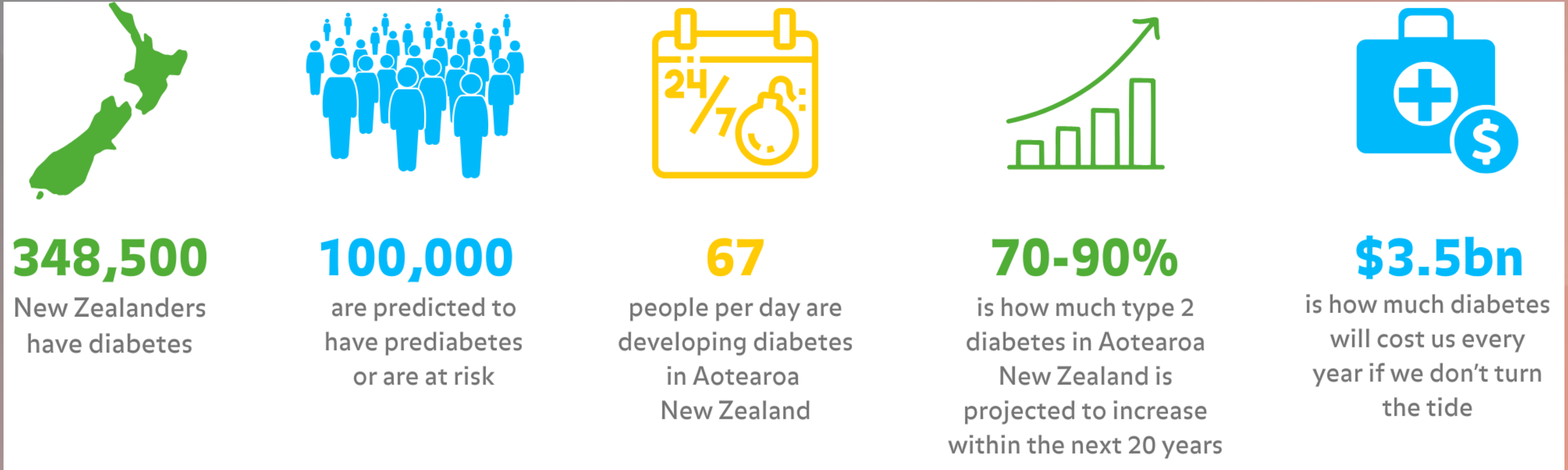
Diabetes Study

Exploratory Analysis & Statistical Inference

What is Diabetes?



348,500
New Zealanders
have diabetes

100,000
are predicted to
have prediabetes
or are at risk

67
people per day are
developing diabetes
in Aotearoa
New Zealand

70-90%
is how much type 2
diabetes in Aotearoa
New Zealand is
projected to increase
within the next 20 years

$3.5bn
is how much diabetes
will cost us every
year if we don’t turn
the tide

Dataset source and description

- ‘diabetes_prediction_dataset.csv’ dataset
- Obtained from Kaggle by author **Mohammed Mustafa**
- Containing 100,000 unique patient records
- Collection of medical and demographic patient data
- Comprising of a mixture of categorical, binary, and continuous variables

Dataset source and description

Features:

- Age
- BMI
- Hypertension
- Heart Disease
- Smoking History
- HbA1c Level
- Blood Glucose Level
- Diabetes (dependant variable)

Dataset source and description

- Tidiness?
- Statistical Measures?
- Authenticity?

Research Intent

Purpose

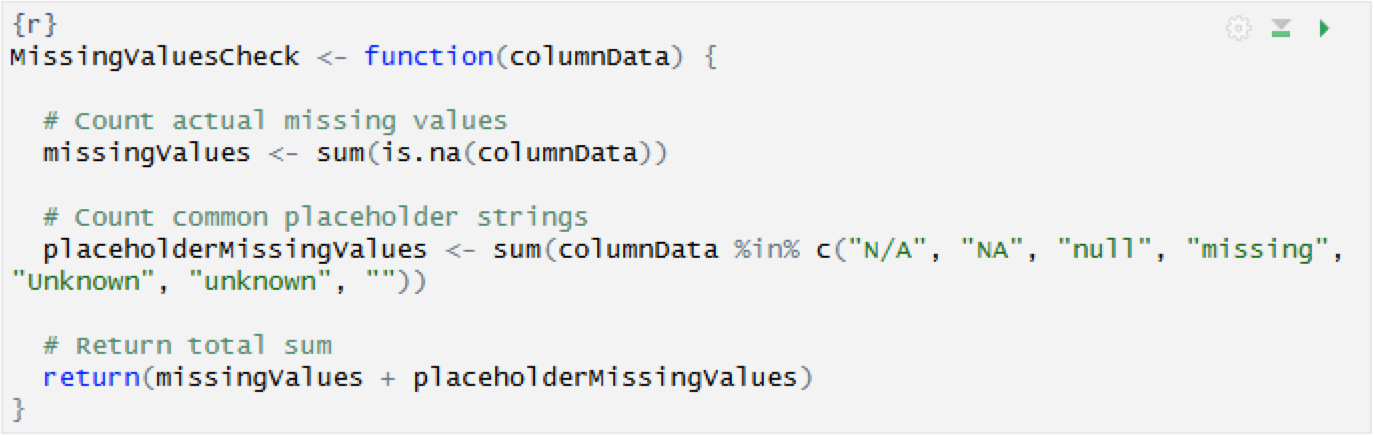
- To explore the relationships between demographic, lifestyle, and health-related factors associated with diabetes, and heart disease
- Identify patterns and correlations between these factors

Data Exploration, Analysis and Testing

- Statistical inference
- Hypothesis testing and p-values
- R Studio
    - Packages:
      “here”
      "ggplot2"
      "dplyr"

Data Cleaning and Preparation

- Missing Values
- Data Validation

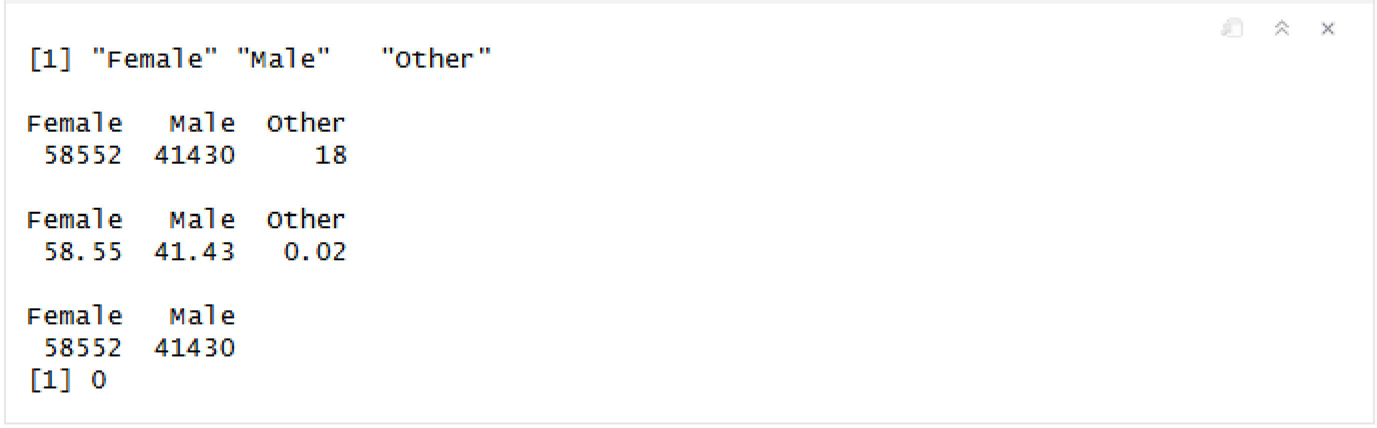


```
{r}
MissingValuesCheck <- function(columnData) {

  # Count actual missing values
  missingValues <- sum(is.na(columnData))

  # Count common placeholder strings
  placeholderMissingValues <- sum(columnData %in% c("N/A", "NA", "null", "missing",
"Unknown", "unknown", ""))

  # Return total sum
  return(missingValues + placeholderMissingValues)
}
```

Data Cleaning and Preparation

Gender

- “Other” values observed
- 0.02% of dataset population

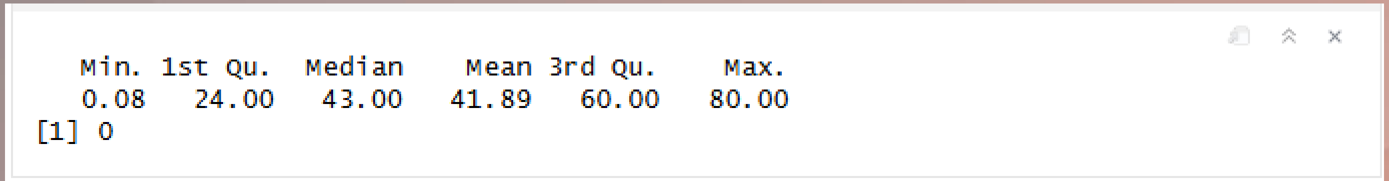


```
[1] "Female" "Male"   "Other"

Female   Male  Other
 58552  41430     18

Female   Male  Other
 58.55  41.43   0.02

Female   Male
 58552  41430
[1] 0
```

Data Cleaning and Preparation

Age

- Plausibility check required
- No values lower than 0
- No values greater than 120

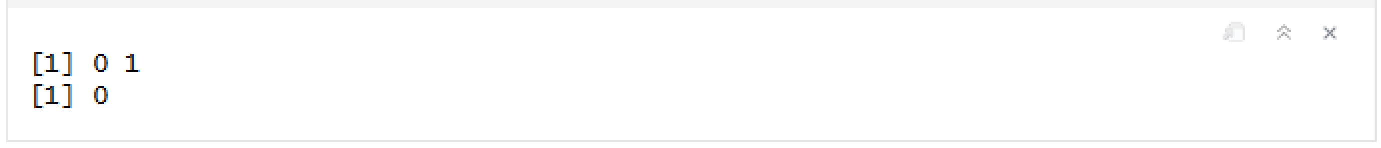


```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.08   24.00   43.00   41.89   60.00   80.00 
[1] 0
```

Data Cleaning and Preparation

Hypertension, Heart Disease and Diabetes.

- Check for non-binary values

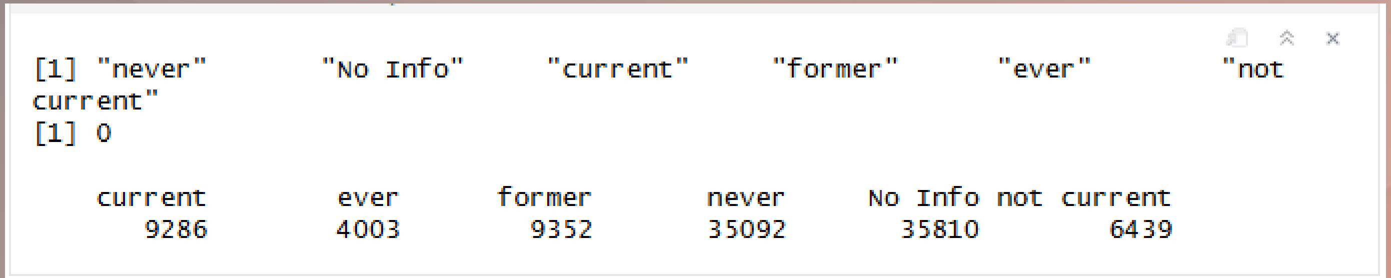


```
[1] 0 1
[1] 0
```

Data Cleaning and Preparation

Smoking History

- “No Info” category identified
- 35% of population
- Potentially overlapping features

