LETTERKENNY INSTITUTE OF TECHNOLOGY

ASSIGNMENT COVER SHEET

Lecturer's Name: Dr **James Connolly**

Assessment Title: CA1 - stroke dataset analysis

Work to be submitted to: Blackboard

Date for submission of work: _____

Place and time for submitting work: _____

To be completed by the Student

Student's Name: **Danfeng Wang**

Class: _____

Subject/Module: Data Science

Word Count (where applicable): **3852**

I confirm that the work submitted has been produced solely through my own efforts.

Student's signature: _____ Date: **16th May 2021**

Notes

Penalties: The total marks available for an assessment is reduced by 15% for work submitted up to one week late. The total marks available are reduced by 30% for work up to two weeks late. Assessment work received more than two weeks late will receive a mark of zero. [Incidents of alleged plagiarism and cheating are dealt with in accordance with the Institute's Assessment Regulations.]

Plagiarism: Presenting the ideas etc. of someone else without proper acknowledgement (see section L1 paragraph 8).

Cheating: The use of unauthorised material in a test, exam etc., unauthorised access to test matter, unauthorised collusion, dishonest behaviour in respect of assessments, and deliberate plagiarism (see section L1 paragraph 8).

Continuous Assessment: For students repeating an examination, marks awarded for continuous assessment, shall normally be carried forward from the original examination to the repeat examination.

Abstract

Smoking contributes to man diseases, we hypothesized that smoking is a risk factor for stroke, which could result in an observable effect on younger the average stroke age. Furthermore, aging is believed to be a robust immutable risk factor for stroke.

We constructed related hypothesises based on a dataset of 3246 patients in 5 years, of which 180 patients have history of stroke. All hypothesis tests in this reports are two-sides test by default, and use alpha = 0.05 level of significance.

Our tests have proved that the age is significantly different in stroke and non-stroke patients group. In stroke patients group, there is a difference in mean age for smoker and non-smoker group, which is statistically significant (p-value = 2.2e-16).

Limited by data distribution, whether the stroke patients group is significantly older than the non-stroke group has not been tested. However, in the stroke patient's subgroup, we found a significant difference in age of smokers and non-smokers. After testing the observations have equal variance ratio in both group, we did one-tailed test, which shows the mean age of patients who do not smoke is significantly different with that of the smoker patients. This can be seen from the descriptive analysis, the mean age of non-smoker group is 70~71, and 62~62 is the mean age of smokers.

Shapiro-Wilk test is used to check the normal distribution of age, average glucose value, and bmi, none of them are normally distributed in the corresponding subgroup and the entire population. Man-Whitney test is used for testing the age different in subgroup of stroke patients. Chi-squared tests are applied for testing the correlations between age, smoking and stroke. We also tested the correlation between average glucose level and age by using Spearman's correlation co-efficient test, and compared the co-efficiency values in stroke and non-stroke group.

Overall, age is strongly associated with stroke. Smoking was associated with younger age at stroke patients group.

The related work has been committed to the below repository
https://github.com/rachel0614/stroke.git

Research questions

1. Is there relationship between age and stroke?
2. Are patients who have history of stroke averagely older than those who have not had stroke?
3. Does smoking have relationship with stroke?
4. Are the age of smoker stroke patients averagely younger than stroke patients who does not smoke?
5. Is there relationship between average blood glucose level and bmi?

Data preparation

Dataset description

The dataset being used in this report is stroke patients' information from 2015 to 2021 (See below table). From this table, we can see the data describes the patient information from aspects of basic information, health status, life styles. The dataset is stored as csv format, We are going to clean and transform the data to make it prepared for the next analysis.

Table 1 The raw dataset and description

No.	Name	Description
1	Id	Unique id
2	Gender	Male / Female / Other
3	Age	Age of the patient
4	Hypertension	Has hypertension or not (1/0)
5	Heart disease	Has heart disease or not (1/0)
6	Ever married	Yes / No
7	Work type	Children / Gov_jov / never_worked / Private / Sel-smployed
8	Residence type	Rural / Urban
9	Avg glucose level	Average blood glucose level
10	BMI	Body mass index
11	Smoking status	Formerly smoked / never smoked / smokes / Unknown
12	Stroke	The patient has had a stroke or not (1/0)
13	Date	Record date

Data clean and transformation

The purpose of this stage is to make sure the dataset for analysis is tidy and structured, and prepared for investigate our research questions. The relevant steps will be done as follows.

Step1 – load the dataset from csv, and store each variable into suitable types. Below table is the type conversion plan. The hypertension, heart disease variables' values are 0 or 1. 0 means patient has no such disease, and 1 has the opposite meaning. In order to know clearly what the data represents, we convert them to *No* and *Yes* accordingly when converting them to factor type.

