

ITBW41 Predictive Analytics Project

Milestone Report

Healthcare Predictions

| Submitted by: | Shermaine Peh (213276Q)\*  Koh Shi Min (201357X)  Zhang Xiang (210896X)  Rawtbhik (210902G) |
| --- | --- |
| Module group: | ITBW41 - 01 |
| Tutor: | Mr Law Chee Yong |
| Report Last Amended: | 7th August 2022 |

**Table of Contents**

[**Data Collection**](#_1xhju17hhnql) **3**

[Rejected Datasets](#_p3tgtp7vv48r) 3

[Stroke](#_7hq17z625yet) 3

[Diabetes](#_onltctwzg7ub) 3

[About Selected Datasets](#_xsdgpj4d1j0e) 4

[Stroke](#_y27l4nk5h2ih) 4

[Diabetes](#_lbyxbp2hr0jx) 5

[**Data Understanding**](#_6za7b16wfx8z) **6**

[Stroke](#_obh4khtxmz6o) 6

[Diabetes](#_s2ryi9ieuxf5) 12

[**Data Preparation / Cleaning**](#_8vjblxj607ms) **30**

[Stroke](#_o6uzmwltzvd6) 30

[With one-hot encoding](#_lmu2atcra9c7) 32

[Need for Balancing](#_w12hn8tsw98p) 33

[Data Balancing](#_d3as63i1gj0w) 34

[Without one-hot encoding](#_v1ogwx54xzac) 35

[Proving Dataset Balanced](#_jlpm6p8mmi7) 37

[Selecting Dataset (One-hot Encoded vs Original)](#_of63yrnfrti) 38

[Partitioning Dataset](#_ce55k0h82e5n) 40

[Diabetes](#_l830tz4qt1kg) 43

[Data Cleaning Overview](#_nu1p5xhg6p4s) 43

[Steps taken for cleaning](#_1sftwoi9al7d) 44

[Selecting Dataset [Balanced Vs Unbalanced]](#_n9n7w278pbm4) 52

[**Timeline**](#_96oorxhiyzuy) **55**

[Updated Gantt Chart](#_gda5k7g479a4) 55

[Next Phases](#_ztqfy39pinp) 55

[**Appendix**](#_31schbbioyl) **56**

[**References**](#_4ya2a84b76az) **63**

# 

# **Data Collection**

As we have found out that stroke and diabetes are on the rise, we have decided to split the task to focus on these two health conditions. Stroke will be done and documented by Shi Min and Shermaine, while Diabetes will be done and documented by ZhangXiang and Rawtbhik. The data collected and prepared are sourced from Kaggle and the individual selected data source can be found below:

* Stroke – [Stroke Prediction|Visualization & Prediction](https://www.kaggle.com/code/aditimulye/stroke-prediction-visualization-prediction/data)
* Diabetes – [Diabetes Prediction | BRFSS2015](https://www.kaggle.com/datasets/cdc/behavioral-risk-factor-surveillance-system)

## Rejected Datasets

### Stroke

1. [Mayo Clinic - STRIP AI](https://www.kaggle.com/competitions/mayo-clinic-strip-ai/data?select=sample_submission.csv)

This dataset, provided by **Mayo Clinic**, was originally the one we wanted to use. However, upon reading their **licensing and agreement statement**, we discovered that they required us to **participate in their competition**. Moreover, the dataset included **image** classification, which was **outside the scope** of our prediction. Hence, in order to **avoid violating their rules**, we decided to find a new dataset.

1. [Mayo Clinic - STRIP AI 2](https://www.kaggle.com/competitions/mayo-clinic-strip-ai/data?select=sample_submission.csv)

This second dataset we wanted was **similar** to the first, but it had **too many unwanted columns which we did not understand**. Furthermore, the additional columns **lacked data explanations**. As a result, we **rejected** this dataset as we believe that we are **unable** to train the best model **without understanding** our data.

1. [HealthCare Problem: Prediction of Stroke Patients](https://www.kaggle.com/code/asaumya/healthcare-problem-prediction-stroke-patients/data)

This final dataset we rejected consisted of **only** ID and stroke occurrence, which was **not enough information** and columns to train a model.

### Diabetes

1. [N. Inst. of Diabetes & Diges. & Kidney Dis - MEHMET AKTURK](https://www.kaggle.com/datasets/mathchi/diabetes-data-set)

This dataset comes from the National Institute of Diabetes and Digestive and Kidney Diseases. This dataset was **rejected** due to **insufficient rows.** Data modelling requires a **minimum** of **1000** rows, but only **768** rows were **found** in this file.

1. [Hospital Frankfurt, Germany - JOHN](https://www.kaggle.com/datasets/johndasilva/diabetes)

This dataset **lacked data definition**, and the source of data mentioned might **not** be **reliable**.

## About Selected Datasets

### Stroke

*This section is documented by Shermaine and Shi Min*

Our stroke dataset is retrieved from [Kaggle](https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset), a **public** data repository for datasets. As part of [McKinsey & Company](https://datahack.analyticsvidhya.com/contest/mckinsey-analytics-online-hackathon/)'s healthcare **hackathon challenge** in 2018, they published this dataset of **electronic health records**, which is now publicly available on Kaggle.

According to the World Health Organisation (WHO), stroke is the **2nd leading cause of death** globally, accounting for 11% of total deaths. As a result, this dataset is used to **predict whether a patient is likely to suffer a stroke** based on the following **input** parameters. In the data, each row contains relevant **information about the patient**, which will help us determine how **strongly** each piece of information affects the **likelihood** of a person getting a stroke.

This dataset contains **5110 EHR observations and 12 attributes**. It has **11 input attributes and 1 output** attribute. The **output response is a binary value**, indicating whether or not the patient has suffered a stroke (1 if the patient has/had a stroke, 0 otherwise). Input attributes in EHRs include patient identifiers, gender and age, along with binary information of hypertension, heart disease, marital status, occupation, place of residence, average glucose level, body mass index, and smoking status.

An electronic health record (EHR) or electronic medical record (EMR), is a repository of patient information. This is a **computer-readable, automated recording of a patient's medical condition**, by four qualified physicians. Each record contains **vitals, diagnoses, and medical exam results.**

Further explanation of our attributes are as follows:

1. id: unique identifier
2. gender: "Male", "Female" or "Other"
3. age: age of the patient
4. hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5. heart\_disease: 0 if the patient has/ has no heart diseases, 1 if the patient has heart disease
6. ever\_married: "No" or "Yes"
7. work\_type: "children", "Govt\_jov", "Never\_worked", "Private" or "Self-employed"
8. Residence\_type: "Rural" or "Urban"
9. avg\_glucose\_level: average glucose level in blood
10. bmi: body mass index
11. smoking\_status: "formerly smoked", "never smoked", "smokes" or "Unknown"\*

### Diabetes

*This section is documented by Rawtbhik and Zhang Xiang*

This **Behavioural Risk Factor Surveillance System (BRFSS)** dataset is administered and supported by CDC's Population Health Surveillance Branch, under the Division of Population Health at the National Centre for Chronic Disease Prevention and Health Promotion.

The BRFSS [dataset](https://www.cdc.gov/brfss/annual_data/annual_2015.html) that we have obtained is an ongoing surveillance system designed to measure behavioural risk factors for the non-institutionalised adult population (aged 18 years of age and older) residing in the United States.

This dataset contains **441,455 responses** and has **330 features (columns)**. These features are either **questions directly asked of participants** or **calculated variables based on individual participant responses**.

There are **features** that include **tobacco** use, **HIV/AIDS** knowledge and prevention, **exercise, immunisation, health status**, **healthy days, health-related quality of life, health care access, hypertension awareness, arthritis burden, chronic health conditions, alcohol consumption, fruits and vegetable consumption**, and seatbelt use.

This data is helpful in identifying the numerous preventive health practices and risk behaviours that are linked to chronic diseases, injuries, and preventable infectious diseases that affect the population.

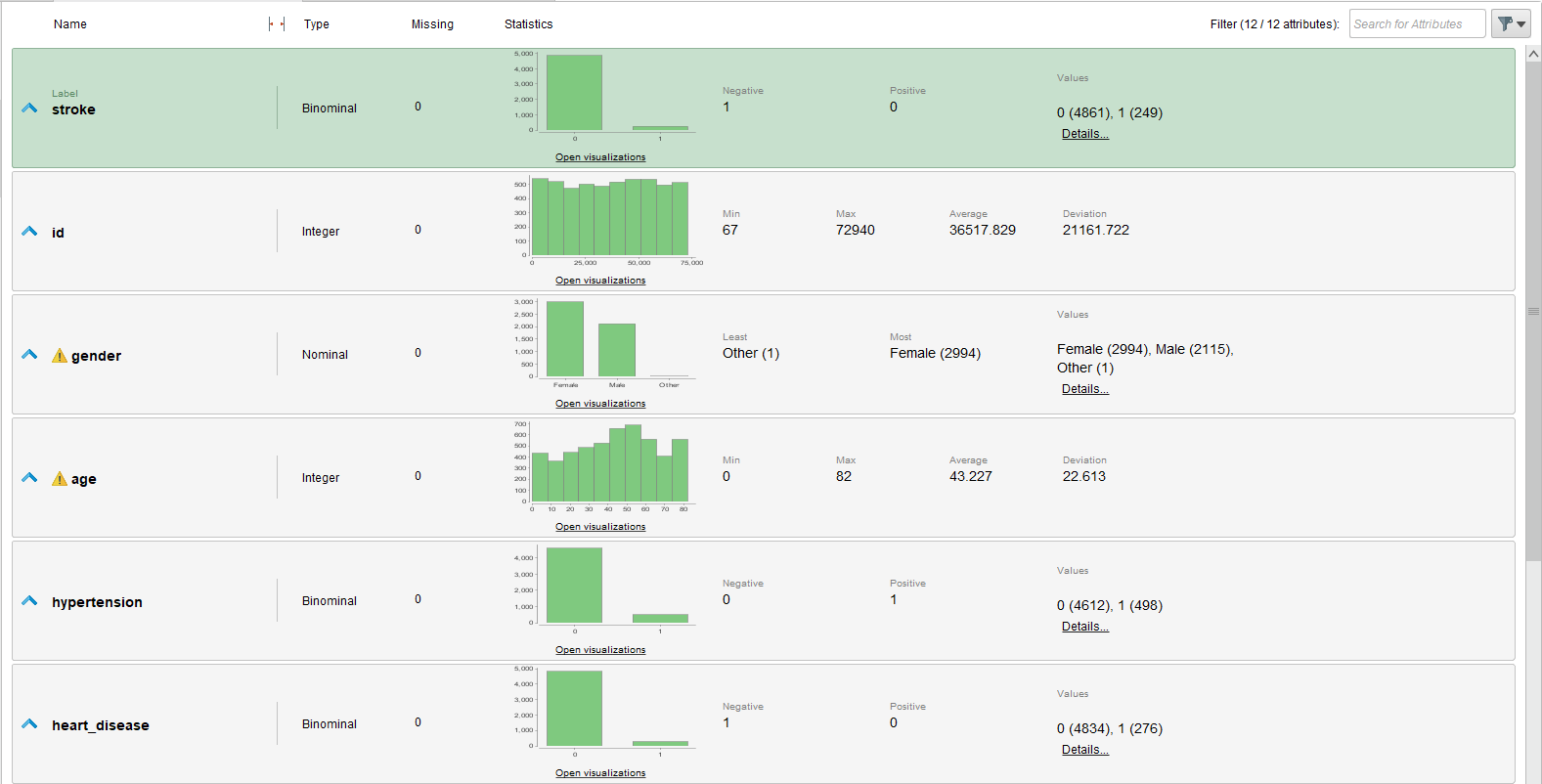
However, based on professional opinions and some diabetes disease research done regarding factors influencing diabetes and other chronic health conditions, we **have reasons to believe that the selected desirable features (columns)** that we have chosen would help us achieve the ideal analysis goal that we have in mind.

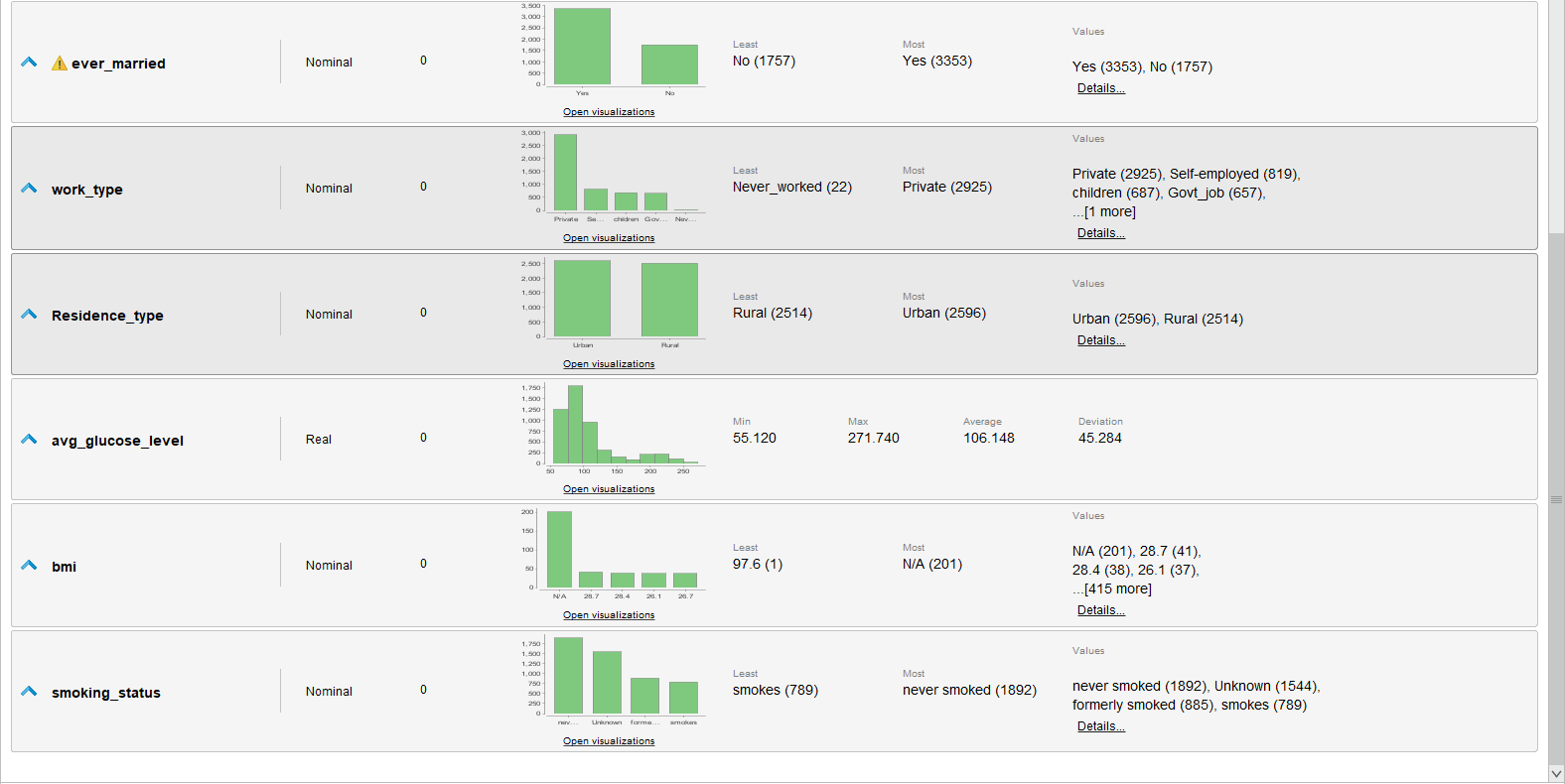
# **Data Understanding**

## Stroke

*This section is documented by Shermaine and Shi Min*

In order to prepare and clean our **raw** data, we first need to **understand** them. By visualising our data, we can better understand what our data is before preparing and cleaning it. With the help of **Tableau and rapid miner**, we have created a visualisation that allows us to detect any **abnormalities** and **insights** into what our data can possibly show through importing our CSV dataset.





Overview of data

1. To better understand the **severity** of **unbalanced** data in our raw data, we have plotted a bar chart.

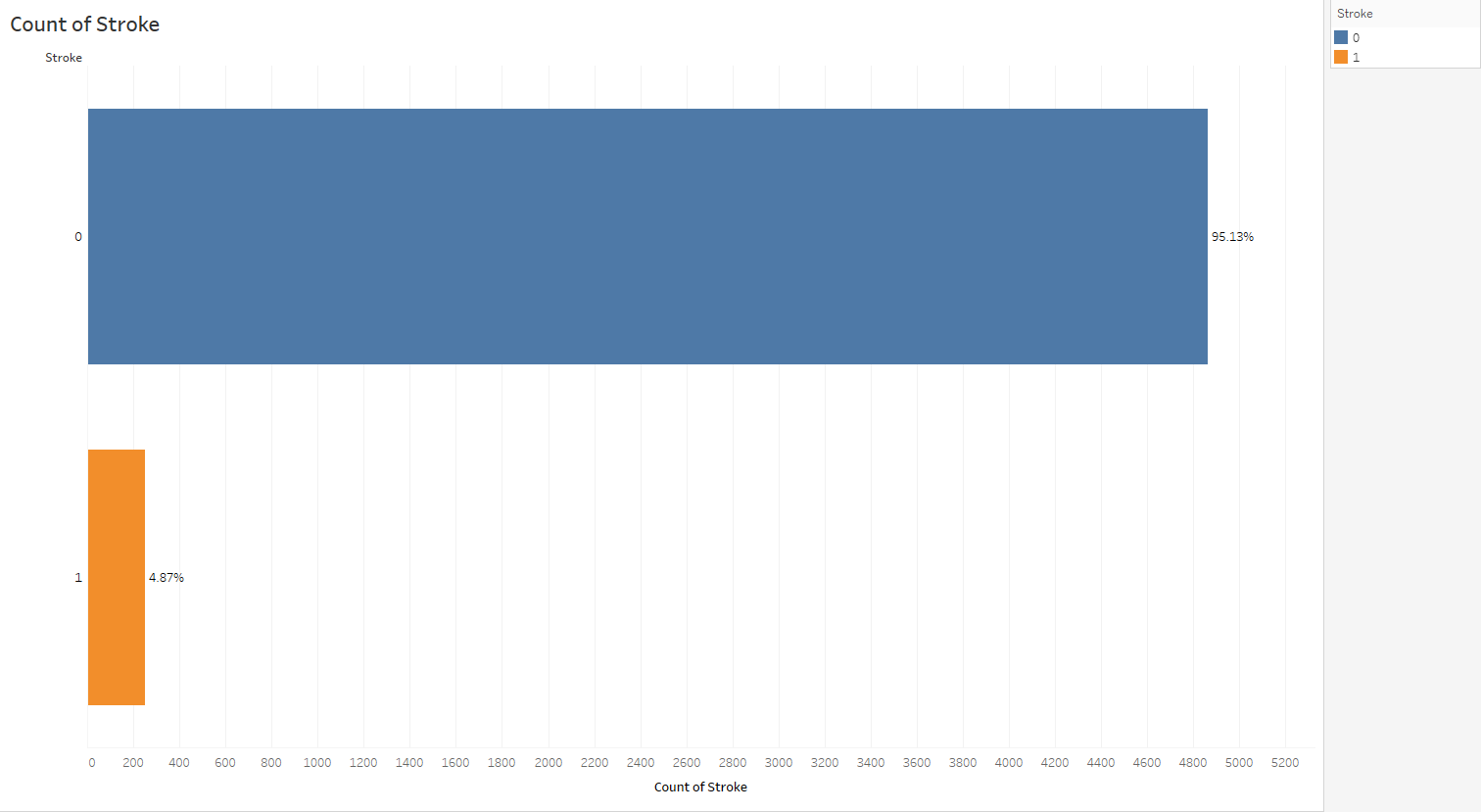


Figure 1 - Count of Stoke

*Figure 1* shows that almost **95%** of our raw data states that an individual does **not** have a stroke, which is our **target variable**. With this result, our model is proven to be **highly unbalanced** and will be misled during the creation of the model as our models will only be accurate in predicting individuals that do not have strokes.

1. As shown in *Figures 2 and 3*, we find that not only is the target variable **unbalanced**, but also the input variables that are binaries such as hypertension and heart disease are about 90% and 95% unbalanced.

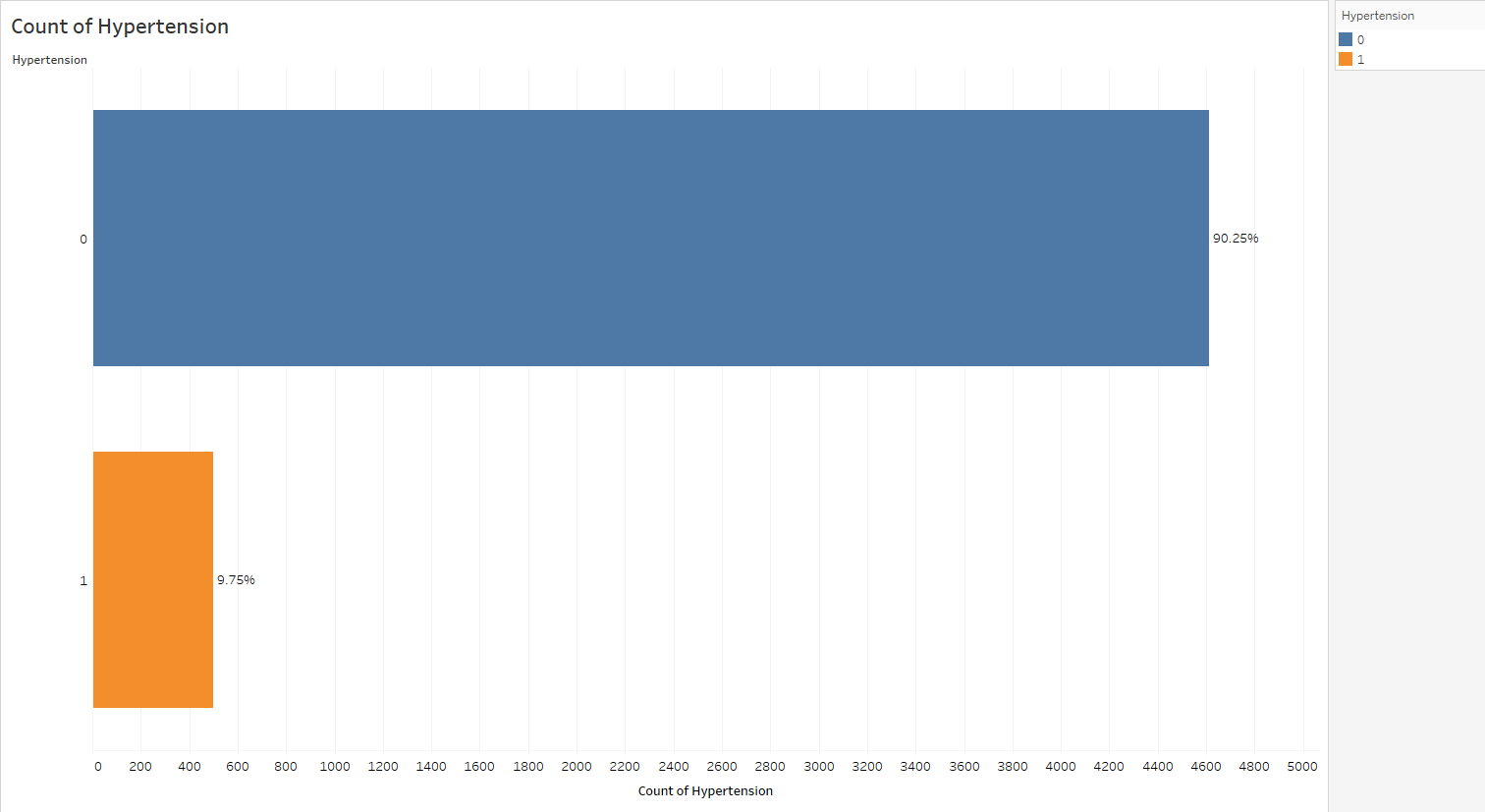


Figure 2 - Count of Hypertension

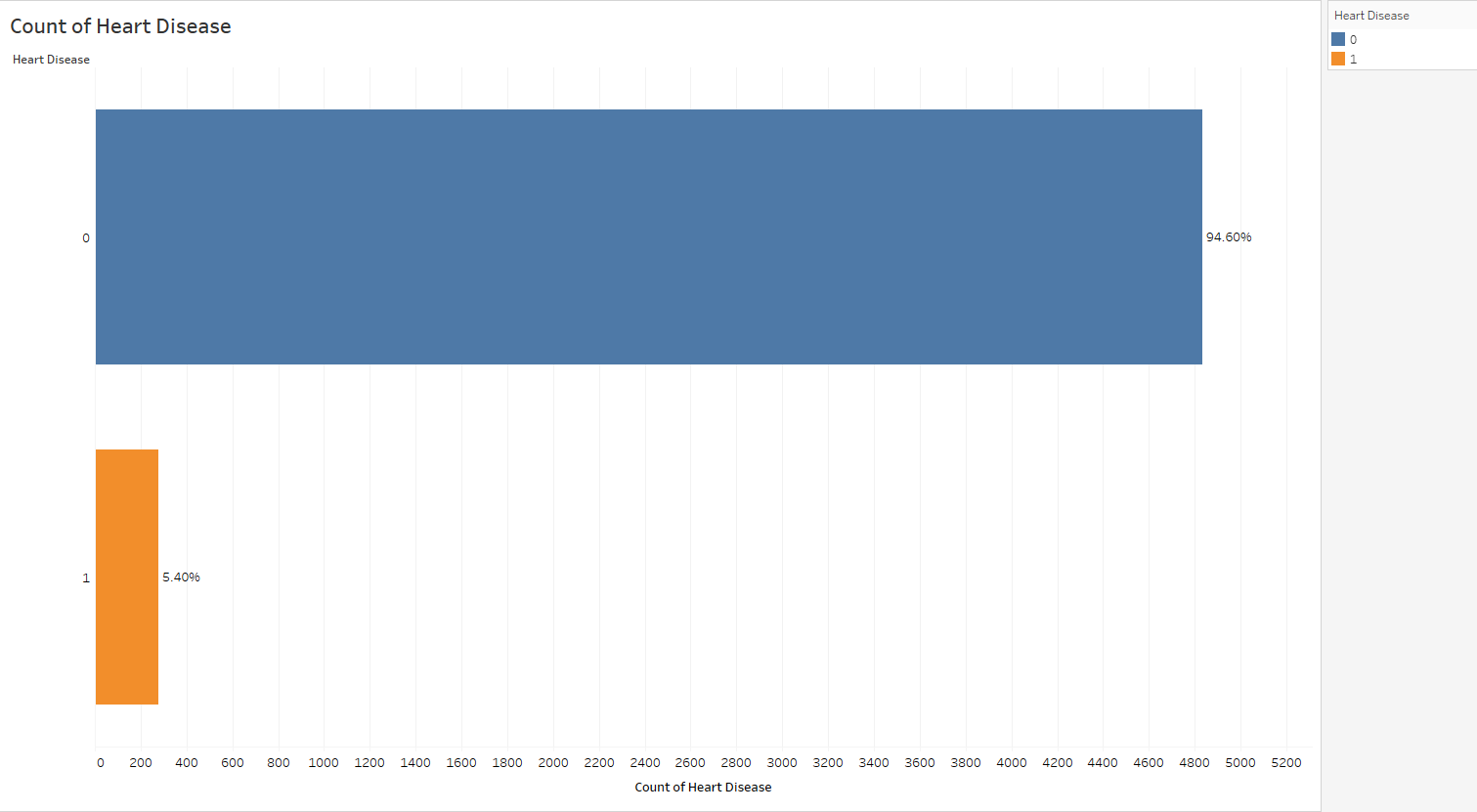


Figure 3 - Count of Heart Diseases

1. Taking a look at *Figure 4*, we find that there are unknown smoking values. Based on the dataset explanation, "Unknown" in smoking\_status indicates that the patient's information is **not available**. Due to the **subjective** nature of smoking and that unknown values do not provide us with any information, we have decided to **remove** smoking’s "unknown" values.

