**STAT 112**

**Stroke Project**

**by**

**Tamerlan Assylkhan,**

**Kirill Kiselev,**

**Aiaru Abdirakhman,**

**Buğra Gülsoy**

**and**

**Hadiya Khan**

**Student IDs:** 2604528; 2601706;

2603553; 2283109; 2602084

2023

1. **Abstract**

The study aims to investigate the problem of the sharp increase in the incidence of strokes, which is becoming a pressing issue in the 21st century. Using a stroke disease dataset, the study performs data manipulations and analyses to answer several research questions and draw conclusions. The data cleaning process includes identifying and removing null values, modifying column names, and replacing extreme values. Exploratory data analysis is then used to answer research questions, such as the effect of smoking status on the probability of getting a stroke. The results suggest that former smokers have a greater chance of stroke than current smokers and non-smokers. The study concludes that individuals should control their BMI, age, smoking status, and hypertension levels to minimize the chances of stroke.

1. **Introduction**

The problem of a sharp jump in the incidence of strokes is becoming one of the most pressing troubles of the 21st century. According to the World Health Organization (WHO), stroke is the second leading cause of death globally, responsible for approximately 11% of total deaths. To study this problem in more detail and be able to predict an attack in patients, many statistical studies have already been conducted and continue to be undertaken.

This project aims to study the stroke\_disease dataset, perform some manipulations on the data for further analysis, answer several research questions and draw conclusions based on the information received.

To implement this project, our team used the Python programming language, namely the pandas libraries for working with data and sns for creating graphs.

1. **Data tidying and Cleaning steps**

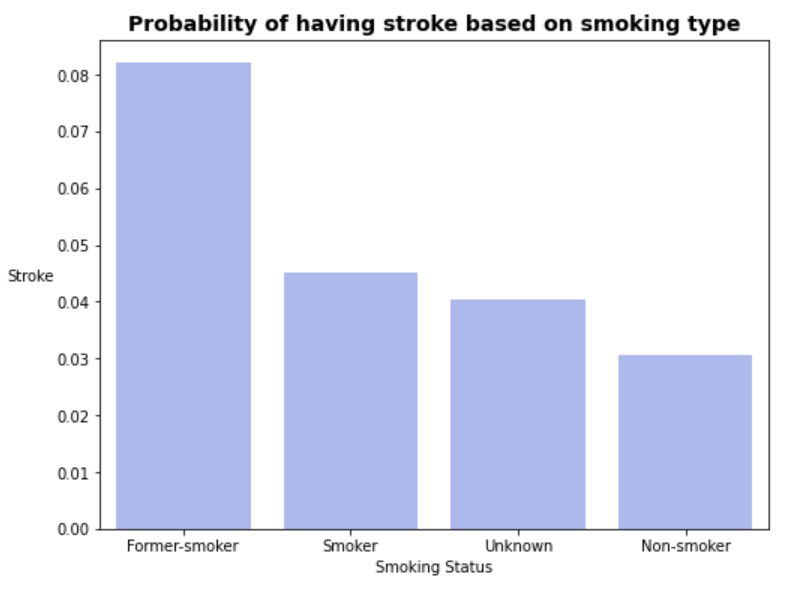
It needs to be cleaned in advance to analyze a data set successfully. Before doing this, thoroughly familiarize yourself with the available data; otherwise, it may lead to incorrect conclusions.

* There are 12 variables presented in the given dataset. 5 of them have object data type (gender, ever\_married, work\_type, Residence\_type, smoking status), and others have numerical float64 data type (id, age, hypertension, heart\_disease, avg\_glucose\_level, BMI and stroke).
* In order to make sure that the data set was imported correctly, we checked the first five and the last five records. There are 1000 records in the stroke\_disease dataset, which is enough to make some conclusions. We can notice that there are some null values in several columns (ever\_married, smoking\_status, etc.), data type (id unnecessarily has float data type), and string case problems ("FeMALe," "Self-employED").
* The next step was to drop the "id" column, which will not benefit further work with the data.
* Column names have been slightly modified for a more straightforward understanding of the meaning of the columns. Firstly, we titled all columns to preserve the integrity of the appearance of the data. Secondly, the dashes between words in column names have been replaced with spaces, which will facilitate further work with axis names in charts.
* Our next step was to learn more about the data set (for example, the number of null values). All the columns have more than a hundred missing values; each requires a personal approach to deal with NA values.
* The first column that our team fixed was Gender. String-case problems were solved, and missing values were filled with mode value (“Female”).
* Afterward, the work type problem was tidied and cleaned. Also, our team fixed some letter case problems and removed dashes between words to increase the dataset's clearness. Our team did the same thing with the Smoking Status column.
* The Hypertension, Heart Disease, and Stroke columns suffered similar problems: all of the columns have some extreme values (999999.0), which were replaced with null values. Also, instead of using zeros and ones as responses, our team preferred using “Yes” and “No” strings since it would make creating graphs easier during the EDA stage. Missing values were replaced with mode values because columns have object data types.
* In other columns, missing values were filled either with mode or mean, which depends on the variable's data type.
* The last but not least part of the data cleaning process was checking the consistency of variables. We assumed that recording errors could occur during the collection of information. For example, we changed the type of work of minors to "child" since child labor is illegal. The same applies to the smoking status of minors.

