Higher Diploma in Data Analytics for Business

Strategic Thinking

Report 3

23rd April 2021

Lecturers:

Mark Morrisey

Graham Glanville

Student:

Dean Watters / SBA20144

Contents Page

Resources

Anaconda Jupyter Notebook Python

- Numpy package
- Pandas package
- Matplotlib package
- Seaborn package

Data Camp

World Health Organization

NHS

Irish Heart Foundation

British Heart Foundation

Additional Research

Note

All coding for this project has been completed through Anaconda Jupyter Notebook in which all the following reports refer to information and insights gained through the attached Python Notebook files.

Research Objectives & Goals

Align research with dataset analysis and exploration aiming to extract interesting and useful trends in relation to cardiovascular diseases.

Build a machine learning model to analyse and test its performance on the specific domain problem. The model of choice is the Logistic Regression Model algorithm, and the domain specific problem is cardiovascular diseases.

Analyse, interpret, and record performance metrics of the Logistical Regression algorithm providing sufficient evidence whether it may or may not be a good model choice for the development, production, and deployment of a large-scale machine learning model to deliver on cardiovascular disease classification and prediction at a regional and national levels in the Republic of Ireland.

Design and develop blue-print national plans for a large-scale database system focusing on feature selection related to cardiovascular diseases. Findings will be presented to the Health Service Executive for analysis and assessment. This will assist in the reshaping of healthcare provision and the development of an updated strategy focused on preventative rather than curative measures at local, regional, and national levels in relation to cardiovascular diseases.

Conducting research on *The Danish digital healthcare model* to highlight the impactful and positive results being achieved in Denmark.

Build argumentative evidence highlighting the need for revolutionary changes required to use existing health data universally to better improve healthcare in the Republic of Ireland. We will be using *The Open Health Data Governance Strategy 2016 and the Open Data Strategy 2017 -2022* as the foundations for this argument.

Research Fundamentals

Following the *CRISP DM Framework to help achieve organized and well formatted research meeting project objectives and gaols.*

Complete a good standard of research that is reproducible in the fields of cardiovascular disease and machine learning model research.

Due the contextual background of this project being disease classification completing truthful and ethical research is a crucial foundation influencing credible and reproducible research in the simultaneous fields of machine learning and disease classification research.

An ability to alter, adapt and readapt when required acting within the best interests of project objectives and goals.

Business Understanding

Once again initial objectives and end goals outlined in previous reports for this project have been readjusted according to the further investigation, research, and findings. Described below are re-adjusted goals of this project along with challenges that lead to a newly developed approach consistent with research fundamentals and goals.

Initial project goals and plans targeted providing findings generated from the application of an optimized trained and tested machine learning model using data engendered within the Republic of Ireland. This model was to be presented to the Health Service Executive for the reshaping of healthcare provision and the development of an updated strategy focusing on preventative rather than curative measures at regional and national levels in relation to cardiovascular diseases. Hopes were that the HSE would move forward with the implementation and deployment of this model at a national level following legislative and regulatory guidelines to be put in place.

However, one of the main challenges included finding a large scale publicly available dataset engendered within the Republic of Ireland specifically related to cardiovascular diseases. This task was unsuccessful as outlined in the coming Data Collection Report. Follow up research sourced the Open Health Data Governance Strategy 2016 which reveals existing challenges that evidently support the challenges encountered in this project. This includes factors such as an abundance of data collated by the HSE which could be safely published as Open Health Data is not currently being made available. Data is being published in pdf and word format, or web applications making it difficult to access for processing and analysis. There is currently not an established role for Open Data management within the Health system. According to this report there are often poor data maintenance and standardisation techniques in place. It was clear that project goals at this stage were disrupted. The Open Data Strategy 2017 – 2022 advocates publicising high value government data for free public use stating this as one of their core objectives in this strategy. However, our findings as outlined in the data collection report suggest otherwise in relation to cardiovascular diseases.

From a research fundaments perspective this hindered the project enormously as how could we make statistical inferences from cardiovascular disease data, not engendered within the Republic of Ireland and use for reformational health strategies related to cardiovascular diseases in Ireland. According to the Irish Heart Foundation (2021) cardiovascular diseases is among the leading causes of death in Ireland. The World Health Organization (2020) informs us that socio-economic factors such as poverty is correlated to an increased chance of developing cardiovascular diseases and that approximately three quarters of world deaths related to cardiovascular diseases occur in low- and middle-income countries. Subsequently, any findings or inferences derived from data not engendered within the Republic of Ireland would lack accuracy, generalize, and remove known underlying determinants associated with cardiovascular disease such as social, economic, cultural change-globalization, urbanization, and population ageing (Fuster & Kelly (2010)). To proceed with a dataset not in align with these set parameters and hope to achieve accurate, impactful results would be irrational and unethical as the contextual background of this project is disease classification. There can be minimal error in the design and development of a machine learning predicational disease model.

Following this understanding and despite strategic alterations implemented to broaden the search scope, no viable results on this aspect were yielded. Details once again are provided in the coming Data Collection Report.

Consequently, and urgently followed was a statutory review meeting which changed the landscape and protocol of this project. Ultimately and unanimously a decision was made by project directors largely influenced by one of the contextual fundaments of this research piece which is to produce high standard reproducible research for future study in the field of disease classification. Therefore,

shifting the focus and emphasis from the building of a finalized machine learning disease predicational model for deployment, rather to the design and development of blue-print national plans of a large-scale database system focusing on feature selection related to cardiovascular diseases which will assist in the reshaping of healthcare provision and the development of an updated strategy focusing on preventative rather than curative measures at regional and national levels in relation to cardiovascular diseases.
Furthermore, the broader scope of this project hopes to influence the design, development, and implementation of a large-scale healthcare - database system based on the Danish digital healthcare model to be implemented at local, regional and national levels. Research is currently being conducted on the Danish healthcare model which systematically involves a large-scale protected network of population based medical databases, which routinely collect data which is consistent of high – quality data collected as a by-product of health care provision (Shmidt t al (2019)). We will outline in the next section the approach and methodology being employed to deliver on the readjusted goals for this project.

Methodology

After researching this area extensively, as anticipated and approximated from prior knowledge working with machine learning algorithms, an evidence based, and informed decision has been made to proceed with a supervised learning algorithm. The algorithm of choice will be the Logistic Regression Model aiming to solve this domain specific problem. We will be utilizing this one specific model of choice, rather than experimentation with alternating algorithms. Below we will outline the ideology, advantages and disadvantages behind this approach which are in align with project objectives, goals, and research fundamentals.

According to Park (2019) the Logistic Regression Model is a good choice when the domain problem is predicting a dichotomous outcome or probable event. According to Peng et al (2002) a logistic regression model is a good machine learning model choice for describing the effect of various categorical and continuous explanatory variables on the probability of occurrences of a dichotomous outcome variable. Royston & Altman (2010) suggests the logistic regression models are commonly used in medicine for patient prediction.

It is felt that a more robust, intensive, and specific study can be complete by adopting this approach as it will create a more in-depth environment to explore and analyse the mechanics, design and statistical inferences produced by the specific model of choice.

This meets one of the main research fundaments which is to produce a high standard research piece which is reproducible for future simultaneous studies in the fields of disease classification and machine learning. The approximated outcome will result in the foundational testing for a large-scale prediction model that will tackle the goal of reshaping healthcare provision and influence the development of an updated strategy focusing on preventative rather than curative measures at regional and national levels in relation to cardiovascular diseases.

It is also important to understand when performing any form of research, it is fundamentally imperative to consider during research phases all consequential outcomes that may arise from decisions and actions. Therefore, outlining a key disadvantage of the approach being adopted above is that focusing on one specific algorithm for this domain specific problem ultimately may be neglecting enhanced and optimal results to be achieved using other reputable algorithms. After consideration and to challenge this concept, the general conception was a commitment to complete a robust, intensive, and specific analysis of various algorithms is not within the realms of the deadlines in place for this project and will be computationally expensive. To ignore this and continue with various alternating algorithms ultimately would depreciate and discredit the overall research being conducted.

We aim to provide evidence whether the Logistic Regression model will be an optimal algorithm choice for the readjusted goals and to answer ultimately on the successful development of the large-scale database system related to cardiovascular diseases whether this algorithm will deliver desirable and ideal results.

Data Understanding

Data Collection Report

As highlighted in the previous report several data repositories were classified into both Primary and Secondary repository resource systems, to identify appropriate cardiovascular disease datasets. Prior knowledge on dataset repositories influenced both segregations based on project goals. Below are both these sources (see Figure (1.0)) including all data repositories included for search analysis.

(Figure 1.0)

Primary Sources	**URL**
Irish Government Data Portal	https://data.gov.ie/ e – health Ireland
UK Government Data Portal	https://data.gov.uk/
UK National Health Service Data	https://digital.nhs.uk/data-and-information
EU Open Data Portal	http://data.europa.eu/euodp/en/data/

Secondary Sources	**URL**
World Bank	https://data.worldbank.org/
World Health Organisation	http://www.who.int/gho/en/
UNICEF	https://data.unicef.org/
Google Public Data Explorer	https://www.google.com/publicdata/directory
Amazon Web Services Open Data Registry	https://registry.opendata.aws/
Pew Research Datasets	http://www.pewinternet.org/datasets/
Kaggle	https://www.kaggle.com/
UCI Machine Learning Repository	https://archive.ics.uci.edu/ml/index.php
Open Data Network	https://www.opendatanetwork.com/
US Medicare Hospital Quality Data	https://data.medicare.gov/data/hospital-compare
Yelp Data	https://www.yelp.com/dataset/challenge
Broad Institute Cancer Program Data	http://portals.broadinstitute.org/cgi-bin/cancer/datasets.cgi
Centres for Disease Control and Prevention	https://www.cdc.gov/datastatistics/
DataHub	http://datahub.io

A strategic plan was developed and implemented at the beginning of this project which envisioned successfully identifying an appropriate dataset fitting to initial dataset requirements. As challenges were approximated in advance such as unsuccessfully identifying data which would potentially hinder project goals, this assumption both influenced and motivated a three seeding setting to be incorporated to counter any challenging factors that may arise. This process includes three separate seedings created for search analysis. Details are outlined below. The first seeding is referred to as *Seeding One CVs* see (Figure (1.1)). This included a thorough search analysis of all repositories included within this seeding.