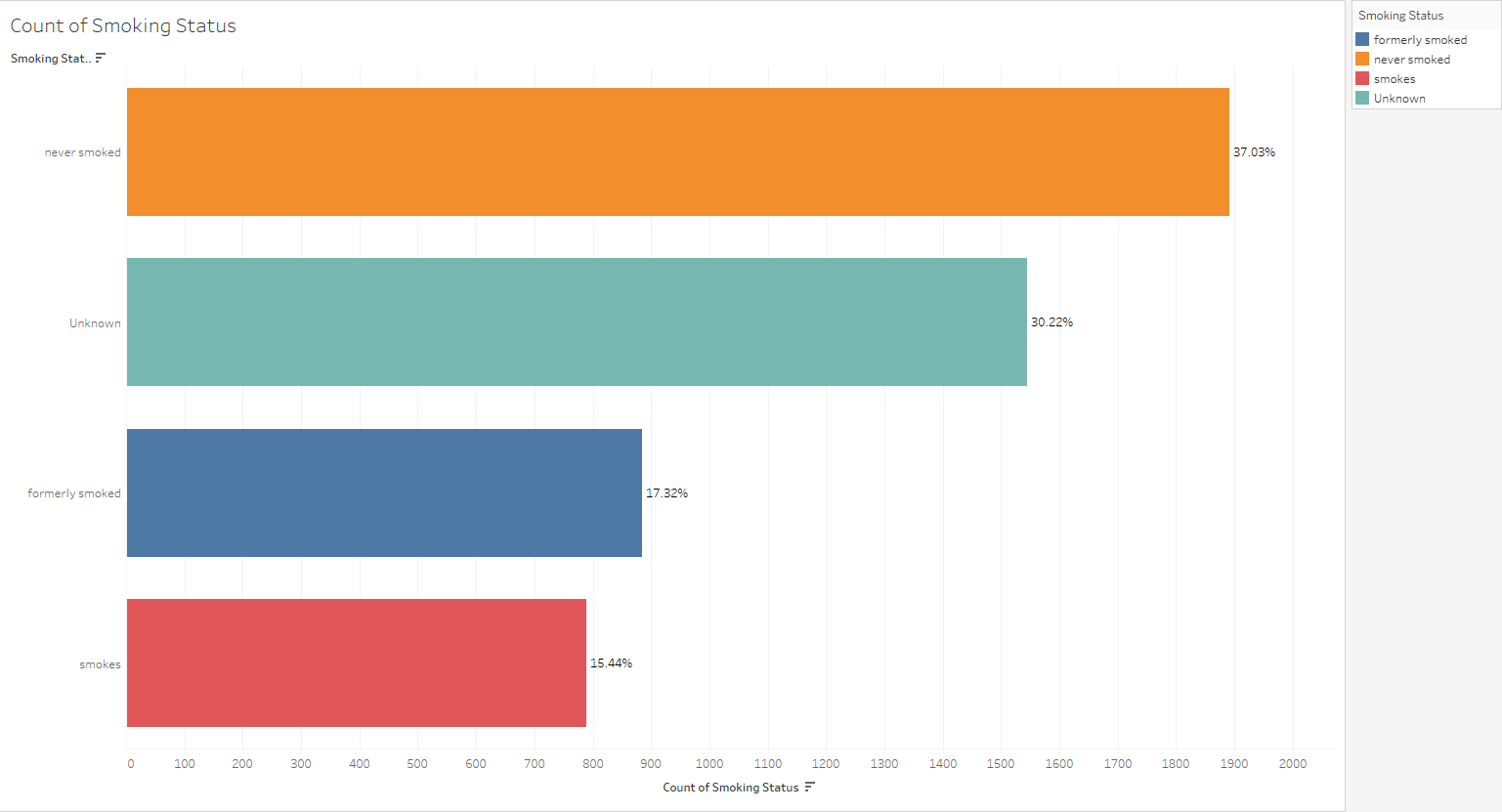


Figure 4 - Smoking Status with Unknown Values

1. Continuing to use visualisations to find **anomalies**, *Figure 5* reveals that the age column contains **decimal** values. As the dataset **did not include an age decimal data explanation** and we are **uncertain** whether it is a wrong value or if it signifies something else, we **removed** ages < 2. However, these values were **automatically removed** when the "unknown" smoking value was removed, hence, **no action** was needed for this section. Here, we find that age also consists of **many unique values.**

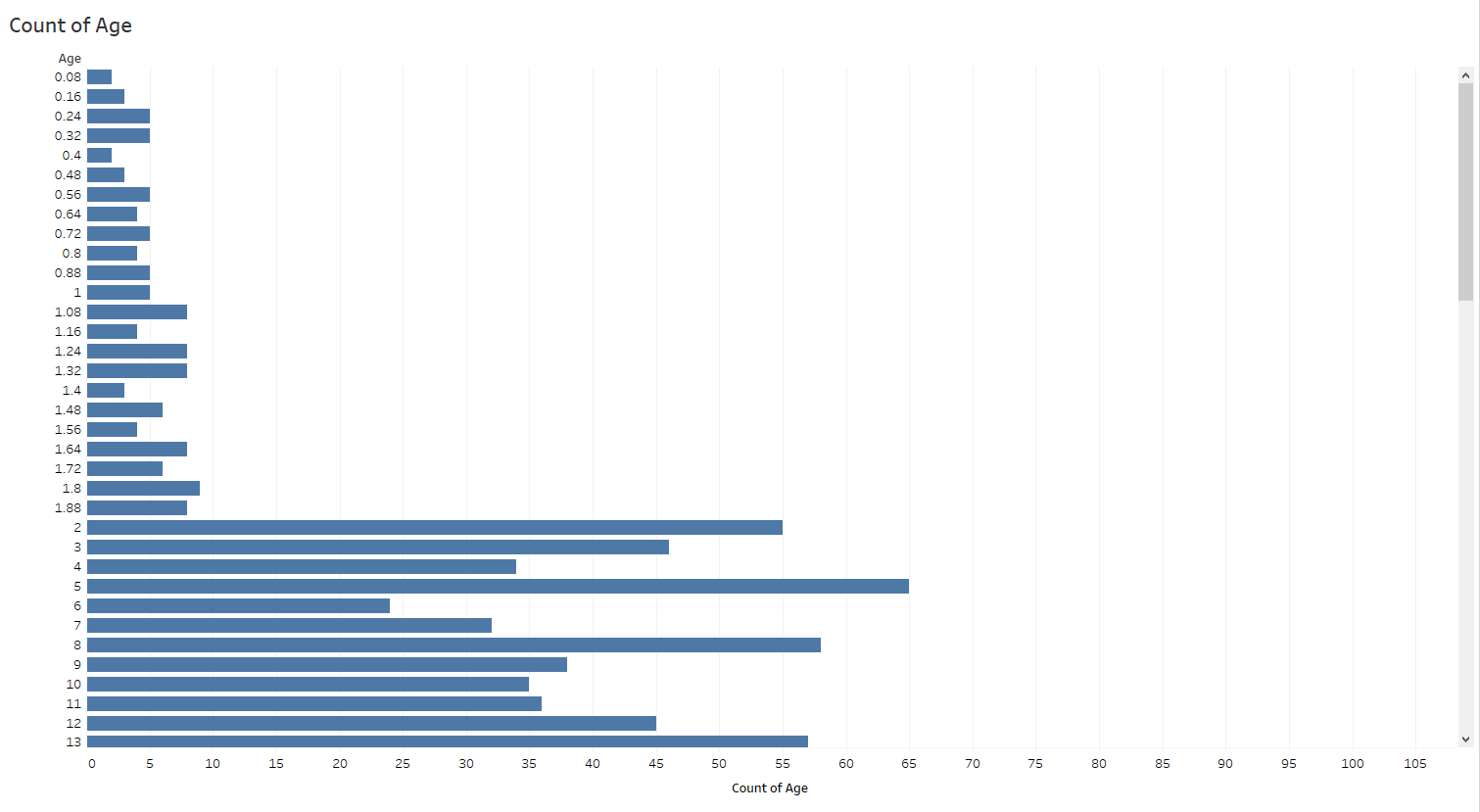


Figure 5 - Age with decimal values

1. The scatter plot in *Figure 6* also shows that there is a very strong **anomaly** in bmi, consisting of **201 missing values**. We have therefore decided to **remove all N/A** values from the bmi and do imputation, using the mean values.

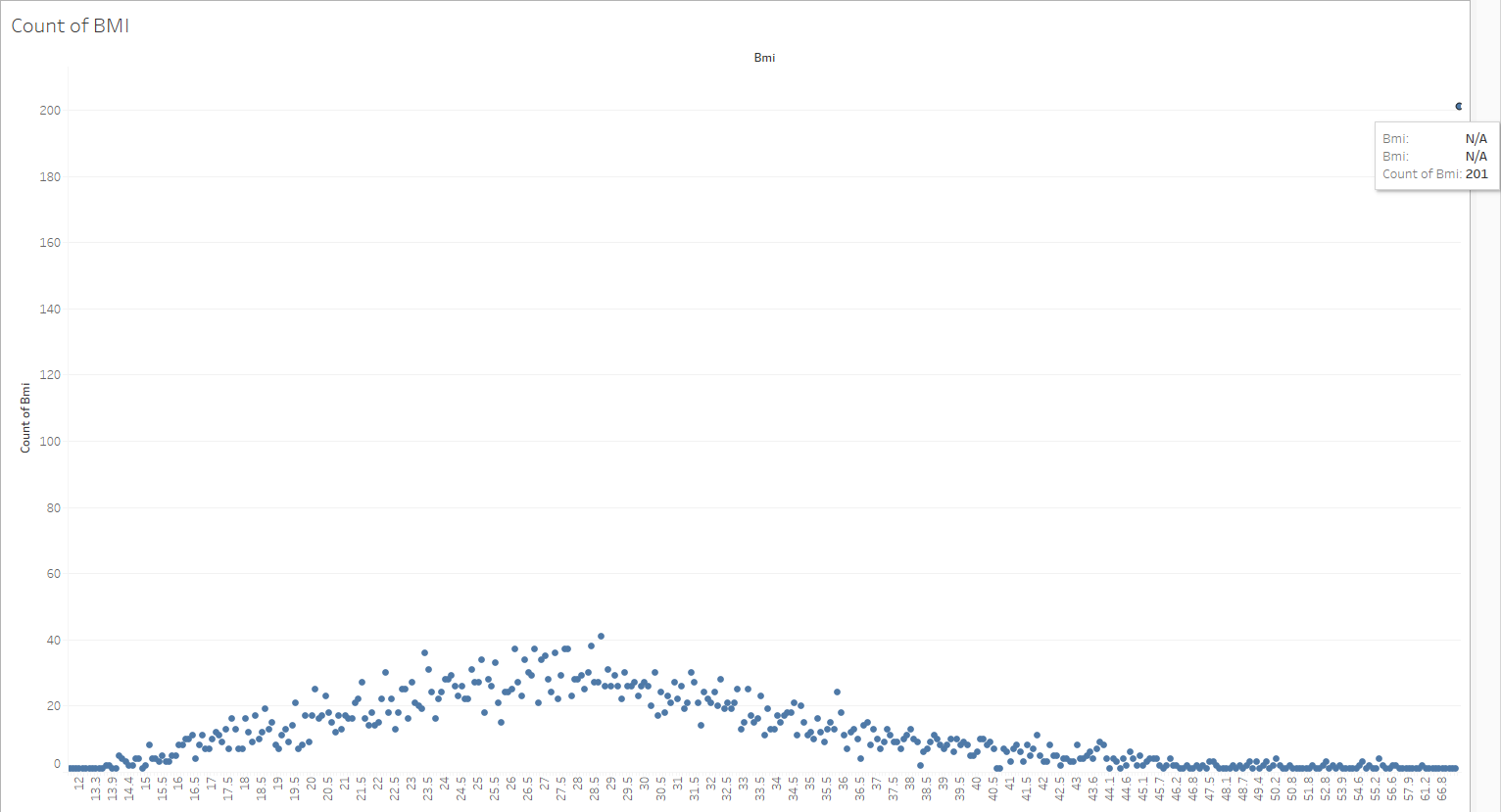
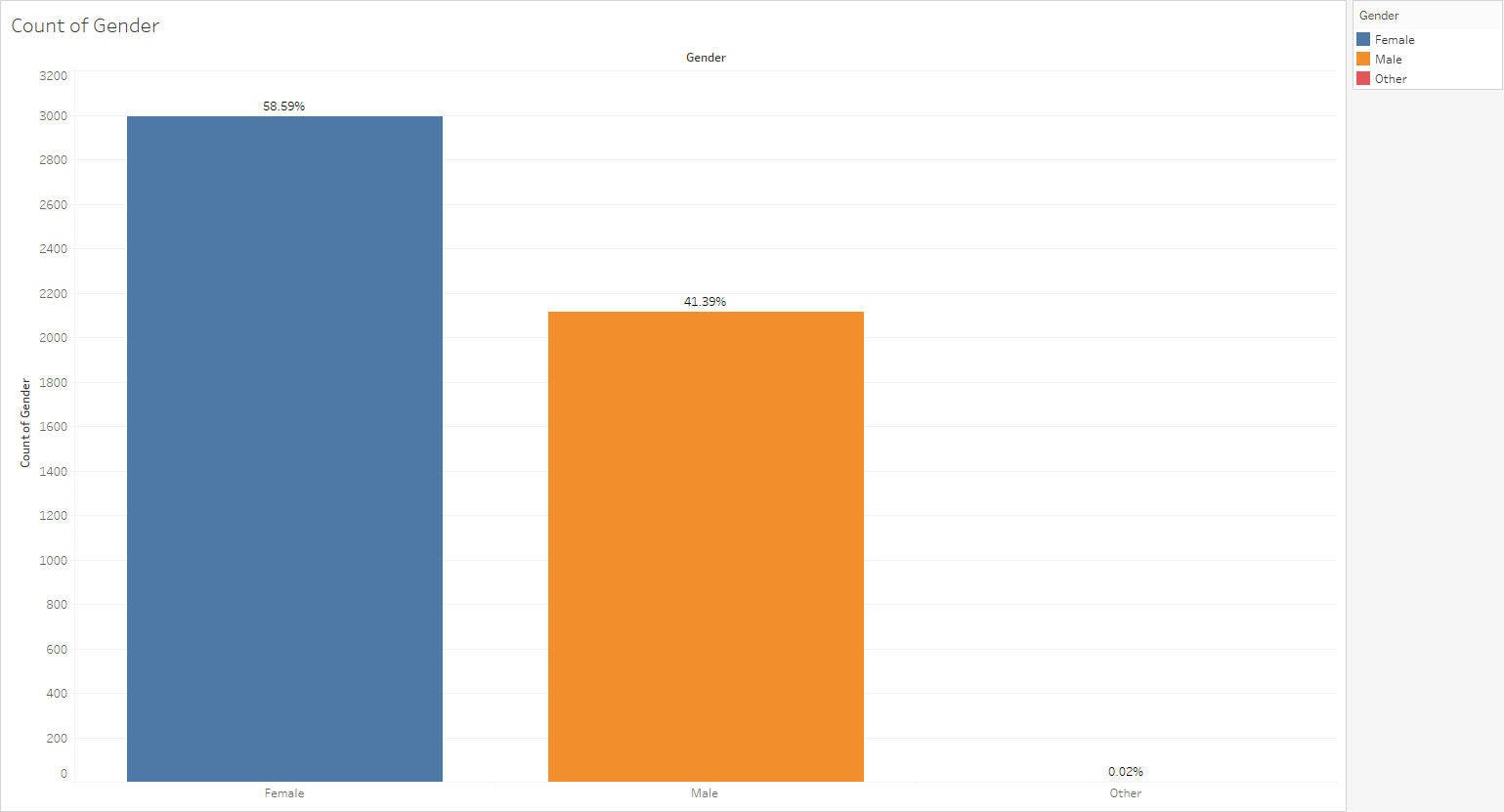


