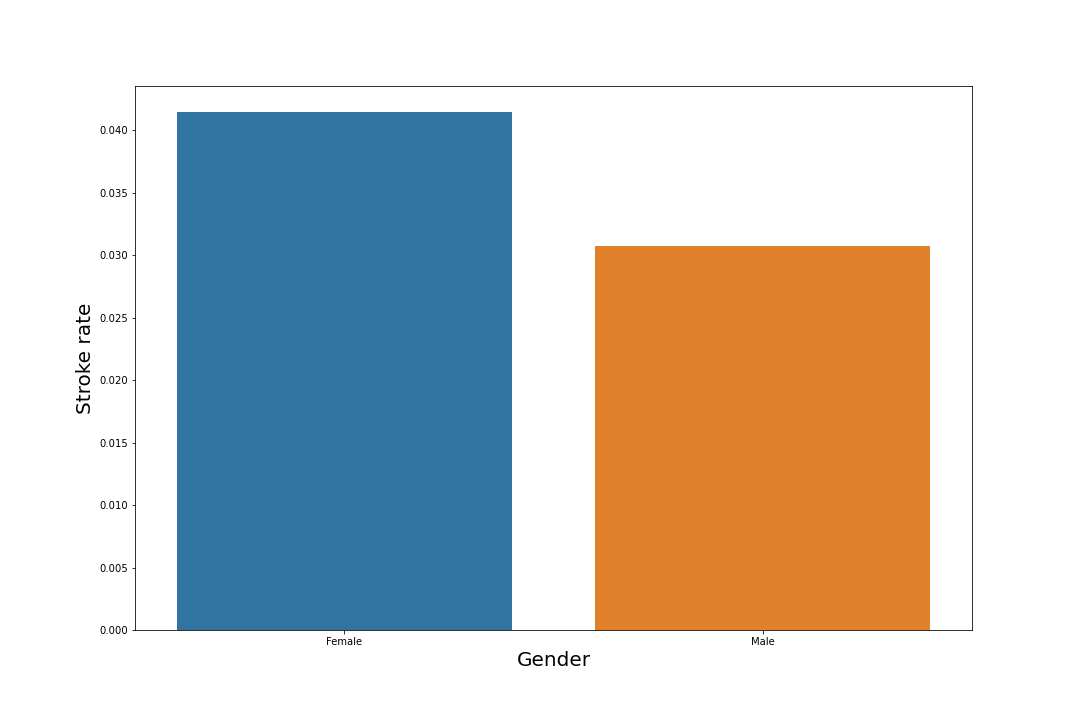
Stroke dataset analysis report

According to WHO, stroke is the second leading cause of death globally, responsible for approximately 11% of total death. Previous work has been done to show that heart disease[[1]](#footnote-1), Tobacco[[2]](#footnote-2), work type[[3]](#footnote-3) could be related to stroke. This data science project uses a dataset collected from Kaggle[[4]](#footnote-4) which includes patients’ gender, age, various diseases, and smoking status etc. to predict whether a patient is likely to get stroke or not.

As stated above, the dataset is obtained from Kaggle website with the author specified that the data source is confidential and can only be used for educational purposes. The dataset contains 5110 rows, each row provides relevant information about a patient including ID, gender, age, hypertension, heart disease, marriage status, work type, residence type, average glucose level in blood, body mass index, smoking status and whether caught stroke or not. Remarkably, there are 201 missing BMI values. It is not a rigorous manner to fill those missing values as their true values are nowhere to be found and considering the small scale of the missing values, the rows with missing BMI are removed.