Figure (1.1)

SEEDING CVs One (Dataset Requirements)
Type: Qualitative/Quantitative Data (Quality Focus High)
Dataset Origins: Republic of Ireland Repositories / UK Repositories
Dataset Size: 20,000 (+) observations/10 columns (min)
Timeframe: 01/01/90 – 01/12/20

However, this yielded unsatisfactory results unsuccessfully identifying cardiovascular disease specific and relational datasets originating in the Republic of Ireland & UK. Concluding Seeding One CVS search analysis was no large scale publicly available datasets linked to cardiovascular disease from Irish / UK repositories were currently available. Furthermore, this prompted project parameters to be extended (Seeding CVs Two) which included broadening the scope of the project from Irish and UK data repository resource systems to EU repository resource systems. This re-adjustment can be seen below (Figure (1.2)).

Figure (1.2)

SEEDING CVs Two
Type: Qualitative/Quantitative Data (Quality Focus High)
Dataset Origins: Republic of Ireland Repositories/ UK Repositories / EU Repositories
Dataset Size: 10,000 (+) observations/10 columns (min)
Timeframe: 01/01/90 – 01/12/20

Primary Sources such as Irish, UK and EU data repository resource systems included an extensive search in the hopes of identifying datasets which meet the requirements set in advance of the Seeding CVs Two search analysis. Although in this seeding the parameters and scope has been broadened to EU repository resource systems this was not considered to be a substantial change which could hinder project end goals. As already mentioned in the data understanding report (see page 02) The World Health Organization states that socio-economic factors such as poverty increases the chance of developing cardiovascular disease. Ireland is a member of the EU, in which the UK have opted to depart from the EU through a process known as Brexit. However, the UK is geographically located close to Ireland more so than other EU neighbouring countries. Therefore, the broadening of these parameters does not deviate far apart in terms of geographical location and economic stability and will have little impact on the application and deployment of an optimized model for this specific domain problem.

Identified was a small-scale dataset approximately 350 observations reported in the previous report, which despite not meeting quality focus standards, it did hold interesting information which will act as the building foundations in this project. This assisted in the identification of specific markers which are considered as key contributors to cardiovascular diseases (see page 10 – 11). This is an important feature of this project. We will touch on this in later reports.
Concluding Seeding CV's two search analysis despite identifying a dataset holding valuable data, on the contrary this search analysis was unsuccessful. It did not meet the size (10,000 observations) within the dataset requirements set for the project.

Furthermore, as suitable datasets could not be identified within Irish, UK and EU repositories this meant the landscape of the project goals were totally disrupted which required a review meeting to analyse the project foundations and goals. As highlighted previously in the data understanding

report these discoveries from Seeding One and Seeding Two led to the end goals of this project being readjusted. The main problem is using a dataset that did not originate in Ireland, UK or EU systems and applying machine learning algorithm prediction results to instil the platform and foundations for preventative rather than curative measures related to cardiovascular diseases in the Republic of Ireland. The research suggests that cardiovascular diseases are linked to socio economic factors as we have mentioned throughout this project as highlighted previously (see page 02). This therefore would be a direct contradiction to move forward and turn a blind eye to such known contributary factors. This influentially hindered and changed the landscape of this project. Therefore, shifting the focus and emphasis from the building of a finalized model for deployment to the design and development of blue-print national plans of a large-scale database system focusing on feature selection related to cardiovascular diseases which will assist in the reshaping of healthcare provision and the development of an updated strategy focusing on preventative rather than curative measures at regional and national levels in relation to cardiovascular diseases.

(Figure 1.3)

SEEDING CVs Three
Type: Qualitative/Quantitative Data (Quality Focus High)
Origins: Primary Sources: Republic of Ireland / UK / EU Repositories Secondary Sources: Refer to previous appendix (0.0) for list of secondary resources included
Dataset Size: 5,000 – (+) observations/10 columns (min)
Timeframe: 01/01/90 – 01/12/20

Through shifting and the expansion once again of the parameters of the project end goals as highlighted above Seedings CVs three (see figure (1.3)) and in the previous data understanding stage, a dataset related to cardiovascular diseases was identified. This met the quality standards of Seeding CVs Three search analysis. Details are outlined in the coming Data Description Report.

Concluding our overall search analysis for identifying datasets relating to cardiovascular disease is there are no datasets publicly available related to cardiovascular diseases within Irish, UK and EU repositories. Contradictory to this is there is a large magnitude of data available online publicly related to cardiovascular diseases. High standard resources include the World Health Organization, the British Heart Foundation which aided in the research foundations of this project. These offer widely available statistics and research available on this subject.

Data Description Report

Identified was a dataset related to cardiovascular diseases through seeding cv's three project requirements and standards as mentioned in the previous data collection report. The data description report will complete an in-depth analysis of the approved dataset.

Original Dataset
Observations / 5110
Instances / 12
float64(3) / int64(4) / object (5)

In total there are 5110 observations, and 12 instances present in this dataset. The 12 instances include id, gender, age, hypertension, heart disease, ever married, work type, residence type, avg glucose level, bmi, smoking status and stroke. Initial indications that some instances will require renaming in the data cleaning phase of this project. The ideology and reasoning for this is to make the dataset more comprehensible and understandable. Furthermore, we made some minor changes before commencing our full investigation and analysis into the dataset. These changes included data type conversions to achieve better insight and analysis in this report (Data Description Report). We will refer to this as the updated dataset.

Updated Dataset
Observations / 5110
Instances / 12
float64(2) / int32(1) / int64(1) / object (8)

Below we take a brief look at all object data types (8) included in the updated dataset along with relevant information they withhold required for analysis.

Gender (object data type)	*Smoking Status (object data type)*	*Work Type (object data type)*	*Ever Married (object data type)*	Residence Type (object data type)
2994 / Female 2115 / Male 1 / Other	1892 / Never smoked 1544 / Unknown 885 / Formerly smoked 789 / Smoke	2925 / Private 819 / Self – Employed 687 / Children 657 / Government Job 22 / Never Worked	3353 / Yes 1757 / No	2596 / Urban 2514 / Rural

Hypertension (object data type)	*Stroke (object data type)*	*Heart Disease (object data type)*
4612 / 0 498 / 1	4861 / 0 249 / 1	4834 / 0 276 / 1

Furthermore, we will observe all other data types.

ID (int64 data type)	*Age (int32 data type)*	*Avg_Glucose_Level (Float 64 data type)*	*BMI (Float 64 data type)*

mean – 36517.829354	mean – 43	mean – 106.147677	mean – 28.893237
standard dev – 21161.721625	standard dev – 23	standard dev – 45.283560	standard dev – 7.854062
min – 67.000000	min – 0	min – 55.120000	min – 10.300000
25% - 11741.2500000	25% - 25	25% - 77.245000	25% - 23.500000
50% - 36932.000000	50% - 45	50% - 91.885000	50% - 28.100000
75% - 54682.000000	75% - 64	75% - 114.090000	75% - 33.100000
max – 72940.000000	max - 82	max – 271.740000	Max – 97.600000

We can conclude the data description report stating that this dataset appears to hold some remarkably interesting data in relation to cardiovascular diseases. We will make further assessments on the quality of the data within the data quality report.

Next, we will complete an important aspect of this project aligning with new set goals. As it has been ruled against building a finalized machine learning model for disease classification as discussed previously, to the design and development of a large-scale database system related to cardiovascular diseases. Feature research will become a significant milestone of this project. Furthermore, this section is an accumulation of two dataset analysis which helped guide the direction of external feature research.

Research Markers

The following markers have been identified through dataset feature exploration and analysis in align with research from credible sources in the field of cardiovascular research. In this section we aim to achieve a good synopsis of relevant research on cardiovascular diseases. Feature research will be a key aspect in assisting in the designing and development of blue-print national plans for a large-scale database system focusing on feature selection related to cardiovascular diseases.

Firstly, as discussed previously the identification of a sample size dataset which comprised of 350 observations and 14 instances did not merit the allocated standards to proceed into stage three and stage four of this project. However, as earlier stated it did hold credible and valuable information regarding features associated with cardiovascular diseases in this dataset. This motivated a robust research practice to be implemented proceeding with the project. Secondly, the identification of a dataset which met project requirements as outlined in the data identification report comprised of 5110 observations and 12 instances played a role directing feature research. Features included in both these datasets provided the foundations for the direction and angle of research exploration and helped produce the following synopsis of research in relation to cardiovascular diseases.

Age

According to the NHS (2020) cardiovascular diseases is more commonly developed in individuals over the age of fifty. Furthermore, stated is that men are more likely to develop cardiovascular diseases at a younger age than women. As age increases so does an individual's chances of developing cardiovascular diseases.

Blood Pressure

According to Lin & Svetkey (2012) the term blood pressure refers to the pressure of blood in your arteries, in which the vessel's role is to carry a person's blood from the heart to the brain and the rest of the body. Furthermore, Lin & Svetkey (2012) informs us that high blood pressure is also known as hypertension. Hypertension indicates that your blood pressure is consistently too high adding pressure onto the heart. If hypertension goes undetected or untreated it can ultimately lead to heart failure, heart attack, stroke, kidney failure and other conditions with are a cause for concern.
High Blood pressure can be developed through poor lifestyle choice such as alcohol abuse, smoking, obesity and can result from other medical and medicines. Genetics and age can also play a role in high blood pressure. According to The World Health Organization (2020) people living in areas of deprivation are at an increased risk of having high blood pressure.