Table 2 Type conversion

No.	Name	Target type and format
1	Gender	Factor
2	Hypertension	Factor with Yes/No labels
3	Heart_disease	Factor with Yes/No labels
4	Ever_married	Factor
5	Work_type	Factor
6	Residence_type	Factor
7	Smoking_status	Factor

8	Stroke	Factor with Yes/No labels
9	Date	Date with format yyyy-mm-dd

The other four columns, which has not been listed in the above table includes id, age, avg_glucose_level, bmi, have been loaded as numeric.

After conversion, we check the structure and sample data to make sure all of the variables have been loaded into the corresponding types and the total amount of the data is 5110 rows with 13 columns.

Step2 – process missing value. Before dealing with the missing data, we calculate and plot the missing values of the dataset first. It can be seen that only bmi variable has a small proportion of missing values (201 rows), we drop it directly.

Step3 – process outliers. Looking from the summary of the dataset, there are rows with *Other* gender, and *unknown* smoking status. After filtering out these outliers by subset function, re-check the data summary to make sure they have been removes properly. Moreover, some factor values are no longer existed because of the filter operation, we refresh the factor level by reconverting them from char to factor type.

Step4 – Based on research questions, smoking_level, diabetes, age_level variables are generated to support the next analysis. Final variables are listed in below table.

Table 3 final variables for analysis

No.	**Name**	**Target type and format**
1	Age	Numeric
2	Bmi	Numeric
3	Obesity	Factor (Yes/No)
4	Smoking_level	Ordinal Factor (never smoked/ formerly smoked / smoke)
5	Stroke	Factor (Yes/No)
6	Avg_glucose_level	Numeric
7	Diabetes	Ordinal Factor (normal, prediabetes, diabetes) According to the rule published by WHO. Avg_glucose_level <= 140 = normal Avg_glucose_level <= 199 = normal Avg_glucose_level > 199 = normal
8	Smoke	Factor(Yes/No)
9	Age_level	Ordinal Factor (under 30,31-40,41-50,51-60,61-70,71 or older)
10	heart_disease	Factor(Yes/No)
11	Ever_married	Factor(Yes/No)

Descriptive statistics

After the dataset has been prepared, descriptive analysis will be applied to give an initial view about the variables correlations. The main variables in this analysis will be described individually. Finally, the normality test of continuous variables will be done with Shapiro-Wilk test approach. The null hypothesis of Shapiro-Wilk test is that *a variable is normally*

distributed in some population. If p-value < 0.05, the null hypothesis is rejected. In other words, the variable is not normally distributed.

Assumption of age dependent variable

Table 4 age distribution by stroke with boxplox

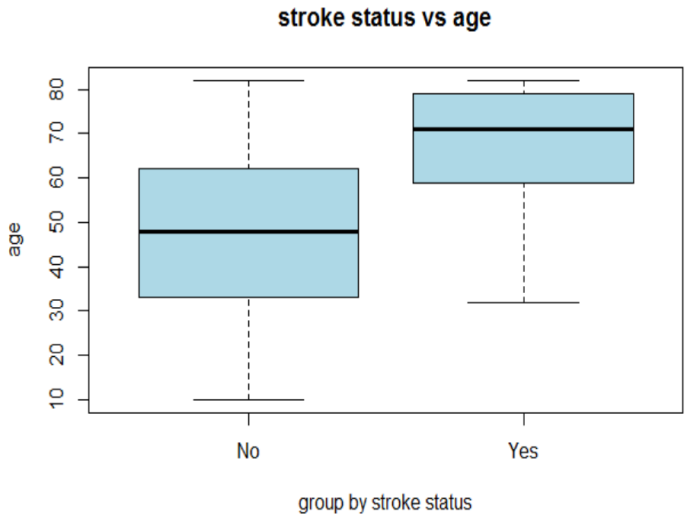
	The median age of stoke group are higher than that of the non-stroke group. We calculate the accurate mean value of both group, the result shows that non-stroke patients are approximately at 47.57 of mean age, while the mean age of stroke group is 68.05. It looks a bell-shaped but not really symmetrical. Both sides are cut off at specific values.
Age – continuous	
Stroke – categorical dichotomous	