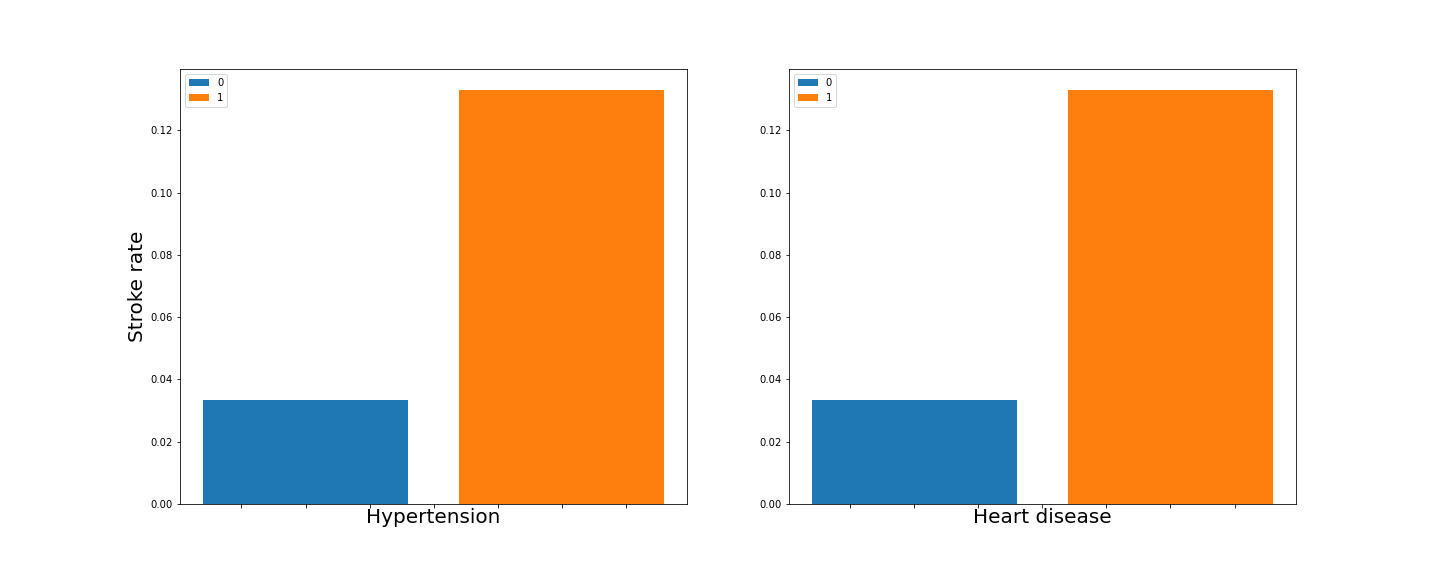
When doing EDA, stroke rate did not show a notable difference in genders. However, the stroke cases were concentrated on age above 50, and rises significantly above 70 years old.