Obesity

According to British Heart Foundation (2020) when an individual is considered over-weight also known as obesity this can lead to many serious health conditions and increase an individual's risk of developing heart or circulatory diseases. Diet and other factors such as medicine or medical conditions can lead to obesity. Furthermore, the British Heart Foundation (2020) states visceral fat refers to the fat that surrounds an individual's internal organs. This fat affects how the hormones work and can raise blood cholesterol levels, increase blood pressure, and increase someone's chances of developing type 2 diabetes. This is intricately linked to heart and circulatory diseases. The body mass index (BMI) is a measurement using both an individual's height and weight to work out whether an individual's weight is healthy. According to NHS (2020) for most adults an ideal BMI is the 18.5 to 24.9 range. Exceptions are granted taking external factors into consideration.

Smoking

Harmful chemicals in cigarettes include carbon monoxide, tar, and nicotine. According to Irish Heart Foundation (2020) having high levels of carbon monoxide in your blood significantly increases your

risk of heart and circulatory diseases. Tar also found in cigarette smoke and can cause cancer. Nicotine increases your heart rate and blood pressure.

Cholesterol

The Irish Heart Foundation (2020) informs us that cholesterol is produced in the liver and is a fatty substance found in an individual's blood. Furthermore, high cholesterol is when an individual has too much cholesterol in the blood. This increases the chances of developing cardiovascular diseases. Cholesterol is carried in the blood by proteins and when these combines, they are known as lipoproteins. Brill (2009) informs us that cholesterol is segregated into two main types high – density lipoproteins (HDL) and non – high density lipoproteins (non – HDL). HDL known as good cholesterol, helps get rid of cholesterol regarded as bad for the individual. Non-HDL is regarded as bad cholesterol as when there is too much non -HDL present, this clogs up walls of the blood vessels causing arteries to narrow which increases an individual's chance of developing cardiovascular diseases. This is known as atherosclerosis. Furthermore, Briill (2009) states that blood contains a type of fat called Triglycerides in which being overweight, bad diet, and alcohol can make you more likely to have a high triglyceride level which also increases the chances of developing cardiovascular diseases. Causations to high levels of cholesterol can be linked to several factors including diet, smoking, lack of activity, ageing, gender, ethnic backgrounds, family history and other medical conditions such as kidney diseases.

Ethnicity

The British Heart Foundation (2020) identify ethnicity as a risk factor to developing cardiovascular diseases with plausible genetic dictation. Different ethnic minority groups are linked to a higher probability of developing cardiovascular diseases in comparison to other ethnic minority groups. There is no clear answer on this topic. Moreover, the NHS (2020) inform us in the United Kingdom, cardiovascular diseases are more common in South Asian, African, and Caribbean backgrounds.

Diabetes

NHS (2020) informs us that diabetes is a condition that can cause significant increases in blood sugar levels. High blood sugar levels can damage blood vessels and is associated with obesity, which increases an individual's chances of developing cardiovascular diseases. The British Heart Foundation (2020) also informs us that diabetes is intricately linked with obesity and a family history of type II diabetes. They also inform us that some ethnic groups currently have a much higher rate of diabetes.

Physical Inactivity

The British Heart Foundation (2020) advise us that physical activity (exercise) can decrease an individual's chances of developing cardiovascular diseases. Physical inactivity can lead to fatty material building up within arteries leading to blood vessels becoming damaged and clogged. This can lead to cardiovascular issues such as heart attack.

This is regarded as a critical part of this project supporting one of the main goals. Feature research will support in the formulation, design, and development of blue-print national plans for a large-scale database system focusing on feature selection related to cardiovascular diseases. This will assist in the reshaping of healthcare provision and the development of an updated strategy focusing on preventative rather than curative measures at regional and national levels in relation to cardiovascular diseases. We recognize these as specific markers which can be considered strong contributary factors to cardiovascular diseases and will be a key factor in formulating the conclusive proposals for feature inclusion along with measurement equipment and data collection suggestions in the concluding parts of this project.

Data Exploration Report

This section completes an in-depth dataset analysis and exploration of the dataset, aligning it with research we have completed over the course of this project. Simultaneously, we will combine some data mining findings with data visuals.

Figure (1.4)
Age Distribution

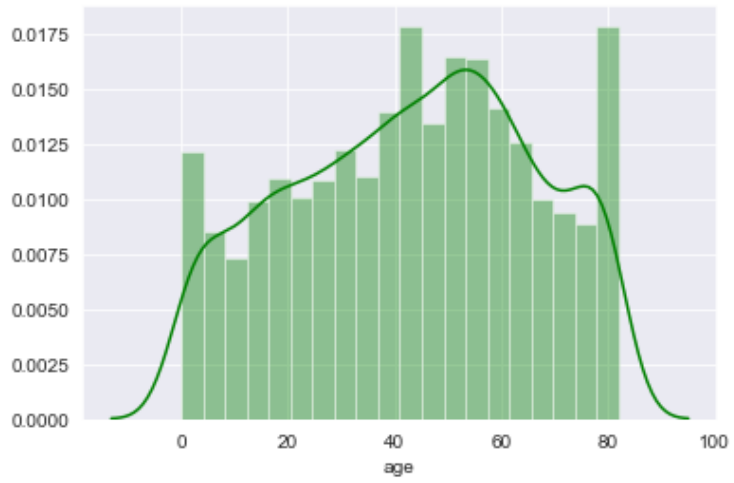


Figure (1.4) histogram/kernel density plot shows the distribution of the age variable. This is an important variable for two purposes. The age variable will be useful to assess the association between higher ranging age groups and cardiovascular diseases and other predisposing factors within this dataset. Additionally, it will also be useful to assess the usability of this dataset in relation to cardiovascular disease analysis which we will go into detail in the following Data Quality Report. From a statistical analysis perspective, we observe that the distribution is marginally right skewed. At this point we are unaware of how this will affect the Logistic Regression algorithm. However, research will be conducted to investigate and analyse if this may impact the modelling process.

Figure (1.5)
Age/Smoking Status

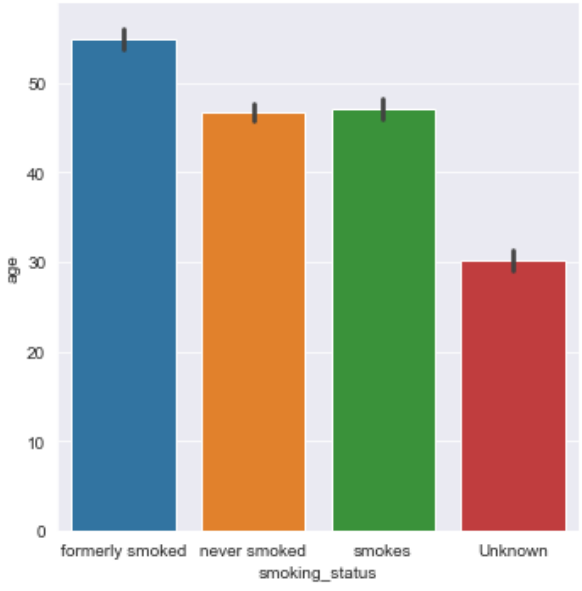
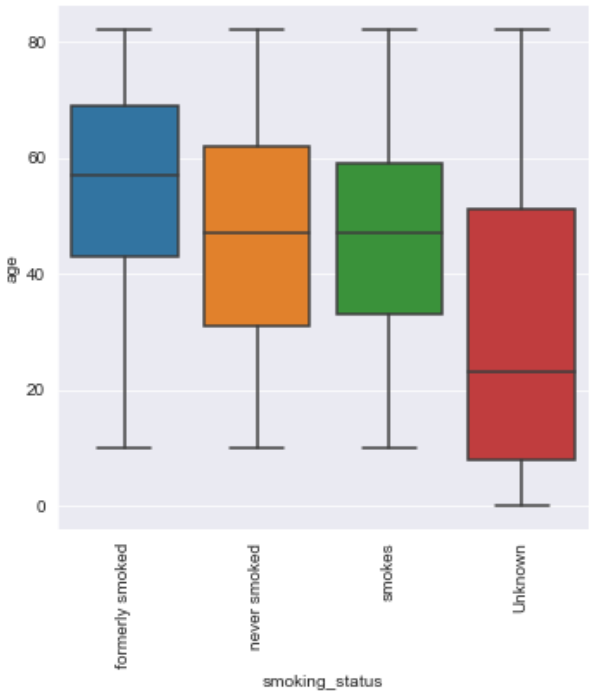


Figure (1.5) which shows a bar plot graphical analysis of smoking status against age ranges. We can see clearly that age categories range in steps of 10 years from 0 years to just below 60 years.
A technical error has been founded after the computation of this visual. The age range against smoking status in this type of plot hold little information in relation to age group and smoking status. We will investigate whether a box plot may be a better fit.
However what is useful from this plot is we can clearly observe the most prevalent and least prevalent smoking related behaviours associated with this dataset. Moreover, formerly smoked which in this case we can approximate the subject to have previously smoked is the most popular category. Furthermore, we can assume that previously smoked means the subject no longer engages in smoking behaviours. Additionally, and interestingly, we can see that both never smoked and smoked categories have similar distribution ranges.

