Stroke Data Set

Nonparametric tests, correlations analysis and regression.

- **Group D**
- **Aya Abouelela**
- **Amel Khirreddine**
- **Kwaku Asamoah Gyimah**
- **Arlizze Faye R. Ongchua**
- **Supervisor: Prof. Elnaz Gholipiour**

Goal: To use non-parametric tests to verify our null and alternative hypotheses as well as run correlation and regression tests to see some patterns across our dataset.

	Unnamed: 0	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	0	Male	58.0	1.0	0.0	Yes	Private	Urban	87.96	39.2	never smoked	0.0
1	1	Female	70.0	0.0	0.0	Yes	Private	Rural	69.04	35.9	formerly smoked	0.0
2	2	Female	52.0	0.0	0.0	Yes	Private	Urban	77.59	17.7	formerly smoked	0.0
3	3	Female	75.0	0.0	1.0	Yes	Self-employed	Rural	243.53	27.0	never smoked	0.0
4	4	Female	32.0	0.0	0.0	Yes	Private	Rural	77.67	32.3	smokes	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...
28911	29060	Female	10.0	0.0	0.0	No	children	Urban	58.64	20.4	never smoked	0.0
28912	29061	Female	56.0	0.0	0.0	Yes	Govt_job	Urban	213.61	55.4	formerly smoked	0.0
28913	29062	Female	82.0	1.0	0.0	Yes	Private	Urban	91.94	28.9	formerly smoked	0.0
28914	29063	Male	40.0	0.0	0.0	Yes	Private	Urban	99.16	33.2	never smoked	0.0
28915	29064	Female	82.0	0.0	0.0	Yes	Private	Urban	79.48	20.6	never smoked	0.0

Non-parametric tests

- Sample of people who suffered a stroke, N=450
- Sample of people who did and didn't suffer a stroke, N= 900
- Dependant variables BMI and Average Glucose level
- Independent variables= Work-type, Heart disease, Gender,
- Ever-married, residence type

We would like to know the effect of work-type on the average glucose levels for people who suffered a stroke

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]
1	The distribution of avg_glucose_level is the same across categories of work_type.	Independent-Samples Kruskal-Wallis Test	,167

Hypothesis Test Summary

	Decision
1	Retain the null hypothesis.

a. The significance level is ,050.

b. Asymptotic significance is displayed.

Independent-Samples Kruskal-Wallis Test Summary

Total N	450
Test Statistic	3,583[a]
Degree Of Freedom	2
Asymptotic Sig.(2-sided test)	,167

a. The test statistic is adjusted for ties.

Pairwise Comparisons of work_type

Sample 1-Sample 2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.[a]
Self-employed-Private	,345	13,465	,026	,980	1,000
Self-employed-Govt_job	34,837	20,053	1,737	,082	,247
Private-Govt_job	34,492	19,036	1,812	,070	,210

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.

Asymptotic significances (2-sided tests) are displayed. The significance level is ,050.

a. Significance values have been adjusted by the Bonferroni correction for multiple

Conclusion= Work type doesn't seem to have an effect on the average glucose level among the people who have suffered a stroke

We would like to know the effect of having a heart disease on the average glucose levels for people who suffered a stroke

Ranks

	heart_disease	N	Mean Rank	Sum of Ranks
avg_glucose_level	0	364	215,98	78617,00
	1	86	265,79	22858,00
	Total	450		

Hypothesis Test Summary

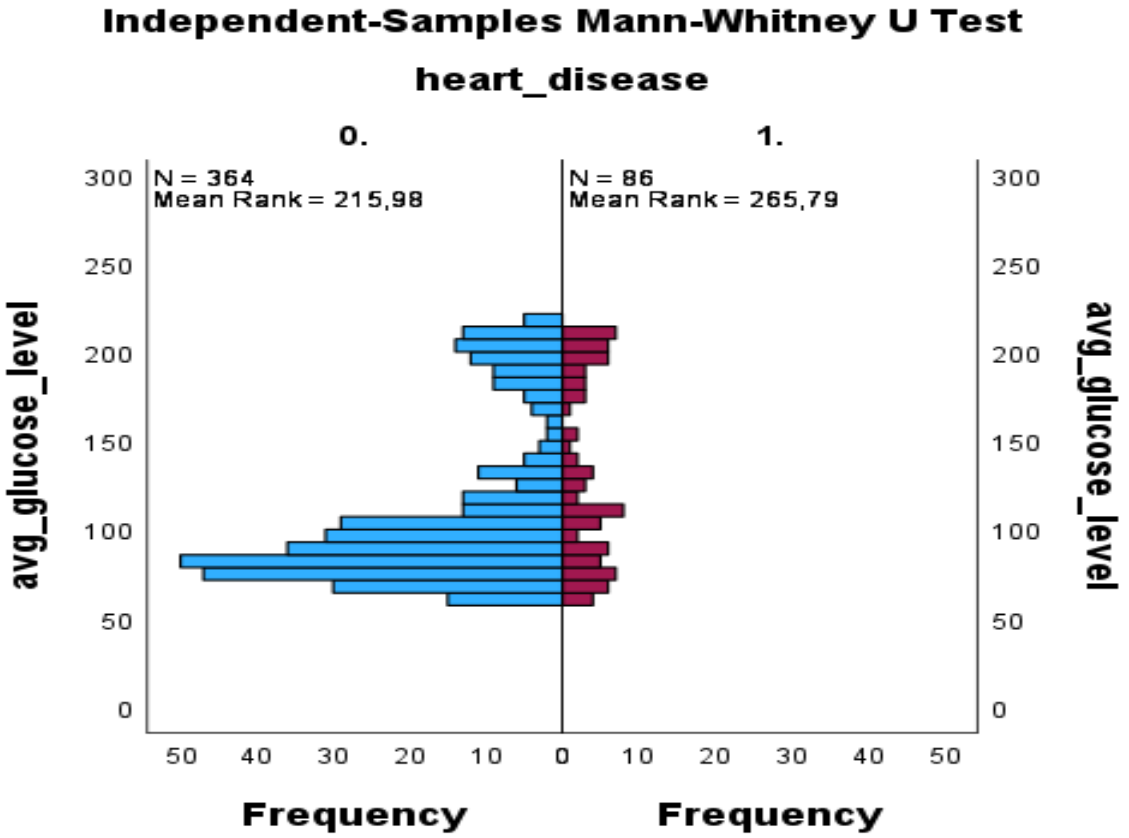
	Null Hypothesis	Test	Sig.[a,b]
1	The distribution of avg_glucose_level is the same across categories of heart_disease.	Independent-Samples Mann-Whitney U Test	,001

Hypothesis Test Summary

	Decision
1	Reject the null hypothesis.

a. The significance level is ,050.

b. Asymptotic significance is displayed.



Conclusion= For people who have suffered a stroke and have a heart disease, they seem to have high average glucose as well.

For people who suffered a stroke, Is BMI the same for male and female ?

Hypothesis Test Summary

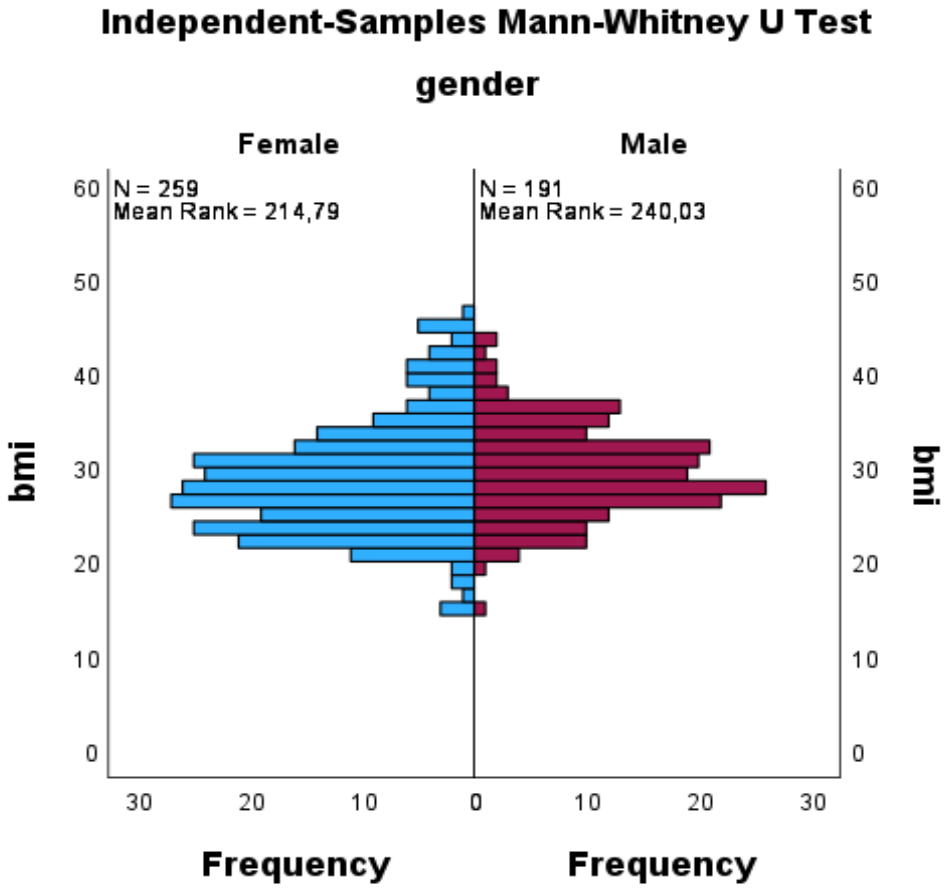
	Null Hypothesis	Test	Sig.[a,b]
1	The distribution of bmi is the same across categories of gender.	Independent-Samples Mann-Whitney U Test	,042

Hypothesis Test Summary

	Decision
1	Reject the null hypothesis.

a. The significance level is ,050.

b. Asymptotic significance is displayed.



