BUAN 6341

**US Stroke Prediction: What is the leading factor that causes stroke to develop?**

Project 4: Xin Yao

horizontal line

# 

# Contents

1. Abstract …………………………………………Page 3
   1. Background …………………………………………Page 3
   2. Goal …………………………………………Page 3
2. Data Description …………………………………………Page 4
3. Exploratory Data Analysis …………………………………………Page 6
   1. Data Introduction …………………………………………Page 6
4. Data Cleaning and Visualization …………………………………Page 9
5. Data Processing …………………………………………Page 12
6. Conclusion …………………………………………Page 14

# 

# Abstract

**Background:** According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. In America, Stroke kills about 140,000 Americans each year and in 2021, 1 in 6 deaths from cardiovascular disease was due to stroke based on the reports from the Centers for Disease Control and prevention (CDCP). To make it more precise, someone in the United States has a stroke every 40 seconds and someone dies of stroke every 4 minutes (UTMBHealth). The top 5 leading causes of a stroke are: high blood pressure, high cholesterol, smoking, obesity, and diabetes.

**Goal:** Our goal is to determine the leading factor that has the most impact on causing a stroke to happen; To determine whether there is a relationship between any of the factors that might contribute to the development of a stroke, and to identify which factor is the leading cause, we analyzed the dataset provided hoping this will help people to maintain a healthy habit and know how much risk they are facing when practicing bad habits.

# Data Description

Data records were provided by the professor. This dataset contains a total of 5110 observations, four were categorical, four binary and four were numerical. Here is a detailed list of all the variables the dataset contains:

|  |  |
| --- | --- |
| **Variables** | **Description** |
| Id | This has numerical values used to represent each unique user who participated in the research |
| Gender | This has three categories of gender: male, female, other |
| Age | This indicates the age that the person experienced a stroke |
| Hypertension | Hypertension is another word for high blood pressure which is often considered as one of the leading causes of stroke. The dataset uses “1” to indicate a Yes and “0” for a No on whether the participant has hypertension or not. |
| Heart Disease | Similar to Hypertension, heart disease is also a leading factor that causes stroke. This also uses “1” to indicate a Yes and “0” for a No on whether the participant has heart disease or not |
| Ever Married | This is to survey participant’s marital status using a “Yes” if they ever married and a “No” for never married. |
| Work Type | This surveyed the type of jobs each participant has. The categories are “Private”, “Self-Employed”, “Government Job”, “Children” and “Never Worked” |
| Residence Type | The location that the participant lives: “Urban” or “Rural”. |
| Average Glucose Level | Their average glucose level. |
| BMI | Stands for Body mass Index, the higher the number is, the more overweight you are. |
| Smoking Status | A categorical variable that indicates whether the participant’s smoking history. “Formerly smoked”, “never smoked”, “Smokes” and “Unknown”. |
| Stroke | This indicates whether the participant had a stroke or not. “1” indicates a Yes and “o” means No. |