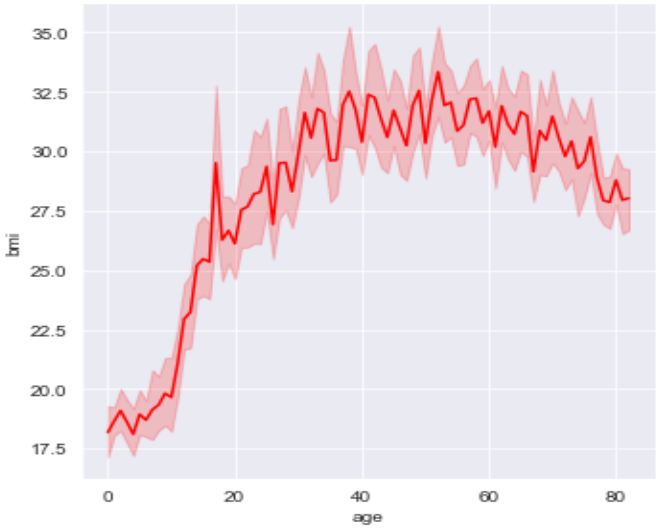
Figure (1.8)
Age/Smoking Status



As we outlined above in which the count plot does not capture the true essence of age ranges associated with smoking status of the subjects included in this dataset. We decided to compute the boxplot to analyse this information (see Figure (1.8)).
This gives much better insight on smoking habits revealing interesting information. From further data mining, interestingly, we found that formal smokers (885) largely range between 40 and 60 years. Never smoked (1892) and smokes (789) range closely together according to the age category 25 to 65. Unknown (1544) contributes to a large proportion of the subjects in this dataset.

Figure (1.9), (2.0), (2.2), (2.3) & (2.5) displays a line plot, which show more than one observation per x value. If a line plot is given multiple observations per x value it will aggregate them into a single summary measure. By default, displayed is the mean value and this is the summary measure we have determined as of interest for analysis. It is important to understand the underlying mechanics of this plot prior to interpretations and evaluations.

Figure (1.9)
Age/BMI



As approximated from research analysis, as we can see from figure (1.9), it is clear BMI on average rises as the age of subjects increase. We can see a slight fall in BMI levels ranging between 32.5 and 27.5 in relation to the 60 – 80 year age category. In the 0 – 40 year age categories there is a sharp rise in BMI level ranging between 17.5 and 32.5. Furthermore, between 0 – 40 years there is a balance in the average BMI levels of subject ranging between 30.0 and 32.5. This is in align with research conducted and shows that the data is consistent with feature research (see page 00)

Figure (2.0)
Age/Average Glucose Levels

As we see can from figure (2.0) which shows age category against average glucose levels, average glucose levels ranging between 80 – 110 remain balanced in the 0 – 40 year age category. Furthermore, within the 40 – 80 year age category we can see a clear upward trajectory in average glucose levels ranging between 100 – 140.
Furthermore, the mean value of the average glucose level is 106.14767.

Figure (2.1)
Age/Gender

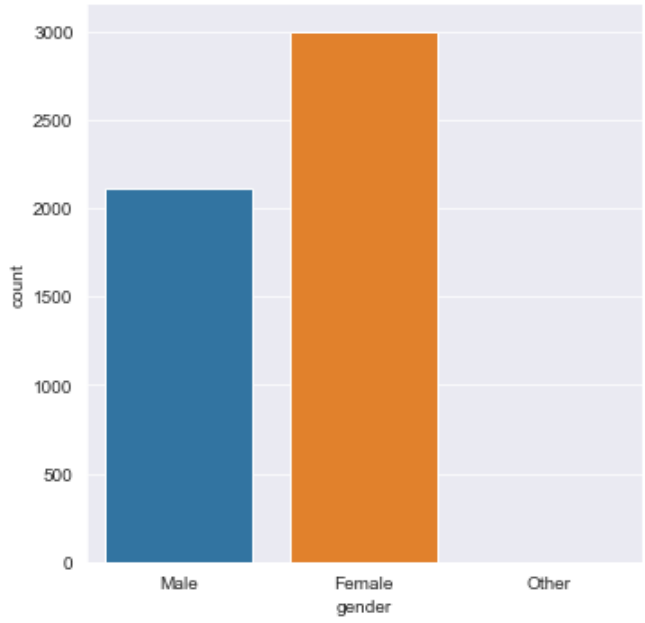


Figure (2.1) which displays a count plot clearly shows that the frequency of males in comparison to females is lower. Approximately 2994 are documented as female whereas approximately 2115 are documented as male, and 1 as other in this dataset.
Documented in the previous report which analysed and explored a smaller dataset, it was envisioned that researching gender equality in relation to cardiovascular disease would be a dominant feature of this project.
However, as we outlined in this report specifically the updated Data Understanding goals have been readjusted and this envision is no longer in align with original proposed objectives and goals.
Despite this being an interesting topic for research, deadlines are in place and focus is required on our specific goals as highlighted in the updated data understanding report.

Figure (2.2)
Age/Stroke

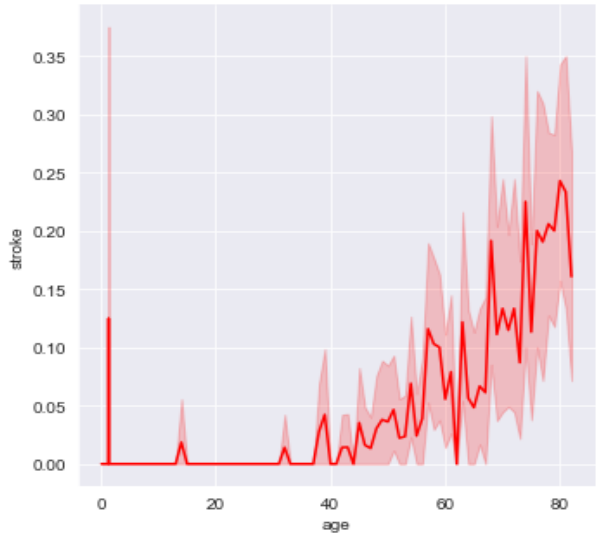


Figure (2.2) displays the average range of subjects who suffered a stroke according to age. Approximately 4861 subjects have not suffered from a stroke and approximately 249 subjects suffered from a stroke. We can see between the age range 40 – 80 years a sharp volatile increase of subjects on average who suffered from strokes. Prior to this subject's age ranging between 0 – 40 years shows on average this phenomenon is uncommon. This is an interesting instance in relation to cardiovascular diseases.

Figure (2.3)
Age/Heart Disease

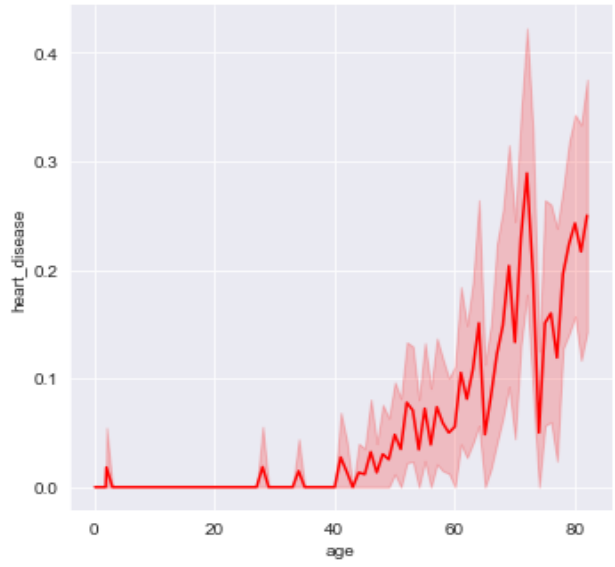


Figure (2.3) displays the average range of subjects who suffered/suffering from heart disease according to age. Approximately 4834 subjects are not suffering from heart disease and approximately 276 subjects currently/previously suffered from heart diseases. We can see between the age range 40 – 80 years a sharp volatile increase of subjects on average currently/previously suffering from heart diseases. Prior to this as approximated the age range between 0 – 40 years shows subjects rarely suffer from heart diseases. Noticeably both stroke and heart disease follow a similar increasing and volatile forward trajectory trend when measured against the age variable. Similarly, we see an increase from forty years and above in both these cases.

Figure (2.4)
Age/Heart Disease

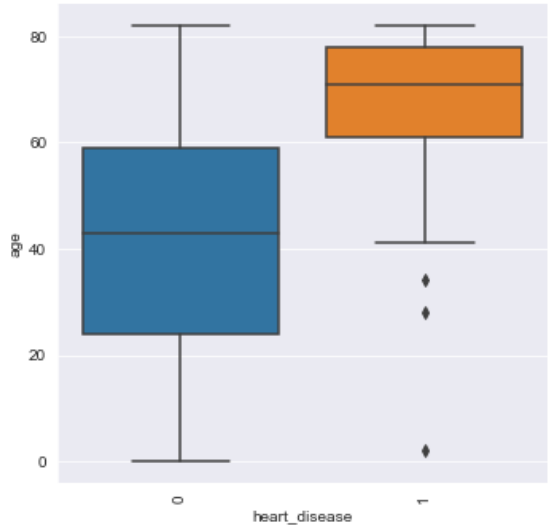


Figure (2.4) displaying the boxplot reinforces our above interpretation of the line plot which displays the average range of subjects with heart disease according to age.
It is evident from this boxplot that subjects in this dataset recorded as suffering from heart disease range in the higher age categories. Once again this is an interesting revelation in this dataset.

Figure (2.5)
Age/Hypertension

Figure (2.5) shows hypertension measured against the age category of subjects in this dataset.
Approximately, 4612 subjects record not having hypertension and approximately 498 subjects record having hypertension. Between the age range 40 – 80 years we see a sharp volatile increase of subjects on average having hypertension. The age range between 0 – 40 years of subjects show hypertension is rare as approximated prior to this visual analysis.

