CPM 252 FOUNDATIONS AND PROGRAMMING FOR DATA ANALYTICS

ASSIGNMENT 1

NAME: SANTTOSH A/L MUNIYANDY

MATRIC NUMBER: 159193

LECTURER IN CHARGE: MISS NAJIHAH IBRAHIM

1 DECEMBER 2023

SEMESTER 1, 2023/2024

CPM252

TABLE OF CONTENTS

LIST OF FIGURE CAPTIONS

ABSTRACTS

This report presents visualizations and its analysis about stroke data which unpacks multiple characteristics and sets over many patients. By using the stroke dataset given, I utilize various data visualization method to obtain insights and trends. The objective of this report was to identify factors that will lead to having stroke. The data was initially visualized through pie chart regarding genders to find the correlation of genders and stroke. Nevertheless, both showed little to negligible differences. Then a bar chart relating frequency of stroke to work type which implicate that self-employed workers are the highest. To find a reason behind this, another bar chart relating frequency of stroke against type of stress was plotted, which backfired with illogical response of moderate stress being the highest. Another effort of plotting line chart relating frequency of stress against types of activity which shows that high activity the most among the three. This does not provide the backbone reasoning, thus a scatter plot relating frequency of strokes against family history stroke was plotted. This in particular gave a reasonable reason to account as an factor to stroke. Next, two lines chart of frequency of stroke against types of heart disease and hypertension was plotted. It was found that the graph of heart disease provided a good reason to add into the factor of stroke. To give support to heart disease line chart, a grouped bar chart was plotted about stroke against smoking habits and alcohol intake. Here, it was concluded that smoking habits affect the heart disease more than alcohol intake. Next the smoking habits was taken account for people with heart disease against frequency of stroke in form of bar chart. It was implicated that all almost the same thus, the smoking habits were taken accounted for as a factor to stroke. Relationship of strokes against BMI and dietary habits was plotted similarly in grouped bar chart. It was analysed that obese people are significantly high in strokes among other groups. This was related with heart disease in line chart which proves that people with obese or high BMI have high amount of heart diseases. Three box plots of three highest diets in obese BMI was created to shows the distributions of average people glucose level. It was shown as the Paleo contains more average people having higher glucose level. This proves as BMI and dietary habits play their part in stroke. Finally, a pie chart of symptoms was created to observe which symptoms highly brings to stroke.

1 INTRODUCTION

1.1 Data

In the current era, the abundance and overnumbered of data becomes a great key information in various domains such as education, business, economics, healthcare, engineering and almost every aspect that runs that capability of modern world. Data is a collection of information gathered using various methods such as observations, measurements, research and analysis. Data consists of numbers, names, pictures, figures, time and much more that can be stored for an important use [https://byjus.com/maths/introduction-to-data/]. The number of data accumulated had over-exceed the limitations and became a challenge for data scientists and data analytics to understand and make a proper decision for respective reasons. There are two types of data which are qualitative and quantitative. Qualitative describes the information in words, simply said it is such as describing a quality of a shirt, religion of a person and type of shoes. Quantitative describes the numerical information such as the weight of the person, price of a shoe and telephone number of any person [https://byjus.com/maths/introduction-to-data].

quantities
numbers
information
What Is Data?
facts
graphs
observations
measurement

Figure 1.1 shows the source of data taken

Since data comes in many forms and formats, it can be in structured or unstructured with each respective unique characteristic. Most of the datasets that I will be frequently mentioning in this report would be structured data. Structured data is categorized as quantitative data and usually had been fit into relational databases and spreadsheets which will be easier to work with [https://www.g2.com/articles/structured-vs-unstructured-data]. The reason for structured data is highly used in industries is because it is highly organized and easily be understood by machine learning. The data produced or organized will be higher quality and consistent which reduces the encounter with error of compromising

the data. Unstructured data is categorized as qualitative data which is difficult to analyse as the data is non-relational or in other words it does not have a proper categorization to group in. As mentioned, the qualitative data could be name of a restaurant, file names and much more. However, it does not bring to a point where unstructured data is difficult and hard to work with, this type of data can be used as a technique to learn a certain trend about a certain product. Typically, businessman frequently use these data to upgrade their product against their challengers.
[https://www.g2.com/articles/structured-vs-unstructured-data]

1.2 Data Visualization

As we venture about the basics of data, we need to know what are the ways to communicate, derive and understand data to obtain useful information and insights, and how are we going to achieve this? We transform these raw data into visually formatted figure where it represents the complex data visually which is commonly known as 'Data Visualization'. Data Visualization is the practice of translating information into a visual context to help humans to obtain insights and information by identifying patterns, trends and outliers in the data sets
[https://www.techtarget.com/searchbusinessanalytics/definition/data-visualization].
Data visualization is an efficiently made bridge for acting as a catalyst for modern technological advancements by making quick and accurate decision making. In the early times during the war, Charles Minard had mapped the Napoleon's invasion of Russia by exploring the size of the army that retreated from Moscow relating to the temperature and time to understand depths of the event [https://www.sas.com/en_my/insights/big-data/data-visualization.html].

It is no doubt that we have uncovered the enhanced decision making, obtaining trends and scientific findings from the exploring the data visualization, thus let's give some examples to further understand how is it done in the modern career to help with various domains such as economics, businesses, education systems, and etc. Spotify, one of the most popular world leading music platform, uses data visualization to always engages with each of their customers experiences. Spotify learns the listening habits of each of their customers by visual representation which then they obtain the listening habits and insights. This key information is then used as trends for the respective customers or the Spotify user to identify their similar tastes of music, top listened artists or music, and etc. There is also a feature in Spotify known as 'Discover Weekly' where it also similarly to identify the patterns and the trends of the user listening genres, songs or preferences which then is aligned into visualization so that Spotify can recommend similar song selections to the respective user through that in-application feature[https://medium.com/@shrunalisalian97/spotify-data-visualization-4c878c8114e]. In the parts of the business sector such as Starbucks also uses data visualization to keep fighting with its contenders. Using the data visualization, it helps Starbucks by keeping on their feet to always update their menu as it analyzes the sales of every item on the menu and understands which of the products is popular and worst among the customers. For example, according to a study, 43% of tea drinkers avoid adding sugar. Starbucks created a new product line of unsweetened iced teas to cater to this market. Also, after discovering that 25% of consumers do not add milk to their coffee, the company launched a new line of black iced coffee without milk [https://www.linkedin.com/pulse/how-starbucks-uses-big-data-levon-hovsepyan-7bprf/].

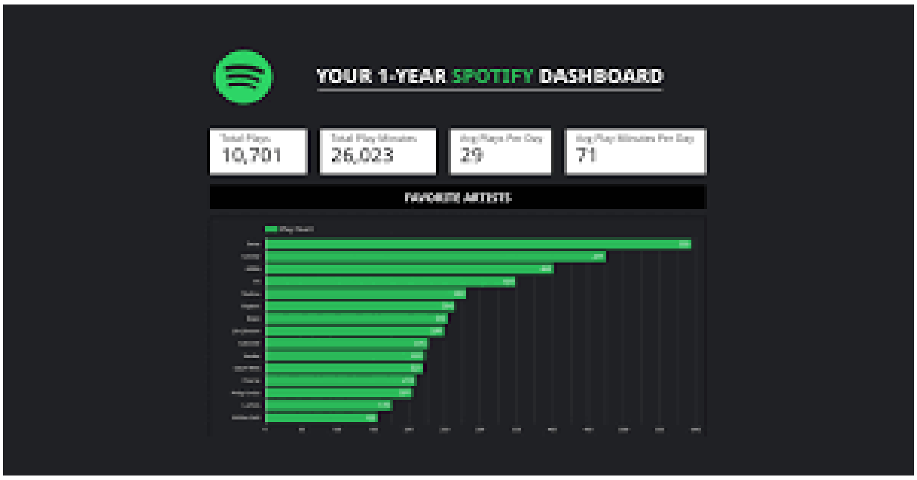


Figure 1.2 shows Spotify Wrapped (feature in Spotify) for each user

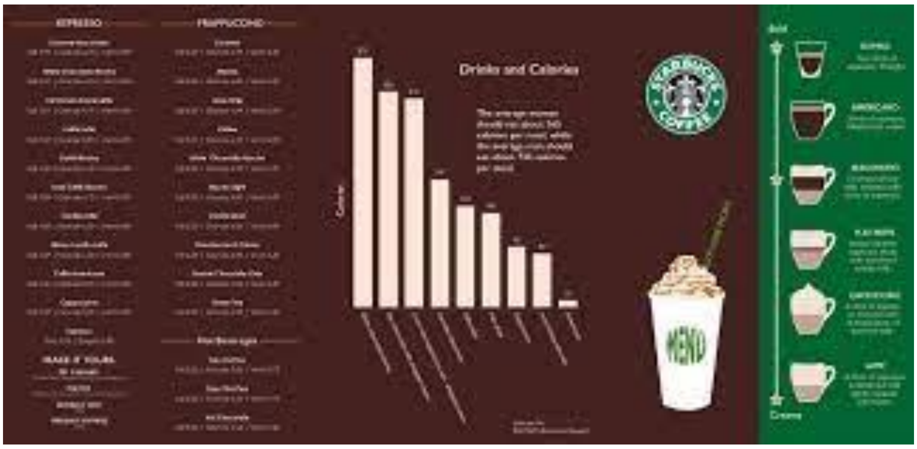


Figure 1.3 shows Starbucks choosing the latest trends for their menu

1.3 Types of Data Visualization

As we have learned of how the data visualization is vastly used in numerous sectors of the world, we will see some common types of data visualization.

1) Bar Charts:
   - Bar chart can either horizontal or vertical but the key point is the values is measured through the length/ height of the bar (the greater the height, the greater the value).

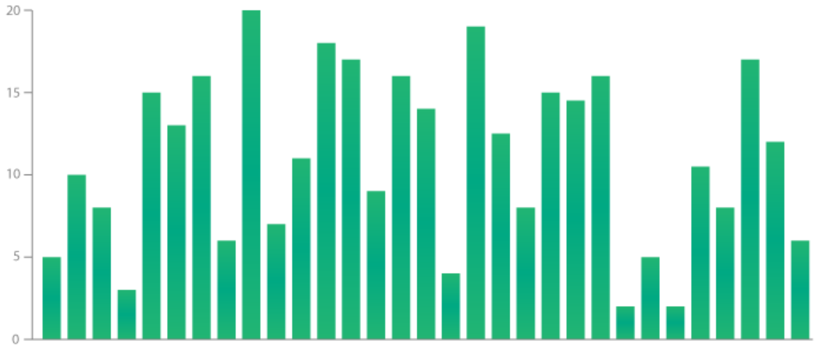


Figure 1.4 shows an example of bar chart

   - Bar charts represents categorical variables, discrete variables or continuous variables which are grouped and the height shows the frequencies of the variables[https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch9/bargraph-diagrammeabarres/5214818-eng.htm].

2) Pie Charts:
   - Pie charts gives the user or the reader a quick idea just by the glance of the visual about the proportional distribution of the data. Pie charts uses percentage distribution for ease of an eye to differentiate easily the different segments of categories in the dataset. Each percentage represent a proportion of each category which when the sum of full circle (sum of all data) equals to 100%[https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch9/bargraph-diagrammeabarres/5214818-eng.htm].

Figure 1.5 shows an example of pie charts

3) Line Chart:

- Line chart is easy to create and also very commonly used in various domains frequently as it is very easy to reveal trends of how data has changed over time. Line charts are comparison of two variables on two x and y axes. The x axis usually is a timescale or sequence of intervals while y axis is a quantitative value. The way to read line charts are easy as the we only need to observe the direction of the lines in which the upward slope means values are increasing while downward slope meaning value are decreasing.
- [https://datavizcatalogue.com/methods/line_graph.html]

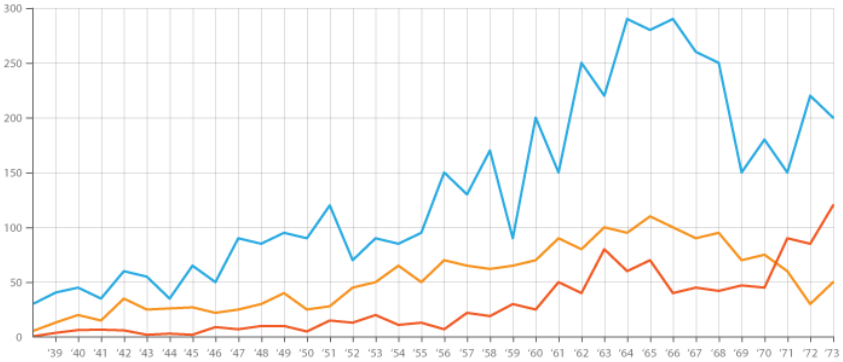


Figure 1.6 shows an example of line chart

4) Scatter plot:

- Scatterplots shows the relationship between the two variables by observing the most density placements of the dots. We can detect correlation between these two variables. Usually, the data points or the dots will not be joined as a line as we more into observing the strength of two variables.
- [https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch9/scatter-nuages/5214827-eng.htm]

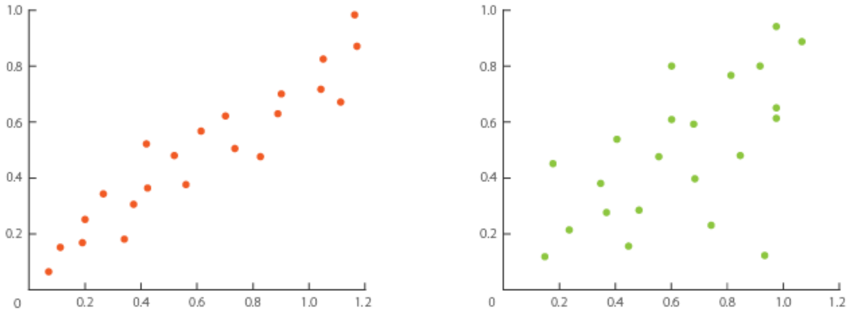


Figure 1.7 shows an example of scatterplot

5) Box and Whisker Plot:

- In other word it is known as ‘boxplot’ where it shows the spread and centers of data set. It is quite convenient to read and display data distribution through the quartiles to see which data is in upper, median or lower quartiles to understand better. To read the box plot can get quite complicated, thus I will try to simplify it. The main part of the box shows the middle portion of the data which known as interquartile range. The top and bottom of the box are known as third quartile (75%) and first quartile (25%) respectively. The centre of the bar contains a line where it represents the median of the data. The lines after the box are whisker where the line below is for minimum value while above the box represent the maximum value [https://www.statisticshowto.com/probability-and-statistics/descriptive-statistics/box-plot/].