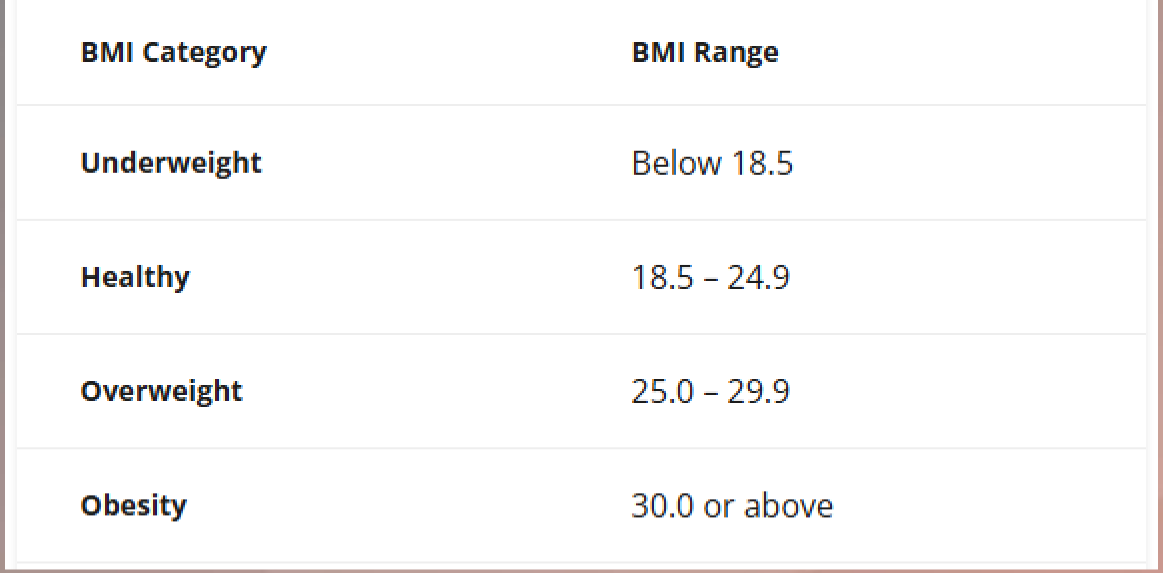


```
[1] "never"        "No Info"      "current"      "former"       "ever"         "not
current"
[1] 0

    current        ever      former       never     No Info not current
       9286        4003        9352       35092       35810        6439
```

Data Cleaning and Preparation

BMI

- Plausibility validation

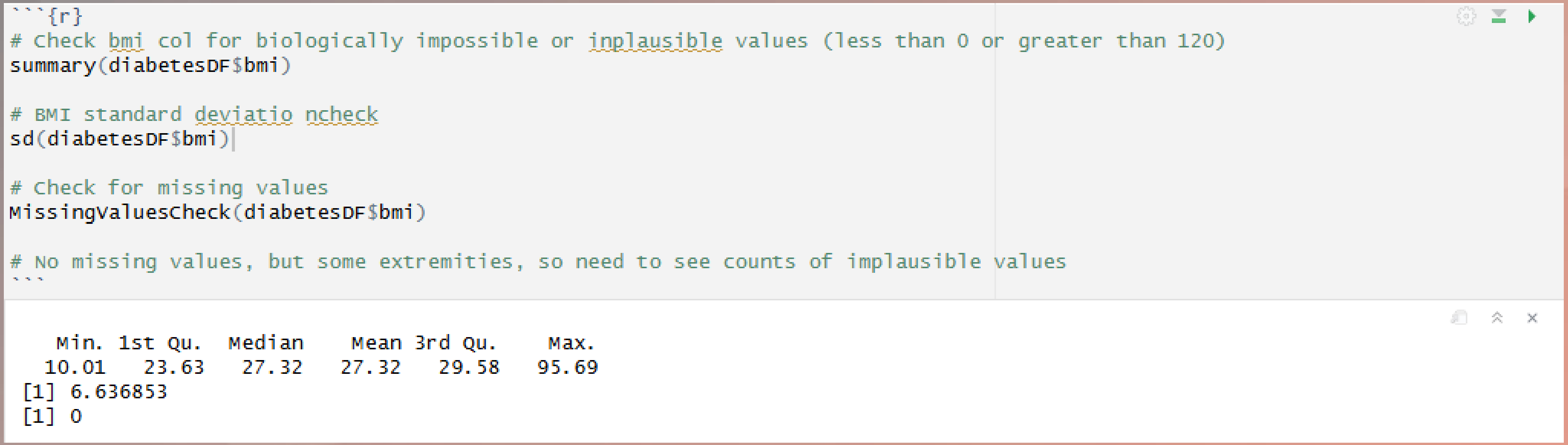


BMI Category	BMI Range
Underweight	Below 18.5
Healthy	18.5 – 24.9
Overweight	25.0 – 29.9
Obesity	30.0 or above

Data Cleaning and Preparation

BMI

- Plausibility validation



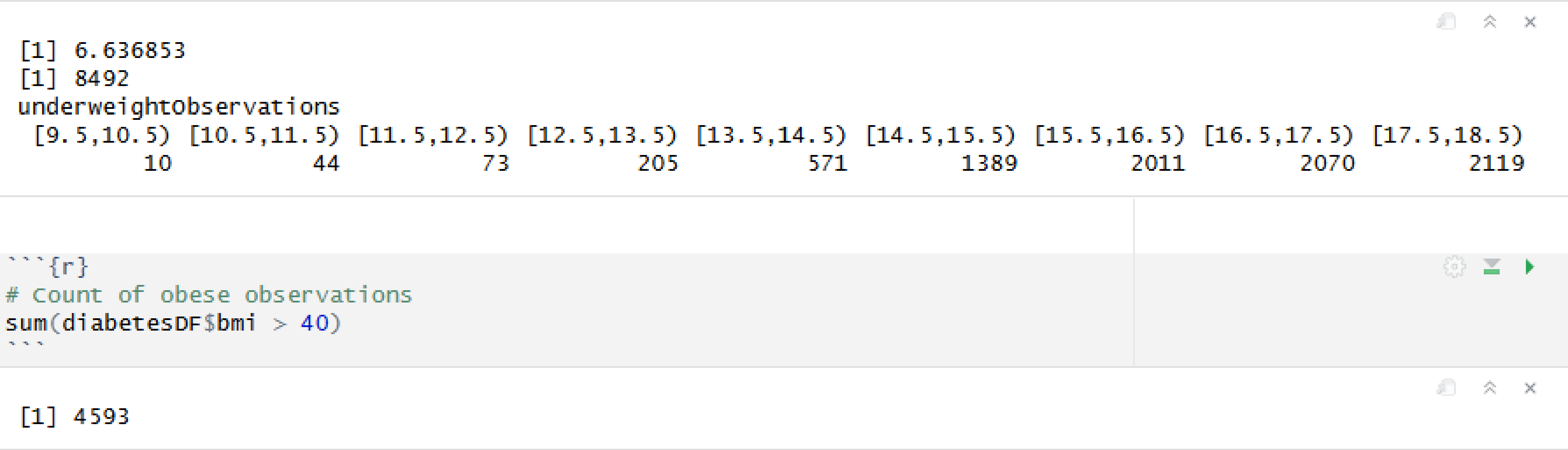
~~~
```{r}
# Check bmi col for biologically impossible or inplausible values (less than 0 or greater than 120)
summary(diabetesDF$bmi)

# BMI standard deviatio ncheck
sd(diabetesDF$bmi)

# Check for missing values
MissingValuesCheck(diabetesDF$bmi)

# No missing values, but some extremities, so need to see counts of implausible values
```
~~~

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  10.01   23.63   27.32   27.32   29.58   95.69 
[1] 6.636853
[1] 0
```

Data Cleaning and Preparation

BMI

- Plausibility validation



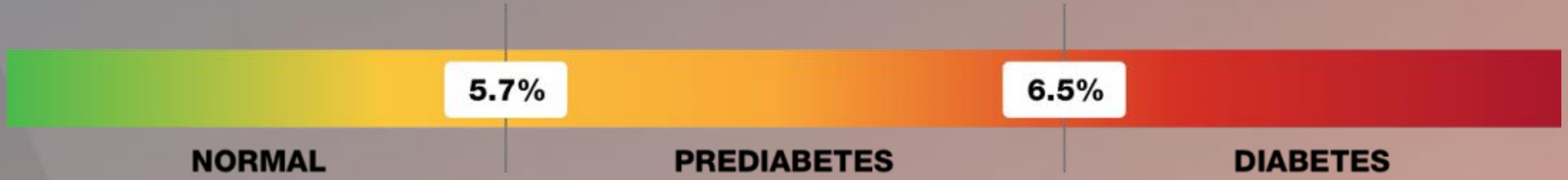
```
[1] 6.636853
[1] 8492
underweightObservations
 [9.5,10.5) [10.5,11.5) [11.5,12.5) [12.5,13.5) [13.5,14.5) [14.5,15.5) [15.5,16.5) [16.5,17.5) [17.5,18.5) 
         10          44          73         205         571        1389        2011        2070        2119 
```

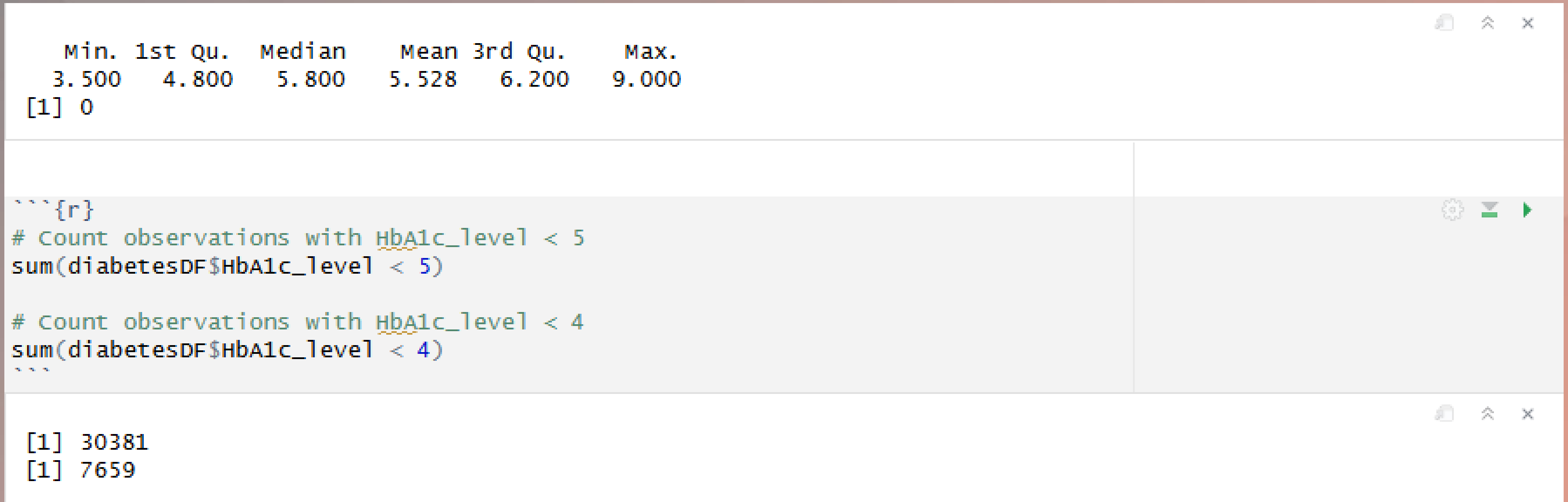
~~~
```{r}
# Count of obese observations
sum(diabetesDF$bmi > 40)
```
~~~

```
[1] 4593
```

Data Cleaning and Preparation

HbA1c Level





```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  3.500   4.800   5.800   5.528   6.200   9.000 
[1] 0
```

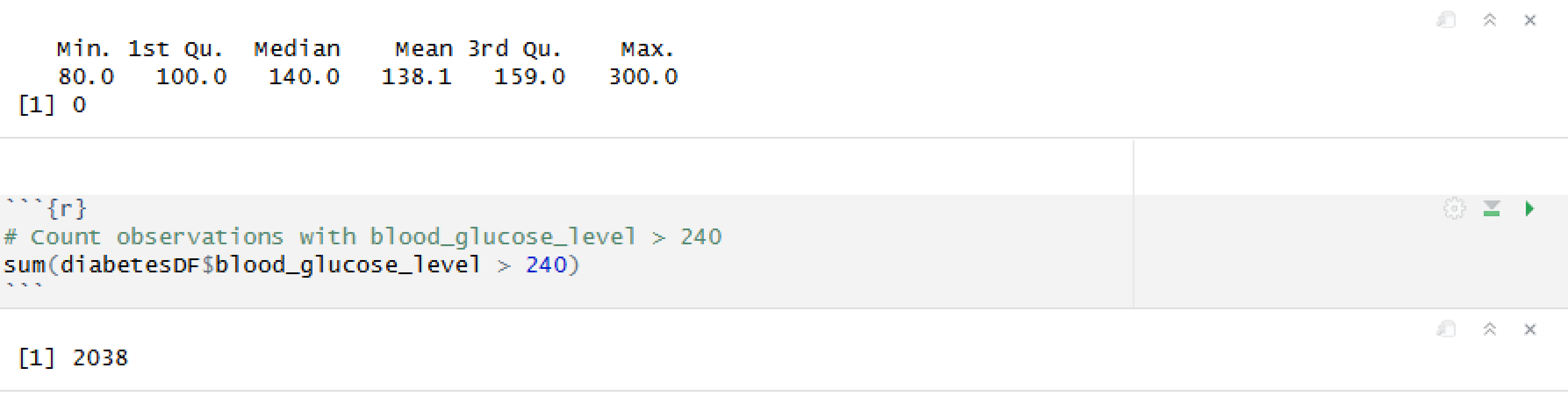
~~~
```{r}
# Count observations with HbA1c_level < 5
sum(diabetesDF$HbA1c_level < 5)

# Count observations with HbA1c_level < 4
sum(diabetesDF$HbA1c_level < 4)
```
~~~