Table 5 age normality check by stroke

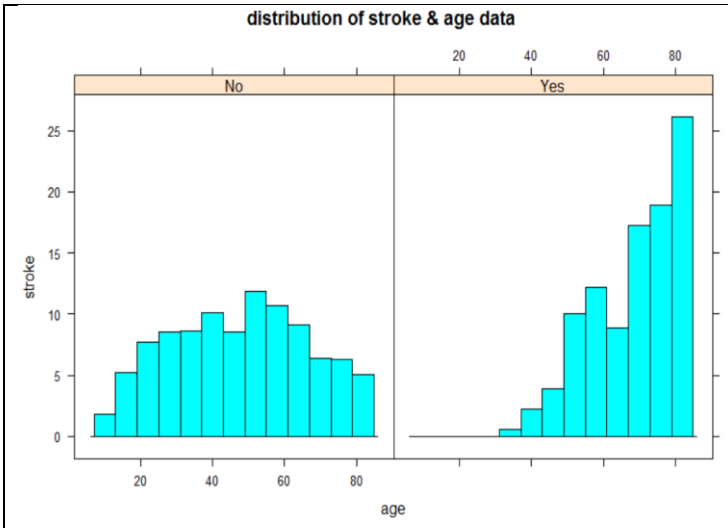
	Check the distribution of patients' age in stroke and non-stroke group. Because our investigation revolves around stroke, so we also split the population by stroke status to check normally distribution for further analysis. The non-stroke group looks a bell-shaped but not really symmetrical, it seems to be normally distributed. The stroke patients group is unlikely to be normally distributed.
Age – continuous	
Stroke – categorical dichotomous	

Table 6 distribution of age for normality check

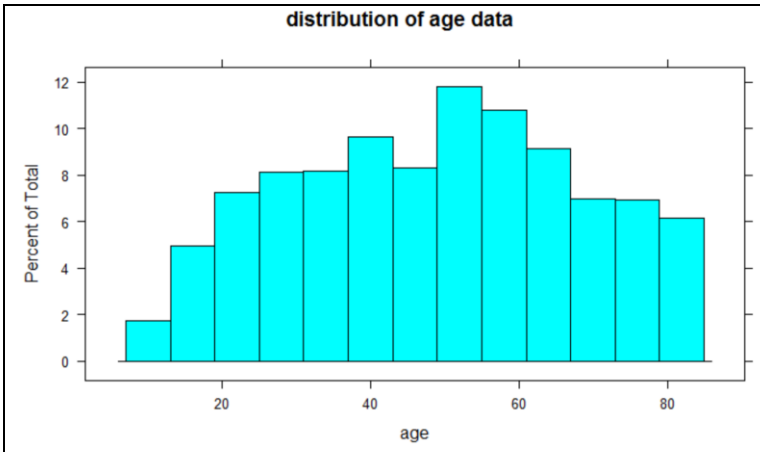
	Check the distribution of patients' age. It looks a bell-shaped but not really symmetrical. Both sides are cut off at specific values. It seems to be normally distributed.
Age – continuous	
Stroke – categorical dichotomous	

Table 7 age distribution for normality check

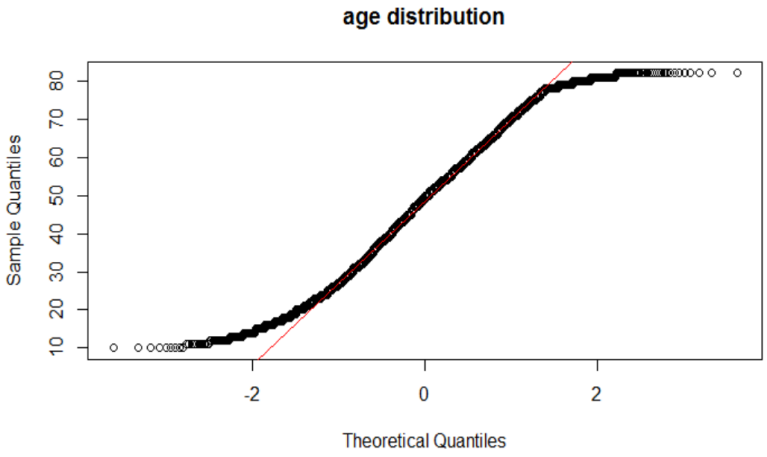
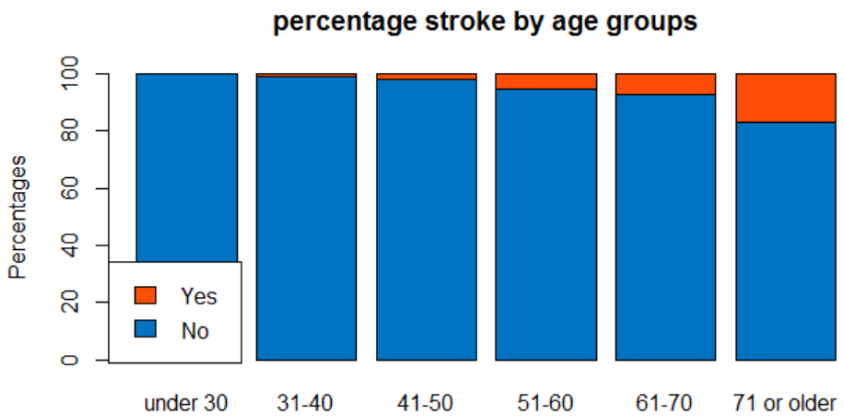
	Check the distribution of patients' age with Q-Q norm with Q-Q line. Though the age from 20 ~ 75 data points distributed along the normal line, the other data points didn't fall on the normal straight line. The data points of age under 20 skewed left and the points of age above 75 skewed right It's unlikely to be normally distributed, but still we need air-tight proof.
Age – continuous	
Stroke – categorical dichotomous	

Table 8 stroke patient distribution in age group



We have cut the age by 6 groups in data preparation stage.