Conclusion=The BMI of males seems higher than the one for females for people who have suffered a stroke

For people who suffered a stroke, does marital status affect the BMI levels

Hypothesis Test Summary

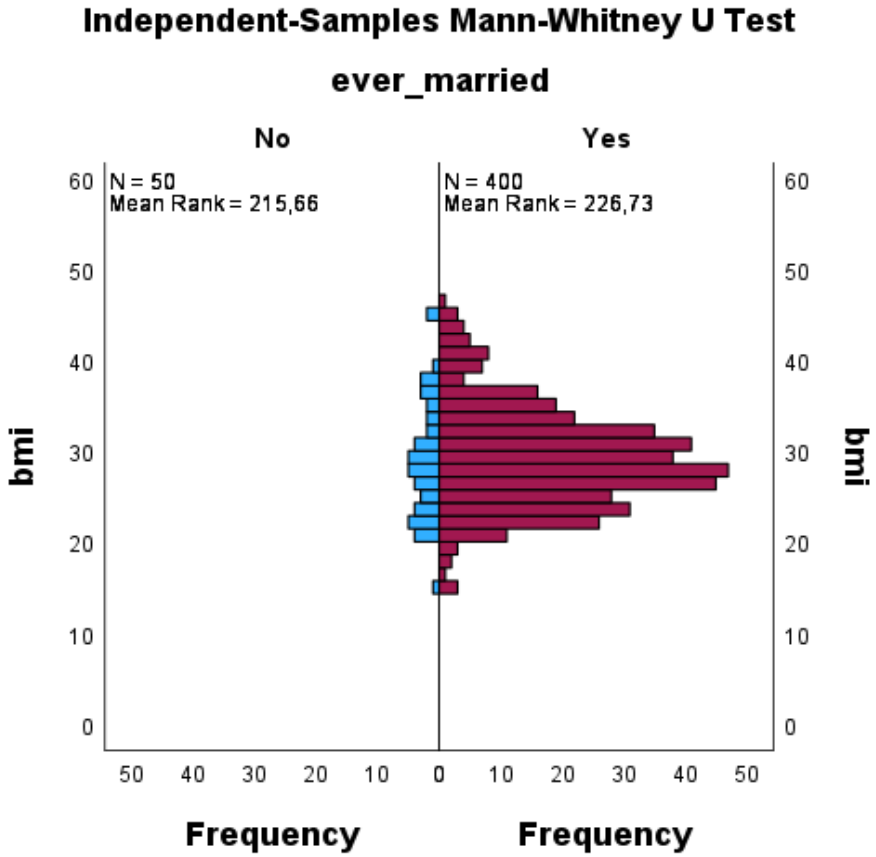
	Null Hypothesis	Test	Sig.[a,b]
1	The distribution of bmi is the same across categories of ever_married.	Independent-Samples Mann-Whitney U Test	,570

Hypothesis Test Summary

	Decision
1	Retain the null hypothesis.

a. The significance level is ,050.

b. Asymptotic significance is displayed.



Conclusion=Marital status does not affect the BMI levels of people who have suffered a stroke

We would like to investigate the people who suffer a stroke, to find out if there is a difference in the BMI among people's smoking status (smokes, formerly smokes and never smokes).

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of bmi is the same across categories of smoking_status.	Independent-Samples Kruskal-Wallis Test	.847	Retain the null hypothesis.

a. The significance level is .050.

b. Asymptotic significance is displayed.

Conclusion = There is no difference in the BMI across different smoking status of people who have suffered stroke.

Independent-Samples Kruskal-Wallis Test

bmi across smoking_status

Independent-Samples Kruskal-Wallis Test Summary

Total N	450
Test Statistic	.331[a]
Degree Of Freedom	2
Asymptotic Sig.(2-sided test)	.847

a. The test statistic is adjusted for ties.

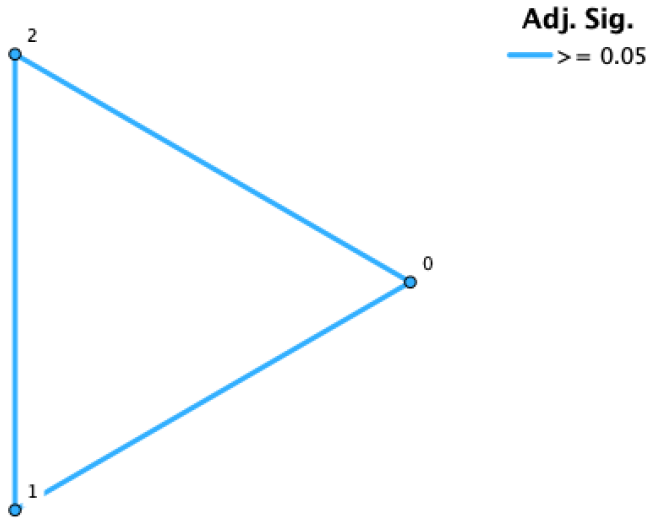
Pairwise Comparisons of smoking_status

Sample 1-Sample 2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.[a]
smokes-never smoked	1.582	16.372	.097	.923	1.000
smokes-formerly smoked	8.499	17.361	.490	.624	1.000
never smoked-formerly smoked	6.917	13.916	.497	.619	1.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .050.

a. Significance values have been adjusted by the Bonferroni correction for multiple tests.

Pairwise Comparisons of smoking_status



Each node shows the sample average rank of smoking_status.

We would like to investigate the people who suffer a stroke, to find out if there is a difference in the average glucose levels among people who reside in rural areas and people who reside in urban areas.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of avg_glucose_level is the same across categories of Residence_type.	Independent-Samples Kruskal-Wallis Test	.953	Retain the null hypothesis.

a. The significance level is .050.

b. Asymptotic significance is displayed.

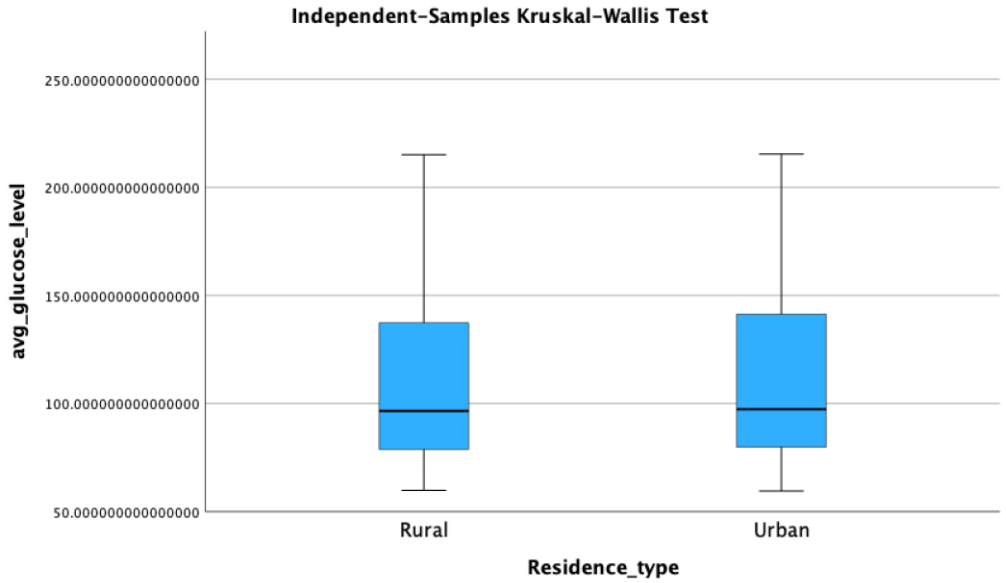
Hypothesis Test Summary

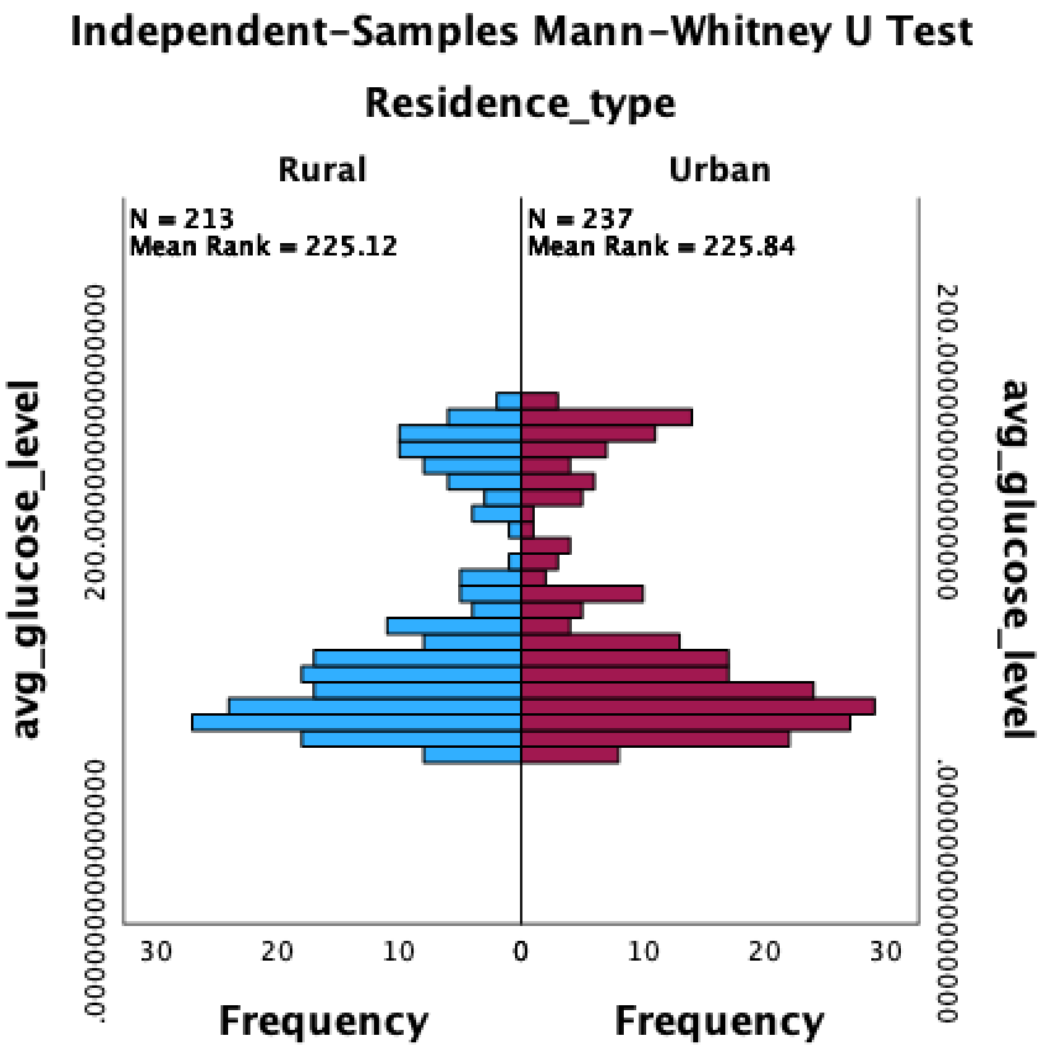
	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of avg_glucose_level is the same across categories of Residence_type.	Independent-Samples Mann-Whitney U Test	.953	Retain the null hypothesis.

a. The significance level is .050.

b. Asymptotic significance is displayed.

