Tyrone Ekhator

Introduction

According to the World Health Organization (2022) research has proven that from the years 1980 to 2014 the number of diabetic patients has gone from 108 million to 422 million; with the increase of patients mostly coming from second and third world countries. Moreover, the dataset that is being used is from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). However, the current dataset being used is limited because it only provides information on diabetic patients who are at least 21, female and of pina Indian heritage; and its objective is to predict if a person has diabetes via diagnostic measures.

# ***Methods***

Dataset

The dataset being used has 768 records; with 769 rows and 9 columns. Furthermore, the data as elaborated by the table below contains information showing the different diagnostic factors that may cause a woman of Pima Indian heritage to become diabetic (with class value 1 being equal to a patient being diabetic whilst 0 meaning no signs of diabetes).

|  |  |  |  |
| --- | --- | --- | --- |
| Field variables | Field Description | Types of variables | Structured data type |
| pregnancies | Pregnancies: Number of times pregnant | Independent variable | numeric |
| Glucose | Glucose: Plasma glucose concentration 2 hours in an oral glucose tolerance test | Independent variable | numeric |
| Blood Pressure | Blood Pressure: Diastolic blood pressure (mm Hg) | Independent variable | numeric |
| Skin Thickness | Skin Thickness: Triceps skin fold thickness (mm) | Independent variable | numeric |
| Insulin | Insulin: 2-Hour serum insulin (mu U/ml) | Independent variable | numeric |
| BMI | BMI: Body mass index (weight in kg/ (height in m) ^2) | Independent variable | numeric |
| Diabetes Pedigree Function | Diabetes Pedigree Function: Diabetes pedigree function | Independent variable | numeric |
| Age | Age: Age (years) | Independent variable | numeric |
| Outcome | Outcome: Class variable (0 or 1) | Independent variable | numeric |