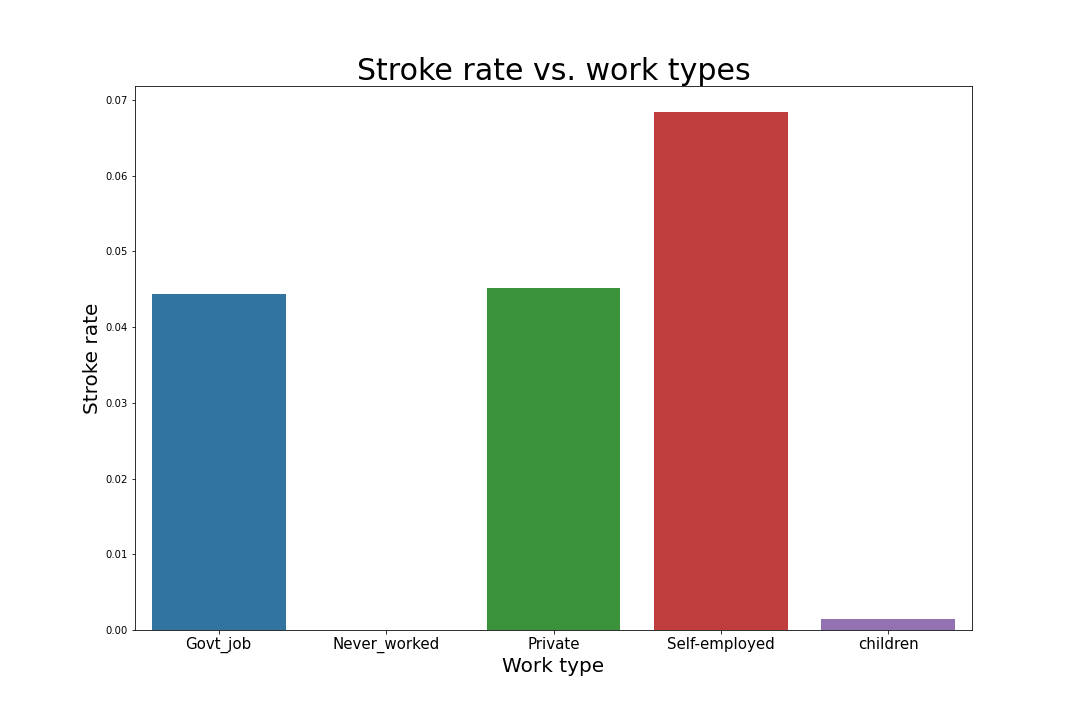
图表, 直方图

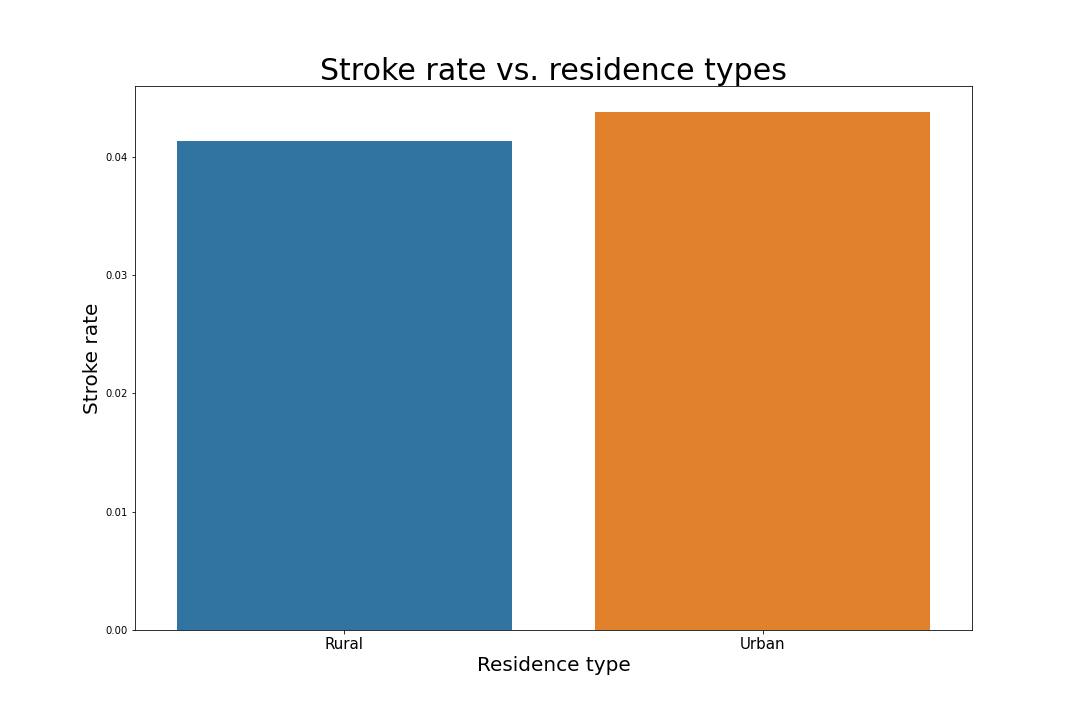
描述已自动生成

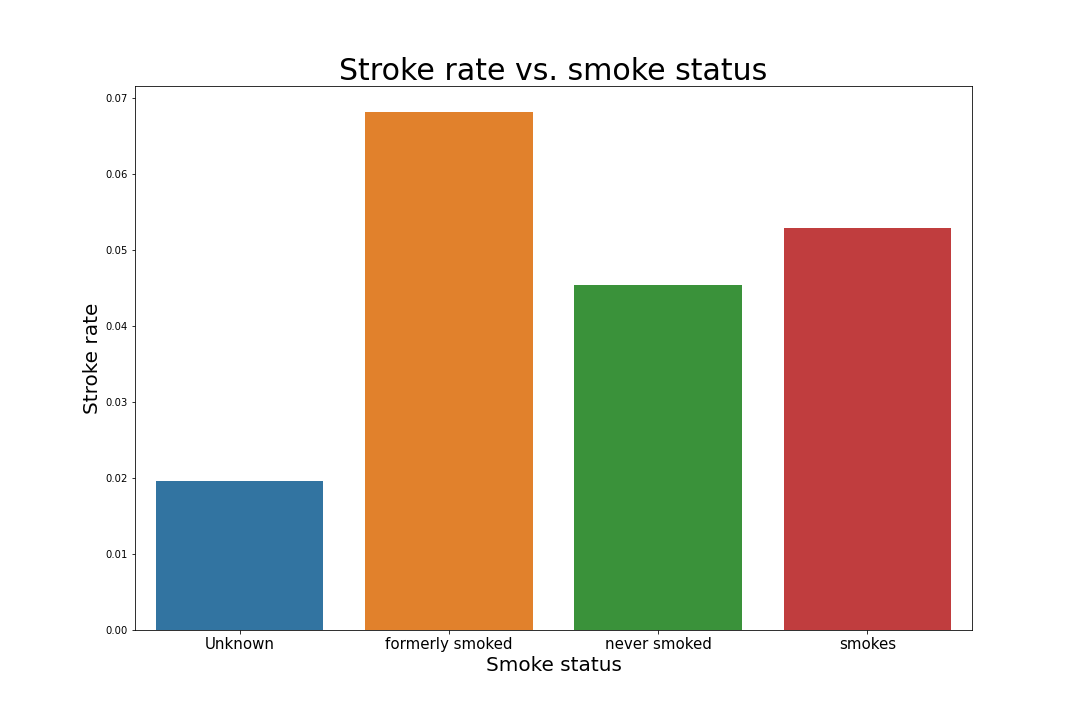
图表, 直方图

描述已自动生成

Unlike gender, hypertension and heart disease showed a heavy impact on stroke rate. Patients who have hypertension or heart disease are over three-times more likely to catch stroke than those who do not. 

Although we mentioned previously that stroke have work-related causes, the dataset only specified five sketchy distinct work types. Thus unsurprisingly, they did not show a prominent difference except for those who have never worked. A more accurate relationship between work type and stroke risk shall be expected if we take risk factors like exposure to long working hours and workplace exposure to air pollution, asthmagens, carcinogens, ergonomic risk factors, and noise into consideration.