Figure 6 - BMI with missing values

1. Last but not least, *Figure 7* shows that Gender consists of **one "other" value**. Due to the **lack of explanation** in the dataset, our **uncertainty** of what this value means and the fact that gender is also a **subjective** attribute, we have decided to **remove the “other” value.**

~~~~

Gender with “Other” value

1. Finally, when looking through our dataset, we find that our dataset also consists of some **data privacy and ethics worries** as follows
   1. **Id [Quasi Identifiers]** → possible risk of **directly** identifying individual
      1. To counter this, we have decided to **anonymise** the ID and create a rowID which has **no relation** to the individual, to **protect** the individual’s **identity**. A **master table** was also created so that an **authorised** individual will be able to **re-identify** this individual if needed.
   2. **Age [Quasi Identifiers]** → with many **specific** values about an individual, a possible risk of **indirectly** identifying the individual may occur.
      1. In order to counter the data privacy & ethics worries, we have decided to **bin** the ages in groups in order to **reduce** the individual’s age **specificity**.

In **summary**, the following columns need to be **cleaned and anonymized** after doing data understanding on the **uncleaned** dataset:

1. gender
2. smoking\_status
3. age
4. bmi
5. id
6. Clean and trim – remove white and trailing spaces to ensure that there are no differences in all values

## 

## Diabetes

*This section is documented by Rawtbhik and Zhang Xiang.*

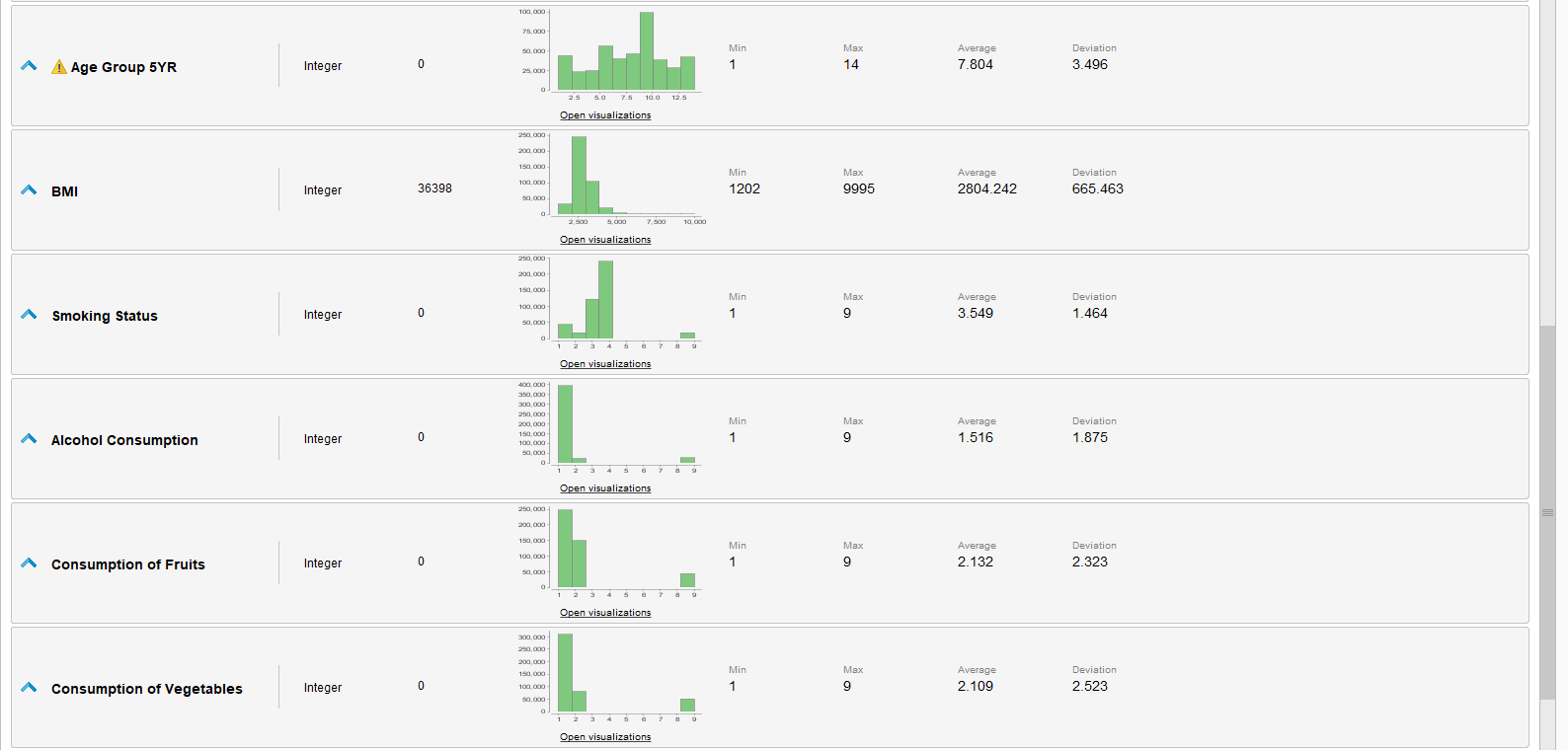
Based on professional opinions and some diabetes disease research done regarding factors influencing diabetes and other chronic health conditions, we **have reasons to believe that the selected desirable features (columns)** that we have chosen would help us achieve the ideal analysis goal that we have in mind.

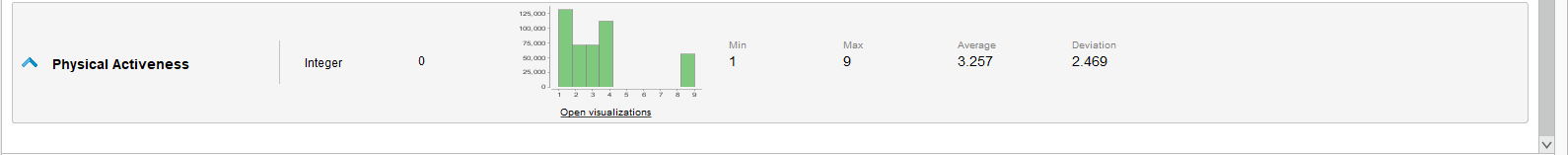
More definitions and references of data can be found in [Diabetes Appendix](#_w9denneu4x1r)

Overview-data









Understanding Variables Individually

1. Diabetes (DIABETE3) [Target Variable]

(Ever told) you have diabetes (If "Yes' and the respondent is female, ask "Was this only when you were pregnant?". If the Respondent says pre-diabetes or borderline diabetes, use response code 4.)

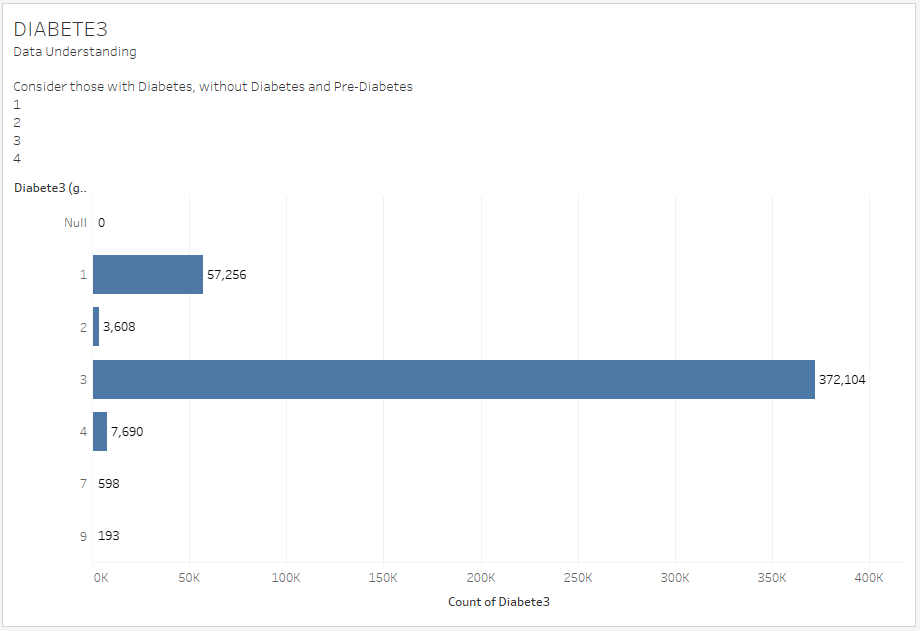


Figure 1 - Diabetes with values 1, 2, 3, 4 & 9

This chart shows us the number of respondents who are diagnosed with diabetes, non-diabetics and borderline diabetes.

* ***Value 1*** is respondents with diabetes (male and female).
* ***Value 2*** are **female respondents only** who say they were told that they were diagnosed with diabetes **during pregnancy.**
* ***Values 3 & 4*** are respondents who **do not have diabetes or are experiencing borderline diabetes.**

Since our target variable is to produce a binary outcome, we will reduce it to 2 values by creating values 1 and 0. 1 for respondents with diabetes and 0 for non-diabetics.

As shown in the chart above it was found that there are **values like “Null”, 7 which represents “Don’t Know”/”Not Sure” and 9 represents “Refused to answer”**. These data recorded down would not generate any useful findings. Only data that is able to tell if a person has or does not have diabetes are considered to support our modelling.

1. Age Group (\_AGEG5YR)

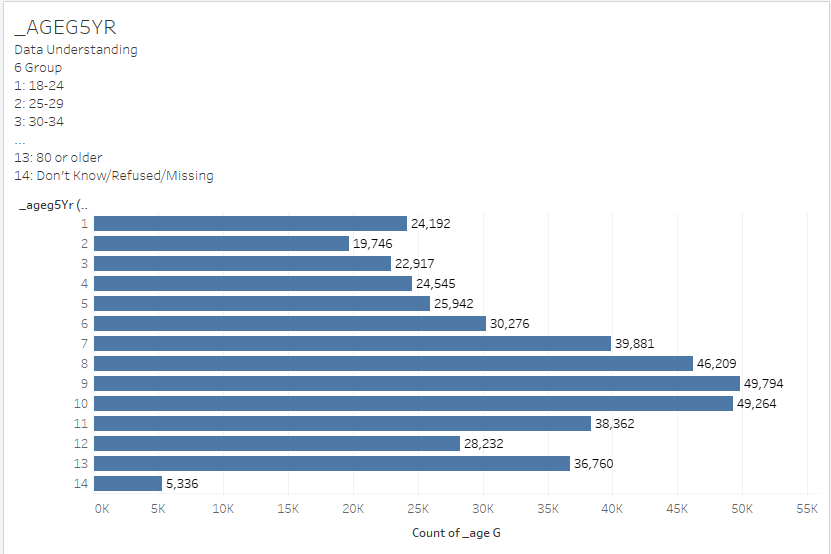


Figure 2 - Fourteen-level age category (group)

This graph **depicts the 14 values** that represent the different age groups, which are **grouped in the range of 5 years interval**. Age 18-24, for example, is represented as value 1, age 25-29 is represented as value 2, and age 80 and older is represented as value 13, just to name a few.

The age group with the greatest number of surveyors is between the ages of 60 and 64, which is a value of 9 because many respondents in this age group gradually develop/detect common health problems.

As shown above, there is an age group (Value 14) that represents "Missing/Don't Know or Responder Refused to Answer." Since this value **generates no insights in our model**, it will be converted to a missing value and removed.

1. BMI (\_BMI5)

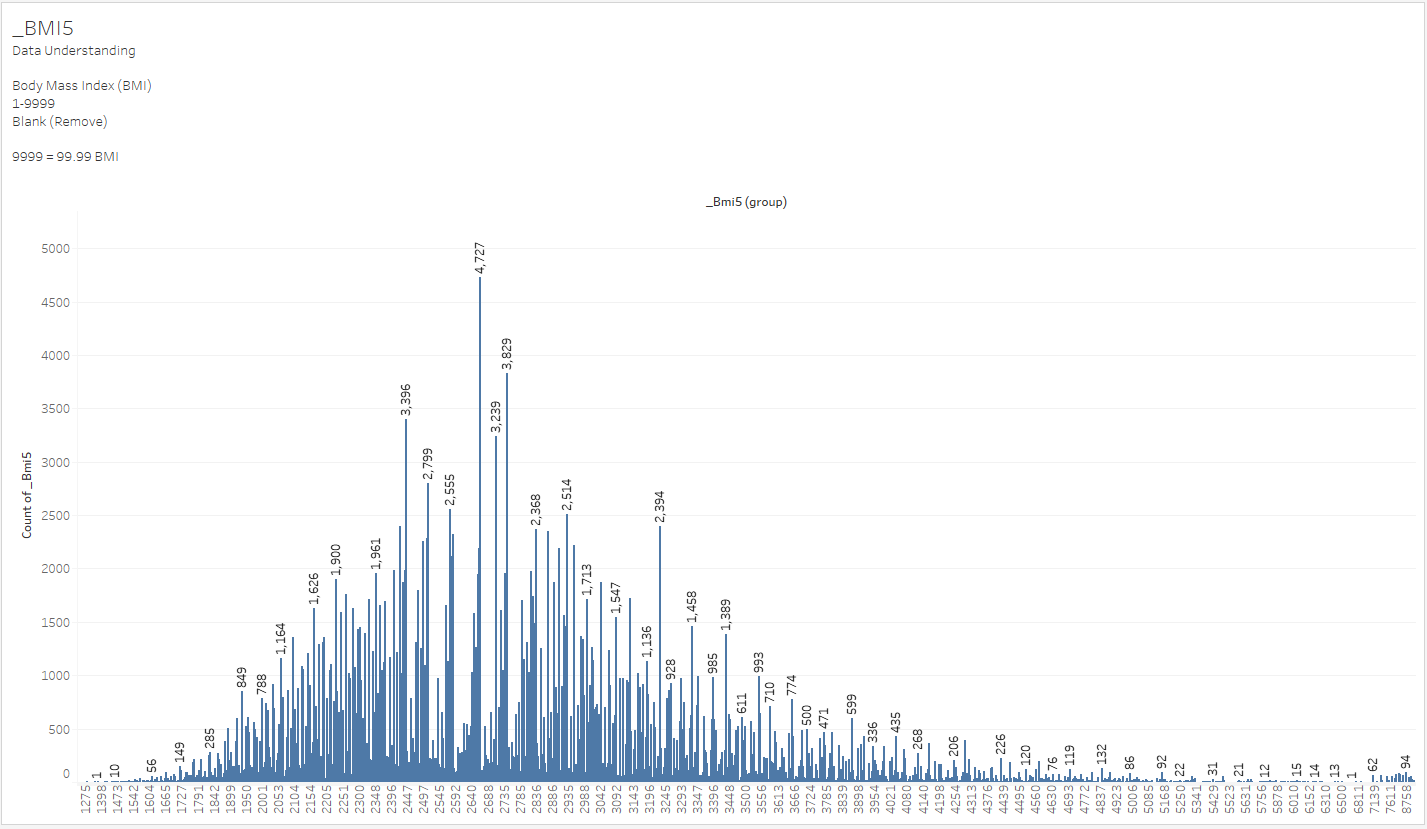


Figure 3 - BMI with inputs in thousands

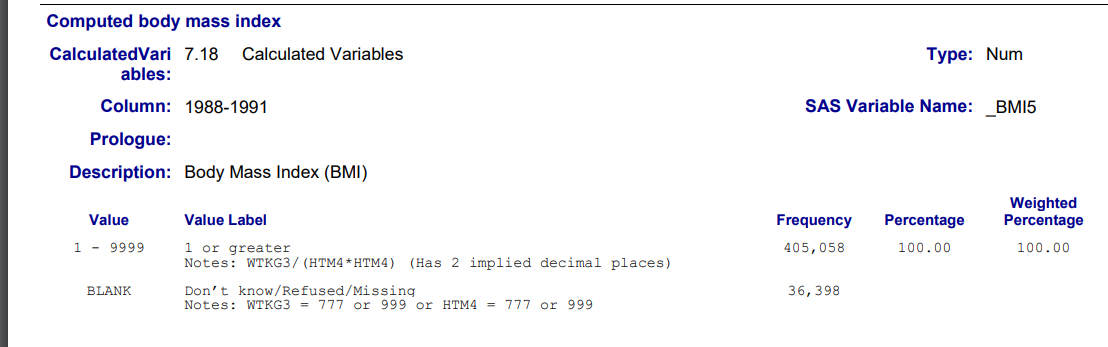


Figure 4 -Proof of BMI with 2 implied decimal places

In Figure 4, It was illustrated and noted in the codebook that the respondents' BMIs are recorded in the thousands (it states “Has 2 implied decimal places”). We then have to divide these numbers by **100** to return them to **2 decimal places**, which would be done via the data cleaning process.

It is clear that there are outliers in the figure above, and for this reason we want to **remove all outliers** from our dataset while **keeping appropriate inputs in our data using KNIME** as part of the cleaning stage.

1. Diet

**\_FRTLT1** – Consume Fruit 1 or more times per day

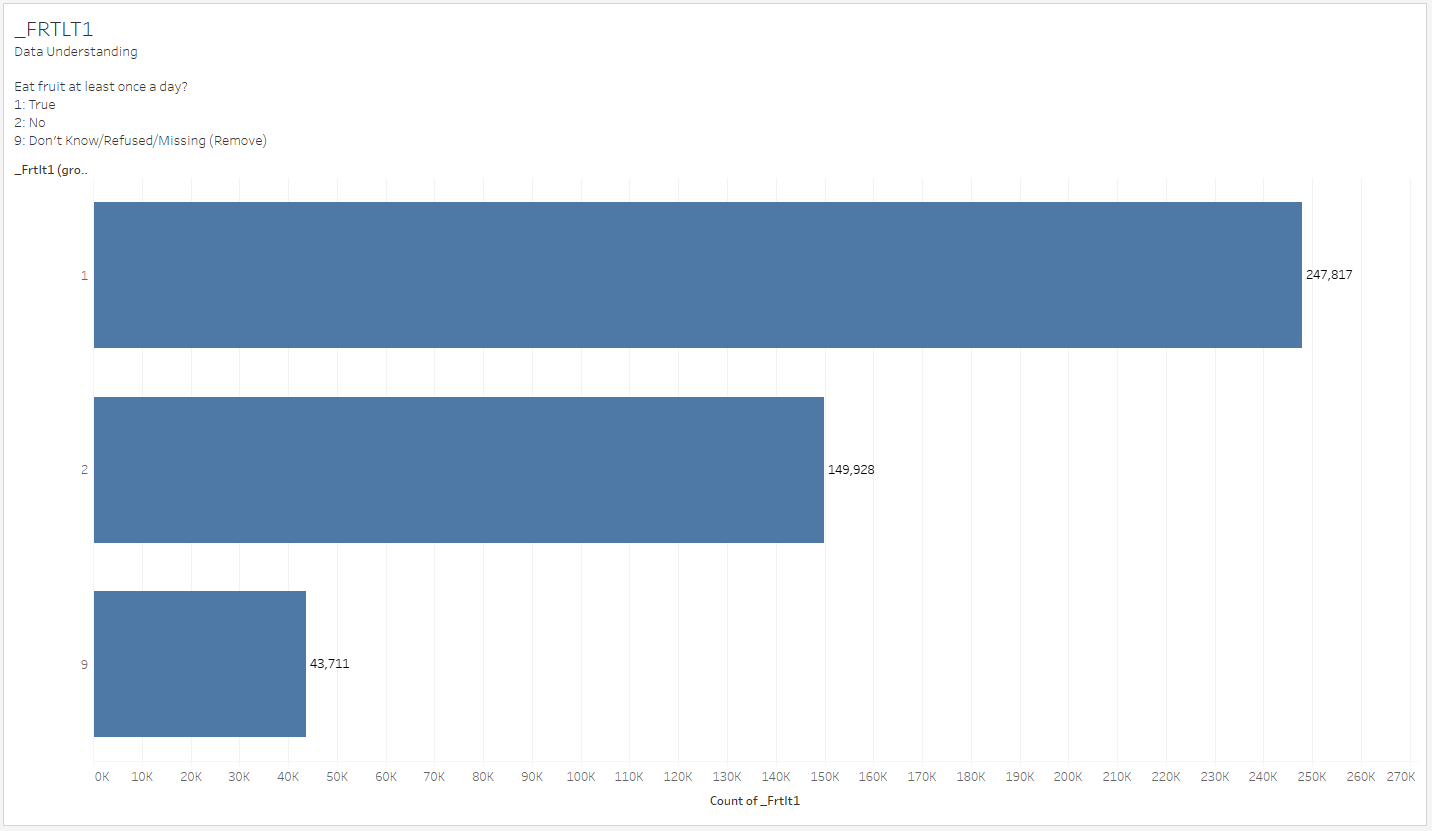


Figure 5 - Consumption of Fruits per day

**\_VEGLT1** – Consume vegetables 1 or more times per day

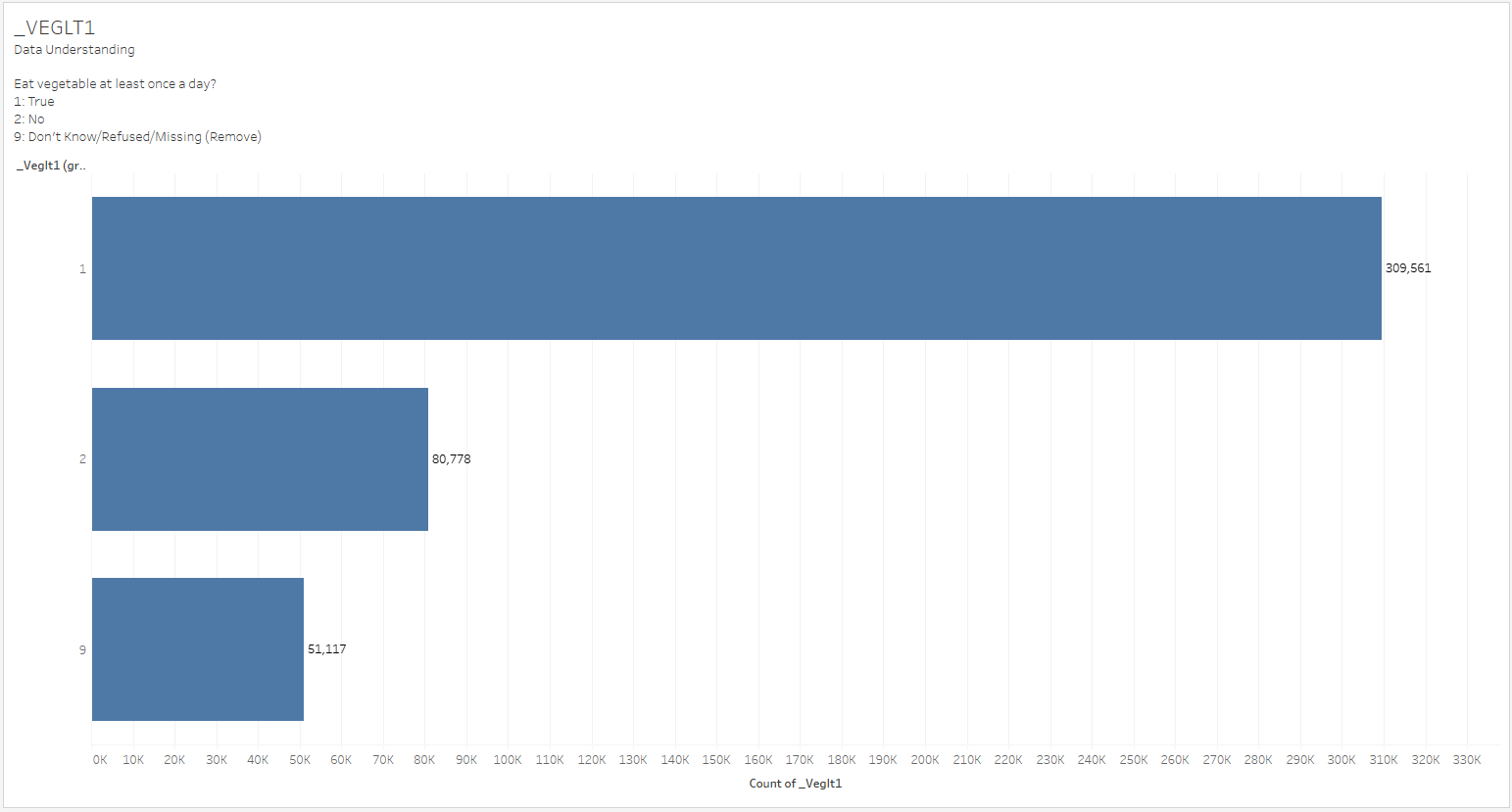


Figure 6 - Consumption of Vegetables per day

**Figures 5 & 6 tell us:**

Majority consumes fruits and vegetables multiple times per day. Respondents who consume fruit and vegetables ***one or more times* per day** are represented by **Value 1,** while respondents who consume food and vegetables **less than once per day** are represented by Value 2.

Responses with the input **"Missing," "Don't Know," or "Refused"** are **assigned the value "9**." As a result, **these values will be removed** to help us improve our modelling. With this, the data would then be **left with values 1 and 2** which **leads us to have a binary outcome**. The value 1 and 2 will be changed to a “True” or “False” or “0” and “1” type of answer.

As we all know the food you put into your body today has an impact on your health tomorrow, according to the saying that goes **"You are what you eat."** After seeking professional input and opinions, **diet is a factor that influences diabetes**. As a result, we chose the variables **\_FRTLT1 and \_VEGLT1** as part of our data modelling.

1. Alcohol Consumption (\_RFDRHV5)

Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)

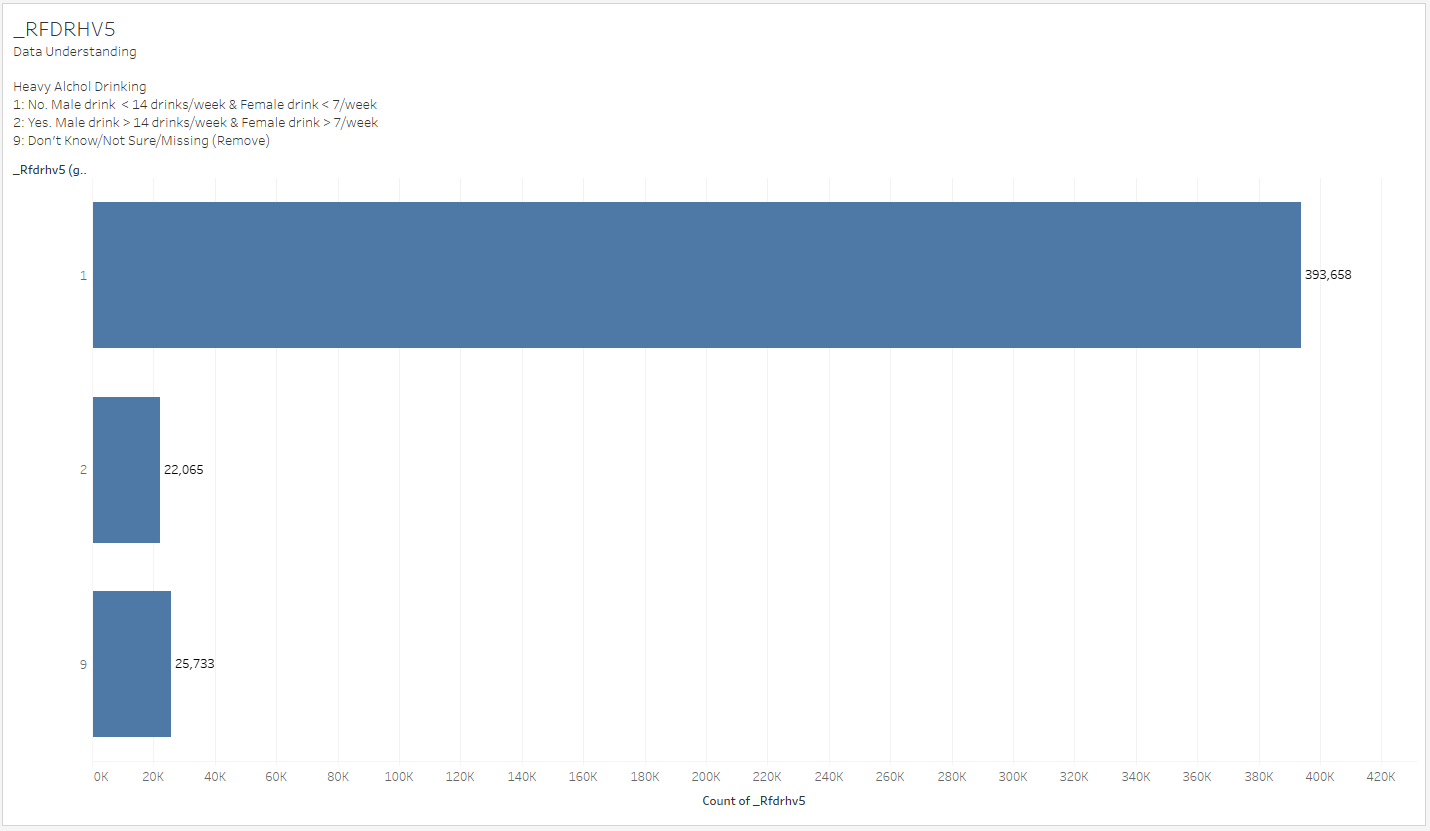


Figure 7 - Alcohol Consumption

According to Figure 7, the **majority of the records are respondents with responses “1”** of **not drinking** alcohol that exceeds **14 drinks per week for *males***and **7 drinks per week for *females***. In addition, the bar chart contains the value **"9," indicating that respondents who answered "Don't Know"/"Not Sure"/"Missing"** will be removed.

Through our research, we discovered that the **combination of diabetes medication and alcohol can result in an insulin shock**. As a result, we chose this variable for our modelling becausedrinking alcohol increases one's risk of developing diabetes.

1. Smoking

Smoke Frequency (SMOKDAY2)

Do you now smoke cigarettes every day, some days, or not at all?