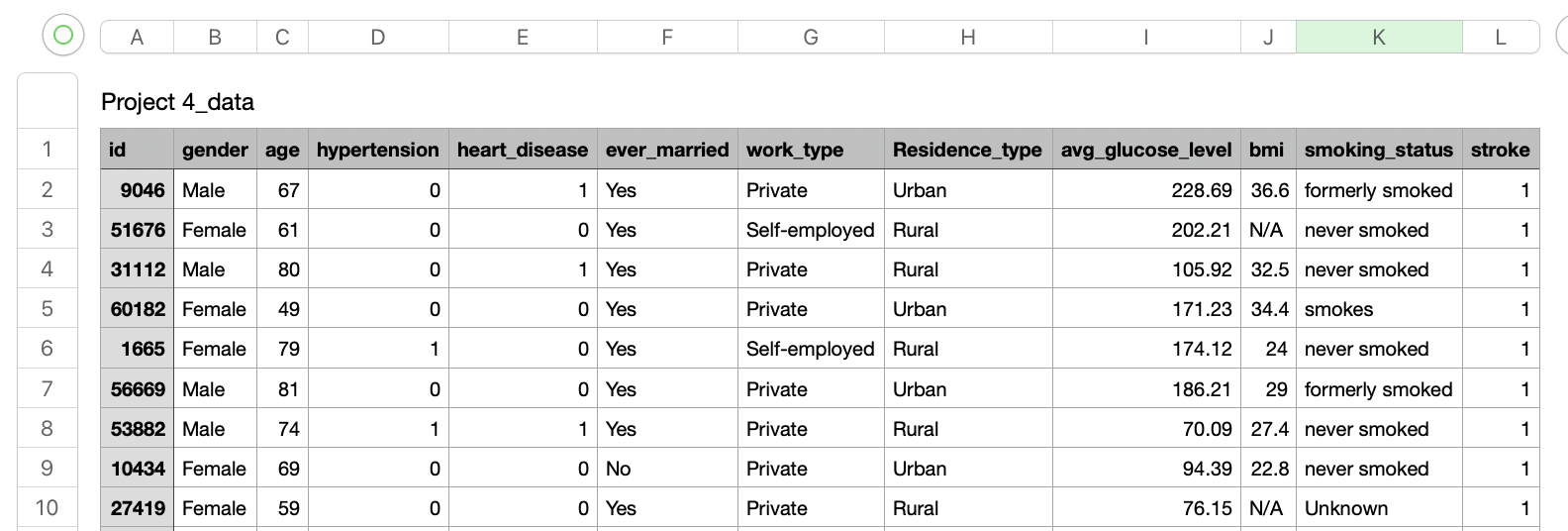
# 

# 

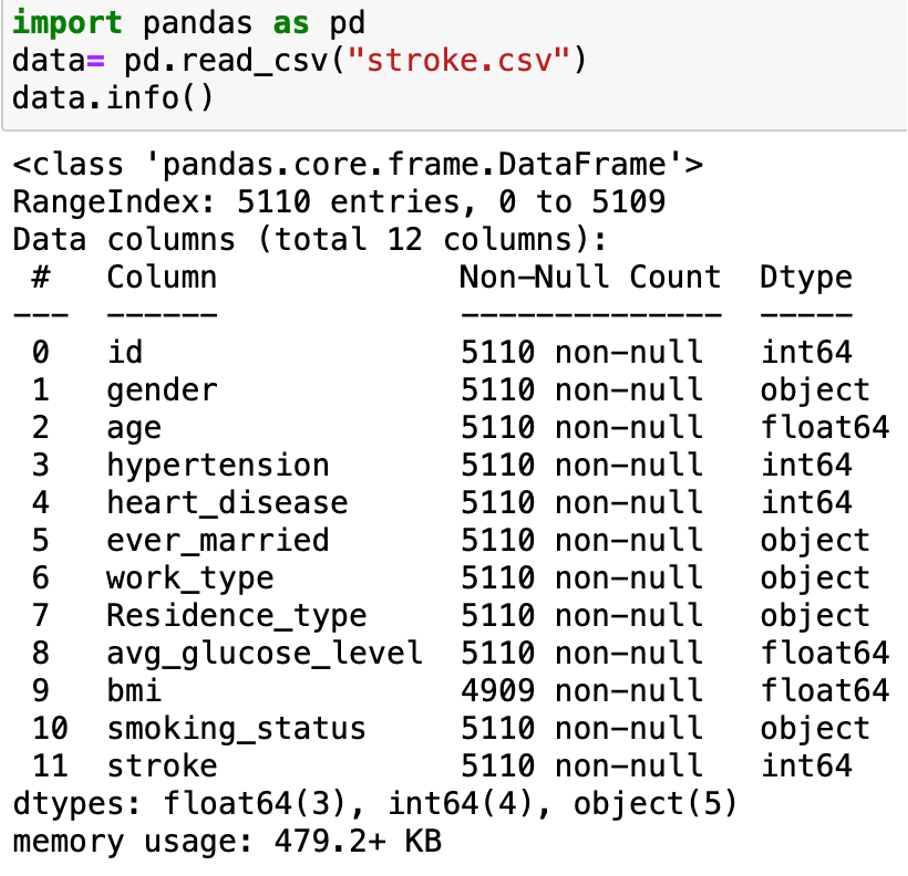
# Exploratory Data Analysis

**Data Introduction:**

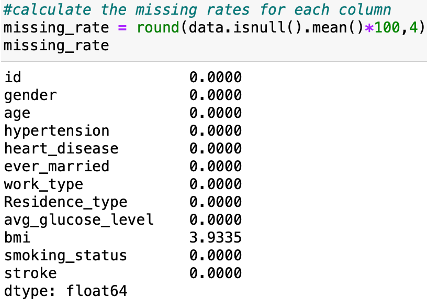
1. Sample of the data:



1. Data types of the variables:

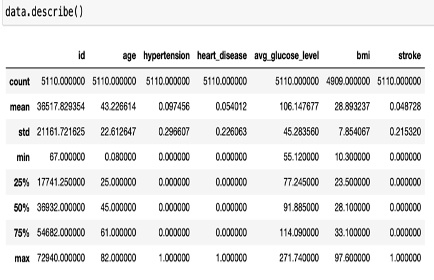


1. Checking for null values, missing values, and duplicated values:





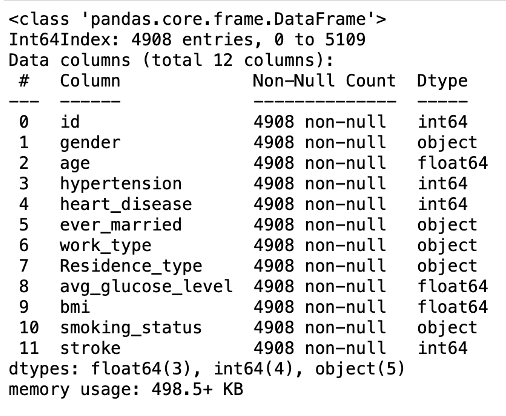
1. Checking the statistic:



Based on the result of the statistics, the ranges among the majority of the numerical variables show a great spread between the lowest and highest values. Variables such as “Hypertension”, “Heart Disease” and “Stroke” are binary which means “1” indicates a Yes and “0” indicates a No. Therefore, from the data statistics, the average age is about 43 years old, average glucose level is 106.18 and average BMI is 28.89.

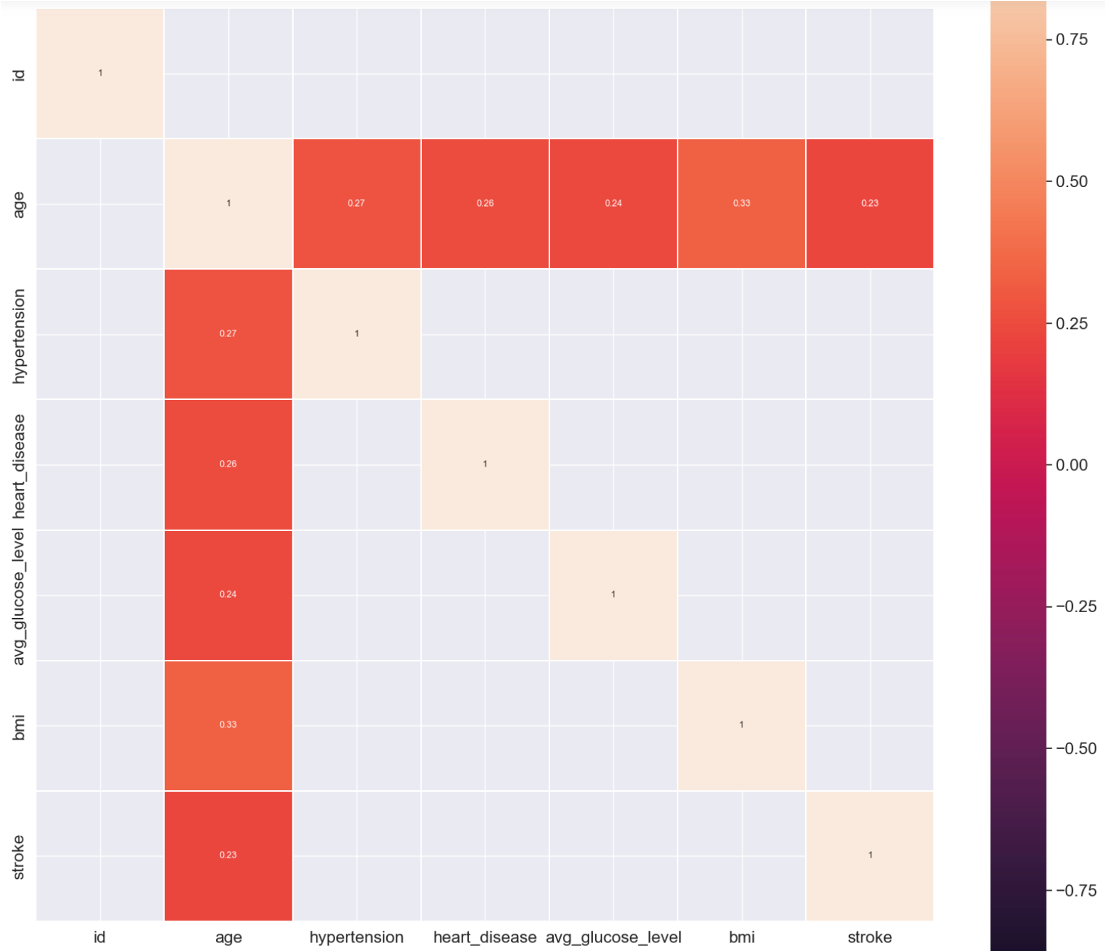
# Data Cleaning and Visualization

1. Getting rid of the Null values:

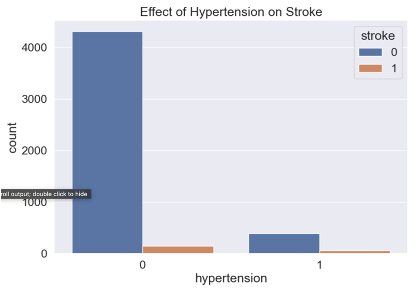
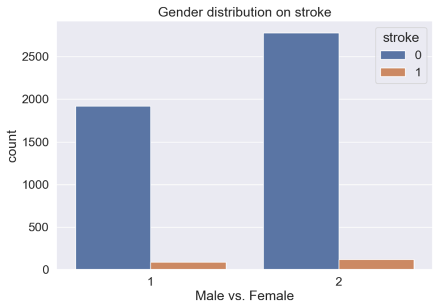


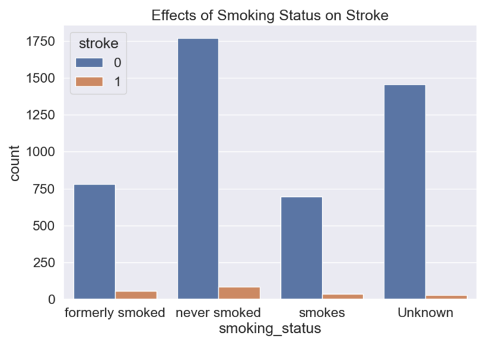
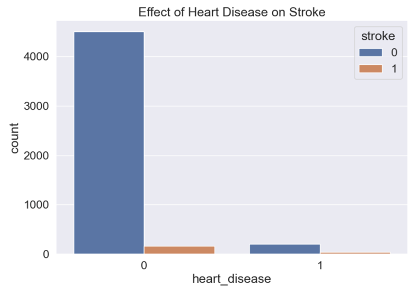
According to the first report generated, 4% of the values were missing for the attribute “BMI”. Therefore, we deleted the null values and an outlier for gender that is labeled as “Other”. Which we are now resulting in 4908 entities.