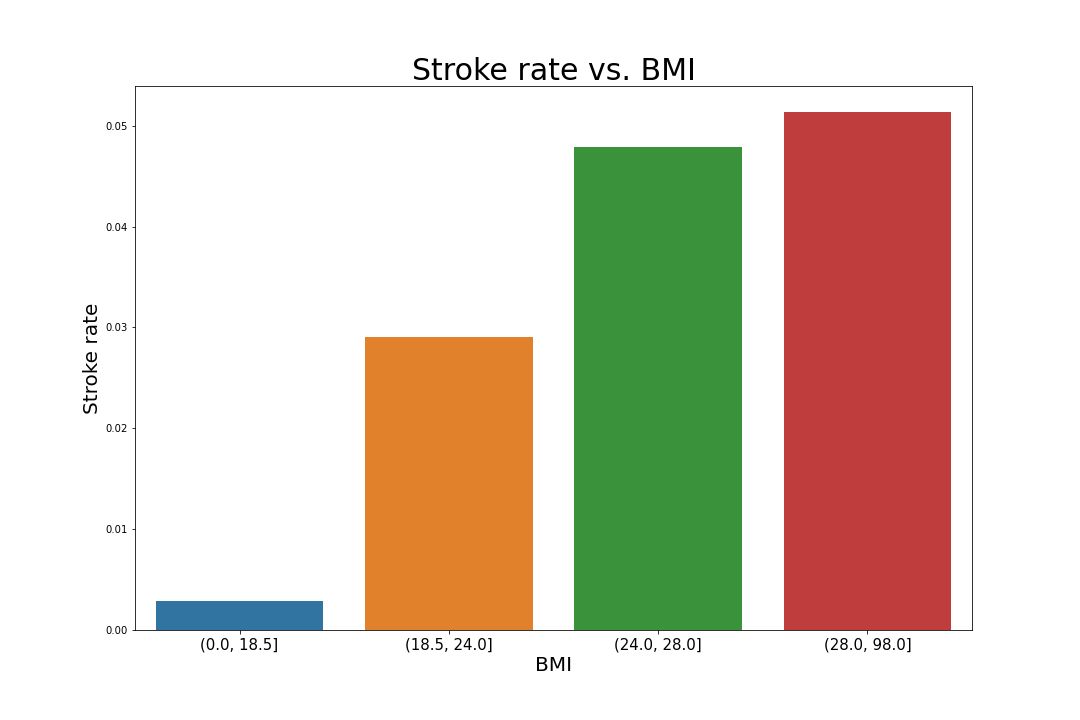
Unanticipatedly the difference in stroke rate between urban and rural residence is neglectable. 

The smoke status should play the most important role in causing stroke for the greatest causes of deaths were chronic obstructive pulmonary disease according to WHO[[5]](#footnote-5). Nevertheless, there is only a small difference between smoker and non-smoker. What could be miss leading is that those who have formerly smoked were shown to have higher chance to catch stroke than those who keeps smoking. The small scale of the dataset could be one of the seasons for this. According to another study by NIH, smokers are nearly 2.5 times more likely to have a stroke than those who never smoked, while former smokers show a similarly lower risk as never smokers[[6]](#footnote-6).

The average glucose level distribution in the stroke population took roughly the same shape as that in the total population. Whereas the BMI showed a strong link with stroke incidence. BMI was cut into four groups, representing underweight, normal weight, overweight and obesity[[7]](#footnote-7). Stroke incidence showed clear increasing rate from underweight to obesity.

