Table1 shows the variables that will be used to answer the hypothesis test1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table 1: Description of variables required for test1** | | | | | |
| **Variable Name** | **description** | **Variable Type** | | **Convert to** | **comments** |
| Gender | Pdf中的描述 | Chr | Categorical  Dichotomous | factor |  |
| Age |  | Num |  |  |  |
| Ever\_married |  | Chr |  | Factor |  |
| Bmi |  | Num |  | Factor | 范围 |
| Date |  | Chr |  | Date |  |
| avg\_glucose\_level |  | Num |  |  |  |
| Hypertension |  | Int |  | Factor |  |
| Stroke |  | Int |  | Factor | Yes/No |

Questions

1. Men and woman differ in getting stroke

H0 - Men and woman do not differ in getting stroke

H1 – Men and woman differ in getting stroke

on the average, the mean age of stroke is 67.73.

H0 – the mean age of man equals to 67.73

H1 – the mean age of man not equals to 67.73

1. smoking has correlation with stroke
2. age has correlation with stroke
3. bmi has correlation with stroke
4. avg\_glucose\_level has correlation with stroke

### Abstract

这部分最后写， 包括 – 检验什么问题，结果是什么，统计方法的简短描述，描述最后的结果包括重要的数据指标比如p-value以及含义，表明检验结果说明了什么

The contribution to the stroke of bmi, hypertension, gender are described in this report. These are all risk factors to stroke. Our goal is to find the patterns of health habit and the risk of getting stroke. We test the correlations between blood pressure, diabetes, BMI and stroke separately with chi-square test of independence. The dependent variable (DV) is stroke which is categorical variable.

Data is collected from ….. from 2015 to 2021.

This data set consists of xx rows and xx columns. For patients records from 2015 to 2021, the following variables are reported: gender (Male, Female), …..It contains categorical variables and continuous variables.

We evaluated the 5000 patients, 50% is woman, 50% is men. Men are younger than woman (数字范围均值+最大值) , 3% of the men has smoke history, 50% of the woman has smoke history. Men had more stroke patients than woman (p-value = 0.4, )

Overall, there is a significant age difference between stroke and non-stroke patients groups. The mean age of stroke group patients is xx which is higher than that of the non-stroke group. So that age is a risk factor contributes to stroke.

### Research question

确认每个问题都能回答

1. Whether there is correlation between gender and stroke.
2. Whether there is correlation between smoking and stroke.
3. Whether there is correlation between age and stroke
4. Whether there is correlation between High average glucose level and stroke
5. Whether there is correlation between smoking and stroke
6. Whether there is correlation between Bmi and stroke.

### 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, before using it, we are going to clean and transform it to make it prepared for the next analysis.

| **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 |

Table 1 Dataset description

#### Data clean and transformation

The purpose of this stage is to make sure the dataset for analysis is tidy and structured. Three stapes will be involved to do this work. After finishing the below steps, descriptive analysis will be applied to give an initial view about the variables correlations.

**Step1** – load the dataset from csv, and store each variable into suitable types. Below table is the type conversion plan.

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **Name** | **Source type** | **Target type and format** |
| 1 | Gender | Char | Factor |
| 2 | Hypertension | Char | Factor with Yes/No labels |
| 3 | Heart\_disease | Char | Factor with Yes/No labels |
| 4 | Ever\_married | Char | Factor |
| 5 | Work\_type | Char | Factor |
| 6 | Residence\_type | Char | Factor |
| 7 | Smoking\_status | Char | Factor |
| 8 | Stroke | Char | Factor with Yes/No labels |
| 9 | Date | Char | Date with format yyy-mm-dd |

Table 2 Type conversion

The other four columns including id, age, avg\_glucose\_level, bmi has 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 201 rows of missing values. We remove it.

**Step3** – process outliers. From the dataset introduction, we have known that there are patients’ data with *Other* gender, and smoking status with *Unknown* type. First, we use summary function to give an overview of the data, and filter out these outliers by subset function. Re-check the data summary to make sure the outliers have been removed. Some factor values are no longer useful because of the filter operation. We remove them by updating the variable types.

**Step4** – get subset. The analysis of this report focus on the risk factors of stroke, *id* and *date* is useless in the data investigation. We remove both columns in the final dataset.

Now the dataset has been prepared. We plot the correlation between the variables. In this report, our main purpose is to check the risk factors of stroke. We split the variables into groups by their content nature, including life-style related variables (gender, age, marital status, work type, residence type), health status related variables (hypertension, heart disease, average glucose value,).

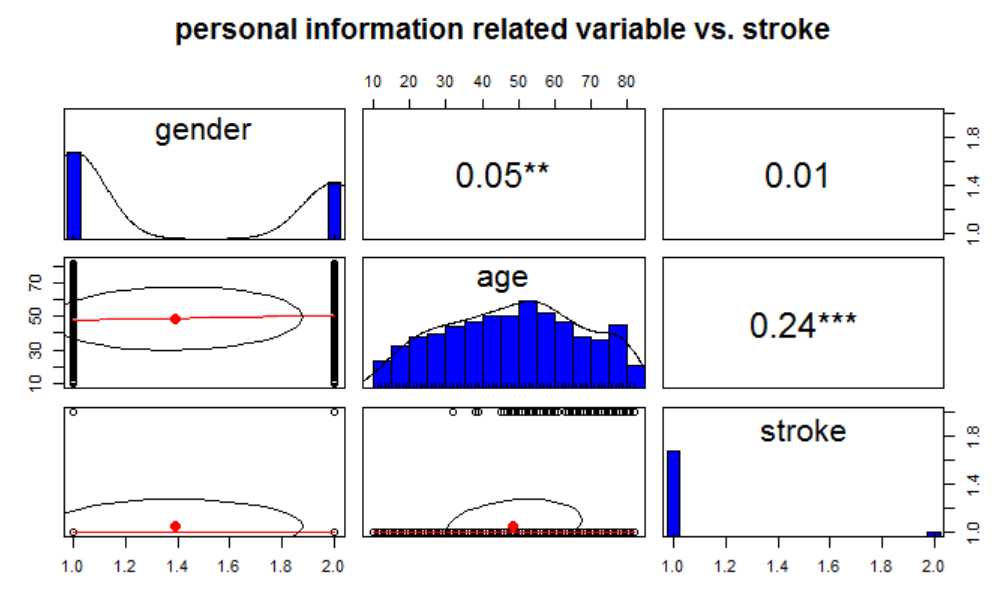


Figure 1 personal information related variables vs. stroke

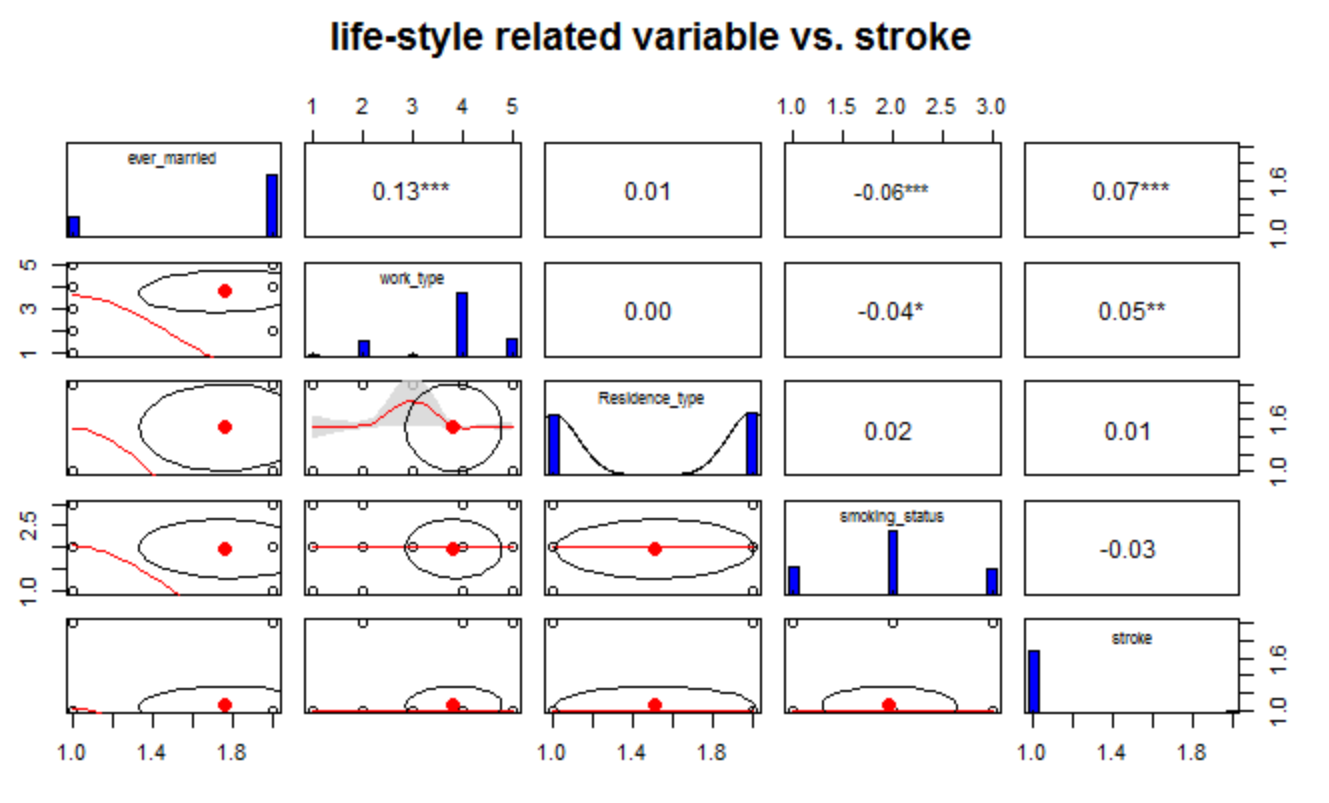


Figure 2 life-style related variables vs stroke

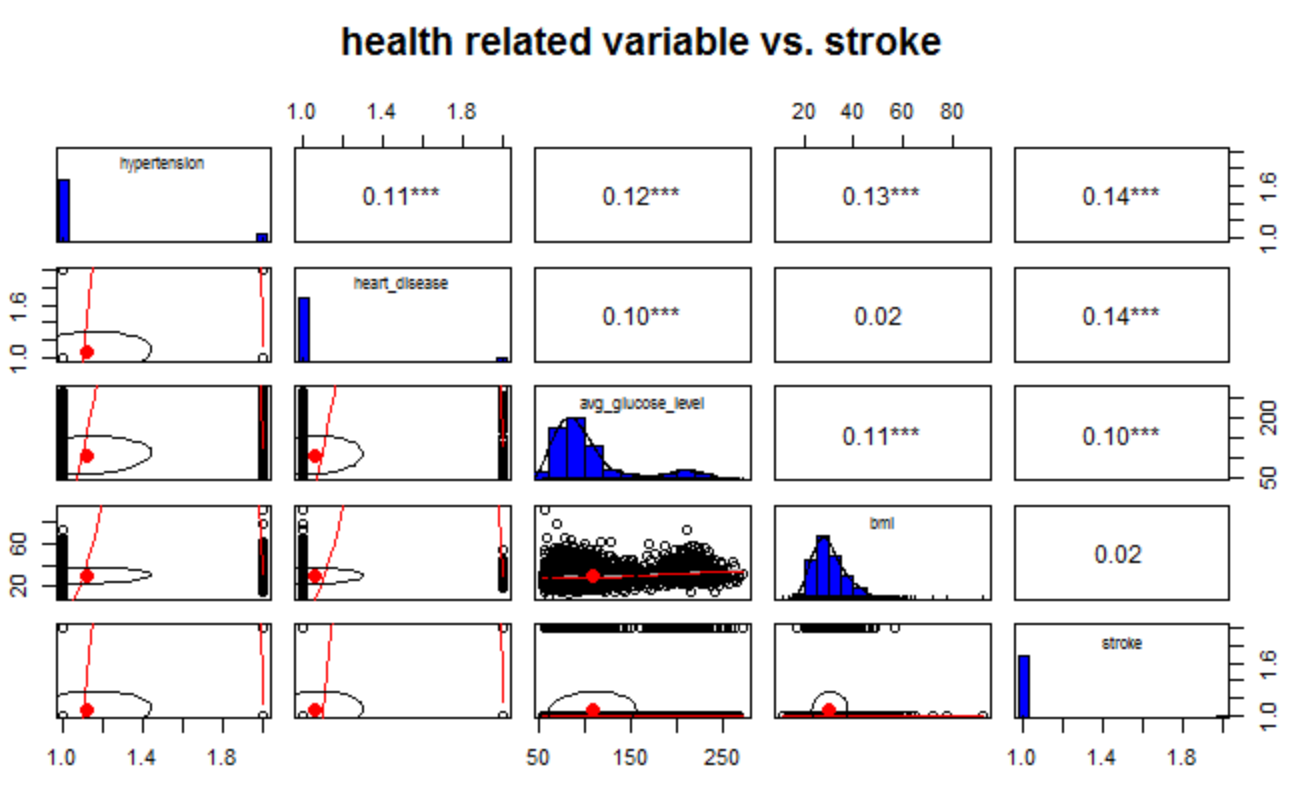


Figure 3 health related variables vs. stroke

用descriptive statistics techniques分析。描述过程，用数据和图表discuss。直到dataset prepared

### Hypothesis testing

Describe the questions in hypothesis test.

Describe the relationship between variables.

Explain how each of the hypothesis test will be valuable when answering each question. 参照问题定义

### Statistical methods

Using the hypothesis test to describe examine each test.

Describe in detail each statistical test by examining the following characteristics

* The structure of the data variables I will examine
* How selected variables enable me to answer the question
* Any assumptions you are making about your data variables such as normality

The below frequency table summarizes two variables from the **stoke** dataset: gender and stroke.

1. Whether there is correlation between gender and stroke.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test1 - Whether there is correlation between gender and stroke | | | | |
| Variable Name | Description | Type | Type | Target Type |
| Gender | Gender of the patient | Char | Categorical | Factor |
| Stroke | Whether the patient has had a stroke in the past or not | Character | Categorical dichotomous | Factor |

H0 – there is no correlation between gender and stroke

H1 – there is correlation between gender and stroke.

There exists gender bias with some diseases. Before making hypothesis test of the correlation of gender and stroke status, we see the data summation first from these two dimensions. Below table shows the number of patients by gender by stroke. Percentage of female patients had stroke in this dataset, is approximately 4.14% (120/2897), and the according percentage of man is about 4.25% (209/4908).

|  |  |  |  |
| --- | --- | --- | --- |
| **Number of patients by gender by stroke** | | | |
| **Gender** | **Non-stroke** | **Stroke** | **Total** |
| **Female** | 2777 | 120 | 2897 |
| **Male** | 1922 | 89 | 2011 |
| **Total** | 4699 | 209 | 4908 |

Judging from the percentage difference between male who got stroke and that of the female group, it appears that there is a slight difference between gender and stroke. However, in order to prove it, we need to do hypothesis test.

We are going to test whether there is correlation between gender and the stroke based on this dataset. Gender is a typical categorical variable, the value of gender consists of Male and Female. Because we found there are outliers in the gender variable in data preparation stage, we remove the row with gender values are neither Male nor Female, and then convert gender from char to factor. Stroke in this dataset means whether the patient ever got stroke or not, we convert the variable into factor and label the original 0 and 1 value with No and Ye respectively.

Both gender and stroke are categorical variables. Gender is independent and stroke we take it as dependent variable. When both the dependent variable and independent variable are categorical, we use Chi-square test to do hypothesis testing.

The null hypothesis in chi-square is that there is no relationship between the independent variable and the dependent variable. So our test is defined as below:

H0 – there is no correlation between gender and stroke status.

H1 – there is correlation between gender and stroke status.

We choose Alpha = 0.05 in this test which means when these two variables have a 0.05 or less probability, there is no relationship between the two variables.

p-value of this test is 0.6805 which indicates weak evidence against the null hypothesis, so that there is no correlation between gender and stroke.

1. Whether there is correlation between smoking and stroke.

H0 – the probability of stroke patients who are smokers equal to that of the non smokers.

H1 – the probability of stroke patients who are smokers not equal to that of the non smokers.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test1 - Whether there is correlation between age and stroke | | | | |
| Variable Name | Description | Type | Type | Target Type |
| Smoke | Whether is a smoker | Character | Categorical dichotomous | Factor |
| Stroke | Whether the patient has had a stroke in the past or not | Character | Categorical dichotomous | Factor |

|  |  |  |  |
| --- | --- | --- | --- |
| **Number of patients by smoking status by stroke** | | | |
| **Smoking** | **Non-stroke** | **Stroke** | **Total** |
| **No** | 1768 | 84 | 1852 |
| **Yes** | 1477 | 96 | 1573 |
| **Total** | 3245 | 180 | 3425 |

1. Whether there is correlation between age and stroke.

H0 – there is no correlation between age and stroke.

H1 – there is correlation between age and stroke

The dependent variable is age, which is continuous variable. Stroke is categorical dichotomous variable. The distribution of stroke is not normally distributed. The dependent variable is not normally distributed in this dataset, so that we use Kruskai-Wallis test of Non-parametric test.

These data samples are independent, and the samples do not affect each other. Using this test, we can decide if the population distribution are identical without assuming them to follow the normal distribution.

The p-value of the test is 2.2e-16 which is obviously less than the significant level 0.05, we can conclude that there are significant age differences between the patients of stroke and non-stroke groups.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test2 - Test2 - Whether there is correlation between hypertension and stroke | | | | |
| Variable Name | Description | Type | Type | Target Type |
| Hypertension | Whether the patient has hypertension | Numeric | Categorical dichotomous | Factor (Yes/No) |
| Stroke | Whether the patient has had a stroke in the past or not | Character | Categorical dichotomous | Factor (Yes,No) |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test3 - Whether there is correlation betweenage and stroke | | | | |
| Variable Name | Description | Type | Type | Target Type |
| Age |  | Numeric | Continuous |  |
| Stroke | Whether the patient has had a stroke in the past or not | Character | Categorical dichotomous | Factor |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test4 - Whether there is correlation between smoking | | | | |
| Variable Name | Description | Type | Type | Target Type |
| Smoking status | "formerly smoked", "never smoked", "smokes" or "Unknown". | Character | Categorical dichotomous |  |
| Stroke | Whether the patient has had a stroke in the past or not | Character | Categorical dichotomous | Factor |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test5 - Whether there is correlation between diabetes 肥胖(从bmi中拆分出是否肥胖) | | | | |
| Variable Name | Description | Type | Type | Target Type |
| BMI | Body mass index | Numeric | Continuous |  |
| Stroke | Whether the patient has had a stroke in the past or not | Character | Categorical dichotomous | Factor |

### ~~Data Analysis~~

We calculated the stroke patients by gender, and compare the outcome by using a t test for continuous variable and chi-square test for categorical variables. All hypothesis tests in this reports are two-sides test, and the significant level is 0.05.

### Result

用R 对variables应用统计学方法回答问题。解释每个测试的输出。解释如何决定每个假设的

XX percent of the sample are woman, the mean age of the sample is xxx years, and woman is sigficantly older than men (55 years vs 50 years). The two groups did not differ in health.

Recognition of stroke

Blood pressure, smoking, and obesity are risk factors of stroke. BMI in a good level (according to WHO) has lower risk in getting stroke.

### Conclusion

讨论每个问题的分析，统计结果表明。对结果明细进行批判。提供深度讨论这个结果表明。