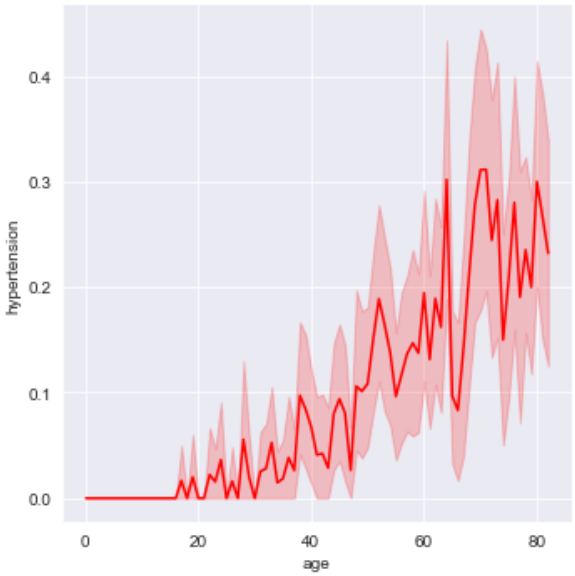
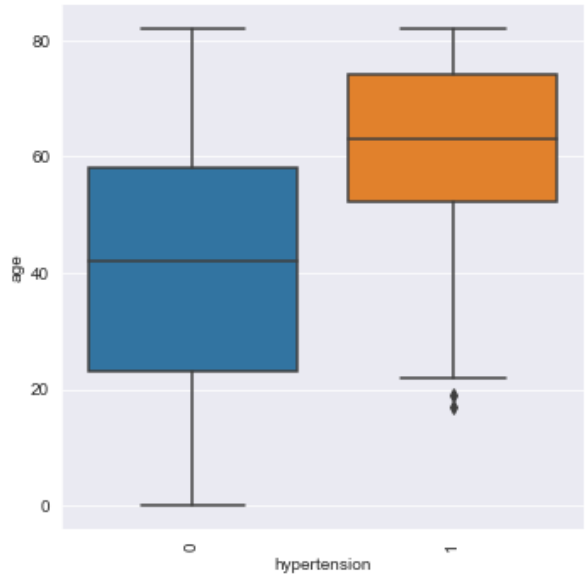


Figure (2.6)
Age/Hypertension



Further, the boxplot display (figure 2.6) reinforces our above interpretation of the line plot which displays the average range of subjects displaying hypertension according to age. We can see those subjects displaying hypertension lie within the higher age range of subjects.

It is clear from the above exploration of the data that subjects ranging in higher age groups are at an increased risk of becoming smokers in their lifetime, having higher blood pressure, having high levels of glucose blood levels. Body mass index as approximated increases as does age. Furthermore, in those aged forty and above we see an increase in subjects suffering from strokes and developing heart diseases. Overall, this has been a successful mining and exploration analysis of the data. The findings are aligned with external research conducted in this project as seen in the feature research section. This alignment contests to the accuracy and validity of the dataset quality.
As mentioned above a more specific analysis of data in relation to gender would have been an interesting direction for the project. Contrasting and comparing gender with known predisposing factors certainly would have been an important aspect of the research. However, as discussed it is not within project objectives and goals.
We have collected valuable information on predisposing factors related to cardiovascular diseases. Furthermore, this data exploration is a steppingstone in relation to design and development of blue-print national plans for a large-scale database system focusing on feature selection related to cardiovascular diseases.

Data Quality Report

On assessment of this dataset and through external research which aided in building up domain specific expertise, an informed and evidence-based evaluation was articulated based on dataset quality. Below we will outline this assessment process.

As highlighted in the data collection report this chosen dataset has met the standardized requirements aligning with newly designed objectives and goals of this project. Successfully identified was a dataset specifically associated with cardiovascular diseases, sourced from Kaggle (https://www.kaggle.com/lirilkumaramal/heart-stroke) which was categorized as Secondary Sources for search analysis. The dataset includes approximately 5110 observations and 12 instances. Origins of this dataset are unknown.

Linking Dataset Features with previous Feature Research (Specific Markers see page 9 - 10)

Age and smoking are instances in this dataset, and they are recognized as specific markers from previous feature research. Higher age groups are associated with an increased chance of developing cardiovascular diseases. Age ranges from zero to eighty-one for subjects in this dataset. This instance is one of the more important features of this dataset as it allows for analysis across known predisposing factors. The smoking instance holds valuable information on smoking behaviours and habits of subjects. There are four different categories including never smoked, previously smoked, currently smokes and unknown. Smoking is associated with increasing chances of developing cardiovascular diseases.

BMI is an instance in this dataset. This instance (body mass index) ranging from a minimum value of 10.3 to a maximum value of 37.6 holds valuable measurements which give excellent insight into a subject's weight. As highlighted previously obesity is associated with increased chances of developing cardiovascular diseases. A further analysis of this instance revealed that a shocking 3329 of subjects are considered overweight with body mass index values greater than twenty-five. The mean age of those subjects is 49. Approximately 1908 are female and 1421 are male. A further in-depth analysis revealed that 1920 subjects are considered obese ranging with body mass index values greater than thirty. The mean age is 50 and approximately 1115 of these subjects are female and the remaining 805 are male. Obesity has also been marked as a specific marker in the previous feature research section.

Average glucose level is an instance in this dataset. Average glucose levels range from a minimum value of 55.12 to a maximum value of 217.74, with a mean value of 106.147677. Hypoglycaemia is a condition in which blood sugar levels is lower than normal which is common in those who have diabetes (NHS (2020)). Insights can be extracted which can give an indication if subjects have diabetes. Diabetes is associated with increased chances of developing cardiovascular diseases and is also marked as a specific marker in the previous feature research section.

Hypertension refers to high blood pressure levels of the subjects in this dataset. This instance represents hypertension either as one for yes or zero for no. Although holding valuable information as hypertension is associated with an increased chance of developing cardiovascular diseases, this instance does not show blood pressure measurements (mm/Hg). Although holding valuable information there is credible and informal information missing that could prove to be an important part of analysis. Blood pressure has also been marked as a specific marker in the previous feature research section.

Gender is an important feature of this dataset as it permits for significant analysis regarding gender association with known predisposing factors of cardiovascular diseases. Stroke instance representing stroke either as 1 for yes or 0 for no hold's valuable information of those subjects who suffered from a stroke in their lifetime. Stroke increases the chances of developing more severe cardiovascular issues in life (British Heart Foundation (2020)). Heart diseases or as we have phrased throughout this

project cardiovascular diseases, a term interchangeably used to represent a range of different heart related diseases. This instance is a vital acquirement of this dataset as it will become our target bivariate variable for our Logistic Regression algorithm.

Dataset Issue's

Highlighted are some small amendments required, that have been altered prior to the Data Cleaning phase. These includes instance renaming and data type conversions as mentioned in the Data Description Report. These minor issues did not discredit the value of the data, amendment simply meant easier readability and understandability of the data. The forthcoming data cleaning report will detail further on these minor data cleaning tasks (see page 30).

Through further analysis of all instances, identified was approximately 4% of the BMI (body mass index) instance encompassing NAN values. No other instances held NAN values. This is an extremely low quantity of missing values and will have little impact in terms of distorting performance and results in the later stages of the machine learning modelling phase. Deleting these NAN values, it can be approximated will have little impact on the inferences to be gained from the dataset analysis. However, considering the contextual background of the project which is disease classification further analysis and appropriate management is required. Further details on this handling these values are outlined in the coming Data Cleaning Report. Maimon & Rokach (2005) emphasizes the importance of analysing data for outliers, as if undetected calculations and statistical inferences may be misleading and biased thus deviating from the true inference. As discussed previously in this report the development of domain specific research encouraged a thorough analysis to detect outliers in which two instances with probable outliers were identified. These included both the Age and BMI instance. Approximately -- which is an extremely low value and will have little impact in terms of distorting performance and results in the later stages of the machine learning modelling phase. As this is a significantly low value, an informed decision has been made to analyse and handle probable outliers in the data cleaning phase. Details and full analysis can be seen in the coming Data Cleaning Report. Considering there are a significantly low proportion of both missing values and probable outliers present, this has been determined as having little impact on the Logistic Regression model performance. Furthermore, this has also boosted confidence on the dataset validity.

Concluding this assessment and evaluation of dataset quality, we can confidently state that this dataset is of good quality and fit for purpose. Minimal problematic issues have been identified to be rectified which will be discussed further in the coming data cleaning report. This dataset holds valuable information and aligns with our specific markers from feature research. Therefore, inputting this data into the Logistic Regression algorithm can be approximated will achieve good results. Finally, this will support in providing sufficient evidence whether this may be a good model choice for the development, production, and deployment of a large-scale machine learning model to deliver on cardiovascular disease classification and prediction at a regional and national levels in the Republic of Ireland.

Moreover, as discussed above we have successfully interlinked dataset analysis and external research (feature research), therefore as envisaged reaching an important milestone of this project. The accumulation of two dataset analysis as discussed in this report and previous reports along with external research has built the foundations and platform for feature propositional inclusion for the large-scale database system related to cardiovascular diseases. This meets one of the main goals of this project which is to design and develop blue-print national plans for a large-scale database system focusing on feature selection related to cardiovascular diseases. This will assist in the reshaping of healthcare provision and the development of an updated strategy focused on preventative rather than curative measures at local, regional, and national levels in relation to cardiovascular diseases.

Note in the concluding section of this project we will deliver blue-print plans documentation in PDF format involving proposals for feature selection for this large-scale database system in relation to cardiovascular diseases. We aim to include suggestive universal equipment, data collection methods along with other suggestive informatic measures. Next, we will look at the data cleaning phase in preparations for our Logistic Regression model.