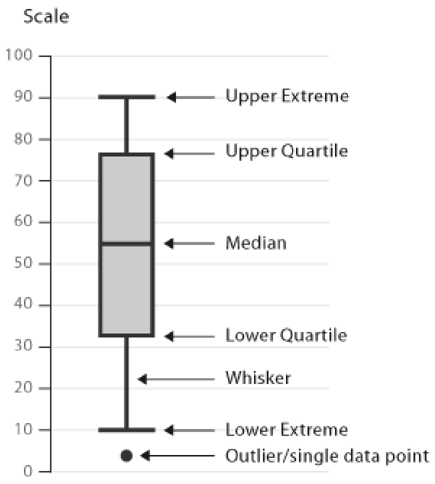


Figure 1.8 shows an example of boxplot

1.4 Objective

There are two main objectives to this report:

PART 1: Choosing the data visualization method:

- By choosing the two out of eight given options of the Python Data Visualization Libraries, we have to compare and highlights the strengths with the weakness of those two respectively. By discussing the ups and the downs of the chosen libraries, we will choose only one of the libraries to visualize the dataset in PART 2.
- I have chosen Matplotlib and Seaborn

PART 2: Data Visualization with given Dataset:

- My dataset that had been given is Stroke dataset. With this my goal is the find' types of people and their lifestyle that most likely to suffer stroke'.
- a) First Objective: to find the relation of external factors such as gender and work type with strokes
- b) Second Objective: to find the relation of major diseases such as heart disease and hypertension with results of stroke
- c) Third Objective: to find the most symptoms among the people that will lead to stroke
- I will use different type of visualizations to differentiate the datasets to enquire the data to my favors. The steps of visualization and the reason of the goal will explained.

2 PART 1: DATA VISUALIZATION METHOD

2.1 Introduction of Matplotlib

Matplotlib is a python library for data visualization for python programming and used by millions of people frequently for work related domains. In the year of 2003, Matplotlib was written by the late John D. Hunter who passed away in August 2012, which then was taken by Thomas Caswell [https://en.wikipedia.org/wiki/Matplotlib]. Initially, John D. Hunter was inspired by the limitations of MATLAB to create Matplotlib. In the early days, John D. Hunter used MATLAB for data analysis and visualizations, in which he praises the beauty of MATLAB to plot visuals with such ease. However, he noticed that as the application grew in complexity to interact with the databases, he began to notice the limitations of MATLAB to manipulate with complex data structures. This is the point where John D. Hunter had decided to take upon his own hands to create a new application similarly to MATLAB but with better plotting capabilities [https://matplotlib.org/stable/users/project/history.html].

2.2 Highlights of Matplotlib

Strengths:

1) Provides a simple way to access a large dataset:
   - Matplotlib has the ability to visualize large datasets by efficiently displaying various types of plots to interpret large volumes of data. Matplotlib also uses libraries such as ‘NumPy’ and ’Pandas’ to handle large datasets and even uses these two libraries that contains many features to aid with processing the datasets.
2) Excellent tool for beginner:
   - Matplotlib contains large amount of documentation that contains numerous of examples and tutorials which act as learning guide for those who are new and interested in using matplotlib for the first choice of data visualization.
3) Advanced Customization:
   - Matplotlib is known to have a complex customization, however with such various customizations, there are elements to add into our visualizations which can give more in depth of what we are going to interpret.

4) Various Types of Plots:

- Matplotlib contains various types of plots ranging from basic line plots to complex 3D visuals to aid the user in having many options. The diversity of plots which with respective functions of displaying trends, patterns and correlations gives the user many ideas and freedom to create visualizations according to their creativity and complexity.

5) Opensource platform:

Matplotlib is an open-source platform that requires no paid license in any form. This is perfect for not only work-heads but also for the students that are having the interests to pursue the career into Data Analytics or career relating with data visualization. Matplotlib can the first amazing choice for the people to experiment and learn many ways to visualize.

[https://www.testgorilla.com/blog/matplotlib-in-python/]

Weakness:

1) Limited Language:

- Matplotlib only requires python language in order to visual. While python can be the easier language to master, it would be convenient to have different languages to compute as well. Python also proven to be slower to compiled languages such as C++ and Java.

2) Overwhelming documentations:

- The documentation can be helpful for the newcomers however, the documentation might be difficult to understand the minute details or to find a specific task. The documentation had been prepared as it mainly done for those with the knowledge of data visualization. Thus, it might take a long time for the newcomers to learn their way in Matplotlib.

3) Steep learning curve:

- Similarly, to previous factor, Matplotlib has way too much freedom in terms of their advance customization. While it is quite better to start simple, it norm to take to next level in terms of visualizing and that can be quite challenging if we are using Matplotlib. Various features and customizations might make great confusion among the users.

4) Limited beauty:

- Matplotlib has very limited aesthetics which appeals as lesser visually striking to the eyes of higher standard people working. This might cause for those with high ranking and knowledge of using data visualization to not choose Matplotlib as it can quite basic and causing them to resort other modern libraries such as Bokeh, Seaborn and etc.

5) Limited exploring tools:

- Matplotlib provides only basic built-in data exploring tools which limits the users to work with the datasets.

[https://www.quora.com/What-is-the-strength-and-weakness-of-the-Python-programming-language]

[https://www.geeksforgeeks.org/difference-between-matplotlib-vs-seaborn/]

2.3 Introduction of Seaborn

Seaborn is a python visualization library which provides high level visuals and interface for statistical data visualization. Seaborn contains various beautiful eye striking and amazing color palettes to make the plots more visually striking. Seaborn was made by Micheal Waskom who started this project back in 2012. Similar to the history of Matplotlib, Micheal Waskom witness the over-extensive of Matplotlib and swore to enhance and also simplify the visual capabilities based on Matplotlib. He was more focused on the aesthetics and the lowering the difficulty for the people to use Seaborn. After the development of Seaborn, it was mainly focused upon statistical visuals with gain the domain of data science.

[https://seaborn.pydata.org/tutorial/introduction.html]

[https://mwaskom.github.io/]

[https://scholar.google.com/citations?user=EIPxNwUAAAAJ&hl=en]

2.4 Highlights of Seaborn

Strength:

1) Various statistical visualization:
   - Seaborn offers numerous statistical visualization that aids with the exploration of the relationships with variables in datasets. With this, it helps out the user to obtain information needed effortlessly.
2) Data Manipulation:
   - Similarly, to Matplotlib, we can use NumPy and Pandas to manipulate the datasets which can be very efficient to data exploration.
3) Easy to plot:
   - Seaborn simplifies the coding to visualize. For example, 'boxplot ()' is all needed to create a boxplot without overextensive coding lines while also maintains the simplicity and aesthetics. Thus, Seaborn can be very straightforward and easy to grasp for the beginners. Even with a high-level library like Seaborn, new users does not require to have the fear of having the knowledge of data visualization to get started.
4) Very visually aesthetic:
   - Seaborn's main key features is that its visually aesthetics to make the visualization very beautiful. Seaborn provides multiple color palettes and features to enhance the plots without altering the main objective of the plot. Most high-ranking users of data visualization uses Seaborn.
5) Rich documentation:
   - Seaborn is packed with many documentations to guide any levels of users to comprehend any explanations needed.

[https://www.datascienceverse.com/data-visualization/seaborn-library-python-perks-of-using-seaborn/]
[https://www.linkedin.com/advice/3/what-advantages-disadvantages-using-seaborn]
[https://www.educba.com/seaborn/]

Weakness:

1) Primarily focused on statistical plot:
   - Seaborn primarily focused only on statistical plots which limits the user to plot something out boundary of statistics. While seaborn can be very beautiful but the very limited plots can cause the users to turn away to different libraries almost immediately.
2) Slow and memory- intensive:
   - If working with large complex datasets, Seaborn takes longer execution times as it uses the Matplotlib backend. This wastes much time that can be used in ways that can be used instead to customize.
3) Interactions are limited:
   - While seaborn are visually expended, the interactive elements that can be added to help with the visualizations are quite limited.
4) Various complex plots:
   - Seaborn does provide numerous statistical plots, however that many amounts of plots it can great headache for the user to even understand certain uses of the plots that can be used of their desire. They might take longer time to understand each plot.
5) Fully relied on Matplotlib:

   -Seaborn is built upon backend of Matplotlib, thus most functionalities are used from the knowledge of Matplotlib. The complexity of Seaborn can cause the user to go back forth from Seaborn to Matplotlib to understand better. This can cause great confusion among the users.

   [https://michaelwaskom.medium.com/three-common-seaborn-difficulties-10fdd0cc2a8b]

   [https://www.linkedin.com/advice/3/what-advantages-disadvantages-using-seaborn#:~:text=2%20Disadvantages%20of%20seaborn,-Seaborn%20is%20not&text=It%20can%20be%20slow%20and,optimized%20for%20performance%20or%20scalability.]

2.5 Comparison of Matplotlib and Seaborn

Matplotlib	Differences	Seaborn
-Matplotlib contains great amount of documentations that can be helpful, however the syntax for Matplotlib can be overwhelming for the beginners	Learning Stages	-Seaborn also contains various documentations and tutorials, but it is very beginner friendly despite its looks of amazing visuals. It requires very simple syntax code to visualize.
-Matplotlib have various customizations options to tweak and experiment with the data. -The options to customize is too broad if we wish to step to next level which overwhelms us with entirety of new stuffs.	Customization	-Seaborn also provides various customizations options to enhance the plots -Seaborn does not overwhelm us too much as it is kept in simplicity if we wish to expend our plot knowledge
-Matplotlib provides various types of plots ranging from basic line plots to 3D plots which gives the user the freedom to express that data in many ways	Plot Limitations	-Seaborn is limited only to statistical plots which does not favor for the users that wish to combine with other type of plots
-Matplotlib is a very basic in terms of its visual and aesthetic which makes it visual plot very bland	Aesthetics	-Seaborn is very creative with the way we can customize the plots in many ways and add many features such as themes and color palettes to make the plot visually striking
-Matplotlib has a lot plots that requires little of knowledge to understand with ease. This gives the users to understand different type of plots and freedom to use them with ease	Complexity of Plots	-Seaborn has a lot plots related to statistical, however they can get complicated and complex to understand which can cause the user to take longer to understand if they wish to use another type of plot

[https://blog.consoleflare.com/matplotlib-vs-seaborn/]
[https://codesolid.com/matplotlib-vs-seaborn/#google_vignette]
[https://ritza.co/articles/matplotlib-vs-seaborn-vs-plotly-vs-MATLAB-vs-ggplot2-vs-pandas/]
[https://www.quora.com/What-are-the-benefits-of-using-Seaborn-over-matplotlib-pandas-for-plotting-data-frames-in-Python]

2.6 Discussion between Matplotlib and Seaborn

After scoping through the strengths, weakness and even comparing the two data visualization libraries, Matplotlib and Seaborn, it is safe to say that both libraries have their own stronghold according to the situation. Matplotlib is amazing as it have many customization, types of plots and flexibility. It also has amazing guides to help any level of user out, however if the beginner user does need more time to understand fully. On the other hand, Seaborn is amazing as well to the extent it has many customizations, aesthetics and even beginner friendly. However, it very limited to statistical plots and does not give much freedom to the user. To understand better, we shall which type of cases works better for the respective Matplotlib and Seaborn.

1) Task requires in-depth and complex visualizations:
   - Certain datasets require complex visualizations to understand better the objectives. Sometimes it needed a 3D or a multi panel figures plotted. For this its better to use Matplotlib as it has all the freedom to customize in-depth it can get.
2) Task requires visually striking plots:
   - Visually striking or aesthetic plots can be an eye pleaser when presenting to those with or without knowledge of data visualization or even makes things easier to interpret and obtaining the information needed. Seaborn is the best choice as it has various ways to make its plots beautiful with its aesthetic settings with simple syntax as well.
3) Real Life Application:
   a) Healthcare Data Analysis:
      - Matplotlib can be used to plot details of the patients relating with the complex medical data and even create image visuals for medicals.
      - Seaborn can be used in this case to analyze and plot the statistical plot for the patients or the trends in certain sickness

[https://www.comptia.org/content/articles/how-is-data-analytics-used-in-health-care#:~:text=The%20top%20categories%20of%20data,learning%20to%20propose%20a%20strategy.]

b) Education System Data:
   - Matplotlib can be used to create the distribution of the grades in the certain year or to relate the complex life of the students with the marks obtained
   - Seaborn can be used to make trends in the student performances correlating with marks

[https://www.learninga-z.com/site/resources/breakroom-blog/data-in-education#:~:text=Data%20analysis%20helps%20teachers%20understand,outputs%20(results%20for%20students).]

c) Marketing Business:
   - Matplotlib can be used to visualize the attribution of customers and the products or simply create a performance trend
   - Seaborn can be used to visualize the distribution of customer background or trend to the sales

[https://uk.indeed.com/career-advice/career-development/data-analysis-for-marketing]

After analyzing the situation of how and when can the respective Matplotlib and Seaborn can be used, it is observed that it is not so different in certain ways but quite differently efficient in their own ways. Both libraries are amazing in their own way and only differentiated according to the type of the situation given. But in my opinion, for my report and task given, I have chosen Matplotlib as my primary data visualization library. The reason is that, in my opinion, Matplotlib have given the freedom for me to choose any types of data visualization plots and with the documentations given that will guide me to aid me extract my information and trend needed.

3 PART 2: DATA VISUALIZATION

3.1 Main Goal

Based on the dataset that I had been given which is the ‘Stroke Dataset’, I had made it my goal to find the types of people that will be most at risk and prone for stroke. I like differentiate this dataset to find the possibility factors for stroke. Firstly, I would like to see the vast difference between the genders. If that shows a significant difference, I would get in depth with the factors that affect the men and female differently. However, if the differences shows are small, I would combine the dataset. Then, I would like to see the background of the people that could affect the stroke such as ‘work type’, ‘heart disease’, ‘hypertension’, ‘smoking habits’, ‘alcohol intake’, ‘bmi’ and ‘dietary habits’. These could possibly show that their characteristics such as written before could play a huge key in triggering stroke. This could help people in recognizing the pattern that they are following which are dangerous to their health.

3.2 Results and Discussion

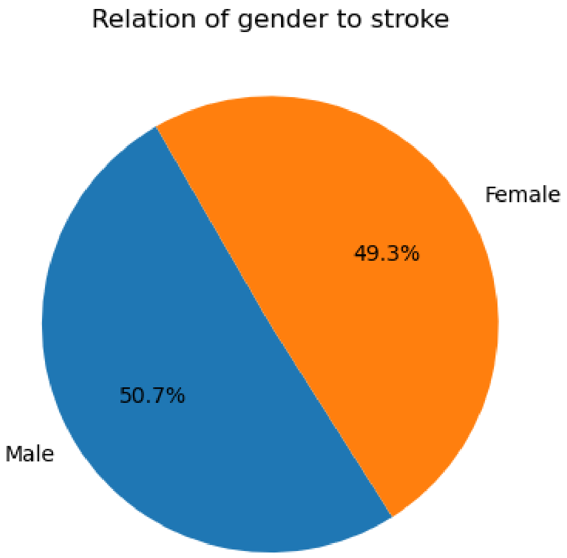


Figure 3.1 shows the relationship of gender to stroke

From the first visual created which is the pie chart to show the differences the gender role played to have more risk to stroke. From the pie chart shown, male population that have stroke are 50.7% while female have 49.3%. The pie chart does show that male is slightly highly than female in terms of getting stroke which brings the meaning that male are more likely at risk to stroke than female. The reason why mal have higher stroke frequency are affected by their lifestyle choices, age and even hormonal differences. Thus, for this dataset we will focus on the lifestyle choices. However, since the differences between the gender are quite small, the stroke dataset will not be divided to study among the lifestyle of each gender but to study the lifestyle of both genders.

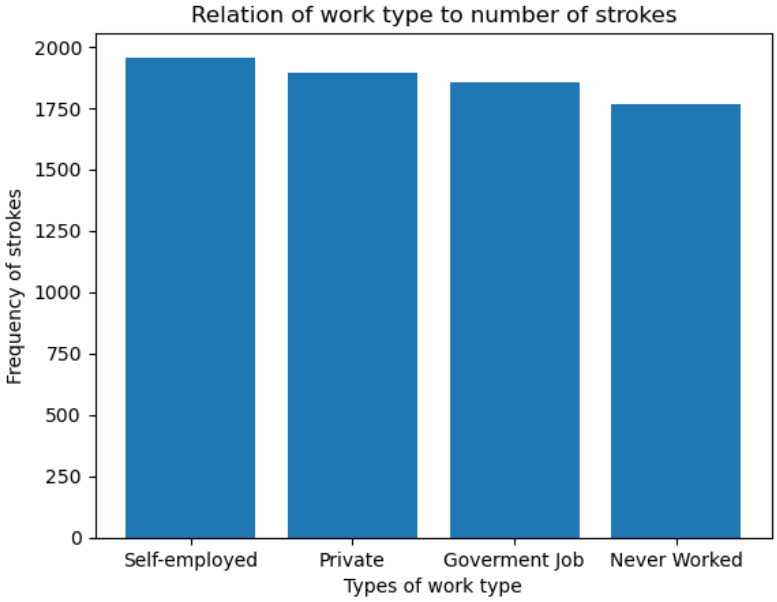


Figure 3.2 shows the relationship of work type to number of strokes

I begin to dive into the what type of lifestyle did people that suffered stroke gone through. Thus, I simply started with their work life as work is a big part of our life that affect us psychologically and physically. From figure 3.2 above, it shows that self-employed workers are more at risk to stroke than private and government job. People that never worked have the lowest number of people that suffered stroke which make logical sense that they never had any pressure or push that leads them to stroke. Then, I studied why does self-employed people have the highest among the four types of work types given, thus I relate the stress for the self-employed workers and observe the frequency of strokes.

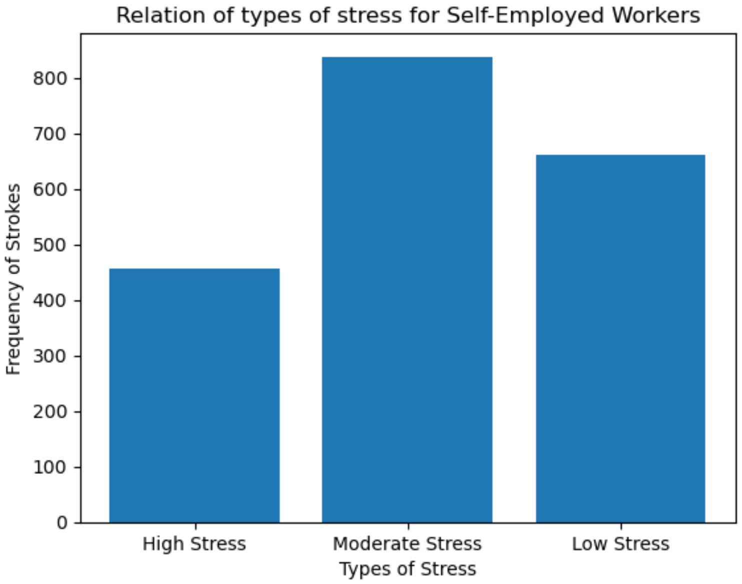


Figure 3.3 shows the frequency of stress against the types of stress for self-employed workers

Based on figure 3.3 above, it shows the more frequent stroke are occurred on moderate stress and less frequent stroke on high stress. Theoretically, this wouldn't make sense, as usually high stress is the common issue for stroke. However, the stress factor did not play a key part in proving the stroke for self-employed workers. Since self-employed workers does not have a permanent work space, this could affect their physical activity to either move frequently or move less as they could be working from home. Thus, I related the physical activity to frequency of stress of self-employed workers.

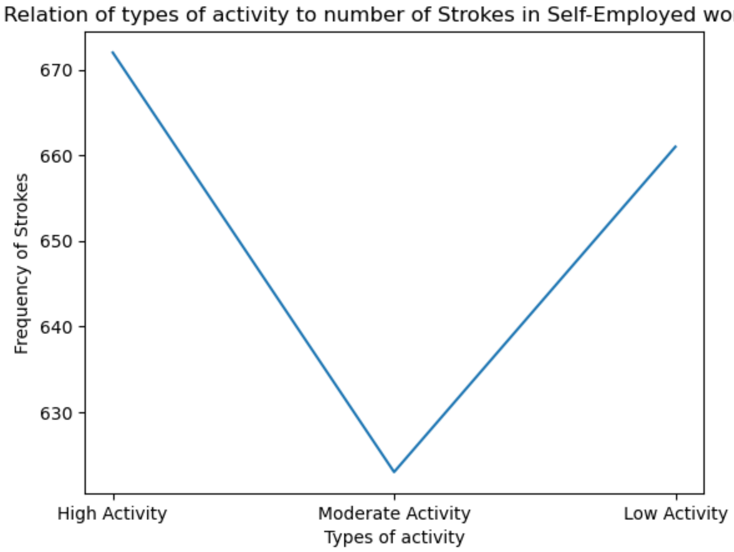


Figure 3.4 shows frequency of strokes against types of activity

Based on graph 3.4 above, self-employed with high activity suffered more strokes than workers with moderate and low activity. These further does not prove that movement can help with stroke. Scientifically proven is that more movement in our daily lifestyle can help more to prevent blood clot which then prevents stroke. Once again this does not factor to in proving that people could suffer stroke from having more or less movement in their lifestyle. Moving on, I considered the possibility that this could play into the major role of genetics.

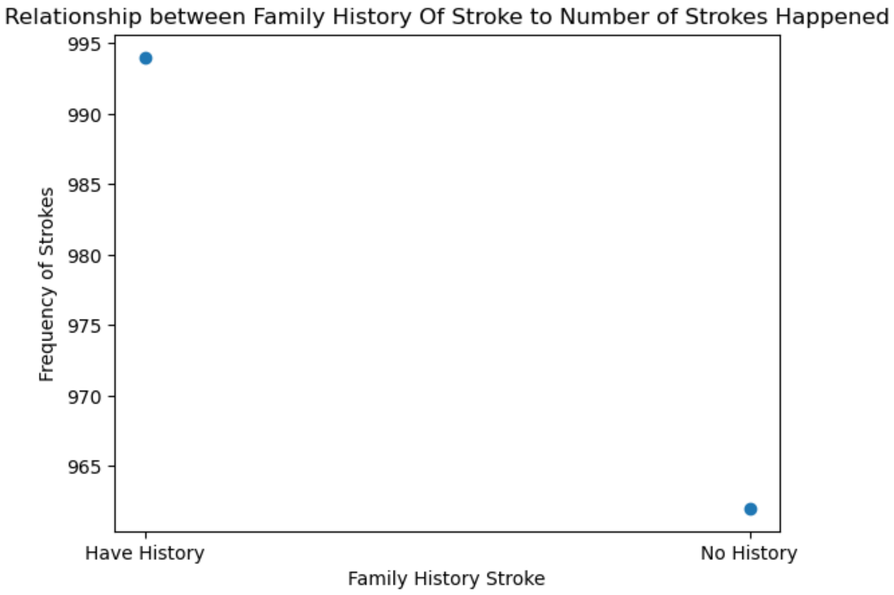


Figure 3.5 shows relation of frequency of strokes to family history stroke

From graph 3.5, it clearly displayed that self-employed workers that have family history strokes are significantly higher than workers with no family history. Thus, from this statement I can safely say that genetics play a huge role in passing the stroke.

Moreover, I dived in the lifestyle of people that suffered stroke, to learn the pattern that would more likely prone themselves having stroke. I started relating the frequency of strokes to heart diseases.

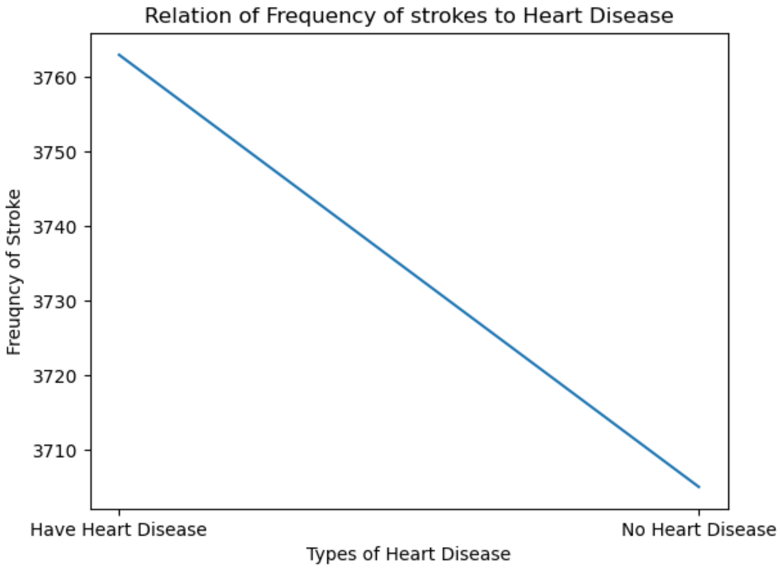


Figure 3.6 shows the relationship for strokes against heart diseases

From figure 3.6, we can observe that people that have heart disease have higher number of strokes than the number of people that have no heart diseases. From this statement, it is safe to say that heart disease is another key that can be used to identify people that more likely to stroke. Similarly, I related the strokes to hypertension.

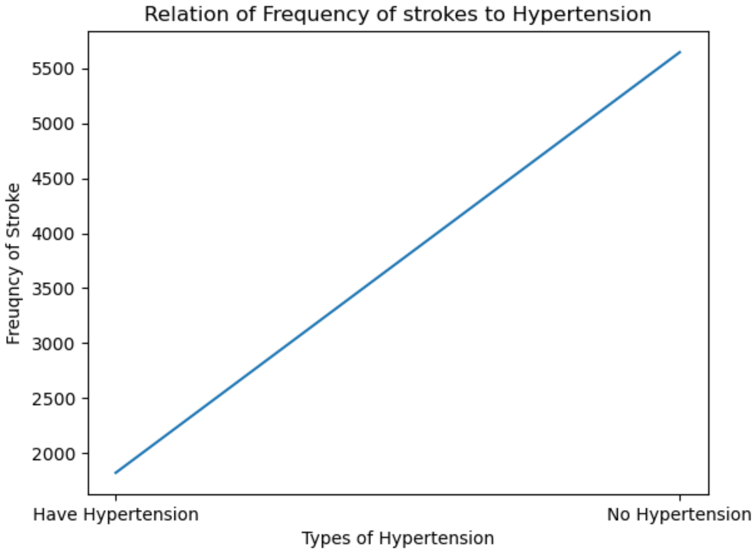


Figure 3.7 shows the relation between strokes and hypertension

From figure 3.7, people with hypertension have lesser number of strokes than people without hypertension. This brings the question that hypertension supposed to trigger the strokes as high blood pressure damages the blood vessels as it strains on the walls. However, from the graph above, the hypertension does not play the key role in determining the possibility of stroke. We will compare why heart disease is more prone to stroke than hypertension. Heat disease directly causes the clotting formation in the blood vessels such as in heart and arteries. The effects are quicker than hypertension which also similarly damages and causes blood clot in the vessels. Heart disease is also a direct disease that effects on the heart while hypertension is high blood pressure which affects throughout the body.

However, from previous figure, we have concluded that heart disease is important to determination of stroke. Thus, I related the factors that could tribute to heart disease such as Body Mass Index, dietary habits, alcohol intake and smoking habits. All these factors are taken into consideration that could lead to heart disease.

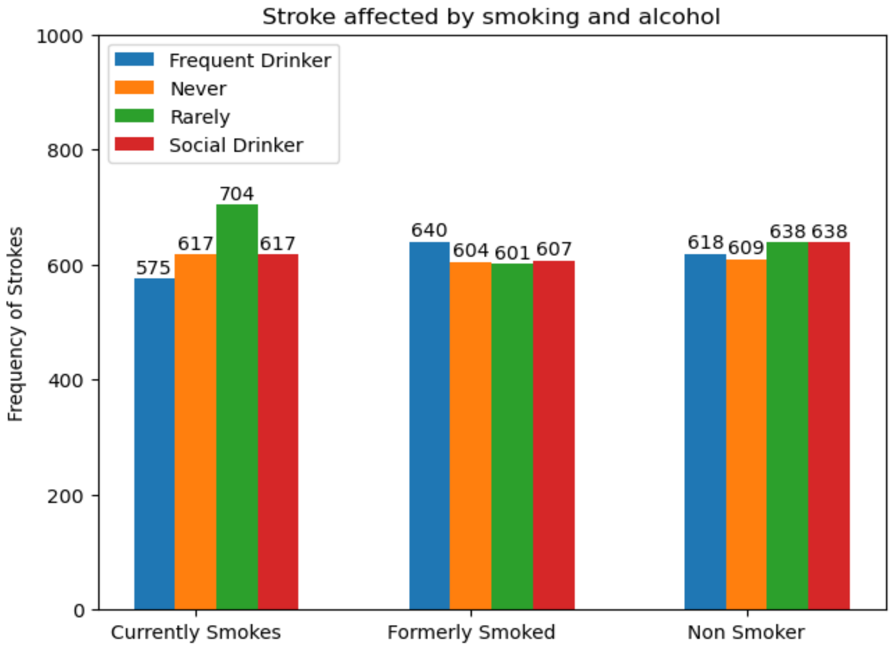


Figure 3.8 shows the relation of smoking habits and alcohol intake to stroke

Based on figure 3.8, it is observed that the group of currently smokes and rarely drinker takes the lead in having the most strokes. Comparing this with the formerly smoked and frequent drinker with second in the lead for most strokes. This shows that the smoking habits influences the risk of stroke more than alcohol intake. Smoke also causes damage to the blood vessels and cause clot formation which increase the risk of stroke. The reason why smoking causes more stroke than alcohol intake is, smoking impact the blood vessels directly which narrows and hardens the blood vessels. This process is very direct unlike alcohol which increases the blood pressure first then follows by the damage of the blood vessels. This shows that alcohol intake is less frequent leads to stroke than smoking. I visualize another bar chart to see the effect of smoking habits on heart disease.

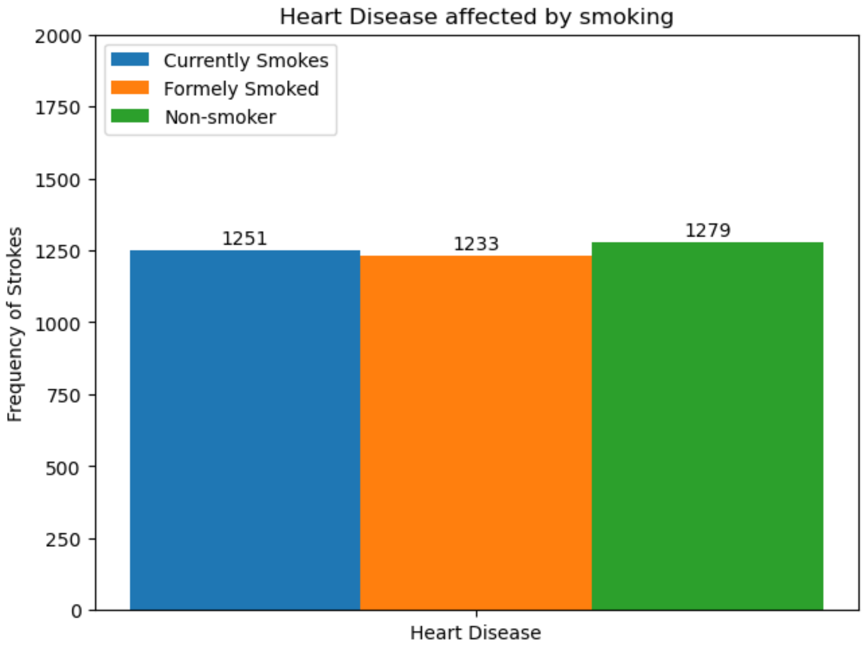


Figure 3.9 shows the relationship of smoking status to heart disease and stroke

Based on graph 3.9, people that are non-smoker have the most strokes which have heart disease compare to other. Although the differences are small, I take in count that smoking habits does effect on the lifestyle which will lead to stroke. Then, I visualized a graph of BMI and dietary habits related against number of strokes.

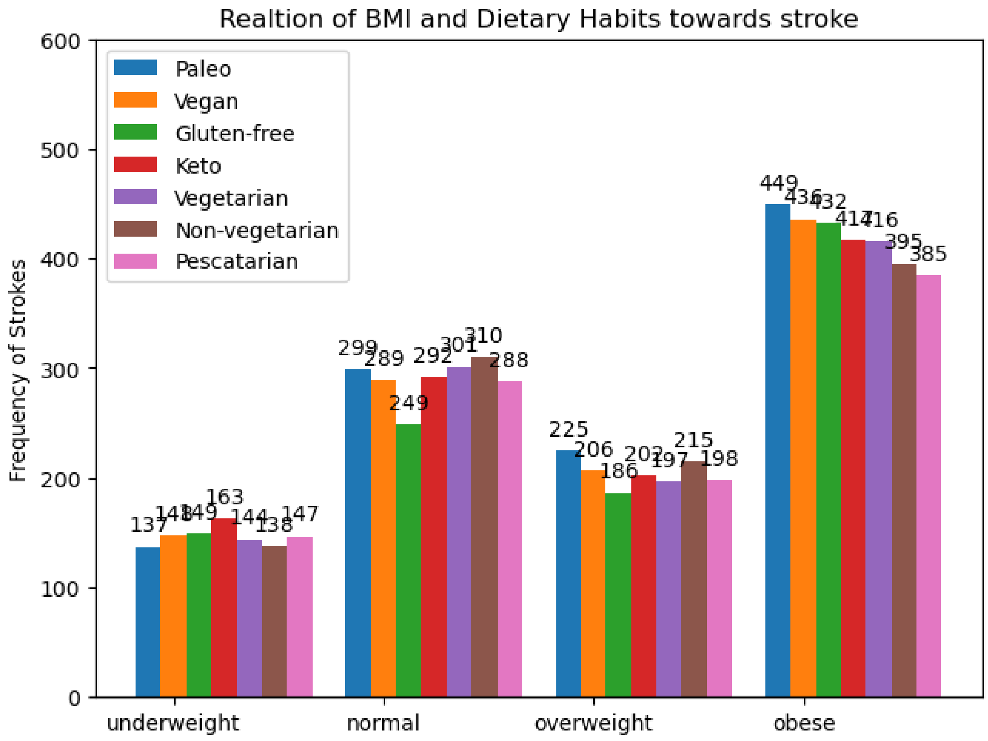


Figure 3.10 shows frequency of strokes against types of diets and BMI

Based on figure 3.10 above, the most number of strokes are in the group of 'Obese' which proves the point of people with high BMI are highly likely to have stroke as obesity can cause inflammation caused by excess fatty tissue. This will cause the blood flow difficult to flow which brings to blockage and thus stroke. On the same figure 3.10, we can analyse that on the group of 'Obese', the top 3 diets leading was Paleo diet, Vegan diet followed by Gluten-free diet. I tried to understand how does these three diets had cause much strokes and high BMI as they could another key factor of stroke. But before that, I want to quickly conclude the BMI does play a factor in knowing the risk for stroke.

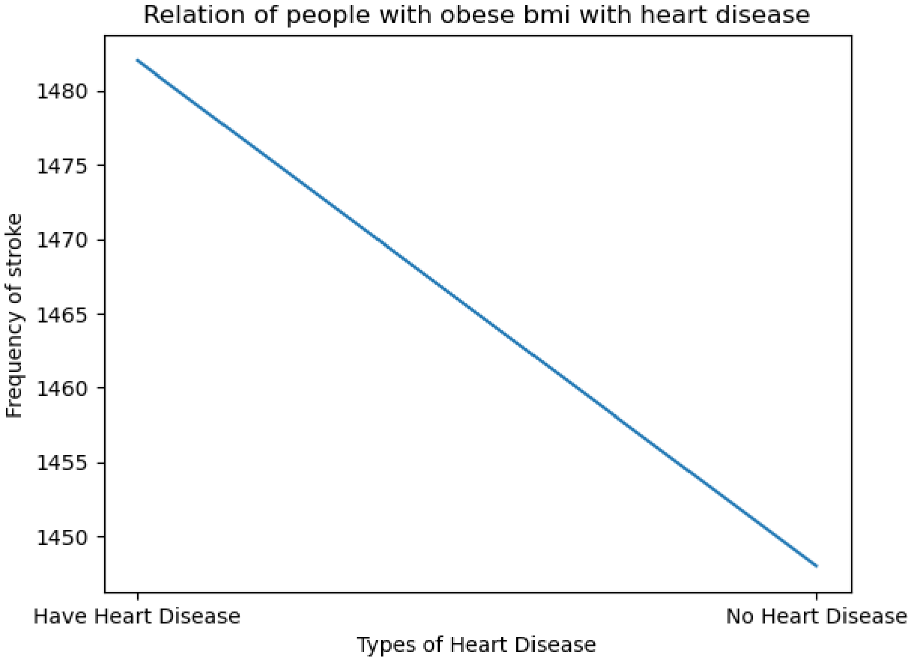


Figure 3.11 shows people with obese BMI with types of heart disease against number of strokes

From the figure 3.11 it is obvious that people with Obese BMI have high number of strokes and having heart disease. This concludes that BMI factor is important in determining heart disease which then to stroke.

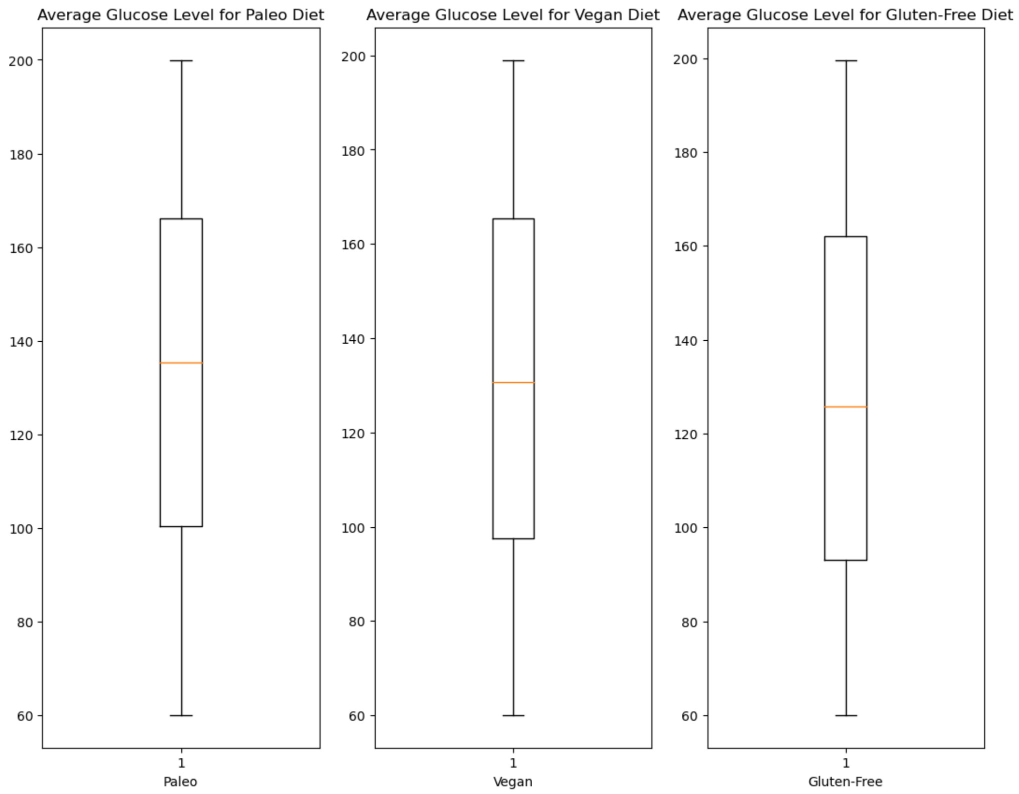


Figure 3.12 shows the boxplots distribution of Paleo, Vegan and Gluten-free diet

Based on the figure 3.12 above, we can observe the median line from Paleo to Vegan and to Glute-free diet slowly declining. We can also tell that all three of them have normal distributions. The decline shows decreasing trend of average glucose level for people that practise those diets. There higher average people with higher glucose level in Paleo diet compare to Vegan and Gluten-free diet. This can be concluded as Paleo diet is more at risk to increase the glucose level in people. High sugar blood associated with increased risk of coronary heart diseases and strokes[https://drc.bmj.com/content/9/1/e001928]. Paleo diet has higher protein and fat intake compare to others which impact highly on insulin levels of the individuals. We can conclude that dietary habits is a key to determining heart diseases and strokes as well. Finally, I categorized the symptoms into 10 different categories which are numbness, headache, difficulty speaking, loss of balance, confusion, weakness, severe fatigue, dizziness, blurred vision and seizures. The reason why I did this is to identify the symptoms that would most likely risk of stroke.

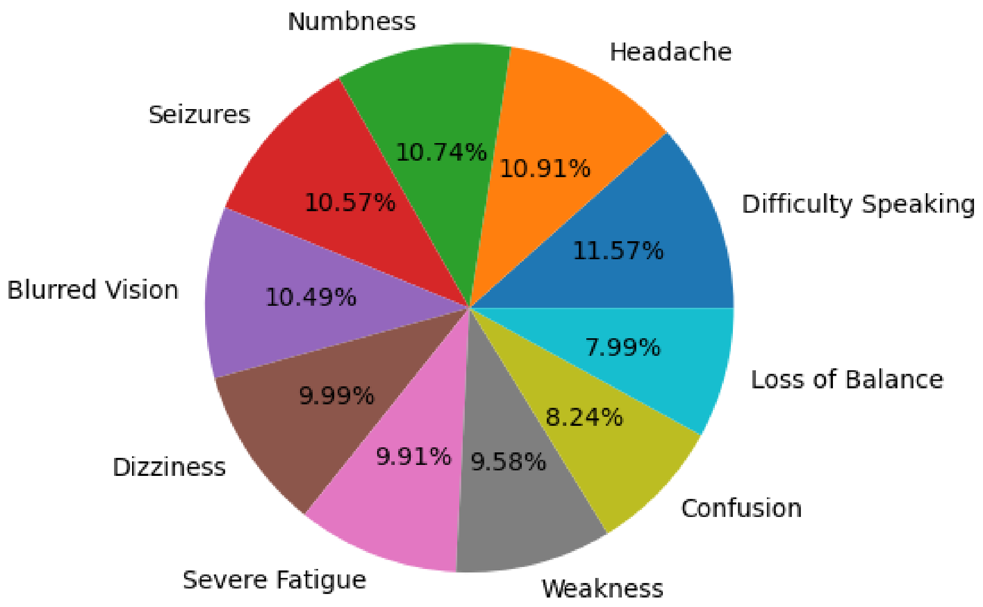
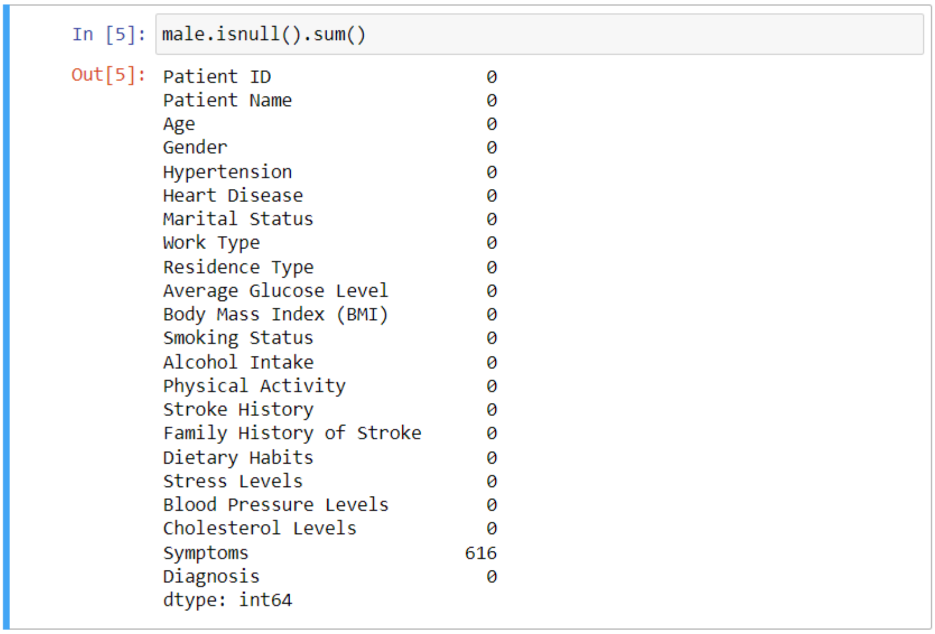


Figure 3.13 shows a pie chart of symptoms

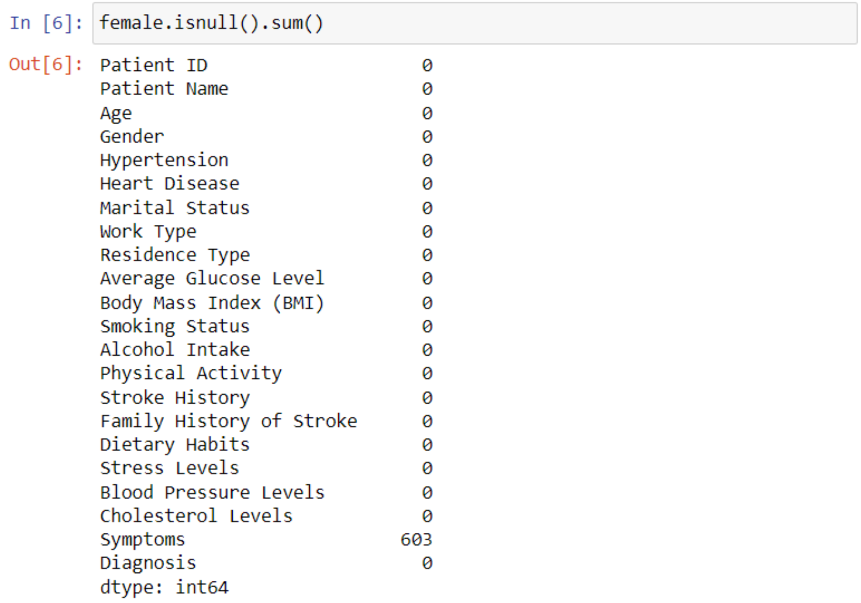
Based on figure 3.13 above, the top three symptoms that high risk of strokes are difficulty speaking, headache and numbness in respective order.

3.3 Steps to build the visualizations

1) Trying to find the missing data of strokes between male and female



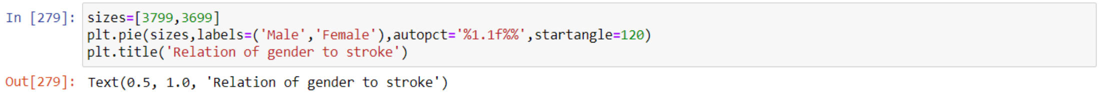
```
In [5]: male.isnull().sum()

Out[5]: Patient ID                      0
        Patient Name                    0
        Age                             0
        Gender                          0
        Hypertension                    0
        Heart Disease                   0
        Marital Status                  0
        Work Type                       0
        Residence Type                  0
        Average Glucose Level           0
        Body Mass Index (BMI)           0
        Smoking Status                  0
        Alcohol Intake                  0
        Physical Activity               0
        Stroke History                  0
        Family History of Stroke        0
        Dietary Habits                  0
        Stress Levels                   0
        Blood Pressure Levels           0
        Cholesterol Levels              0
        Symptoms                      616
        Diagnosis                       0
        dtype: int64
```



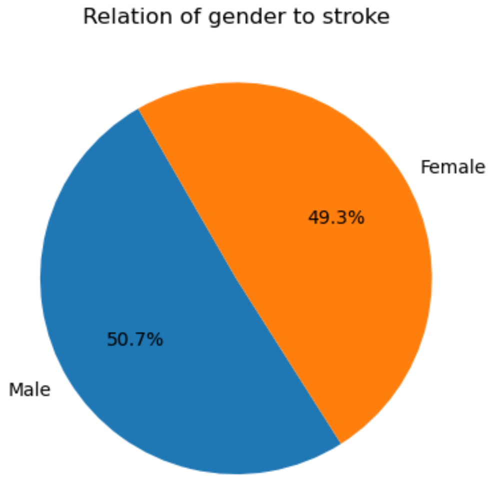
```
In [6]: female.isnull().sum()

Out[6]: Patient ID                      0
        Patient Name                    0
        Age                             0
        Gender                          0
        Hypertension                    0
        Heart Disease                   0
        Marital Status                  0
        Work Type                       0
        Residence Type                  0
        Average Glucose Level           0
        Body Mass Index (BMI)           0
        Smoking Status                  0
        Alcohol Intake                  0
        Physical Activity               0
        Stroke History                  0
        Family History of Stroke        0
        Dietary Habits                  0
        Stress Levels                   0
        Blood Pressure Levels           0
        Cholesterol Levels              0
        Symptoms                      603
        Diagnosis                       0
        dtype: int64
```

We can see that there are no missing data in all columns except the ‘symptoms’ column

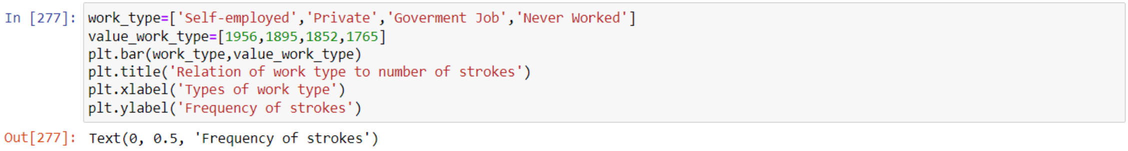
2) Pie chart of male and female distributions of stroke



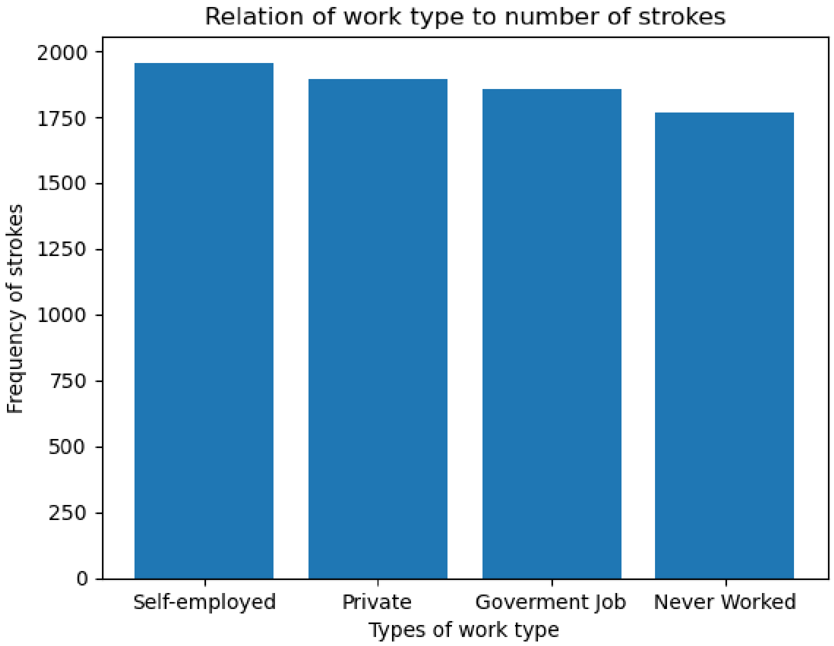
```
In [279]: sizes=[3799,3699]
          plt.pie(sizes,labels=('Male','Female'),autopct='%1.1f%%',startangle=120)
          plt.title('Relation of gender to stroke')

Out[279]: Text(0.5, 1.0, 'Relation of gender to stroke')
```



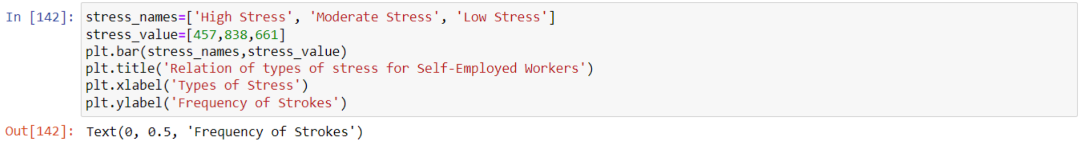
3) Bar chart of work types with frequency of stroke



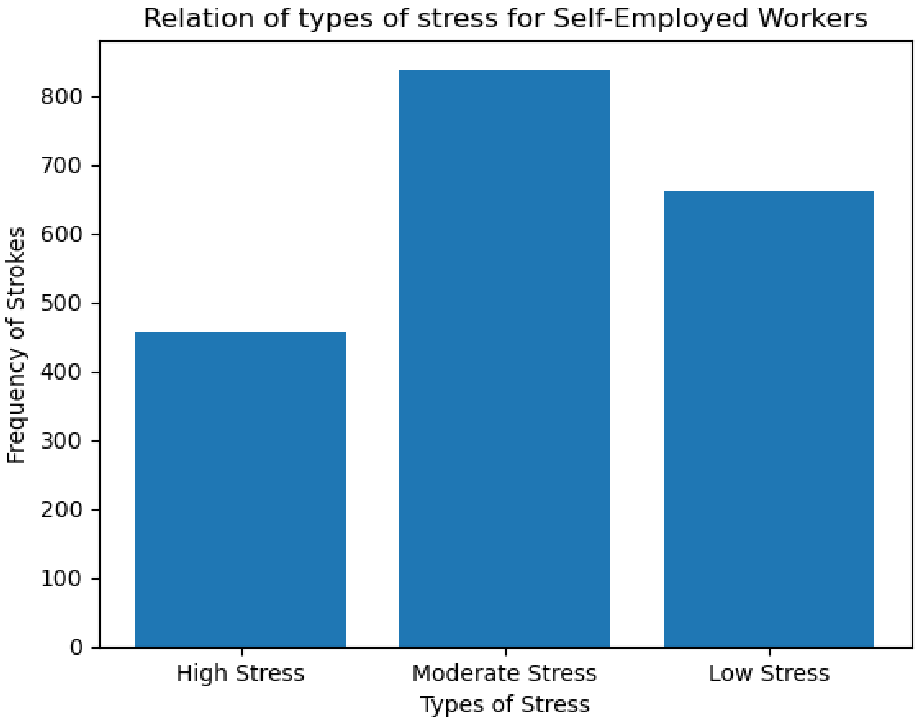
```
In [277]: work_type=['Self-employed','Private','Goverment Job','Never Worked']
          value_work_type=[1956,1895,1852,1765]
          plt.bar(work_type,value_work_type)
          plt.title('Relation of work type to number of strokes')
          plt.xlabel('Types of work type')
          plt.ylabel('Frequency of strokes')
Out[277]: Text(0, 0.5, 'Frequency of strokes')
```



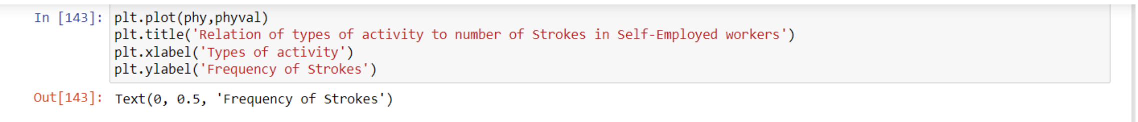
4) Bar chart of types of stress for self-employed workers with frequency of stroke



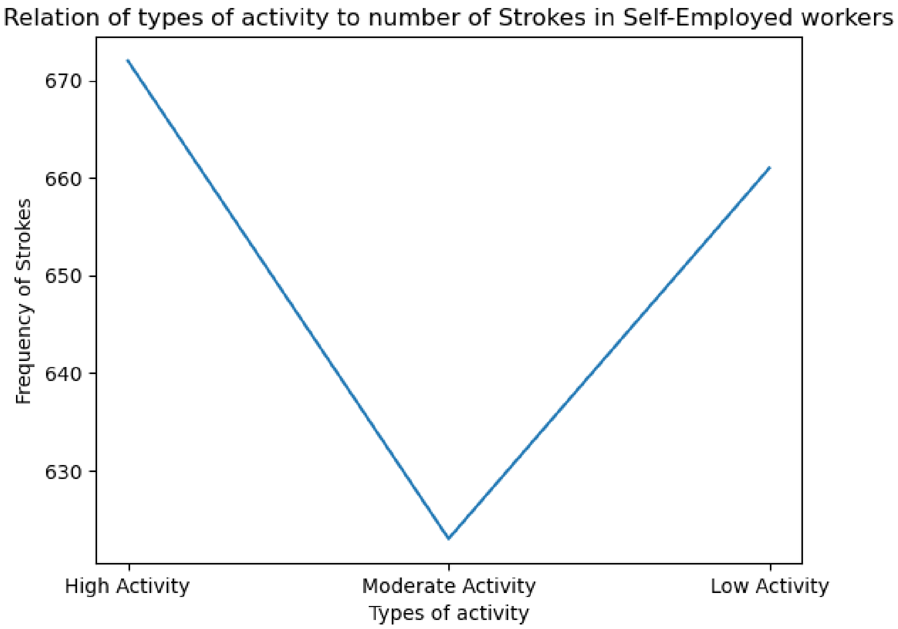
```
In [142]: stress_names=['High Stress', 'Moderate Stress', 'Low Stress']
          stress_value=[457,838,661]
          plt.bar(stress_names,stress_value)
          plt.title('Relation of types of stress for Self-Employed Workers')
          plt.xlabel('Types of Stress')
          plt.ylabel('Frequency of Strokes')
Out[142]: Text(0, 0.5, 'Frequency of Strokes')
```



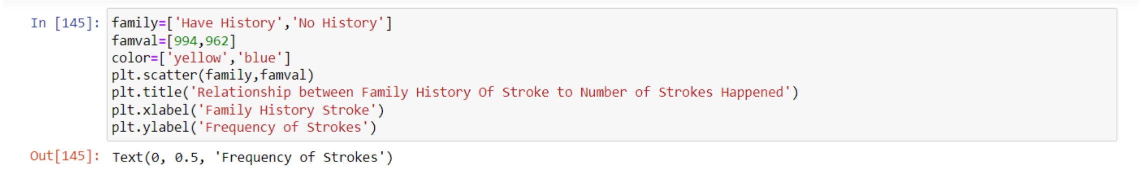
5) Line chart of activity types for self-employed workers against frequency of strokes



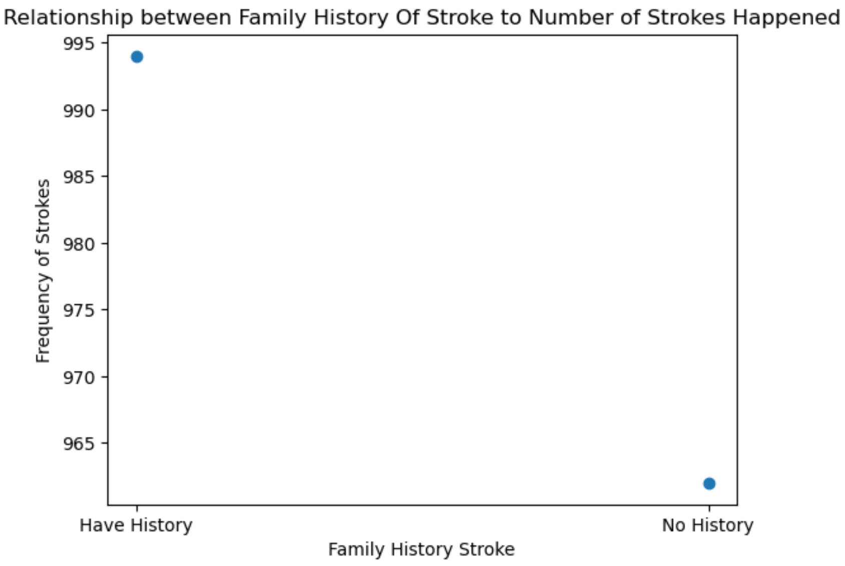
```
In [143]: plt.plot(phy,phyval)
          plt.title('Relation of types of activity to number of Strokes in Self-Employed workers')
          plt.xlabel('Types of activity')
          plt.ylabel('Frequency of Strokes')

Out[143]: Text(0, 0.5, 'Frequency of Strokes')
```



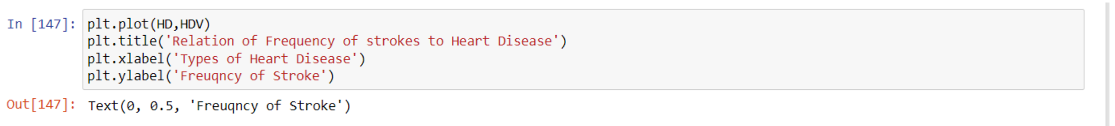
6) Scatter plot of family history with strokes with frequency of stroke



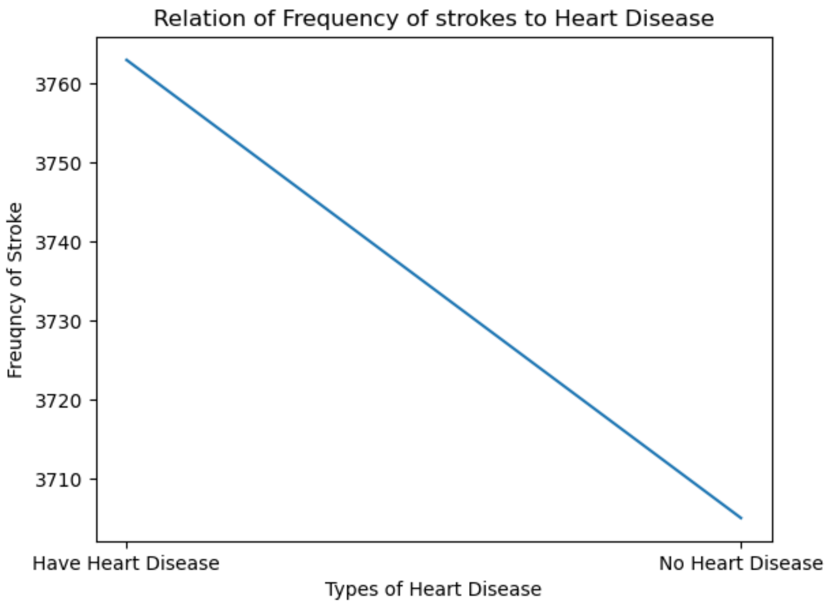
```
In [145]: family=['Have History','No History']
          famval=[994,962]
          color=['yellow','blue']
          plt.scatter(family,famval)
          plt.title('Relationship between Family History Of Stroke to Number of Strokes Happened')
          plt.xlabel('Family History Stroke')
          plt.ylabel('Frequency of Strokes')
Out[145]: Text(0, 0.5, 'Frequency of Strokes')
```



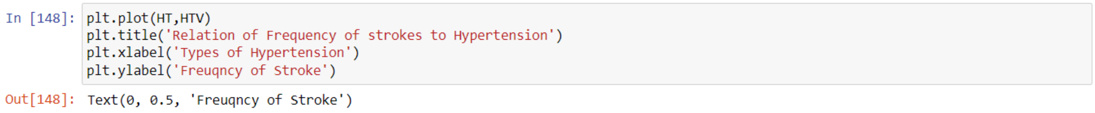
7) Line chart of types of heart disease against frequency of strokes



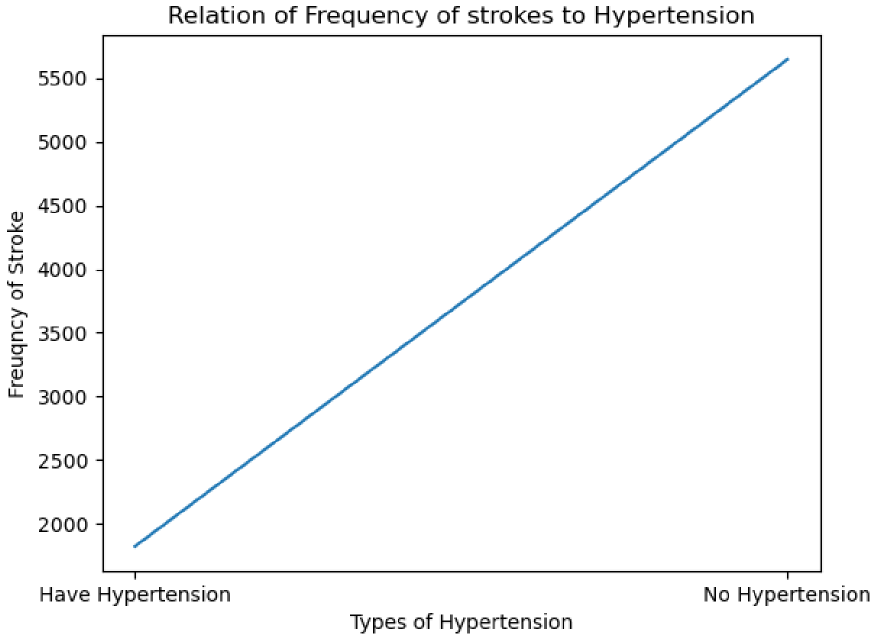
```
In [147]: plt.plot(HD,HDV)
          plt.title('Relation of Frequency of strokes to Heart Disease')
          plt.xlabel('Types of Heart Disease')
          plt.ylabel('Freuqncy of Stroke')
Out[147]: Text(0, 0.5, 'Freuqncy of Stroke')
```



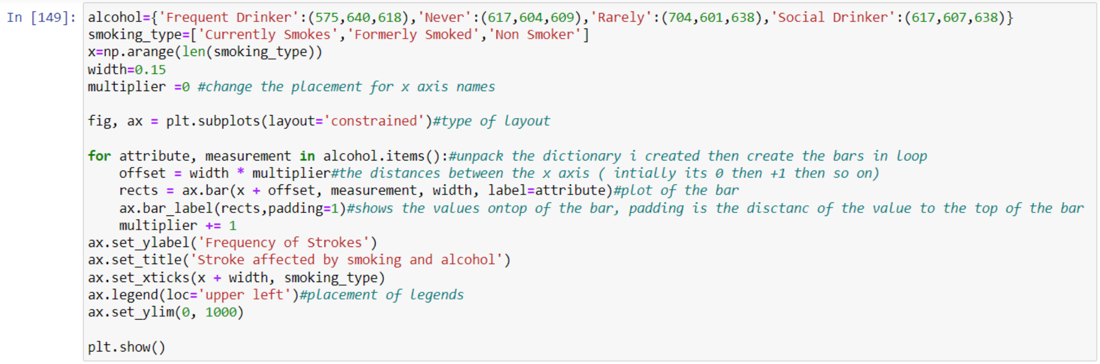
8) Line chart of hypertension types with frequency of strokes



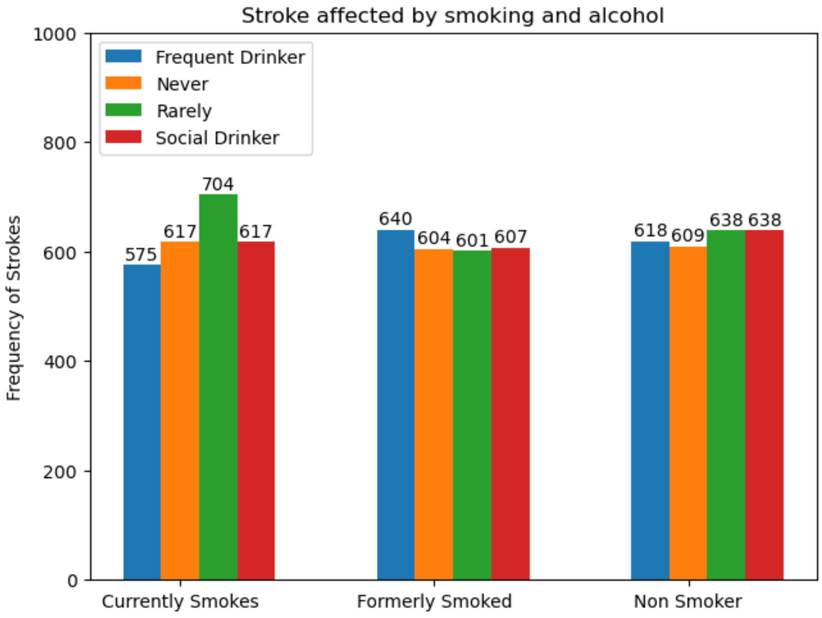
```
In [148]: plt.plot(HT,HTV)
          plt.title('Relation of Frequency of strokes to Hypertension')
          plt.xlabel('Types of Hypertension')
          plt.ylabel('Freuqncy of Stroke')

Out[148]: Text(0, 0.5, 'Freuqncy of Stroke')
```



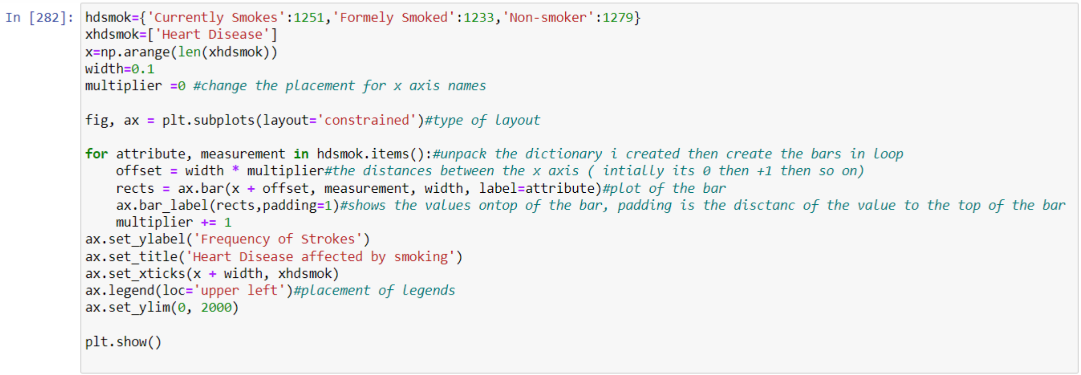
9) Stacked bar chart relating alcohol intake with smoking types against frequency of strokes



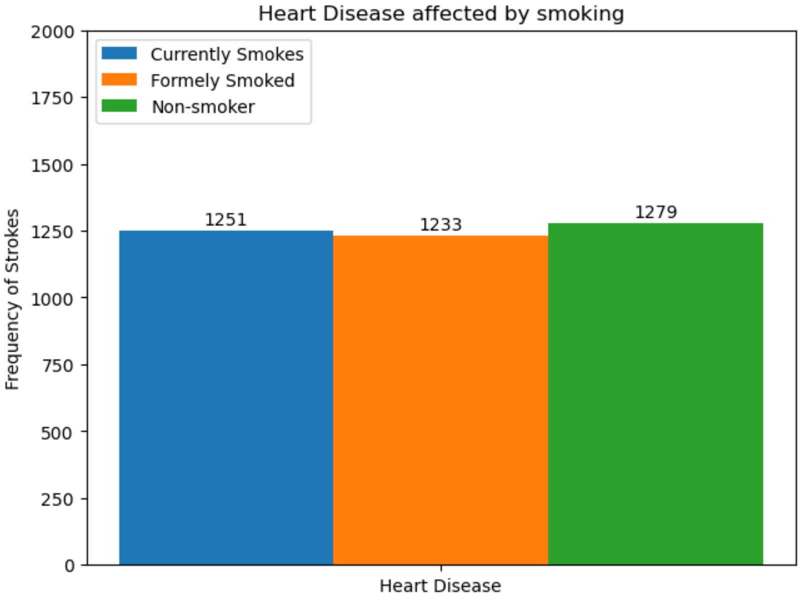
```
In [149]: alcohol={'Frequent Drinker':(575,640,618),'Never':(617,604,609),'Rarely':(704,601,638),'Social Drinker':(617,607,638)}
          smoking_type=['Currently Smokes','Formerly Smoked','Non Smoker']
          x=np.arange(len(smoking_type))
          width=0.15
          multiplier =0 #change the placement for x axis names

          fig, ax = plt.subplots(layout='constrained')#type of layout

          for attribute, measurement in alcohol.items():#unpack the dictionary i created then create the bars in loop
              offset = width * multiplier#the distances between the x axis ( intially its 0 then +1 then so on)
              rects = ax.bar(x + offset, measurement, width, label=attribute)#plot of the bar
              ax.bar_label(rects,padding=1)#shows the values ontop of the bar, padding is the disctanc of the value to the top of the bar
              multiplier += 1
          ax.set_ylabel('Frequency of Strokes')
          ax.set_title('Stroke affected by smoking and alcohol')
          ax.set_xticks(x + width, smoking_type)
          ax.legend(loc='upper left')#placement of legends
          ax.set_ylim(0, 1000)

          plt.show()
```



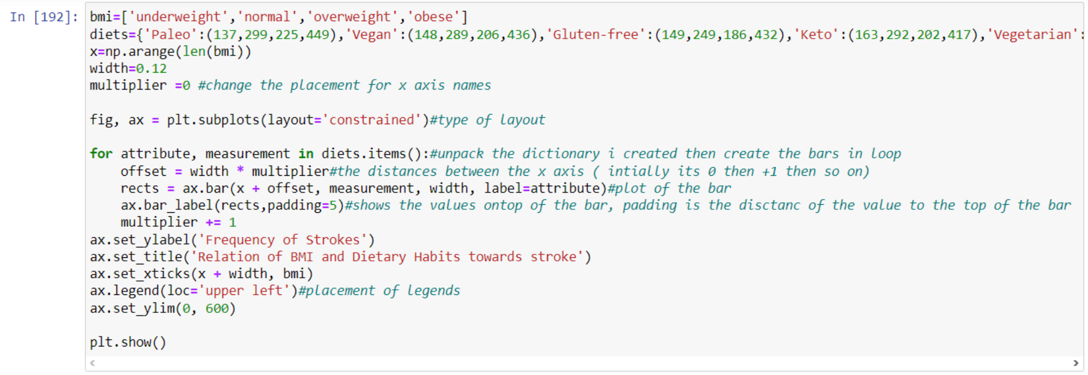
10) Bar chart relating heart disease with smoking against frequency of strokes



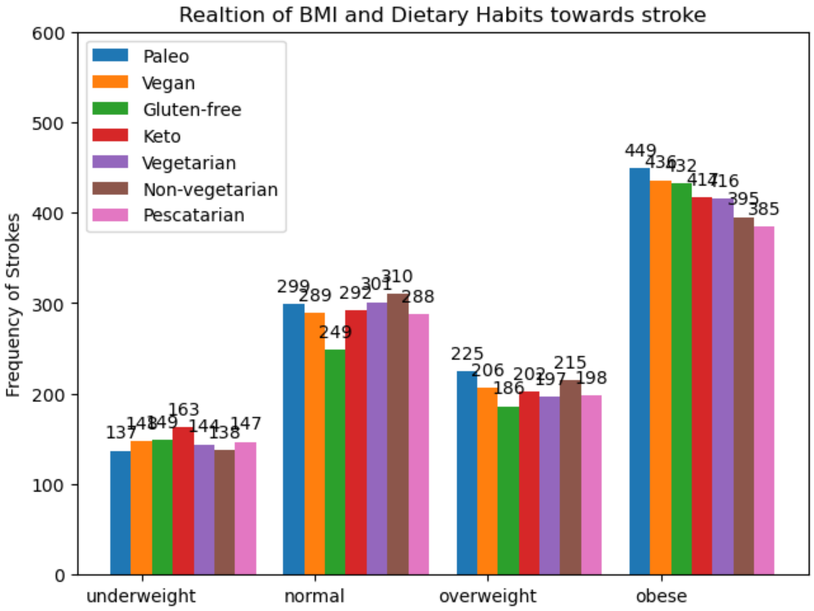
```
In [282]: hdsmok={'Currently Smokes':1251,'Formely Smoked':1233,'Non-smoker':1279}
          xhdsmok=['Heart Disease']
          x=np.arange(len(xhdsmok))
          width=0.1
          multiplier =0 #change the placement for x axis names

          fig, ax = plt.subplots(layout='constrained')#type of layout

          for attribute, measurement in hdsmok.items():#unpack the dictionary i created then create the bars in loop
              offset = width * multiplier#the distances between the x axis ( intially its 0 then +1 then so on)
              rects = ax.bar(x + offset, measurement, width, label=attribute)#plot of the bar
              ax.bar_label(rects,padding=1)#shows the values ontop of the bar, padding is the disctanc of the value to the top of the bar
              multiplier += 1
          ax.set_ylabel('Frequency of Strokes')
          ax.set_title('Heart Disease affected by smoking')
          ax.set_xticks(x + width, xhdsmok)
          ax.legend(loc='upper left')#placement of legends
          ax.set_ylim(0, 2000)

          plt.show()
```



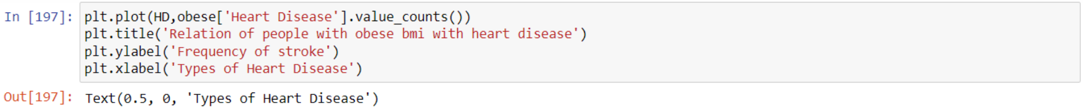
11) Stacked bar chart relating BMI and dietary habits against frequency of strokes



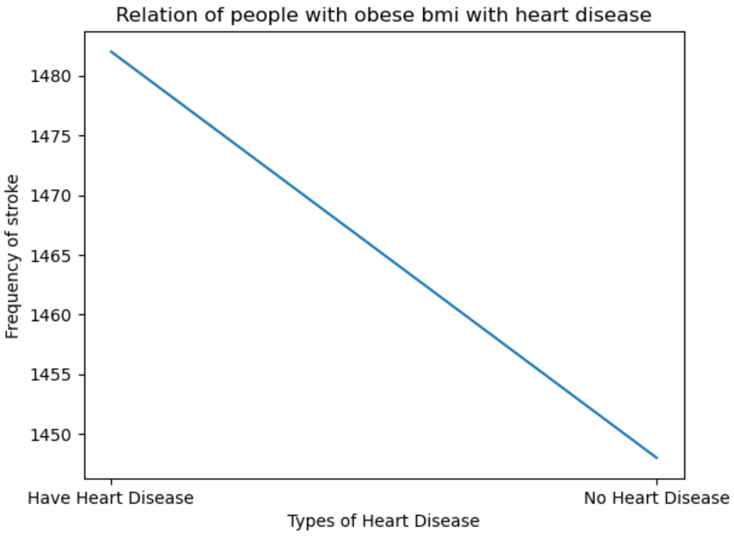
```
In [192]: bmi=['underweight','normal','overweight','obese']
          diets={'Paleo':(137,299,225,449),'Vegan':(148,289,206,436),'Gluten-free':(149,249,186,432),'Keto':(163,292,202,417),'Vegetarian':
          x=np.arange(len(bmi))
          width=0.12
          multiplier =0 #change the placement for x axis names

          fig, ax = plt.subplots(layout='constrained')#type of layout

          for attribute, measurement in diets.items():#unpack the dictionary i created then create the bars in loop
              offset = width * multiplier#the distances between the x axis ( intially its 0 then +1 then so on)
              rects = ax.bar(x + offset, measurement, width, label=attribute)#plot of the bar
              ax.bar_label(rects,padding=5)#shows the values ontop of the bar, padding is the disctanc of the value to the top of the bar
              multiplier += 1
          ax.set_ylabel('Frequency of Strokes')
          ax.set_title('Relation of BMI and Dietary Habits towards stroke')
          ax.set_xticks(x + width, bmi)
          ax.legend(loc='upper left')#placement of legends
          ax.set_ylim(0, 600)

          plt.show()
```



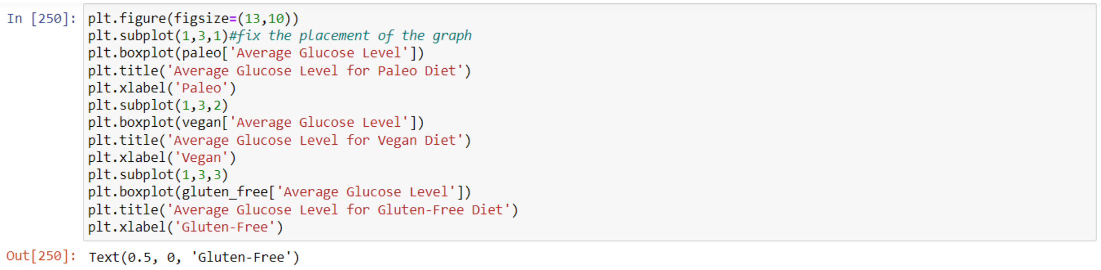
12) Line chart relating the people of Obese BMI with heart disease against frequency of strokes



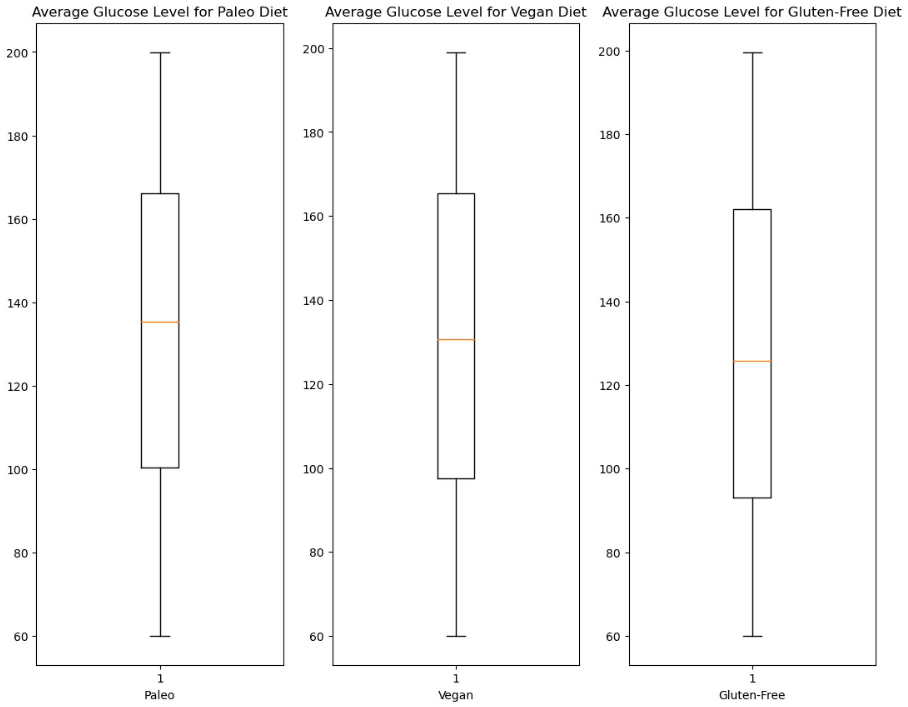
```
In [197]: plt.plot(HD,obese['Heart Disease'].value_counts())
          plt.title('Relation of people with obese bmi with heart disease')
          plt.ylabel('Frequency of stroke')
          plt.xlabel('Types of Heart Disease')

Out[197]: Text(0.5, 0, 'Types of Heart Disease')
```



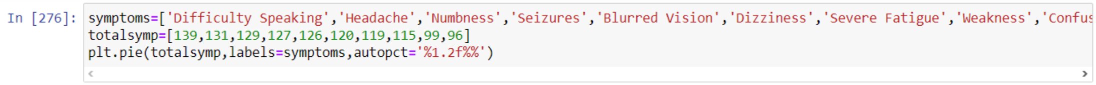
13) Box plots of average glucose level with Paleo, Vegan and Gluten-free diet



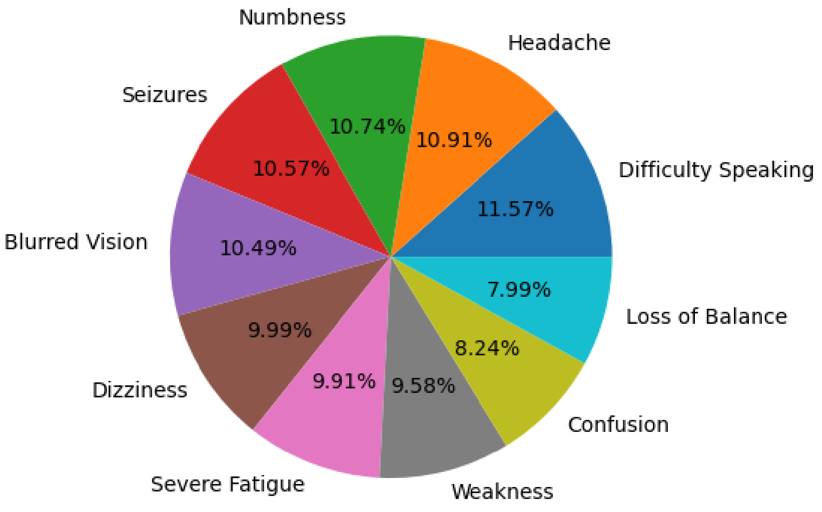
```
In [250]: plt.figure(figsize=(13,10))
          plt.subplot(1,3,1)#fix the placement of the graph
          plt.boxplot(paleo['Average Glucose Level'])
          plt.title('Average Glucose Level for Paleo Diet')
          plt.xlabel('Paleo')
          plt.subplot(1,3,2)
          plt.boxplot(vegan['Average Glucose Level'])
          plt.title('Average Glucose Level for Vegan Diet')
          plt.xlabel('Vegan')
          plt.subplot(1,3,3)
          plt.boxplot(gluten_free['Average Glucose Level'])
          plt.title('Average Glucose Level for Gluten-Free Diet')
          plt.xlabel('Gluten-Free')

Out[250]: Text(0.5, 0, 'Gluten-Free')
```



14) Pie chart that distributes all of the symptoms



```
In [276]: symptoms=['Difficulty Speaking','Headache','Numbness','Seizures','Blurred Vision','Dizziness','Severe Fatigue','Weakness','Confus
          totalsymp=[139,131,129,127,126,120,119,115,99,96]
          plt.pie(totalsymp,labels=symptoms,autopct='%1.2f%%')
```



Recommendation:

1. Identify the risk factor of Genetic stroke. The research into the components of stroke and how is it can be mutated must be priotised since this issue is one of the most difficult cases. Government or private sectors could invest in treatment strategies by analyzing an individual’s genetic profile. Eventhough, genetics play a hard role in keeping the stroke, there are other possibilities that could aid the individuals such as changing their lifestyle, environment, hobbies, diets and exercises.
2. Heart disease is often based on the lifestyle that was led by the individual’s choices. One of the lifestyles is smoking habits. Government can increase their efforts in organizing campaigns to stop smoking. Government also could ban or increase the prices of smoke to reduce smokes consumptions. This will discourage the people with lower income or sensitive to their money for considering dropping the smoking habits.
3. BMI is very essential to ensure the overall health and reducing the risk heart diseases. There must campaigns held to help people in managing their diets and eating style. There are dieticians who are able to guide the community into the healthy eating style. These dieticians are expert in knowing to control the types of food that are able to control the glucose level as well. To maintain the healthy BMI, one must always regularly exercise and have healthy sleep patterns.
4. For the individuals that have frequent symptoms of difficulty speaking, headache and numbness, they must be given a higher priority for treatments. There must a campaign held to give awareness of the symptoms that lead to stroke which will be educating the people in need of help.

4 CONCLUSION

In conclusion, the findings that I was able to obtain was:

- Gender does not take account to shows significant difference in number of strokes
- Self-employed workers are most number workers to have suffered stroke compare to private, government and never worked people.
- Stress and types of activity are not major factor that divine into stroke
- Family history of stroke is largely played factored in knowing the risk for stroke
- Heart disease is a more dangerous disease than hypertension for risk to stroke
- Smoking habits affects the heart disease more than alcohol intake
- Both Body Mass Index (BMI) and dietary habits are taken as factors that lead to both heart disease and stroke
- Dietary habits heavily impact onto glucose level which then cause heart disease.
- The top three symptoms to look out for are difficulty speaking, headache and numbness

The purpose of my data visualization was to find the factors that was led people in having stroke. During the process of visualizing, I found that certain visualization goes against the law of logic which withheld my decision-making. However, from the results that I have obtained, my findings can be used as a remarkable factor to help the people in fixing their habits and lifestyle for the sake of having a healthy and stroke-free life.

REFERENCES

1. BYJU'S:
   https://byjus.com/maths/introduction-to-data
2. Devin Pickel (3 MAY 2023). Structured vs Unstructured Data:
   https://www.g2.com/articles/structured-vs-unstructured-data
3. Kate Brush & Ed Burns (December 2022). Data Visualization:
   https://www.techtarget.com/searchbusinessanalytics/definition/data-visualization
4. Sas. Data Visualization What it is and why it matters:
   https://www.sas.com/en_my/insights/big-data/data-visualization.html
5. Shrunali Suresh Salian (3 March). Spotify Data Visualization:
   https://medium.com/@shrunalisalian97/spotify-data-visualization-4c878c8114e
6. Levon Hovsepyan (29 October 2023). How Starbuck uses Big Data:
   https://www.linkedin.com/pulse/how-starbucks-uses-big-data-levon-hovsepyan-7bprf/
7. Statistics Canada. Bar chart:
   https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch9/bargraph-diagrammeabarres/5214818-eng.htm
8. The Data Visualization Catalogue. Line Graph:
   https://datavizcatalogue.com/methods/line_graph.html
9. Statistics Canada. Scatter Plot:
   https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch9/scatter-nuages/5214827-eng.htm
10. Statistics How To. Box Plot: How to read one and how to make one in Excel, TI-93, SPSS:
    https://www.statisticshowto.com/probability-and-statistics/descriptive-statistics/box-plot/
11. Wikipedia. Matplotlib:
    https://en.wikipedia.org/wiki/Matplotlib
12. Matplotlib. History:
    https://matplotlib.org/stable/users/project/history.html
13. TestGorilla. What is Matplotlib in Python? Top 10 advantages of Matplotlib you should know:

https://www.testgorilla.com/blog/matplotlib-in-python/

14. Azib Farooq (13 April 2023). What is strength and weakness of python programming language: https://www.quora.com/What-is-the-strength-and-weakness-of-the-Python-programming-language
15. Hacksight (20 August 2022). Difference between Matplotlib vs Seaborn.: https://www.geeksforgeeks.org/difference-between-matplotlib-vs-seaborn/
16. Michael Waskom. An introduction to seaborn:https://seaborn.pydata.org/tutorial/introduction.html
17. Michael Waskom (2023). Micheal Waskom, PHD: https://mwaskom.github.io/
18. Micael Waskom (22 Febuary 2021). Three common Seaborn difficulties: https://michaelwaskom.medium.com/three-common-seaborn-difficulties-10fdd0cc2a8b
19. Linkedln. What are the advantages and disadvantages of using seaborn over other visualization libraries: https://www.linkedin.com/advice/3/what-advantages-disadvantages-using-seaborn#:~:text=2%20Disadvantages%20of%20seaborn,-Seaborn%20is%20not&text=It%20can%20be%20slow%20and,optimized%20for%20performance%20or%20scalability
20. Console Flare (22 May 2023). Differences between Matplotlib vs Seaborn: https://blog.consoleflare.com/matplotlib-vs-seaborn/
21. Aizhamal Zhetigenova. Matplotlib vs Seaborn: https://codesolid.com/matplotlib-vs-seaborn/#google_vignette
22. RitzaArticles. Matplotlib vs Seaborn vs Plotly vs MATLAB vs ggplot2 vs pandas: https://ritza.co/articles/matplotlib-vs-seaborn-vs-plotly-vs-MATLAB-vs-ggplot2-vs-pandas/
23. Biswajit Panda, Quora. What are the benefits of using Seaborn over matplotlib/pandas for plotting a dataframe :https://www.quora.com/What-are-the-benefits-of-using-Seaborn-over-matplotlib-pandas-for-plotting-data-frames-in-Python
24. CompTIA. How is data analytics used in health care?: https://www.comptia.org/content/articles/how-is-data-analytics-used-in-health-

care#:~:text=The%20top%20categories%20of%20data,learning%20to%20propose%20a%20strategy

25. Iris Garner. Data in Education: https://www.learninga-z.com/site/resources/breakroom-blog/data-in-education#:~:text=Data%20analysis%20helps%20teachers%20understand,outputs%20(results%20for%20students)
26. Indeed Editorial Team (16 May 2023). How to use data analysis for marketing: https://uk.indeed.com/career-advice/career-development/data-analysis-for-marketing
27. Dr Hilde Kristin Refvik Riise. Casual blood glucose and subsequent cardiovascular disease and all-cause mortality among 159731 participants in Cohort of Norway: https://drc.bmj.com/content/9/1/e001928