1. **Exploratory Data Analysis**

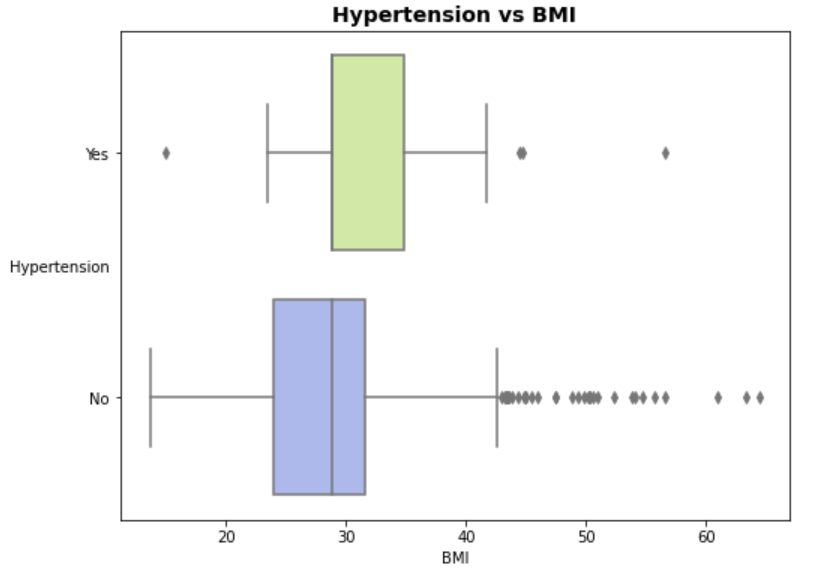
After thoroughly cleaning the data set, our team moved on to compiling research questions of interest to us, finding the answers using descriptive statistics and graphs, and interpreting the results.

* How is the probability of getting a stroke affected by a person's smoking status?

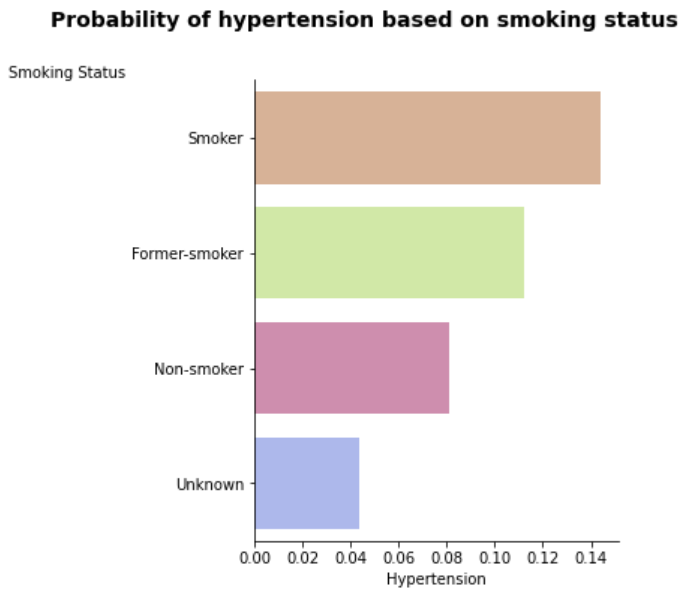


Based on the bar graph above, we can say that former smokers have a greater chance of stroke than those who currently smoke and do not smoke. Approximately 8% of former smokers in the available data set suffered a stroke. At the same time, about half as many smokers experienced such a thing. Unsurprisingly, only 3% of non-smokers have experienced a heart attack. Thus, the answer to the research question posed will be that the probability of getting a heart attack depends on whether a person smokes or not.

* Do the BMI value and smoking affect the appearance of human hypertension?

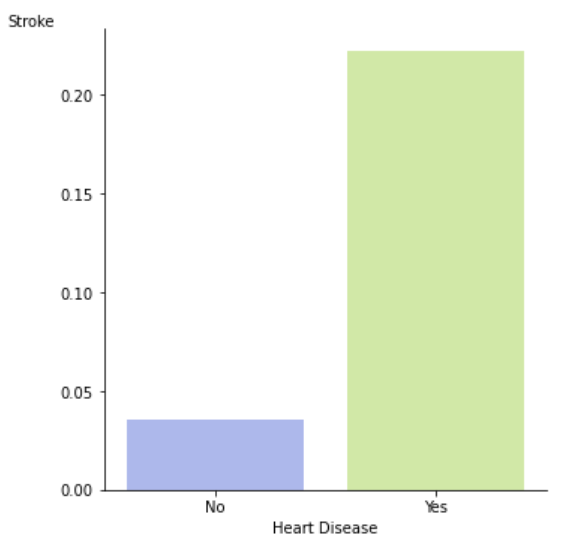


The diagram above shows the distribution of BMI among people with and without hypertension. Despite the fact that the median BMI value among both groups is approximately the same (about 30), we can still draw some conclusions. Firstly, the values of the first quartile and median in hypertensive patients are the same, which tells us that the distribution of BMI in people with hypertension is strongly skewed to the right. The third quartile of people with hypertension is much higher than those without it (35 and 32, respectively). The lower and upper "whiskers" among people with hypertension are also noticeably higher. Due to the above-described differences between the two boxes, it can be concluded that, on average, people with hypertension have a higher BMI than those who do not have hypertension. Answering the research question, there is a correlation between a person's BMI indicator and hypertension. We can also assume that an increase in BMI increases the likelihood of hypertension.



A bar chart was constructed to study the effect of smoking on the likelihood of hypertension in a person. As we can see, the proportion of people with hypertension among smokers is the largest, followed by the ratio of former smokers (0.14 and about 0.11, respectively). At the same time, only 8% of non-smokers have hypertension, significantly less than the above values. Thus, we can conclude that smoking affects the appearance of hypertension in a person, which is not surprising since smoking negatively affects the work of the heart.

* Is there a connection between a person having heart disease and a stroke?

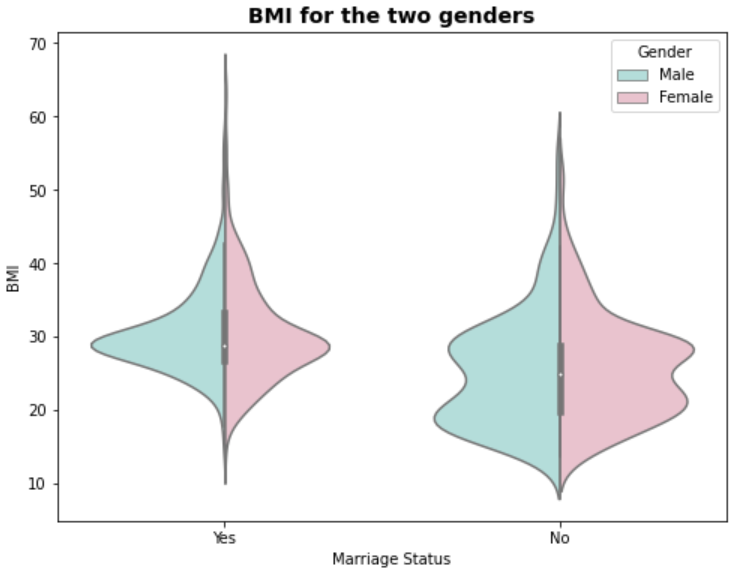


In order to answer the above research question, our team constructed a bar chart. As you can see, almost 20% of people who had a stroke had or have heart disease, while just 2.5% of people who did not have a stroke had heart problems. Thus, answering a research question, there is an apparent link between a stroke and heart disease.

* How do the average glucose level and BMI values between people differ depending on their marriage status?

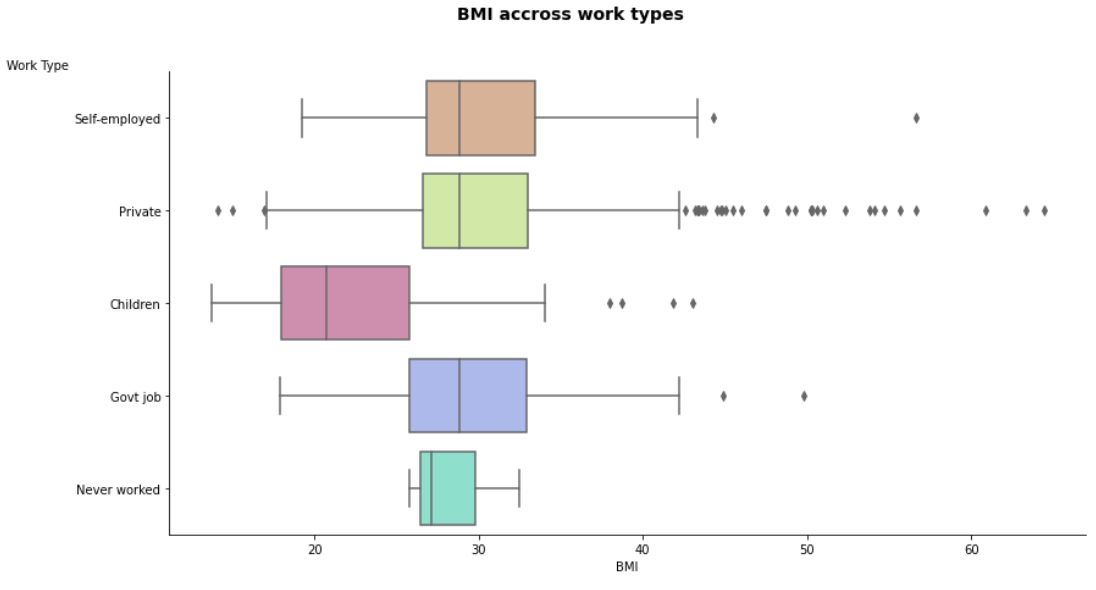


As we can see from the violin plot, the average glucose level in married people is noticeably higher than in unmarried people (100 and 90, respectively). You can also notice that the interquartile range in married people is slightly more extensive, indicating a somewhat broader distribution than unmarried people's. It is also worth mentioning a large number of sharply different values in both groups, which may indicate the presence of people with diabetes in the data set. The distribution of the average glucose level of married people is bimodal (100 and 200), while unmarried people have unimodal distribution. Thus, the answer to the research question will be that the status of marriage affects the growth of average glucose levels in people. We can assume that married people generally follow a healthy diet less.



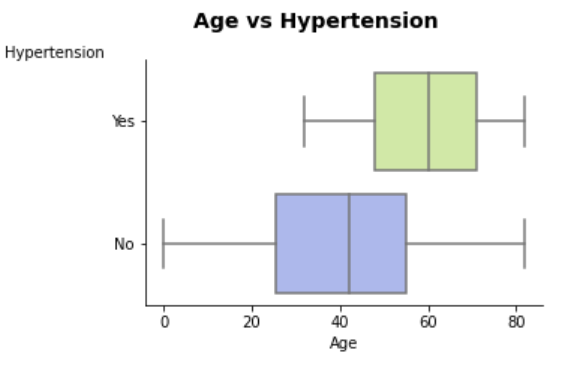
As seen from the violin plot above, the median BMI of married people is slightly higher than that of those who are not married (30 and 25, respectively). Also, the "whiskers" and quartiles of the values of unmarried people are greater than those of married people. The BMI values of married people are skewed to the right, while those of unmarried people are skewed to the left. The BMI value of married people has unimodal distribution, while it is bimodal for those who are not married. Based on information from the boxplot, married people often have higher BMI levels than single people, which age or lifestyle changes after marriage can explain.

* How does people's BMI differ depending on their work type?

