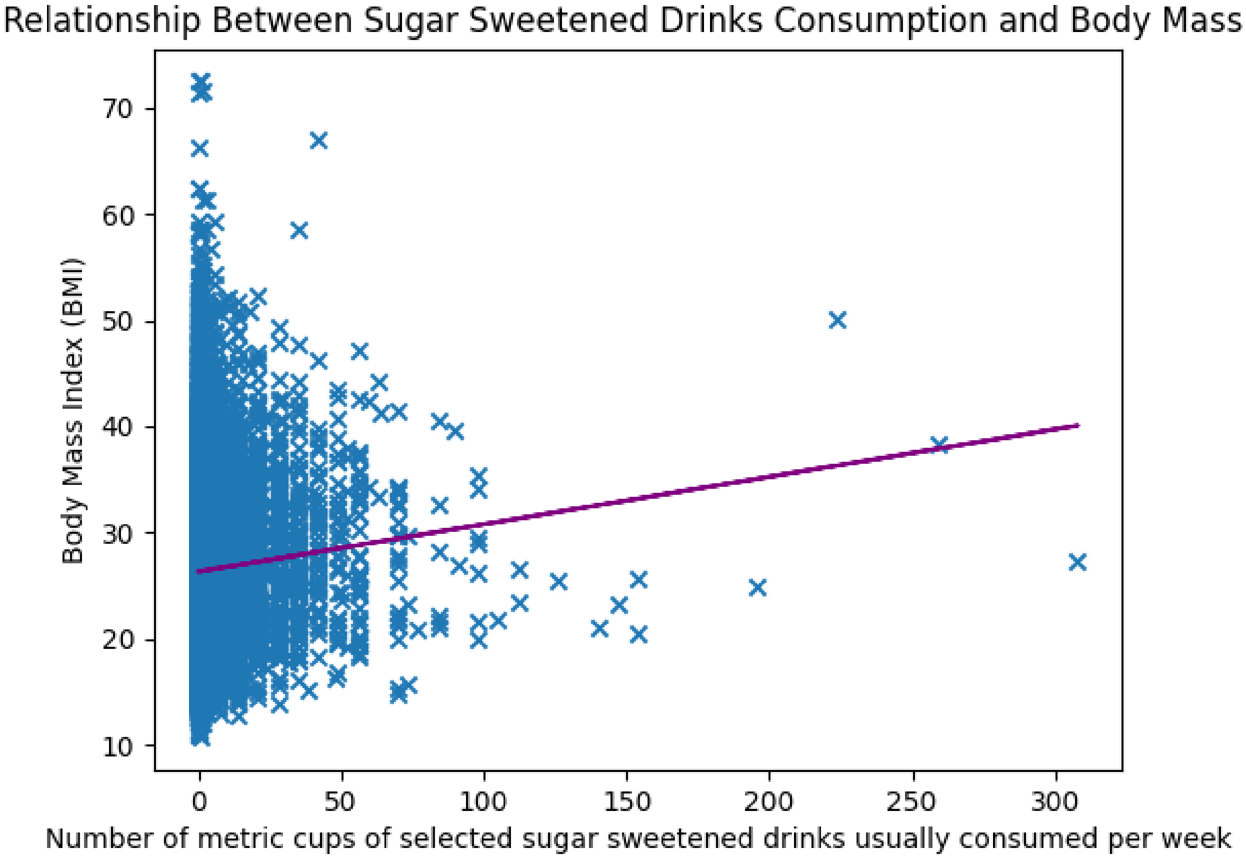
There is no clear evidence that simply meeting dietary requirements (incidentally or by design) has a clear effect on BMI, as the BMI means (as can be seen in the table) and the boxplot ranges (as can be seen on the graph) do not vary largely between each level in RDI2013. Unexpectedly, the table indicates that the demographic with the lowest mean BMI are the participants who only met the fruit consumption guideline and not the vegetable guidelines. However, the graph does indicate that the demographics who met both guidelines and met the vegetable guideline only had similar maximum BMI values, lower than the groups who met the fruit guideline only or did not meet either guideline. The groups who met the fruit guideline only or didn’t meet either guideline had lower minimum BMIs, which may represent the underweight people who undereat, have eating disorders, or simply do not meet their minimum energy intake.

RDI2013	BMISC Mean
1	25.942592
2	27.776511
3	25.635973
4	27.353401
Overall Aggregate	26.424420



Pertinent to this conclusion, the data is limited as it does not specify the level of consumption or nutritional intake, and only shows that a certain quota has been met. To gain a deeper understanding of how fruit and vegetable consumption affects BMI, more specific information about what and how much the participant has consumed would be necessary.

However, there is evidence of a moderate positive correlation between higher sweetened drink consumption and higher BMI values, as can be seen in this graph where the regression line has a positive gradient.



However, the data examined in this report is only limited to a single snapshot in time, and more research or interviews need to be continuously obtained over a period of time in order to establish if there is an element of causation between sweetened drink consumption and a higher BMI.

Overall conclusion:

Despite the analysis investigating multiple different possible areas that may significantly affect BMI, the results found within this document are quite similar. Almost all areas of investigation found some weak relationships between the variable of interest and BMI. For instance the increase in sugar sweetened drink intake, cigarettes, and alcohol all have a weak positive relationship with BMI, while increased frequency of physical exercise has a weak negative relationship with BMI. These findings are all in accordance with the summaries and charts produced earlier in the document. Given these results it may suggest to the Australian Government particular points of interest for health legislation in order to ensure optimal physical wellbeing for the Australian population. Furthermore, other interesting insights such as the persisting low level of physical exercise, significant portion of the population still not meeting dietary standards, and the continued use of cigarettes may suggest possible corporate opportunities for companies to use to enact positive social change for better physical health of their customers. Or in fact to the Australian public at large, the data reveals insights into behaviours that they should avoid in order to decrease their chances of increasing their BMI to an unhealthy level.

Despite these valuable results, there too are significant limitations within the data set used and summaries and graphs produced. These limitations include significant omission of data through the cleaning process, inability to distinguish a conclusive relationship between the lifestyle and BMI, inability to control for the large amount of confounding variables associated with BMI and the recorded variables only accounting for a limited time frame of respondents' habits. Thus, to dispute the results produced, the limitations of the analysis limit the ability to make sound conclusions from the data. Therefore, it should be stated that in order for the stakeholders to better address the physical wellbeing of the Australian public more data and analysis will need to be created to increase the conclusivity of the results.