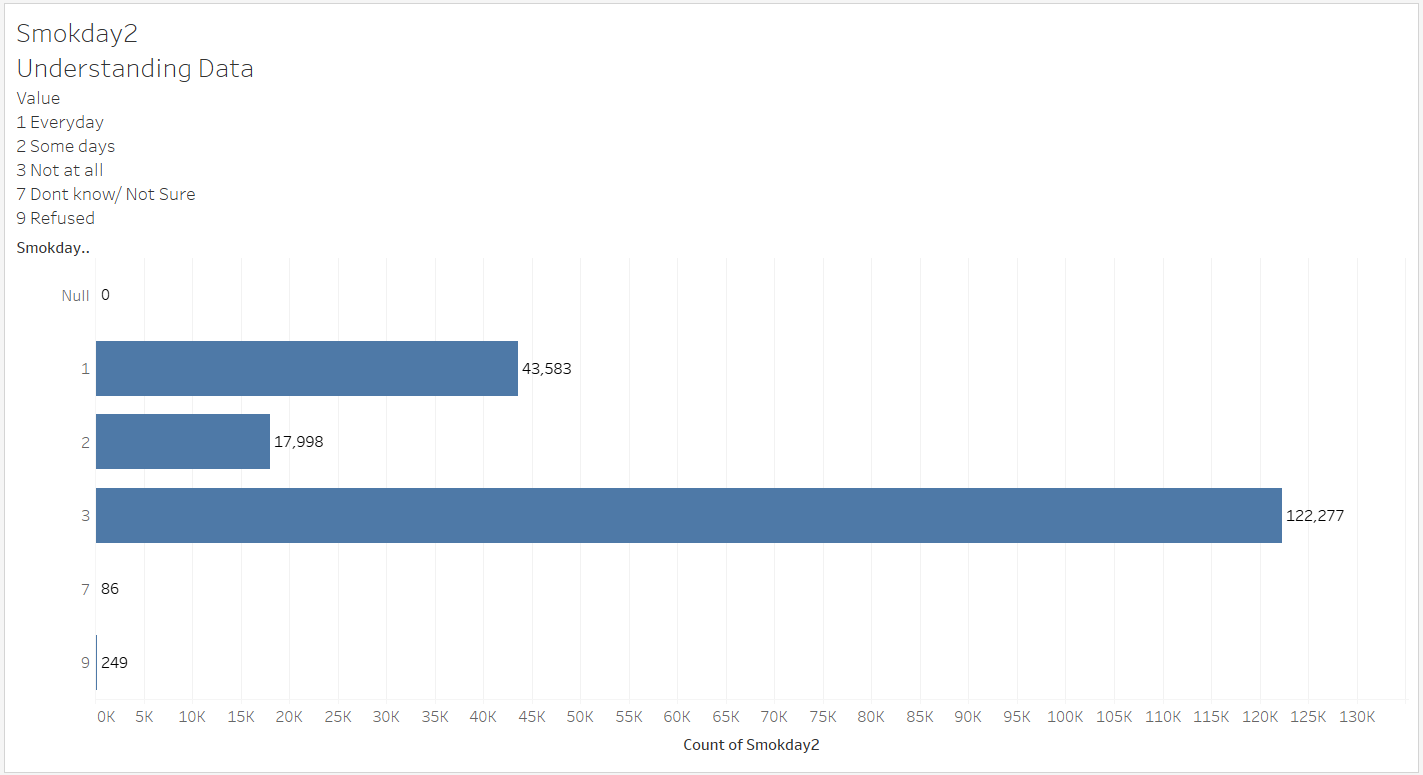


Figure 9 - Smoking Frequency per day

Figure 9 shows that the majority of respondents do not smoke now ("3”). Following that would be respondents who **smoke every day (“1”) with the second-highest number of responses,** which accounts for approximately **⅓ of those who do not smoke**. Lastly, ("2") with the smallest group of 17,998 respondents who smoke **some days**. Values such as "Null," "7," and "9" are considered unnecessary, and **rows with input "Null," "7," and "9" will be removed**.

Smoking Status (\_SMOKER3)

Four-level smoker status: Everyday smoker, Someday smoker, Former smoker, Non-smoker

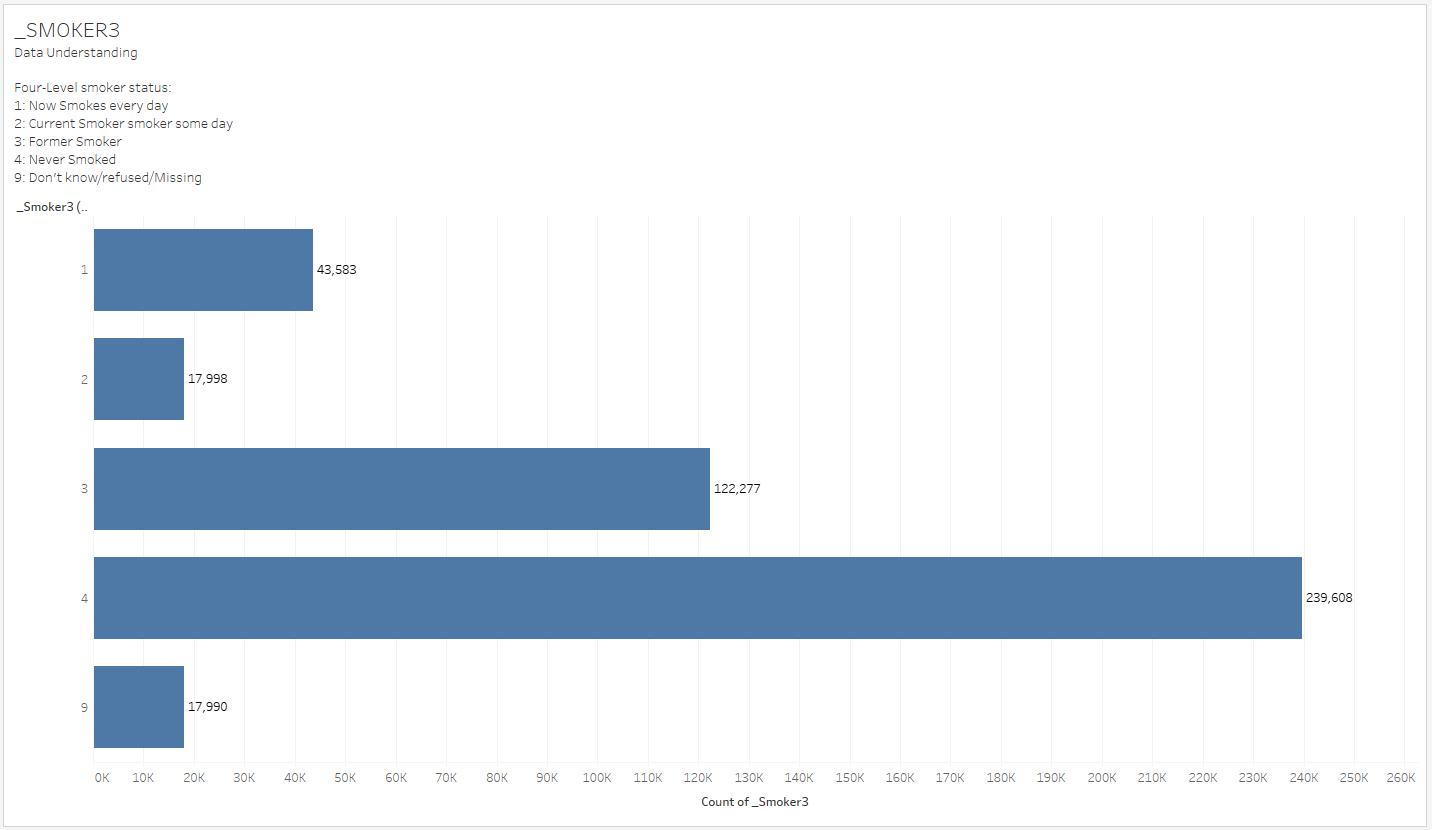


Figure 10 - Responders Smoking Status

Figure 10 shows **five categories** of responses, including the ("9") that represents "Don't Know"/"Missing" and will not be considered. Most records of respondents who **do not smoke are ("4")** and **former smokers("3")** are about half the population of **nonsmokers("4")**, followed by smokers who **smoke daily("1")** and smokers who **smoke some day(“2”)**.

*Summary for Figures 9 and 10:*

It is critical to obtain information about an individual's smoking status and frequency. **Suggestions for a lifestyle change** that would be **made easier** with the help of this type of data for respondents suffering from diabetes.

1. Physical/Mental Health

Physical Activeness (\_PACAT1)

Physical Activity Categories

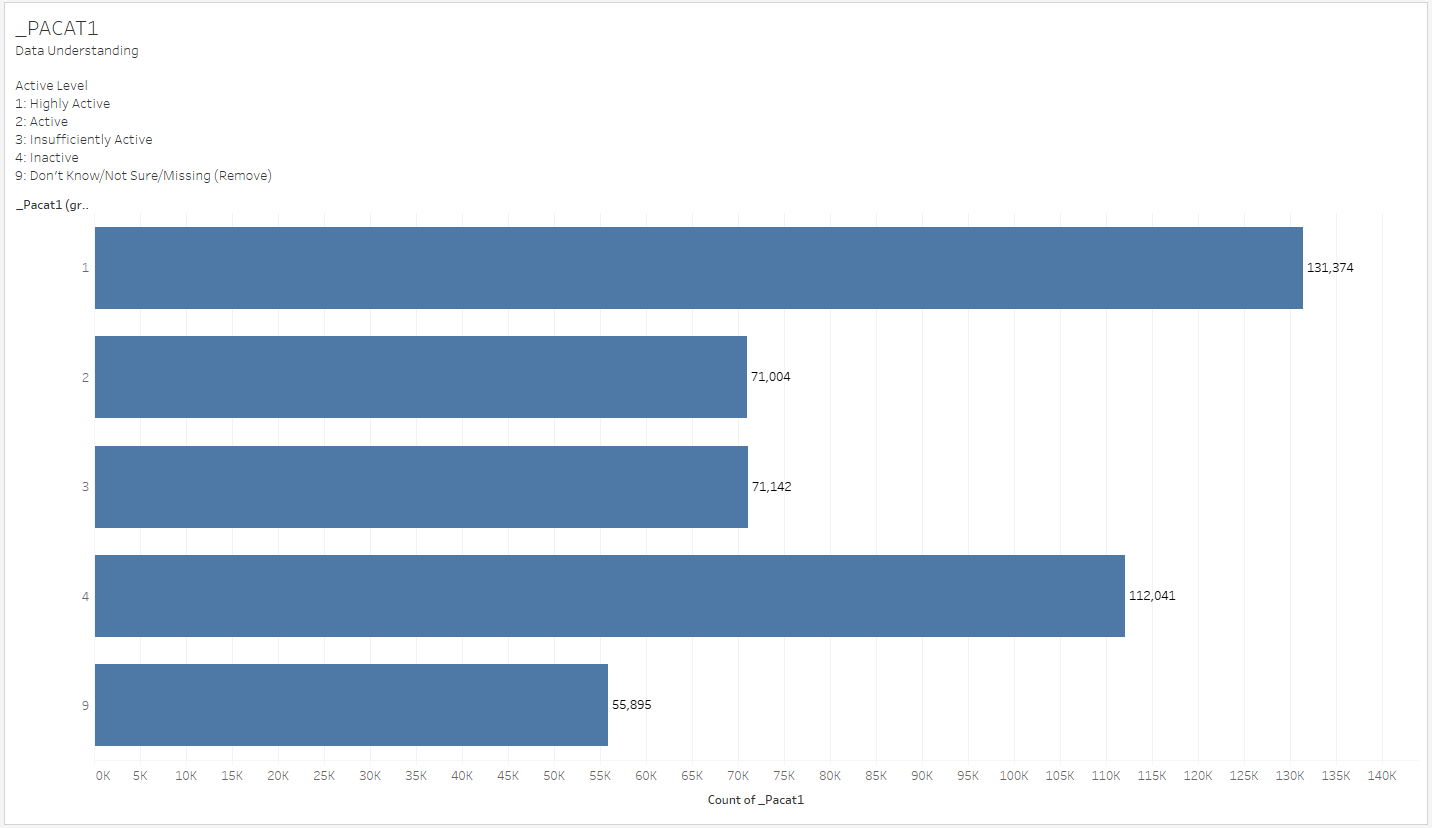


Figure 11- Physical Activeness

Figure 11 shows that there are **5 categories of responses**, with the largest group of 131,374 respondents saying they are **highly active("1")** in physical activities, followed by the next largest group of 112,041 respondents saying they are **inactive("4")**, and respondents who said they are **active("2")** and **insufficiently active("3")** having roughly equal size of data around 71,000. A **response of "5"** is found, which means **"Don't Know," "Not Sure," or "Missing,"** and rows that consist of these values will be removed.

Difficulty in concentration, remembering and decision making. (DECIDE)

Because of a physical, mental, or emotional condition, do you have serious difficulty concentrating, remembering, or making decisions?

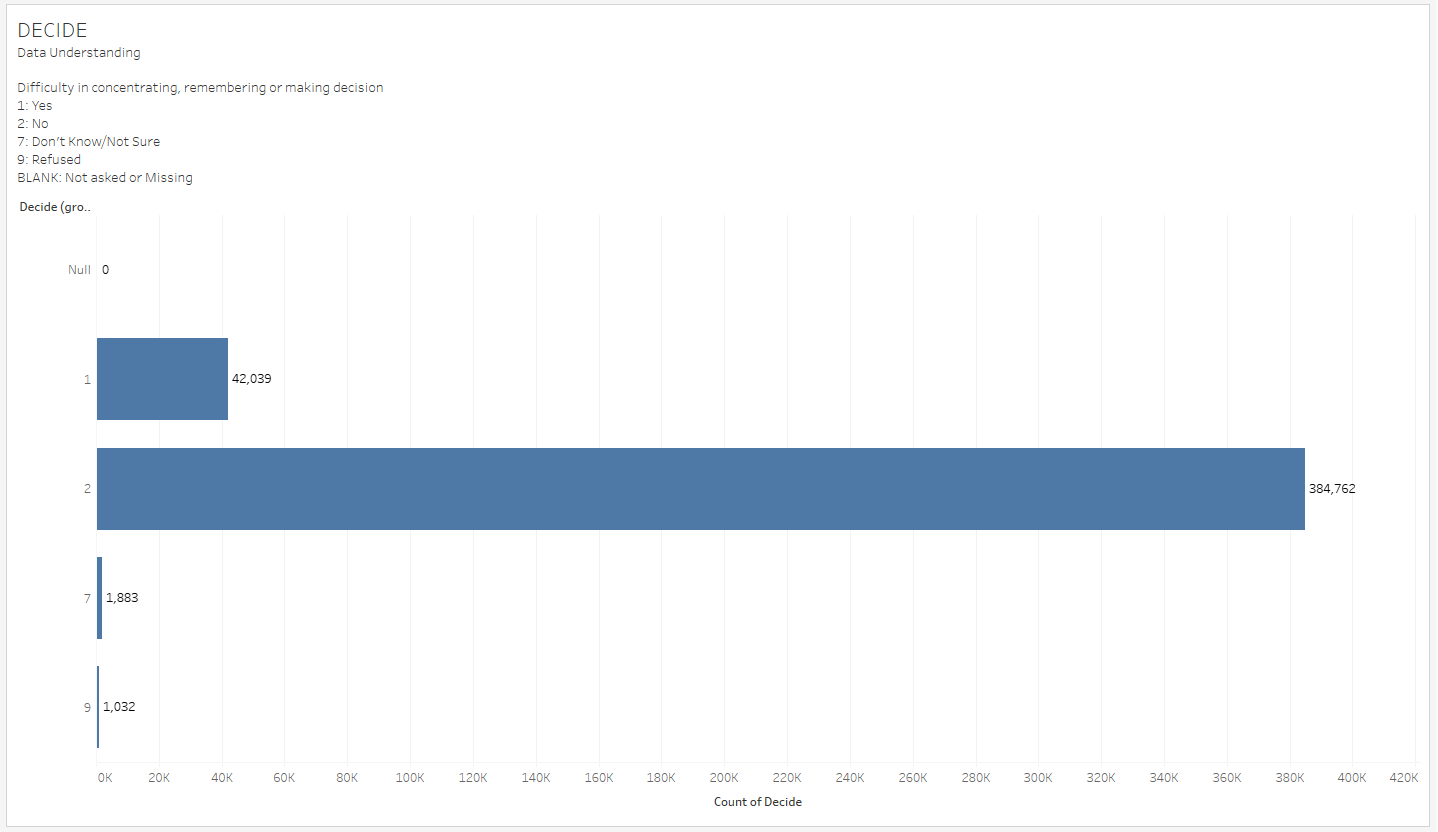


Figure 12 - Difficulty in concentration, remembering and decision making.

Figure 12 shows that the majority of respondents have no trouble remembering, concentrating, or making decisions. **Value "1" indicates difficulty** in concentrating, remembering, and making decisions. The **value "2" indicates that there is no difficulty** concentrating, remembering, or making decisions. The **values "7" and "9" represent "don't know/not sure" and "refused"** to answer respectively. When we begin modelling, this will be removed because it does not provide any insightful results.

Difficulty in walking (DIFFWALK)

Do you have serious difficulty walking or climbing stairs?

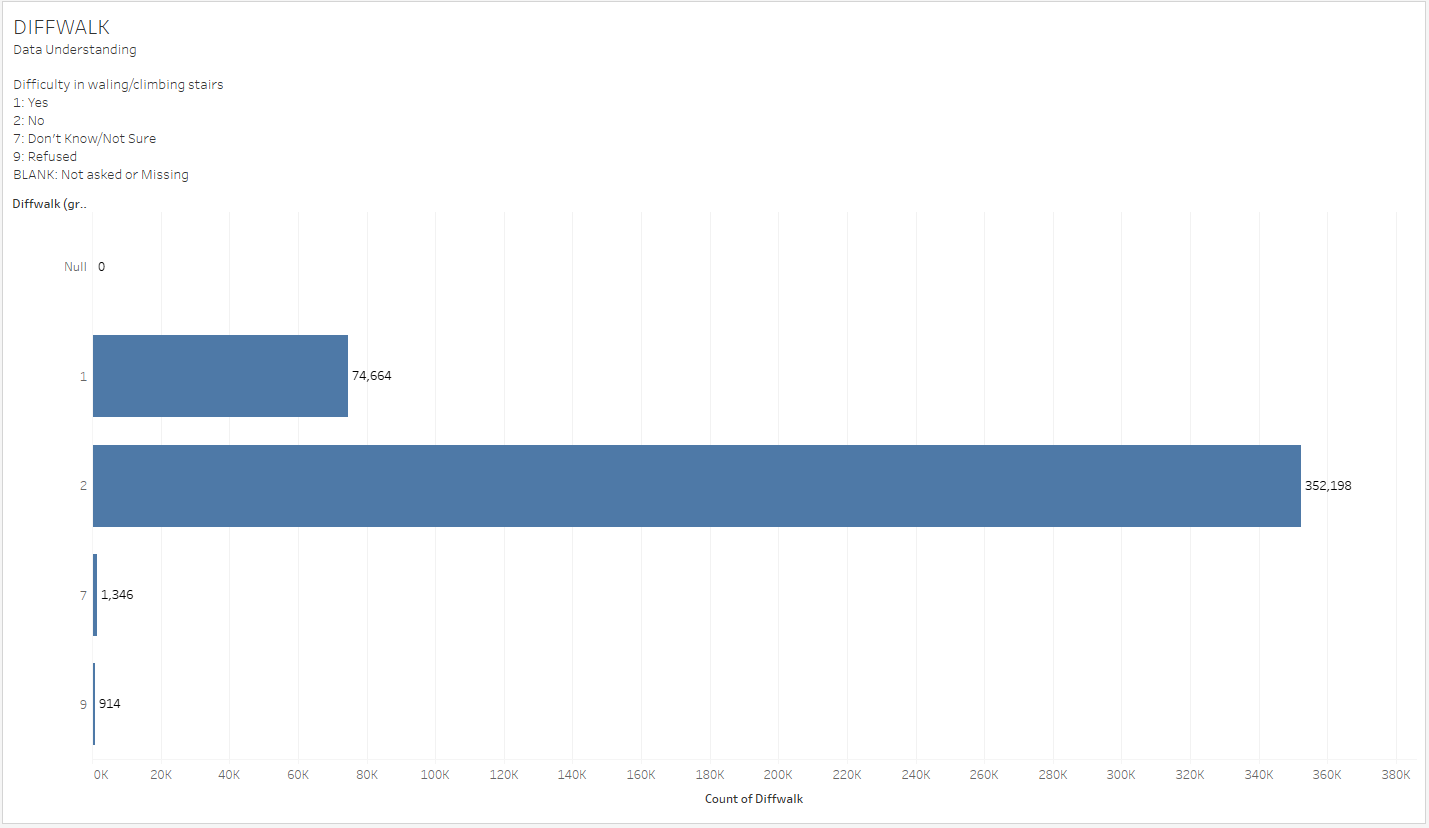


Figure 13 - Difficulty in walking or climbing up stairs

Figure 13 shows that the majority of respondents have no trouble walking or climbing stairs because Value "2" is No, which means they have no trouble walking or climbing stairs. The value "1" indicates that respondents have difficulty walking or climbing stairs. The values "7" and "9" represent **"Don't know/Not sure" and "Refused"** to answer, respectively.

Depressive Disorder (ADDEPEV2)

(Ever told) yourself that you have a depressive disorder, including depression, major depression, dysthymia, or minor depression?

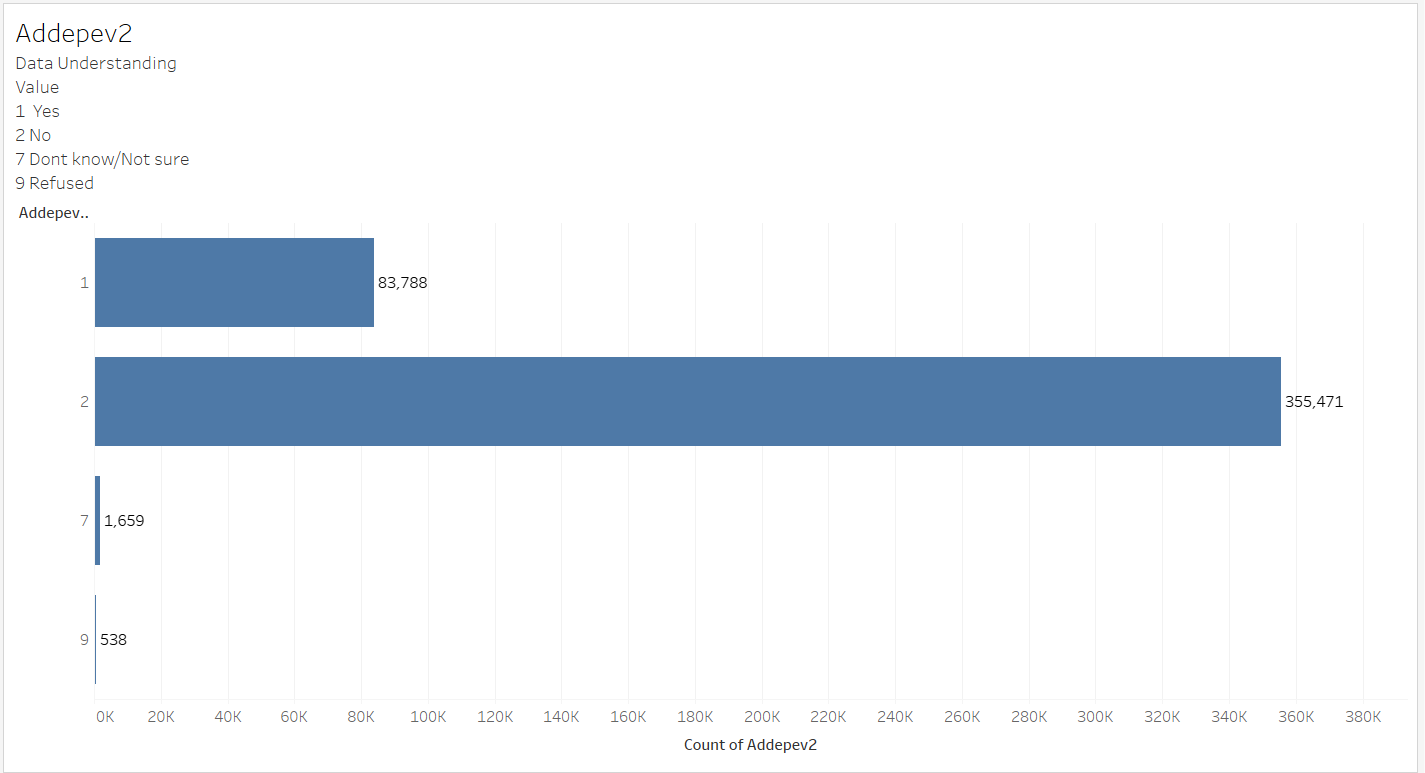


Figure 12- Having Depressive Disorder

According to Figure 12, 355,471 of all respondents do not have a depressive disorder. Value "1" indicates that respondents have a depressive disorder. Value "2" indicates that respondents do not have a depressive disorder. Similarly, the values "7" and "9" represent **"Don't know/Not sure" and "Refused"** to answer.

*Figures 9, 10, 11, and 12 are summarised as follows:*

Physical and mental health has always been important indicators of a person's risk of developing diabetes. Activeness, difficulty walking, and concentration/remembering are all valuable variables among the 330 columns. Whether respondents are in recovery or in a life-or-death situation, mental health can either improve or worsen illness status. We discovered that data on health status is critical to the success of our model.

1. High Blood Pressure (\_RFHYPE5)

Adults who have been told they have high blood pressure by a doctor, nurse, or other health professional

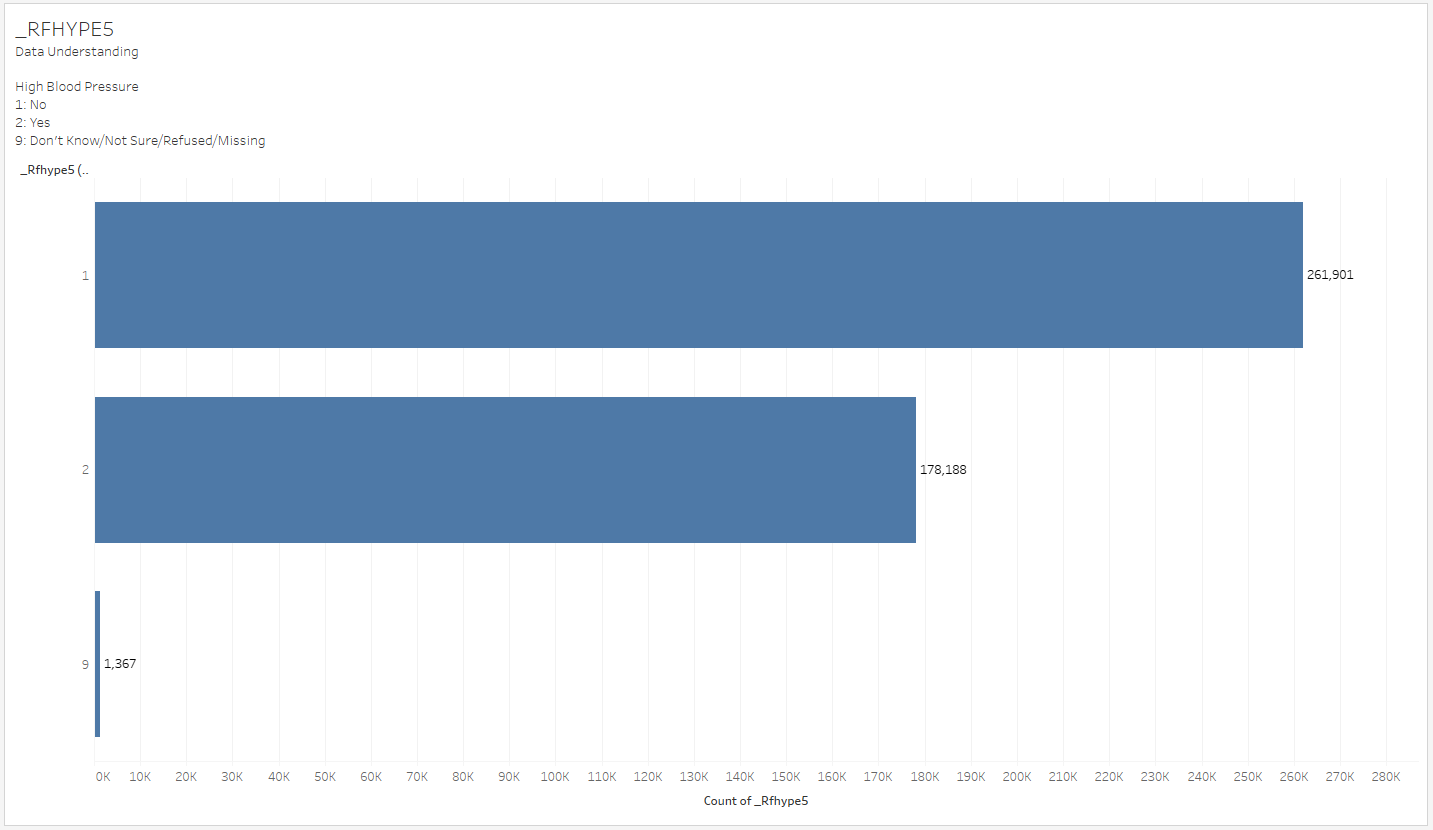


Figure 13- Reported to have High Blood Pressure

Figure 12 shows a response of "9" which represents "Don't Know","Not Sure","Refused"/"Missing" and will be removed, leaving a binary set of data with 261,901 records indicating that the respondents have no high blood pressure ("1") and 178,188 respondents with high blood pressure ("2").

During the cleaning process, binary records would be changed by standardising the value 0 to mean "no" and 1 to mean "yes." The reason is that in the codebook, binary record value 1 can be defined as "yes" or "no" in different columns.

1. High Cholesterol (\_RFCHOL)

Adults who have had their cholesterol checked and have been told by a doctor, nurse, or other health professionals that it was high

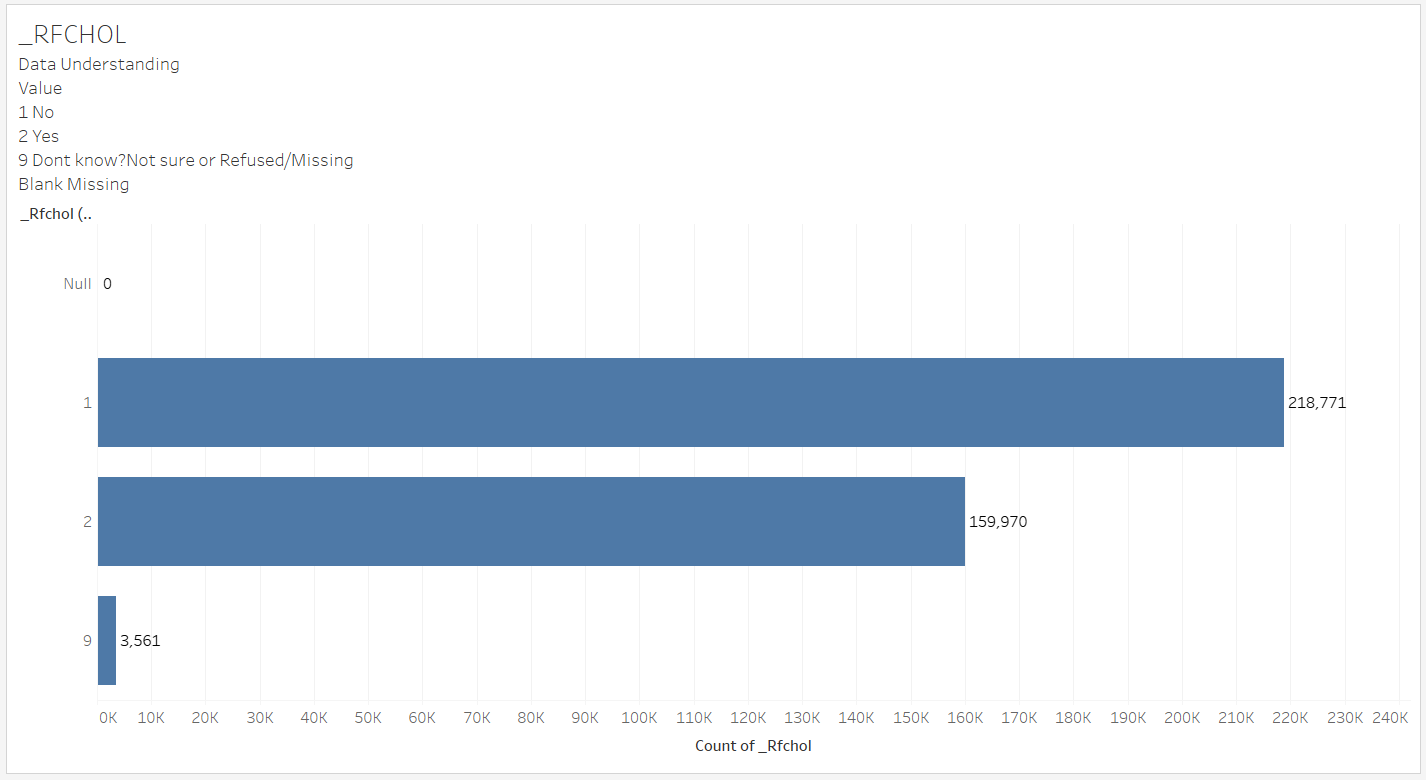


Figure 14- Reported to have High Cholesterol before

Figure 14 shows that 210,771 respondents were not informed that their cholesterol level was high by a doctor, nurse, or other health professional. 3,561 respondents do not know, are unsure, refused to answer, avoided answering or were not asked this question, resulting in missing values. Respondents who said no to being told they have high cholesterol are assigned the value "1." Respondents who said yes to being told they have high cholesterol are assigned the value "2." Value "9" represents respondents who have refused to answer, do not know/are unsure of an answer, or have not been asked the question, resulting in the input being missing.

1. Other Chronic Health Conditions

Coronary Heart Disease (CVDCRHD4)

(Ever told) you had angina or coronary heart disease?

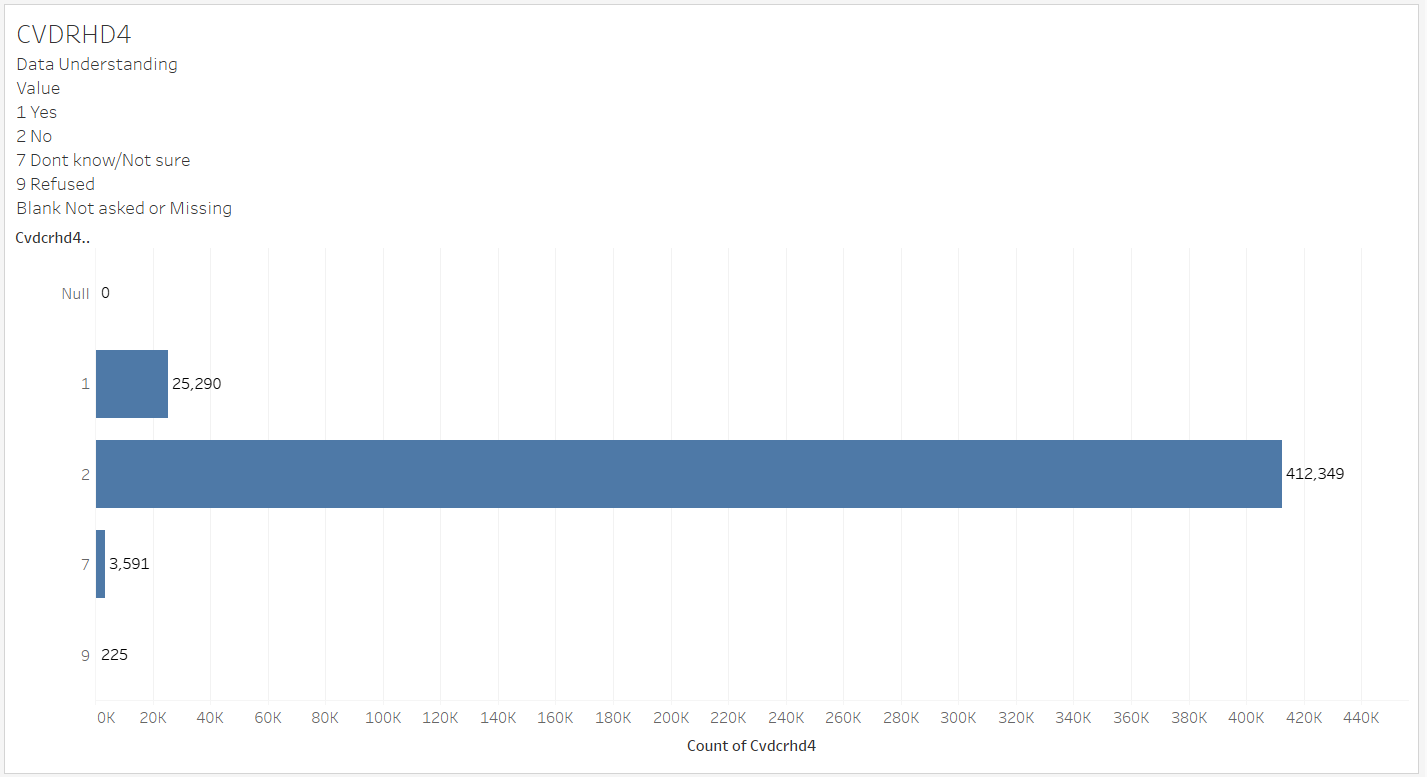


Figure 15- Had a past history of coronary heart disease?

Figure 15 shows that at least 85 percent of all respondents did not have coronary heart disease. Only 25,290 respondents were diagnosed with coronary heart disease. Respondents who said they **have coronary heart disease** are represented by the value "1." Respondents who answered no, indicating that they **do not have coronary** heart disease, are represented by the value "2." The values "7" and "9" stand for **"Don't know/Not Sure" and "Refused"** to answer, respectively.

Arthritis (HAVARTH3)

(Ever told) you have some form of arthritis, rheumatoid arthritis, gout, lupus, or fibromyalgia?

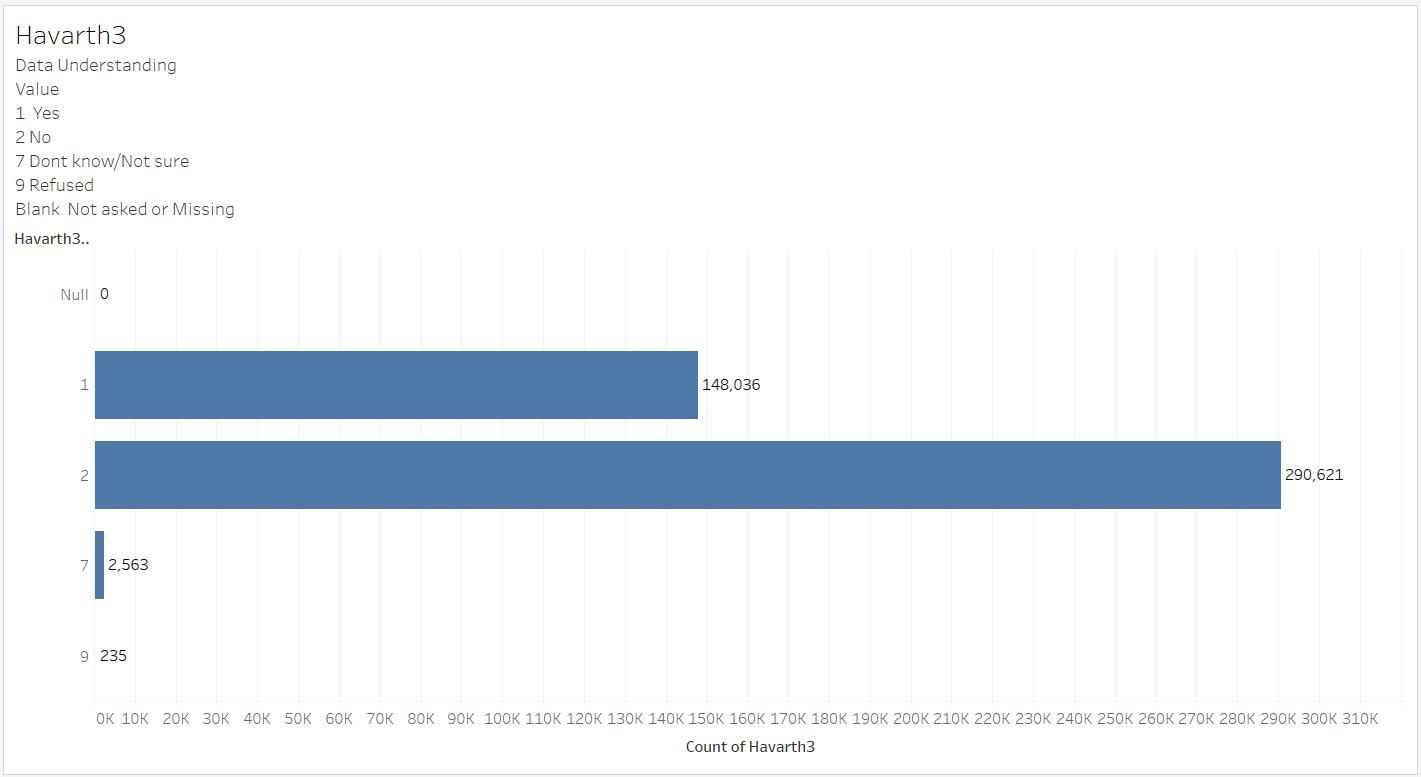


Figure 16- Arthritis Diagnosis ( rheumatoid arthritis, gout, lupus, or fibromyalgia)