Data Preparation

Introduction

According to Squire (2015) data cleaning commonly referenced as data cleansing or data wrangling is a critical step in the analytics process. In advancement of the identification of specific datasets for this project, preparations were made in approximation of data cleaning problems arising. Prior knowledge and experience working with data influenced a wide-ranging plan to be designed established on both common data cleaning problems in the realms of machine learning and analysis of the specific datasets in question. Details of this plan are described below.

Furthermore, Nausman & Henschal (2010) stress to improve the quality and effectiveness of data various complex tasks may be required to correct data errors and inconsistencies. McCallum (2013) informs us that data wrangling involves the preparation and validation of data, which commonly occurs prior to the primary analysis. This research aided in motivating the data wrangling plan to be designed and implemented in this project. Osbourne (2013) tells us that cleaning data broadly defines a large spectrum of processes needed to happen before any attempt of analysis, mining, visualizations, or machine learning.

Subsequently, the plan proposal for data cleaning tasks have been divided into two separate categories. We will define and refer to these as primary and secondary cleaning tasks. This approach and ideology we feel will aid in prioritizing and ensuring no required tasks are missed. The *Primary Cleaning Tasks* will include missing value's, duplicate value's, correlation analysis, and outliers. Moreover, Chu & Ilyas (2019) highlights the Importance of understanding the processes involved in data cleaning and emphasizes that cleaning data does not solely involve removing erroneous data but other critical tasks. Other tasks included handling incomplete, inaccurate, irrelevant, corrupt, or incorrectly formatted data. The process also includes tasks such as merging data and other. *Secondary Cleaning Tasks* will include any data cleaning tasks identified that fall outside off the above stated Primary Cleaning Tasks category. Further details are outlined on these specific tasks in the following data cleaning report.

Data Cleaning Report

Primary Cleaning Tasks

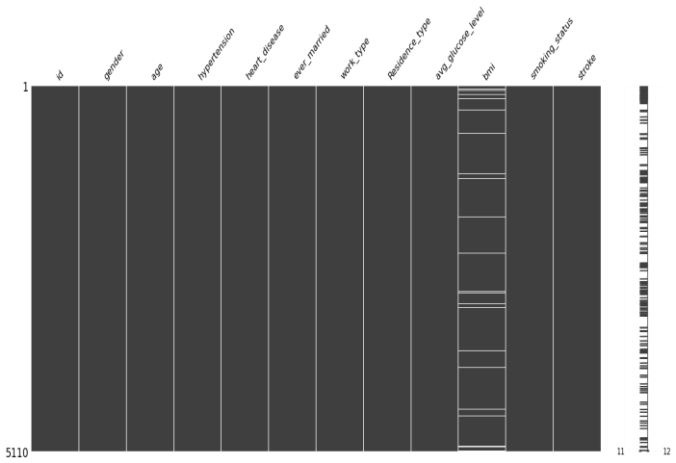
Missing Data

Missing data can be defined as data that is not stored within a variable or otherwise unobserved values that would be evocative for analysis consequently obscuring meaningful value. McKnight et al (2007) advises that the frequency of missing data can promote problems that affect the interpretation of research results and therefore ultimately hindering our understanding of the research and phenomena being examined. Graham (2013) & Kang (2013) further amplifies this suggesting that missing data will have significant implications within research when drawing inferences and conclusions from data with missingness.

It is an important feature of research analysis to understand the underlying factors behind the missingness. It is an important process to both identify and categorizing the types of missingness prevalent in the data. According to Hedeker & Gibbons (2006) there are three types of missingness. MCAR (Missing Completely at Random) in which missingness has no relationship between any values, observed or missing. MAR (Missing at Random) in which there is a systematic relationship between missingness and other observed data, but not the between the missing data. MNAR (Missing Not at Random) in which there is a relationship between missingness and its values, missing or non – missing. Furthermore, Hedeker & Gibbons (2006) highlight the importance of identifying the missingness type which helps to narrow the methodologies that can utilized to deal with missingness appropriately and effectively.

We analysed and investigated whether any missing – data was prevalent in the dataset. This required a heatmap to visualize any missingness present as seen below in (figure 2.7) and other alternative measures and checks to set in place in which we will discuss in this report.

Figure (2.7)



As we can see the heatmap revealed that one variable encompasses missing values and subsequently is a cause for concern that requires assessment. Initial analysis indicates and suggests that there are no distinctive patterns between the missingness within the observed or unobserved data.

Previous knowledge and an understanding of the fundamentals behind missing data allows us to highlight that the heatmap may well in fact hide missingness in variables that have a small proportion of missing data. According to Leke & Marwala (2019), the research available suggests that missingness will have a negative impact on the performance levels of designated machine learning algorithms thus reducing optimal results. This concept determines the next steps to be

taken to further analyse the missingness within the dataset in which a count comprehension was computed and utilized. This revealed that exactly 201 values where missing in the BMI (body mass index) variable which is approximately 4% of the variable in question. Furthermore, there were no missing values in any other variable as was approximated from initial visualizations of the heatmap. This was a positive finding and inspired confidence for project progression and for promising results in the later stages of the project.

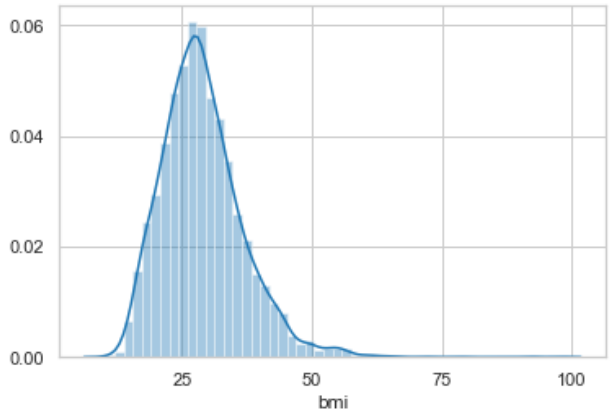
However, taking into consideration the project context which ultimately is to build a disease prediction model there can be minimal error in data preparations to be in-putted into the machine learning model. This thinking further encouraged a more in-depth analysis of the missing values to determine the missingness type and methodologies to be considered.
This in-depth analysis revealed a distinctive pattern between the missing values sorted in ascending order in which the relationship identified showed that the missingness within this variable ranged within the high ranging values. Concluding from this finding was that the missingness could be precisely determined as (MNAR) missing not at random. Following this finding a closer inspection and thought lead to the following methodologies being considered for imputation as discussed below.

Imputation Technique Analysis

Below in figure (2.8) is a histogram/kernel density plot which shows the univariate distribution of the BMI (body mass index) variable encompassing missing values. We can evidently confirm this distribution follows a normal distributed curve from an analysis of this visualization. As outlined within our *Imputation Technique Analysis* on our Jupiter Notebook Python environment, techniques considered and analysed include mean, median and mode imputations.

It is worthy to mention that other imputation techniques were considered such as forward, backward fill and regression imputation techniques. These were ruled out when it was discovered that the proportion of missing values was founded to be low and considering the BMI (body mass index) variable which from an earlier analysis of observations showed no distinctive direction in values.

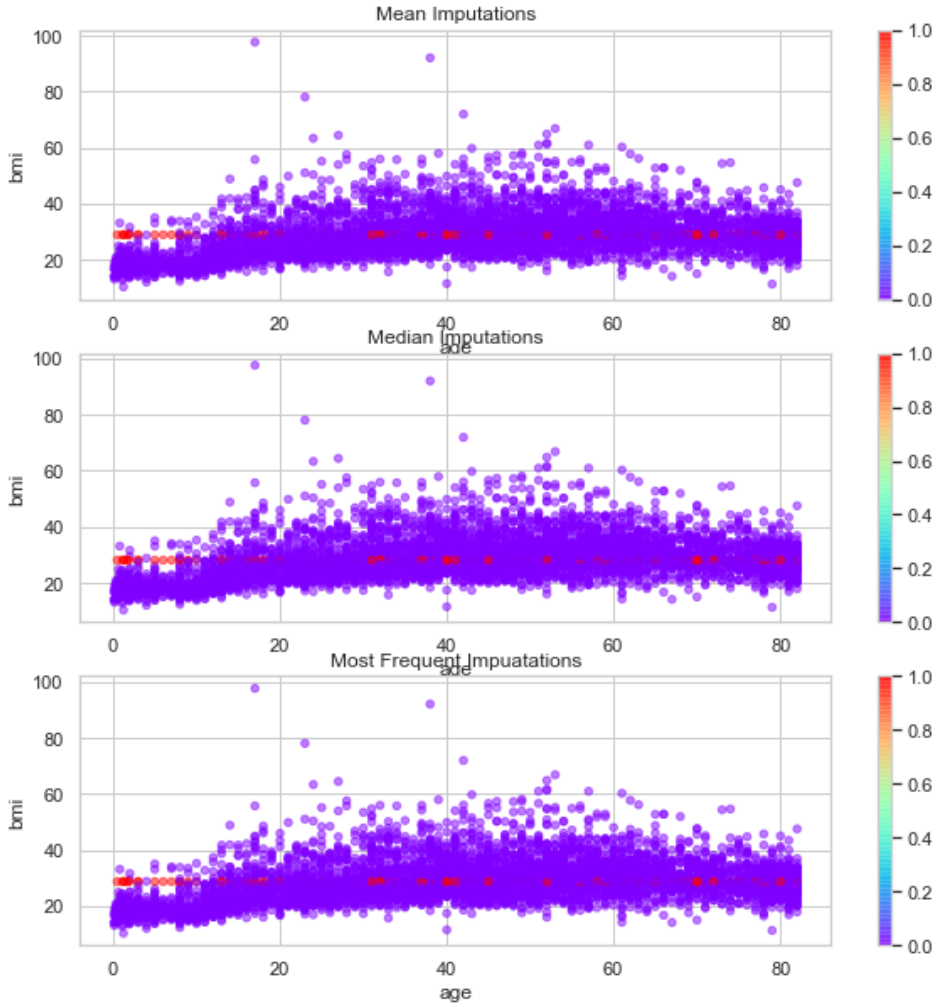
Figure (2.8)



Furthermore, as the original distribution holds approximately 4% of missing value's and follows a normal distribution curve, therefore the imputations techniques mentioned above will not deviate far from the normal distribution curve. This statement requires an understanding of the statistical foundations which inform us that a gaussian distribution (normal distribution) curve will include the mean, median and mode positioning at the peak of the curve. Theoretically this means these summary statistics in question are simultaneously corresponding values or they are approximately close to the same values not deviating far from each other.

Moreover, it is important to mention that as the proportion of missing value's is so low approximately 0.5% of the complete dataset, deleting or removing would hypothetically not discredit results to be achieved in the machine learning model stage. However, considering the contextual background of this project which is to build a disease predicational model there can be zero lenience for erroneous practices.
Therefore, we will be proceeding with testing imputation techniques such as the mean, median and mode imputation and making an informed decision on which procedure to proceed with that best fits our analysis.

Figure (2.9)



Following the trial and analysis of all the above imputation techniques, it is clear from figure (2.9) that all three imputation techniques result in corresponding if not exact values. This finding as approximated prior to all three imputations clearly show that any of these techniques would be a satisfactory imputation technique based on our findings.

As mentioned above and harmoniously throughout this project considering the contextual background of this project, which is building a disease predicational model to provide evidence for national plans, specifics are crucial. Furthermore, we will specifically analyse the effect of each individual imputation technique on the univariate variable in question. This process will involve visual analysis of each imputation against the non – imputed BMI variable. We will evaluate distribution distortion.

Figure (3.0)

Mean Imputation

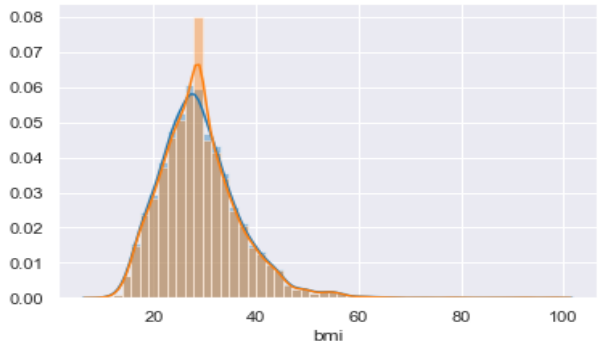


Figure (3.1)

Median Imputation

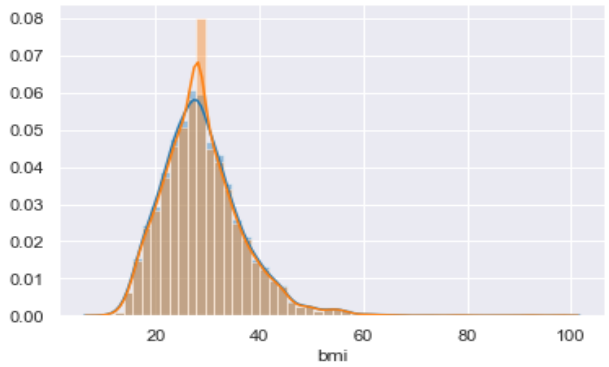
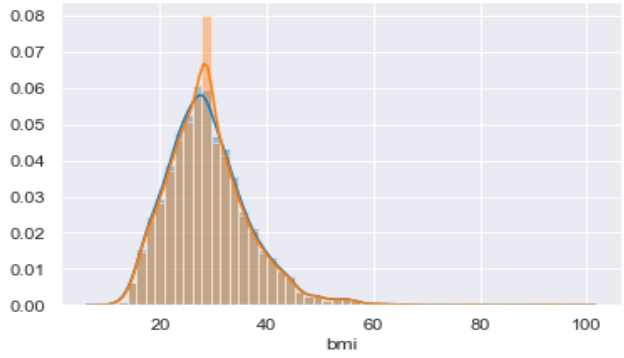


Figure (3.2)

Mode Imputation



Post analysis we can conclude that all three imputation techniques distort the original distribution minimally and equally or close to equal. We can therefore confidently conclude and close our analysis on missing data and proceed with our informed and evident based technique of choice. After strong consideration we have decide to proceed with mean imputation technique as we feel this will hold the normal distributed curve well.

Duplicate Value's

According to Nauman & Herschel (2010) duplicated values can be demarcated as values where identical information is repeated across multiple rows for some or all columns in a dataset. These can be broken down into two sub-categories:
Complete Duplicate Values refers to exact copies of records which are considered relatively easy to detect in comparison too incomplete duplicate values. Incomplete Duplicate Values also known as (Fuzzy Duplicates) can display slight or even large discrepancies within individual data values. Furthermore, incomplete duplicates are difficult to detect especially within larger volumes of data (Anderson & Semmelroth (2015)). Nauman & Herschel (2010) informs us that problems that arise from duplicated values include an increase in unnecessary expenses, a decrease in data usability, incorrect performance indicators and they constrain comprehension of the data in question and its value.

We completed an analysis to identify both complete and incomplete duplicate values in the dataset using code utilized for this specific and common data problem that arises when working within data environments. The code will identify if any duplicated values are present and which category they may fall within. Furthermore, if duplicated values are identified and falling within the specific categories mentioned above, the focus will then be deciding the best course of actions to take. Through research and experience working with duplicated values below we discuss and highlight various techniques and processes for dealing with duplicated. Moreover, we will provide evidence-based research for decisions to be made when necessary.

The analysis showed that there are no duplicated values present in the dataset. There was no further attention or action warranted. This was a positive finding as it motivated confidence moving forward with this specific dataset. One of the goals of this project is to provide evidence on how well a logistic regression model would deliver in the overall aims which is cardiovascular disease prediction.

Statistical Concepts & Correlation Analysis

It is important to understand various statistical techniques and terminology when working with data and within data environments. Discussed below are various statistical concepts which have strong underlying fundamentals behind machine learning algorithms. Moreover, we will focus specifically on correlation and its implications within data as will be discussed below.

V*ariance* can be defined as the average of the squared differences from the mean. This value indicates essentially how spread out the data points are from the mean. The *Standard Deviation* refers to the average distance of a data point from the mean value. A low standard deviation would indicate data points are clustered around the mean whereas a large standard deviation would indicate they are widely spread-out around the mean. Other statistical measures that illustrate the spread of data include the *Range* and *Z- Score*. Furthermore, regarding sample populations the standard deviation of multiple means is called the *Standard Error* (Bruce (2015)).
Covariance is an important statistical measurement used to analyse the linear relationship between two variables. In simplified terminology covariance describes how two data points co-vary i.e., how they change together. A positive value indicates a direct or increasing linear relationship whereas a negative value indicates a decreasing linear relationship. It is important to understand that covariance does not indicate the strength of a relationship rather the direction of the relationship (Kingsley & Robertson (2020)).

According to Archdeacon (1994) *Correlation* is a bivariate analysis that measures both the strength and direction of association between two variables. When analysing the correlation coefficient relationship, strength will vary between (– 1) and (+ 1). A value of + 1 or – 1 indicates a perfect degree of association whereas a value deviating towards or achieving a value of zero indicates a weaker/no relationship. The sign of the correlation coefficient indicates the direction of the relationship, (+) indicating a positive relationship whereas (–) indicates a negative relationship. Another important note to consider is that the correlation value is always between (- 1 / + 1) in which the scale is independent of the scale of the variables in question. This is not the case for covariance in which the scale is dependent on the variables in question.
Additionally, Kingsley & Robertson (2020) tells us that within statistics there are four common measurements of correlation which include Pearson Correlation, Kendall Rank Correlation, Spearman Correlation, and the Point – Biserial Correlation. Moreover, Kingsley & Robertson (2020) add Pearson r correlation is one the most popular and widely used statistical correlation measurements in quantifying the degree of the relationship or association between two variables.

In statistics, multicollinearity also known as collinearity is phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy. Subsequently, in this situation, the coefficient estimates of the multiple regression model may alter erratically in response to small changes in the data or model.

A research based and informed decision has been made to proceed with using the Pearson's R Correlation metric to analyse and measure the relationship of quantitative variables existing in the specific data collected. The decision process included an analysis of research on statistical fundamentals and more advanced methods as can be observed in the synopsis above. Furthermore, this helped to solidify prior statistical knowledge and consequently a threshold value for Pearson's correlation was set in advance of commencing preparations for the logistic regression model. The threshold value set in advance for correlation was (< = 0.7 / > = 0.7). Any value's ranging outside our original threshold value will be analysed and one instance will be removed prior to advancing the processed data into the machine learning algorithm.
The approach throughout the entirety of this project is to return to specific stages as deemed necessary to tweak parameters ultimately to stimulate better results aligned with project objectives and goals.

Previous knowledge and work with statistical measurements allows us to approximate this may possibly be a key factor in exponentially improving results in the algorithm model and testing phases in stage four.

Correlation Map

	ID	*Age*	*Hypertension*	*Heart Disease*	*Avg_Glucose_Level*	*BMI*	*Stroke*
ID	1.000000	0.003538	0.003550	-0.001296	0.001092	0.003084	0.006388
Age	0.003538	1.000000	0.276398	0.263796	0.238171	0.333398	0.245257
Hypertension	0.003550	0.276398	1.000000	0.108306	0.174474	0.167811	0.127904
Heart Disease	-0.001296	0.263796	0.108306	1.000000	0.161857	0.041357	0.134914
Avg_Glucose_Level	0.001092	0.238171	0.174474	0.161857	1.000000	0.175502	0.131945
BMI	0.003084	0.333398	0.167811	0.041357	0.175502	1.000000	0.042374
Stroke	0.006388	0.245257	0.127904	0.134914	0.131945	0.042374	1.000000

After analysing whether any numerical values may be correlated via the Pearson's correlation metric no returned values breach the threshold value (< = 0.7 / > = 0.7). The maximum returned value (BMI/AGE 0.333398) is significantly low considering it is the max correlated value. The minimum returned value (Average Glucose Level/ID 0.001092) is significantly low.
Concluding this analysis is there are no numeric values correlated meaning all the above instances can be inputted into the Logistic Regression algorithm. This removes the probability of the phenomena multicollinearity potentially occurring and causing issues within the algorithm.

Outliers

Hawkins (1980) tells us that an outlier can be defined as any observation that deviates so far from other observations such as to arouse scepticism that it was generated from a deviating mechanism. Maimon & Rokach (2005) informs us that outliers can be thought of as any data points that range significantly greater or smaller than other data points in a dataset. Hawkins (1980) explains that a more in-depth examination of a sample containing outliers would show up such characteristics or large gaps between outlying and inlying observations and that deviation between the outliers and the group of inliers, as measured on some suitability standardized scale. Below we will discuss and outline the measures to be taken to both preparations and deal with outliers that may be present in the dataset.

Age/Outlier Investigation

The first instance of interest was that containing the age of subjects. This instance was highlighted on the computation of summary statistics of the numeric values, revealing the min value of this instance age being 0.
Envisioned were two possible explanations. The first being that those subjects with a value of 0 for age may be representative of NAN values in this dataset and therefore concluding outliers are present in this instance. Secondly, the dataset includes children including those below the age of one.
To determine which of the two proposed explanations is more probable we further analysed the age instance. This can be seen in the attached Outlier Detection and Analysis Jupyter Notebook.
Uncovered were some interesting facts regarding age.
45 subjects in total were under the age of 1. If these were to be determined as outliers it is a relatively small proportion. Additionally, 320 of the subjects were under and aged 5 years. Moreover, 856 of subjects were under 18 years of age. However, this did not exactly conclude our investigation. A further analysis revealed that of the 856 of subjects under age 18, their bmi levels and average glucose levels would indicate levels are matching to someone under the age of 18 years. This indicates that the age instance does not hold outliers as described above.
Therefore we can state with evidence that this dataset includes teenagers and children some below the age of one years.

BMI/Outlier Detection

Prior to our analysis of the age instance in accordance too outliers, the *BMI* instance was also highlighted as possibly a cause for concern. This was once again highlighted through the computation of summary statistics which revealed the min value in the BMI instance was 10.3. As the BMI is a good indicator whether an individual is under or over-weight which is associated and regarded as a predisposing factor in relation to cardiovascular diseases, this was marked as a specific marker for research. The research conducted on BMI as highlighted earlier shows that any value for an adult below 18.5 is regarded as underweight.
As a value of 10.3 for BMI is regarded as extremely low if not in accurate for an adult this rose suspicion whether outliers where present in this instance.
As we concluded above that children and teenagers account for approximately 856 of subjects, therefore BMI ranging below the ideal value for adults which is 18.5 – 24.9, we were confident at this point that outlier occurrence was improbable.
However, we decided to further investigate this by analysing both BMI less than 18.5 within subjects aged below 18 years. Returned were approximately 299 subjects for analysis. We compared this to the total number of subjects ranging below the value of 18.5 for BMI which returned approximately 337 subjects. The difference being 38 subjects and a close analysis indicated that these subjects ranged above 60 years approximately 17 subjects. The difference being 21 in which we can determine that approximately 17 subjects' range between 18 and 60 years of age and are regarded

as underweight.
Therefore, closing the investigation into probable outliers present in the BMI instance, this has been determined not to be the case. Moreover, these findings influence optimized model predictions and results to be achieved in the model testing phase of the project.

Secondary Cleaning Tasks

As discussed earlier in the Data Cleaning Report cleaning tasks were segregated into primary cleaning tasks and secondary cleaning tasks.
We have now moved onto secondary cleaning tasks for this project and as outlined above they include any tasks that fall outside of missing value's, duplicate value's, correlation analysis, and outliers.

Data Type Conversions

We have identified some minor issues as highlighted in the previous data description report regarding specific data types in which several data type conversions are required.

Age instance was converted from float data type to int64 data type.
Hypertension instance was converted from int64 data type to object data type. This instance ranged from zero to one. Zero indicating no hypertension and one indicating hypertension.
Stroke instance was converted from int64 data type to object data type. This instance ranged between zero to one. Zero indicating a subject has not suffered from a stroke in their lifetime and one indicting they have suffered from a stroke in their lifetime.
Heart Disease instance was converted from int64 data type to object data type. This instance also ranged from zero to one. Zero indicating subjects who do not have heart diseases and one indicting those subjects that do. This is one of the more important conversions as it will become the target bivariate variable for the Logistic Regression Algorithm.

The ideology behind these conversions is an ability to generate better analytical results and machine learning modelling results. If unchanged these specific data types could cause problems for data analysis, manipulations, and the machine learning model phase (Logistic Regression Algorithm).

Column Renaming

To increase readability and understandability of the dataset proceeding with renaming specific instances was a measure taken. Below we will detail these amendments to the original data.

Original Instance Name	Updated Instance Name
id	Identification Number
gender	Gender
age	Age
hypertension	Hypertension
heart disease	Heart Disease
ever married	Marriage Status
work type	Work Type
residence type	Residence Type
avg glucose level	Average Glucose Level
bmi	Body Mass Index
smoking status	Smoking Status
stroke	Stroke

Planned for the data preparation stage was data encoding for the machine learning algorithm in stage four. However, due to time constraints this section has been pushed forward. Rational for encoding choice will be in full. Concluding the data preparation stage of this project we have successfully prepared data in preparations for the Logistic Regression Algorithm.

References

Anderson, A., Semmelroth, D (2015) Statistics for Big Data for Dummies. New Jersey, John Wiley & Sons

Archdeacon, T (1994) Correlation and Regression Analysis A Historians Guide. Wisconsin, The University of Wisconsin Press

British Heart Foundation (2020) Available at: https://www.bhf.org.uk/informationsupport/risk-factors/diabetes (Accessed: 10 January 2020)

Bruce, P (2015) Introductory Statistics and Analytics: A Resampling Perspective. Canada, John Wiley & Sons

Chu, X., Ilyas, F (2019) Data Cleaning. Georgia, ACM Books

CIO HSE (2016) Open Health Data Governance Strategy. Ireland: Office of the CIO HSE

Department of Public Expenditure and Reform (2017) Open Data Strategy 2017 – 2022. Ireland; Department of Public Expenditure and Reform

Fuster, V., Kelly, B (2010) Promoting Cardiovascular Health in the Developing World: A Critical Challenge to Achieve Global Health. Washington DC. The National Academies Press.

Graham, J (2012) Missing Data Analysis and Design. New York, Springer Publications

Hawkins, D (1980) Identification of Outliers: Chapman & Hall

Hedeker, D., Gibbons, R (2006) Longitudinal Data Analysis. New Jersey, John Wiley & Sons

Irish Heart Foundation (2020). Available at:https://irishheart.ie/heart-and-stroke-conditions-a-z/cardiovascular-disease/ (Accessed: 1 January 2021)

Kang, H (2013) "The prevention and handling of the missing data''. Korean Journal of Anesthesiology, Vol. 64, n. 5, pp. 402-406, PMC3668100. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100/ (Accessed 01st January 2021)

Kingsley, B., Robertson, J (2020) Research Methods and Statistics in Psychology Made Simple. New York, Oxford University Press

Leke, C., Marwala, T (2019) Deep Learning and Missing Data in Engineering Systems. Switzerland: Springer

Lin, PH., Svetkey, L (2012) Nutrition Lifestyle Factors and Blood Pressure. London, Taylor and Francis Group

Maimon, O., Rorack, L (2005) The Data Mining and Knowledge Discovery Handbook. USA, Spring Science & Business Media Inc

McCallum (2013) Bad Data Handbook: Cleaning up the Data so you can get back to work. USA, O'Reilly Media Inc

McKnight, P., McKnight, K., Sidami, A (2007) A Gentle Introduction. New York, The Guilford Press

Nauman, F., Herschel, M (2010) An Introduction to Duplicate Detection. Unknown, Morgan & Claypool Publishers

NHS (2020) Available at: https://www.nhs.uk/conditions/cardiovascular-disease/ (Accessed: 20 December 2020)

Osbourne, J (2013) Best Practices in Data Cleaning. London, Sage Publications Inc

Park, A (2019) Python Machine Learning: A Complete Guide For Beginners on Machine Learning and Deep Learning. Italy: Amazon Italia Logistica.

Peng, C.Y.J., Lee, K.L., Ingersoll, G.M. (2002) 'An Introduction to Logistic Regression Analysis and Reporting' *The Journal of Educational Research*, Volume 96(1), pp. 3–14

Royston P, Altman DG. (2010) Visualizing and Assessing Discrimination in the Logistic Regression Model, *Stat Med*. doi: 10.1002/sim.3994. PMID: 20641144.

Schmidt, M., Johannesdottie, S, A., Adelborg, K., Sundboll, J., Laugesen, K., Ehrenstein, V., Sorenson, H, T (2019) 'The Danish Health Care System and Epidemiological Research: from Healthcare Contacts to Database Contacts Records', Dove Press Journal: Clinical Epidemiology, pp. 563 – 591. doi: 10.2147 / CLEP. S5179083

Squire, M (2015) Clean Data. Birmingham, Packt Publishing

The World Health Organization (2020). Available at: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (Accessed: 11 Novemeber 2020)