Stroke Data Set

Normality tests

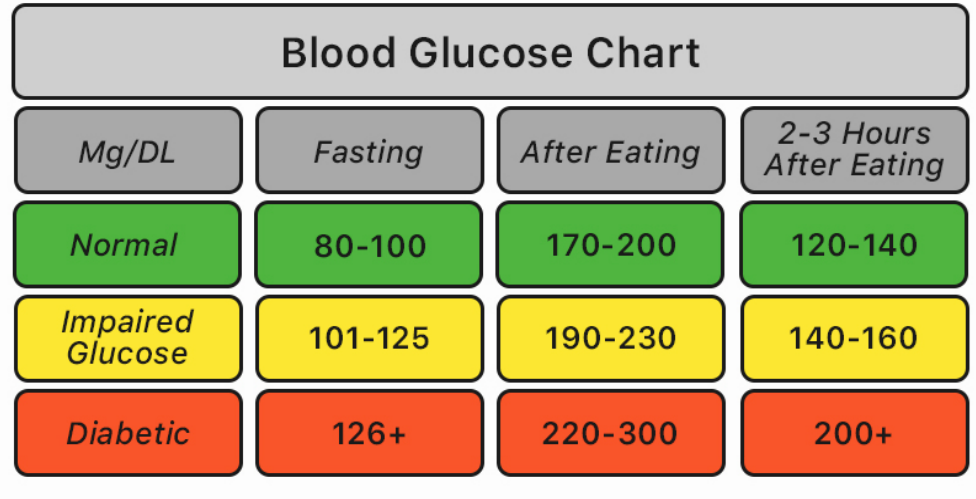
- **Group D**
- **Aya Abouelela**
- **Amel Khirreddine**
- **Kwaku Asamoah Gyimah**
- **Arlizze Faye R. Ongchua**
- **Supervisor: Prof. Elnaz Gholipiour**

Goal of the project: checking the normality of some of our variables, e.g. bmi, avg glucose levels and age for both females and males using visual and statistical tests.

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

Alternative hypothesis: Our avg glucose levels are not normally distributed

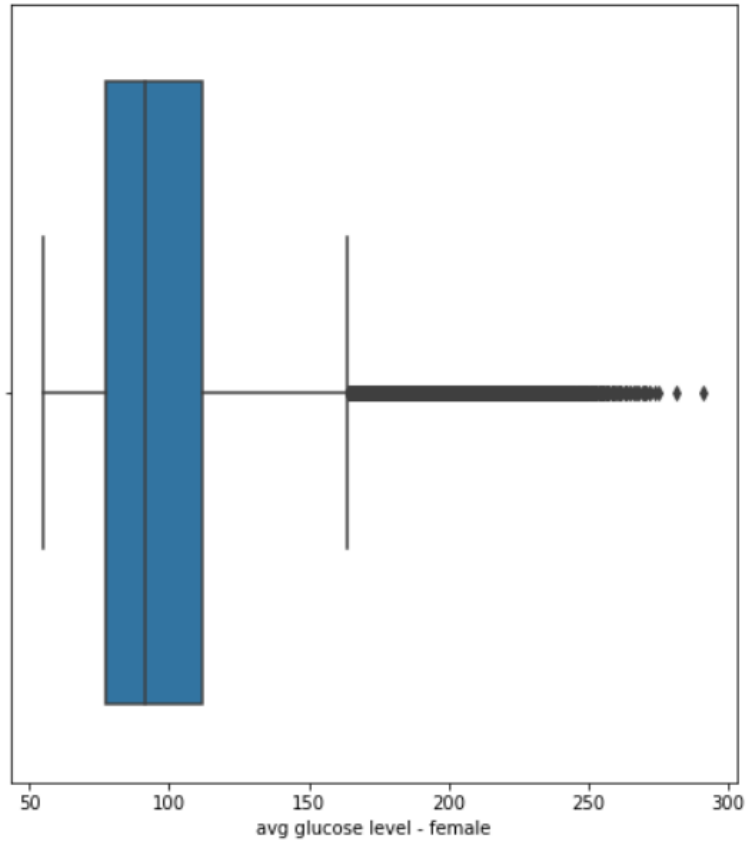
- **The collected data doesn't specify if the numbers were taken at fasting, directly after eating or a while after eating.**
- **The number also don't show the category of people which the sample is drawn from them, e.g, if they are normal, prediabetic or diabetic.**
- **Some ranges overlap e.g, a normal glucose 2-3 hours after eating overlaps with a diabetic glucose level at fasting.**

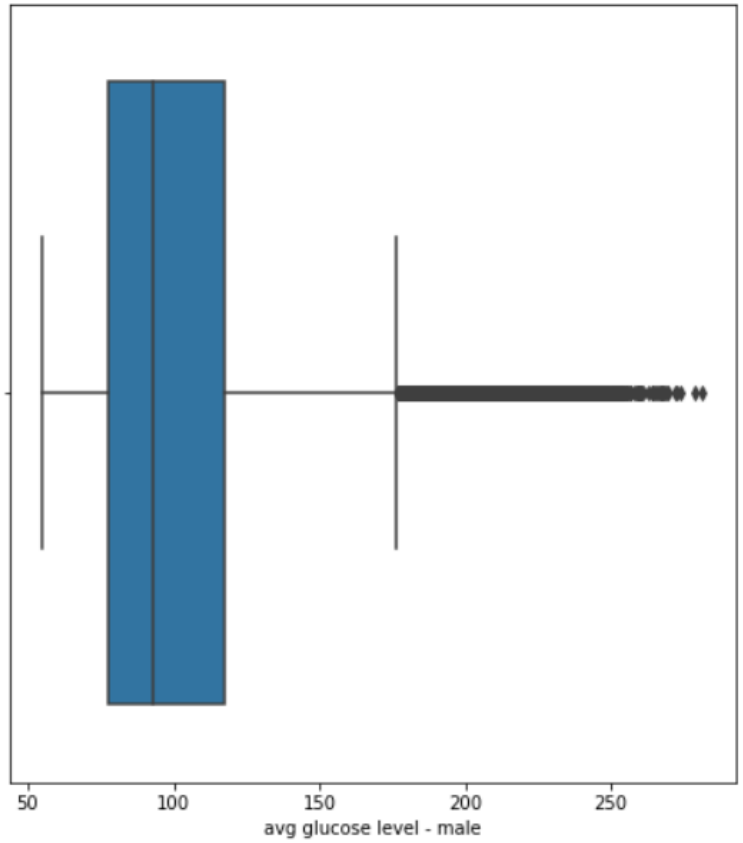


Blood Glucose Chart			
Mg/DL	*Fasting*	*After Eating*	*2-3 Hours After Eating*
Normal	80-100	170-200	120-140
Impaired Glucose	101-125	190-230	140-160
Diabetic	126+	220-300	200+

Ref: https://www.lark.com/resources/blood-sugar-chart

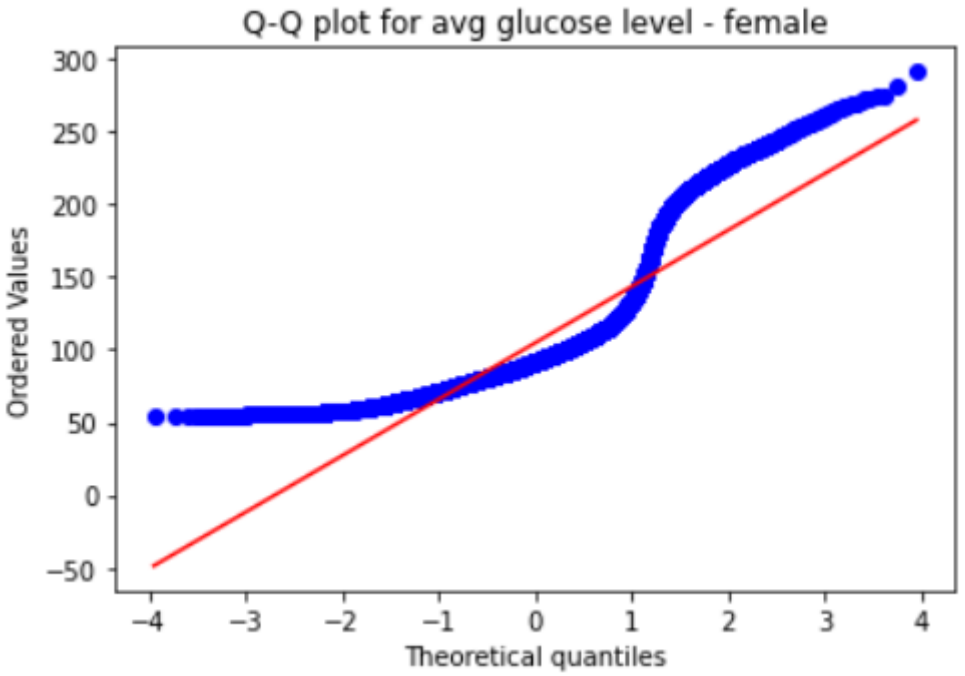
Avg glucose levels; before removing the outliers box-plots