Figure 16 shows that 290621 respondents **do not have** arthritis. Arthritis has been **diagnosed** in 148036 of the respondents. Value "1" represents respondents who answered **yes** when asked if they had been informed that they had some form of arthritis. Value "2" represents respondents who answered **no** and were not informed that they had some form of arthritis. The values "7" and "9" stand for **"Don't know/Not Sure" and "Refused"** to answer, respectively.

Kidney Disease (CHCKIDNY)

(Ever told) you have kidney disease? Do NOT include kidney stones, bladder infection or incontinence. (Incontinence is not being able to control urine flow.)

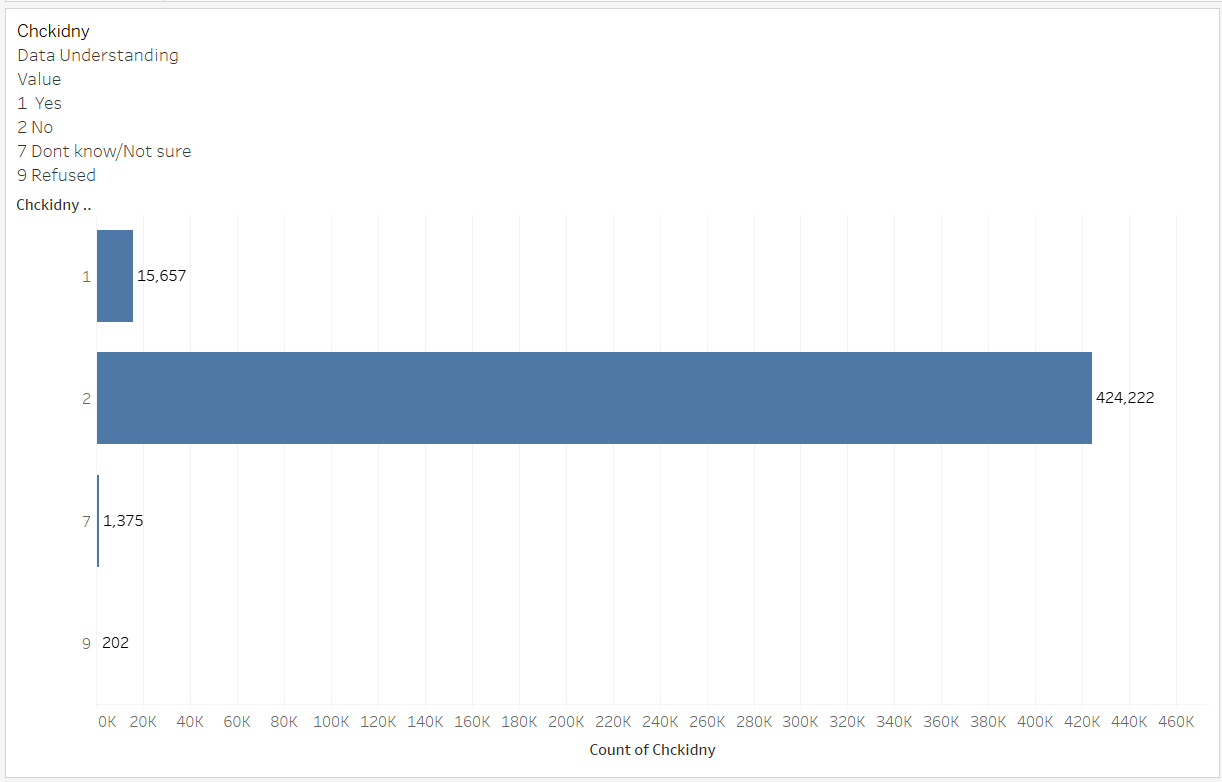


Figure 17- Down with having kidney disease

According to Figure 17, 424222 respondents did **not have kidney disease**. There are 15657 respondents who have found out they **have kidney disease**. Respondents who indicated that they were aware of their kidney disease were assigned the value "1." Value "2" represents respondents who answered no to the question are those who do not have kidney disease. The values "7" and "9" stand for **"Don't know/Not Sure" and "Refused"** to answer, respectively.

In summary, respondents who do not suffer from a depressive disorder are less likely to suffer from a chronic illness. Because the mind influences the body, whether the body is prone to easily becoming ill or being diagnosed with a common health problem.

1. Conclusion

From data understanding, we have discovered that numerous variables have inputs such as **"BLANK", "Don't Know"/"Not Sure", and "Refused to answer"**. There are also values other than **"7," "9," and "BLANK"** that contain inputs such as **"Don't know," "Refused," and "Missing."** This data that has been collected yields no useful results and only data with no missing values across the row would be considered valuable for supporting our modelling. In addition, these values would be **removed by changing them all to missing values** and **removing all rows that contain missing values** during the data cleaning process. We will also change the majority of variables/columns that are **binary to 1 and 0**, which represent **Yes/True** and **No/False**, respectively which will be explained during the data cleaning process.

In summary, the following columns need to be cleaned after doing visualisation on the **uncleaned** dataset:

1. Diabetes (DIABETE3)
2. Diets (\_FRTLT1, \_VEGLT1)
3. Smoking (SMOKDAY2, SMOKER3)
4. High Blood Pressure ( \_RFHYPE5)
5. High Cholesterol (\_RFCHOL)
6. Alcohol Consumption (\_RFDRHV5)
7. Other Chronic Health Conditions (CVDSTRK3, CVDRHD4, HAVARTH3)
8. General/Mental Health (DIFFWALK, PACAT1, DECIDE, ADDEPEV2)
9. Demographics (SEX, AGEG5YR, \_BMI5)

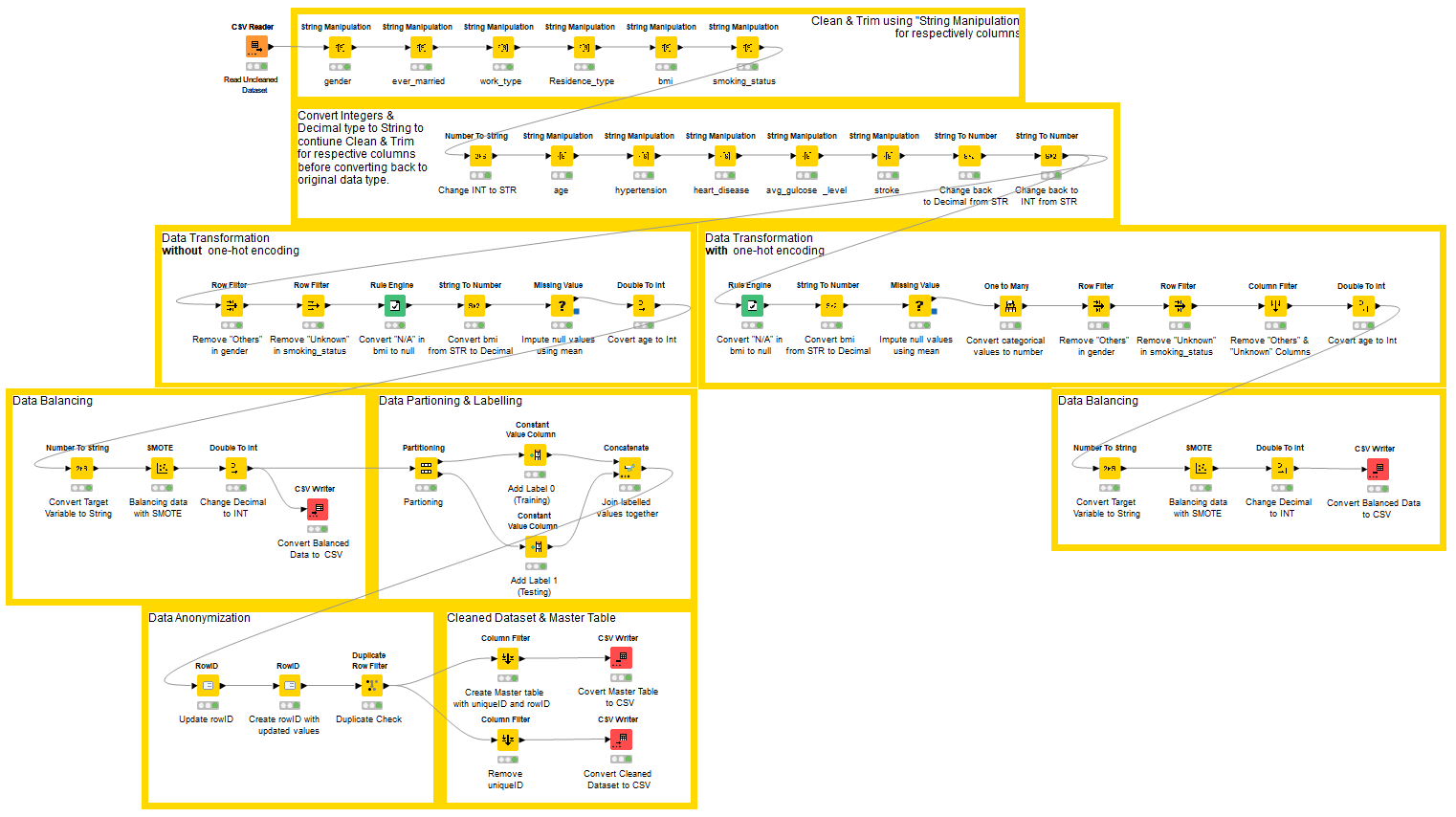
# 

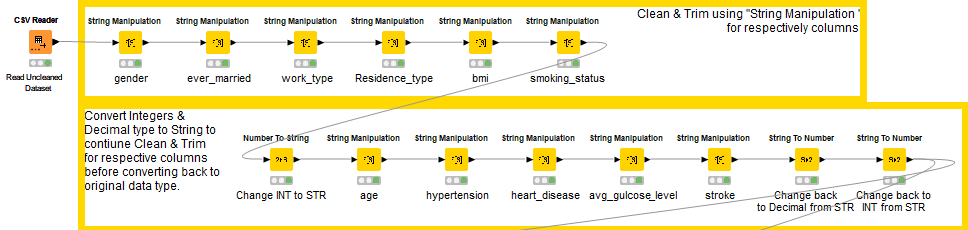
# **Data Preparation / Cleaning**

## Stroke

*This section is documented by Shermaine and Shi Min*

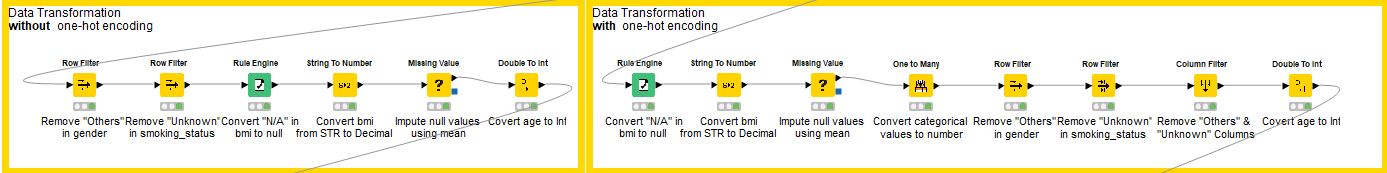
Our data is cleaned and prepared for modelling with **KNIME**. The following is an overview of the KNIME **workflow**.



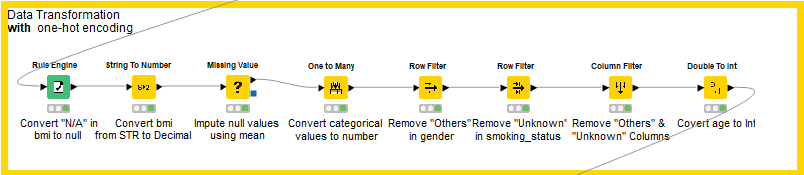
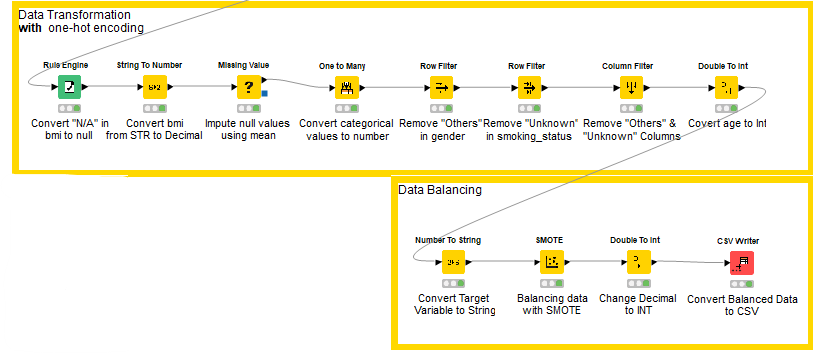
To remove **white and trailing space**, we will focus on the top 2 annotations.

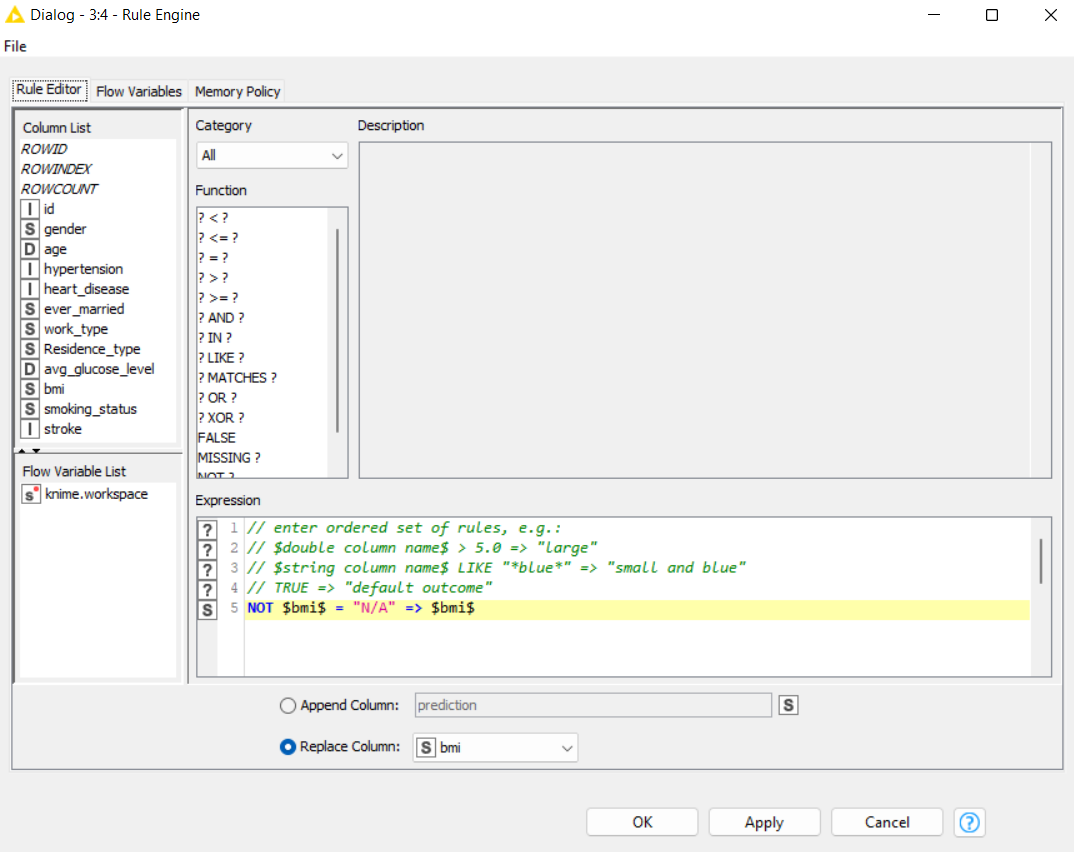
In order to ensure the **cleanliness and integrity of the dataset,** after importing an uncleaned dataset into KNIME with the "**Read CSV**" node, a best practice would be to **clean and trim** the existing columns. To accomplish this, we used the "**String Manipulation**" node. With the "**Number to String**" node, we convert columns of **decimal or integer data types to strings**. By using the "**String to Number**" node, we converted data types that had been **cleaned and trimmed** back to their **original data types**.

From Data Cleaning, we will discuss **Data Transformation, Data Balancing, and Data Anonymization**. There will be two sections in this area, one **with one-hot encoding** and one **without one-hot encoding**, to see if there is a **difference in accuracy** based on the column type and length. Using one-hot encoding, string data types are converted into numerical data types instead of allowing SAS Viya to manually convert them, which enables us to achieve the most accurate result and reduce the possibility of errors.



### **With** one-hot encoding



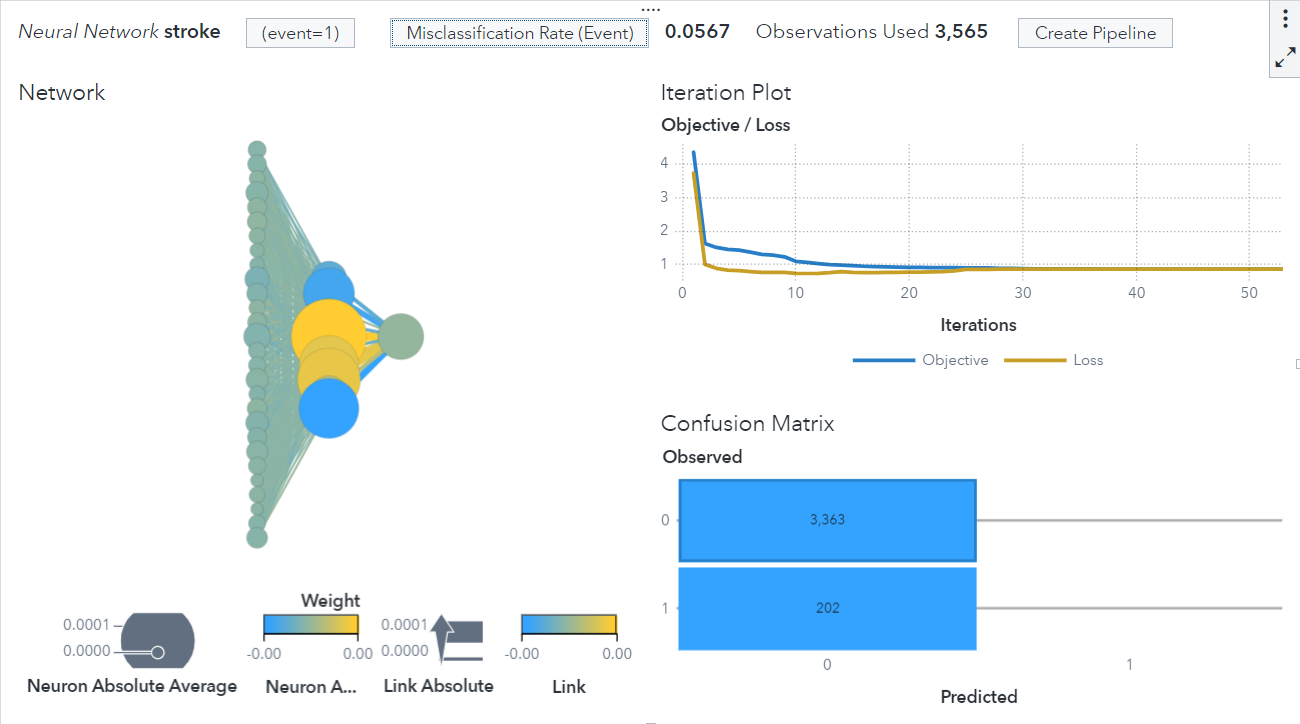
Before using one-hot encoding, we converted N/A values in bmi to **null, using the** “Rule Engine” node with the rule stating that if bmi is not “N/A”, **return the bmi value,** and if it is N/A, return null, with this statement: 

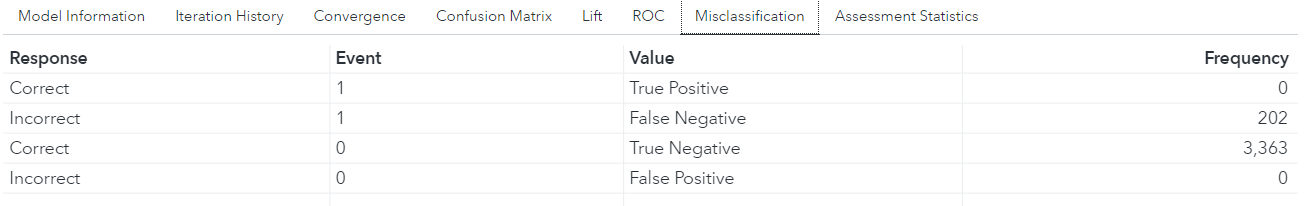
In KNIME, a null is represented using a “?” symbol. After converting N/A values to null, we converted the bmi data type to decimal and used the average bmi across the column for imputation.

Afterwards, we did **one-hot encoding** using KNIME’s **“One to Many”** node to convert **categorical** values to **numbers**. This gives us **individual categorical values** as columns with 0 and 1 values indicating true or false at each row. In our data understanding, we mentioned **removing** rows that contained **“others” in gender, and “unknown” in smoking\_status**, so we removed these rows **after** one-hot encoding. As a next step, we also further **removed “others” and “unknown” columns**, which were converted from one-hot encoding.

Before data Balancing, we decided to test our data out in SAS Viya to see if it is balanced.

#### **Need for Balancing**

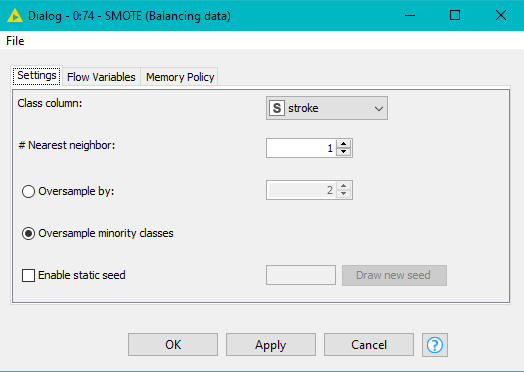




The cleaned dataset without balancing has a **misclassification** rate of only **0.0567**, meaning it has an **accuracy of about 94%**. However, this high level of accuracy is **not due to a good model,** but rather to an **unbalanced** dataset of **3363:202** of the target values, with 94% of our target values focused on patients **without stroke**. In addition, the **true positive and false positive** rates are both **zero**, further proving its **unbalance**.

When working with **imbalanced datasets**, most machine learning techniques will **ignore** them, leading to **poor performance** in the **minority** class, since the dataset always predicts the **majority** class. Therefore, we have decided to **balance** the dataset to be able to provide a model of the **highest accuracy.**

#### **Data Balancing**



After data Transformation, we decided to use **KNIME’s “SMOTE”** node to do **data balancing**. Generally, there are **two** common ways to balance datasets, namely **randomly undersampling** the **majority** class or **oversampling** the **minority** class. However, as our minority class only has **202** rows, undersampling may cause our dataset to have **too little data** for us to train, hence we chose to use **SMOTE** to oversample.

*What is SMOTE?*

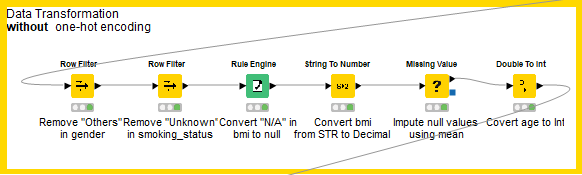