1. Finding the correlations between each variable:

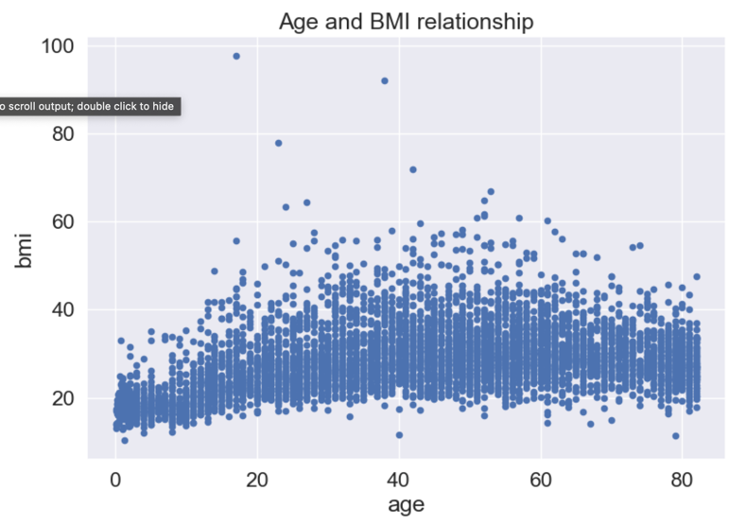


Heatmap, or in other words, “correlation matrix” is a graphical representation of data where values are depicted by color. In this case, the higher the correlation is, the darker the color is. Based on the heatmap, those numerical variables are more related to people’s age than any other factors. With all the variables taken into consideration, BMI and age are seen as more correlated with a correlation coefficient of 0.33.

1. 



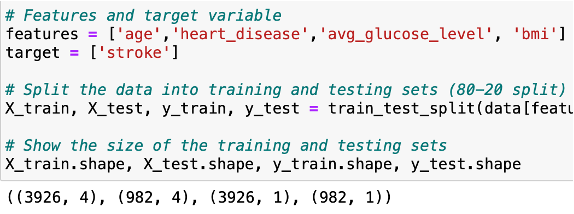
There are few things we’ve noticed based on these graphs. First, there is a huge difference between the number of people who had a stroke and the number of people who didn’t have a stroke. This could cause the training dataset and testing sets to be highly skewed with these imbalanced values. Second, the number of people who had a stork while experiencing other factors are proportionally related to the survey population. This means that when the number of people in a certain category is higher than the other, it proportionally will have more people experiencing a stroke. Lastly, having a stroke or not doesn't seem to depend on one factor, rather it’s possibly a combination of multiple factors that led to the development of a stroke to happen.

1.  

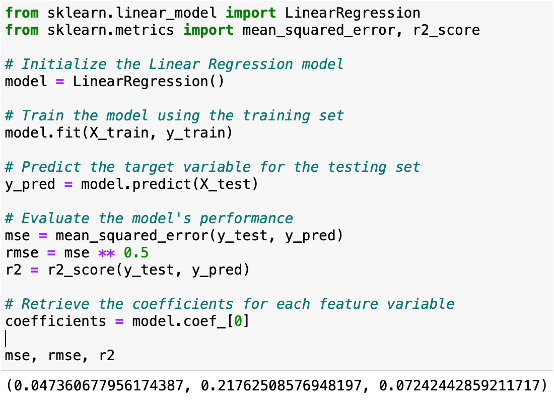
As we can see from the above graphs, there is not a clear trend in relation between age and BMI, and age and glucose level. This could be because the data contains more values on who never had a stroke than who had a stroke which skewed the data, or also because there is not a recognizable trend that could be used to predict who will develop a stroke.

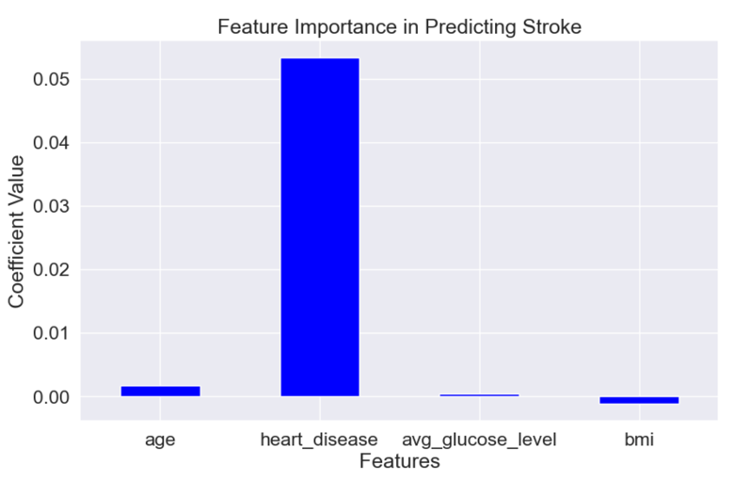
# Data Processing:

1. To represent the unseen data and avoid data leakage, we split the data into training and testing datasets based on an eight/twenty ratio. Resulted in 3926 in training and 982 in testing.