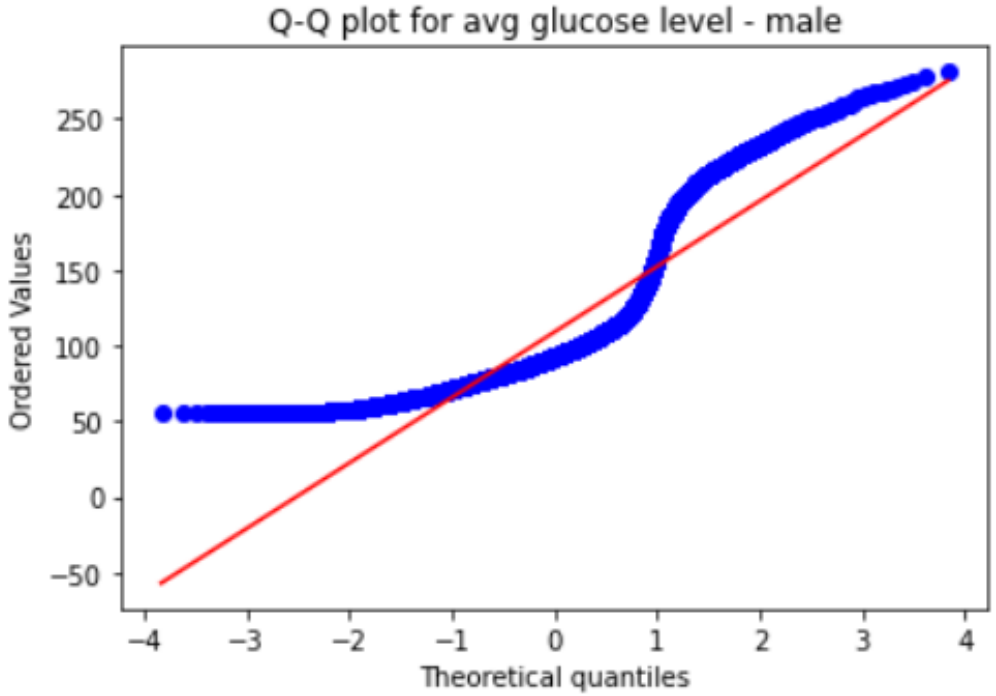




	Mean/Median/Mode
Female	104.58/91.57/73.0
Male	109.26/92.95/83.1

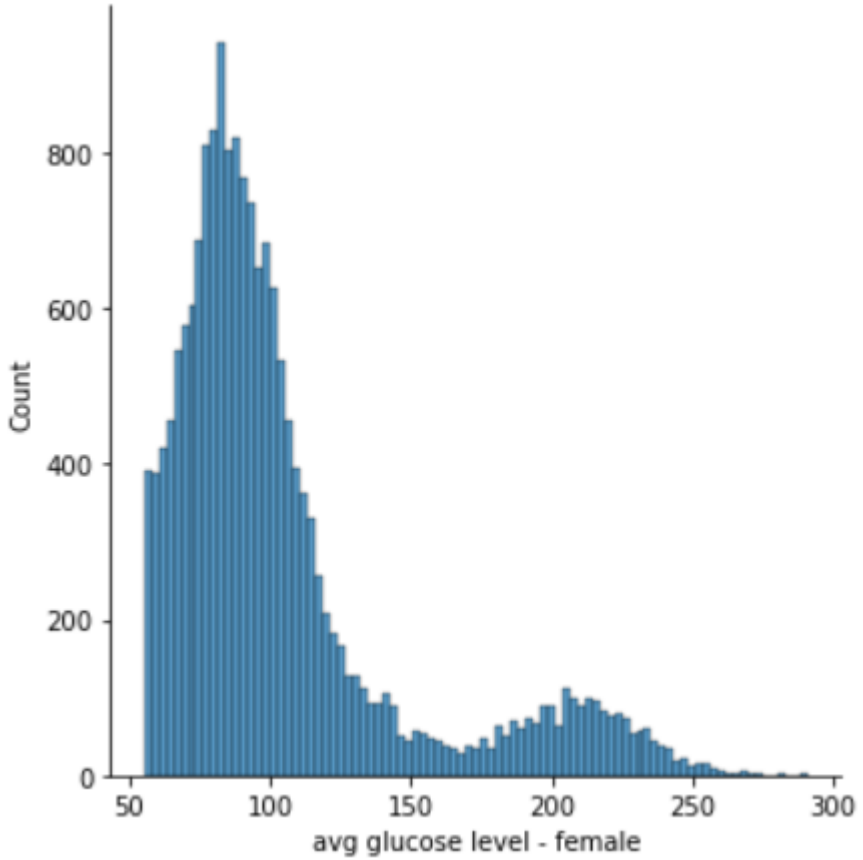
Avg glucose levels; before removing the outliers
Q-Q plot

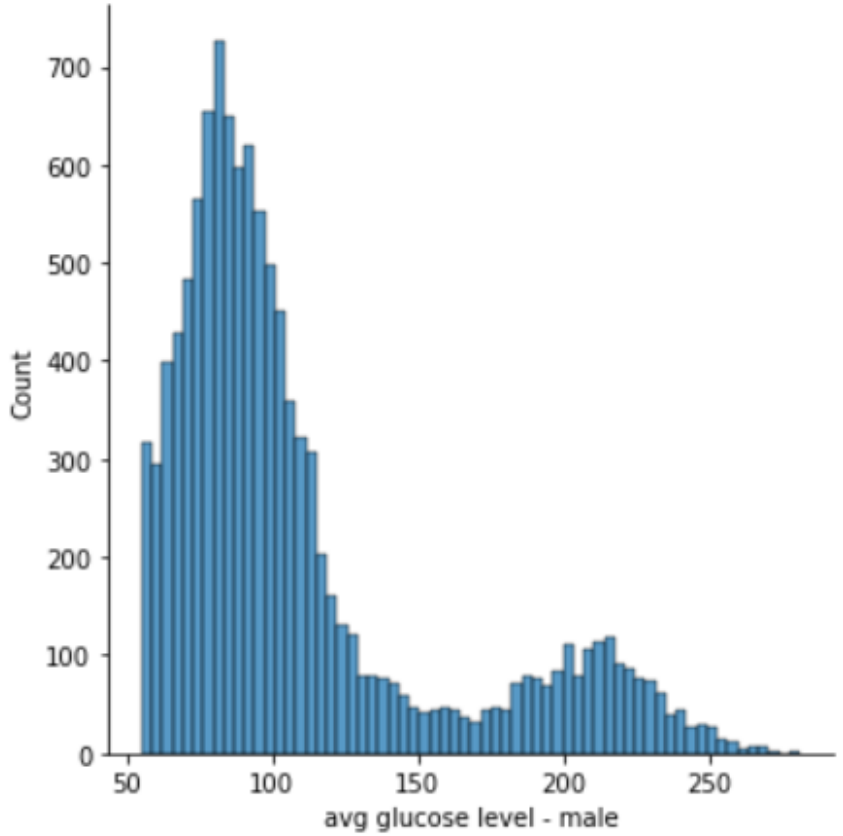




Avg glucose levels; before removing the outliers histogram plot

Clearly not bell-shaped :D





Avg glucose levels; before removing the outliers
Statistical Methods

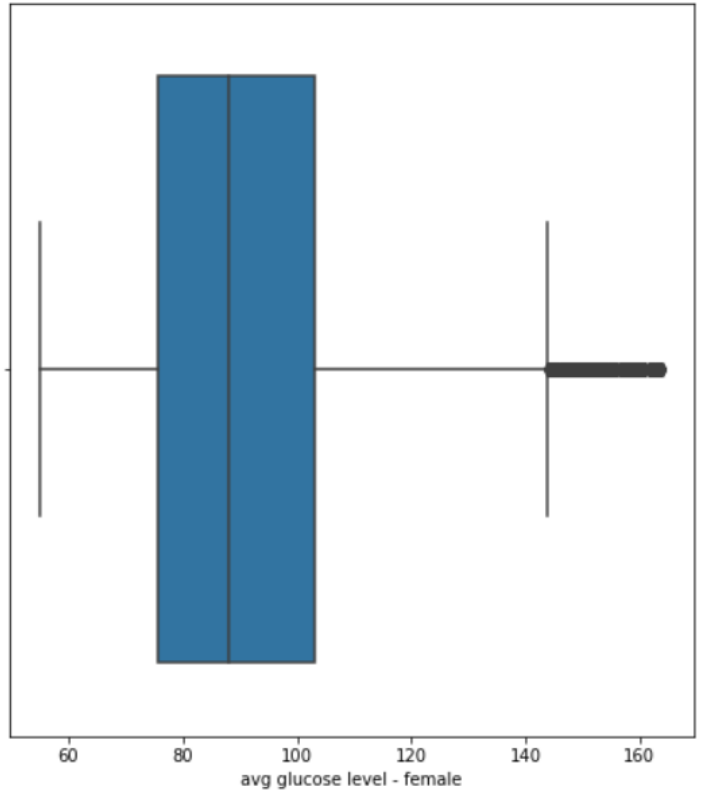
```
1  ## Statistical tests - Shapiro-Wilk
2  from scipy.stats import shapiro
3  result, p = shapiro(df_female_glucose)
4  print('avg glucose level female','z=%0.3f, p = %0.3f\n' % (result, p))
```

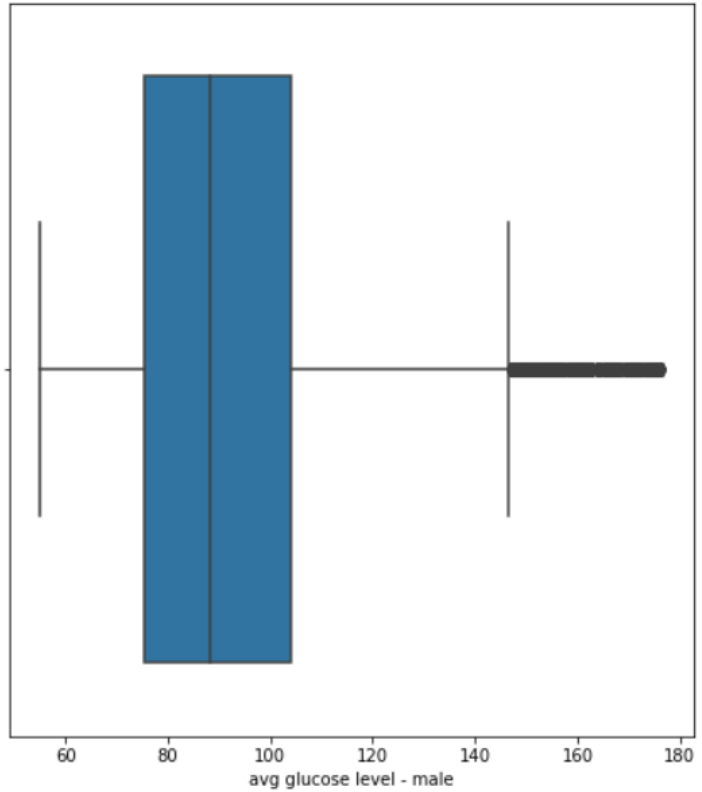
```
avg glucose level female z=0.801, p = 0.000
```

Test	Female	Male
Shapiro-Wilk	Z = 0.801, p = 0.000	Z = 0.8111, p = 0.000
Skewness	Z = 64.964, p = 0.000	Z = 46.652, p = 0.000
Kurtosis	Z = 29.728, p = 0.000	Z = 14.917, p = 0.000
Kolmogorov–Smirnov	Z = 1.000, p = 0.000	Z = 1.000, p = 0.000
Anderson-Darling	Z = 1223.94	Z = 797.411

Anderson darling critical values = [0.576 0.656 0.787 0.918 1.092]
Anderson darling statistical significance = [15. 10. 5. 2.5 1.]