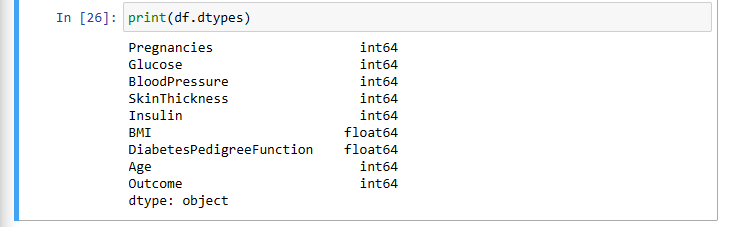
# ***Analysis and Results***

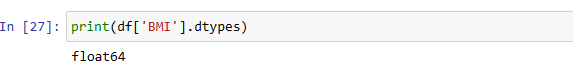
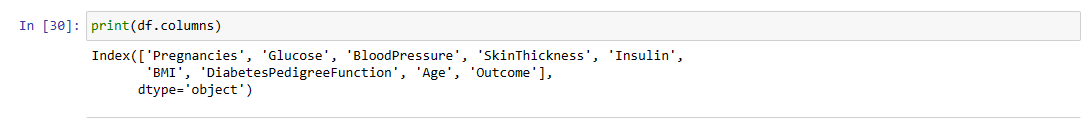
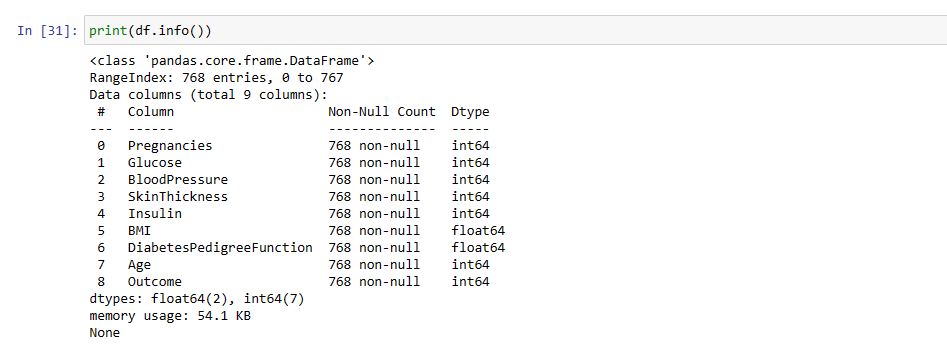
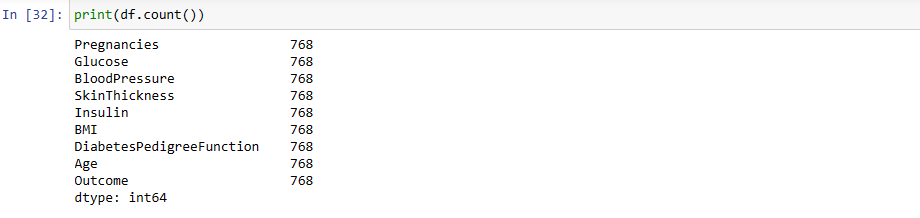
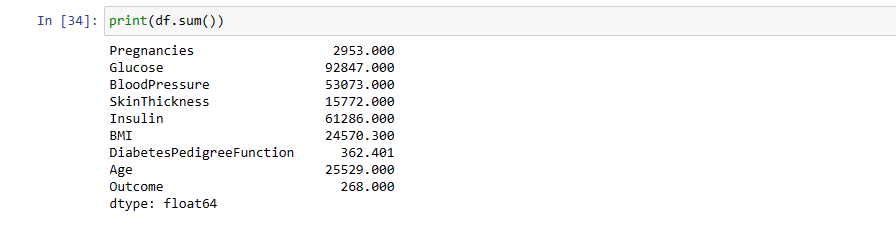
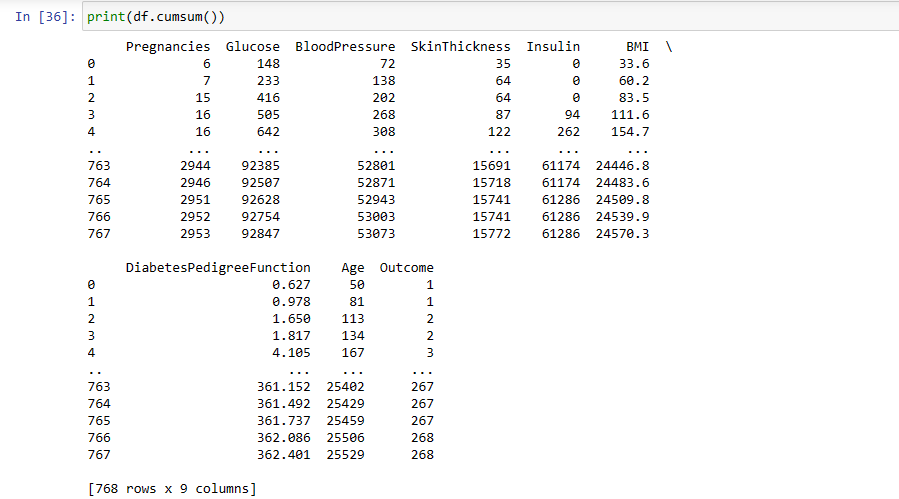
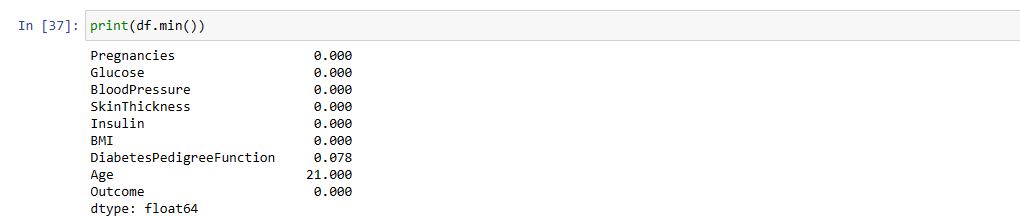
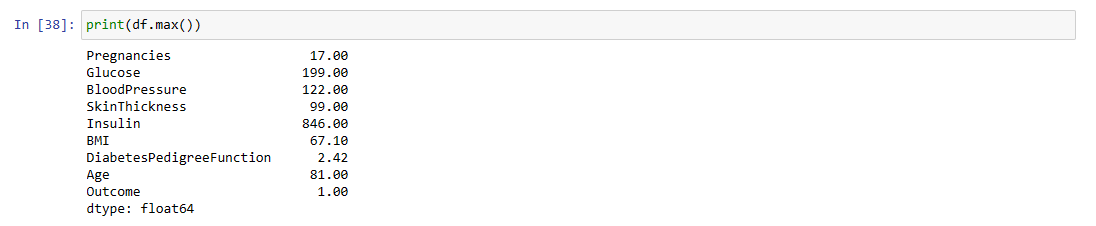
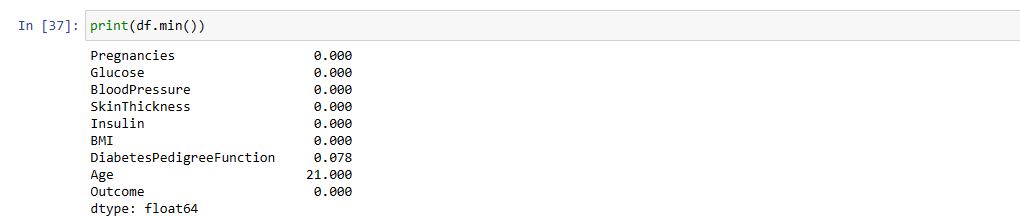
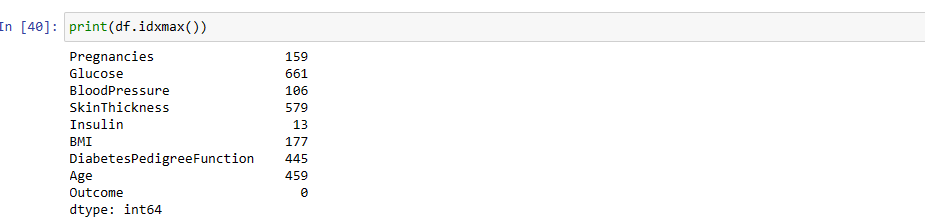
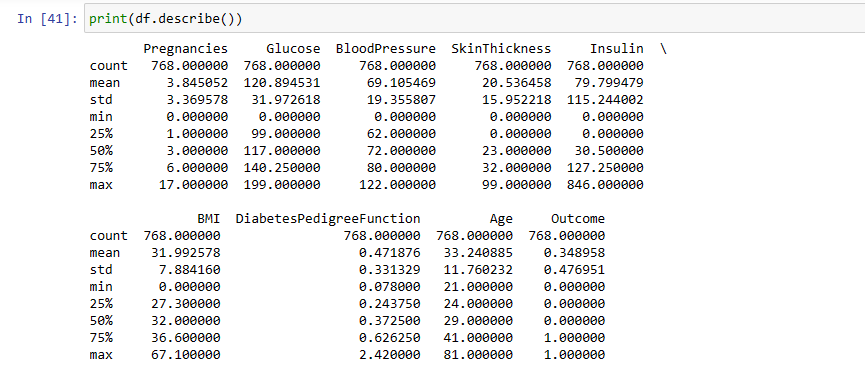
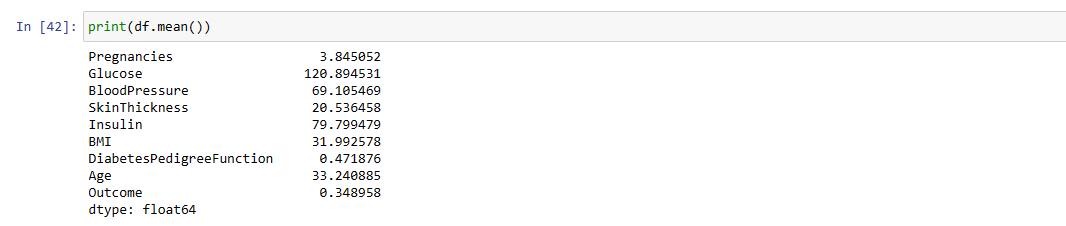
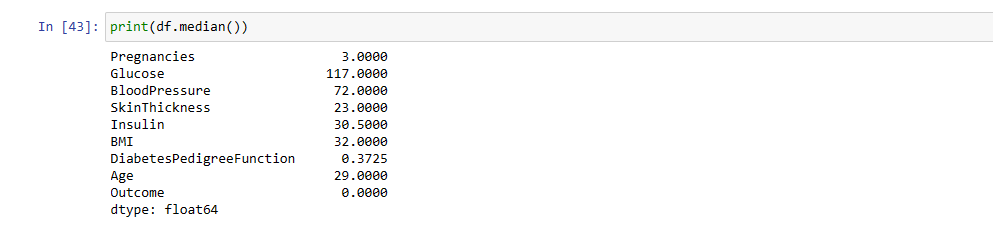
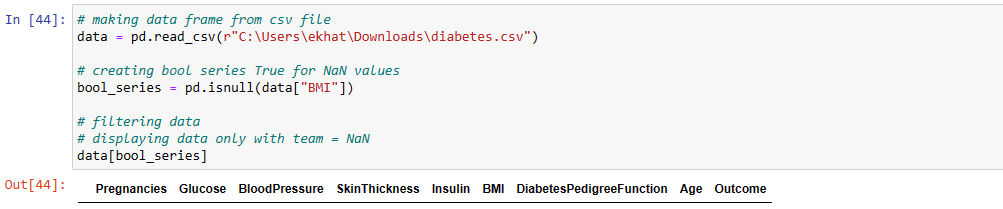
Dataset Preparation

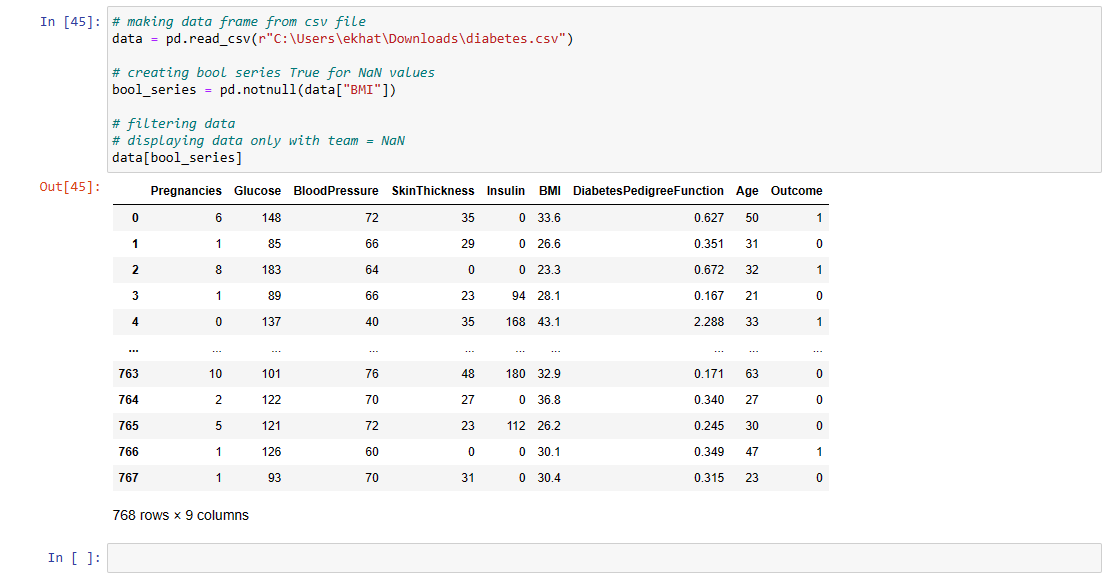
This section of the report shall be looking at dataset preparation. Prior to processing and analysis, raw data must be cleaned and transformed, which is known as data preparation. It is a crucial stage before processing and frequently entails reformatting data, fixing data, and integrating datasets to enrich data. Furthermore, dataset preparation is necessary in order to get rid of potential errors that exist from running an unedited bad dataset.

Next, this report shall be discussing the techniques used to assess the quality of the data and also checking for null values (code snippets will be added).

1. Check the data type: This is done by first importing pandas as pd; followed by importing the csv file and printing the df.dtypes. This technique is for showing the data type of each column.



1. Checking the data type of a specific column:
2. Check the shape: the code below shows that the dataset has 768 rows and 9 columns.
3. Df.index:
4. Df.columns: allows to display all the columns 
5. Df.info: This code gives us basic information on the dataset that being the index, column names, non-Null count (number of non-NA values), Dtype(datatype).
6. Df.count: This shows the number of non-NA values in each column.
7. DF.sum: the sum of values in each column.
8. Df.cumsum: this displays the cumulative sum of each row in every column of are dataset.
9. Df.min: The shows the minimum value of each column in the dataset.
10. Df.max: Displays the maximum value of each column in the dataset.
11. Df.idxmin: shows the index number of the minimum value of each index.
12. DF.idxmax: it shows the index number of the maximum value of each column.
13. Df.describe: this function displays the count, mean(average), standard deviation, min value, lower percentile (25%), middle percentile (50%), upper percentile (75%) and max value of every column in the dataset.
14. Df.mean: shows the average value of each column.
15. Df.median: shows the middle number (middle percentile – 50%) of each column.
16. Isnull(): aids in showing all the columns that have null values. Likewise, the code written displays an empty table meaning that there are no null values in the dataset. 
17. Notnull(): aids in showing all the columns that have no null values which ends up displaying a full dataset.

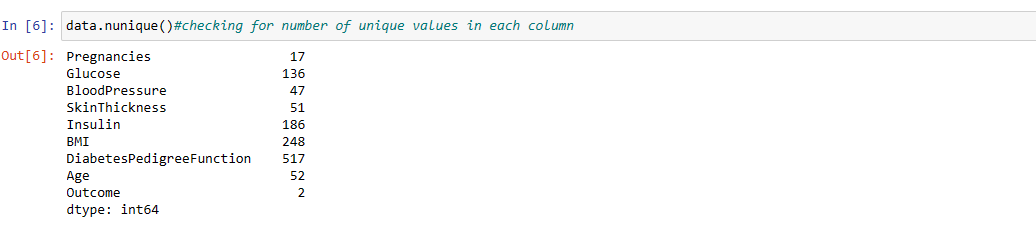
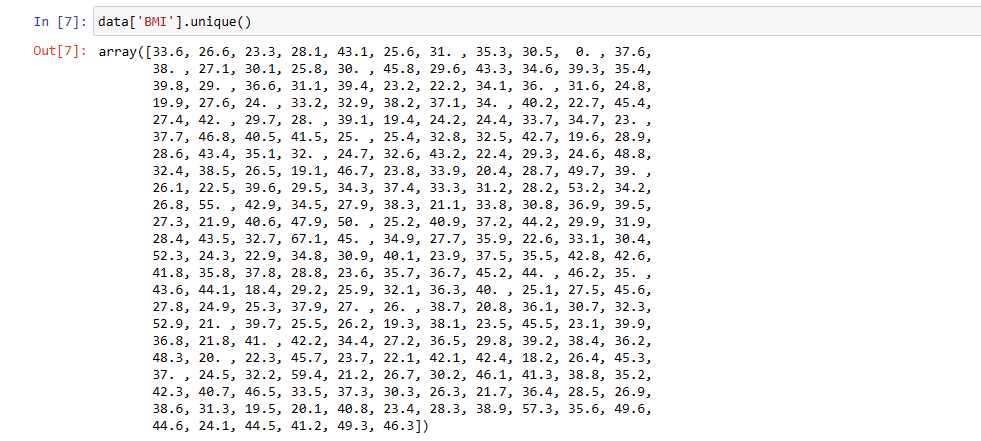
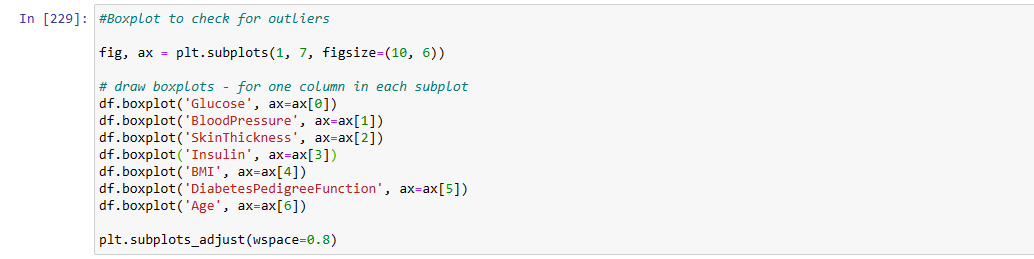