This stack chart shows the proportion of stroke in each age group. It seems the older age group have more stroke patients than the younger age group, and the proportion of stroke patients increased by age.

Age_level – categorical ordinal (6 levels)

Stroke – categorical dichotomous

Table 9 sample count by age group and stroke

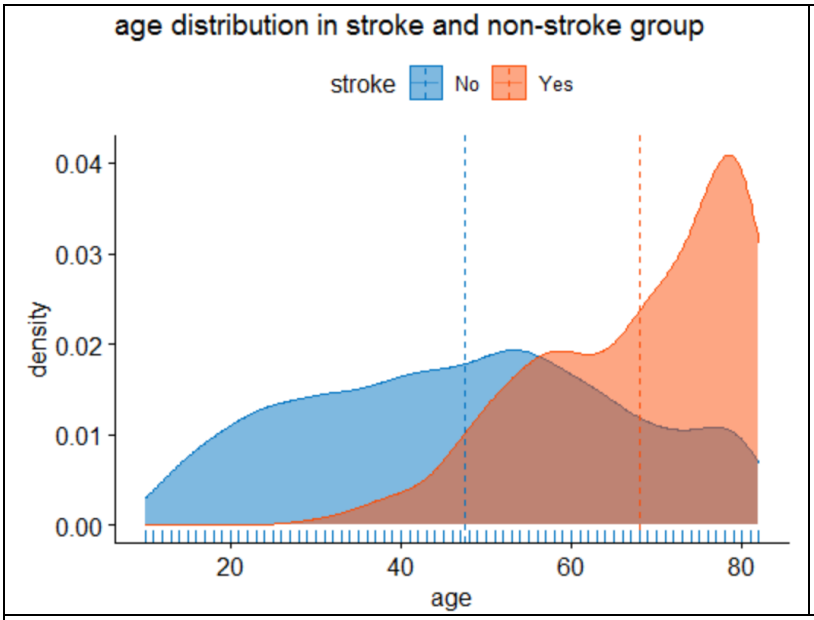
stroke	under 30	31-40	41-50	51-60	61-70	71 or older	Sum
No	652	495	550	627	450	472	3246
Yes	0	5	10	33	35	97	180
Sum	652	500	560	660	485	569	3426

From the contingency table of *age_level* and stroke, we see each group have 5 or over five samples except group of under 30, which is under 20% of the categories. This is for testing whether the sample fulfiled the assumption of Chi-squared test.

Age_level – categorical ordinal (6 levels)

Stroke – categorical dichotomous

Table 10 age distribution density plot by stroke group

	Visualize the underlying distribution of patient age by density group. The marks along the x axis indicates the data location. This graph gives us an evidence of the median value of non-stroke group is obviously different with that of the stroke patient group.
Age – continuous	
Stroke – categorical dichotomous	

According to the Shapiro-Wilk test result, the p-value of non-stroke group, stroke group and entire population is 6.236922e-23, 4.133363e-09, 8.500619e-25 respectively, which are all far smaller than 0.05. In this case, we conclude that age is not normally distributed in our dataset, neither in both stroke and non-stroke group.

---

average glucose variable

Table 11 distribution of average glucose level by stroke status for normality check