SMOTE, also known as **Synthetic Minority Oversampling Technique**, is an oversampling technique where **synthetic samples** are generated for the **minority** class. This algorithm helps to overcome the **overfitting** problem posed by **random** oversampling. It focuses on the **feature space** to generate **new instances** with the help of **interpolation** between the **positive** instances that lie together. In our case, we chose to use the “SMOTE” node configuration and **oversample** by the **minority** class, whose **stroke is 1** so that we can get **more values** for the **minority** classes and can make **better predictions** when it comes to **modelling**. Overall we were produced with a total of **6726** columns from a previous of **3565, with** a **50:50** amount on our **target variable, stroke.**

In KNIME, **SMOTE** only accepts **string** variables, so we **converted** our **target** variable, stroke, to a **string** value, using the **"Number To String"** node, **before** running the SMOTE configuration.

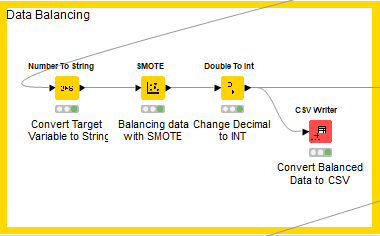
**After** running the “SMOTE” node, we also found out that **some previous data types** have been **changed**, hence we used the **“Double to Int”** node to **convert** respective columns **back** to their **initial data types.**

### **Without** one-hot encoding





For data Cleaning and data Transformation **without** one-hot encoding, the **order of steps taken differs**, but the **reasons** remain the **same**. First, we **remove others and unknown values** from **gender and smoking\_status** columns using KNIME's **"Row Filter"** node. Following that, we converted **N/A** values in bmi to null using the **"Rule Engine"** node and converted the **data type** to **decimals** so that we can do **imputations** based on the **mean** of the bmi columns. Additionally, we converted **age to integer as decimals were removed** from age columns when smoking\_status unknown was removed.

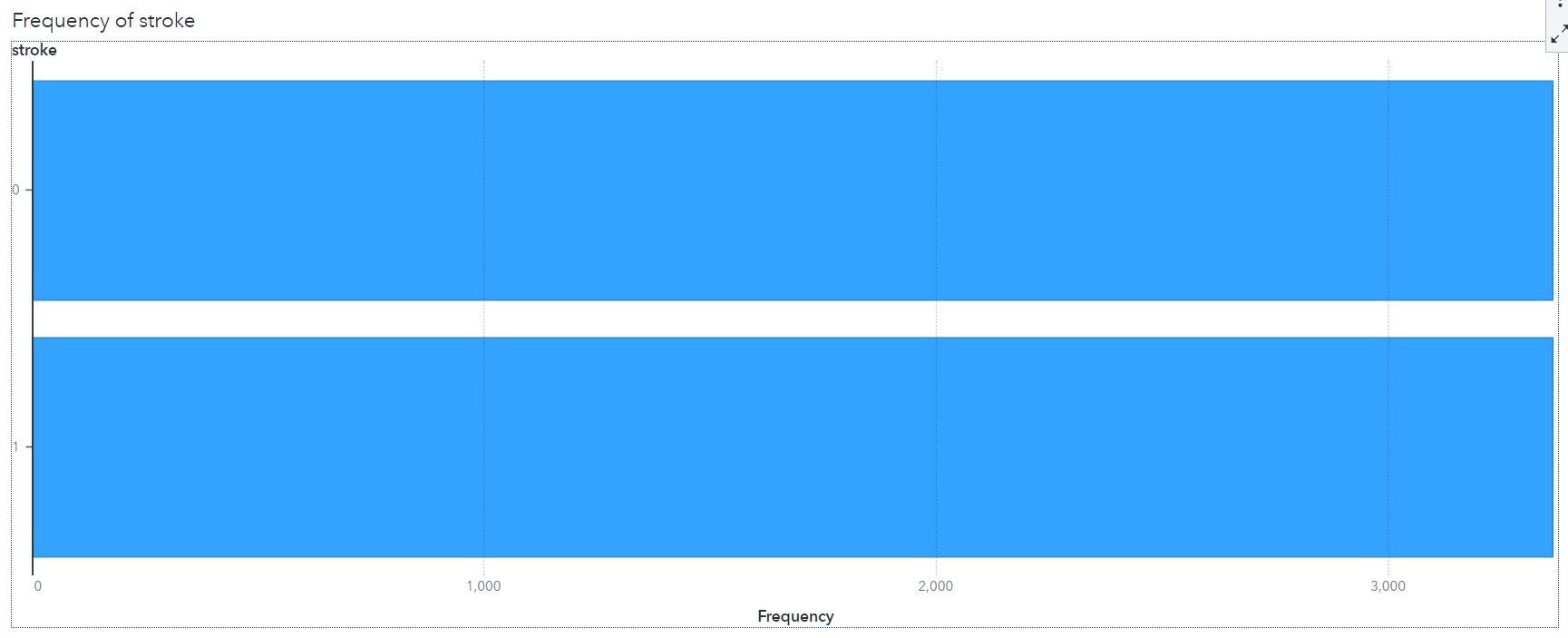


As part of **data Transformation,** we also decided to use the **KNIME "SMOTE"** node for **data balancing, which follows exactly the same steps as one-hot encoding.**

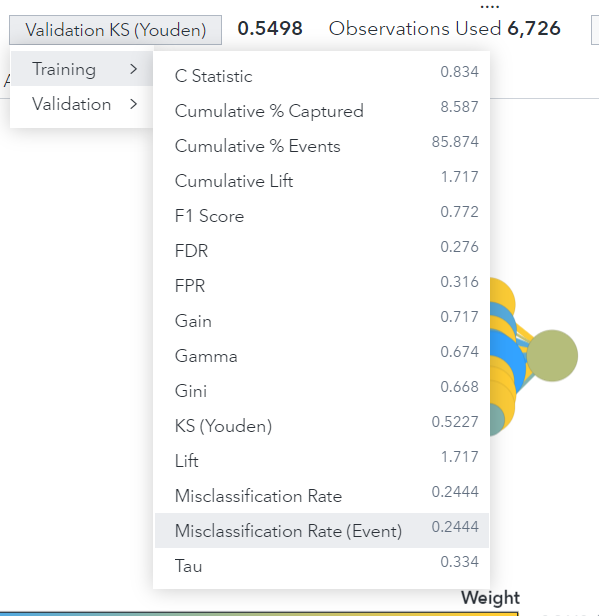
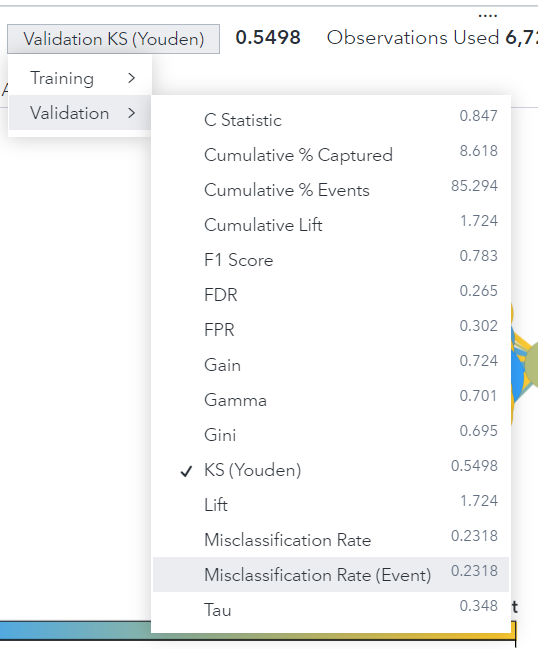
**Before partitioning and anonymising** our dataset without one-hot encoding, we used **SAS Viya** to test the **accuracy** of the **2 different datasets**. Based on the data **output** for both with and without one-hot encoding, we **compared the usability and accuracy** of both to **decide** on our **final dataset** for training and validation.

#### 

#### **Proving Dataset Balanced**



## 

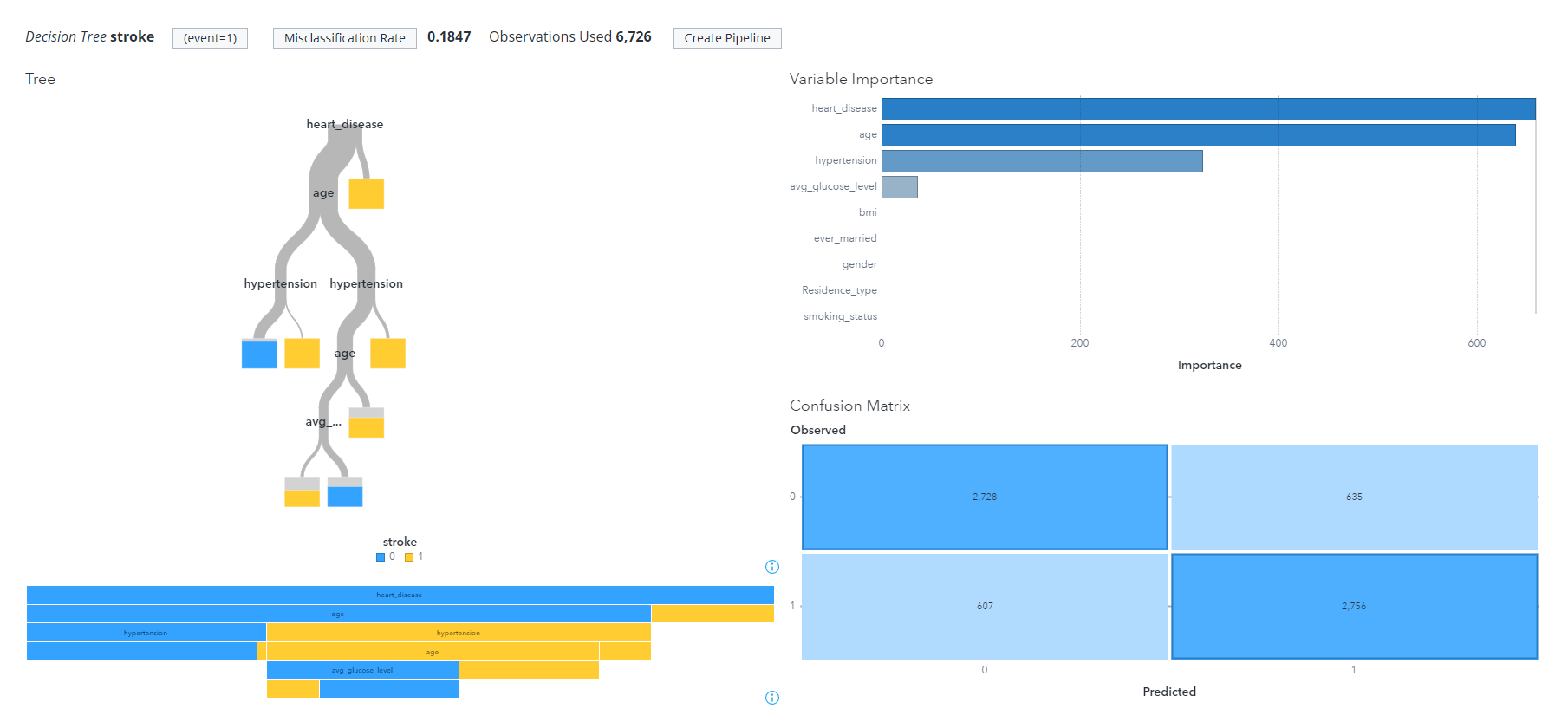
 

**After balancing** with SMOTE & partitioning our dataset, the above further **proves** our dataset is **balanced**, with a bar graph showing the **same amount of values for both values in the target variable** and the misclassification rate.

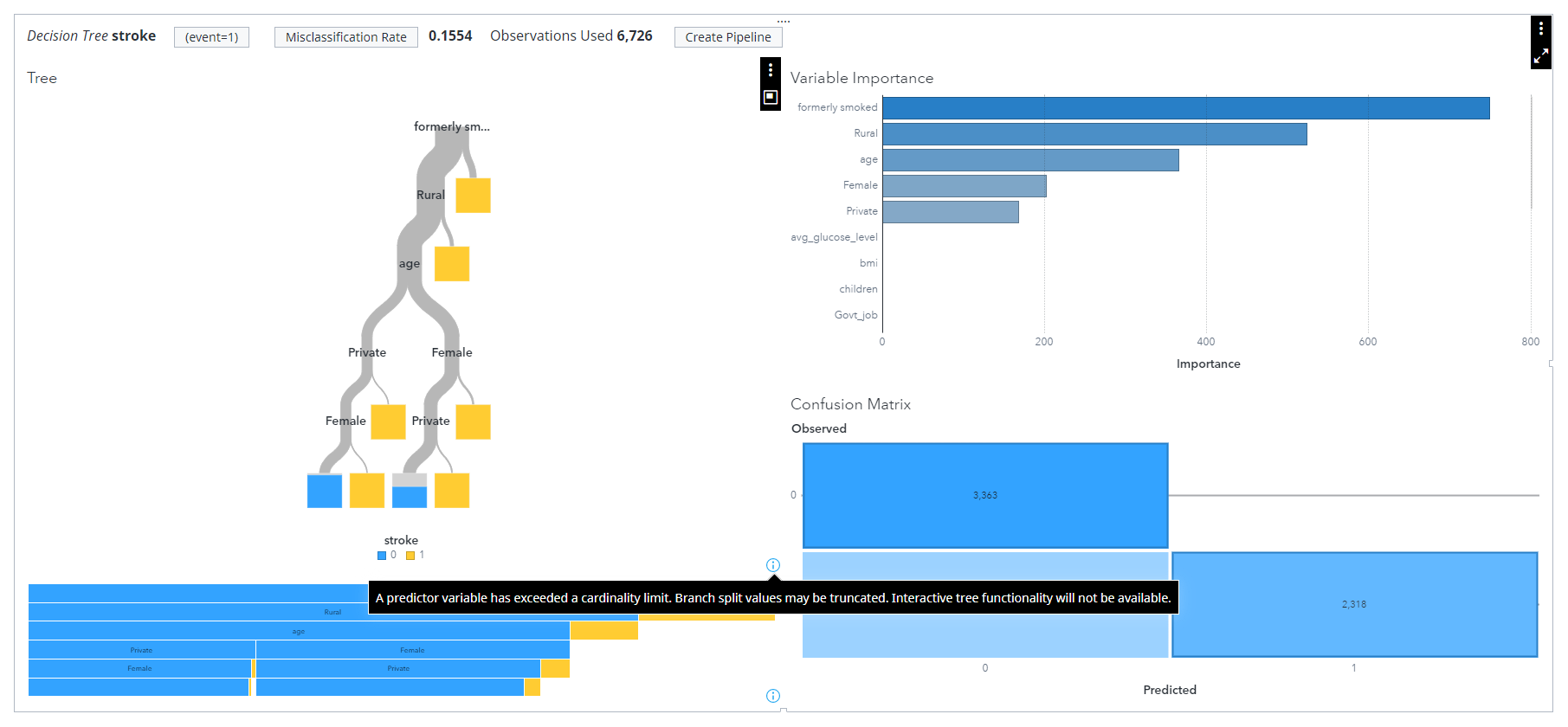
From the bar chart, we can see that both 0 & 1 values of stroke have an **equal frequency of 50%** of **3363** each. Additionally, the **misclassification of 0.2140**, though **lower in accuracy,** it proves that our data is **not biased.** Finally, both the **misclassification** rate of **training and validation is similar**, with 0.2444 and 0.2318, hence proving that they are **not overfitted.**

### **Selecting Dataset (One-hot Encoded vs Original)**

By **converting our categorical values to numerical** with one-hot encoding, we tested both original and one-hot encoded datasets together, to **determine** which dataset was more **accurate**. Based on comparisons between balanced with SMOTE with one-hot encoding and without one-hot encoding datasets, we made the **final decision** to use the dataset **without one-hot encoding. Prior** to inputting the variables, we **converted** the **target variable** and the **one-hot encoded values** to **categorical**.

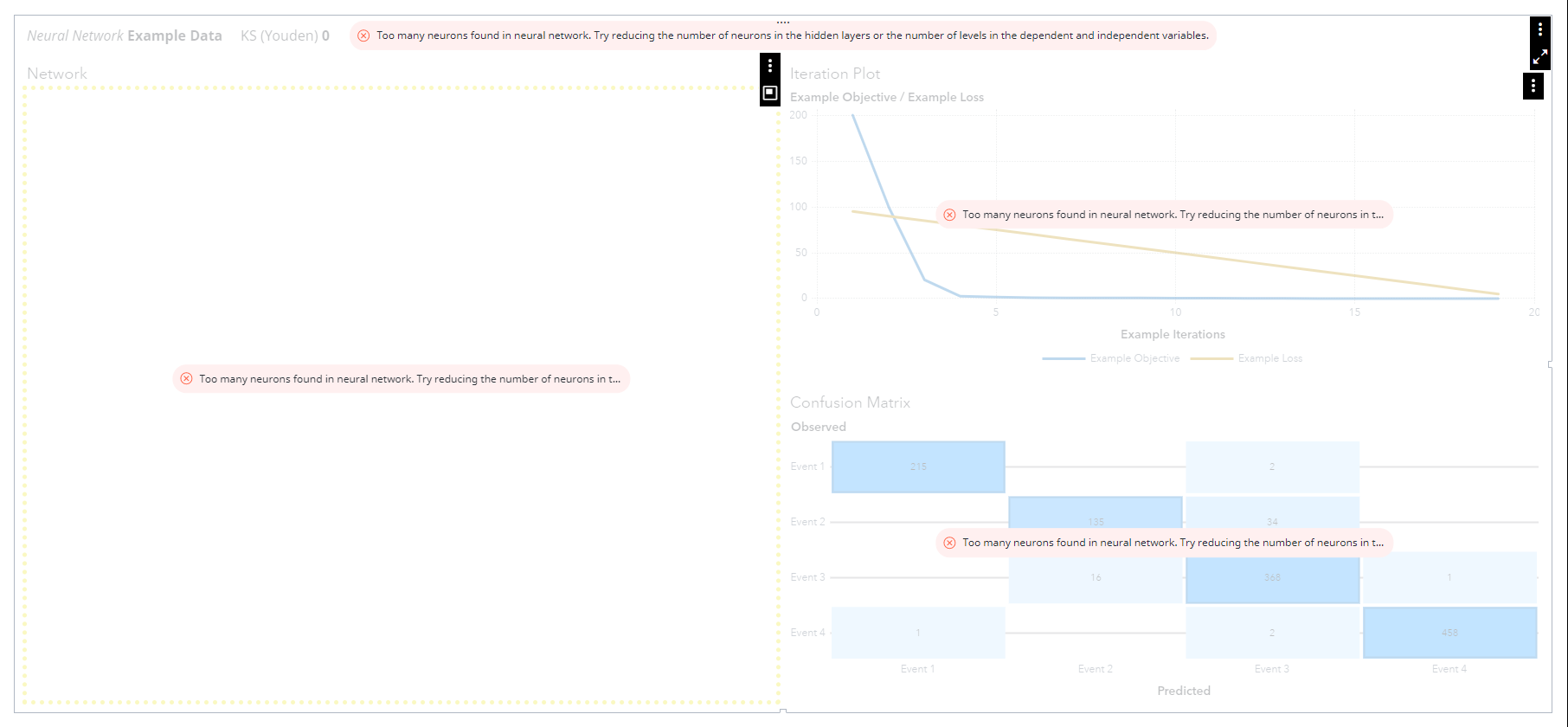


Decision Tree [Balanced with SMOTE without one-hot encoding]

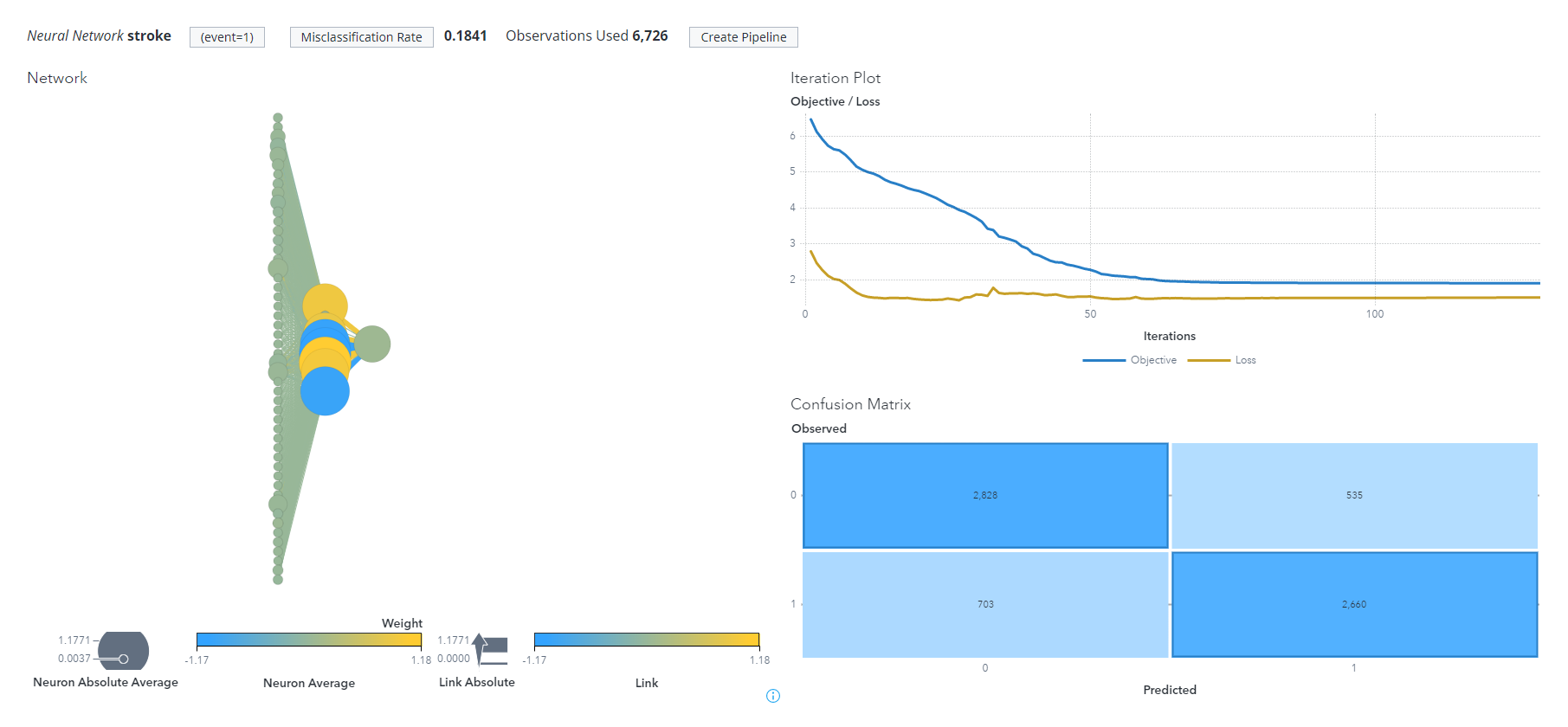


Decision Tree [Balanced with SMOTE with one-hot encoding]

From the above 2 Decision Tree models, we see that the **one-hot encoded** dataset has a **higher** accuracy of **84.5%** as compared to the **non-one-hot encoded** dataset with an accuracy of **82.5%** accuracy. However, the one-hot encoded dataset provides us with an **error** which says that **“a predictor variable has exceeded a cardinality limit”**, as well as a **missing false positive.**



Neural Network [Balanced with SMOTE with one-hot encoding]



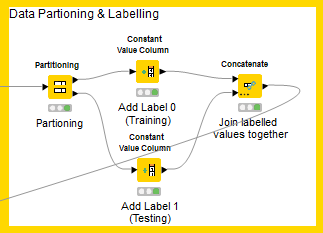
Neural Network [Balanced with SMOTE without one-hot encoding]

From the next 2 Neural Network models above, we also see that a **neural network model** is **unable** to be **built** with the one-hot encoded dataset as there are **too many neurons.** However, the **original** non-one-hot encoded dataset is able to **build the model normally.**

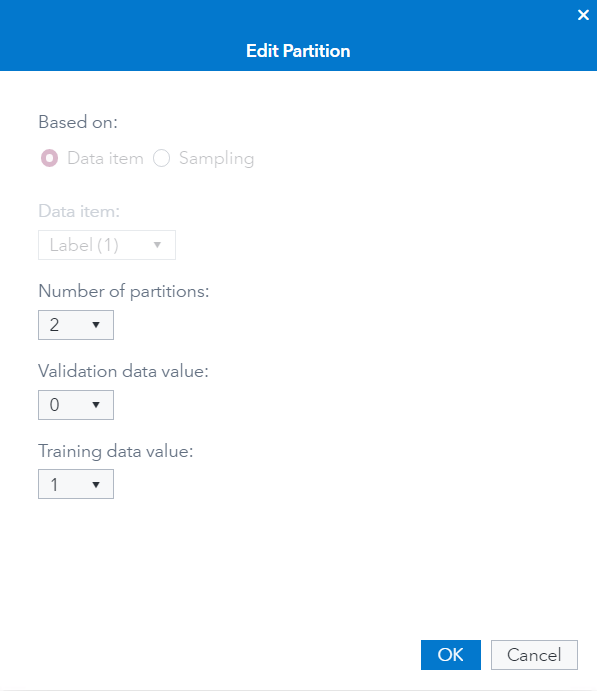
***Conclusion***

Based on the **misclassification** rate for both datasets, we found **similar accuracy** between the two, with the **one-hot encoded** dataset with **slightly higher accuracy**. However, due to a **large number** of **variables** in the one-hot encoded dataset, **some models are unable to model it**. As a result of this main **setback**, our team decided to use the dataset that **isn't one-hot encoded.**

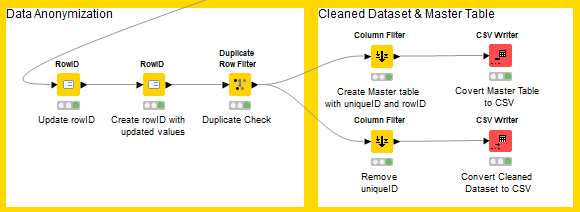
### **Partitioning Dataset**



After selecting the dataset, we did **data partitioning and data labelling** using a **“Partitioning”** node to **ensure** that our **training and validation** data will have a **balanced** amount of our target variables. With this, we managed to **partition our data into 80% training data and 20% testing data**, which is also proven to be the **most effective** in training a model. We then **labelled** our partitioned data into **0s and 1s**, using the **“constant value”** node, where **0 represents training data and 1 represents testing data**. Afterwards, we **combined** the two labelled datasets into one, using the **“Concatenate”** node, so that we can get a **single dataset** to be inputted into SAS Viya which will then be used to **partition** our data and **validate** our model.



By using **SAS Viya’s new data item,** we partitioned our dataset by using the **label data item**, which is what we have **previously split & labelled** in KNIME so that we can **compare our training and validation models.**



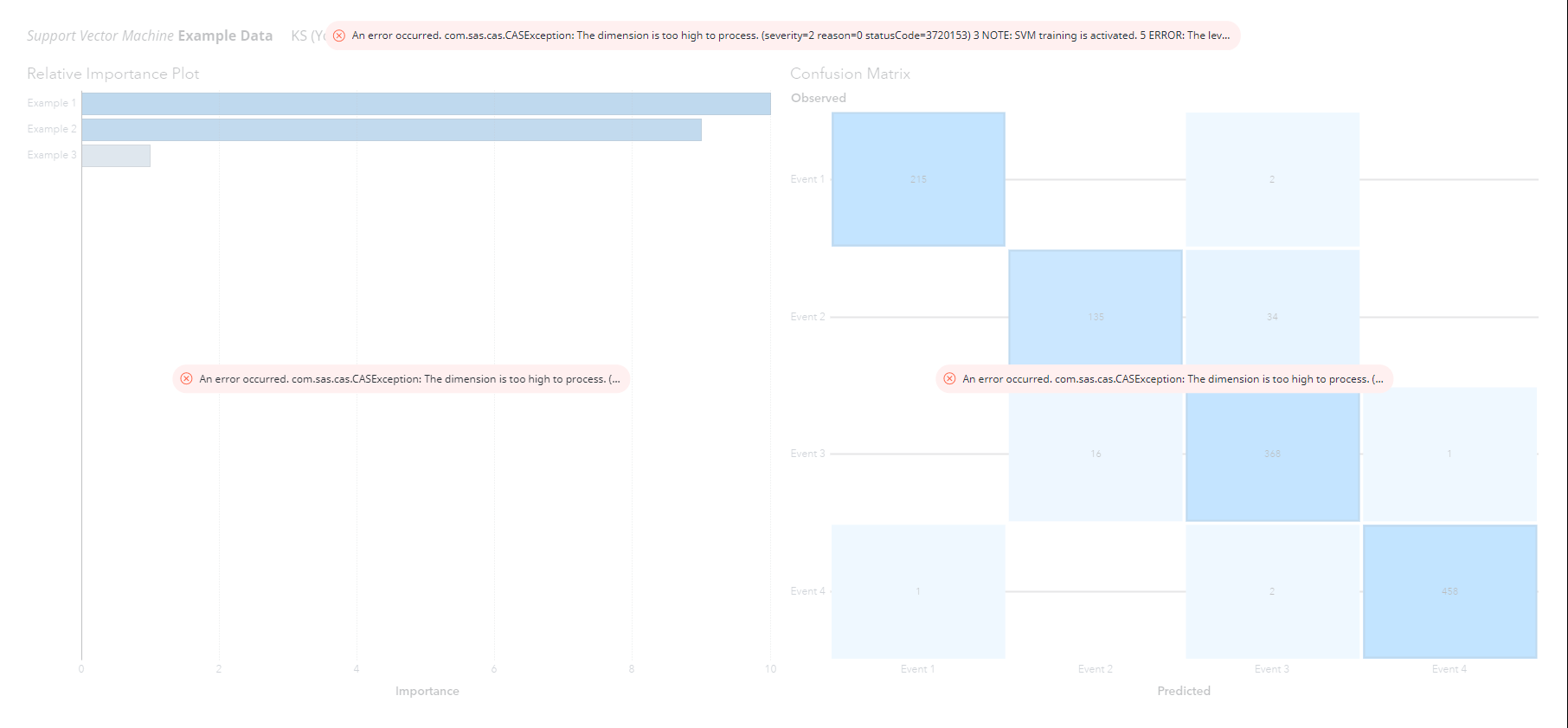
Our dataset was **anonymized** after data partitioning and labelling as it contains **unique identifiers** that can **immediately identify** an individual, such as **ids**, as well as **explicit identifiers** that are **somewhat** able to **identify** an individual, such as **age**. The first **"RowID" node updated existing KNIME-generated row ids**, and the **second "RowID" node created a column based on the updated KNIME-generated row ids.** To ensure that there are **no duplicates** in the dataset, we use the **"Duplicate Row Filter"** before **anonymizing the age column.** Afterwards, we **removed** the **id** column from the dataset and **saved it into a CSV file,** and at the same time, **created a master table** without any columns except **id and rowID** for **future reference** and **linking** of individuals by **authorised personnel**.