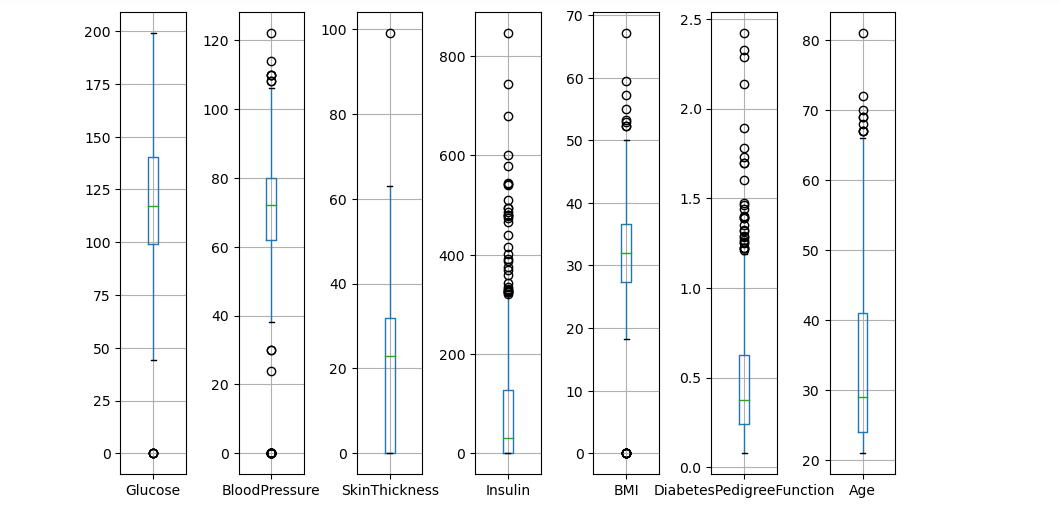
Lastly, the dataset used is actually a subset of a larger dataset gotten from National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK).

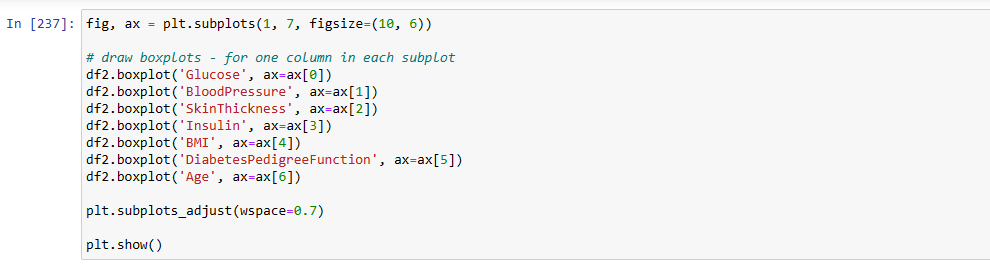
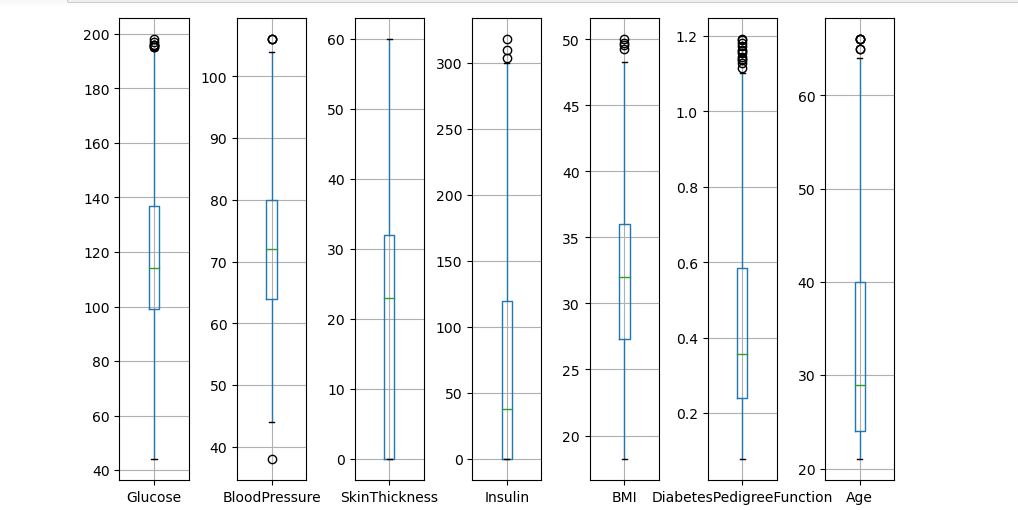
Exploratory Data Analysis

Now the question is what is Exploratory Data Analysis? Exploratory data analysis (EDA) can be defined as a data exploration technique to understand various aspects of data(edureka! ,2020). Furthermore, data scientists utilise exploratory data analysis (EDA), which frequently makes use of data visualisation techniques, to examine and analyse data sets and summarize their key properties. It makes it simpler for data scientists to find trends, identify outliers or test hypotheses by determining how to modify data sources to achieve the answers they desire (IBM,2022).

However, before proper visualisation techniques are used the programmer must check for unique values and outliers.

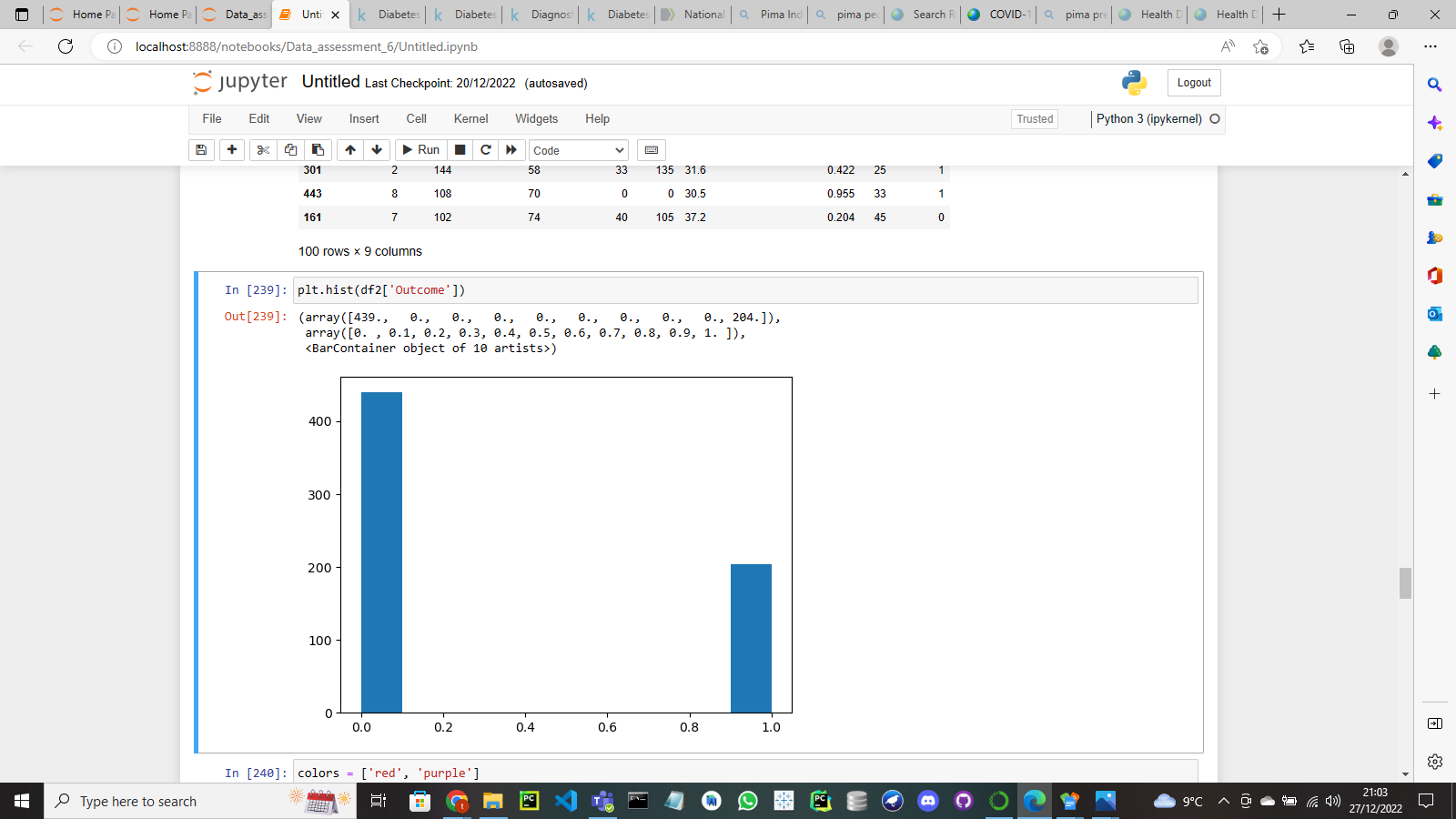
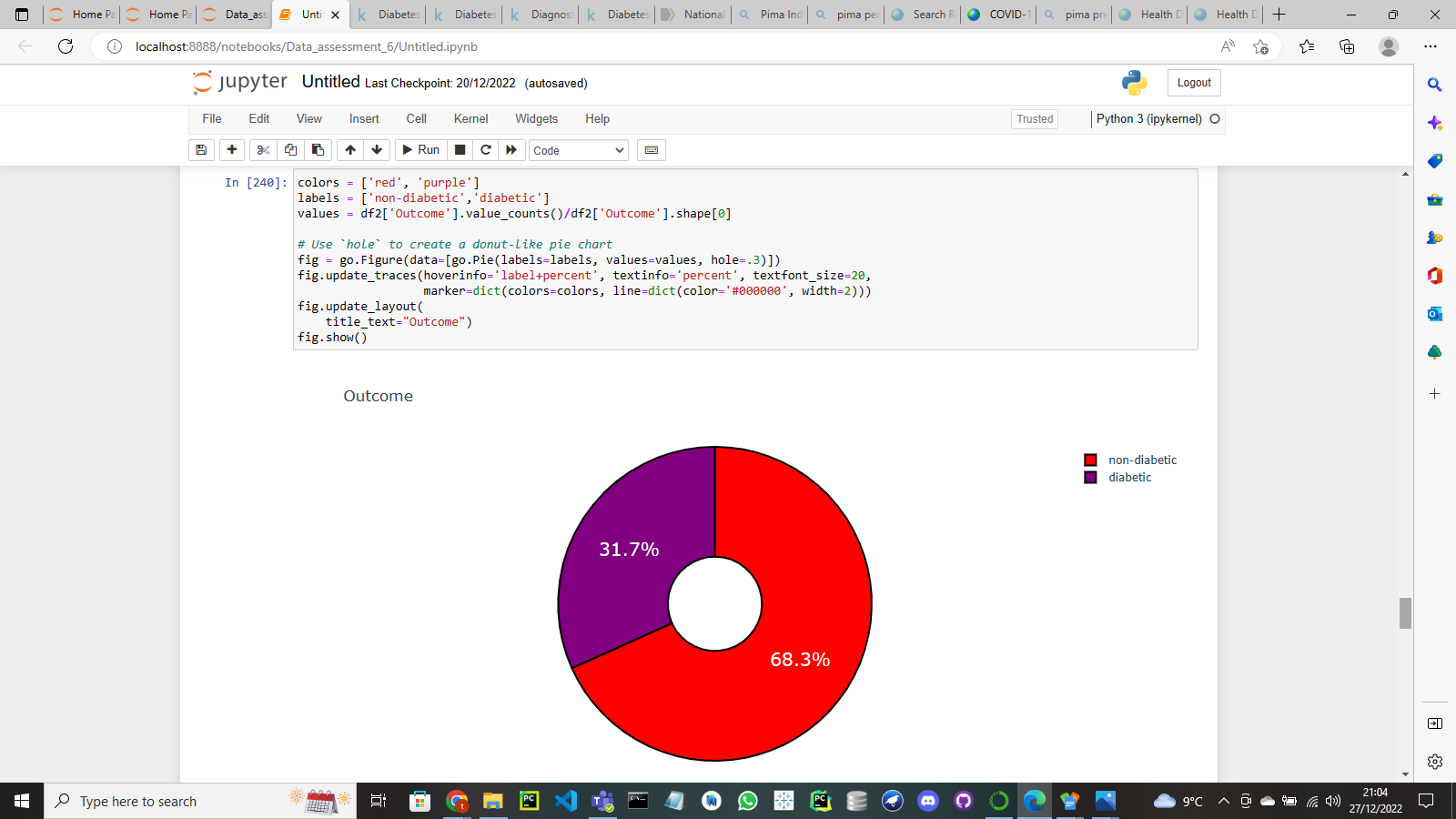
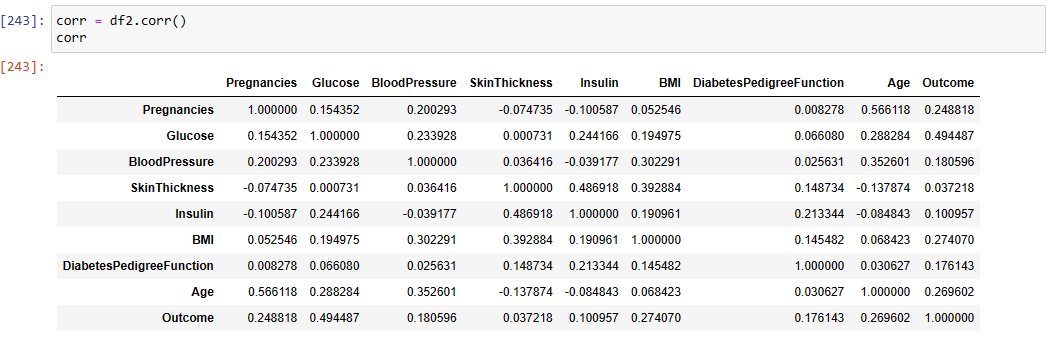
1. Unique values: Theses are the non-redundant values that exist in our dataset. Furthermore, I have also checked for the number of unique values that exist in ever column and have also written a line code that allows one to see the unique values in a specific column.
   1. 
   2. 
2. Outliers: Outliers can be defined as datapoints that differs vastly form other datapoints; and the best way to check for outliers is by plotting a boxplot for each column.
   1. The boxplots below show us the outliers in the dataset that really affect the accuracy of the EDAs and models that shall later be done in the report.

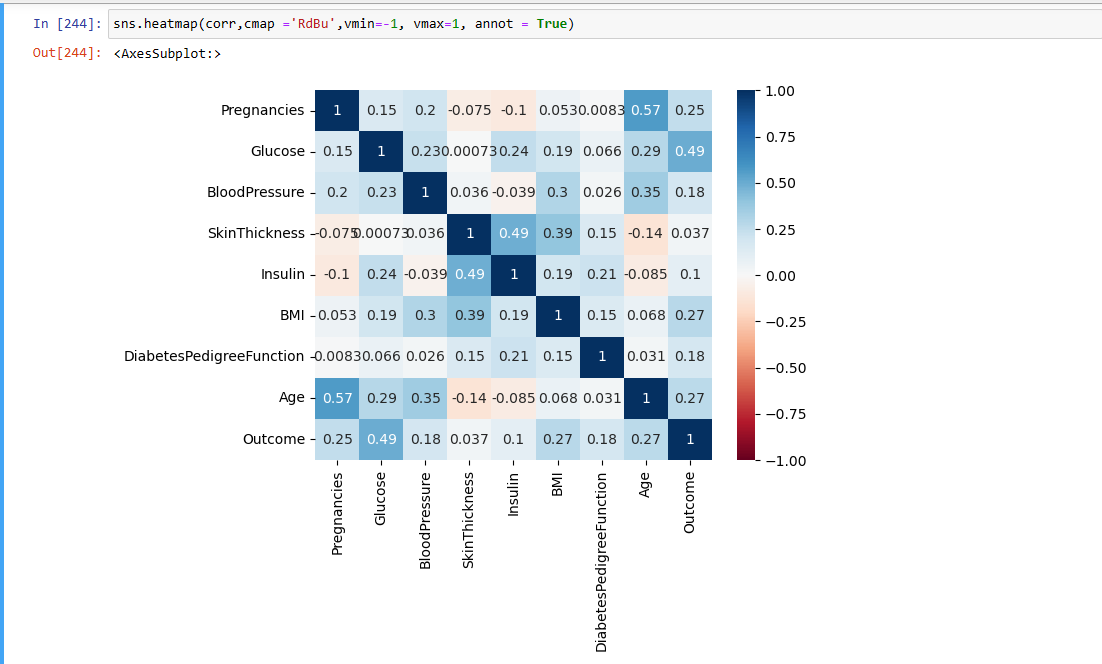


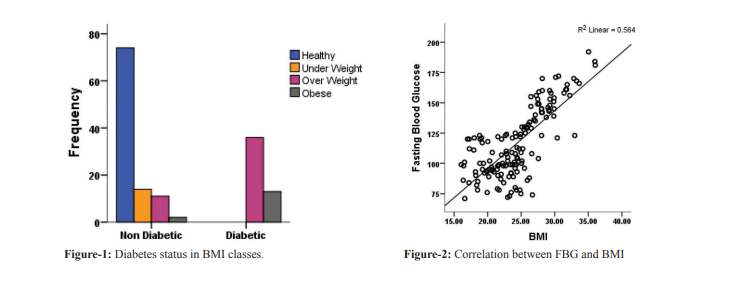
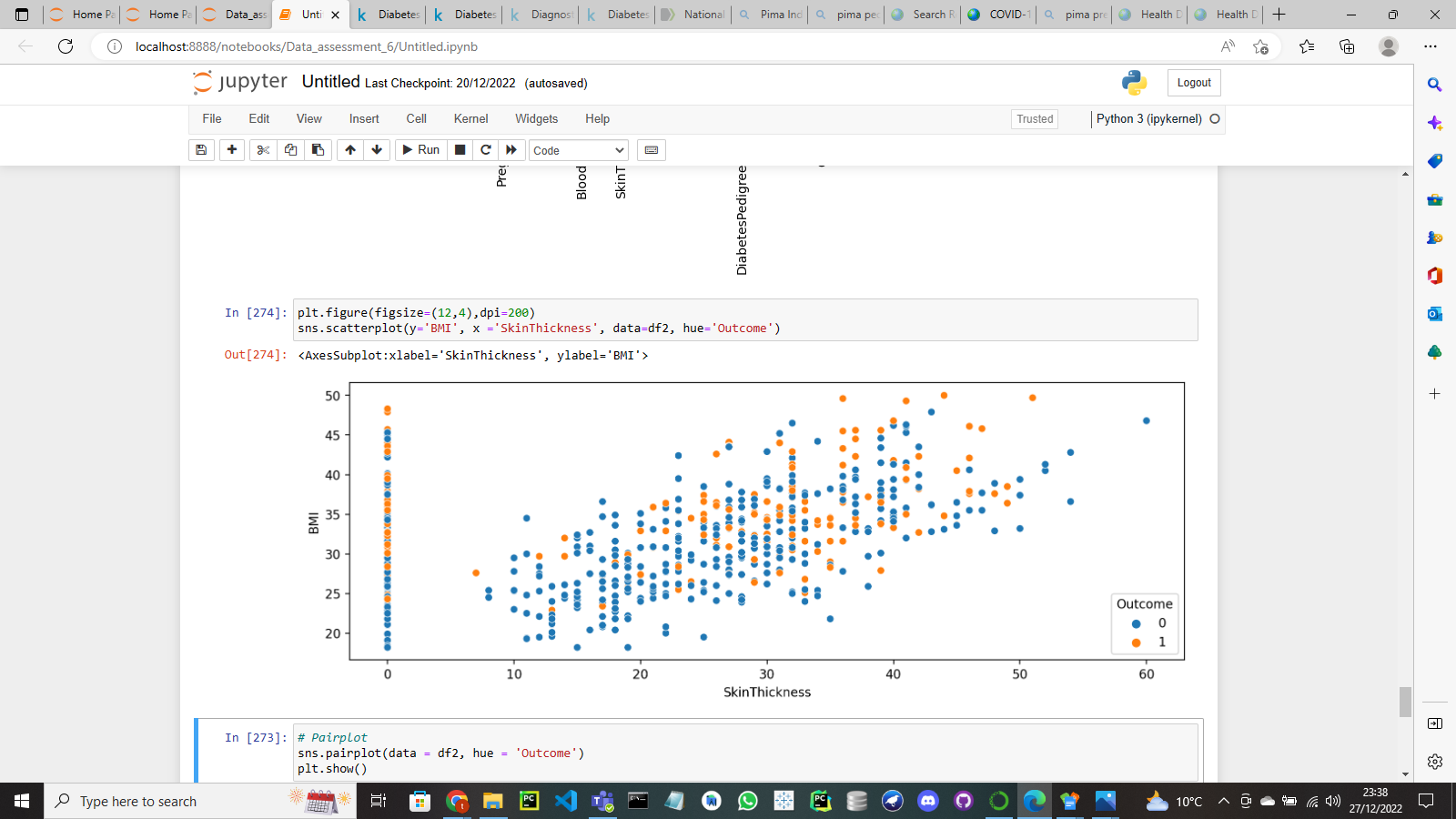
* 1. The next step is to remove as many outliers as possible. The code snippet is placed below:
     1. This basically removes all values less than the minimum and more than the maximum.
     2. Code snippet for new box plots: As one can see below there are still outliers but not to the same degree as before which will improve the accuracy of our models and analyses. 