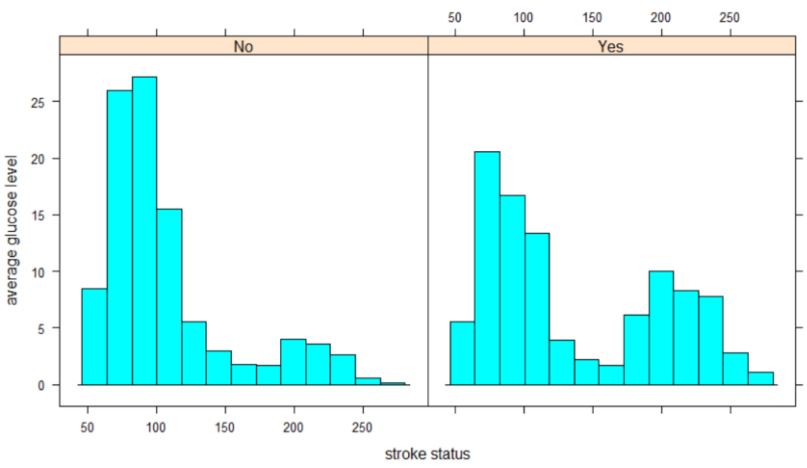
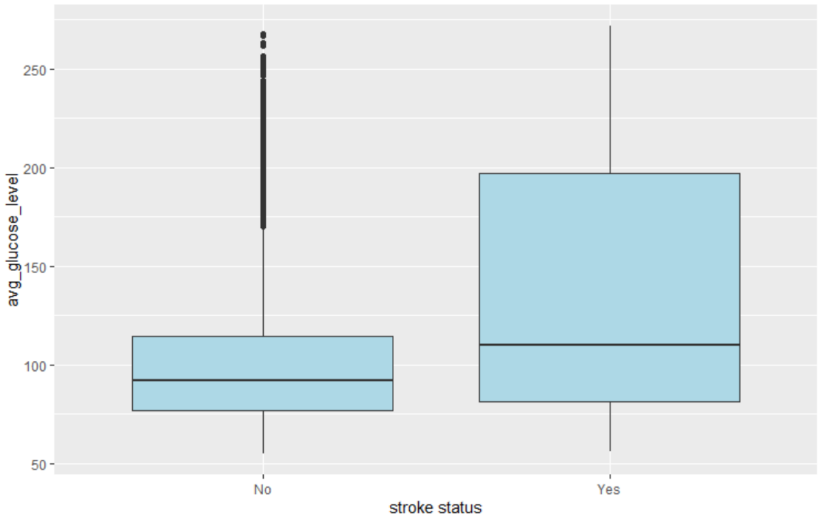
	Check the distribution of glucose level in stroke and non-stroke group. The non-stroke group looks unlikely normally distributed. But it seems both group follow a similar shape.
Age – continuous	
Stroke – categorical dichotomous	

Table 12 distribution of average glucose level by stroke status

<table>
<tr><td> 
</td><td>The side-by-side boxplots seem to indicate many outliers for non-stroke group.

The mean glucose level in stroke group is higher than 125 which is close to the prediabetes level.</td></tr>
<tr><td colspan="2">Avg_glucose_level – continuous</td></tr>
<tr><td colspan="2">Stroke – categorical dichotomous</td></tr>
</table>

According to the Shapiro-Wilk test result, the p-value of average glucose level is 2.212745e-53 in entire population, which is smaller than 0.05. In this case, we conclude that average glucose level is not normally distributed.

smoking status variable

Table 13 sample distribution by stroke by smoking status

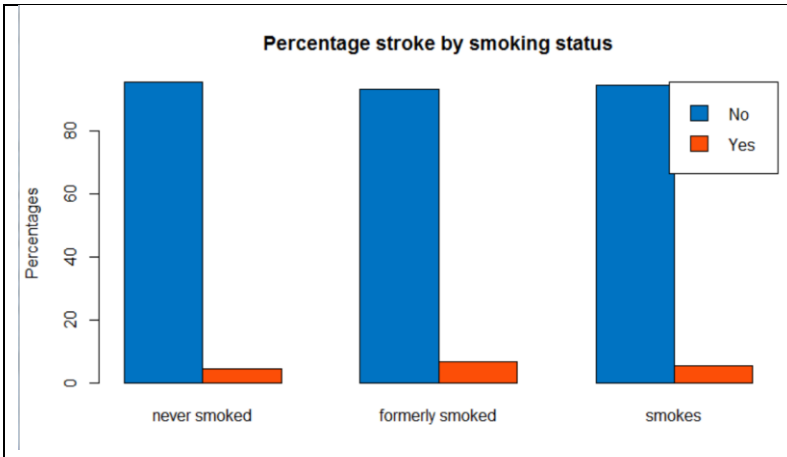
	The stacked bar chart shows the proportion of stroke and non-stroke patients in each smoking status group.
Stroke – categorical dichotomous	
Smoking status – categorical ordinal	