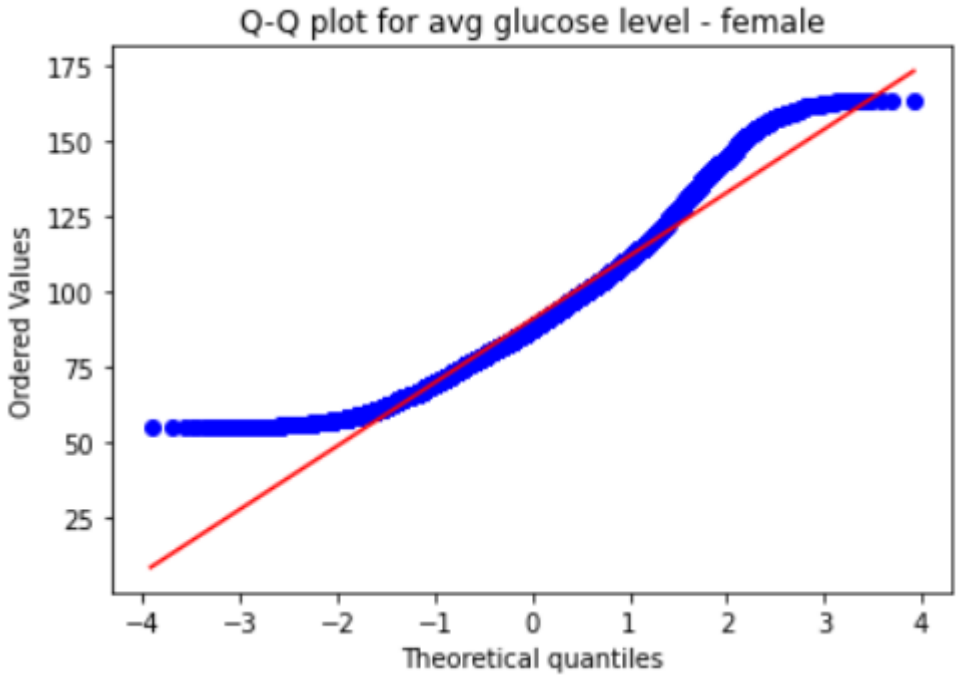
Avg glucose levels; after removing the outliers box-plots

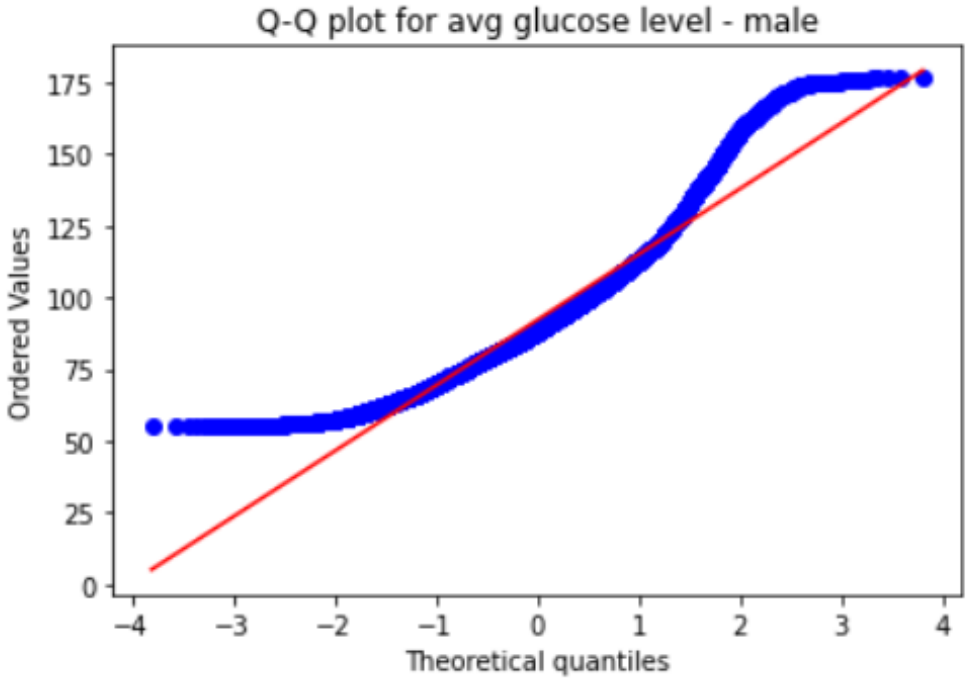




	Mean/Median/Mode
Female	90.98/88.11/73.0
Male	92.14/88.33/83.1

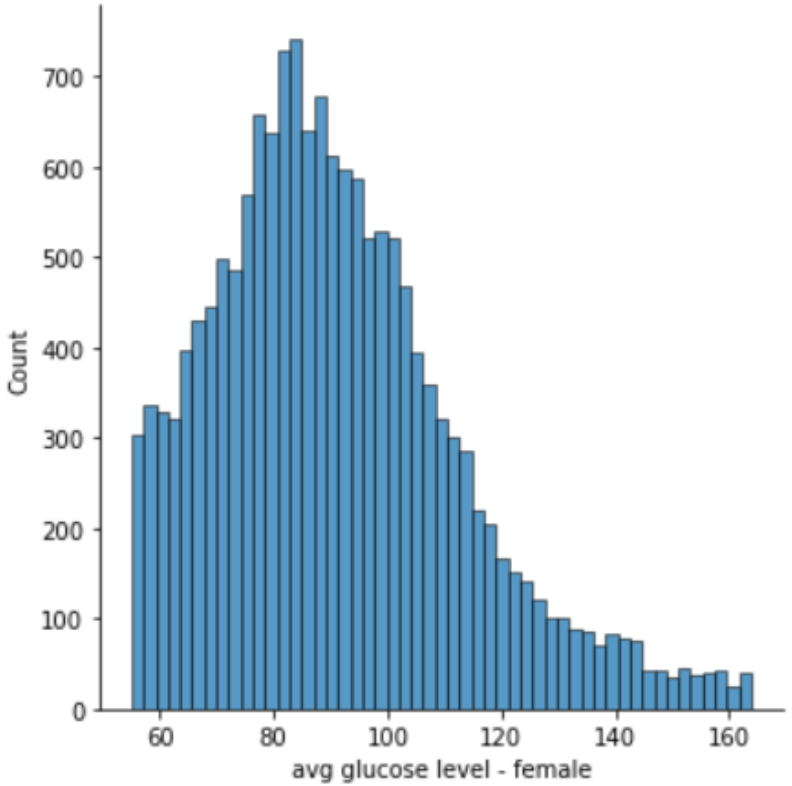
Avg glucose levels; after removing the outliers
Q-Q plot

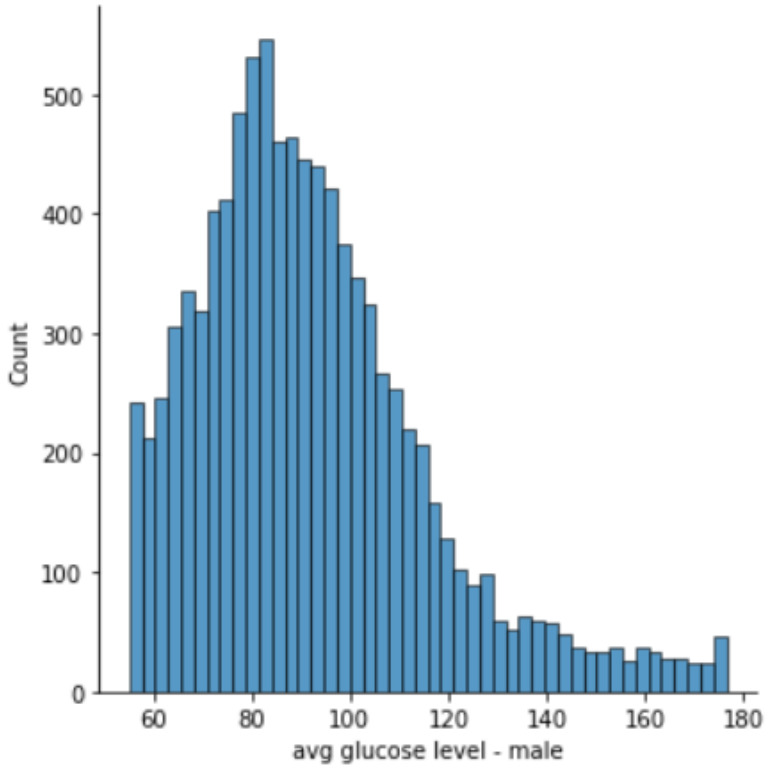




Avg glucose levels; after removing the outliers histogram plot

still not bell-shaped :D





Avg glucose levels; after removing the outliers
Statistical Methods

Test	Female	Male
Shapiro-Wilk	Z = 0.960, p = 0.000	Z = 0.935, p = 0.000
Skewness	Z = 34.956, p = 0.000	Z = 34.807, p = 0.000
Kurtosis	Z = 10.273 , p = 0.000	Z = 15.840, p = 0.000
Kolmogorov–Smirnov	Z = 1.000, p = 0.000	Z = 1.000, p = 0.000
Anderson-Darling	Z = 125.964	Z = 134.843

Anderson darling critical values = [0.576 0.656 0.787 0.918 1.092]
Anderson darling statistical significance = [15. 10. 5. 2.5 1.]

- Since our data is not normal, one cannot use the z-score to calculate the cumulative probability.

- Cumulative probability can be calculated by direct division of the number of entries of the desired range over the total number of entries.
- Example, what is the percentage of males in our dataset with avg glucose levels between 80 and 120?

```
: 1 df_male_glucose.loc[(df_male_glucose >= 80) & (df_male_glucose <= 120)].size
: 5303
```

```
1 df_male_glucose.size
11145
```

- (5303/11145) * 100 = 47.58 %

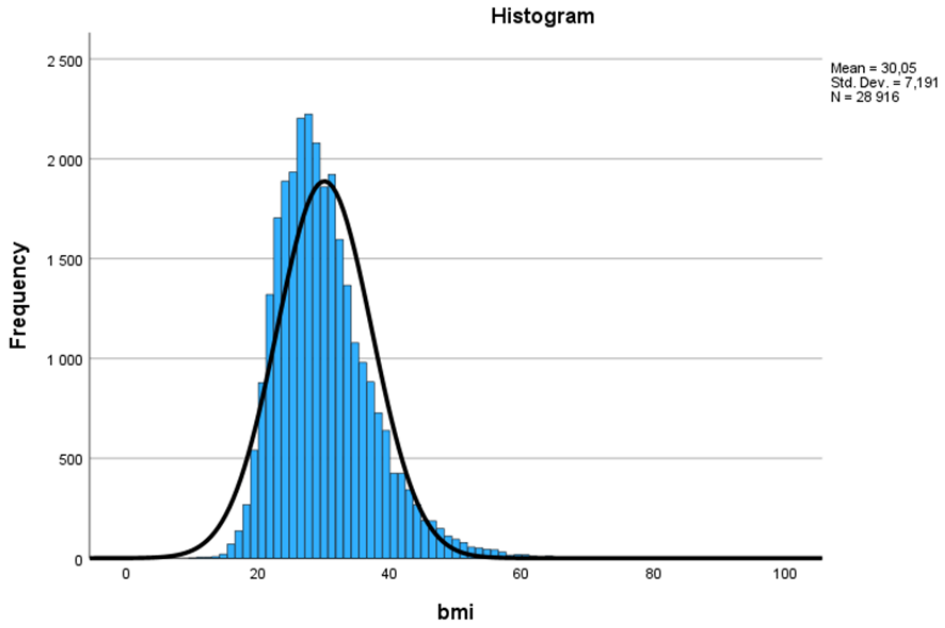
Stratified Sampling

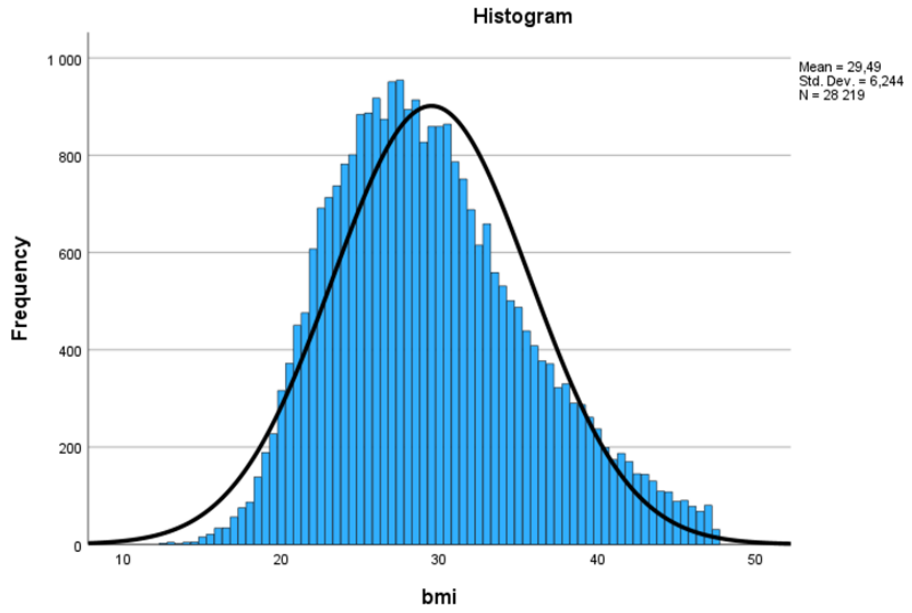
- Most normality tests work best for smaller sample size.
- Total number of entries for the avg glucose levels for males are 11145 (quite a large dataset!)
- Of those 11145, 60% have avg glucose level <= 100, 22% have avg glucose level between 101 and 140 and 18% have avg glucose level > 140.
- We took a stratified sample of 100 entries with the same corresponding percentages.
- Shapiro-Wilk gave a p-value of `1.0757815749329325e-09`

Null Hypothesis

BMI dataset is taken from a population that follows a normal distribution for both Male and Female.

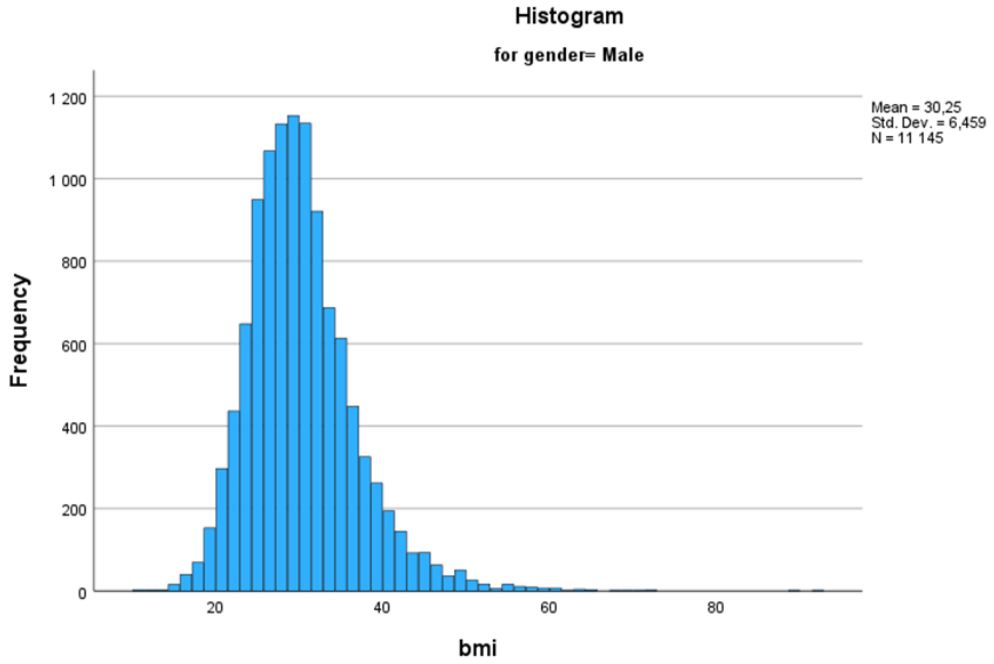
BMI histogram [With/without outliers]

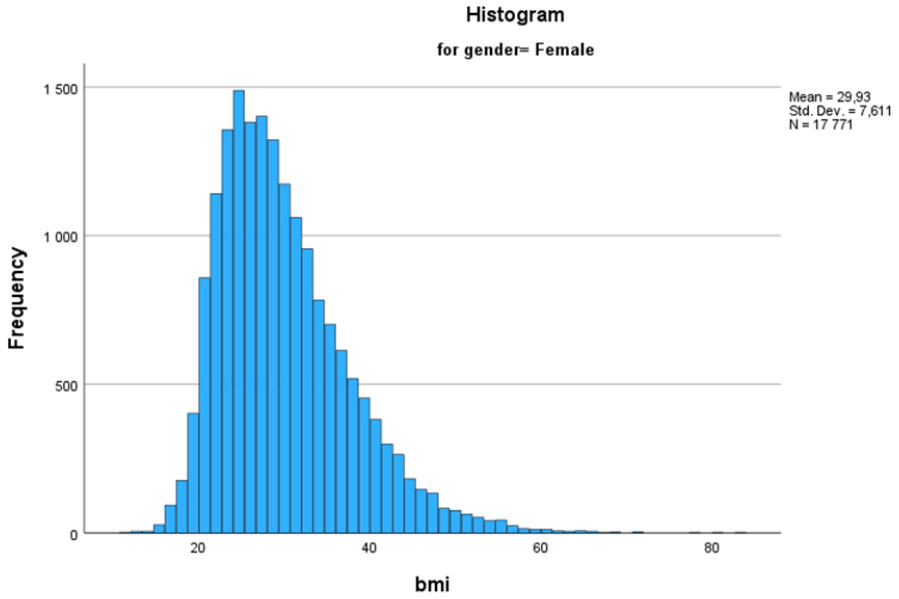




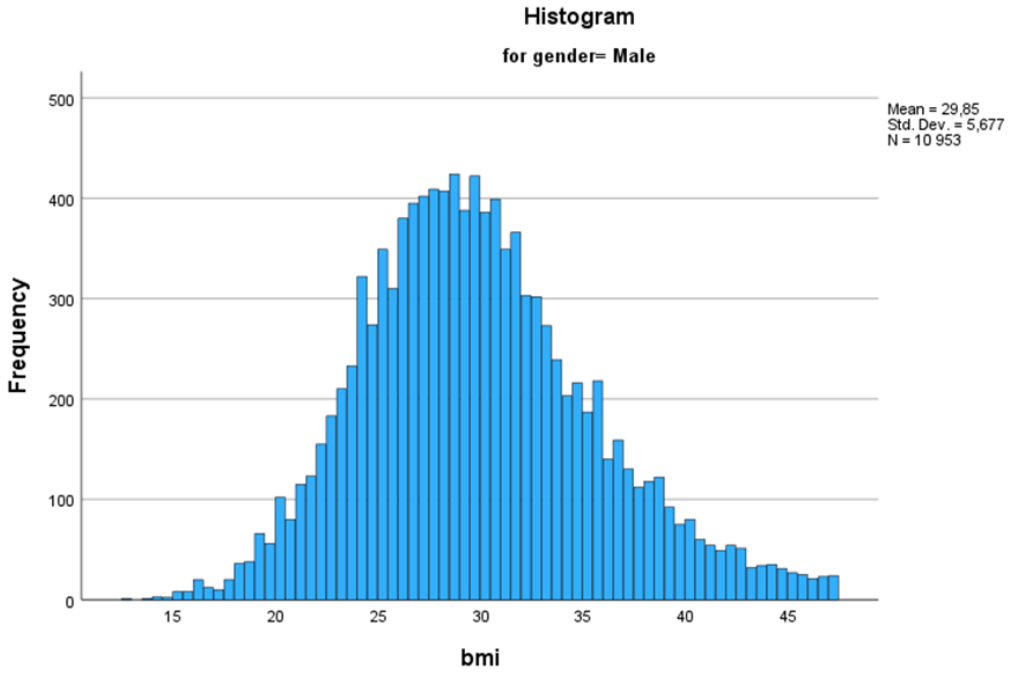
Ideally for a normal distribution this Histogram should look symmetric around the mean of the distribution, in this case, this distribution appears to be skewed to the right [Positively skewed] but close to normality

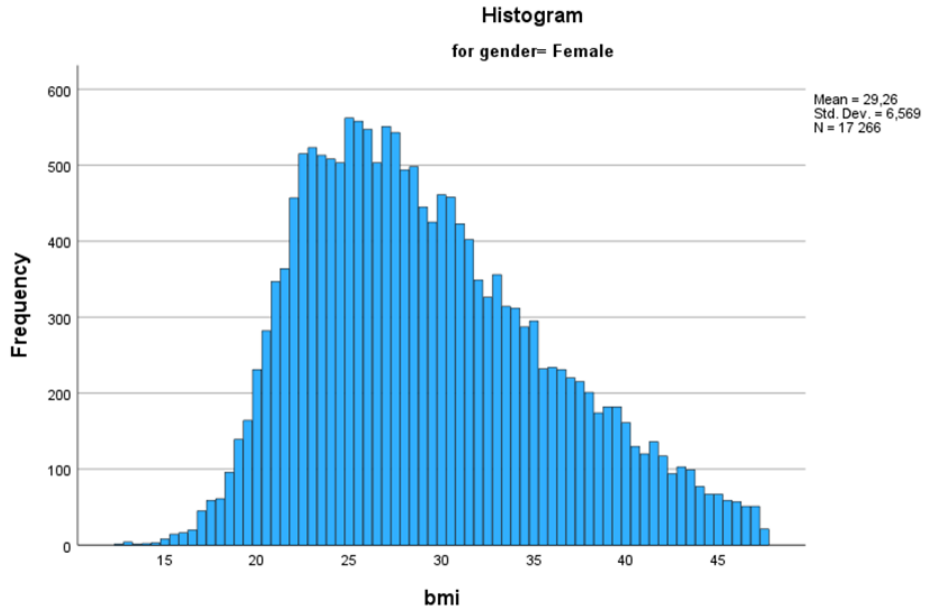
BMI histogram for Male/Female [With Outliers]



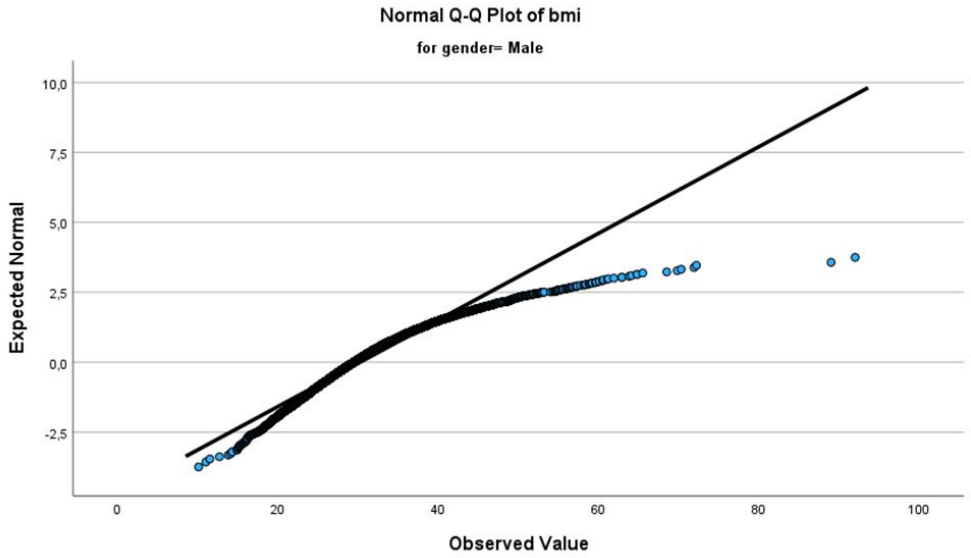


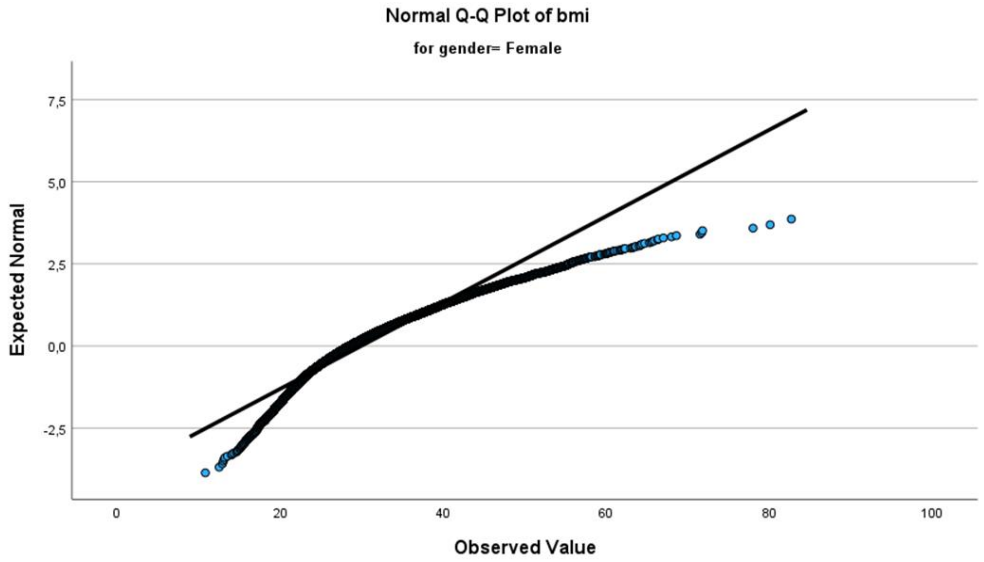
BMI histogram for Male/Female [Without Outliers]





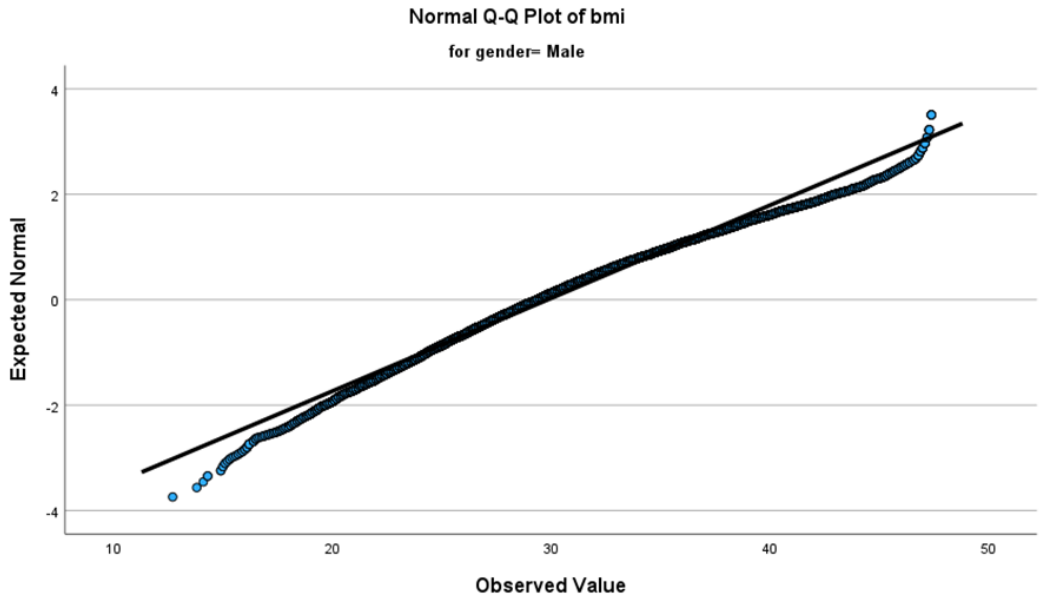
BMI QQ plot for Male and Female [With Outliers]

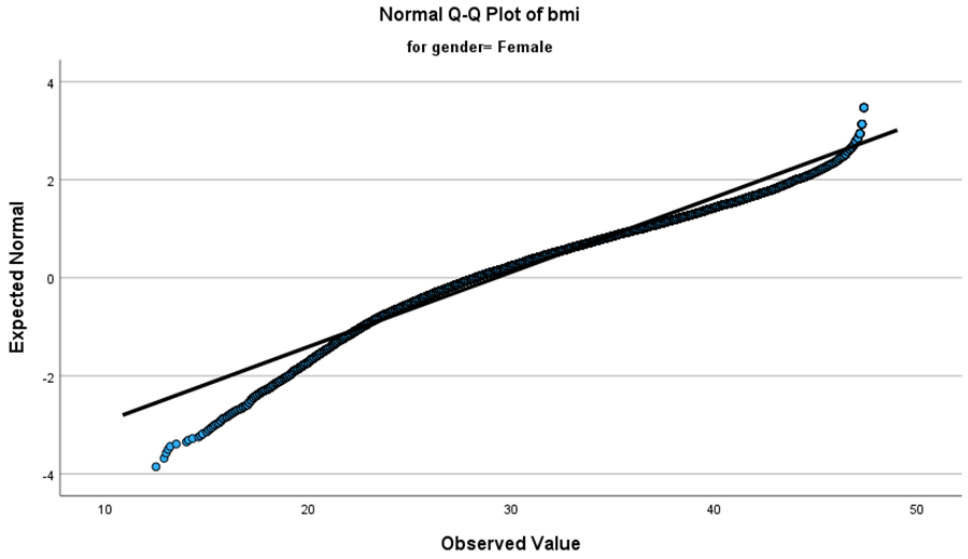




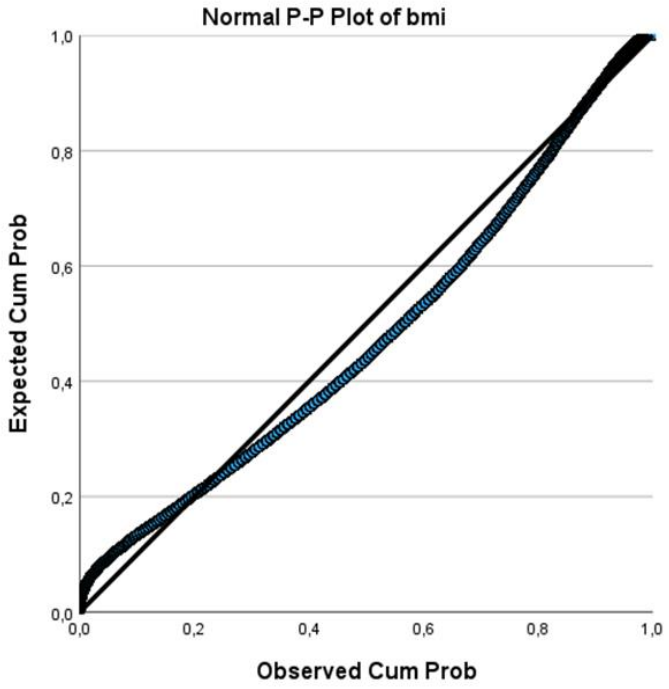
For the QQ plot, the data for BMI are plotted against a theoretical normal distribution in a way that the points should form an approximate straight line, departure from this straight line indicates departure from normality, we are seeing here a points diverging from the line above and below with some outliers lying away from the line

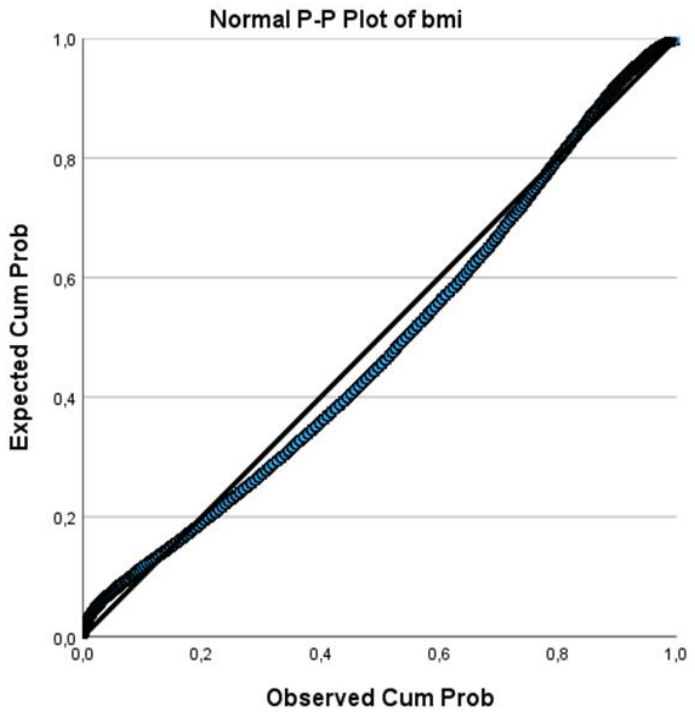
BMI QQ plot for Male and Female [Without Outliers]





BMI PP plot [with/Without Outliers]





Descriptive Statistics for BMI [With /without outliers]

Descriptives

	gender			Statistic	Std. Error
bmi	Female	Mean		29,93	,057
		95% Confidence Interval for Mean	Lower Bound	29,81	
			Upper Bound	30,04	
		5% Trimmed Mean		29,44	
		Median		28,50	
		Variance		57,931	
		Std. Deviation		7,611	
		Minimum		11	
		Maximum		83	
		Range		72	
		Interquartile Range		10	
		Skewness		1,044	,018
		Kurtosis		1,589	,037
	Male	Mean		30,25	,061
		95% Confidence Interval for Mean	Lower Bound	30,13	
			Upper Bound	30,37	
		5% Trimmed Mean		29,90	
		Median		29,50	
		Variance		41,715	
		Std. Deviation		6,459	
		Minimum		10	
		Maximum		92	
		Range		82	
		Interquartile Range		8	
		Skewness		1,151	,023
		Kurtosis		3,793	,046

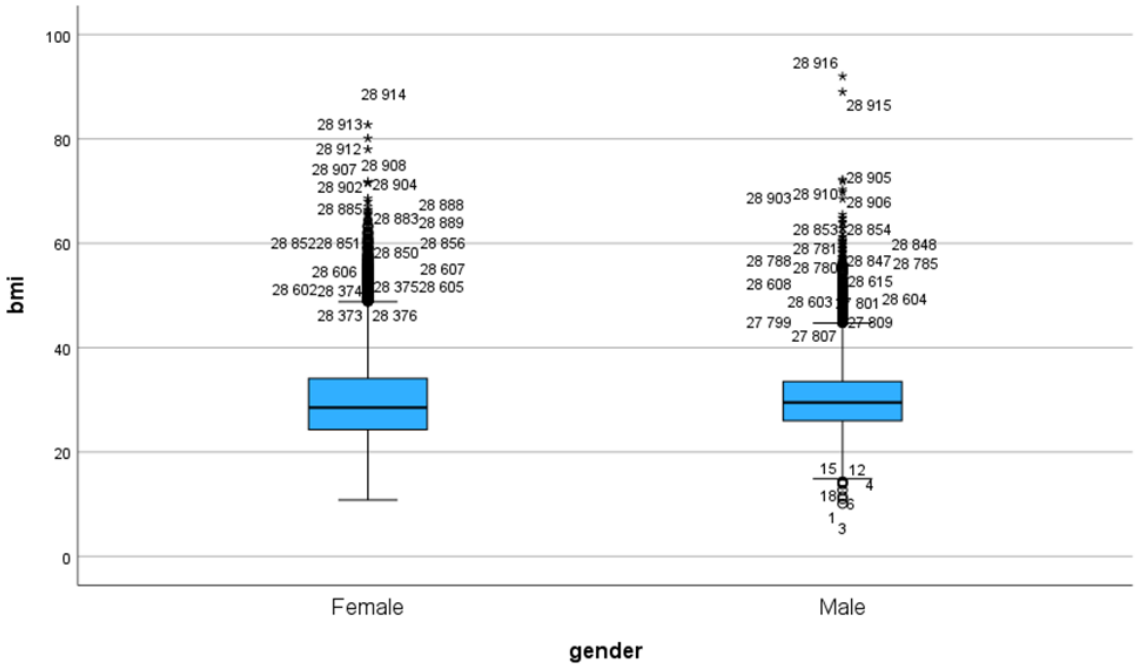
Descriptives

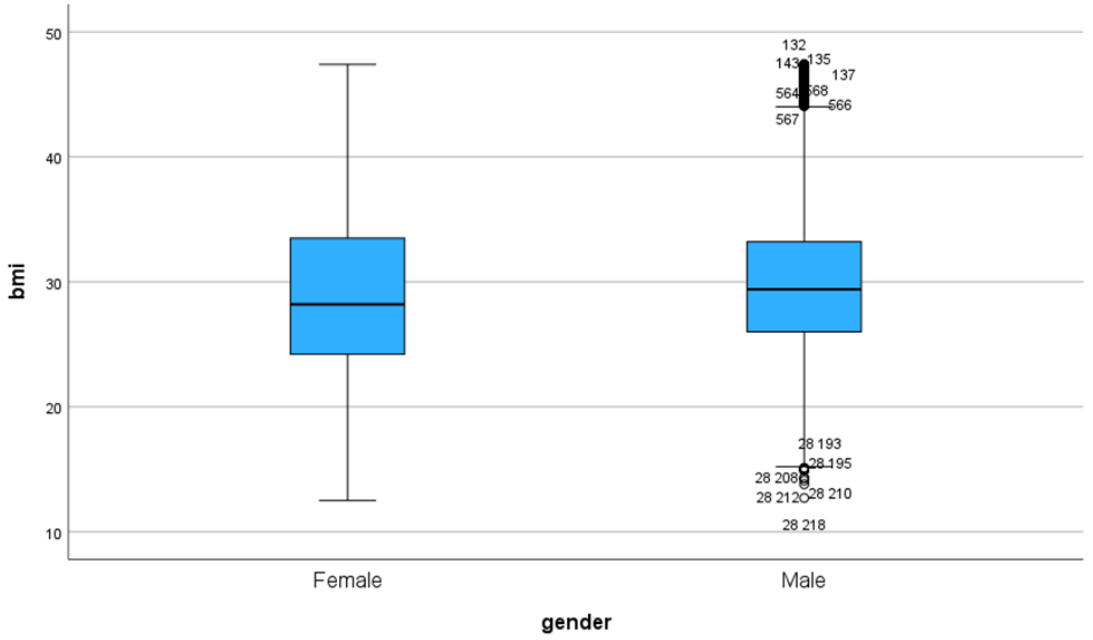
	gender			Statistic	Std. Error
bmi	Female	Mean		29,26	,050
		95% Confidence Interval for Mean	Lower Bound	29,16	
			Upper Bound	29,36	
		5% Trimmed Mean		29,01	
		Median		28,20	
		Variance		43,146	
		Std. Deviation		6,569	
		Minimum		13	
		Maximum		47	
		Range		35	
		Interquartile Range		9	
		Skewness		,541	,019
		Kurtosis		-,314	,037
	Male	Mean		29,85	,054
		95% Confidence Interval for Mean	Lower Bound	29,75	
			Upper Bound	29,96	
		5% Trimmed Mean		29,69	
		Median		29,40	
		Variance		32,226	
		Std. Deviation		5,677	
		Minimum		13	
		Maximum		47	
		Range		35	
		Interquartile Range		7	
		Skewness		,435	,023
		Kurtosis		,151	,047

Z scores [Outliers]
Skewness, M/50.04
Kurtosis, M/82.45
Skewness, F/58
Kurtosis, F/42.9

Z scores [w/o Outliers]
Skewness, M/18.91
Kurtosis, M/3.212
Skewness, F/28.47
Kurtosis, F/8.48

BMI Box plot for gender Male and Female [With/without outliers]





Kolmogorov/Smirnov normality test with/without outliers

Tests of Normality

		Kolmogorov-Smirnov[a]		
	gender	Statistic	df	Sig.
bmi	Female	,078	17771	<,001
	Male	,069	11145	<,001

a. Lilliefors Significance Correction

Tests of Normality

		Kolmogorov-Smirnov[a]		
	gender	Statistic	df	Sig.
bmi	Female	,065	17266	<,001
	Male	,043	10953	<,001

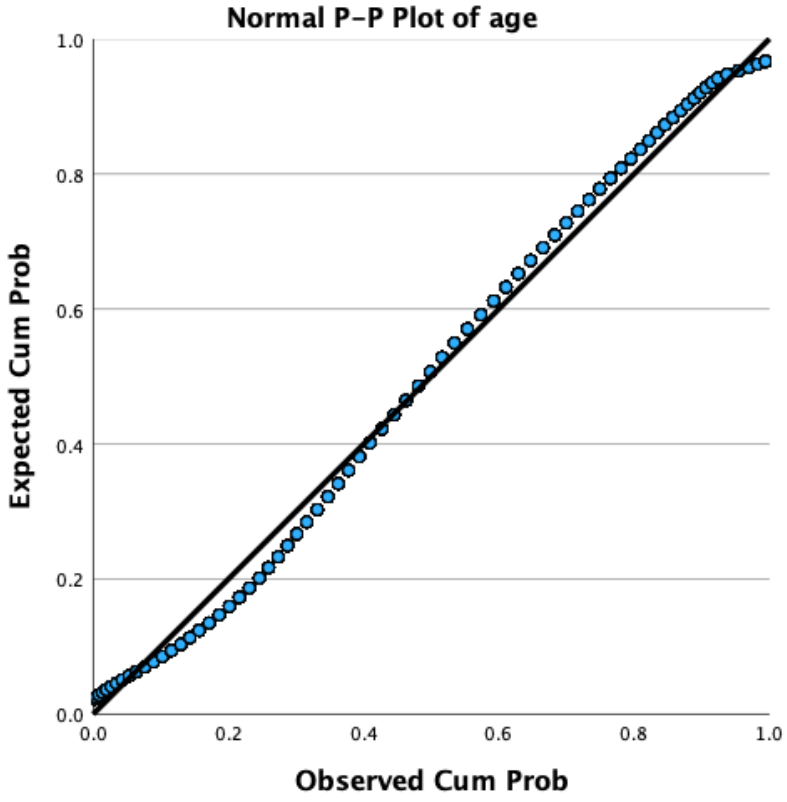
a. Lilliefors Significance Correction

Variable: Age

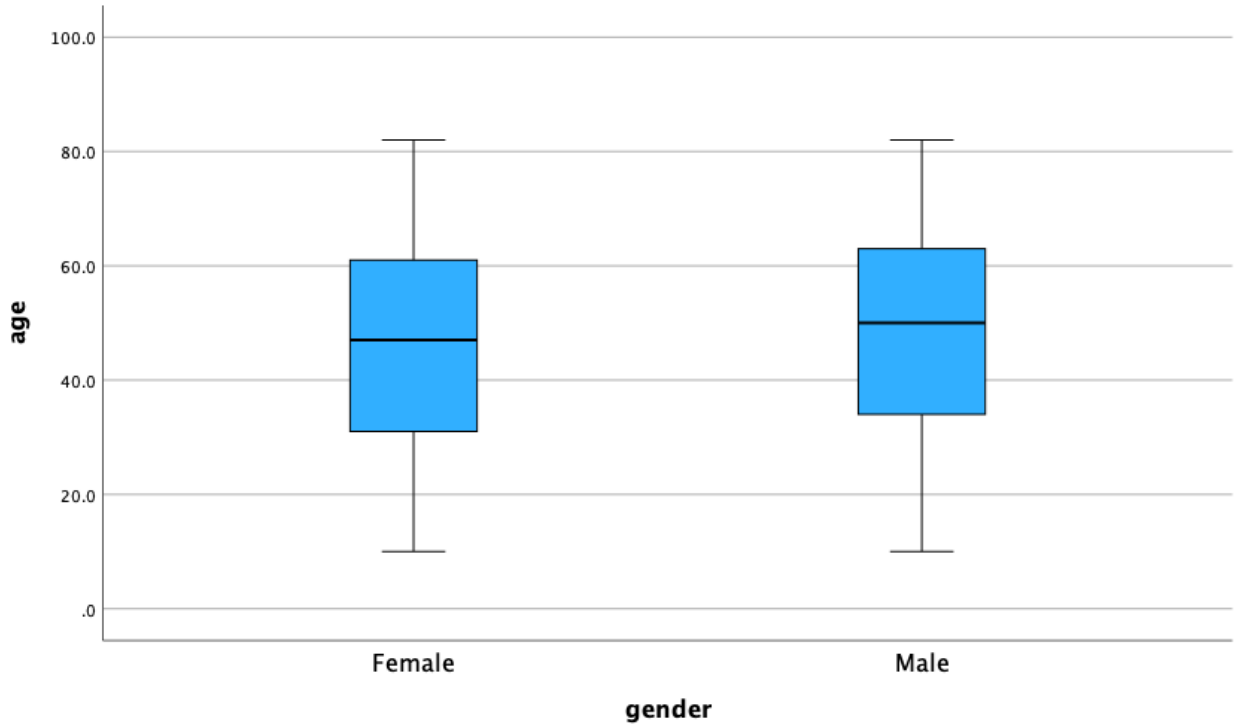
Alternative Hypothesis:

Age is not normally distributed in the Stroke dataset.

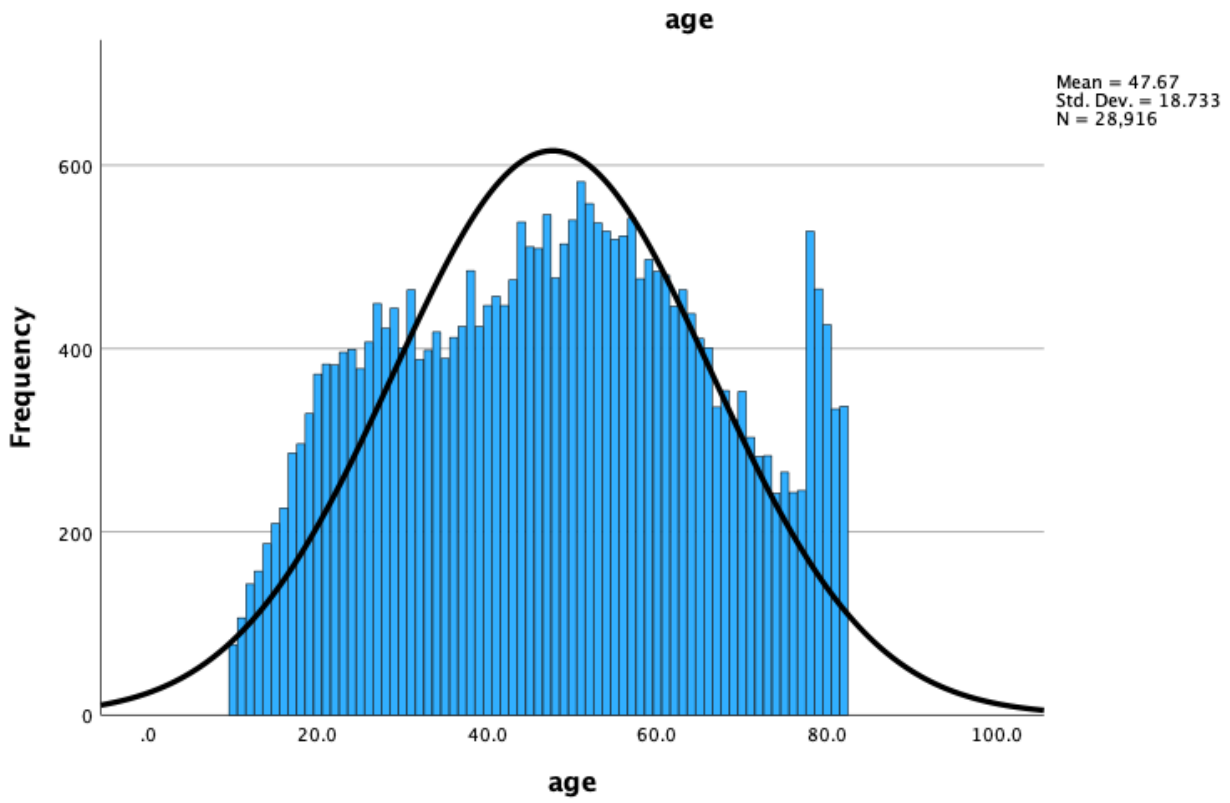
P-P plot for Age

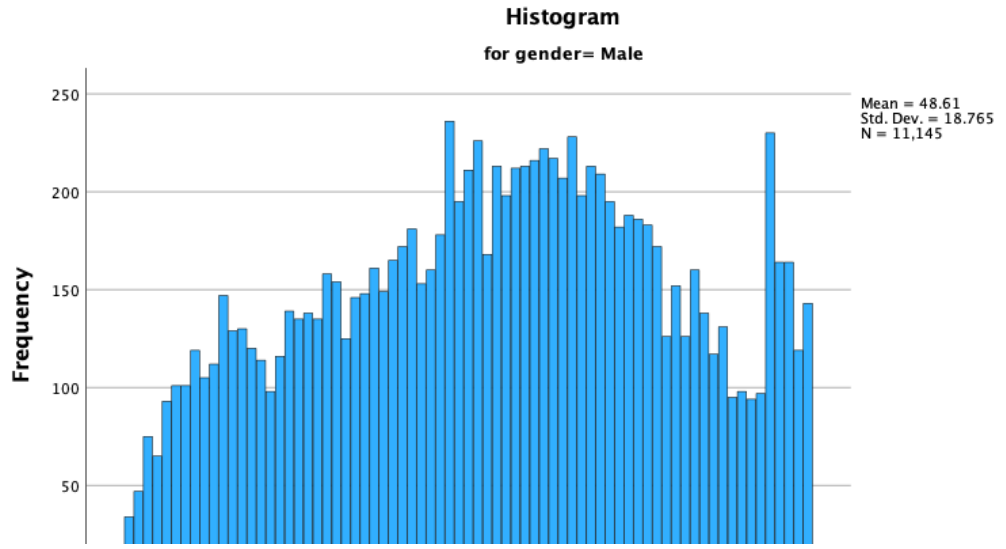


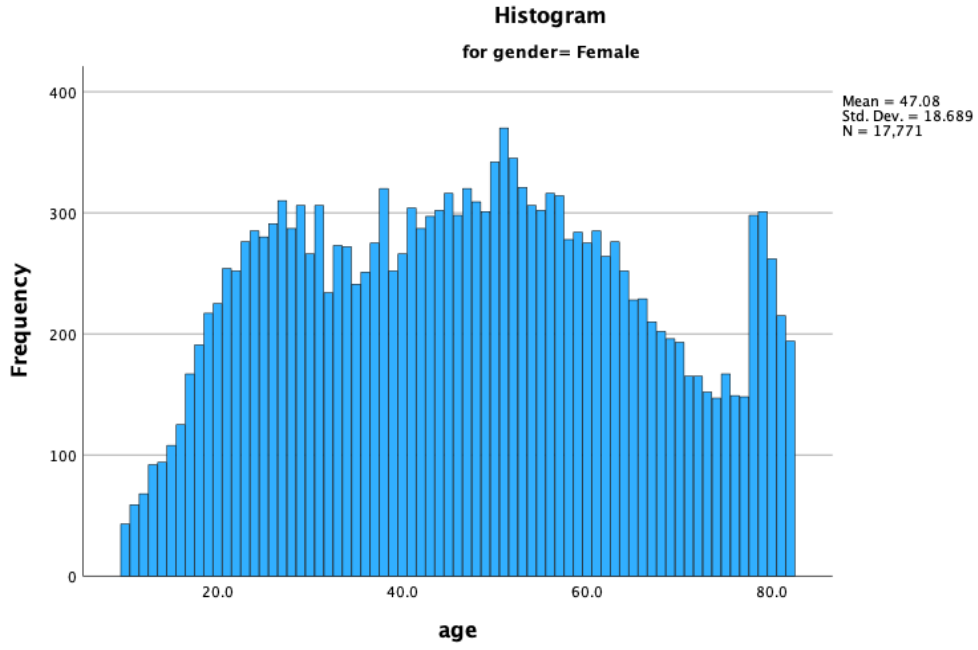
Box plot for Age



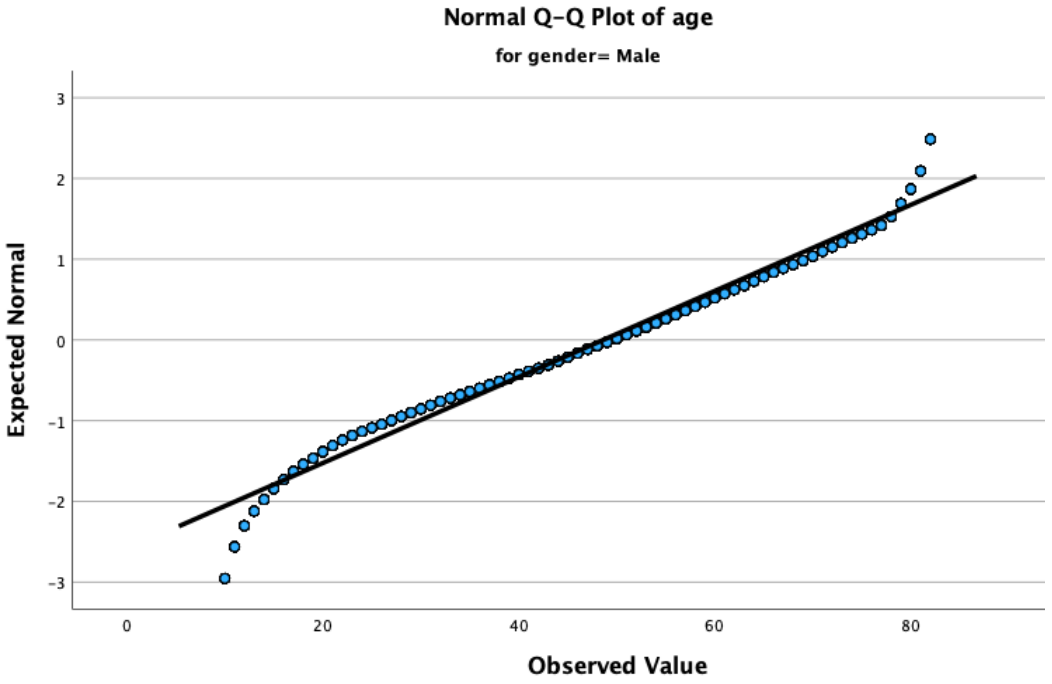
Histogram for Age

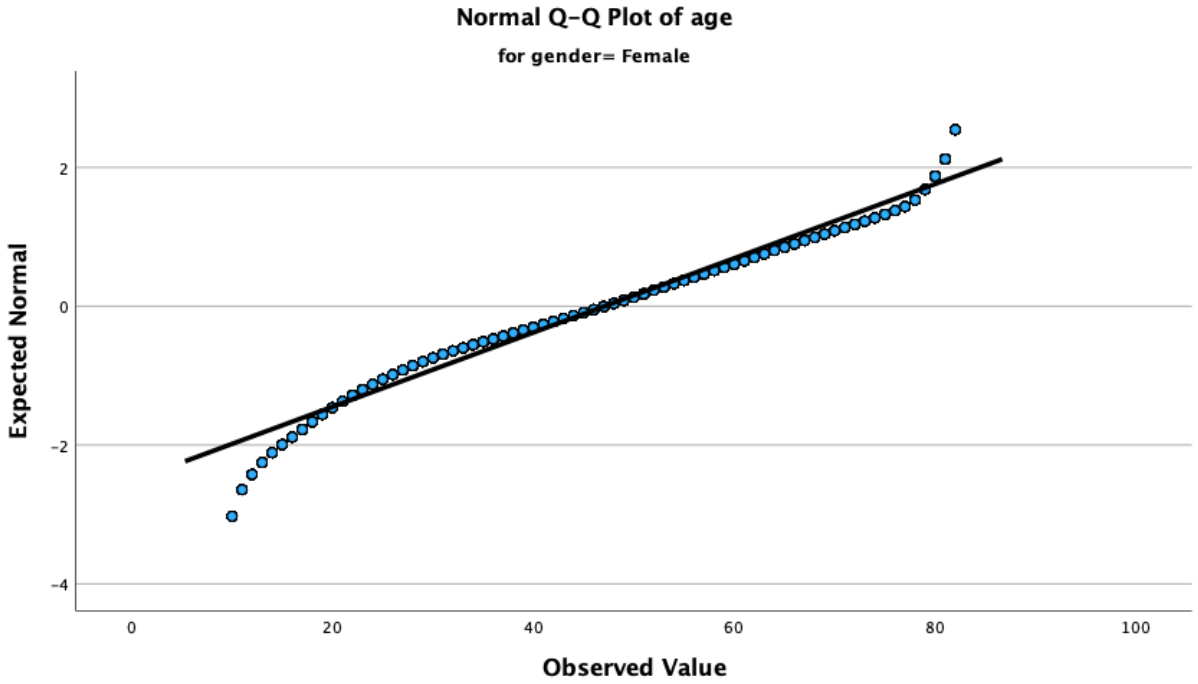






Q-Q plots for Age





Statistics

		age	gender
N	Valid	28916	28916
	Missing	0	0
Mean		47.668	
Std. Error of Mean		.1102	
Median		48.000	
Mode		51.0	
Std. Deviation		18.7327	
Skewness		-.003	
Std. Error of Skewness		.014	
Kurtosis		-.968	
Std. Error of Kurtosis		.029	
Minimum		10.0	
Maximum		82.0	

Tests of Normality

		Kolmogorov-Smirnov[a]		
	gender	Statistic	df	Sig.
age	Female	.059	17771	<.001
	Male	.042	11145	<.001

a. Lilliefors Significance Correction

Descriptives

	gender			Statistic	Std. Error
age	Female	Mean		47.079	.1402
		95% Confidence Interval for Mean	Lower Bound	46.804	
			Upper Bound	47.354	
		5% Trimmed Mean		47.013	
		Median		47.000	
		Variance		349.280	
		Std. Deviation		18.6890	
		Minimum		10.0	
		Maximum		82.0	
		Range		72.0	
		Interquartile Range		30.0	
		Skewness		.070	.018
		Kurtosis		-.983	.037
	Male	Mean		48.608	.1777
		95% Confidence Interval for Mean	Lower Bound	48.259	
			Upper Bound	48.956	
		5% Trimmed Mean		48.765	
		Median		50.000	
		Variance		352.117	
		Std. Deviation		18.7648	
		Minimum		10.0	
		Maximum		82.0	
		Range		72.0	
		Interquartile Range		29.0	
		Skewness		-.119	.023
		Kurtosis		-.914	.046

Z-scores:

- Skewness: Female: 0.070/0.018 = 3.89
- Kurtosis: Female: -0.983/0.037 = -26.57
- Skewness: Male: -0.119/0.023 = -5.17
- Kurtosis: Male: -0.914/0.046 = -19.87

All of the Z-scores are greater than +/-1.96. From this, we can assume that our age (variable) data is approximately not normal.

Although the visual tests such as the Histogram, Box plots, Q-Q and P-P plots show that our data could have some normality, the statistical tests such as Kolmogorov-Smirnov, Z-scores strongly supports that the age variable in our data is not normally distributed.

Conclusion

- Our variables under investigation (avg glucose levels, age and bmi) do not follow a normal distribution according to the statistical tests.
- Some variables, however, like bmi and age show a close proximity to normality when the visual tests are used.
- There can be a discrepancy between the visual and statistical tests when the sample size is large.
- Shapiro-wilk for example doesn't work for a large sample size in SPSS and it might be unreliable in python.
- According to medical articles, bmi in both genders is always positively skewed and only the avg glucose levels while fasting can be normal.
- Sampling the age variable might make the statistical tests converge to normality (needs to be investigated).