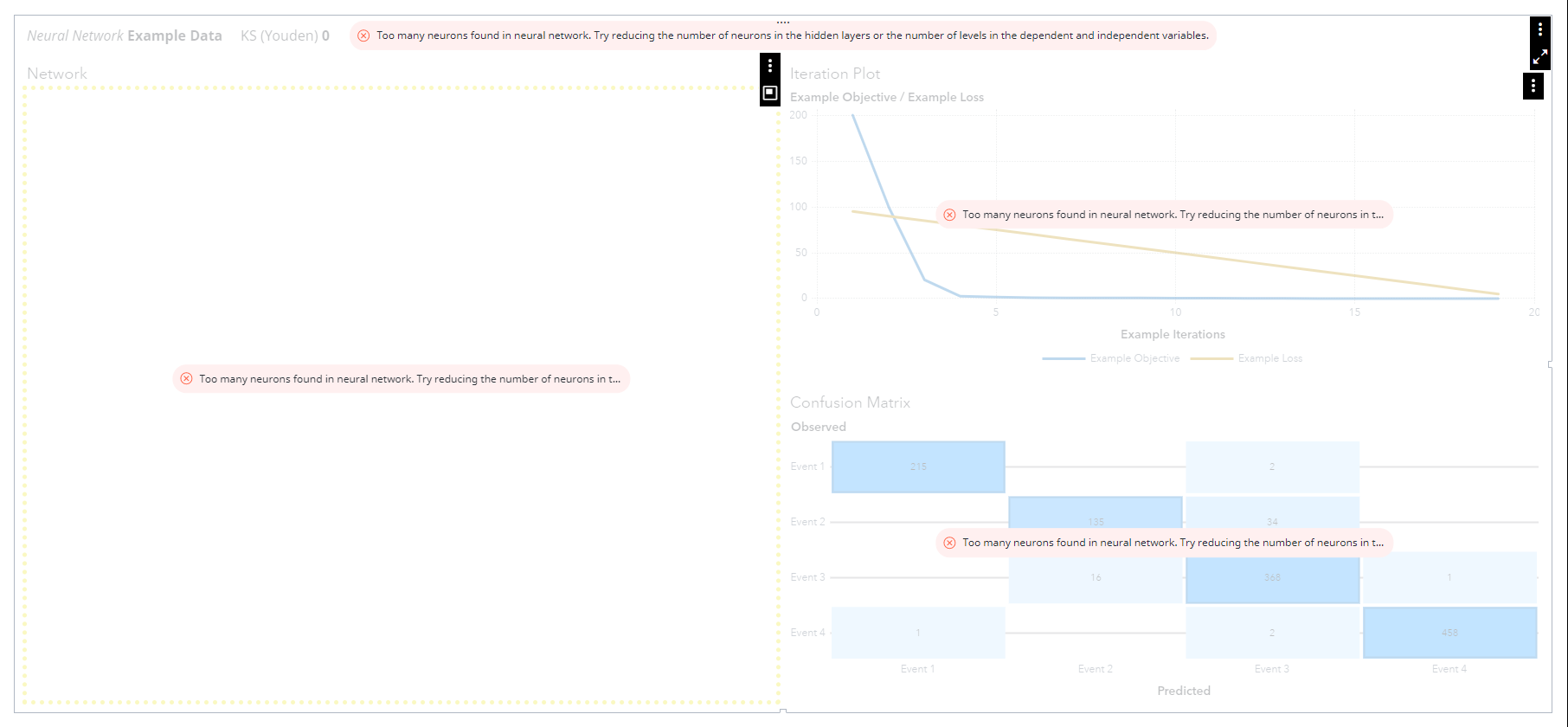
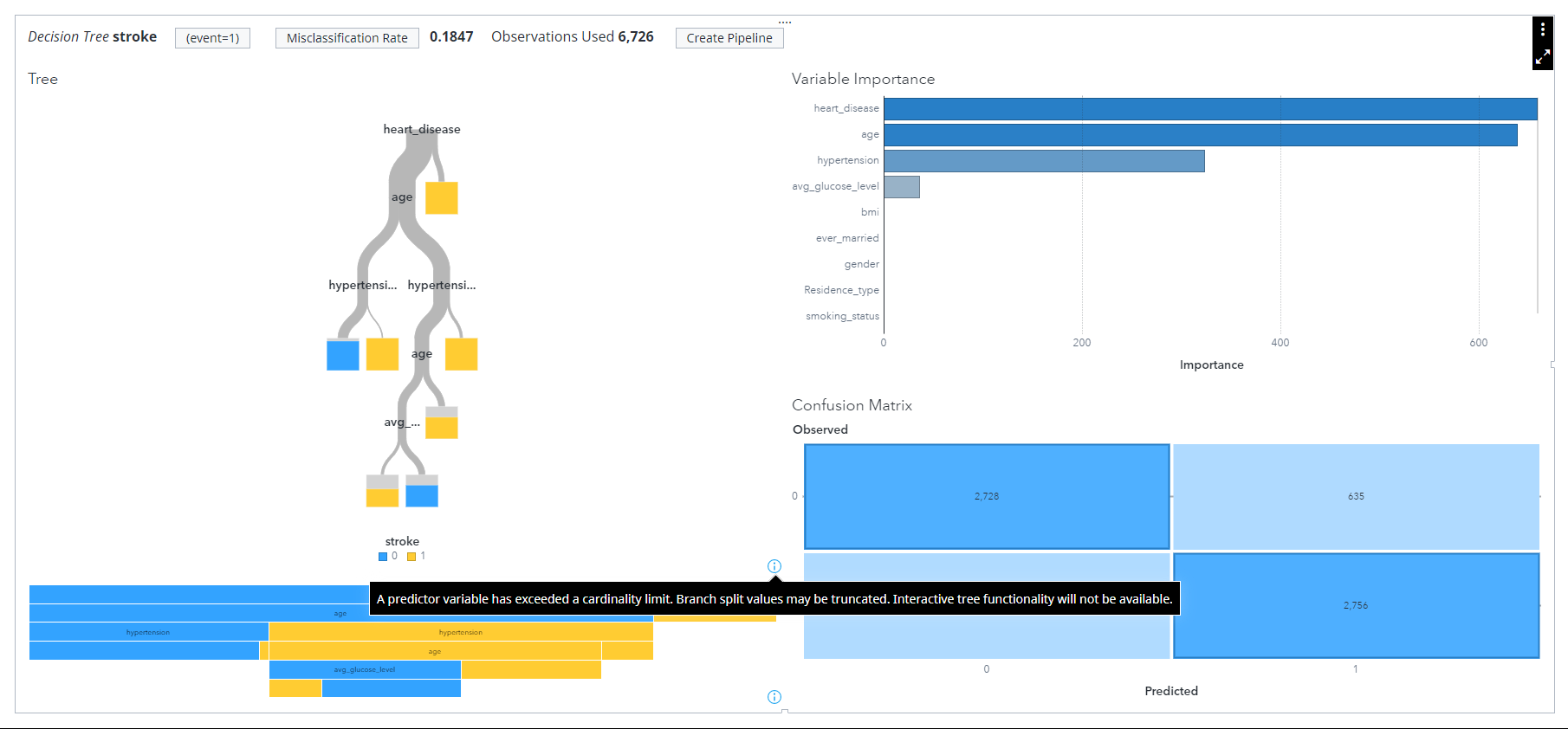
***Reason for Age not being anonymized***

After partitioning and selecting a dataset without one-hot encoding, we wanted to ensure the dataset could be used in different categorical algorithms such as Logistic Regression, Support Vector Machine, Decision Trees and Neural Networks. Hence, we performed a comparison of age binned in 10s and unbinned age using different models

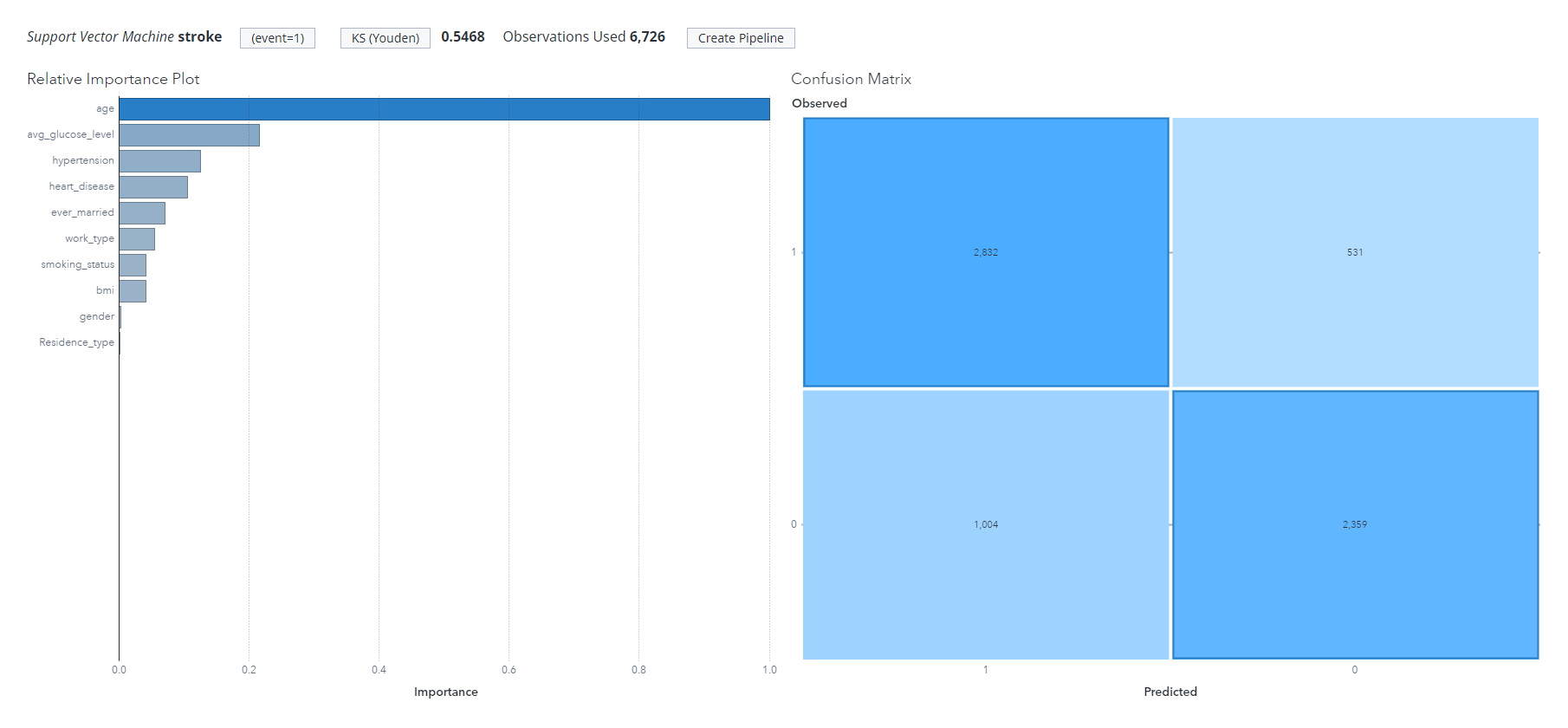
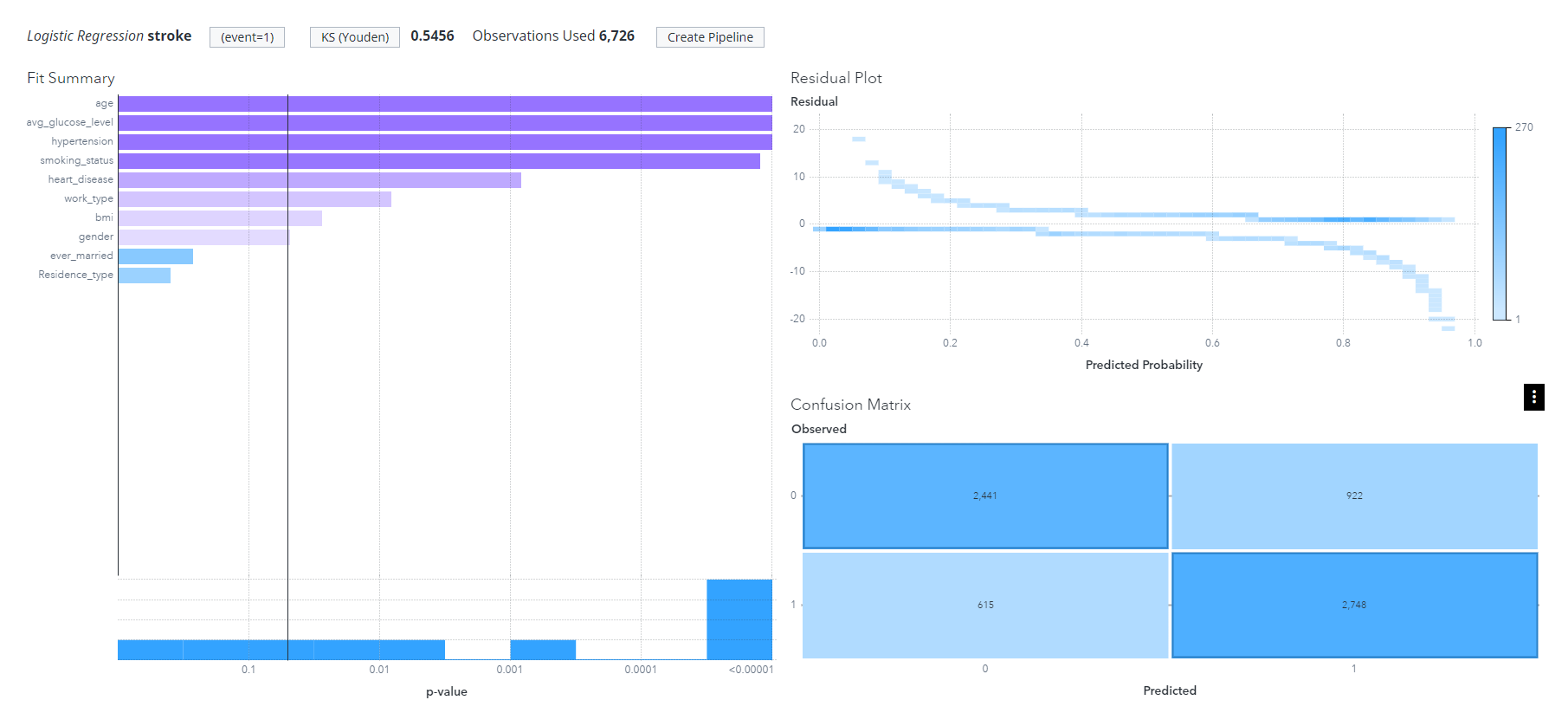
For **binned** age, we found that **Logistic Regression, Support Vector Machine and Neural Network** have returned us with **errors** such as **“Modelling aborted because the number of parameters exceeds the specified limit..”**, **“...The dimension is too high to process..”** and **“Too many neurons found in neural network…”** as shown in the below 4 Figures.

With **Decision Tree,** it has produced the following output with an error stating that **"A predictor variable has exceeded a cardinality limit. The branch split value may be truncated. Interactive tree functionality will not be available.”**

On the other hand, on the **unbinned** age, we have found that **all 4 category algorithms, Logistic Regression, Support Vector Machine, Decision Tree and Neural Network has no issues at all** and display visualisations for us to **better understand** how the model works and which features are the important one in predicting stroke as shown from below.





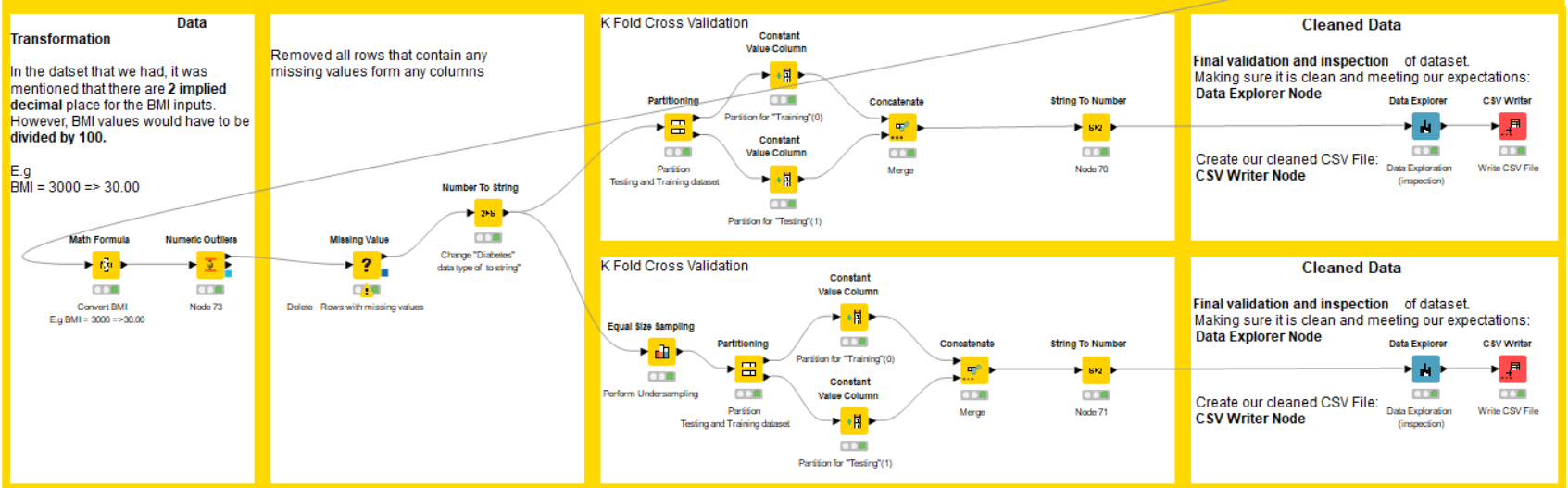
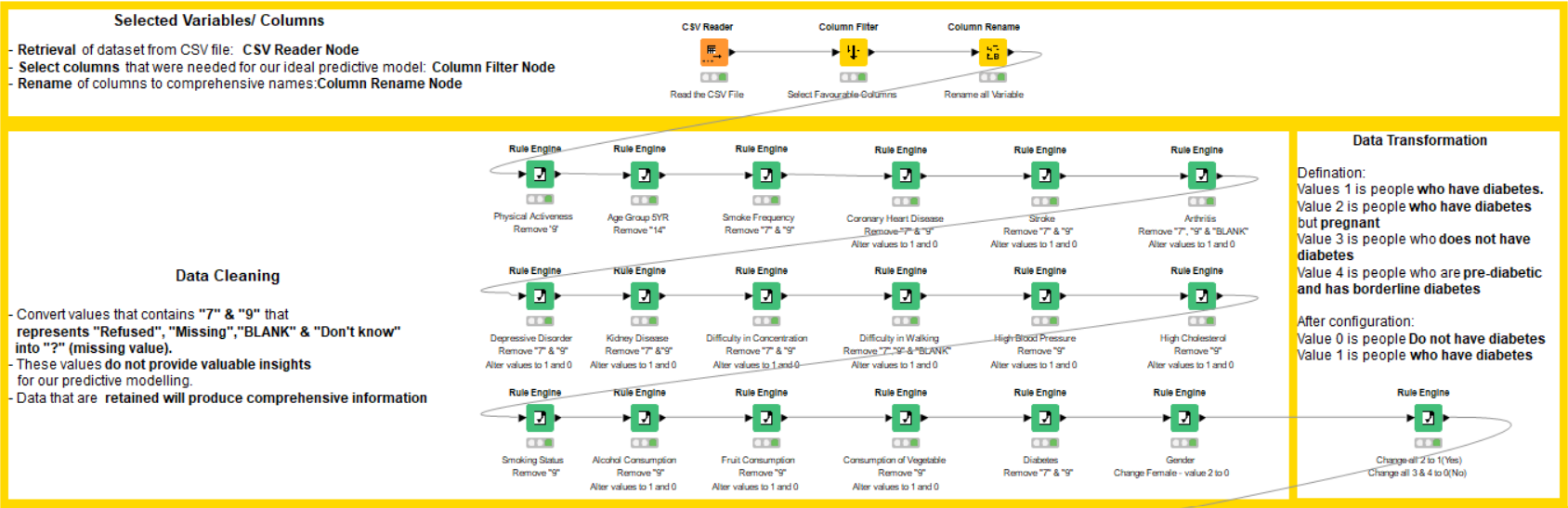
Based on the output given, we can **conclude** that as much as we want to **ensure data privacy** when doing **modelling**, we are unable to do it as it seems like **SAS Viya** platform is **limiting** us to what we can do. Hence, we decided to use the **unbinned age dataset** for better accuracy and modelling.

## Diabetes

### **Data Cleaning Overview**

Under the section on data cleaning, KNIME was used to clean our dataset which contains **441,455 responses** and has **330 features (columns).** After cleaning, only **130,260 responses and 20 features** remain including an additional column “Partition” added to partition our data into “Testing” and “Training” Set.

Below is an overview of the process of data cleaning:



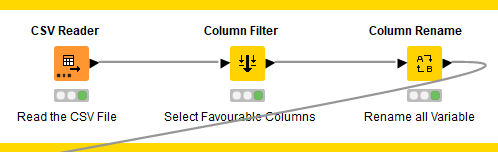
#### **Steps taken for cleaning**

1. Variable Selection
2. Data Transformation
3. Generate an Unbalanced dataset with Partitioning.
4. Generate a Balance dataset with Partitioning
5. Generate Clean Dataset

During data understanding, relevant variables are selected and noticing that there are responses in the dataset that will affect our modelling.

E.g. Responses such as Null/“Refused”/”Don’t Know” and Missing Values. These values will be taken into account in the cleaning process.

**Step 1: Variable Selection**



Feature Selection Consideration:

* By Logic
* By Theory (Professional Input & Research on Journals and documentary)

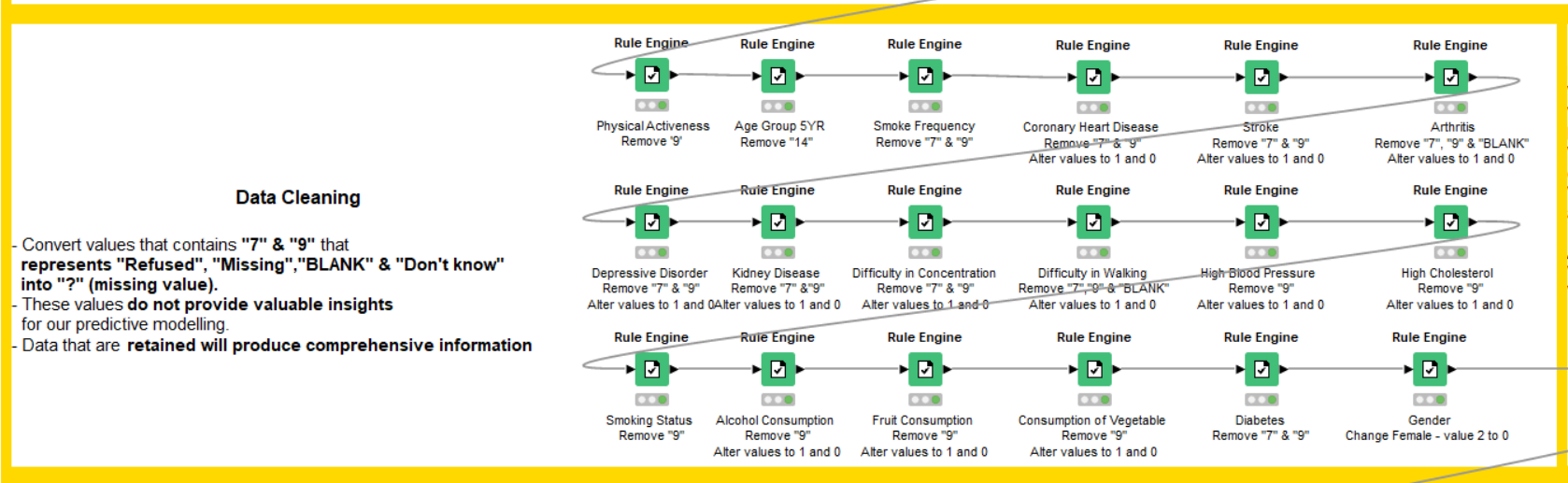
With modelling, we would also perform feature selection not just by **Logic** and **Theory**, we will look into methods like **Wald Chi-Square,** variables with P-value larger than 0.05 we might not have considered during the modelling process. Since our dataset is mostly **categorical/binary**, we would not consider methods like **correlation analysis** for feature selection.

First, the uncleaned dataset is imported into KNIME using the **"CSV Reader"** node and then we use the **"Column Filter"** node to **select 19 variables** that we chose during the data understanding process with the help of professional input. According to the codebook report, the columns in the dataset that we obtained are named using a "code," such as \_BMI5 => Body Mass Index. As a result, before delving deeper into cleaning, we use the **"Column Rename"** node to rename all the columns into comprehensive column names.

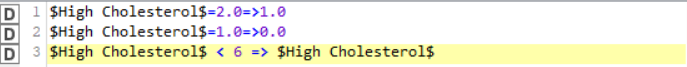
**Rename of Variables Selected** (Column Rename Node)

| **Categories** | **Original Name of Variable/Column** | **New Name of Variable/Column** |
| --- | --- | --- |
| Diabetes | DIABETE3 | Diabetes |
| High Blood Pressure | \_RFHYPE5 | Blood Pressure |
| High Cholesterol | \_CHOLCHK | High Cholesterol |
| BMI | \_BMI5 | BMI |
| Smoking | SMOKDAY2  SMOKER3 | Smoke Frequency  Smoking Status |
| Other Chronic Health Conditions | CVDSTRK3  CVDCRHD4  HAVARTH3  CHCKIDNY | Stroke  Coronary Heart disease  Arthritis  Kidney Disease |
| Diet | FRTLT1  \_VEGLT1 | Consumption of Fruits  Consumption of Vegetables |
| Alcohol Consumption | \_RFDRHV5 | Alcohol Consumption |
| Physical/Mental Health | DIFFWALK  \_PACAT1  DECIDE | Difficulty in Walking  Physical Activiteness  Difficulty in concentration/remembering |
| Demographic | SEX  \_AGEG5YR | Gender  Age group 5YR |

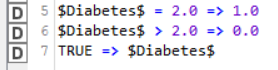
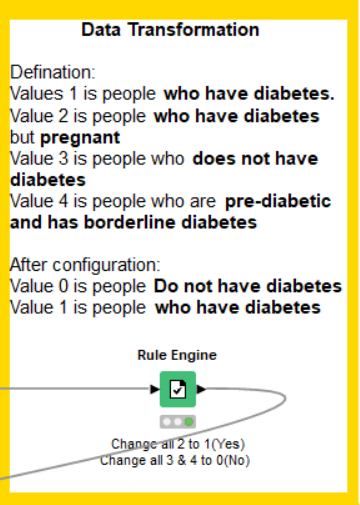
**Step 2: Data Transformation**



We used the "Rule Engine" node to convert values such as "7", "9", and "14" which represent **Null/"Refused"/"Don't Know"/Missing into "?"** which means **"Missing Value"**. According to the dataset's resources, a codebook report is included. **It specifies that some variables with values 1 represent "Yes." However, some variables with 1 representing "No."** Binary numbers, as we all know, are widely used in digital and computer circuits and are represented by a logic **"0" => No** and a logic **"1" => Yes**. As a result, we **set the input variables to 1 and 0**. The rule engine's formula for removing "7," "9,", and "14", also manipulation of values “1” and “2” are shown below.

**Rule Engine**

The formula above will only **return values less than 6**, otherwise, it will **convert it to a "?"** **which is a missing value in KNIME**. Additionally, **standardise every binary column so that "1" represents "Yes" and "0" represents "No”.**



Following that, we took care of the “Response” variable “Diabetes” by transforming values “1”, “2”, “3” and “4”(Codebook definition) to “1” and “2”(Own definition) with the use of “Rule Engine” nodes. Since “3” and “4” are respondents without diabetes both values are changed to “0”. Respondents with diabetes are “1” and “2” and it has been changed to “1”.

Definition based on BRFSS codebook:

“1”: Respondents who have diabetes.

“2”: Respondents who have diabetes but are pregnant.

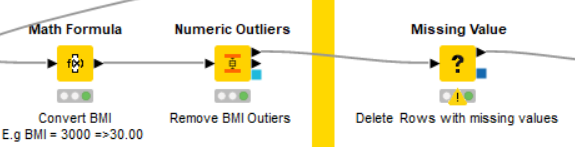
“3”: Respondents who do not have diabetes.

“4”: Respondents who are pre-diabetic and have borderline diabetes.

Own Definition

“0”: Respondents does not have diabetes.

“1”: Respondents who have diabetes.



On the left side, the BMI in the dataset mentioned that it is multiplied by 100 to remove decimal places, thus we had to divide the input by 100 according to the dataset’s source information with the “Math Formula” node. 

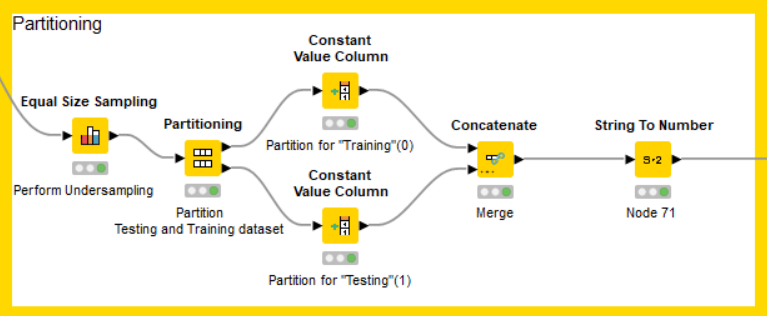