Conclusion = There is no difference in the average glucose levels across the different types of residence of people who have suffered stroke.


Independent-Samples Kruskal-Wallis Test
avg_glucose_level
250.000000000000000
200.000000000000000
150.000000000000000
100.000000000000000
50.000000000000000
Rural
Urban
Residence_type


Independent-Samples Mann-Whitney U Test
Residence_type
Rural
Urban
N = 213
Mean Rank = 225.12
N = 237
Mean Rank = 225.84
avg_glucose_level
200.000000000000000
.000000000000000
30
20
10
0
10
20
30
Frequency
Frequency
avg_glucose_level

We would like to investigate if there is a difference in the BMI among people with stroke and people without stroke?

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of bmi is the same across categories of stroke.	Independent-Samples Kruskal-Wallis Test	.323	Retain the null hypothesis.

a. The significance level is .050.

b. Asymptotic significance is displayed.

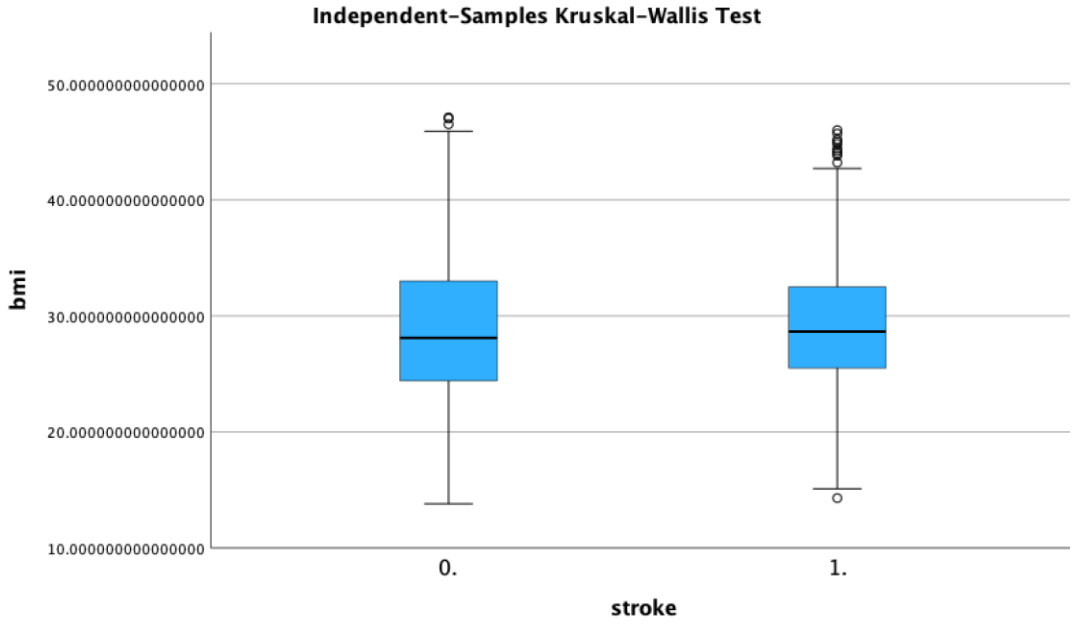
Hypothesis Test Summary

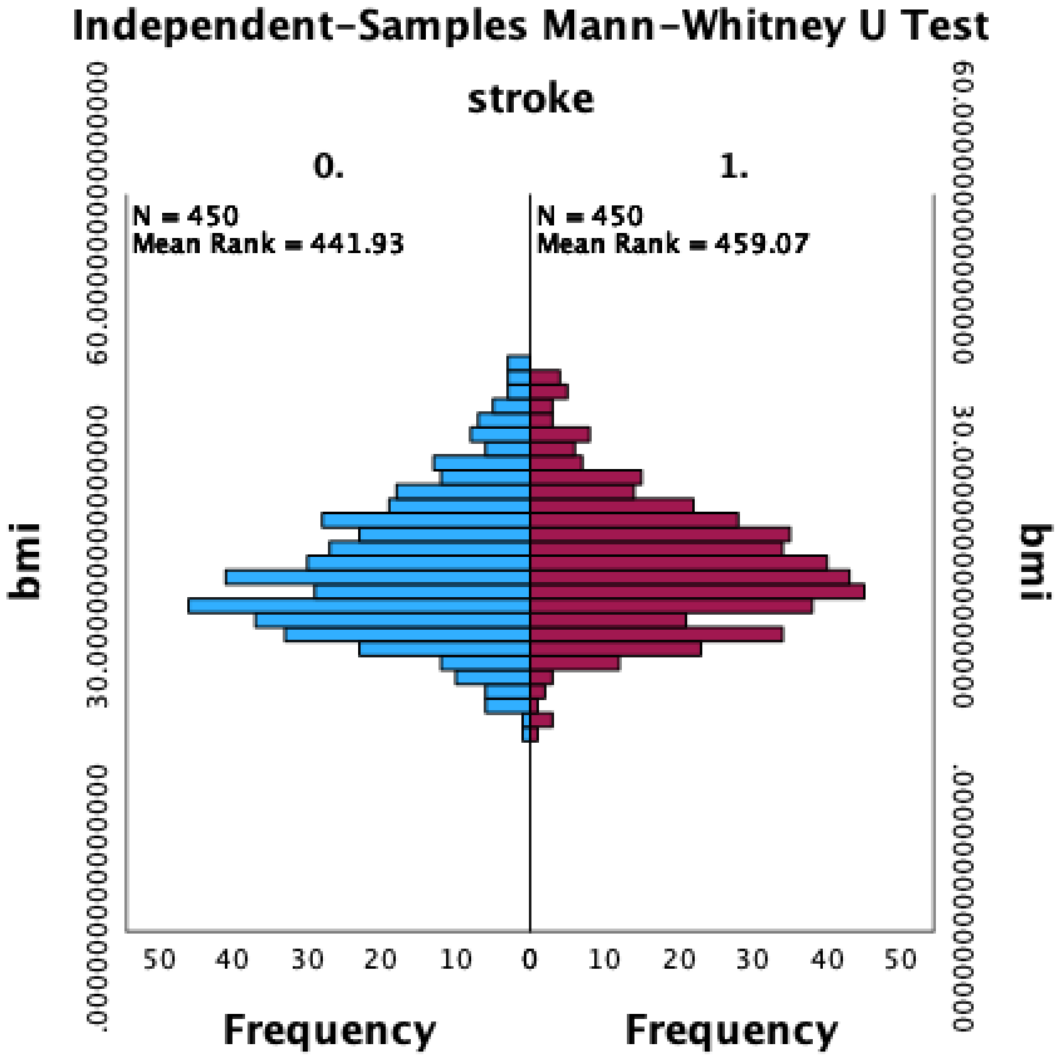
	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of bmi is the same across categories of stroke.	Independent-Samples Mann-Whitney U Test	.323	Retain the null hypothesis.

a. The significance level is .050.

b. Asymptotic significance is displayed.

Conclusion = There is no difference in the BMI across the people with stroke and people without stroke.


Independent-Samples Kruskal-Wallis Test
bmi
50.000000000000000
40.000000000000000
30.000000000000000
20.000000000000000
10.000000000000000
0.
1.
stroke


Independent-Samples Mann-Whitney U Test
stroke
0.
1.
N = 450
Mean Rank = 441.93
N = 450
Mean Rank = 459.07
bmi
60.000000000000000
30.000000000000000
.000000000000000
50 40 30 20 10 0 10 20 30 40 50
Frequency
Frequency

We would like to investigate if there is a difference in the average glucose level among people with stroke and people without stroke?

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of avg_glucose_level is the same across categories of stroke.	Independent-Samples Kruskal-Wallis Test	.009	Reject the null hypothesis.

a. The significance level is .050.

b. Asymptotic significance is displayed.

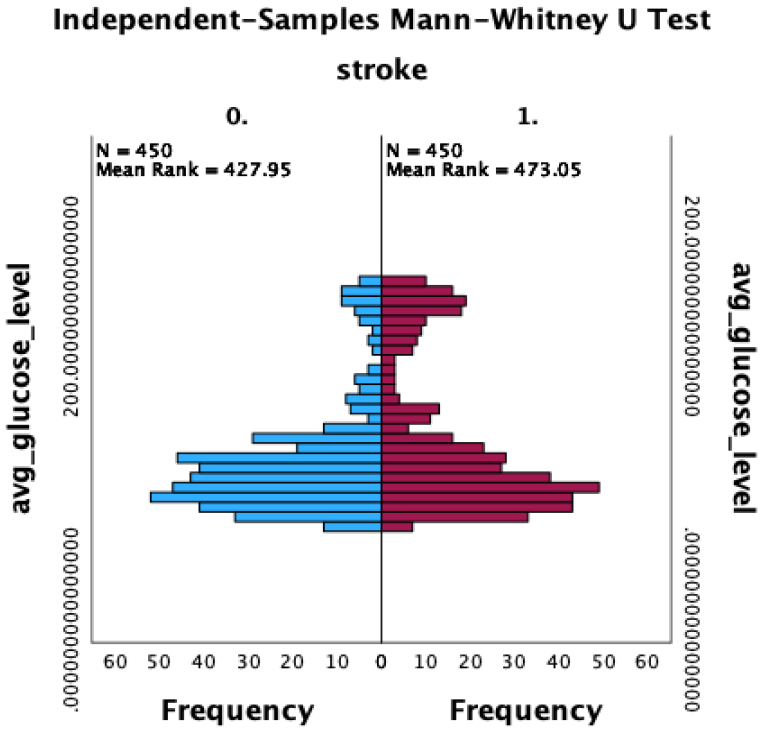
Hypothesis Test Summary

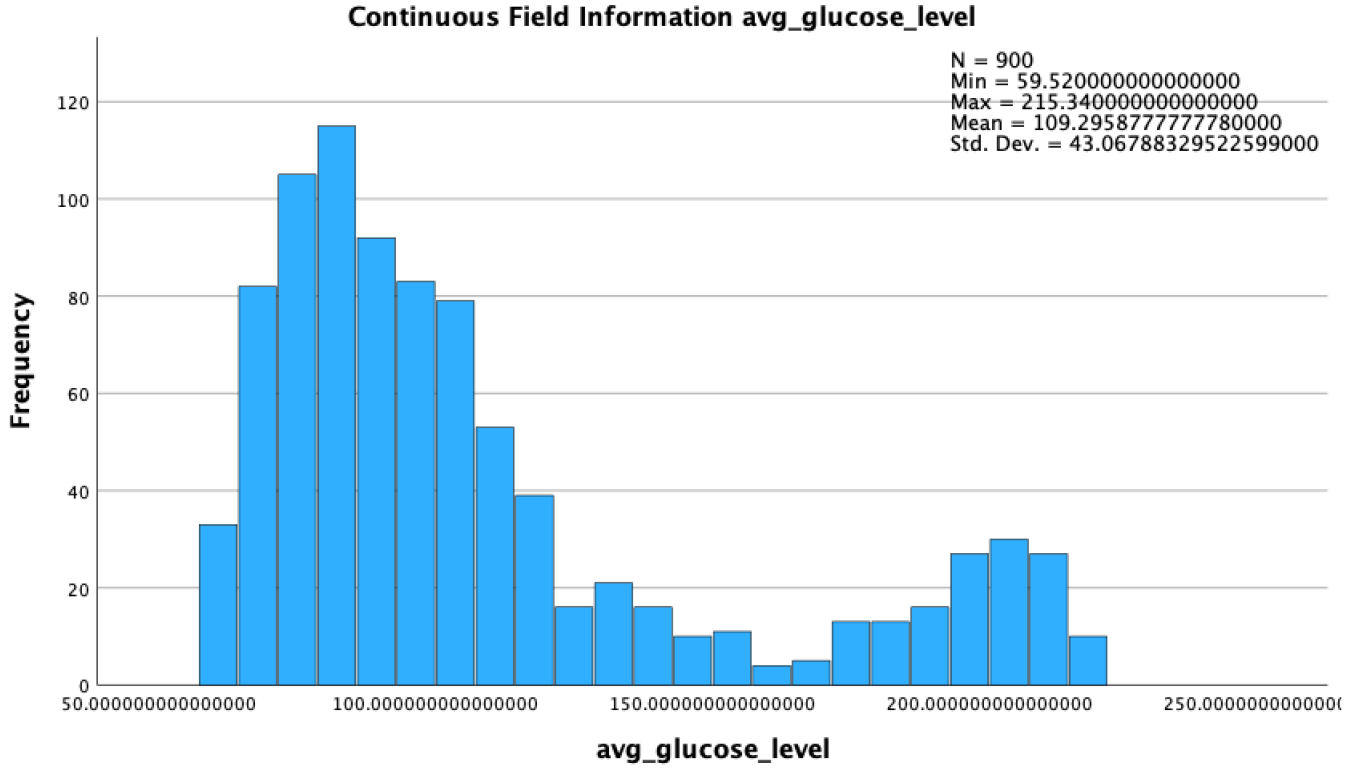
	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of avg_glucose_level is the same across categories of stroke.	Independent-Samples Mann-Whitney U Test	.009	Reject the null hypothesis.

a. The significance level is .050.

b. Asymptotic significance is displayed.

Conclusion = There is a difference in the average glucose levels across the people with stroke and people without stroke. The average glucose levels of people with stroke seem to be a little bit higher than people who did not suffer a stroke.


Independent-Samples Mann-Whitney U Test
stroke
0.
1.
N = 450
Mean Rank = 427.95
N = 450
Mean Rank = 473.05
avg_glucose_level
avg_glucose_level
60 50 40 30 20 10 0 10 20 30 40 50 60
Frequency
Frequency

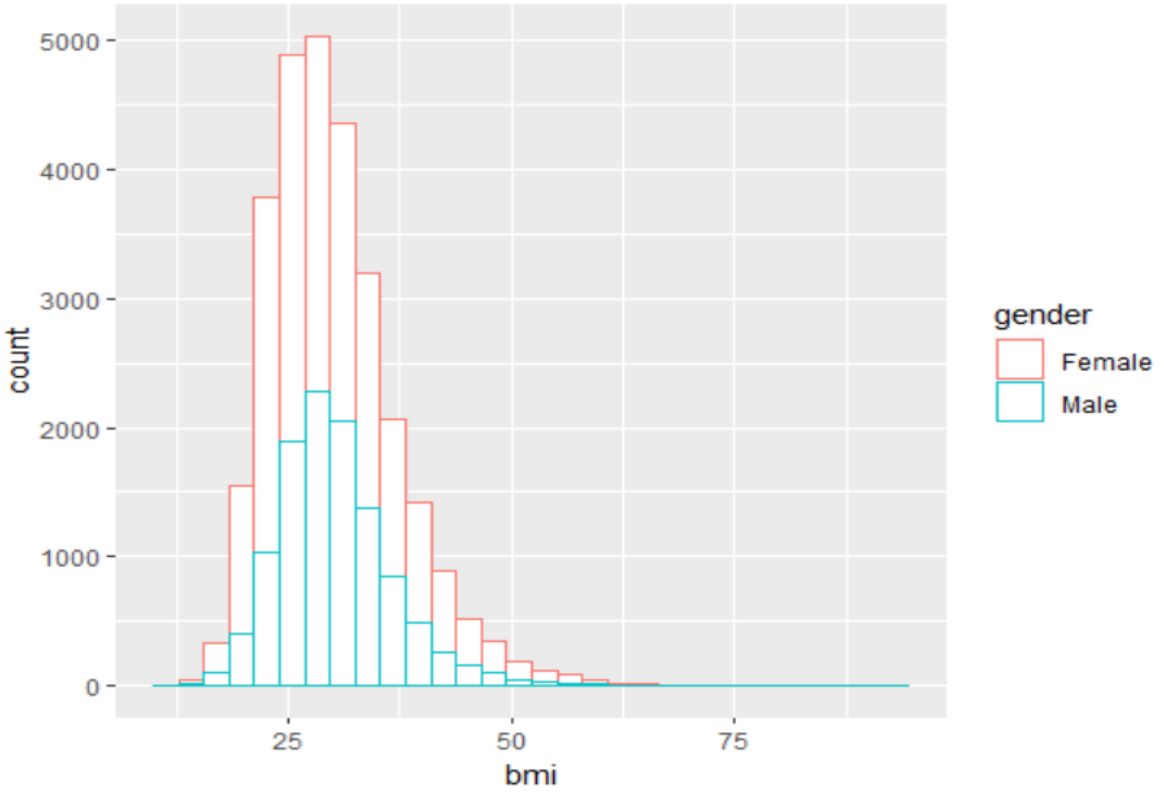

Continuous Field Information avg_glucose_level
N = 900
Min = 59.520000000000000
Max = 215.340000000000000
Mean = 109.295877777780000
Std. Dev. = 43.06788329522599000
Frequency
120
100
80
60
40
20
0
50.000000000000000
100.000000000000000
150.000000000000000
200.000000000000000
avg_glucose_level

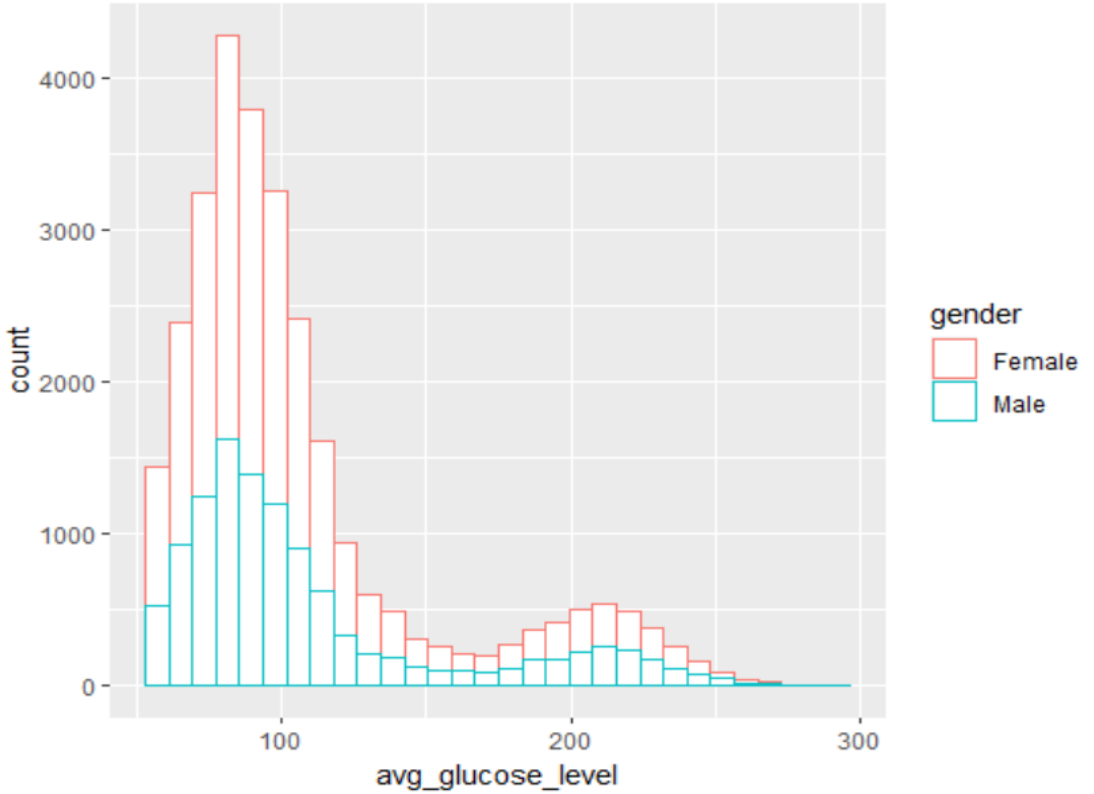
Correlation Analysis, Pearson or Spearman???

Before considering pearson or spearman, we checked the following in our data set.

- **The variables of interest need to be using a continuous scale.** Our variables Glucose lvl, BMI and Age are continuous
- **The pair of variables should have a linear relationship**, which we will check with a scatterplot.
- **There should be no spurious outliers.** We already took out the outliers
- **The variables should be normally or near-to-normally distributed.** We will check with kolmogorov-smirnov test.

BMI & Avg Glucose Lvl should be normally or near-to-normally distributed In order to use Pearson

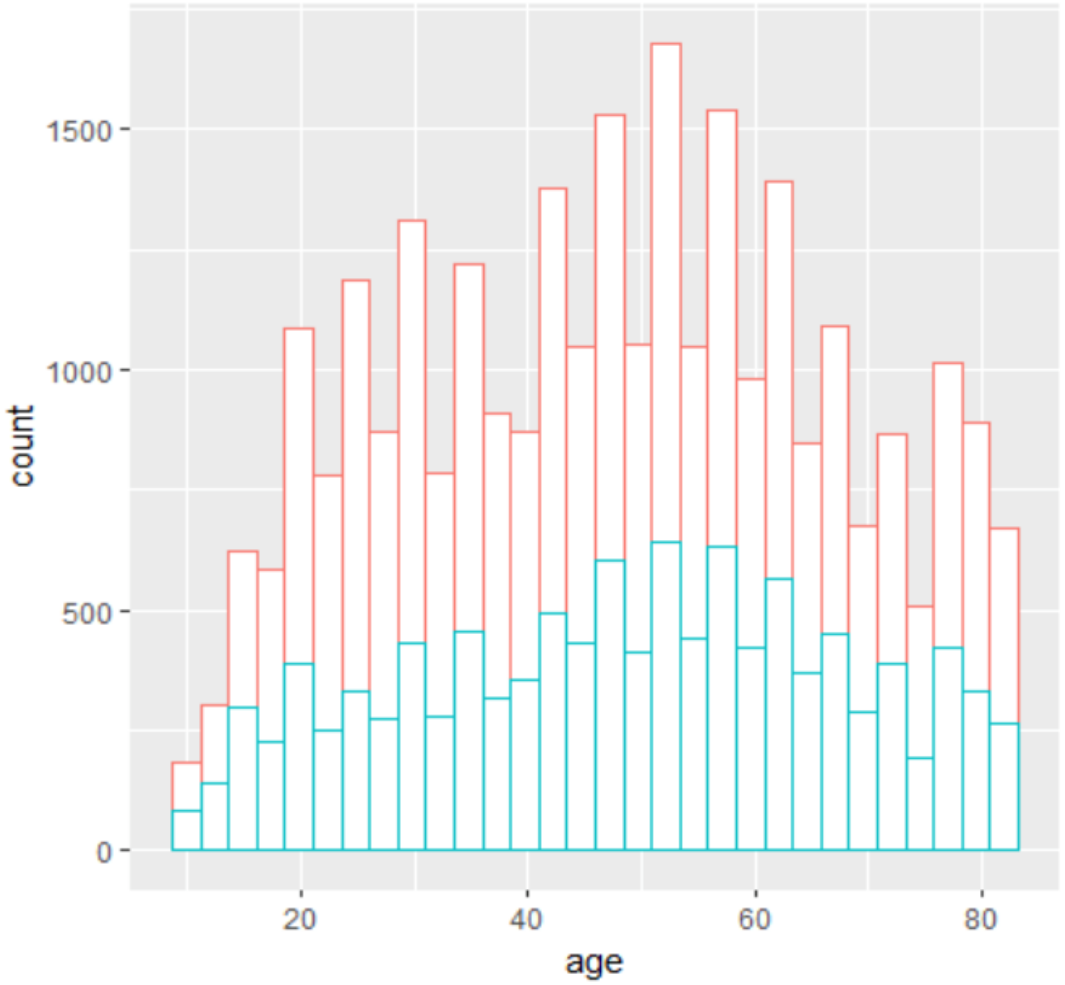





```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  strokenew$bmi
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  strokenew$avg_glucose_level
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

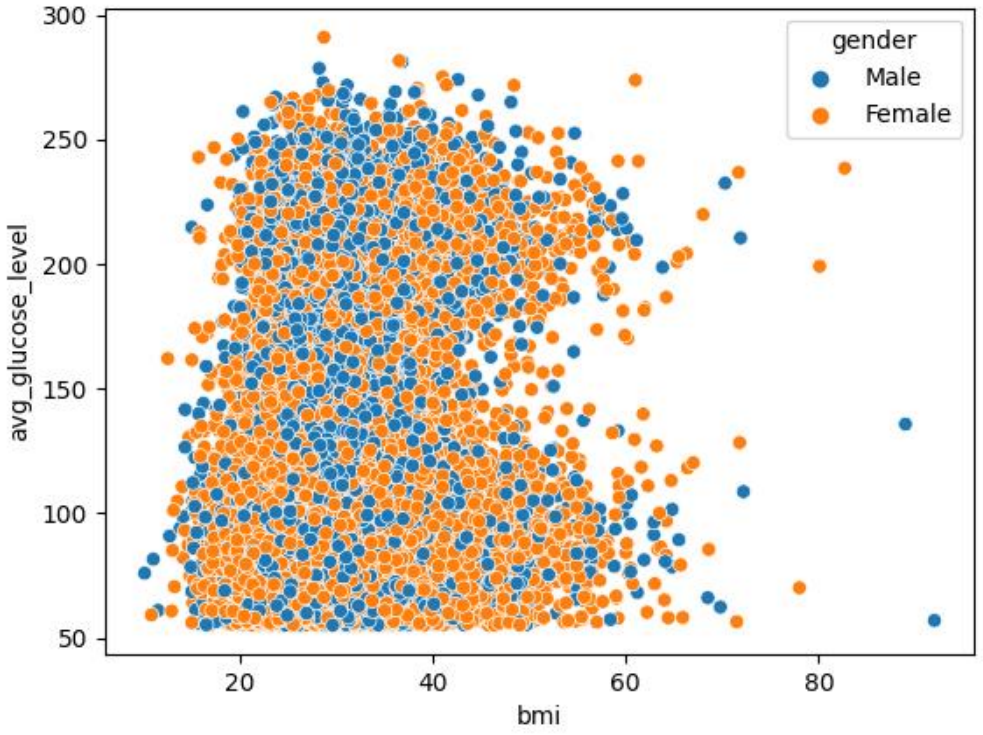
Age should be normally or near-to-normally distributed In order to use Pearson




```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  strokenew$age
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

We will perform a **Spearman** correlation since all the variables are non-normal based on the histograms and kolmogorov tests.

Average Glucose Lvl VS BMI

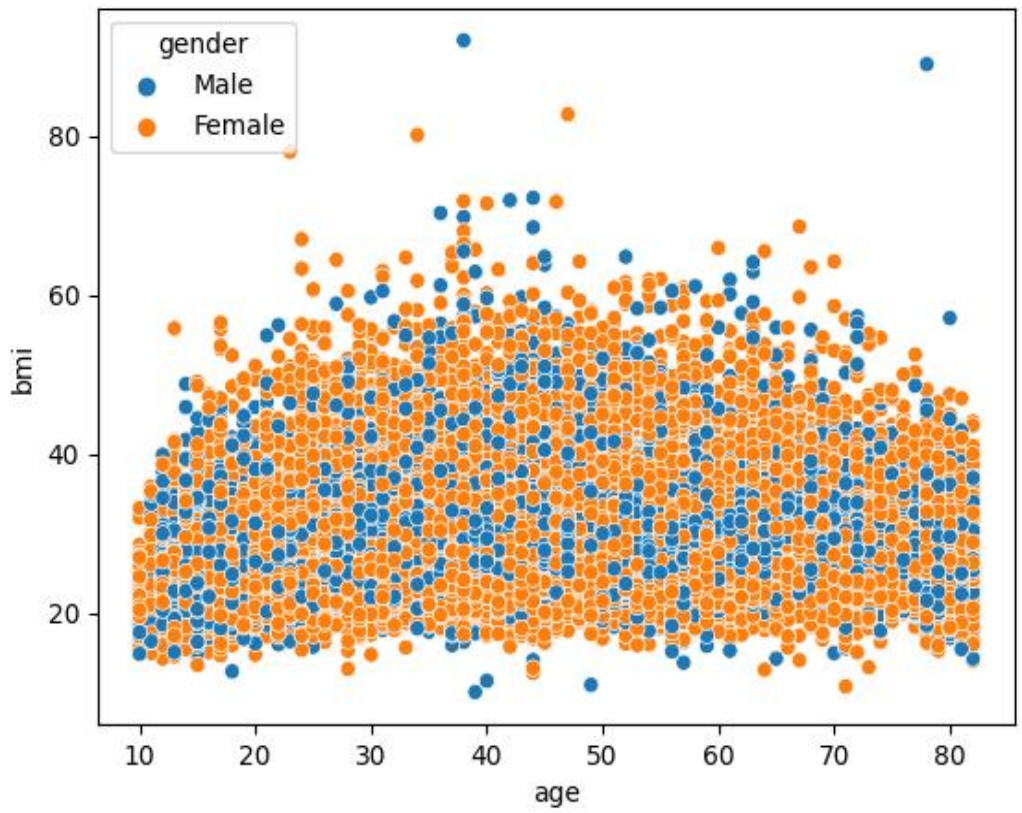


Here from the scatter plot, we realize there is no linear relationship. Let's go on to perform a correlation test.

```
        Spearman's rank correlation rho

data:  strokenew$bmi and strokenew$avg_glucose_level
S = 3.5659e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1150708
```

Also from the spearman's rank correlation test above, there is a weak positive correlation between the two variables. The p value also suggests that.

BMI VS AGE

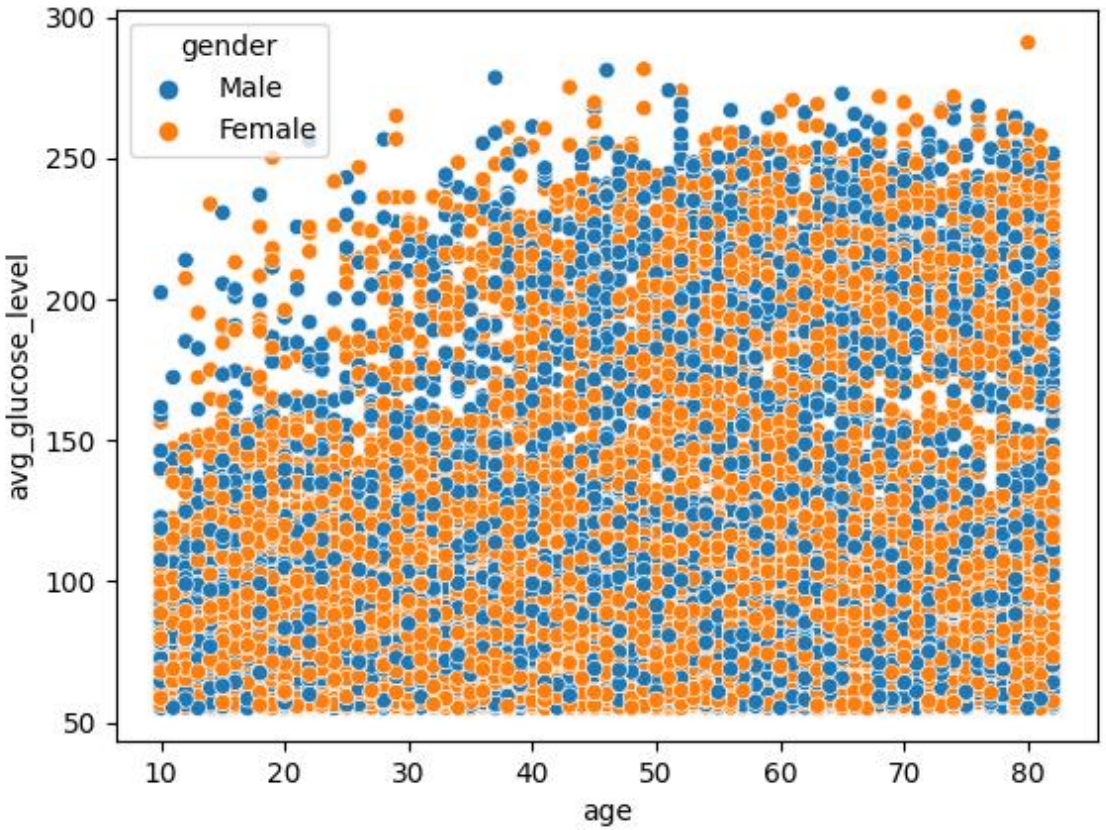


Looking at the scatter plot, we realize there is no linear relationship. Let's go on to perform a correlation test.

```
        Spearman's rank correlation rho

data:  strokenew$age and strokenew$bmi
S = 3.4539e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1428829
```

From the spearman's rank correlation test above, there is a weak positive correlation between the two variables. The p value also suggests that.

Avg Glucose lvl VS Age

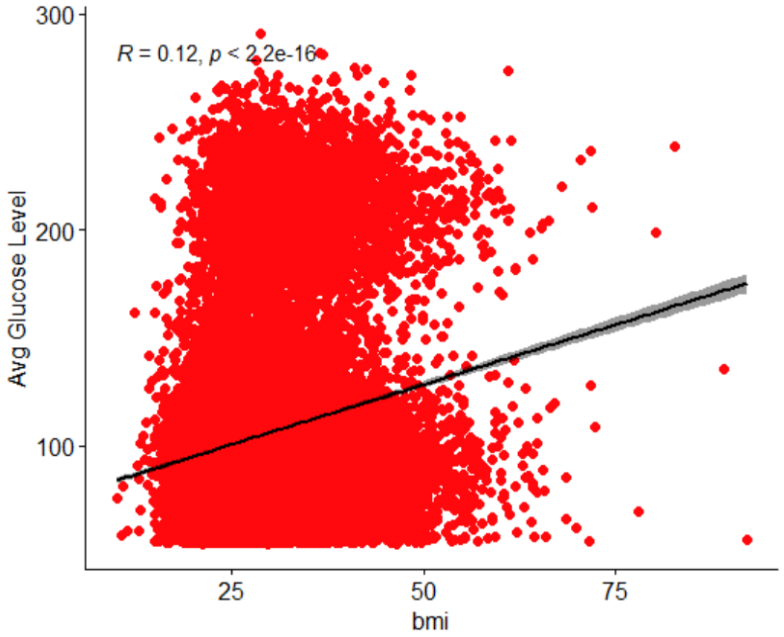


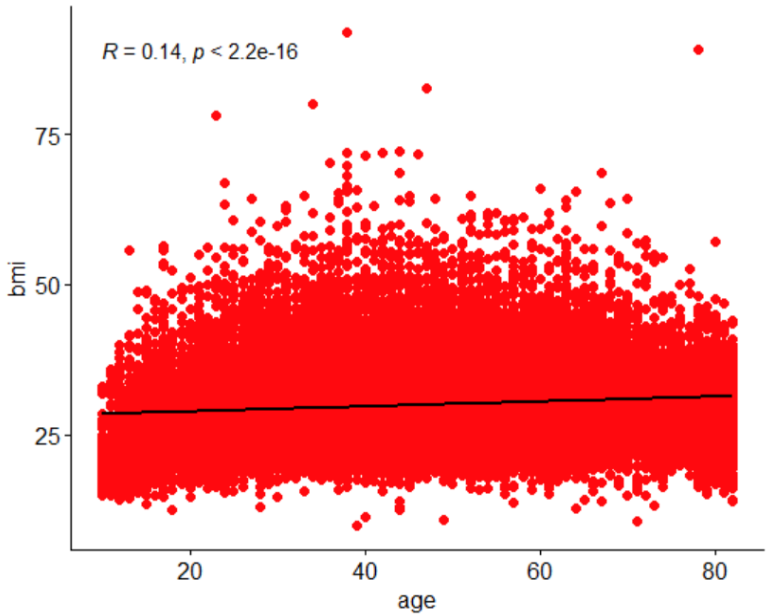
Here from the scatter plot, we realize there is no linear relationship. Let's go on to perform a correlation test.

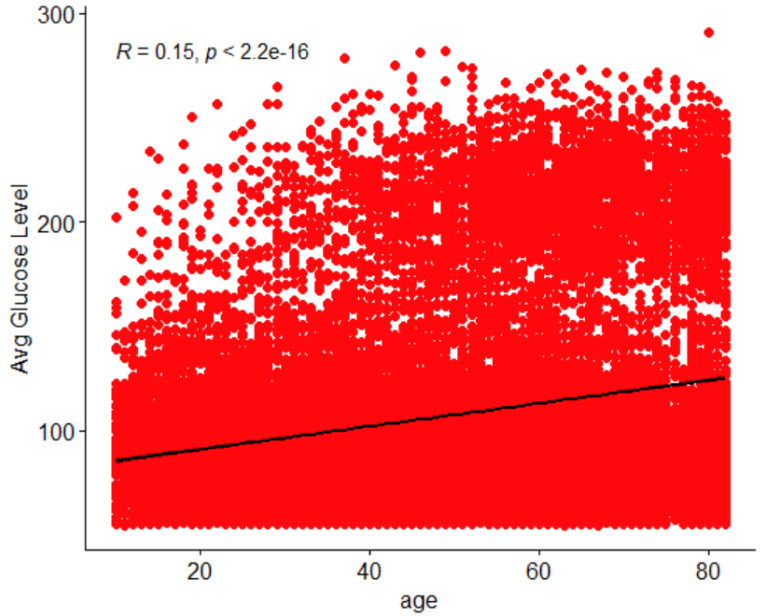
```
        Spearman's rank correlation rho

data:  strokenew$age and strokenew$avg_glucose_level
S = 3.4224e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1506901
```

From the spearman's rank correlation test above, there is a weak positive correlation between the two variables. The p value also suggests that.

Summary







Spearman Corr	Avg Glucose lvl	BMI	Age
Avg Glucose lvl	1	0.12	0.15
BMI	0.12	1	0.14
Age	0.15	0.14	1

Reliability Test

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.204	.398	11

Here the measurements coming from our questionnaire are not reliable because the Cronbach's alpha is less than 0.70

Looking at the inter-item correlation, we have almost all the correlation less than 0.4, which is still a problem for us. This could be due to a low number of questions or poor correlation between items. In our case I feel, some items should be **revised or discarded**

Inter-Item Correlation Matrix

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
gender	1.000	-.040	-.037	-.097	-.024	.020	-.006	-.050	-.022	-.092	-.013
age	-.040	1.000	.258	.247	.547	-.007	.003	.228	.107	.060	.154
hypertension	-.037	.258	1.000	.118	.131	-.018	-.003	.154	.130	.009	.079
heart_disease	-.097	.247	.118	1.000	.095	-.035	-.004	.138	.023	.063	.106
ever_married	-.024	.547	.131	.095	1.000	.094	.005	.118	.143	.070	.048
work_type	.020	-.007	-.018	-.035	.094	1.000	.011	-.013	.084	.045	-.022
Residence_type	-.006	.003	-.003	-.004	.005	.011	1.000	-.003	-.003	.006	.002
avg_glucose_level	-.050	.228	.154	.138	.118	-.013	-.003	1.000	.177	.026	.075
bmi	-.022	.107	.130	.023	.143	.084	-.003	.177	1.000	.021	-.004
smoking_status	-.092	.060	.009	.063	.070	.045	.006	.026	.021	1.000	.011
stroke	-.013	.154	.079	.106	.048	-.022	.002	.075	-.004	.011	1.000

Correlation analysis: Cramer's V

- Cramer's V is a measure of the strength of association between two nominal variables, e.g., heart disease and stroke.

```python
def cramers_V(var1,var2) :
    confusion_matrix =np.array(pd.crosstab(var1,var2, rownames=None, colnames=None)) # Cross table building
    chi2 = stats.chi2_contingency(confusion_matrix)[0]
    n = confusion_matrix.sum()
    return np.sqrt(chi2 / (n*(min(confusion_matrix.shape)-1)))
```

- Cramer's V output gives a value between 0 and +1.

Heart disease + Stroke	Hypertension + Stroke	Marital status + Stroke	Work type + Stroke	Smoking status + Stroke
0.105068	0.078699	0.0474955	0.051203	0.027549

- **Conclusion:** there doesn't seem to be much correlation between our targeted factors and the probability of getting a stroke.

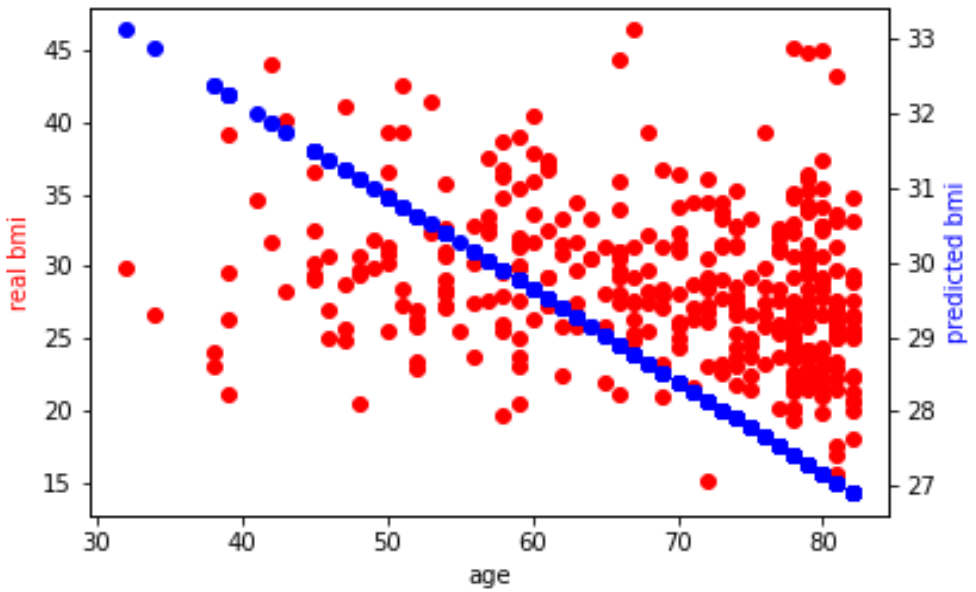
Linear regression: is there a linear relation between the age and BMI of those that suffered a stroke?

OLS Regression Results

Dep. Variable:	bmi	R-squared:	0.073
Model:	OLS	Adj. R-squared:	0.070
Method:	Least Squares	F-statistic:	28.57
Date:	Fri, 03 Feb 2023	Prob (F-statistic):	1.60e-07
Time:	15:54:16	Log-Likelihood:	-1125.0
No. Observations:	365	AIC:	2254.
Df Residuals:	363	BIC:	2262.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	37.0946	1.622	22.875	0.000	33.906	40.284
age	-0.1242	0.023	-5.345	0.000	-0.170	-0.079

Omnibus:	22.941	Durbin-Watson:	2.034
Prob(Omnibus):	0.000	Jarque-Bera (JB):	26.768
Skew:	0.561	Prob(JB):	1.54e-06
Kurtosis:	3.709	Cond. No.	409.



Conclusion: BMI doesn't vary linearly with age for people that suffered a stroke. Most BMI values are concentrated towards old age.

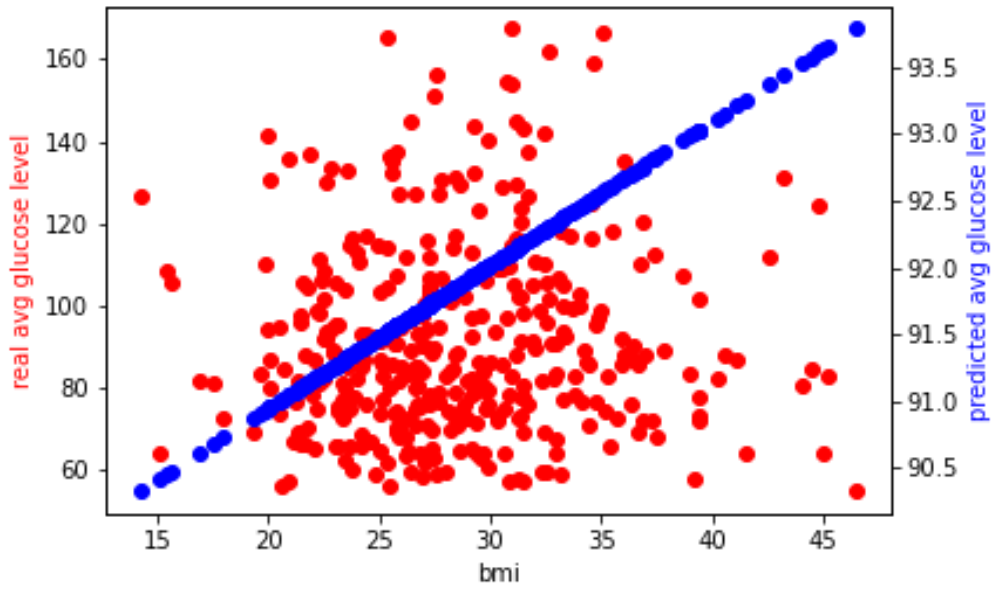
Linear regression: is there a linear relation between the bmi and avg glucose levels of those that suffered a stroke?

OLS Regression Results

Dep. Variable:	avg_glucose_level	R-squared:	0.001
Model:	OLS	Adj. R-squared:	-0.002
Method:	Least Squares	F-statistic:	0.2331
Date:	Fri, 03 Feb 2023	Prob (F-statistic):	0.630
Time:	16:09:03	Log-Likelihood:	-1666.3
No. Observations:	365	AIC:	3337.
Df Residuals:	363	BIC:	3344.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	88.7933	6.475	13.713	0.000	76.060	101.527
bmi	0.1075	0.223	0.483	0.630	-0.330	0.545

Omnibus:	41.455	Durbin-Watson:	1.796
Prob(Omnibus):	0.000	Jarque-Bera (JB):	52.219
Skew:	0.896	Prob(JB):	4.58e-12
Kurtosis:	3.470	Cond. No.	154.



Conclusion: the average glucose level doesn't vary linearly with the BMI for people that suffered from a stroke.

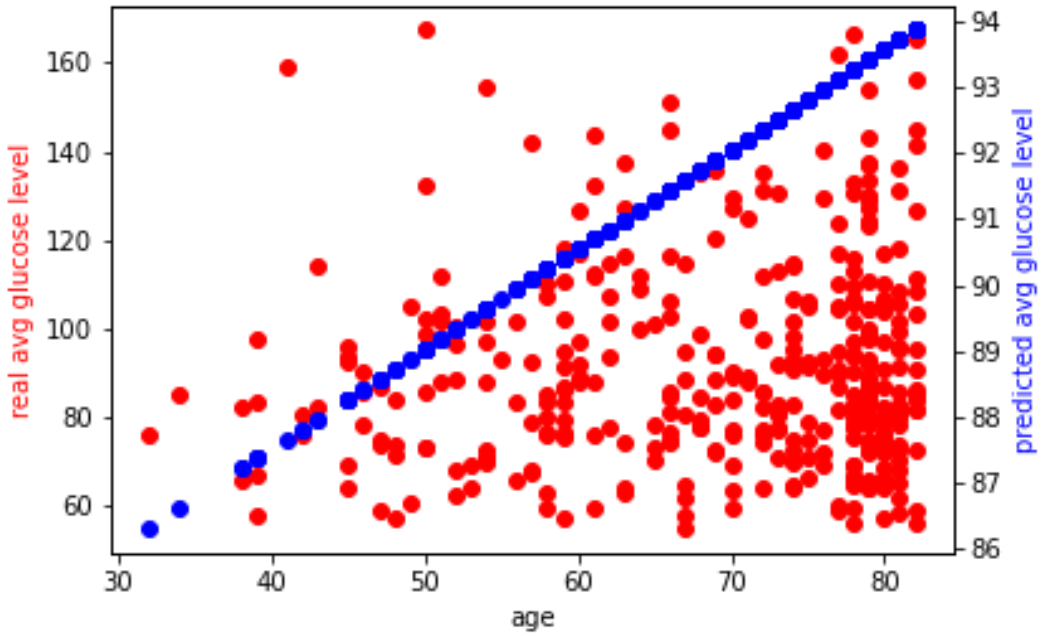
Linear regression: is there a linear relation between the age and avg glucose levels of those that suffered a stroke?

OLS Regression Results

Dep. Variable:	avg_glucose_level	R-squared:	0.006
Model:	OLS	Adj. R-squared:	0.003
Method:	Least Squares	F-statistic:	2.189
Date:	Fri, 03 Feb 2023	Prob (F-statistic):	0.140
Time:	16:18:23	Log-Likelihood:	-1665.4
No. Observations:	365	AIC:	3335.
Df Residuals:	363	BIC:	3343.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	81.4737	7.127	11.431	0.000	67.458	95.490
age	0.1511	0.102	1.479	0.140	-0.050	0.352

Omnibus:	41.794	Durbin-Watson:	1.797
Prob(Omnibus):	0.000	Jarque-Bera (JB):	52.741
Skew:	0.895	Prob(JB):	3.53e-12
Kurtosis:	3.512	Cond. No.	409.



Conclusion: the avg glucose level doesn't vary linearly with age for people that suffered a stroke.

Multi-linear regression: is there a multi-linear relation between the age, BMI and avg glucose levels of people that suffer a stroke?