Table 14 smoking status and age of stroke patient group boxplot

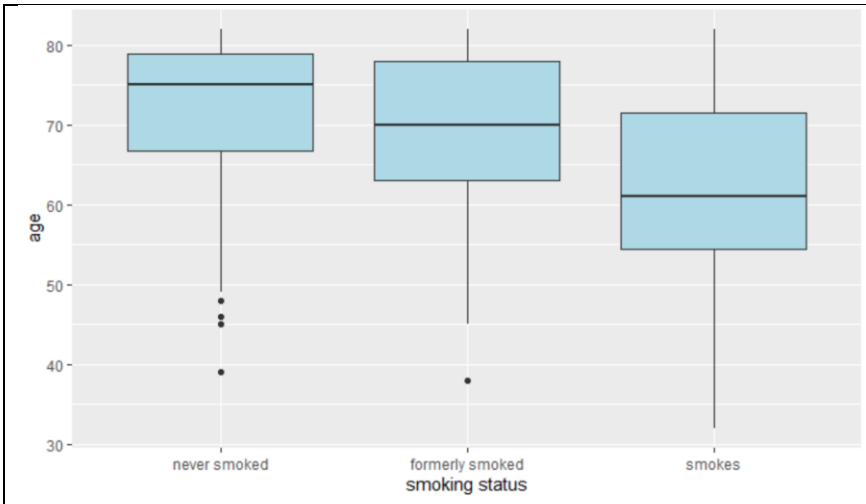
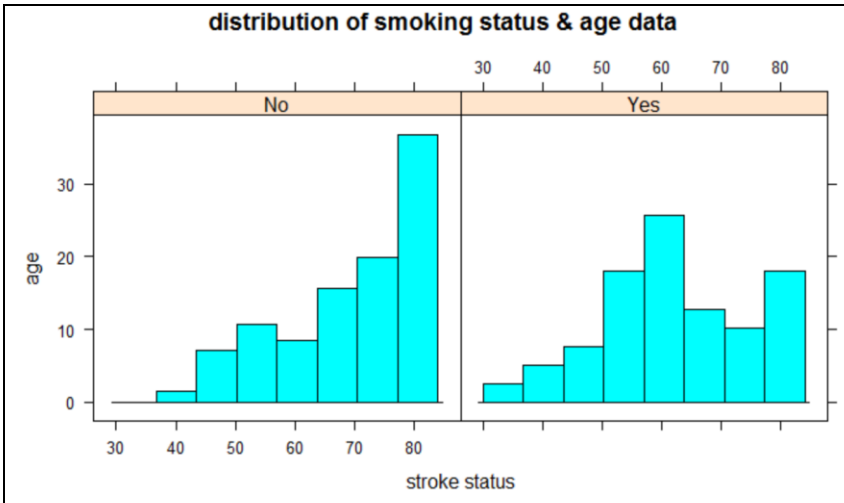
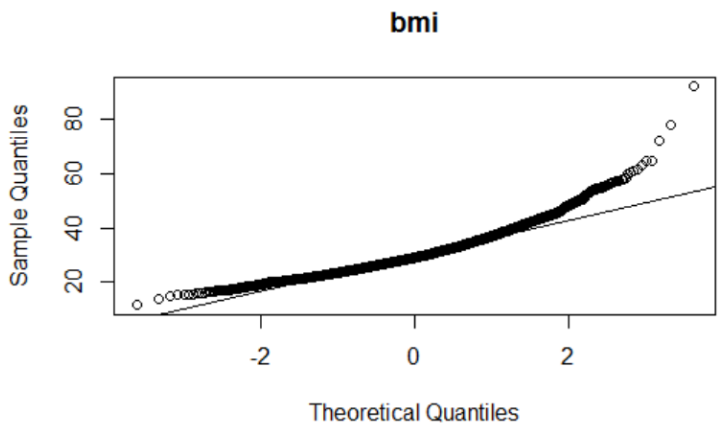
	Though there are outliers distributed in *never smoked* and *formerly smoked* group, that means some younger patients got stroke in both groups, the median age of stoke group seems to be related with smoking degree. The *never smoked* group has highest median age and the *smokes* group has the youngest age, the *ever smoked median age is at the middle of them.*
Stroke – categorical dichotomous	
Smoking status – categorical ordinal	

Table 15 age distribution in stroke patients group by smoking status

	Check the distribution of smoking status of stroke patients group. It looks smoker and non-smoker group has different distribution shape.
Smoke – categorical dichotomous	
Stroke – categorical dichotomous	

bmi variable

After plotting, we did Shapiro-Wilk test to get a precise result of whether the bmi is normally distributed in the entire population. The p-value is 2.855762e-35, which rejected the null hypothesis of normal distribution test. The bmi is not normally distributed in entire population.

<table>
<tr><td> 
</td><td>Check the distribution of patients' bmi with Q-Q norm with Q-Q line.

It's unlikely to be normally distributed, but still we need air-tight proof.</td></tr>
<tr><td colspan="2">bmi – continuous</td></tr>
<tr><td colspan="2"></td></tr>
</table>

According to the Shapiro-Wilk test result, the p-value of bmi 2.855762e-35 in entire population, which is smaller than 0.05. In this case, we conclude that bmi is not normally distributed.

Hypothesis testing and statistical methods

In this section, standard notation of hypothesis based on each research question will be described in each table, as well as the definition and introduction of related variables, and the assumptions and corresponding test method.

The detail explanation of each hypothesis is described below the table.

Question 1

Question - Is there relationship between age and stroke?		
Variable	**Type**	**Description**
Age_level	Independent categorical ordinal	Factor (under 30, 31-40, 41-50,51 or older)
Stroke	Independent variable	Factor (Yes/No)
Hypothesis Test		
H0	age and stroke are independent	
H1	age and stroke are dependent	
Method	Chi-squared test	
Assumption	The levels of stroke category (Yes / No) variable is mutually exclusive, and the patients in both groups are individual patients, two groups data are independent. No more than 20% of the expected count are less than 5 and all individual expected counts are 1 or greater[1].	