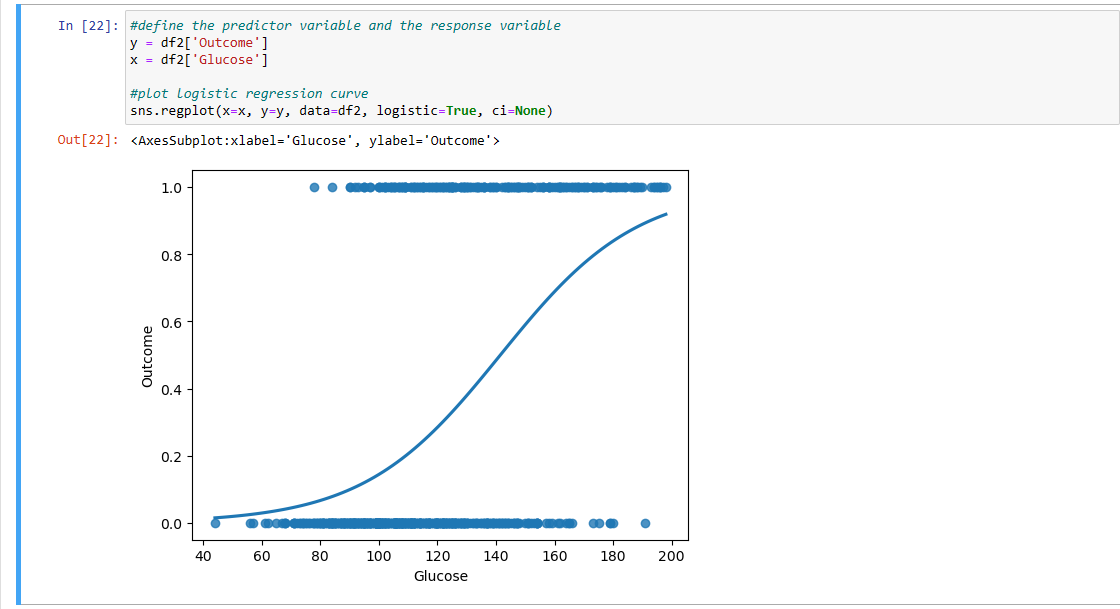
Now that the dataset is setup, I shall create EDAs based on type of EDA which shall be explained below:

Types of EDA

1. Univariate graphical EDA: Univariate analysis is the simplest form of analysis. This is because this analysis only deals with one variable and hence doesn't focus on relationships but on simply summarising and finding patterns in the data. Additionally, this report shall contain different univariate graphs which shall be explained below:
   1. Histogram: the histogram shows that the data has more non-diabetic women compared to diabetic. 
   2. Pie chart: The pie chart shows the percentage ratio between diabetic and non-diabetic women. Moreover, the pie chart below was gotten from another researcher whom shall be in the reference list.
2. Multivariant analysis: This can simply be defined as the relationship between two or more variables; and the analysis being discussed shall be used for the dataset in order see the trends that exist within it. Additionally, real world evidence (for instance- data from WHO) shall be used to back up the trends that exist within the dataset. Lastly, the two main multivariant analyses that are going to be used are listed below:
   1. Heatmap: a representation of data in the form of a map or diagram in which data values are represented as colours (Oxord,2022). Moreover, the heatmap allows the viewer to see trends and the correlation between different variables.



* + 1. Age- Number of times pregnant: The heat map shows a strong relationship between age and number of times a woman gets pregnant. Additionally, according to Pima County (2022) it has been recommended for Pima women to take the COVID-19 vaccine to protect their pregnancy.
    2. Outcome – Glucose (plasma glucose concentration): Outcome is the likeliness of a woman being diabetic whilst glucose level is the amount of blood sugar a human has. Furthermore, the blood sugar level is gotten by taking an oral glucose tolerance test which aids the doctor in diagnosing for Diabetes and checking insulin levels (Diabetics.co.uk,2022).
    3. Outcome – BMI: A pattern exists between these two variables. Research has proven that type-2 diabetes is caused by massive weight gain (BMI over 30: obesity) and physical inactivity. (WHO,2022) (CDC,2022).
    4. Glucose – BMI: A slight correlation exist between a person's BMI and blood sugar levels. For instance, a study on elderly people’s bacterial divisions in their intestine showed that there was positive correlation between glucose levels and BMI (Sepp, E. et al 2014). Likewise, another study was done in India via a blood test where the serum was separated from the blood cells found that patients (150 subjects, comprising of 103 females and 47 males.) who had a BMI over 24.9 and blood sugar levels above 126 mg/dl were found to be diabetic. The diagrams below are not mine but from a fellow researcher:However as discussed by the researcher Asians to tend to be more prone to get diabetes even if their BMI is in the overweight category which basically explains that race is also factor to consider when looking at data relating to diabetes (Neelam et al.2017).
    5. BMI – Triceps skin fold thickness: When one observes the heat map there is an explicit correlation between BMI and Triceps skin fold thickness. Reason being while BMI can give a rough estimation of whether a person is overweight or not; the skin fold test aids the researcher in getting the exact body fat percentage of an individual (which still enables the scientist to classify their patients via a body fat percentage chart). For example, a study was done from 1985 to 2003 where by athletic adolescents (33,896 in total; with their ages ranging from 11-19 years old with average age being 15) were checked for obesity via the skinfold test and also the BMI test in order to observe which test is more efficient in detecting obesity. Now, after the research was done it was discovered when checking for obesity using the BMI scale about 13.31% of adolescent athletes were obese and those classified as such approximately 62% of them were seen as a false positive-meaning they were classified as obese even though they were not- by the skinfold test. However, the data also explains that athletes whom were not in the obese category by checking the BMI scale had a 99% probability of not being categorized as obese via the skinfold method. In summary, the researcher concluded that the BMI is less accurate than the skinfold method for checking obesity; reason being the BMI had difficulty in differentiating body fat and lean body mass ergo increasing the likelihood of a person being categorized as obese (Etchison, W. C. et al. 2011).
    6. Insulin – Glucose: Reason for relationship is because the fatty acids that make up bodily fat are produced by glucose. An increase in BMI will cause an increase in lipid production, which will increase body weight when blood glucose levels rise. Additionally, the beta cells of the islets of Langerhans in the pancreas generate insulin, which acts through a particular cell receptor of insulin sensitive cells to increase glucose absorption into the cell; and being an anabolic hormone, insulin causes the body to conserve energy and signals the production of fat. Insulin resistance rises along with BMI, raising blood glucose levels in the body as a result (Neelam et al.2017).
  1. Scatter plot: This is a graph that shows the relationship between two variables (bivariate analysis). Moreover, the scatter plot in this report will show a graphical representation of a pair of variables that we discussed earlier.
     1. BMI – Skin Thickness: The plot below shows that as a person's BMI increases their skin fold follows suit. Moreover, the data points are closely packed together because of the strong correlation that exist between the 2 variables. The graph also has hue allowing the viewer to distinguish between diabetic and non-diabetic.

1. Classification: This is the last way the dataset will be assessed. Classification is a supervised learning technique that puts new data into different categories on the basis of the training data. Those categories can be groups or classes. Furthermore, the classification model that shall be used for the dataset is the logistic regression model. This is a model used to predict binary outcome; like in the case of the dataset diabetic or non-diabetic.As seen by the chart above the data points are categorised with the ’S’ curve showing the probability of an outcome based on glucose level.

Limitations and challenges

The only challenge experienced was the dataset was limited to only females of Pima Indian heritage which does not provide a diverse enough sample size to confirm the accuracies the analyses done for the current dataset.