OLS Regression Results

Dep. Variable:	avg_glucose_level	R-squared:	0.008
Model:	OLS	Adj. R-squared:	0.003
Method:	Least Squares	F-statistic:	1.515
Date:	Fri, 03 Feb 2023	Prob (F-statistic):	0.221
Time:	16:33:47	Log-Likelihood:	-1664.9
No. Observations:	365	AIC:	3336.
Df Residuals:	362	BIC:	3348.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	73.6192	11.139	6.609	0.000	51.714	95.525
age	0.1774	0.106	1.672	0.095	-0.031	0.386
bmi	0.2117	0.231	0.918	0.359	-0.242	0.665

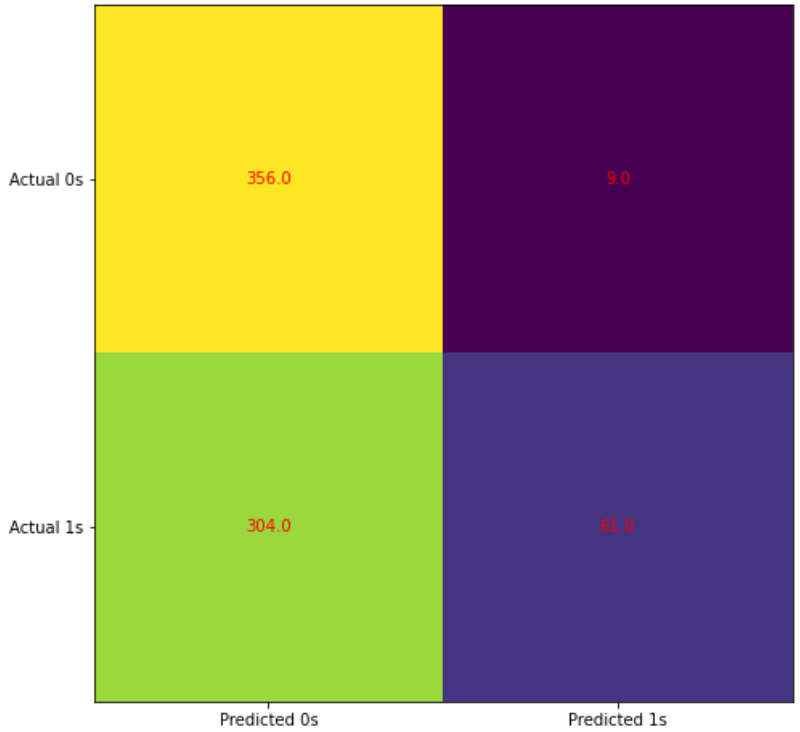
Omnibus:	40.878	Durbin-Watson:	1.794
Prob(Omnibus):	0.000	Jarque-Bera (JB):	51.277
Skew:	0.885	Prob(JB):	7.33e-12
Kurtosis:	3.490	Cond. No.	688.

Conclusion: the avg glucose level doesn't vary linearly with age and BMI for people that suffered a stroke.

Logistic regression: Does having a heart disease lead to having a stroke?

Model:	Logit	Pseudo R-squared:	0.043
Dependent Variable:	stroke	AIC:	970.6670
Date:	2023-02-03 16:47	BIC:	975.2600
No. Observations:	730	Log-Likelihood:	-484.33
Df Model:	0	LL-Null:	-506.00
Df Residuals:	729	LLR p-value:	nan
Converged:	1.0000	Scale:	1.0000
No. Iterations:	6.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
heart_disease	1.9136	0.3571	5.3592	0.0000	1.2138	2.6135

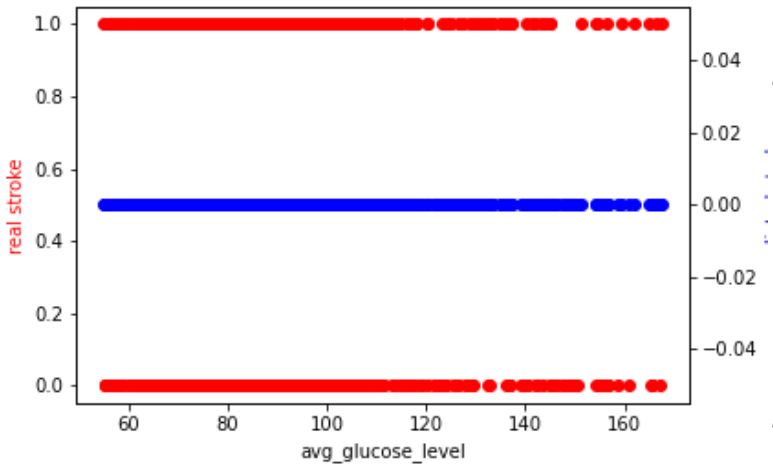


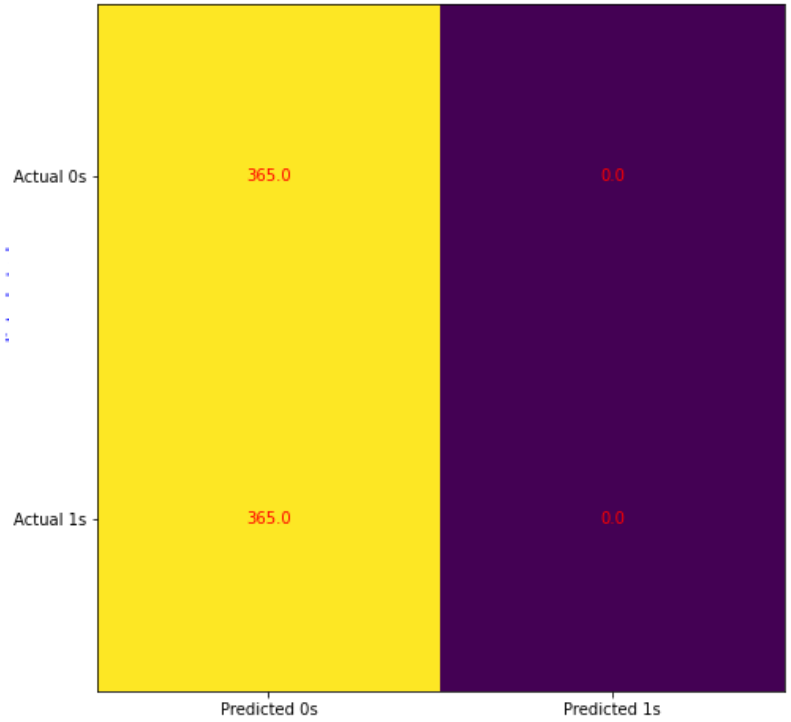
Conclusion: the model fails to correctly predict the probability of having a stroke based on the status of heart disease. More information is needed.

Logistic regression: is the avg glucose level a contributing factor to having a stroke?

Model:	Logit	Pseudo R-squared:	0.000
Dependent Variable:	stroke	AIC:	1013.9181
Date:	2023-02-03 17:16	BIC:	1018.5111
No. Observations:	730	Log-Likelihood:	-505.96
Df Model:	0	LL-Null:	-506.00
Df Residuals:	729	LLR p-value:	nan
Converged:	1.0000	Scale:	1.0000
No. Iterations:	2.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
avg_glucose_level	-0.0002	0.0008	-0.2772	0.7817	-0.0017	0.0013



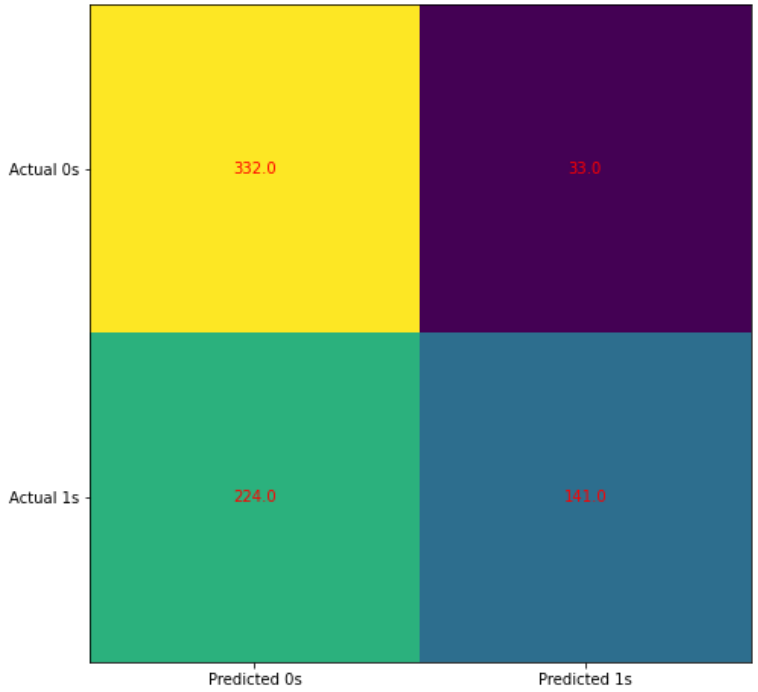


Conclusion: the logistic regression fails to predict the probability of getting a stroke based on a person's avg glucose level. 50% of the cases were predicted wrong.

Logistic regression: are the avg glucose levels + hypertension + BMI + heart disease contributing factors to having a stroke?

Model:	Logit	Pseudo R-squared:	0.096
Dependent Variable:	stroke	AIC:	923.1729
Date:	2023-02-03 17:24	BIC:	941.5451
No. Observations:	730	Log-Likelihood:	-457.59
Df Model:	3	LL-Null:	-506.00
Df Residuals:	726	LLR p-value:	7.4941e-21
Converged:	1.0000	Scale:	1.0000
No. Iterations:	6.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
hypertension	1.5372	0.2406	6.3877	0.0000	1.0655	2.0088
heart_disease	2.0352	0.3719	5.4725	0.0000	1.3063	2.7641
avg_glucose_level	-0.0019	0.0027	-0.7153	0.4744	-0.0071	0.0033
bmi	-0.0077	0.0088	-0.8796	0.3791	-0.0249	0.0095



Conclusion: the model fails to predict the probability of getting a stroke for 30% of the cases when info on their blood sugar, hypertension, heart disease and BMI is used as input.

Conclusion

- Since our data is taken at one point in time, we could not use any paired tests, e.g., Wilcoxon or Friedman tests.
- While BMI didn't show any variation between people w/o stroke, the average glucose levels are significantly higher in those that suffered a stroke.
- There seem to be a negligible correlation relation between (age, BMI, avg glucose levels) for people that suffered a stroke as well as a negligible correlation between the probability of getting a stroke and most of our nominal variables.
- Both linear and logistic regressions failed to predict the proper relation between our variables.
- More thorough survey and dataset are required to draw conclusions about the factors leading to getting a stroke.