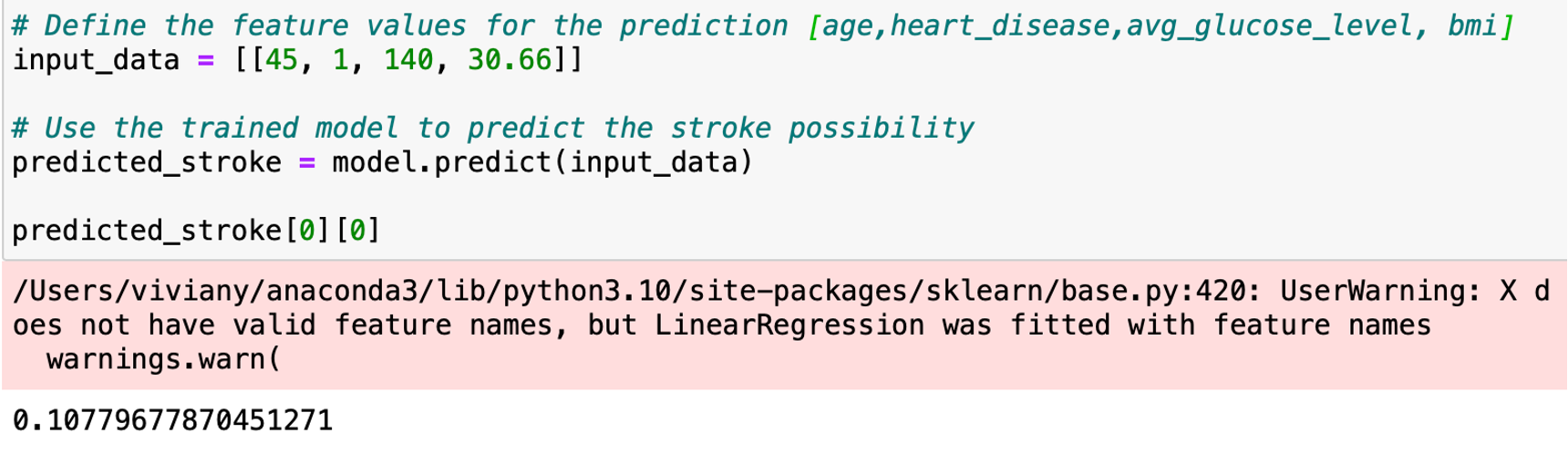


1. Secondly, we then use the linear regression model to calculate the mean squared error, root mean squared error and r^2. Based on the result, mse= 0.047, rmse= 0.218, r^2 = 0.072. A low mse means that the data points were close to the regression line and less error, which makes the prediction more precise. A low r2 means that the independent variables, in this case, “age”, “heart disease”, “glucose level”, and “bmi” don't have too much variation in them.



1. Third, we generated a graph to represent the importance in predicting whether a person will get a stroke or not. Based on the rank the most important feature is heat disease, then comes their age, followed by that is BMI, and the last one is the average glucose level. 
2. These coefficients represent the general trend of each variable. For 1 year increase in age, the likelihood of having a stroke will increase by 0.0017. The possibility of having a stroke if you have heart diseases will increase by 0.053. For 1 unit increase in glucose level, the likelihood of having a stroke will increase by 0.00045. For 1 unit increase in BMI, the likelihood of having a stroke will decrease by 0.00113

****

1. When predicting the accuracy score by randomly assigning each feature a number, the accuracy rate of predicting whether the person will have a stroke is 0.946. 

# Conclusion

To conclude, due to skewed dataset that caused by the number of participant who never had a stroke is larger than the number of participants who had a stork, the overall trend and distribution doesn’t show a significance difference in identifying which factor affects the participants the most in terms of developing a stroke. However, according to the National Institute of Child Health and Human Development, about 795,000 people in the United States have strokes and about 610,000 of these are first strokes. Compare that number to the entire US population, it is 0.24% of the population who will have a stroke. Therefore, when predicting the number of strokes that will happen, a lower percentage is relatively reasonable.