Many diseases have significant different in age, gender, and other group identity. We use this hypothesis to test whether there is a correlation between patients age and stroke.

From the *Table 5 age normality check by stroke*, we see stroke patient with over 5 samples only appears after age of roughly 45 years old. In order to meet the assumption of Chi-squared test, we use ordinal *age_level* rather than continuous *age* variable to test the age correlation with stroke, which is exactly being able to fulfil the above assumption of Chi-square test (see *Table 9 sample count by age group and stroke*).

[1] McHugh, M.L., 2013. The chi-square test of independence. *Biochemia medica*, *23*(2), pp.143-149.

Question 2

Question - Are patients who have had stroke older than those who have not had stroke		
Variable	**Type**	**Description**
Age	independent continuous variable	Numeric
Stroke	Independent ordinal variable	Has been converted to ordinal by the smoking degree, which is *never smoker, formerly smoked, smoke*
Hypothesis Test		
H0	the age of stroke and non-stroke patients group are equal	
H1	the age of stroke and non-stroke patients group are not equal	
Method	Mann-Whitney test	
Assumption	Both stroke patients and non-stroke patients are independent, and do not affect each other. The distribution of patient age in both groups have a similar shape, in other words, they have equal variance.	

First we visualize the data to see if the age distribution in both group, and whether it's normally distributed. Then we prove the normality by Shapiro-Wilk test. According to the test result, age is not normally distributed variable in both group. So, we use *Wilcoxon* test to confirm whether the age in both groups are equal.

Furthermore, according to the *Table 4 age distribution by stroke* and *Table 10 age distribution density plot by stroke group* , it seems the mean age of patients in stroke group is higher than that of the non-stroke group. This test might help us to know if stroke is age-related illness. Summarise data *tapply* function to see the exact mean age in both groups. Then use F test to test the variance is equal in both group. If this assumption is met, we do *Wilcoxon* one-tailed test to check if the age of stroke is older than the age of non-stroke.

Question 3

Question - Does smoking have relationship with stroke?		
Stroke	Dependent categorical variable	2 levels factor (Yes/No)
Smoking level	Independent categorical ordinal variable	Converted to ordinal factor by the degree of smoking, which is *never smoker, formerly smoked, smoke*
Hypothesis Test		
H0	smoking has no correlation with stroke	
H1	smoking has correlation with stroke	
Method	Chi-squared test	
Assumption	No more than 20% of the expected count are less than 5 and all individual expected counts are 1 or greater	

The levels of stroke category (Yes / No) variable is mutually exclusive, and the patients in stroke and non-stroke group are individual patients, both groups data are independent.

Both variables in this hypothesis test are categorical, So, we use *Chi-squared* test to confirm the hypothesis.

Question 4

Question – Are the average age of smoker stroke patients younger than stroke patients who does not smoke?		
Age	Independent continuous variable	Numeric
Stroke	Get subset of stroke patients as sample data (stroke=Yes)	2 levels factor (Yes/No)
smoke	Independent categorical variable	2 levels factor (Yes/No) Converted from *smoking status* where *smokes* group is Yes and the other two groups are No.
Hypothesis Test		
H0	in patients who have ever had stroke, age of smoker and age of non-smoker are equal	
H1	in patients who have ever had stroke, age of smoker and age of non-smoker are not equal	
Method	Mann-Whitney test	
Assumption	If the variance of both groups is equal..	

The data applied in this hypothesis test is the patients who have had stroke.

We apply Mann-Whitney test because both variables are categorical which are not from normal distribution.

From the *Table 14 smoking status and age of stroke patient group boxplot*, we can see the age distribution by smoking status (never smoked / formerly smoked / smokes). In order to see roughly if both group follow the same distribution shape, we plot *Table 15 age distribution in stroke patients group by smoking status.*

We check the variances in both groups by F test. If it's true, we do one-tailed Mann-Whitney test : in stroke patients group, smokers is younger than non-smokers.

Question 5

Question - Is there relationship between average blood glucose level and bmi?		
Average glucose level	Independent continuous variable	Numeric
Stroke	Get subset of stroke patients and non-stroke patients as sample respectively	2 levels factor (Yes/No)
Age	Independent continuous variable	Numeric
Hypothesis Test		
H0	average glucose level has no correlation with bmi	
H1	average glucose level has a correlation with bmi	
Method	Spearman's Correlation Co-efficient	
Assumption	Age is not normally distributed Average glucose level is not normally distributed	

Average glucose level and age in the stroke subgroup or non-stroke subgroup are not normally distributed, so that we select Spearman's Correlation Co-efficient test to check if there is relationship between these two continuous variables.