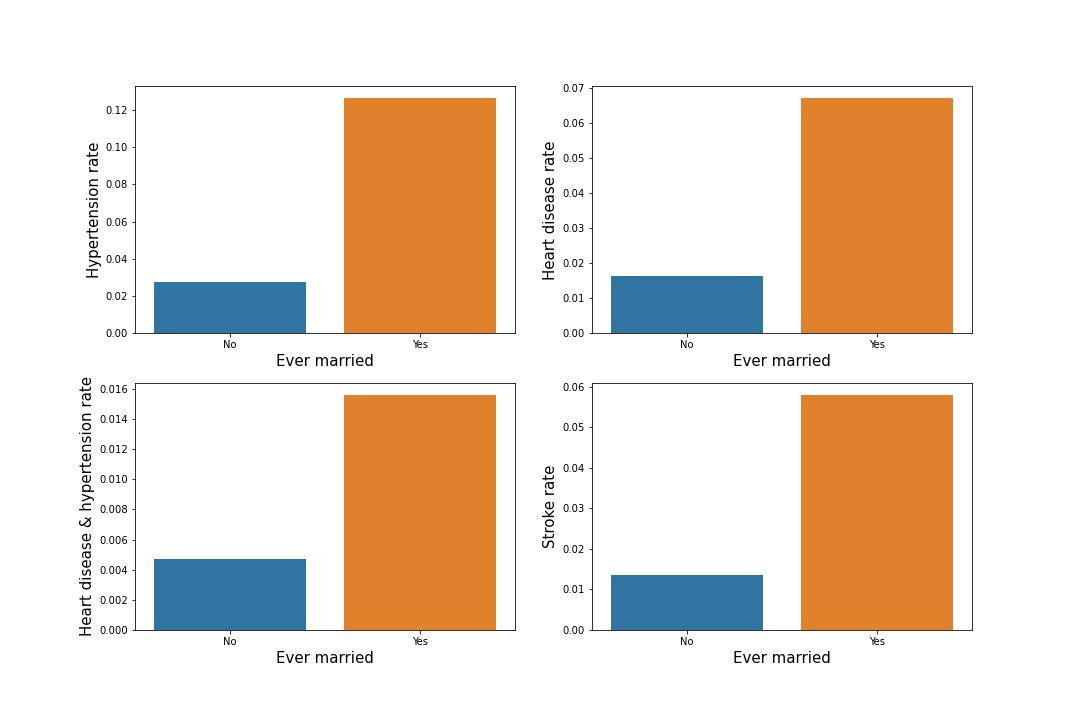
图表, 直方图

描述已自动生成



Last but not least, features were selected from the dataset. Every column except “id” and “stroke” were taken to be features. After standardizing and one-hot encoding, together with the target “stroke”, the data was passed into Random Forest model and Logistic regression model. Both models gave over 95% accuracy in predicting stroke.

You have probably noticed by now, the column “ever\_married” was neglected. That is because of the severity of marriage that made it the grand finale. People who have married are over three times more likely to catch stroke, hypertension and heart disease than those who have not married.



Deficiencies of this study are obvious. The EDA only shows what lies in the dataset but did not say anything about their causal relationships. There are some data seem unrelated to stroke by EDA. Therefore, hypothesis testing could be carried out to examine whether those columns are related to stroke or not. Without those noise, the models could be more well fitted. Additionally, principal component analysis might be of help given the dataset is 12-dimentional.

1. World Health Organization. (n.d.). *Active ageing: Moving hearts for health*. World Health Organization. Retrieved November 16, 2021, from https://www.who.int/news/item/27-09-2002-active-ageing-moving-hearts-for-health. [↑](#footnote-ref-1)
2. World Health Organization. (n.d.). *Tobacco knowledge summaries: Tobacco and stroke*. World Health Organization. Retrieved November 16, 2021, from https://www.who.int/publications/i/item/WHO-NMH-PND-CIC-TKS-16.1. [↑](#footnote-ref-2)
3. World Health Organization. (n.d.). *WHO/ilo: Almost 2 million people die from work-related causes each year*. World Health Organization. Retrieved November 16, 2021, from https://www.who.int/news/item/16-09-2021-who-ilo-almost-2-million-people-die-from-work-related-causes-each-year. [↑](#footnote-ref-3)
4. https://www.kaggle.com/fedesoriano/stroke-prediction-dataset [↑](#footnote-ref-4)
5. World Health Organization. (n.d.). WHO/ilo: Almost 2 million people die from work-related causes each year. World Health Organization. Retrieved November 16, 2021, from https://www.who.int/news/item/16-09-2021-who-ilo-almost-2-million-people-die-from-work-related-causes-each-year. [↑](#footnote-ref-5)
6. U.S. Department of Health and Human Services. (2020, June 10). NIH study links cigarette smoking to higher stroke risk in African Americans. National Institutes of Health. Retrieved November 18, 2021, from https://www.nih.gov/news-events/news-releases/nih-study-links-cigarette-smoking-higher-stroke-risk-african-americans. [↑](#footnote-ref-6)
7. U.S. Department of Health and Human Services. (n.d.). Calculate your BMI - standard BMI Calculator. National Heart Lung and Blood Institute. Retrieved November 19, 2021, from https://www.nhlbi.nih.gov/health/educational/lose\_wt/BMI/bmicalc.htm. [↑](#footnote-ref-7)