Conclusion

In summary, what one can get from analysing the dataset is that the probability of a person being diabetic is based on their BMI, Insulin and Glucose levels which can be determined by taking certain tests in hospitals such as the oral glucose test. Moreover, after comparing work with other researchers it was discovered that other factors such as race played in the probability of a person being diabetic.

References

IBM (2022) Exploratory *Data Analysis.* Available at: <https://www.ibm.com/topics/exploratory-data-analysis> (Accessed: 01 January 2023)

ProjectPro (2022)*Exploratory Data Analysis in Python-Stop, Drop and Explore.* Available at: <https://www.projectpro.io/article/exploratory-data-analysis-in-python-stop-drop-and-explore/427#toc-10> (Accessed: 01 January 2023)

# GeeksforGeeks(2022) *Exploratory Data Analysis (EDA) – Types and Tools.* Available at: <https://www.geeksforgeeks.org/exploratory-data-analysis-eda-types-and-tools/> (Accessed 01 January 2023)

# Statology(2022) *How to Plot a Logistic Regression Curve in Python* Available at: <https://www.statology.org/plot-logistic-regression-in-python/> (Accessed 01 January 2023)

Daibetes.co.uk (2022) ***Oral Glucose Tolerance Test*** Available at: <https://www.diabetes.co.uk/oral-glucose-tolerance-test.html> (Accessed 10 June 2022)

# Centers for Disease Control and Prevention (2022) *Assessing Your Weight*

Available at:<https://www.cdc.gov/healthyweight/assessing/index.html#:~:text=If%20your%20BMI%20is%20less,falls%20within%20the%20obese%20range>. (Accessed 30 December 2022)

WHO(2022) *Diabetes* Available at: <https://www.who.int/news-room/fact-sheets/detail/diabetes> (Accessed 30 December 2022)

Sepp, E., Kolk, H., Lõivukene, K., & Mikelsaar, M. (2014). ‘Higher blood glucose level associated with body mass index and gut microbiota in elderly people’. *Microbial ecology in health and disease*, 25 Doi: <https://doi.org/10.3402/mehd.v25.22857>

Gupta, S., & Bansal, S. (2020). ‘Does a rise in BMI cause an increased risk of diabetes?’, *Evidence from India. PloS one*, 15(4), Doi: <https://doi.org/10.1371/journal.pone.0229716>

Freedman, D. S., Ogden, C. L., Blanck, H. M., Borrud, L. G., & Dietz, W. H. (2013). ‘The abilities of body mass index and skinfold thicknesses to identify children with low or elevated levels of dual-energy X-ray absorptiometry-determined body fatness’. *The Journal of pediatrics*, *163*(1), 160–6.e1.doi: <https://doi.org/10.1016/j.jpeds.2012.12.093>

# Etchison, W. C., Bloodgood, E. A., Minton, C. P., Thompson, N. J., Collins, M. A., Hunter, S. C., & Dai, H. (2011). ‘Body mass index and percentage of body fat as indicators for obesity in an adolescent athletic population’. *Sports health*, *3*(3), 249–252. doi:<https://doi.org/10.1177/1941738111404655>

# Liddle, K., O'Callaghan, M., Mamun, A., Najman, J., & Williams, G. (2012). ‘Comparison of body mass index and triceps skinfold at 5 years and young adult body mass index, waist circumference and blood pressure.’ *Journal of paediatrics and child health*, 48(5), 424–429. Doi: <https://doi.org/10.1111/j.1440-1754.2011.02247.x>

# Agrawal et al. (2017) ‘Correlation between Body Mass Index and Blood Glucose Levels in Jharkhand Population’, *International Journal of Contemporary Medical Research,*4(8), 1633-1636.doi: <https://www.ijcmr.com/uploads/7/7/4/6/77464738/ijcmr_1592.pdf>

# Starmer(2018)- **StatQuest: Logistic Regression. Available at:** [**https://youtu.be/yIYKR4sgzI8?list=TLPQMDEwMTIwMjNadwro5ShJRQ**](https://youtu.be/yIYKR4sgzI8?list=TLPQMDEwMTIwMjNadwro5ShJRQ) **(Accessed 30 December 2022)**

# McDonald (2021)- **Matplotlib Boxplots | Creating Single and Multiple Boxplots in Python. Available at:** [**https://youtu.be/IpMhMUcmLJw?list=TLPQMDEwMTIwMjNadwro5ShJRQ&t=81**](https://youtu.be/IpMhMUcmLJw?list=TLPQMDEwMTIwMjNadwro5ShJRQ&t=81) **(Accessed 01 January 2023)**

# Eigen B from HomeWorkers(2021)- **How to remove outliers in Python? | For multiple columns | Step by step ♥ . Available at:** [**https://youtu.be/Vc4cXIAa69Y?list=TLPQMDEwMTIwMjNadwro5ShJRQ**](https://youtu.be/Vc4cXIAa69Y?list=TLPQMDEwMTIwMjNadwro5ShJRQ) **(Accessed 01 January 2023)**

Sharma(2022)- **scatter plot using seaborn, matplotlib and pandas | python. Available at:** [**https://youtu.be/E4hjPFpDlCk?list=TLPQMDEwMTIwMjNadwro5ShJRQ**](https://youtu.be/E4hjPFpDlCk?list=TLPQMDEwMTIwMjNadwro5ShJRQ) **(Accessed 01 January 2023)**

# **Kanthety(2021)- SCATTER PLOT IN MATPLOTLIB || MATPLOTLIB || PYTHON PROGRAMMING.**

Available at: <https://youtu.be/PcDKxsMCpx8?list=TLPQMDEwMTIwMjNadwro5ShJRQ> **(Accessed 01 January 2023)**

**McDonald (2022)- Seaborn Heatmap - How to Visualise Correlations and Data With Heatmaps in Python. Available at:** [**https://youtu.be/J7cd1-g1O7A?list=TLPQMDEwMTIwMjNadwro5ShJRQ**](https://youtu.be/J7cd1-g1O7A?list=TLPQMDEwMTIwMjNadwro5ShJRQ) **(Accessed 01 January 2023)**

**LadyP(2018)- Introduction to Univariate Analysis. Available at:** [**https://youtu.be/WFMfNLUyG8E?list=TLPQMDEwMTIwMjNadwro5ShJRQ**](https://youtu.be/WFMfNLUyG8E?list=TLPQMDEwMTIwMjNadwro5ShJRQ) **(Accessed 01 January 2023)**

# **edureka!(2020)- Exploratory Data Analysis (EDA) Using Python | Python Data Analysis | Python Training | Edureka. Available at:** [**https://youtu.be/-o3AxdVcUtQ?list=TLPQMDEwMTIwMjNadwro5ShJRQ**](https://youtu.be/-o3AxdVcUtQ?list=TLPQMDEwMTIwMjNadwro5ShJRQ) **(Accessed 01 January 2023)**

# **codebasics(2020)- Outlier detection and removal using percentile | Feature engineering tutorial python # 2. Available at:** [**https://youtu.be/7sJaRHF03K8?list=TLPQMDEwMTIwMjNadwro5ShJRQ**](https://youtu.be/7sJaRHF03K8?list=TLPQMDEwMTIwMjNadwro5ShJRQ) **(Accessed 01 January 2023)**