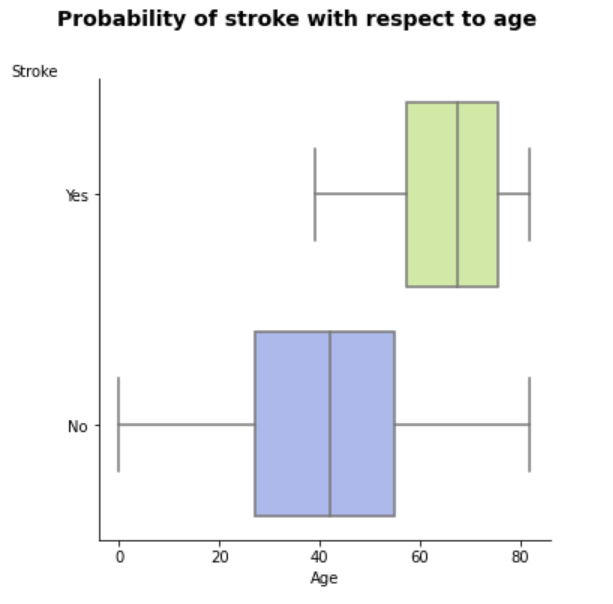


Based on the boxplot above, some interesting observations can be made. Firstly, the median BMI values for self-employed, private workers, and government workers are the same (about 29). At the same time, children and people who do not work have significantly less BMI (20 and 27, respectively). This may be because children and teenagers are primarily thinner than adult working people. It is also worth noting that all values are shifted to the right, which the current trend towards becoming overweight among people can explain. Answering the question, it would be accurate to say that the value of BMI does not depend on a person's work type, only if we are not talking about children and the unemployed.

* Is the appearance of hypertension and stroke in a person related to his age?

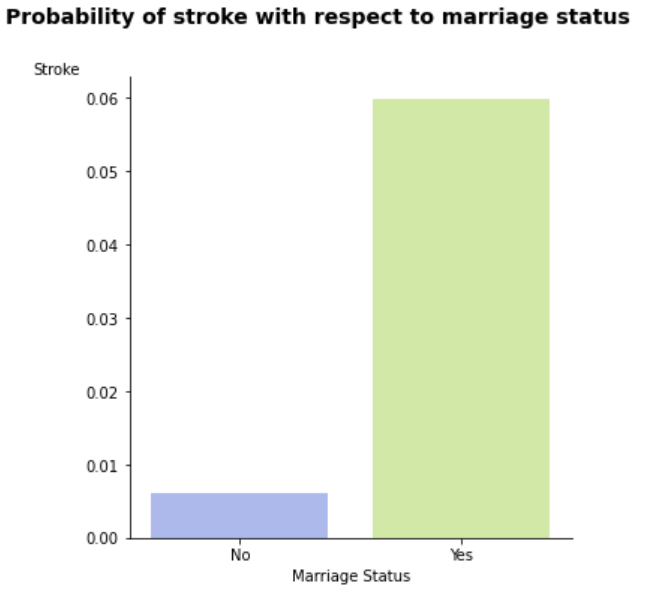


A box plot was built to study hypertension's dependence on a person's age. As we can see, the median age of people with hypertension is 60 years, while the median age of people without hypertension is slightly above 40 years. Both boxes are distributed symmetrically, but people without hypertension have a lower IQR. The answer to the research question is that hypertension in humans develops with age, which is not surprising because, with age, people's heart problems only get worse.



The situation is almost the same with the age of people who survived the blow. The median age of people who have experienced a heart attack is somewhere between 60 and 70 years, while the median age of people who have not had a stroke is just over forty. Thus, the probability of a heart attack depends on the person's age.

* Does the person's marital status affect the probability of a stroke?



From the information from the bar chart, it follows that 6% of married people from our dataset experienced a heart attack, while 12 times fewer unmarried people experienced the same thing. At the same time, it would be wrong to say that it is a very fact of marriage that increases the likelihood of a heart attack. This may be because married people are often no longer at a young age, and as we know from the previous question, the probability of a heart attack directly depends on a person's age.

1. **Conclusion**

Previous studies (1) show that hypertension is the single most important risk factor for stroke. In the research, we analyzed the relationship between stroke and hypertension with various parameters with the aim of finding factors that could be affecting stroke chance.

Firstly, We found that smoking status is significantly related to stroke. That being a former smoker statistically increases the likelihood of suffering from a stroke. In addition, we found that smoking and age increase the chance of hypertension. Smokers and elders have been found to have a higher chance of hypertensiın compared to non-smokers and younger people.

Secondly, we investigated the body mass index of individuals. We found that body mass index is related to individuals' marriage and work type status. Analysis showed that married individuals have statistically higher BMI compared to singles. On the other hand, our analysis found no statistically significant difference in BMI levels between different work types. Moreover, we found that the BMI of individuals increases the risk of hypertension.

Therefore, individuals who smoke and/or have hypertension are in danger of suffering a stroke attack. Since we found that BMI and Age are related to hypertension. BMI and Age indirectly increase the chances of stroke. Thus, individuals should control their BMI, age, smoking status, and hypertension levels to minimize the chances of stroke.

1. **References**

* Dubow, J., Fink, M.E. Impact of Hypertension on Stroke. Curr Atheroscler Rep 13, 298–305 (2011). https://doi.org/10.1007/s11883-011-0187-y