The p-value of Spearman's Correlation Co-efficient does not mean the strength of correlation. Co-efficiency value rho can be checked in both group, which indicate the strength of correlation, then we compare co-efficiency in both patient and non-patient group.

Result

Question 1 – The assumptions of the test are all met. The p-value is 2.2e-16 (scientific form). Therefore, at the 5% significance level, we have strong evidence to reject the null hypothesis that the age has no correlation with stroke. In other words, there is an association between age and stroke.

Question 2 - The assumption of this hypothesis are met except the shape of the distribution for the two groups are roughly same. The distribution shape of age in both stroke and non-stroke patient group are visually different. We test variances by F test. The p-value of F test is 5.485e-13, so that we accept that ration of variances is not equal to 1. It's not suitable to do one-tailed test of whether stroke patients is older than non-stroke patient[2].

Instead, we only test whether the age in both group are equal. The Mann-Withney test on both sides result shows, the p-value is 2.2e-16. Therefore, at the 5% significance level, we reject the null hypothesis, and we conclude that age in stroke and non-stroke patient group are significantly different.

Question 3 – The assumptions are all met. The p-value of Chi-squared test is 0.04996. Therefore, at 5% significance level, the null hypothesis is typically rejected but not with as much confidence as it would be if the p-value were below 0.01. We conclude that smoking are likely having correlation with stroke.

Question 4 – The assumption is met because the p-value of F test is 0.4383 , which is greater than 0.05, so that the ratio of variances of both groups is equal. In the next one-tailed test of whether the non-smoker is older than smoker in the population of stroke patients, the p-value is 0.001007. So, the alternative hypothesis is accepted

Therefore, at 5% significance level, we conclude that in the population of stroke patients, non-smoker is significantly older than smokers.

Question 5 – The assumption are met, both variables in this test are not normally distributed. At 5% significance level, we reject the null hypothesis because the p-value of this Spearman correlation coefficient test in stroke patient group is 6.81e-05, and 2.669e-08 in non-stroke group. So that there is a correlation between average glucose level and bmi in both groups. Though the p-values are all pretty small, it doesn't mean that the relationship between both variables are strong. The rho value in the test of stroke patients group is 0.2923862, which is relatively weak (co-efficiency value 0.21~0.40 is considered to be week)

---

[2] *Assumptions of the Mann-Whitney U test | Laerd Statistics* (no date). Available at: https://statistics.laerd.com/statistical-guides/mann-whitney-u-test-assumptions.php.

[3], while the rho is only 0.09742713 in the non-stroke group, that means the relationship can be ignored.

Conclusion

The research is based on the *stroke* dataset which consist of patient's information gathered from 2000 to 2021. There are 3426 observations with 11 variables after cleaning. 180 of them have ever had stroke. The variables applied in this investigation include stroke, age, smoking, average glucose value, etc.

From this research, we have been able to answer all the predefined research questions. We found age is a risk factor for stroke, and patients who have had stroke seems to have a significant difference in age than those who have no stroke.

We have seen a visually significant difference by plot - the mean age of non-stroke patient group is 47~48, while the mean age of stroke patient group is 68~69. However, limited by the sample distribution, we did not test whether the stroke patients age is significantly higher than that of the non-stroke patients. In the F test of the two population, the p-value is 5.485e-13, which rejected the null hypothesis that the two populations have the same variance. When the variance of two groups are not equal, the interpretations of differences between groups would be difficult.

Moreover, smoking is also a factor for stroke, but it was not supported as strongly as we expected (p-value=0.04996). However, seeing from the stroke patient population, we found there is a significant difference in age between smokers and those who do not smoke, and the non-smoker is significantly older than smokers. From the calculation, mean age of non-smokers in stroke patients group is 69~70, while the same indicator in smoker group is only 62~63, which is obviously younger than the other groups.

We also did two Spearman's Correlation Co-efficient test to check the correlation between average blood glucose level and bmi in the stroke and non-stroke group respectively. The correlation between these two variables in stroke patient group are considered relatively weak. Comparatively, the coefficient value in non-stroke group is negligibly, which is only about 0.09.

The study mainly focussed on age, smoking and the relationship with stroke. Due to space and time limitation, we have ignored the testing of some other important factors of stroke, such as hypertension, heart disease, and etc.

In conclusion, increasing age and smoking are both associated with stroke. Among the stroke patients, smoker shows a significant difference in age than non-smokers, they are more likely to have stroke in younger age than non-smokers.

---

[3] Selala, M., Senzanje, A. and Dhavu, K. (2019) 'Requirements for sustainable operation and maintenance of rural small-scale water infrastructure in Limpopo Province, South Africa', Water SA, 45(2 April). doi: 10.4314/wsa.v45i2.16.