```
[1] 30381
[1] 7659
```

Data Cleaning and Preparation

Blood Glucose Level

- Plausibility check



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   80.0   100.0   140.0   138.1   159.0   300.0 
[1] 0
```

~~~
```{r}
# Count observations with blood_glucose_level > 240
sum(diabetesDF$blood_glucose_level > 240)
```
~~~

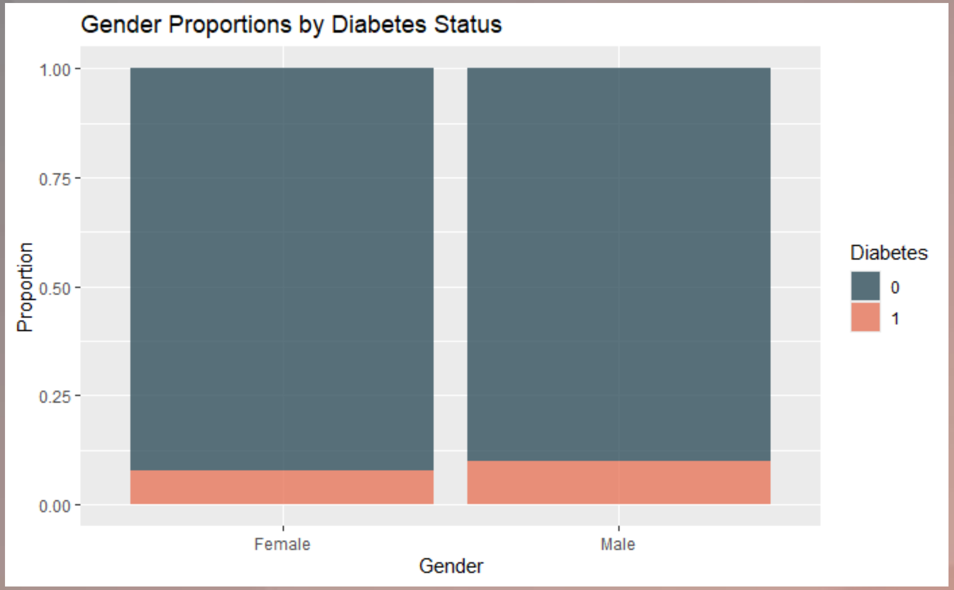
```
[1] 2038
```

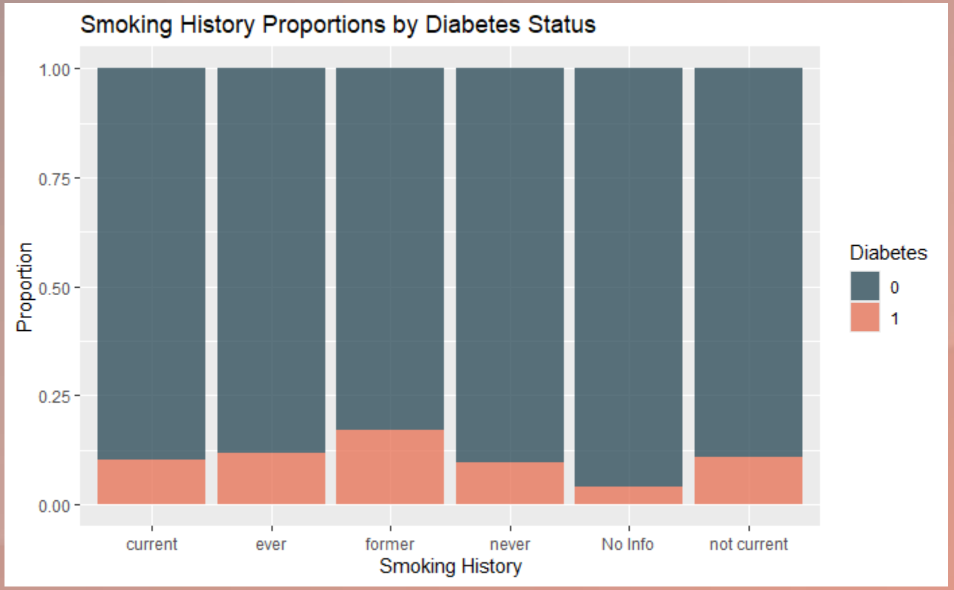
Descriptive Exploration

Categorical & Binary Features

Descriptive Exploration

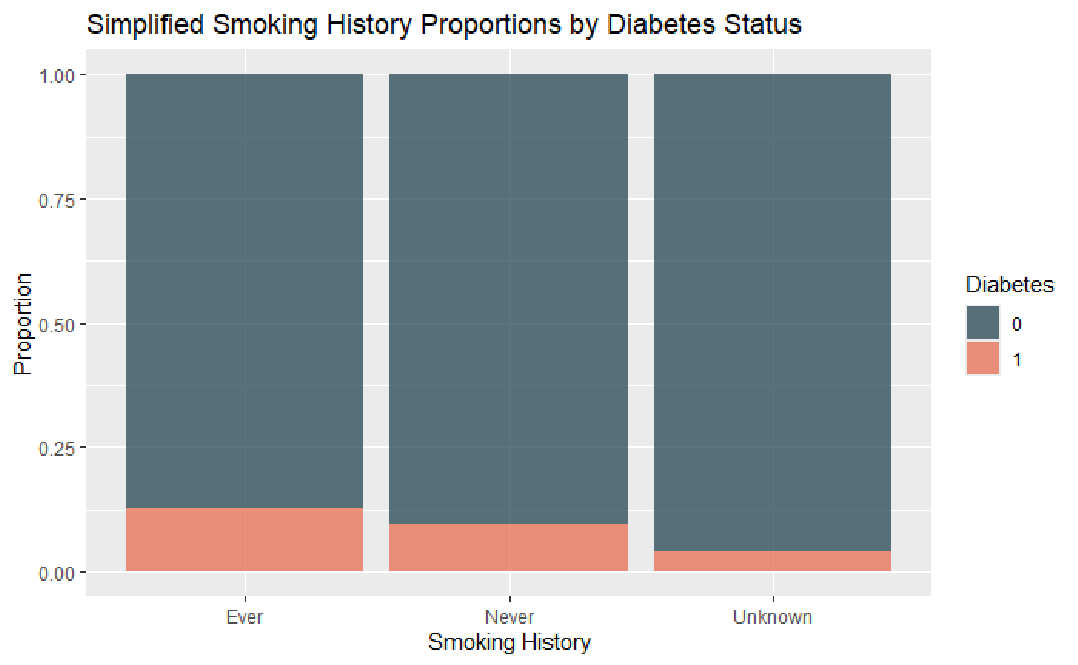
Categorical & Binary Features





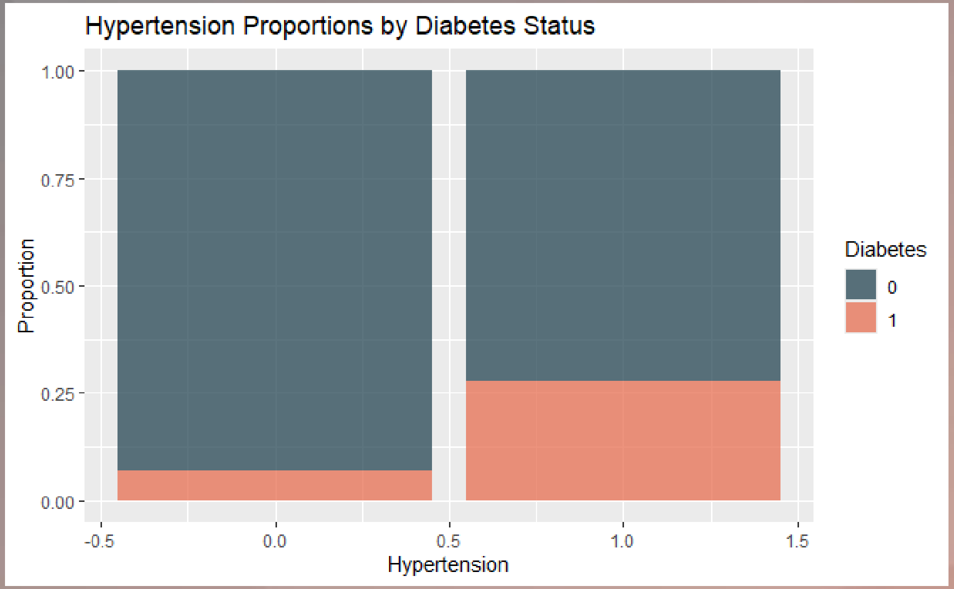
Descriptive Exploration

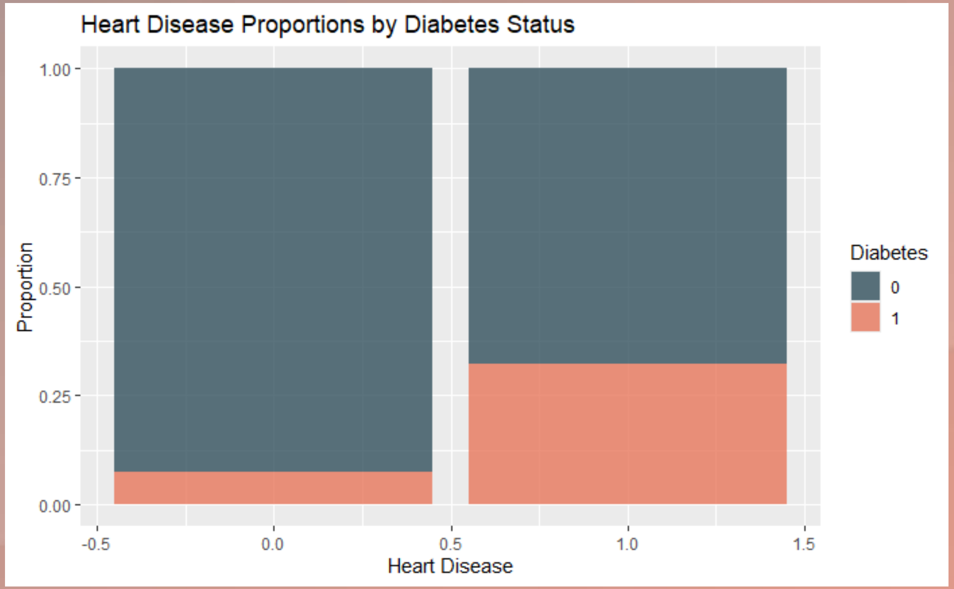
Categorical & Binary Features



Descriptive Exploration

Categorical & Binary Features



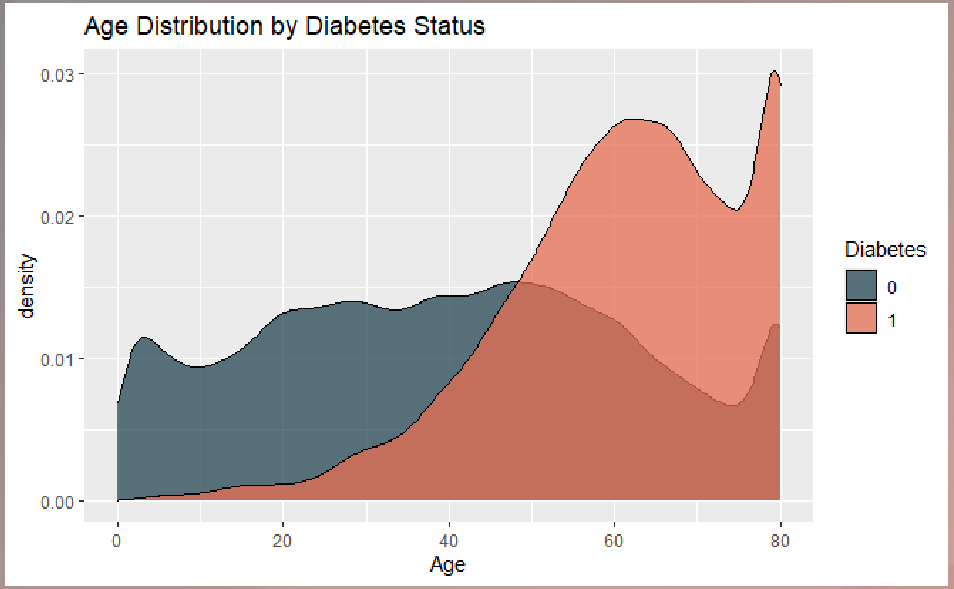


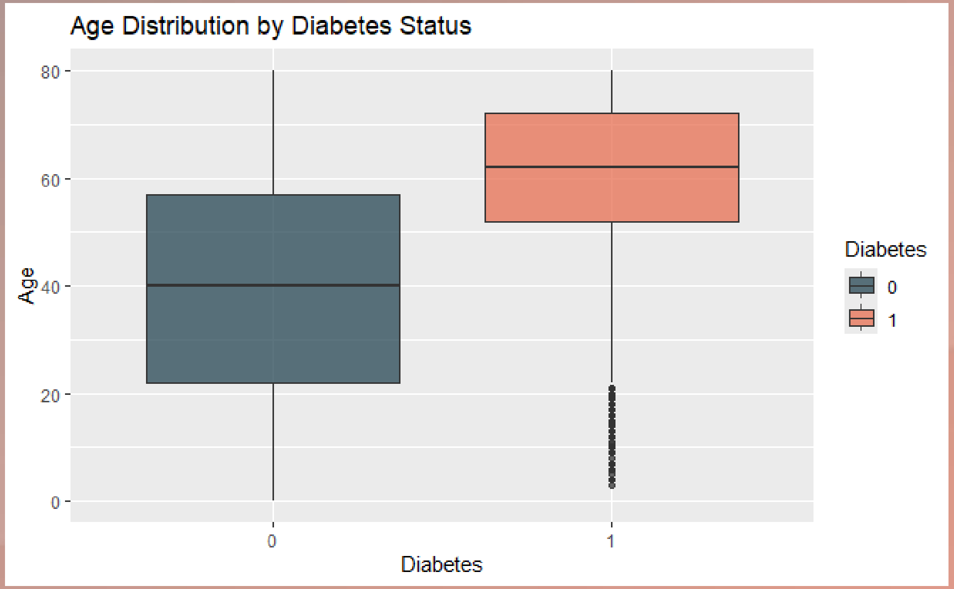
Descriptive Exploration

Continuous Features

Descriptive Exploration

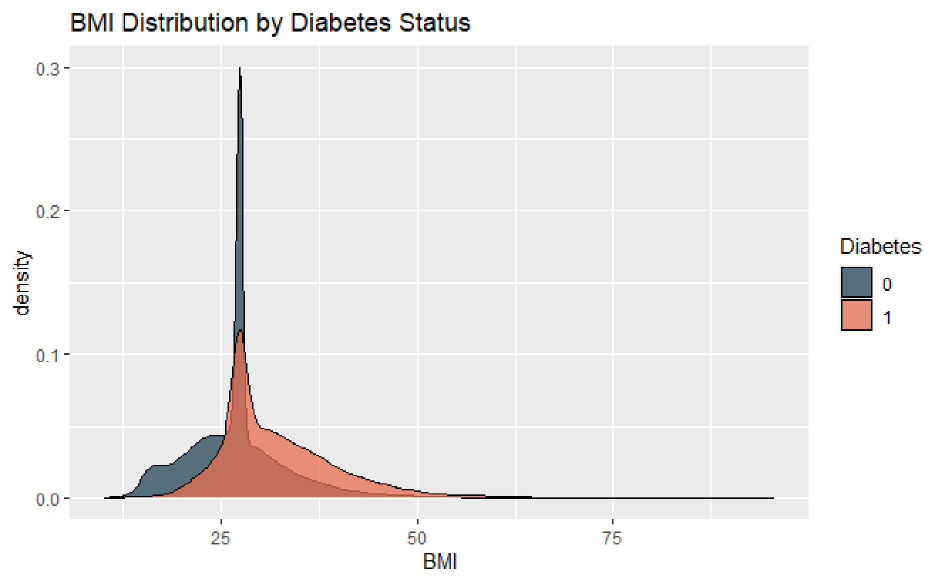
Continuous Features

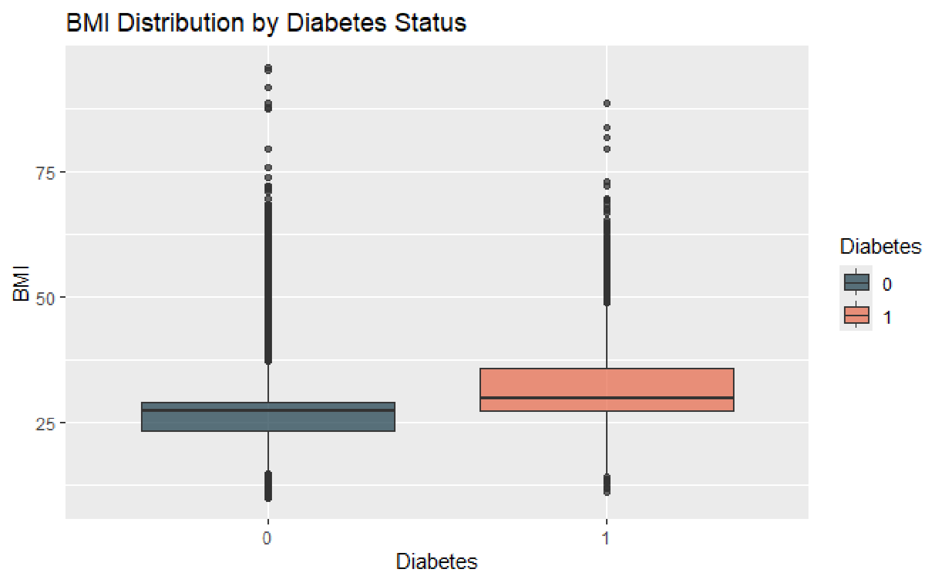




Descriptive Exploration

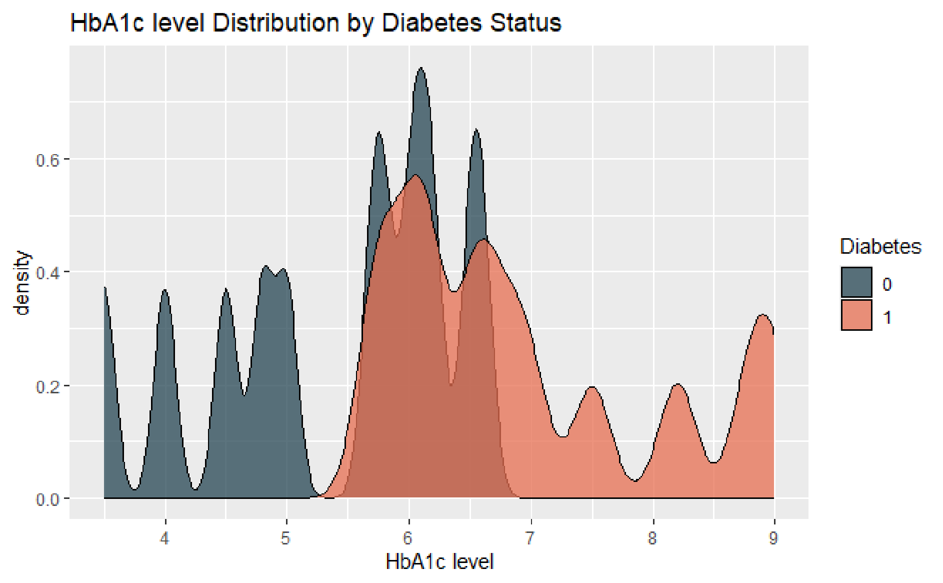
Continuous Features



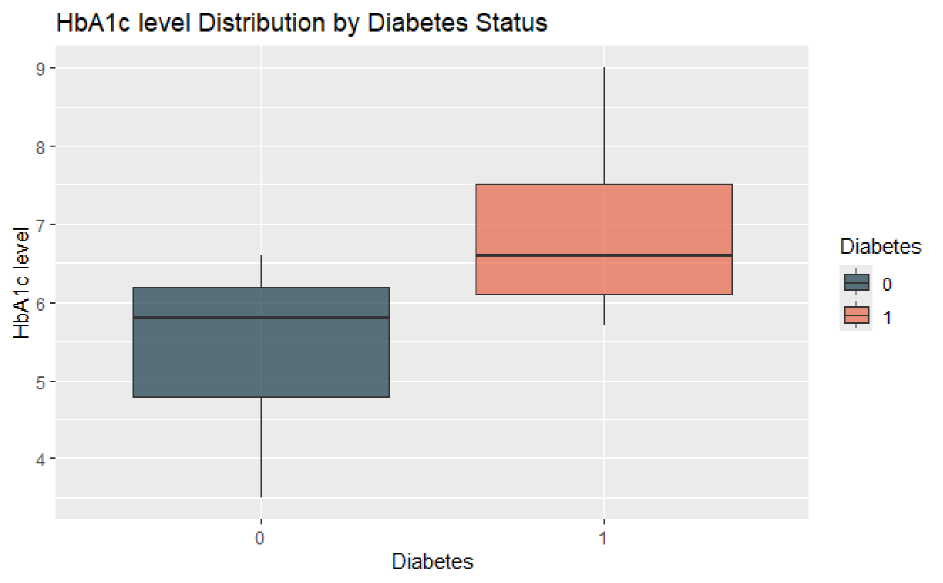


Descriptive Exploration

Continuous Features



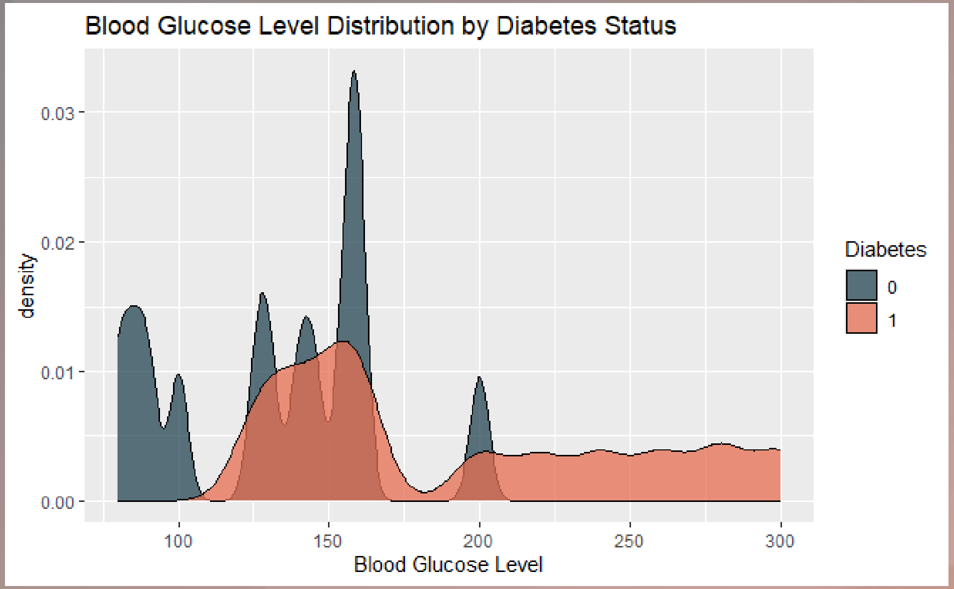
HbA1c level Distribution by Diabetes Status

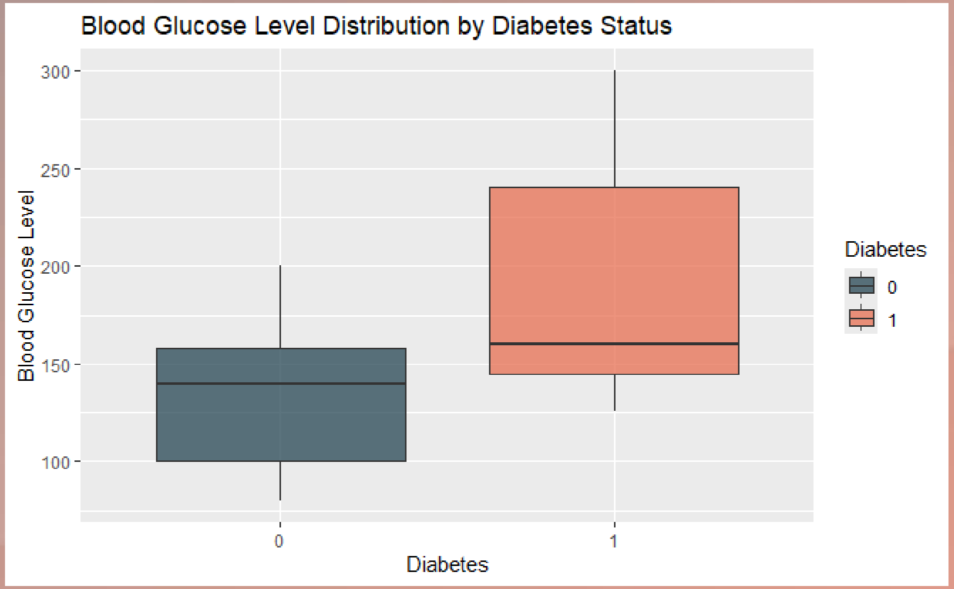


HbA1c level Distribution by Diabetes Status

Descriptive Exploration

Continuous Features



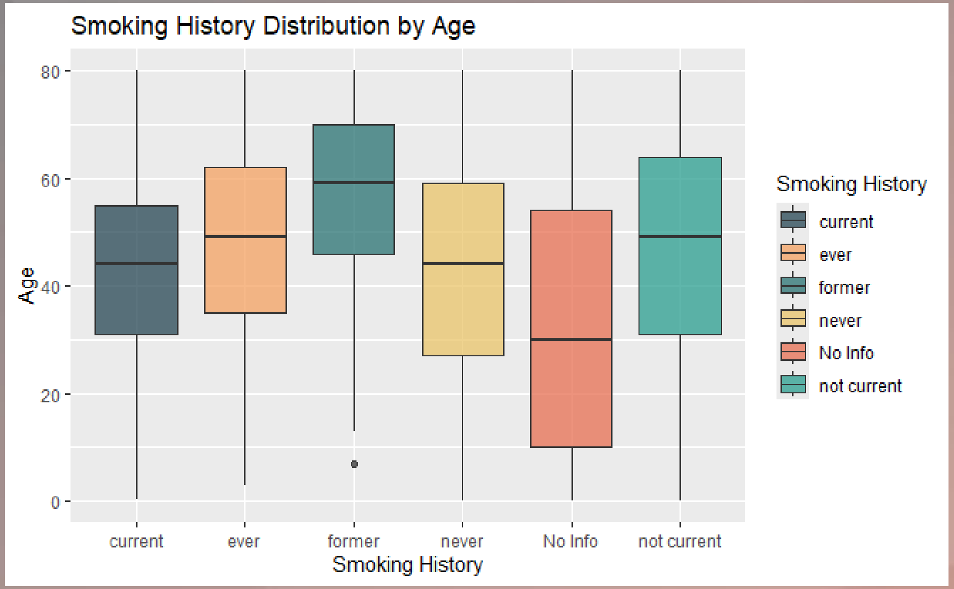


Descriptive Exploration

Inter-feature relationships

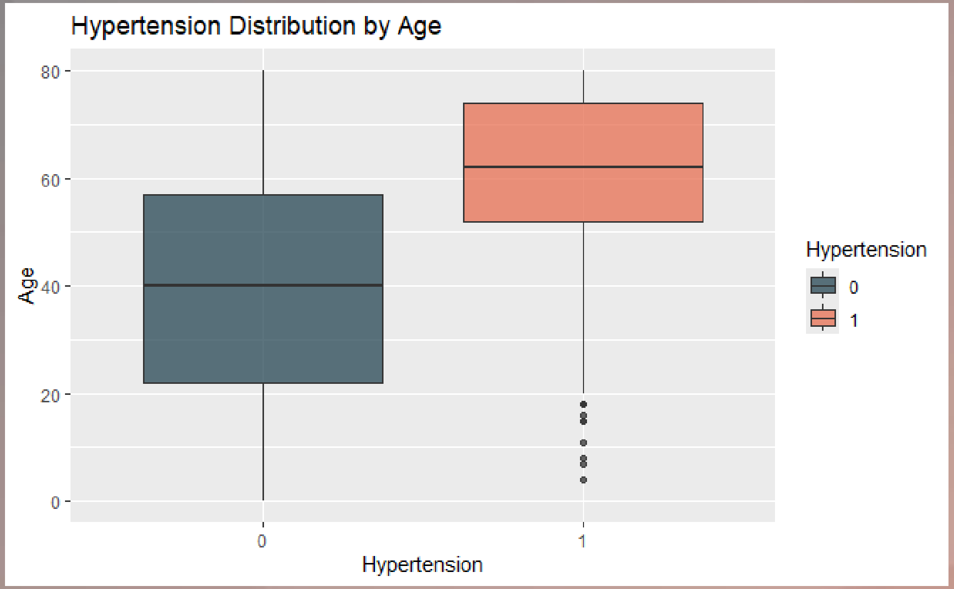
Descriptive Exploration

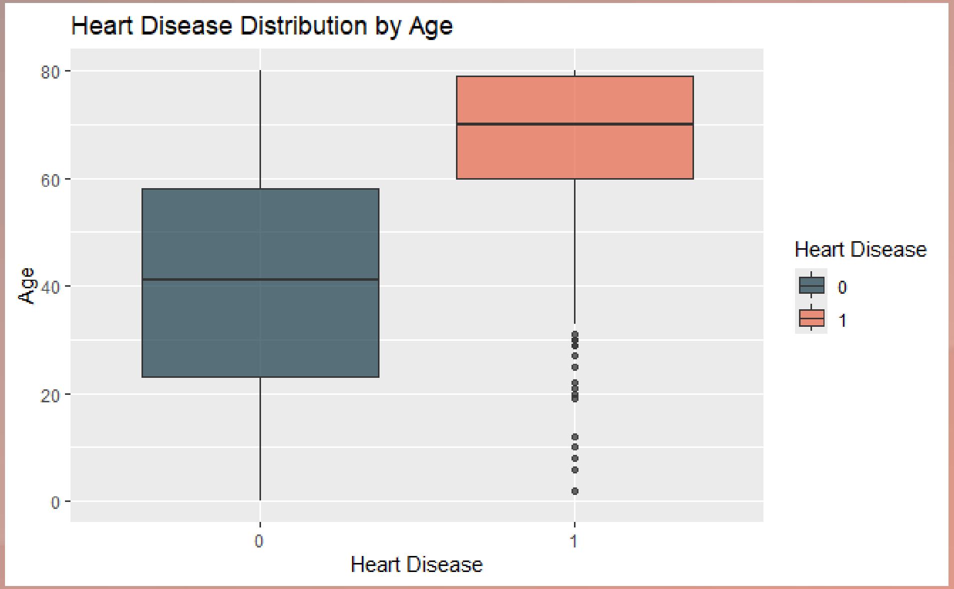
Inter-feature relationships



Descriptive Exploration

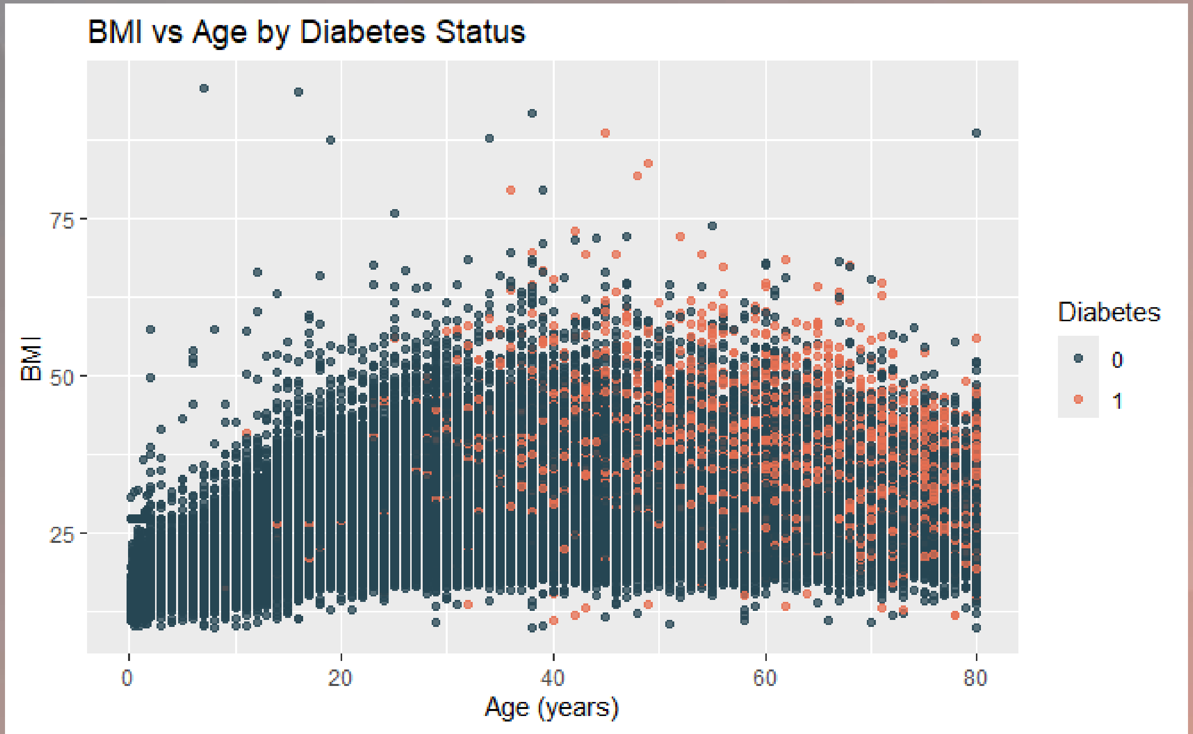
Inter-feature relationships





Descriptive Exploration

Inter-feature relationships

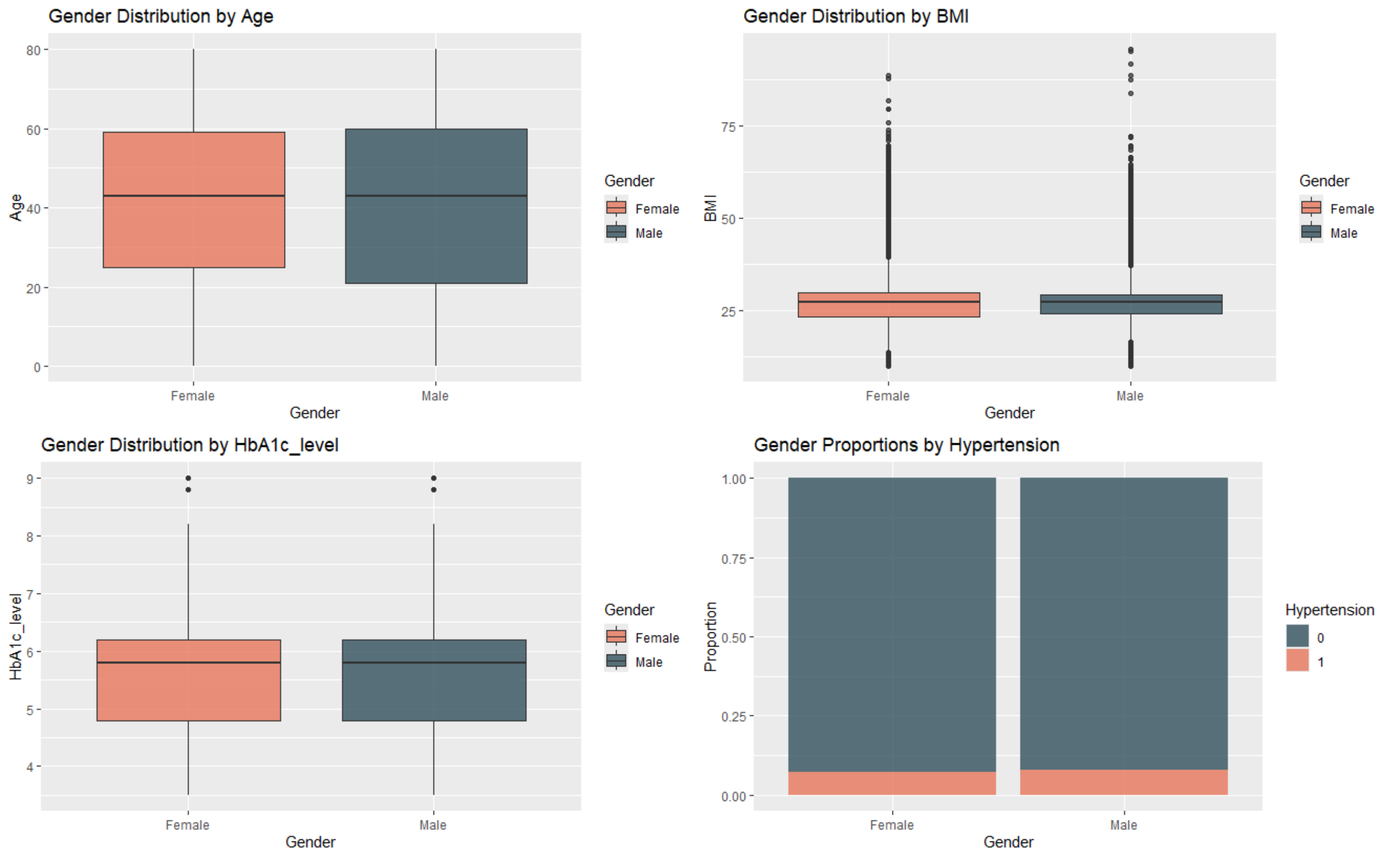


Descriptive Exploration

Inter-feature relationships

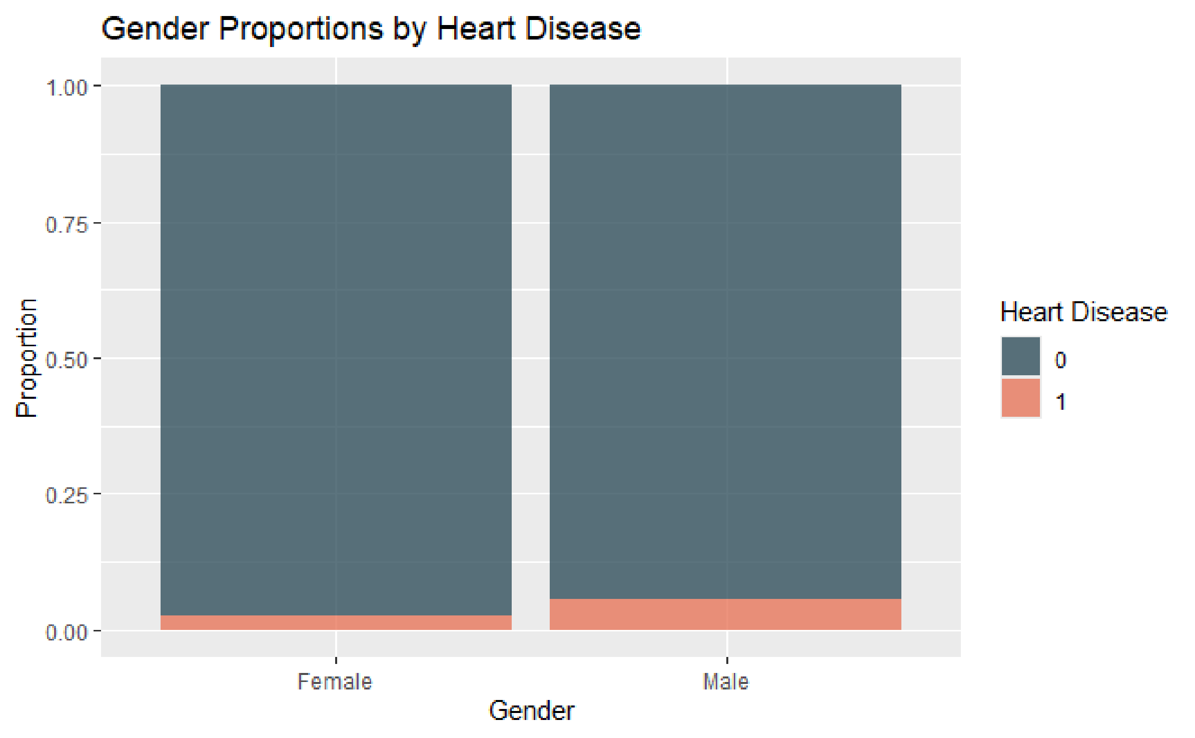
Descriptive Exploration

Inter-feature relationships



Descriptive Exploration

Inter-feature relationships

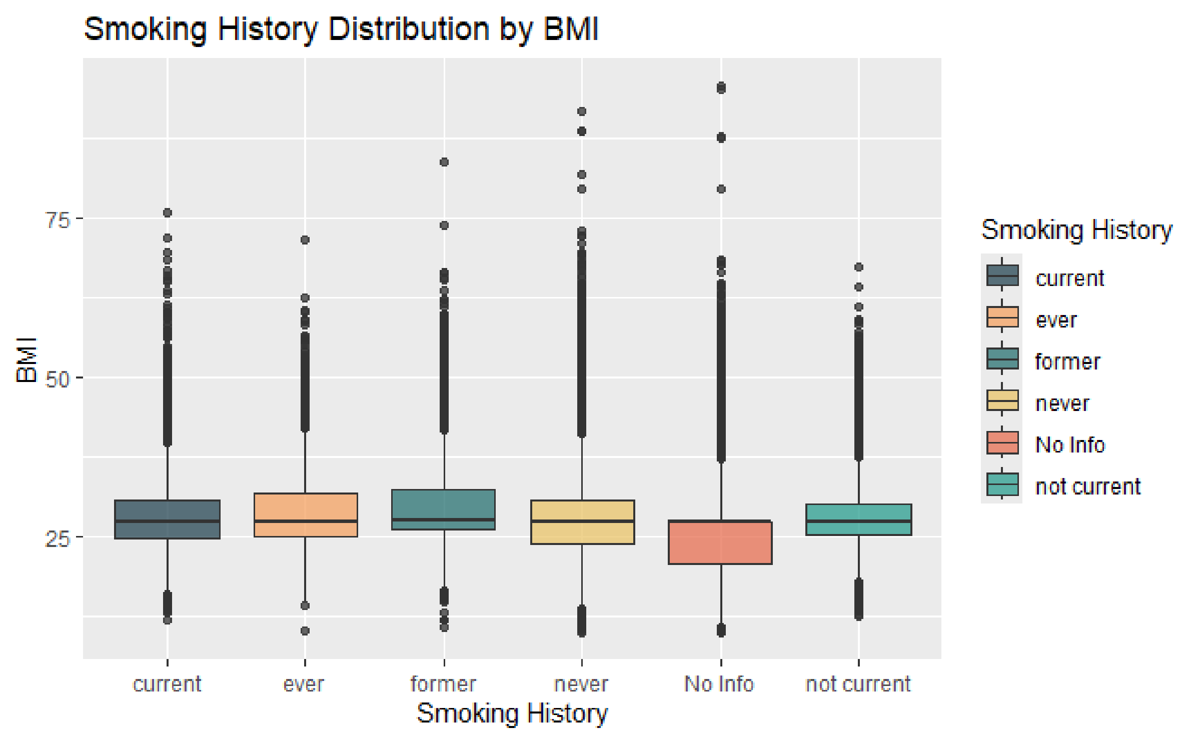


Descriptive Exploration

Inter-feature relationships

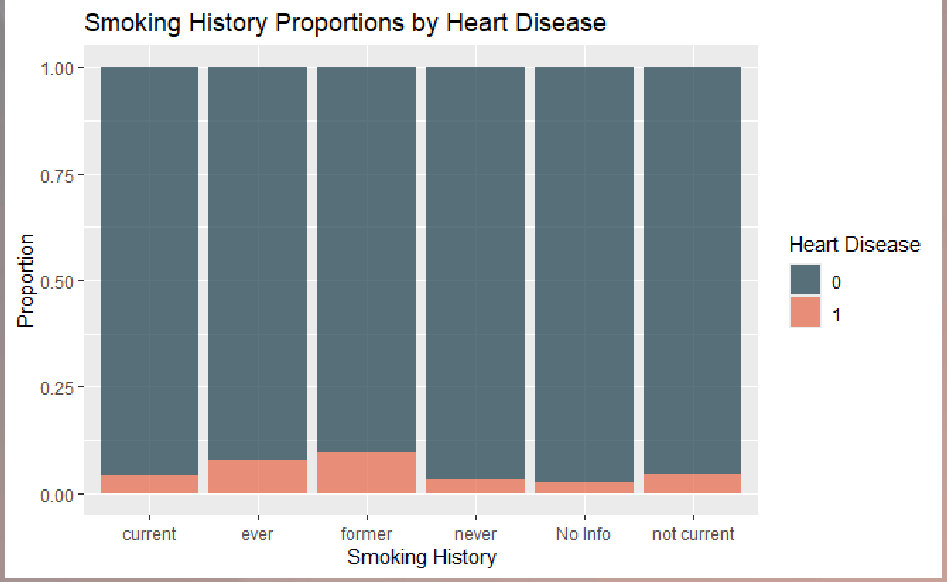
Descriptive Exploration

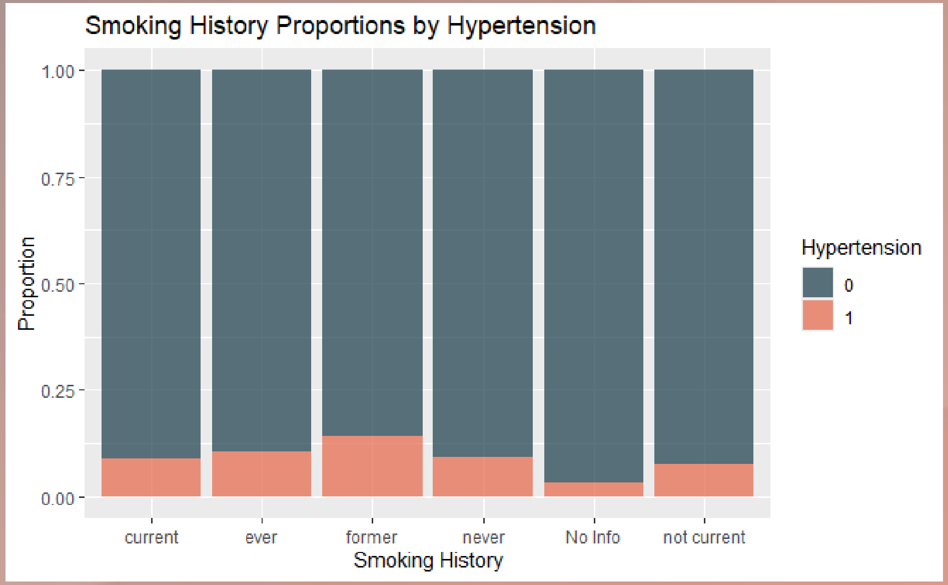
Inter-feature relationships



Descriptive Exploration

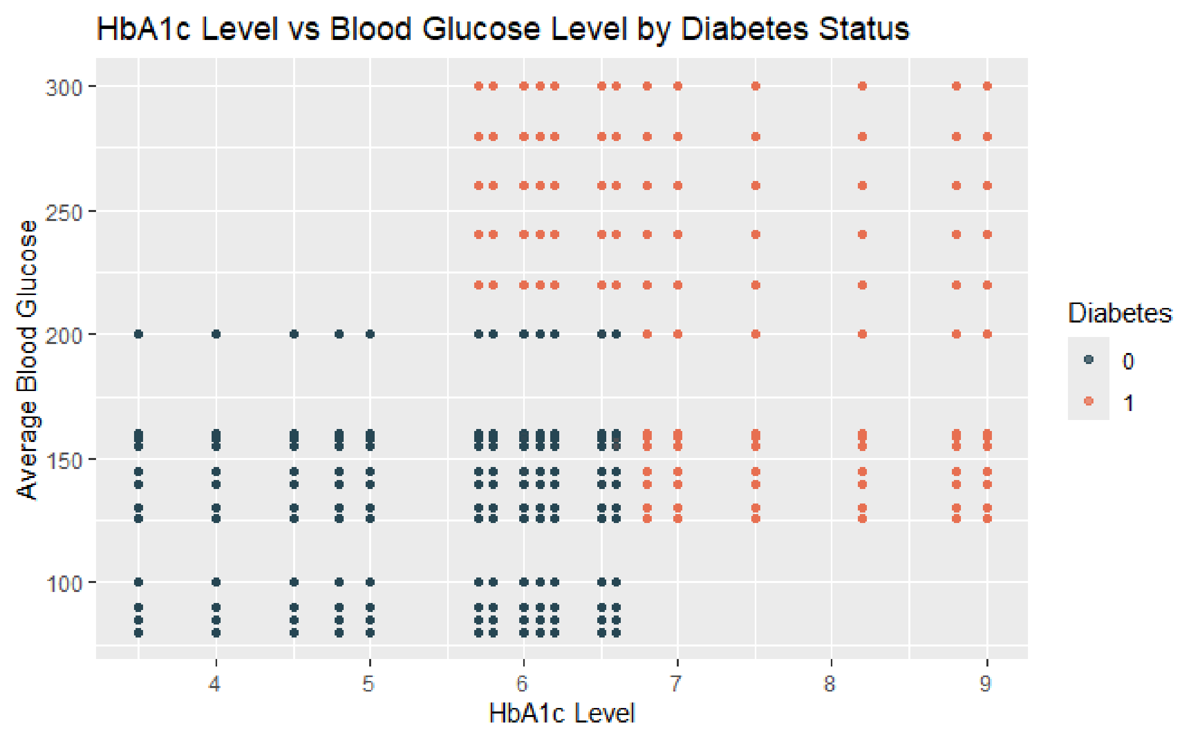
Inter-feature relationships



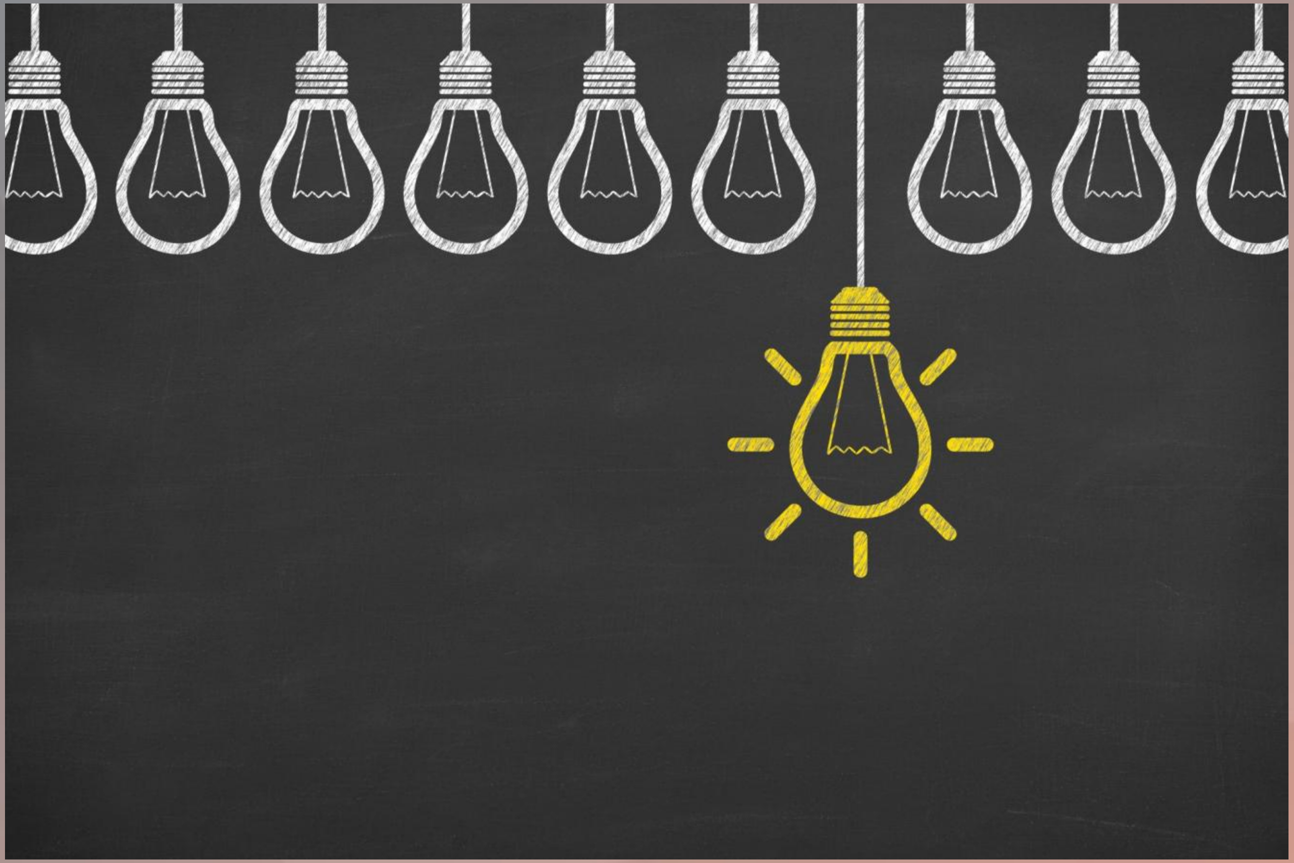


Descriptive Exploration

Inter-feature relationships



Analysis Summary



Inferential Analysis

- Utilising hypothesis testing methods
- 5% significance level
- p-values below this considered statistically significant
- Consistent with common health-related research practices

Hypothesis Testing

Hypotheses:

- BMI differs between diabetic and non diabetic individuals
- Heart disease presence differs between gender
- Heart disease presence differs between smoking status
- Heart disease and diabetes occur independently

Hypothesis Testing

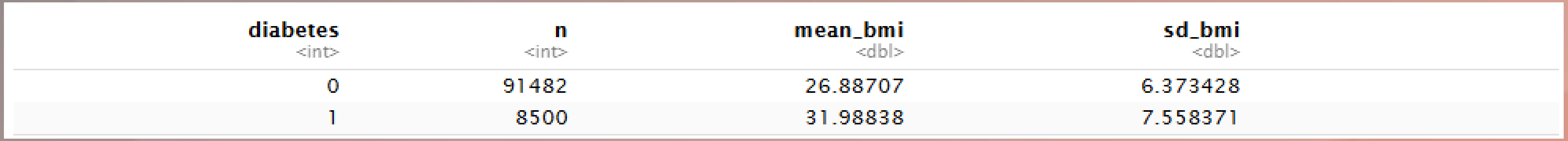
Hypothesis 1: BMI Mean differs between diabetics and non-diabetics.

- **H_0:** The BMI mean is the same for diabetics and non-diabetics
- **H_1:** The BMI mean differs between diabetics and non-diabetics

Hypothesis Testing

Hypothesis 1: BMI Mean differs between diabetics and non-diabetics.

- **H_0:** The BMI mean is the same for diabetics and non-diabetics
- **H_1:** The BMI mean differs between diabetics and non-diabetics

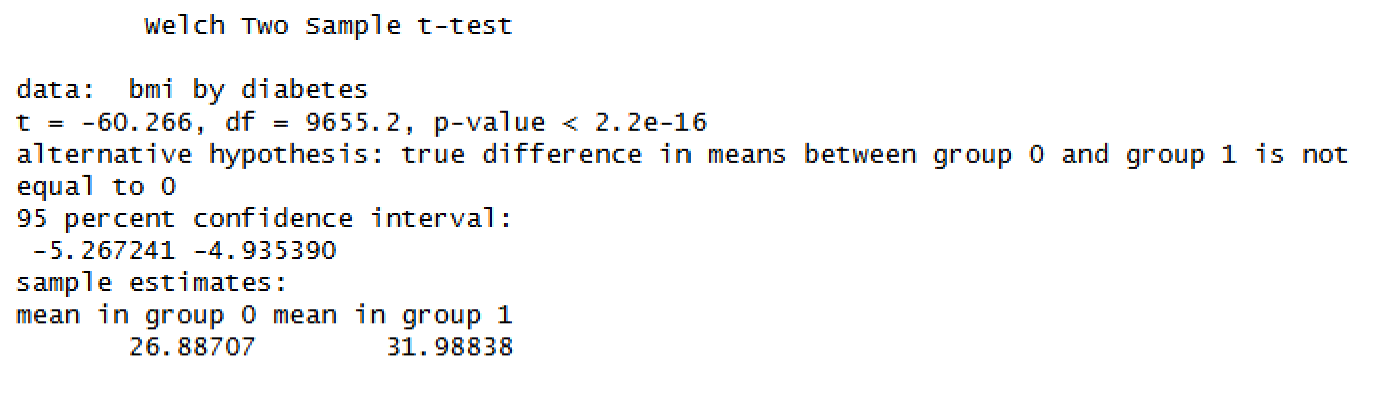


diabetes	n	mean_bmi	sd_bmi
<int>	<int>	<dbl>	<dbl>
0	91482	26.88707	6.373428
1	8500	31.98838	7.558371

Hypothesis Testing

Hypothesis 1: BMI Mean differs between diabetics and non-diabetics.

- **H_0:** The BMI mean is the same for diabetics and non-diabetics
- **H_1:** The BMI mean differs between diabetics and non-diabetics

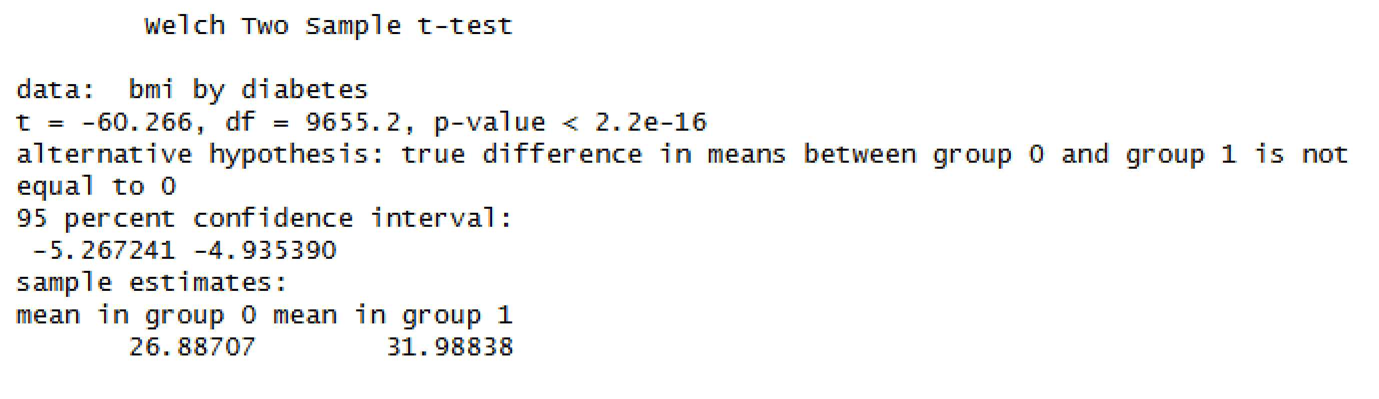


```
        Welch Two Sample t-test

data:  bmi by diabetes
t = -60.266, df = 9655.2, p-value < 2.2e-16
alternative hypothesis: true difference in means between group 0 and group 1 is not
equal to 0
95 percent confidence interval:
 -5.267241 -4.935390
sample estimates:
mean in group 0 mean in group 1
       26.88707        31.98838
```

Hypothesis Testing

Hypothesis 1: BMI Mean differs between diabetics and non-diabetics.

- ~~**H_0:** The BMI mean is the same for diabetics and non-diabetics~~
- **H_1:** The BMI mean differs between diabetics and non-diabetics



```
        Welch Two Sample t-test

data:  bmi by diabetes
t = -60.266, df = 9655.2, p-value < 2.2e-16
alternative hypothesis: true difference in means between group 0 and group 1 is not
equal to 0
95 percent confidence interval:
 -5.267241 -4.935390
sample estimates:
mean in group 0 mean in group 1 
       26.88707        31.98838 
```

Hypothesis Testing

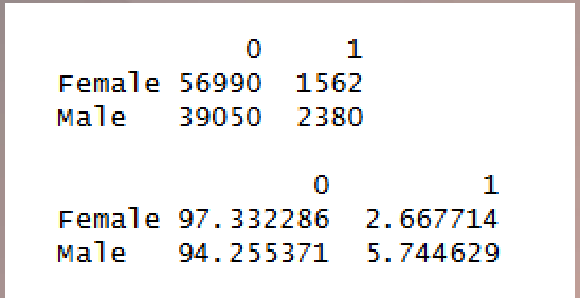
Hypothesis 2: Proportion of heart disease cases differs between genders.

- **H_0:** The proportion of heart disease cases is the same across genders
- **H_1:** The proportion of heart disease cases differs between genders.

Hypothesis Testing

Hypothesis 2: Proportion of heart disease cases differs between genders.

- **H_0:** The proportion of heart disease cases is the same across genders
- **H_1:** The proportion of heart disease cases differs between genders.

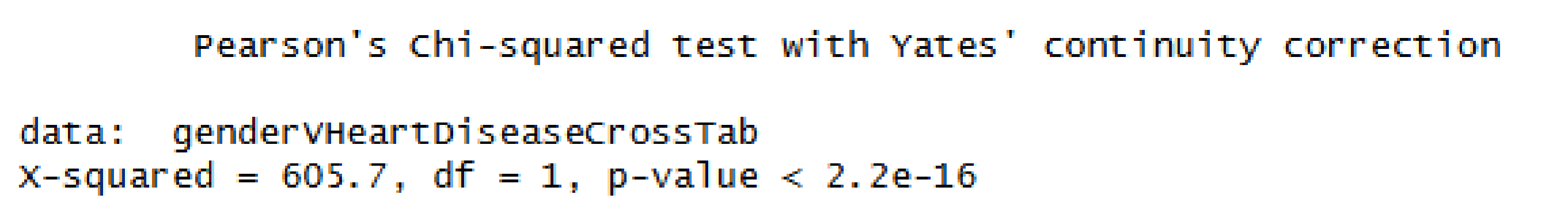


```
            0     1
Female 56990  1562
Male   39050  2380

               0         1
Female 97.332286  2.667714
Male   94.255371  5.744629
```

Hypothesis Testing

Hypothesis 2: Proportion of heart disease cases differs between genders.

- **H_0:** The proportion of heart disease cases is the same across genders
- **H_1:** The proportion of heart disease cases differs between genders.

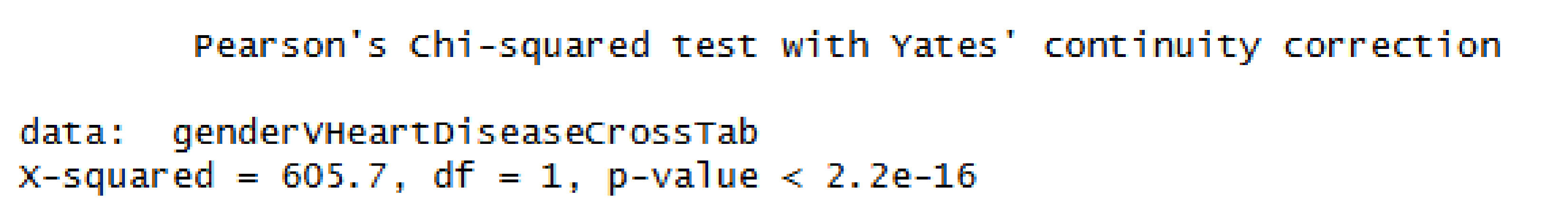


```
	Pearson's Chi-squared test with Yates' continuity correction

data:  genderVHeartDiseaseCrossTab
X-squared = 605.7, df = 1, p-value < 2.2e-16
```

Hypothesis Testing

Hypothesis 2: Proportion of heart disease cases differs between genders.

- ~~**H_0:** The proportion of heart disease cases is the same across genders~~
- **H_1:** The proportion of heart disease cases differs between genders.



```
        Pearson's Chi-squared test with Yates' continuity correction

data:  genderVHeartDiseaseCrossTab
X-squared = 605.7, df = 1, p-value < 2.2e-16
```

Hypothesis Testing

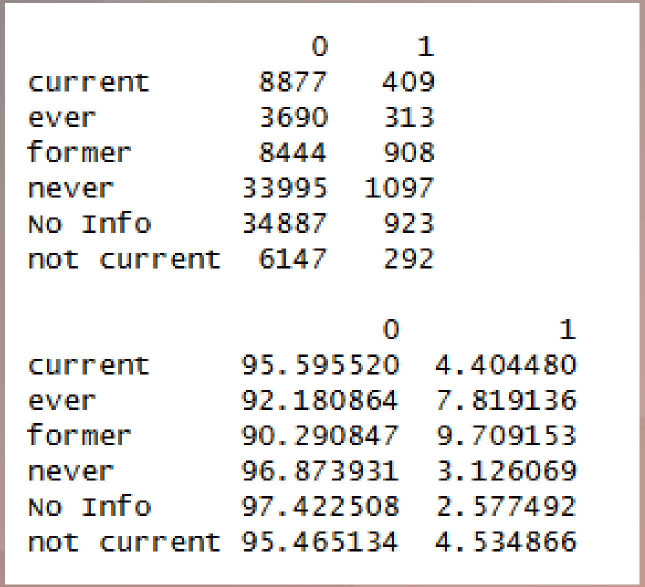
Hypothesis 3: Proportion of heart disease cases differs between smoking histories.

- **H_0:** The proportion of heart disease cases is the same across smoking histories
- **H_1:** The proportion of heart disease cases differs between smoking histories.

Hypothesis Testing

Hypothesis 3: Proportion of heart disease cases differs between smoking histories.

- **H_0:** The proportion of heart disease cases is the same across smoking histories
- **H_1:** The proportion of heart disease cases differs between smoking histories.

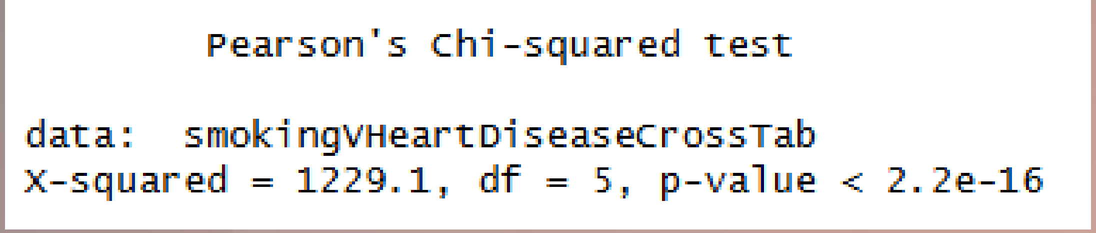


```
                0     1
current      8877   409
ever         3690   313
former       8444   908
never       33995  1097
No Info     34887   923
not current  6147   292

                    0         1
current     95.595520  4.404480
ever        92.180864  7.819136
former      90.290847  9.709153
never       96.873931  3.126069
No Info     97.422508  2.577492
not current 95.465134  4.534866
```

Hypothesis Testing

Hypothesis 3: Proportion of heart disease cases differs between smoking histories.

- **H_0:** The proportion of heart disease cases is the same across smoking histories
- **H_1:** The proportion of heart disease cases differs between smoking histories.

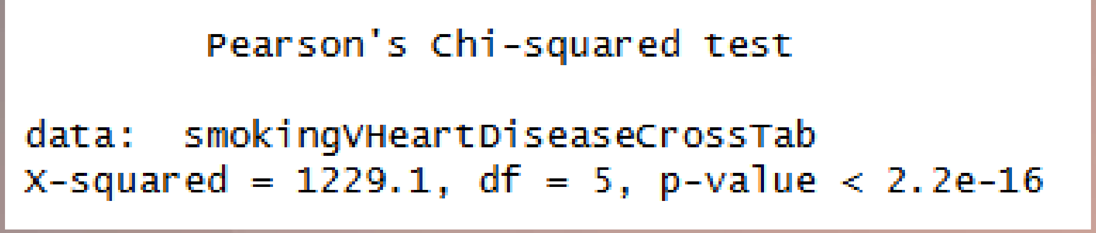


```
        Pearson's Chi-squared test

data:  smokingVHeartDiseaseCrossTab
X-squared = 1229.1, df = 5, p-value < 2.2e-16
```

Hypothesis Testing

Hypothesis 3: Proportion of heart disease cases differs between smoking histories.

- ~~**H_0:** The proportion of heart disease cases is the same across smoking histories~~
- **H_1:** The proportion of heart disease cases differs between smoking histories.



```
        Pearson's Chi-squared test

data:  smokingVHeartDiseaseCrossTab
X-squared = 1229.1, df = 5, p-value < 2.2e-16
```

Hypothesis Testing

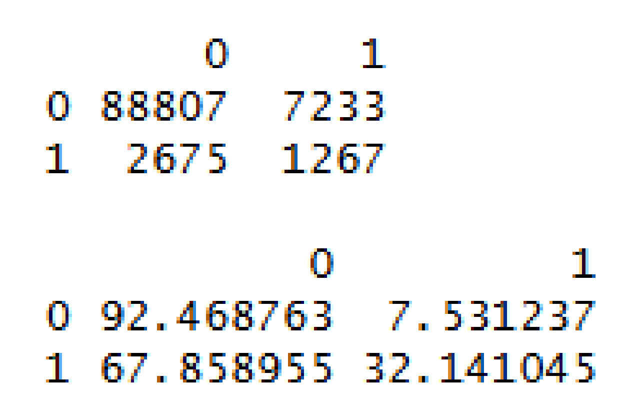
Hypothesis 4: Heart disease and diabetes are independent conditions

- **H_0:** Heart disease and diabetes are independent conditions
- **H_1:** Heart disease and diabetes are not independent

Hypothesis Testing

Hypothesis 4: Heart disease and diabetes are independent conditions

- **H_0:** Heart disease and diabetes are independent conditions
- **H_1:** Heart disease and diabetes are not independent

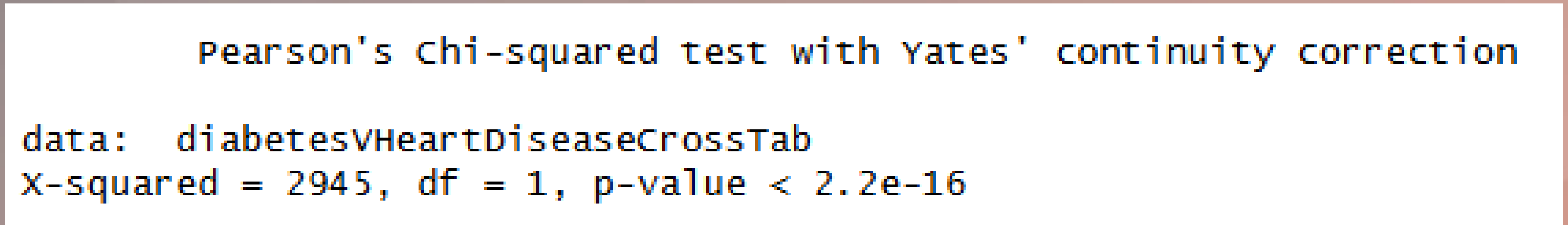


```
      0     1
0 88807  7233
1  2675  1267

          0         1
0 92.468763  7.531237
1 67.858955 32.141045
```

Hypothesis Testing

Hypothesis 4: Heart disease and diabetes are independent conditions

- **H_0:** Heart disease and diabetes are independent conditions
- **H_1:** Heart disease and diabetes are not independent

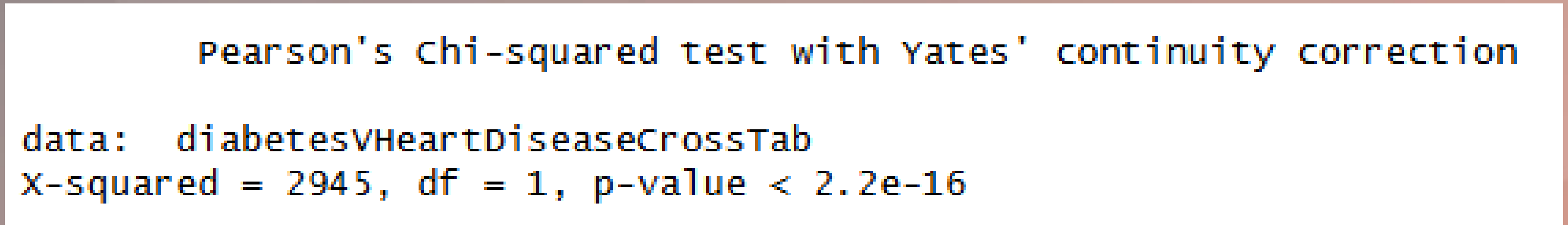


```
        Pearson's Chi-squared test with Yates' continuity correction

data:  diabetesVHeartDiseaseCrossTab
X-squared = 2945, df = 1, p-value < 2.2e-16
```

Hypothesis Testing

Hypothesis 4: Heart disease and diabetes are independent conditions

- ~~**H_0:** Heart disease and diabetes are independent conditions~~
- **H_1:** Heart disease and diabetes are not independent



```
        Pearson's Chi-squared test with Yates' continuity correction

data:  diabetesVHeartDiseaseCrossTab
X-squared = 2945, df = 1, p-value < 2.2e-16
```

Conclusion(s)

- Diabetic individuals exhibited a substantially higher BMI mean than non-diabetics ($p < 0.05$)
- Heart disease prevalence was significantly higher among males ($p < 0.05$)
- Heart disease prevalence was moderately higher among individuals who reported any smoking history ($p < 0.05$)
- Diabetes and heart disease were found to be strongly related ($p < 0.05$)

Future Analysis

- Dataset authenticity
- Dataset features
- Predictive modelling
- Logistic Regression

References

- https://www.diabetes.org.nz/
- https://www.nhlbi.nih.gov/calculate-your-bmi
- https://pmc.ncbi.nlm.nih.gov/articles/PMC1890993/
- https://diabetes.org/about-diabetes/a1c
- https://www.cdc.gov/diabetes/treatment/treatment-low-blood-sugar-hypoglycemia.html#:~:text=If%20your%20blood%20sugar%20drops,treat%20severely%20low%20blood%20sugar
- https://www.mayoclinic.org/diseases-conditions/diabetes/diagnosis-treatment/drc-20371451
- https://diabetes.org/living-with-diabetes/treatment-care/hyperglycemia