During the data analysis process, it was discovered that **BMI has outlier values**. A **"Numeric Outliers"** node is then used to **remove rows with BMI outliers**. The outlier is calculated using a **standard Interquartile range multiplier of 1.5**, which sets the **lower bound BMI to 12.975 and the upper bound BMI to 41.655**; any BMI value that falls outside of this range is considered an outlier and is removed.

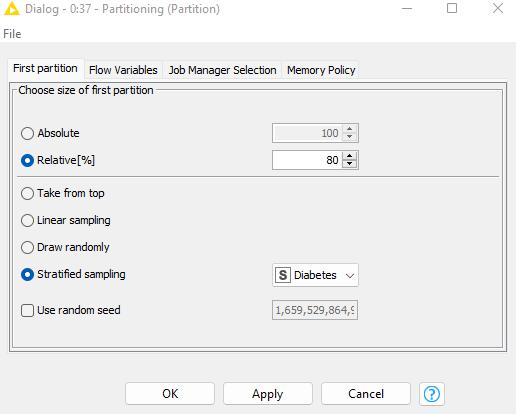
We previously stated that we will convert values such as "7," "9," and "14," which represent Null/"Refused," "Don't Know," and "Missing," into "?" A "Missing Value" node on the right was assigned to **remove all rows with missing values ("?")** in any of the columns.  Missing Value Node

**Step 3: Generate balanced dataset with partitioning**

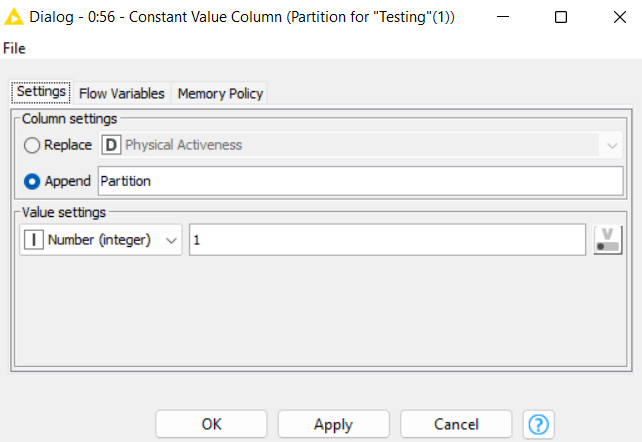
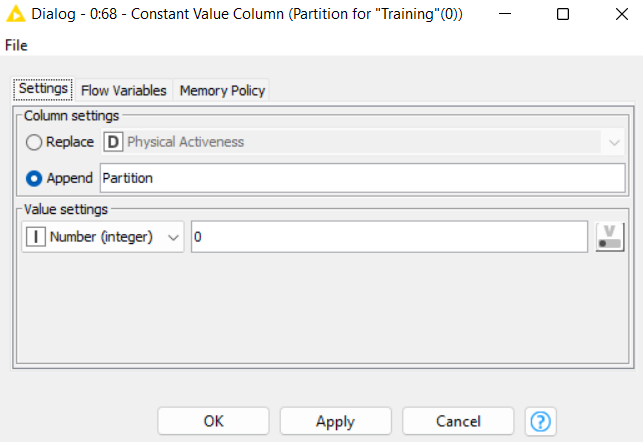


From left to right:

1. To ensure proper data balancing and partitioning, we had to **convert our target variable to strings**.
2. Because the dataset was large enough, we decided to use an **"Equal Size Sampling"** node to perform **undersampling** in order to **generate a balanced dataset**. By reducing the record with those without "Diabetes," this node **equalises** the sample ofRespondents with "Diabetes" and those without "Diabetes."
3. With the "Partitioning" node, we partitioned our dataset based on our target variable "Diabetes," with **80% for "Training"** and **20% for "Testing."**

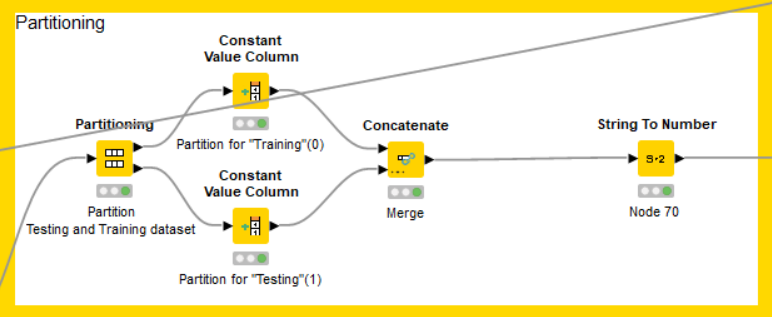


1. The **"Training" set** was separated from the **"Testing" set** as we partitioned our dataset with the **"Partitioning"** node. To determine which row in SAS Viya is used for training and testing, two nodes called **"Constant Value Column"** nodes are used to **create a new column called "Partition" with values of "0" and "1",** which represent **training and testing, respectively.**



1. Lastly, merge the partitioned data back as a whole with a “Concatenate” node.

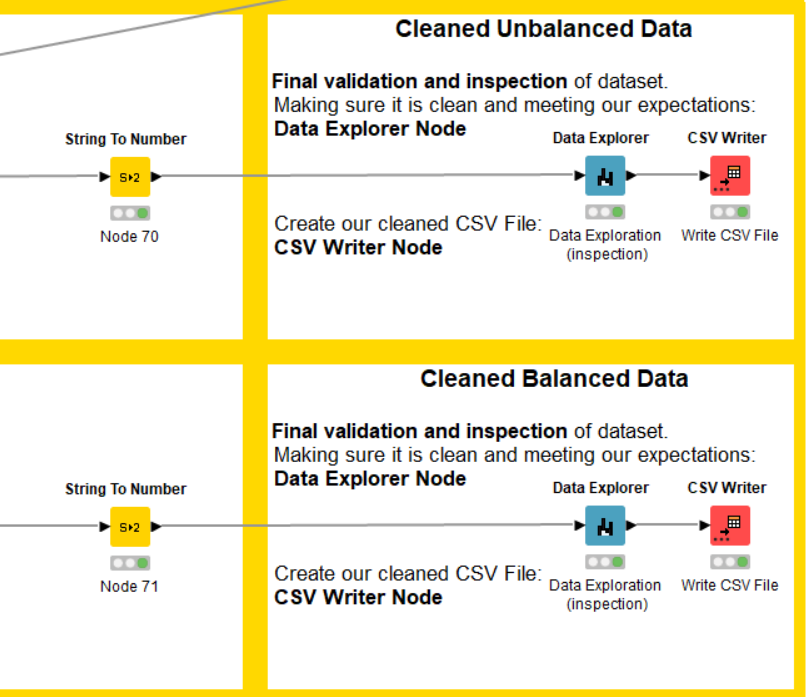
**Step 4: Generate unbalanced dataset with partitioning**



From left to right:

1. To ensure proper data balancing and partitioning, we had to **convert our target variable to strings**.
2. With the "Partitioning" node, we partitioned our dataset based on our target variable "Diabetes," with **80% for "Training"** and **20% for "Testing".**
3. The **"Training" set** was separated from the **"Testing" set** as we partitioned our dataset with the **"Partitioning"** node. To determine which row in SAS Viya is used for training and testing, two nodes called **"Constant Value Column"** nodes are used to **create a new column called "Partition" with values of "0" and "1",** which **represent training and testing, respectively**.

**Step 5: Generate Clean Dataset**

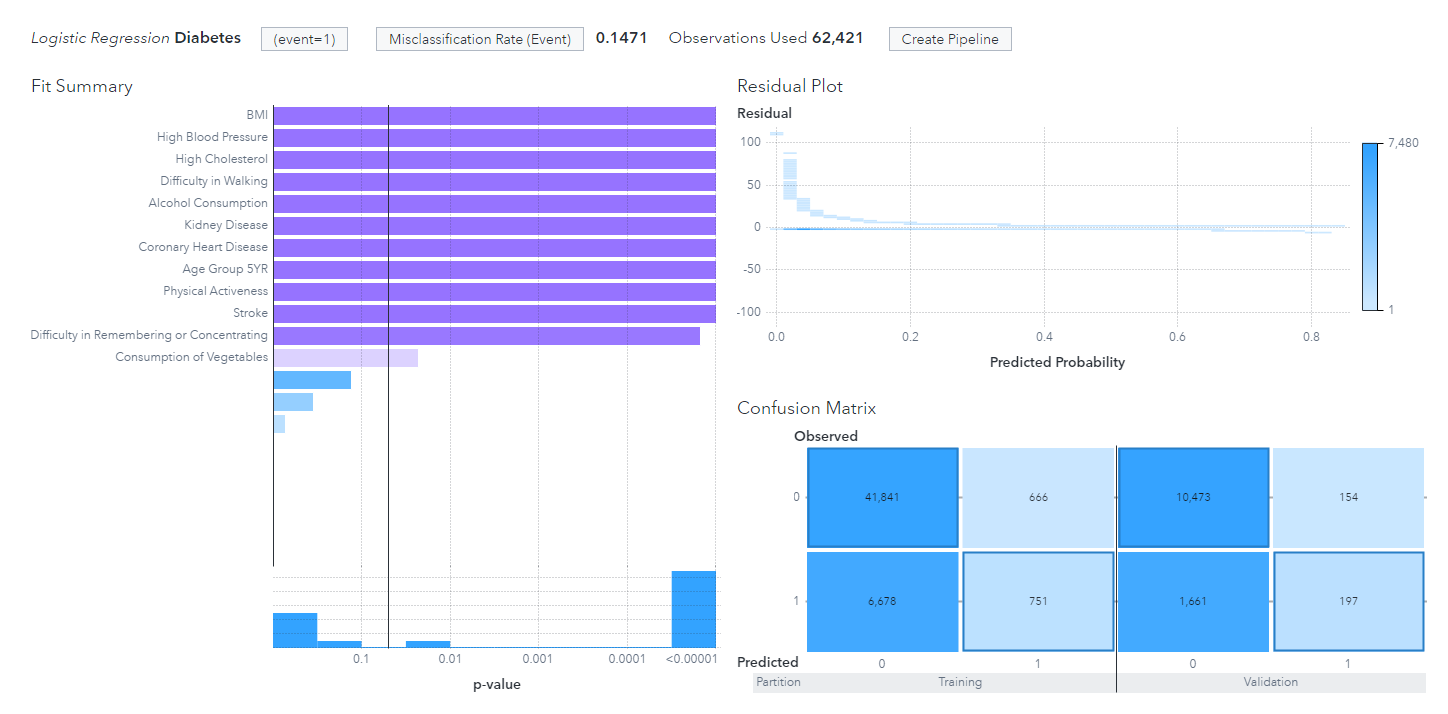


In the final stage, we had to convert strings back to numbers using the **“String to Number”** node in the final stage, and we used the **"Data Explorer"** node to see if our cleaning phase had met our needs and goals. We then used the **"CSV writer"** node to **generate our cleaned balance dataset with undersampling and cleaned unbalanced dataset**.

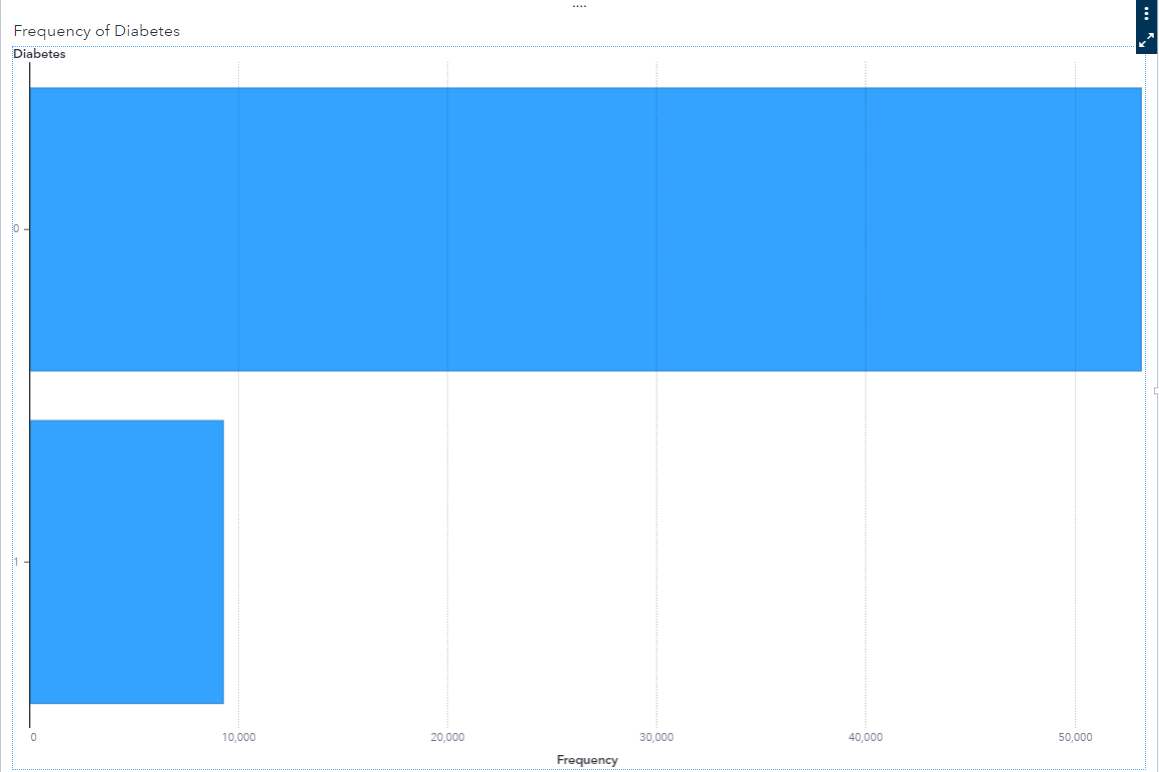
### 

### **Selecting Dataset [Balanced Vs Unbalanced]**

Unbalanced Dataset

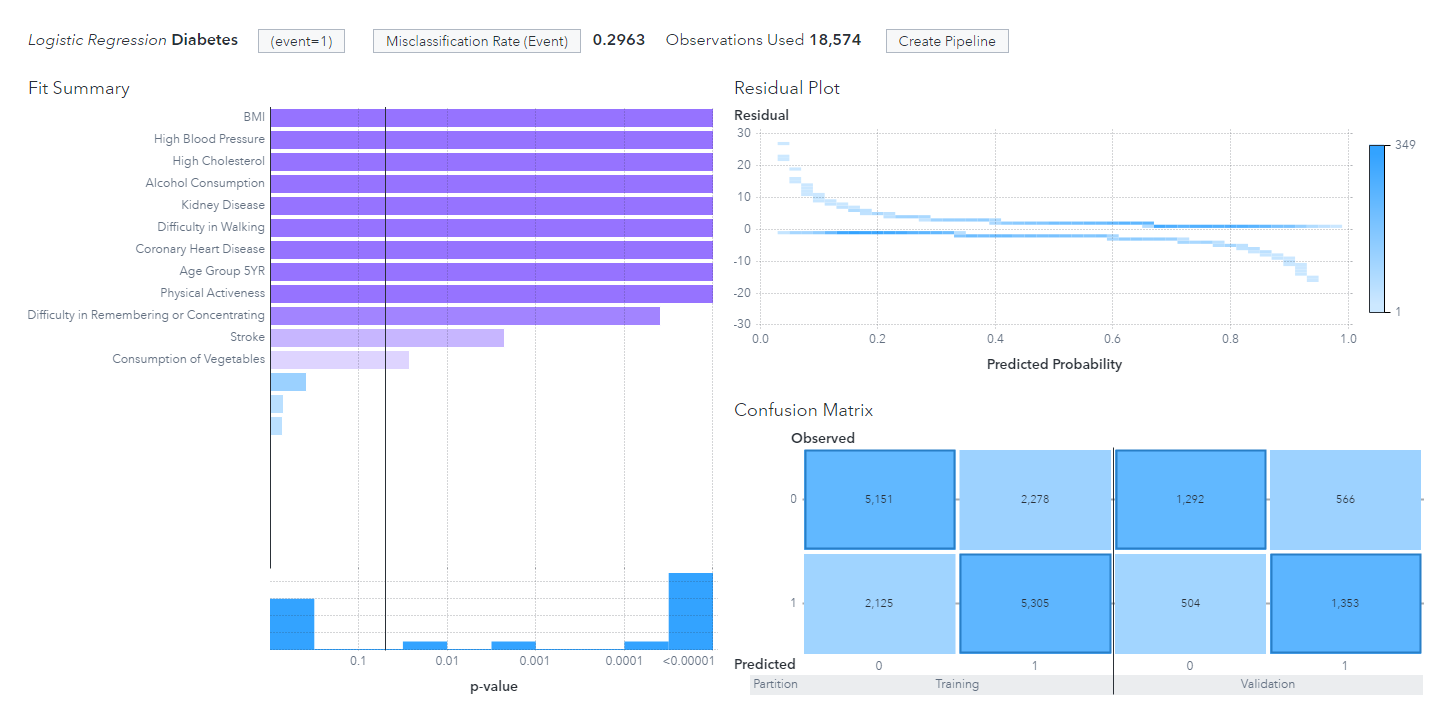


The **misclassification** of the unbalanced data is 0.1471, which means that the chance of classifying incorrectly is 14.71%, which is low.

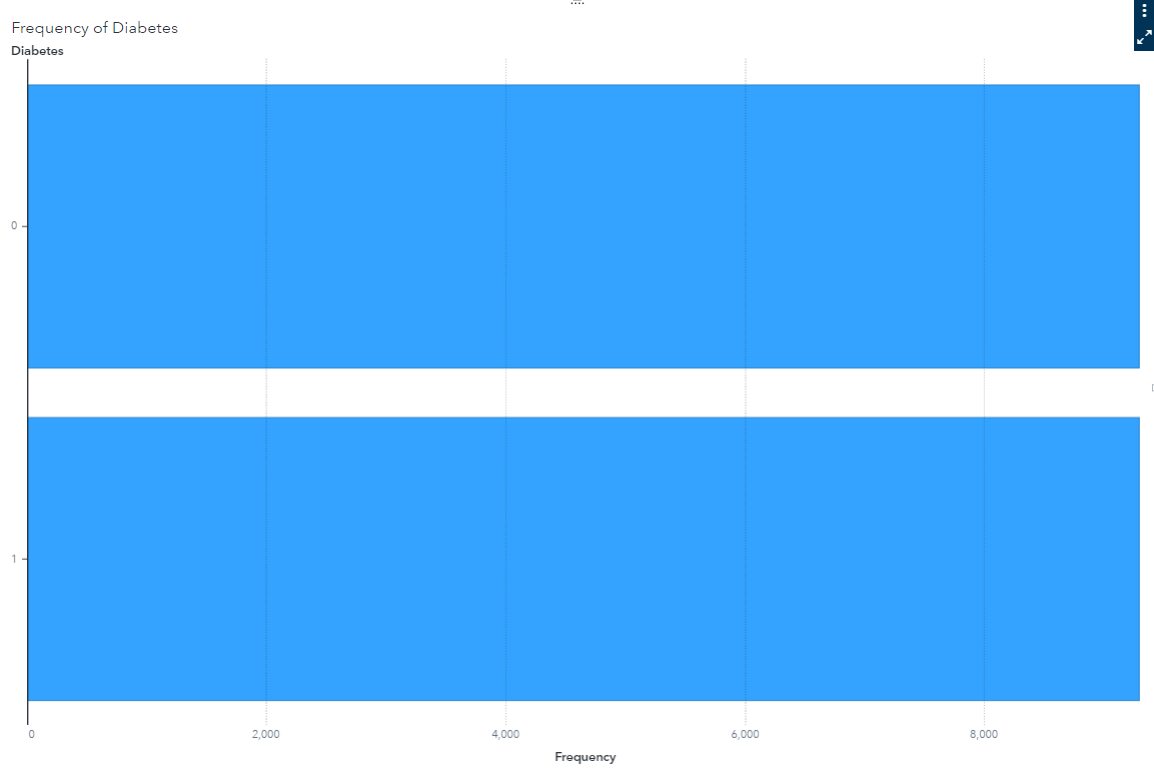


The bar chart above shows that the dataset is highly unbalanced with only 9,287 with data and 53,134 without diabetes (Border-line diabetes/No diabetes). This means that data will be trained mainly on the majority class while the minority class will be slightly ignored.

Balanced Dataset



The **misclassification** of the balanced data is 0.2963, this means that the chance of classifying “Diabetes” wrongly is 29.63%, which is higher than the unbalanced dataset.



From the bar chart above, it shows that the dataset is relatively balanced with 9,287 records for both diabetes and non-diabetes. This means that data will be trained fair and square. Hence, causing the model training to be more accurate.

***Conclusion***

We see that performance for the unbalanced dataset is performing better through misclassification rate. However looking at the confusion matrix of the unbalanced dataset, it was found that the unbalanced dataset is only good at identifying those without “Diabetes” as the number of records of “no diabetes” in the unbalanced dataset is about **5 times larger** than the record of those with “diabetes”. As previously stated, the unbalanced dataset will be trained primarily on the majority class (those without diabetes), while the minority class (those with diabetes) will be marginally ignored in identifying "Diabetes” variable records.

Therefore, after modelling to test both balanced and unbalanced datasets, we decided to use the **Balance Dataset** for our modelling.

# 

# **Timeline**

## **Updated Gantt Chart**



As was previously shown in our proposal report, our team has allocated the tasks accordingly. Generally, the majority of data understanding and cleaning was done in pairs, without much individual work, and only separated by stroke (Shermaine & Shi Min) and diabetes (Zhang Xiang & Rawtbhik), with some focus on each task breakdown.

Nevertheless, some timelines have been adjusted due to the fact that cleaning may have taken longer than anticipated. Furthermore, we also added additional steps along the way, such as data privacy ethics cleanup, testing different models, balancing, partitioning, making comparisons and annotations, which took up even more time. Hence, this is the updated timeline for our milestone report.

## **Next Phases**

Following the completion of Data Collection and Preparation, our next phase, Modelling, involves feature selection, model creation and explanation, comparing them, and tuning and testing the models to build the model of the highest accuracy. Finally, we will also come together to write our report, determine the best model for each topic and prepare for the presentation.

# **Appendix**

Diabetes Data Dictionary

SEX

## 

\_PACAT1

## 

\_BMI5

## 

\_AGEG5YR

## 

DIABETE3(Response Variable)

## 

CVDSTRK3

## 

\_FRTLT1

## 

\_VEGLT1

## 

\_RFDRHV5

## 

DECIDE

## 

DIFFWALK

## 

\_SMOKER3

## 

SMOKDAY2

## 

\_RFHYPE5

## 

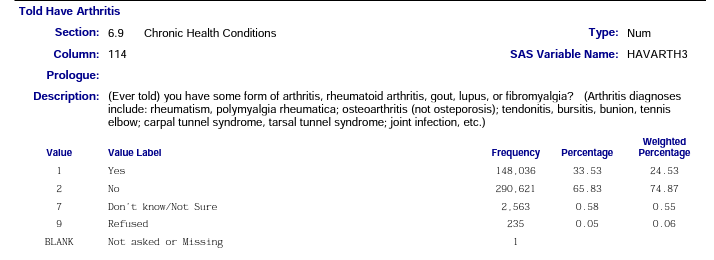
\_RFCHOL

## 

CVDRHD4

## 

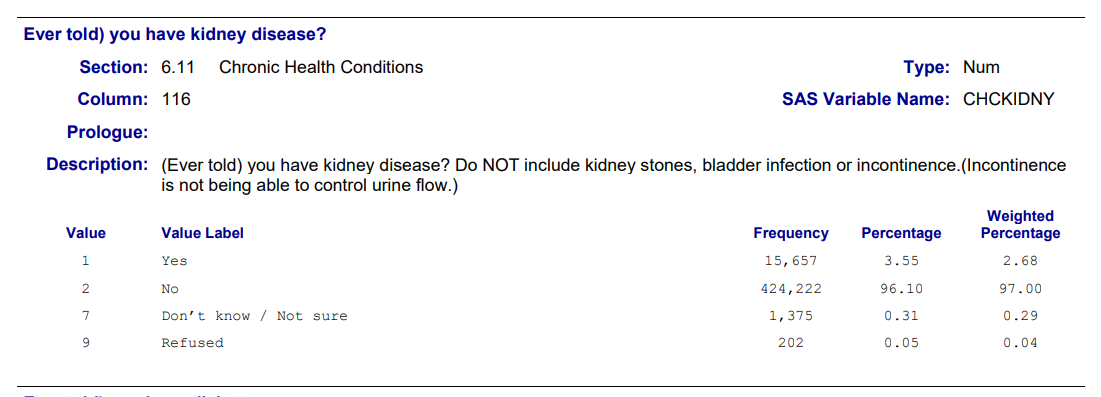
HAVARTH3



ADDEPEV2

## 

CHCKIDNY



# **References**

**Dataset Reference:**

1. About BRFSS Dataset:

<https://www.cdc.gov/brfss/annual_data/annual_2015.html>

1. Definition of BRFSS variables: <https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf>
2. Summary of BRFSS Dataset: [https://www.cdc.gov/brfss/annual\_data/2015/Summary\_Matrix\_15\_version12.](https://www.cdc.gov/brfss/annual_data/2015/Summary_Matrix_15_version12.html)[html](https://www.cdc.gov/brfss/annual_data/annual_2015.html)

**Other Research Reference:**

1. BRFSS Dataset:

<https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset>