Exploratory Data Analysis on the Diabetes Dataset

Markus Saint-Pettersen

Introduction

The diabetes dataset is a CSV file containing medical information about 768 patients being screened for diabetes mellitus. All the patients in the dataset are Pima Indians, an ethnic group native to Arizona with historically high rates of diabetes. Specifically, the dataset only contains information about females between the ages of 21 and 81. The dataset was collected by the US National Institute of Diabetes and Digestive and Kidney Diseases in or around 1988.

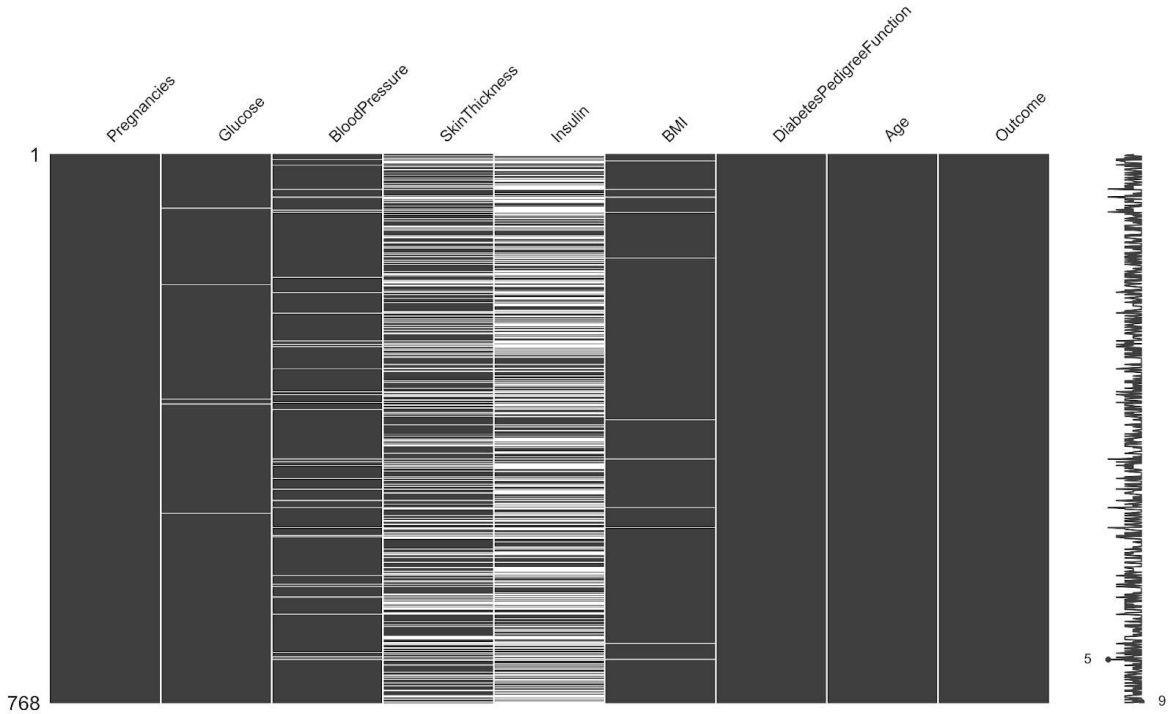
There are 9 columns containing a variety of information such as age, number of pregnancies and body mass index (BMI). Finally, an outcome variable, either 1 or 0, shows the diabetes status of the patient. The 'DiabetesPedigreeFunction' column is a measure of family history of diabetes.

Data Cleaning

The initial data was in relatively good condition and did not require extensive cleaning. Missing values were represented by 0s throughout the dataset. Care had to be taken when replacing these missing values with NaN because not all 0s represented missing values. Zeros were used in the 'Outcome' column to indicate that the patient did not have diabetes, and in the 'Pregnancies' column if the patient had had no pregnancies.

Missing Data

After replacing zeros with NaN in the 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin' and 'BMI' columns, the dataset had 652 missing values. This corresponded to 9.43% of the total data. The majority of the missing data were in the 'SkinThickness' and 'Insulin' columns (227 and 374 entries respectively). The missing values are displayed graphically below:



Before imputing the missing data, any rows containing more than 3 missing values were dropped. This ensures that the remaining data is more reliable. 7 rows were dropped for containing too many missing values, leaving 761 rows.

Summary of imputed values

The remaining missing values were imputed by taking the average of the non-missing values for each category (median or mean). A summary of the missing values and imputations are shown in the table below:

Column	Column data type	Number of missing values	Imputed value	Notes
Glucose	Integer	5	122	Mean
BloodPressure	Integer	28	72	Mean
SkinThickness	Integer	220	32	Mean
Insulin	Integer	327	125	Median
BMI	Float	4	32.46	Mean

Data organisation

Most of the values in the dataset are quantitative. To aid with analysis, categorical values can be derived from ranges of discrete values. This will allow comparisons to be made between different groups.

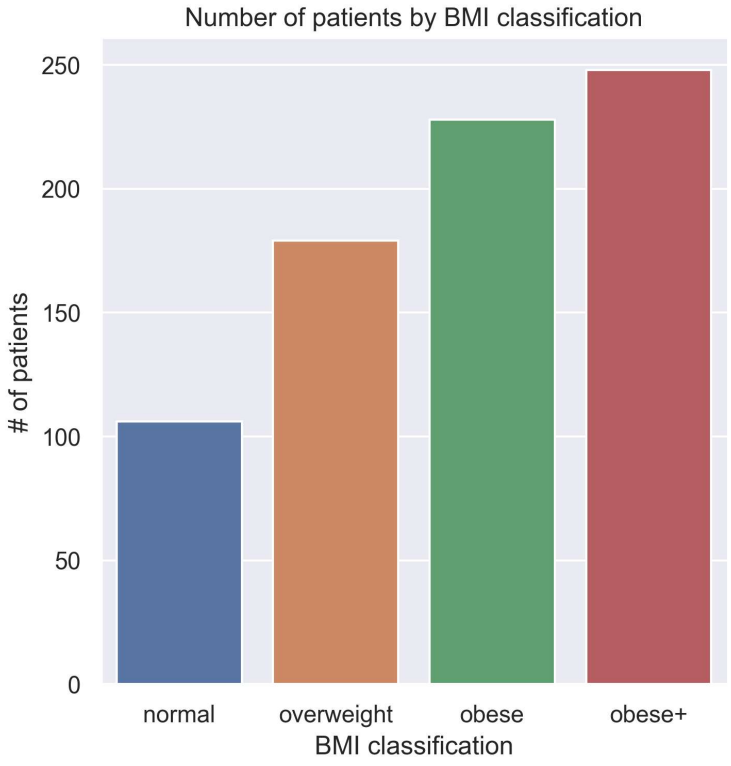
BMI

Body mass index (BMI) is calculated by dividing a person's weight (in kilograms) by their height (in metres) squared. The World Health Organisation (WHO) classifies adults as underweight (BMI < 18.5), overweight (BMI ≥ 25) or obese (BMI ≥ 30)[1].

However, the data for BMI did not fit evenly into these categories so the following changes were made:

Category	BMI ranges (kg/m^2)	Number of patients
Normal	Below 25	106
Overweight	Between 25 and 30	179
Obese	Between 30 and 35	228
Obese+	35 and over	248

The categories are displayed visually below:



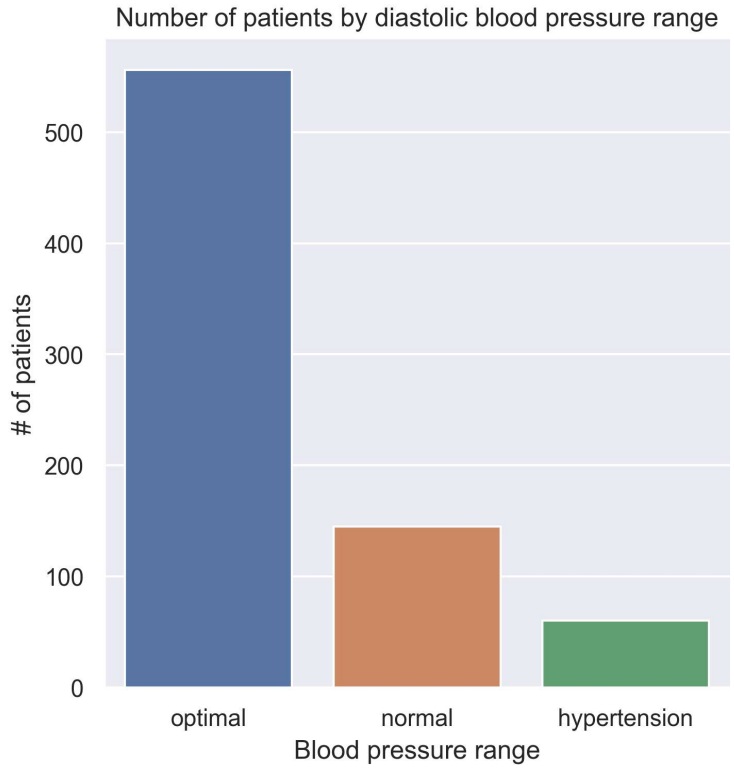
[1] SuRF Report 2, 2005; page 22; https://apps.who.int/iris/bitstream/handle/10665/43190/9241593024_eng.pdf

Blood pressure

Blood pressure is the pressure exerted by blood on the blood vessel walls. In the dataset, blood pressure refers to diastolic blood pressure, the minimum pressure between two heart beats. The European Society of Cardiology and the European Society of Hypertension categorise blood pressure in the following way (simplified below)[2]:

Category	BP range (mmHg)	Number of patients
Optimal	Below 80	556
Normal	Between 80 and 89	145
Hypertension	Above 89	60

The categories are displayed visually below:



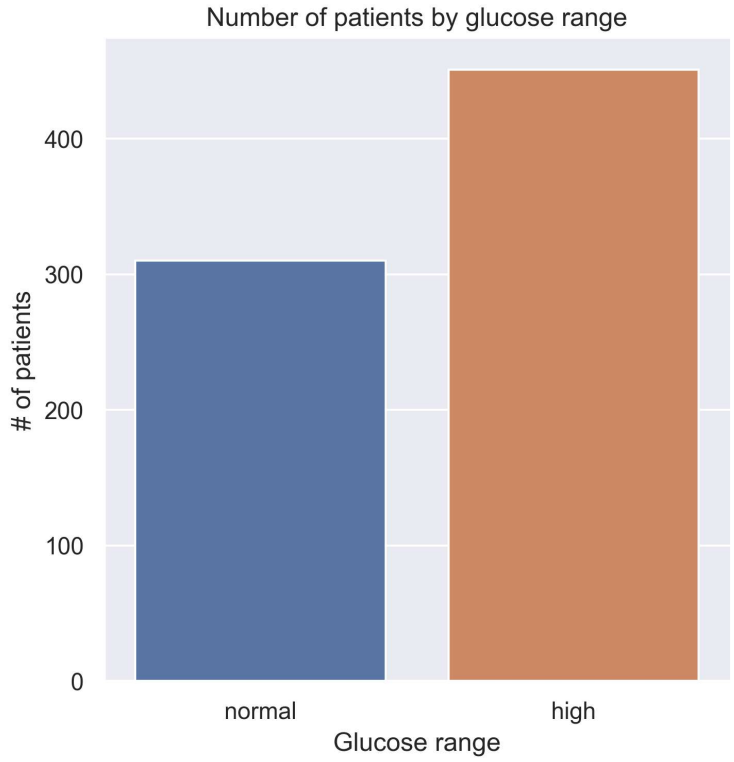
[2] ESC/ESH Guidelines, 2018; shown in Table 3; https://academic.oup.com/eurheartj/article/39/33/3021/5079119

Glucose

Glucose is the plasma glucose level tested two hours after eating. Normal glucose levels are given as between 70 and 110 mg/dL[3]. Only 11 patients had 'low' glucose levels, so the patients were divided into either normal or high categories.

Category	Plasma glucose range (mg/dL)	Number of patients
Normal	110 and below	310
High	Above 110	451

The categories are displayed visually below:



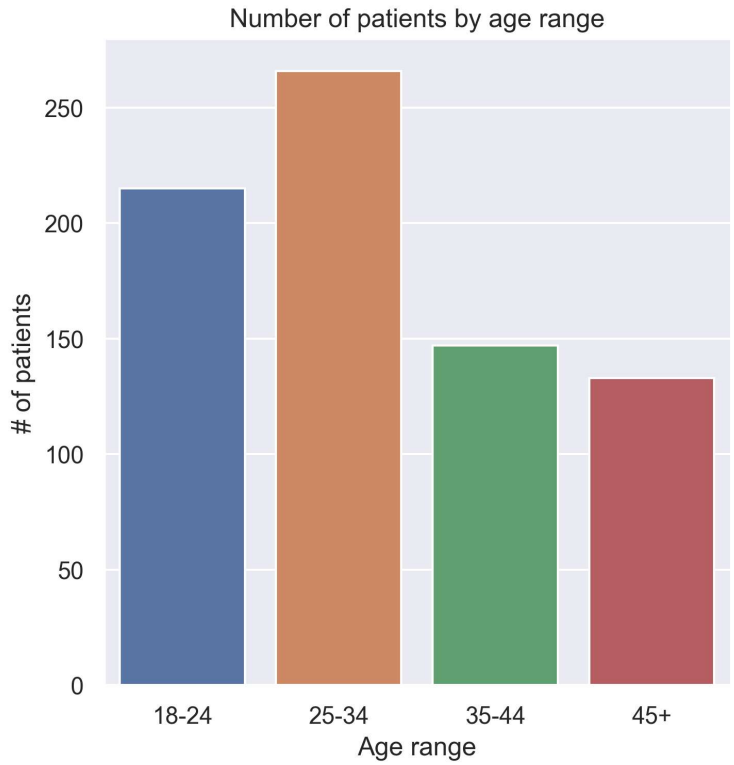
[3] Retrieved from http://www.bloodbook.com/ranges.html

Age

Patients can also be divided into age ranges.

Age range	Number of patients
18-24	215
25-34	266
35-44	147
45+	133

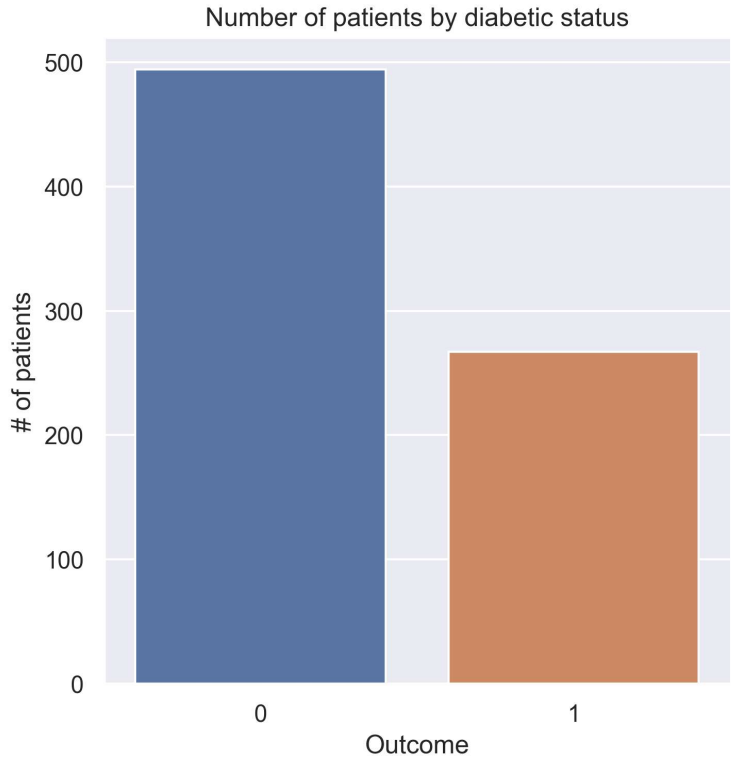
The categories are displayed visually below:



Data Stories and Visualisation

What are the characteristics of patients who are diabetic? In what ways are they different from patients who are not diabetic?

First, what is the general rate of diabetes in the dataset?



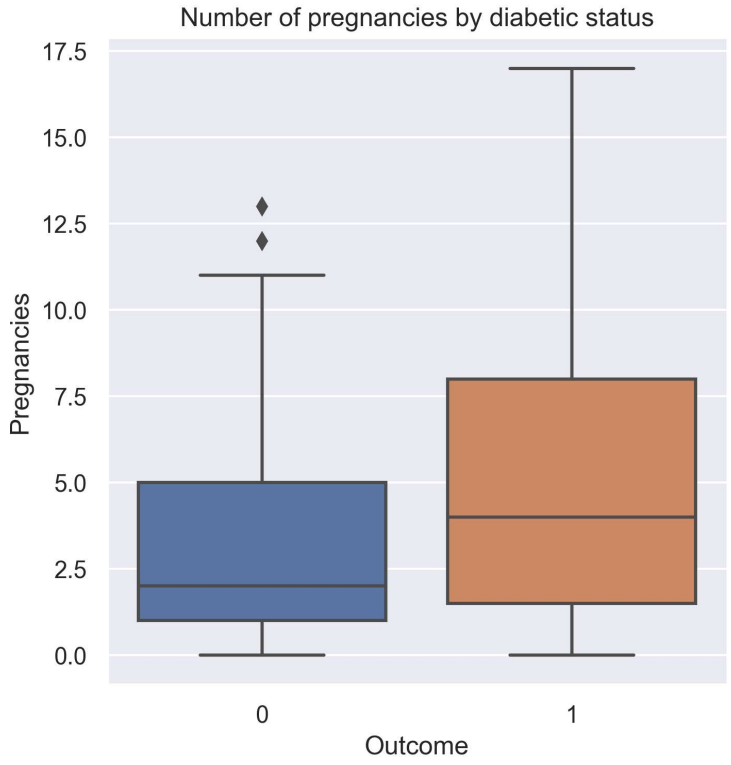
Overall, 267 out of 761 patients were diabetic, or 35.1%.

Diabetic outcome

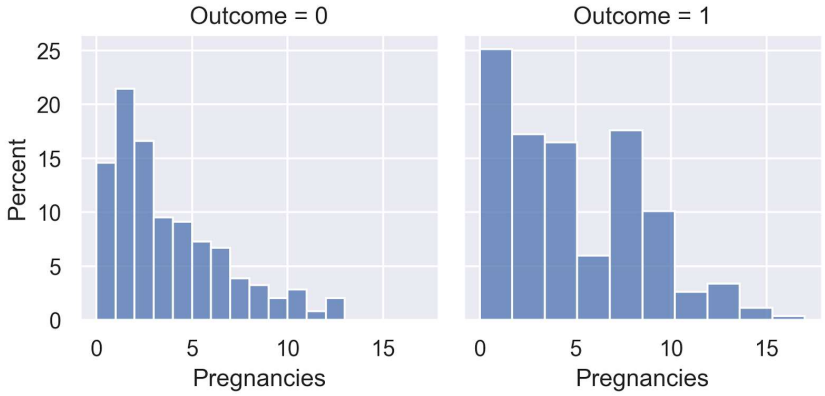
How is the rate of diabetes influenced by other variables in the dataset, and which categories have the highest risk for diabetes?

Number of pregnancies

Is diabetic outcome affected by the number of pregnancies? Below the boxplot shows the distribution of the number of pregnancies for each outcome.



Non-diabetics tend to have fewer pregnancies than diabetics (median: 2 vs. 4; mean: 3.29 vs. 4.85). However, the number of pregnancies is a lot more variable in diabetics. The interquartile range for non-diabetics is smaller (4.0 vs. 6.5). The difference in the means of each sample is significant based on Welch's t-test.

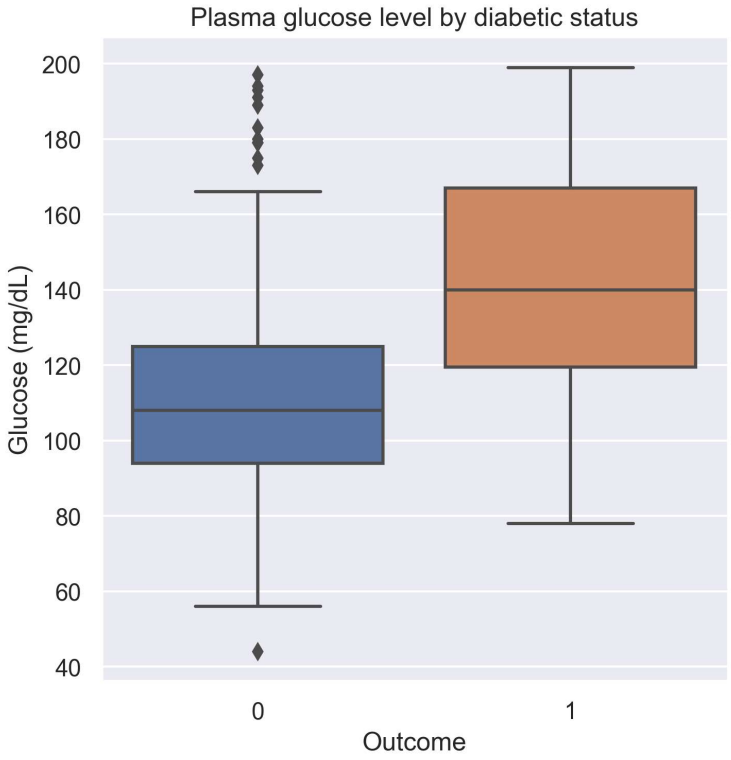


The histograms above also show the distribution. Both groups are right skewed, but the skewness is greater in the non-diabetic group. A large percentage (25%) of diabetics have had zero pregnancies. This

suggests that there may be missing values in the ‘Pregnancies’ column too. However, it would be difficult to determine which are real zeros and which are missing values.

Plasma glucose level

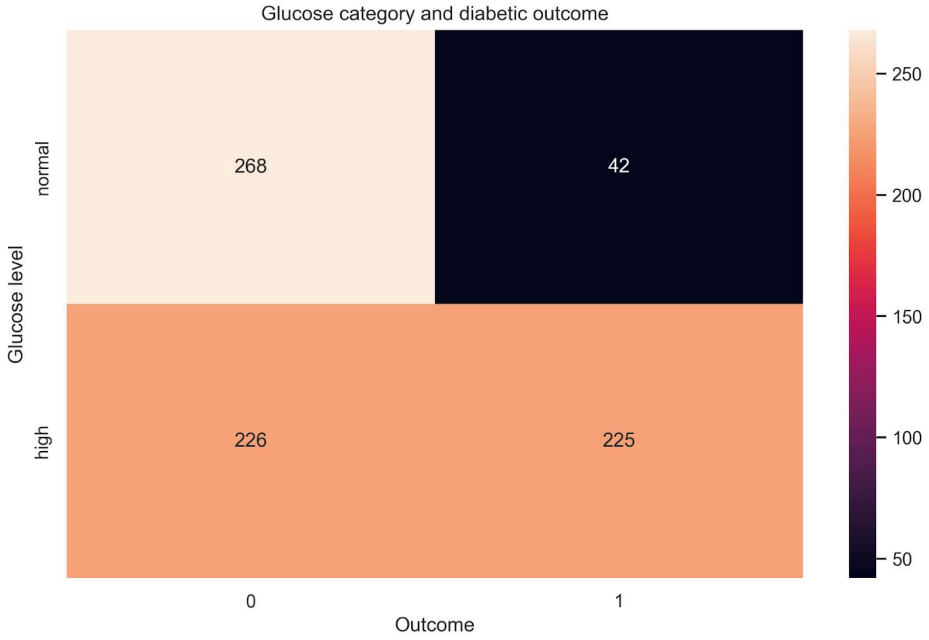
It is expected that patients with diabetes would have higher glucose levels than non-diabetics.



Non-diabetics tend to have lower plasma glucose levels than diabetics (median 108 vs. 150 mg/dL; mean 110.94 vs. 142.26 mg/dL). Similar to pregnancies, diabetics also show more variation. Interquartile range for non-diabetics is 31 mg/dL and for diabetics it is 48 mg/dL. The difference in means was also significant based on Welch’s t-test. However, there are some extreme outliers in the non-diabetic group.

Relative risk of diabetes by glucose levels

It is possible to calculate the relative risk of having high glucose levels by comparing the diabetes rate in that group with the general rate for the whole dataset (35.1%).



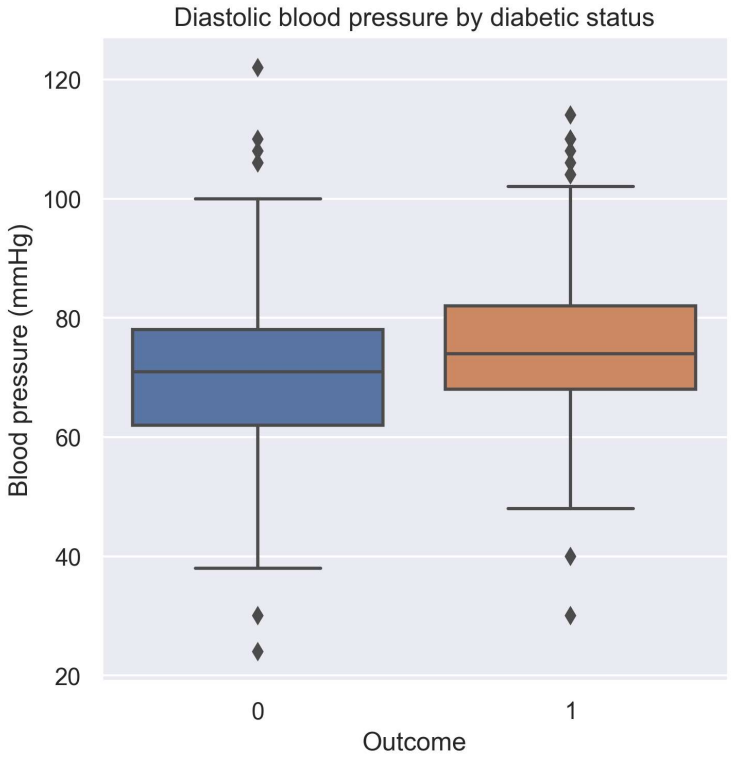
Very few patients with normal levels of plasma glucose were diabetic. Conversely, almost half of those with high levels were. This shows that half of patients with high glucose levels go on to develop diabetes.

	Positive cases	**Total cases**	**Percentage positive**	**Relative risk**
Normal	42	310	13.5%	0.39
High	225	451	49.9%	1.42

Patients with normal plasma glucose levels have a reduced risk of diabetes, while those with high levels are 1.42 times as likely to have diabetes than the general population.

Blood pressure

High blood pressure is generally used as a level to measure overall health, but can it be an indicator of diabetes?



The distribution for blood pressure is more equal between the two groups. Non-diabetics have slightly lower blood pressure (median 71 vs. 74 mmHg; mean 70.91 vs. 75.13 mmHg). Unlike previous comparisons, the interquartile ranges are very similar. For non-diabetics 16 mmHg and diabetics 14 mmHg.

Relative risk of diabetes by blood pressure

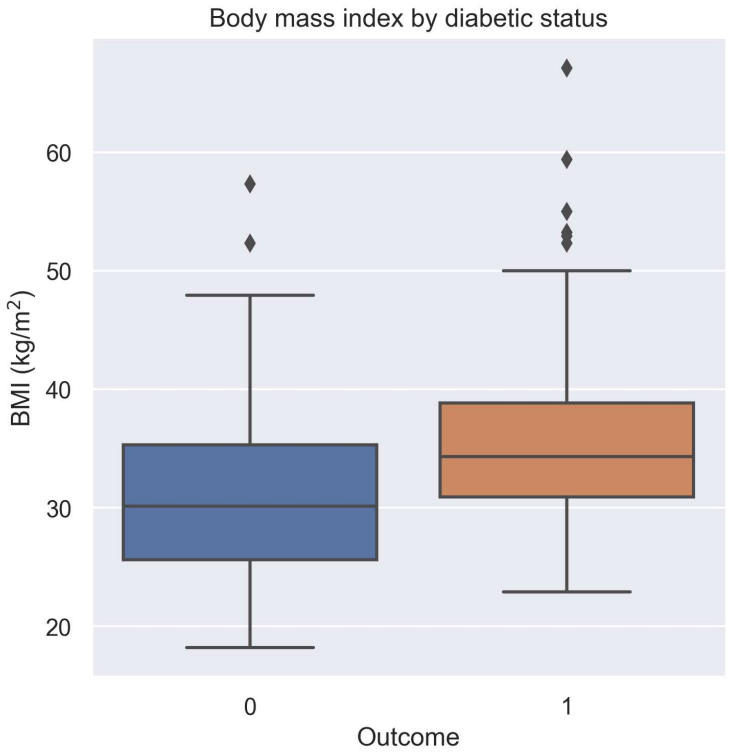
As before, it is possible to calculate relative risks based on the categories for blood pressure.

	Positive cases	**Total cases**	**Percentage positive**	**Relative risk**
Optimal	177	556	31.8%	0.91
Normal	61	145	42.1%	1.20
Hypertension	29	60	48.3%	1.38

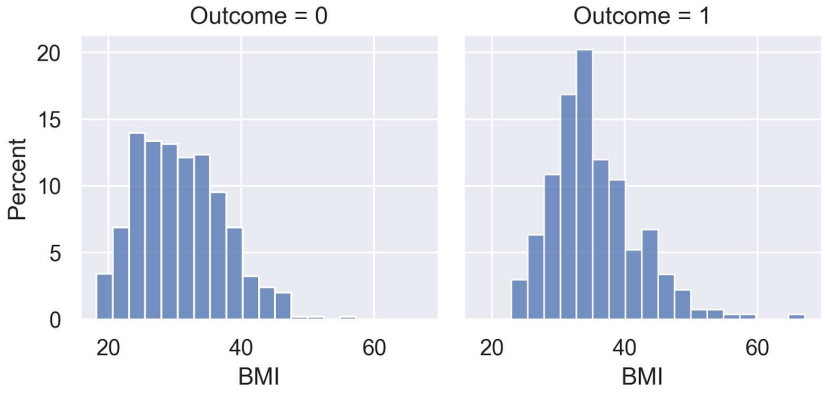
Optimal blood pressure is associated with a slightly lower risk of diabetes, while normal and hypertension have an increased risk, 1.20 and 1.38 times respectively.

BMI

Body mass index is also used as a general measure of health. Is it associated with diabetes in the dataset?



The distributions for BMI look similar to those for blood pressure. Non-diabetics tend to have lower BMIs than diabetics (median 30.15 vs. 34.35, mean 30.87 vs. 35.40). The distributions between the two groups are also fairly equal, but for diabetics the distribution is more skewed to the right. The difference in means was also statistically significant.



Relative risk of diabetes by BMI

The following table shows the relative risks of obesity for diabetes.

	Positive cases	Total cases	Percentage positive	Relative risk
Normal	7	106	6.6%	0.19
Overweight	40	179	22.3%	0.64
Obese	102	228	44.7%	1.28
Obese+	118	248	47.6%	1.36

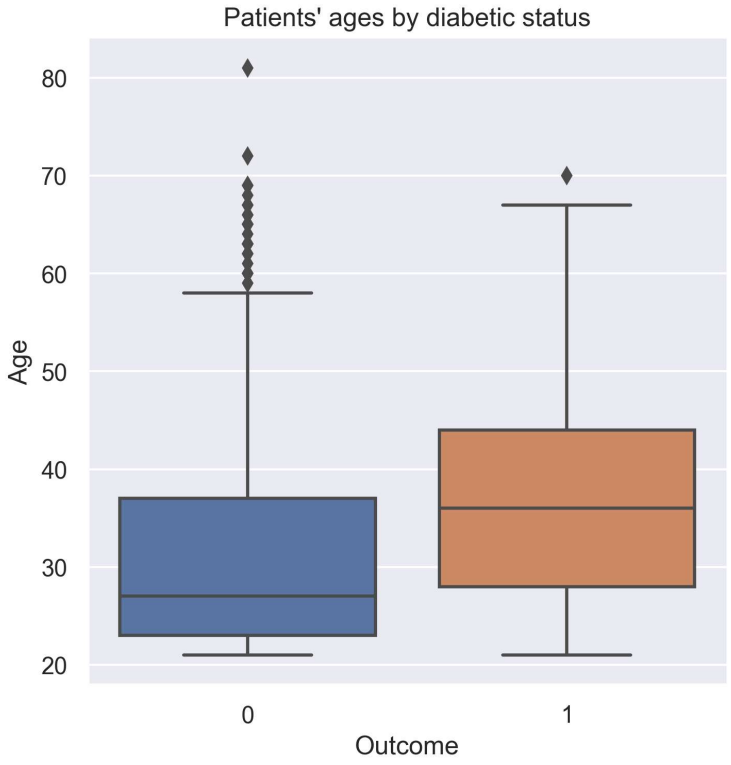
The relative risks increase with BMI. However, the difference between ‘Obese’ and ‘Obese+’ is quite small. A t-test shows that these two groups are not statistically different for outcome. The risks can be reanalysed with the obese groups combined into one.

	Positive cases	Total cases	Percentage positive	Relative risk
Obese*	220	476	46.2%	1.32

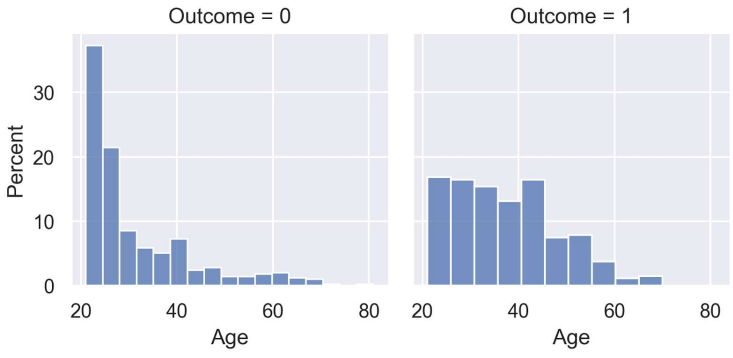
Normal or overweight patients have a reduced risk of diabetes. Obese patients are 1.32 times more likely to have diabetes.

Age

The final variable is age. A lot of diseases are associated with ageing. Are diabetics on average older than non-diabetics?



Non-diabetics tend to be quite a lot younger than diabetics (median 27 vs. 36; mean 31.3 vs. 37.1). This difference in means is statistically significant. The data for both groups have a similar distribution, but the data is more skewed for non-diabetics; participants tend to skew younger. However, some of the oldest patients are not diabetic. These histograms are similar to the histograms for pregnancies.



Relative risk of diabetes by age group

The same relative risks can be calculated for the different age groups.

	Positive cases	Total cases	Percentage positive	Relative risk
18 to 24	31	215	14.4%	0.41
25 to 34	94	266	35.3%	1.01
35 to 44	76	147	51.7%	1.47
45 and older	66	133	49.6%	1.41

Relative risk of diabetes is lowest for the youngest patients. Patients aged between 25-34 have about average risk of diabetes. After 35, there is a sharp increase of risk, but this risk stays constant after 45 (as shown by a t-test). This suggests that around 50% of the Pima population will develop diabetes, and the most common age range to do so is between 35 and 44 years old. It also suggests that patients without diabetes after 45 are not likely to develop diabetes in the future.

As with BMI, the risks can be reanalysed:

	Positive cases	Total cases	Percentage positive	Relative risk
35 and older	142	280	50.7%	1.45

This means the highest risk factor for diabetes is age, namely being over 35.

Summary of risk factors for diabetes

A patient with high risk of diabetes will likely fall into the following categories (from highest to lowest risk):

Characteristic	Level	Relative risk
Age	Older than 35	1.45
Glucose level	High	1.42
Blood pressure	Hypertension	1.38
BMI	Obese	1.32

In the dataset, 25 patients meet all these criteria. Of the 25 patients, 19 patients are diabetic (76%). This shows that other factors can also influence the outcome.

Conversely, a low risk patient would have the following characteristics:

Characteristic	Level	Relative risk
BMI	Normal	0.19
Glucose	Normal	0.39
Age	18-24	0.41
Blood pressure	Optimal	0.91

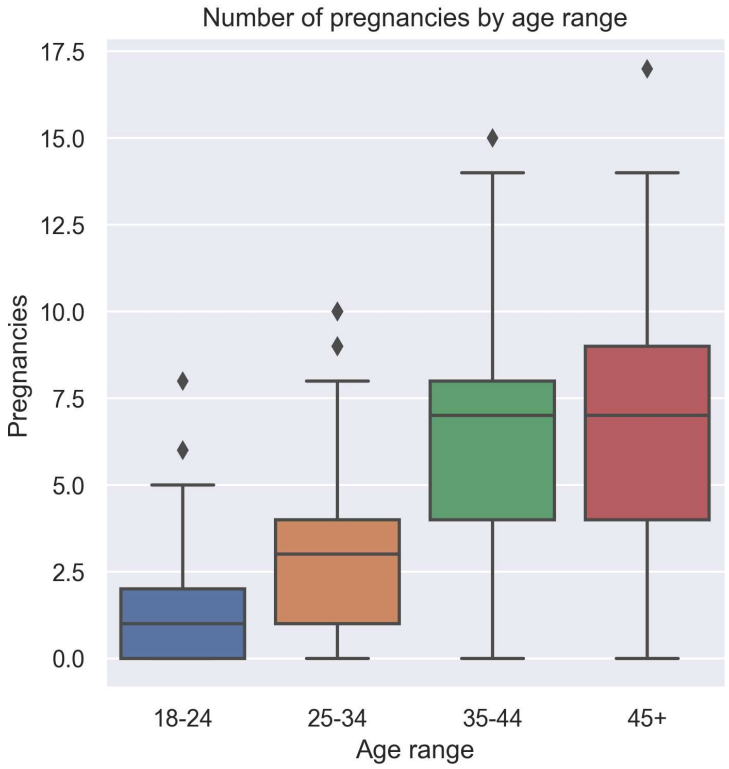
24 patients in the dataset meet these criteria, from these, 1 patient is diabetic (4%).

Other insights from the dataset

Although the main purpose of this analysis was to investigate the risk factors of diabetes, the dataset contained a lot of information that could provide more insight.

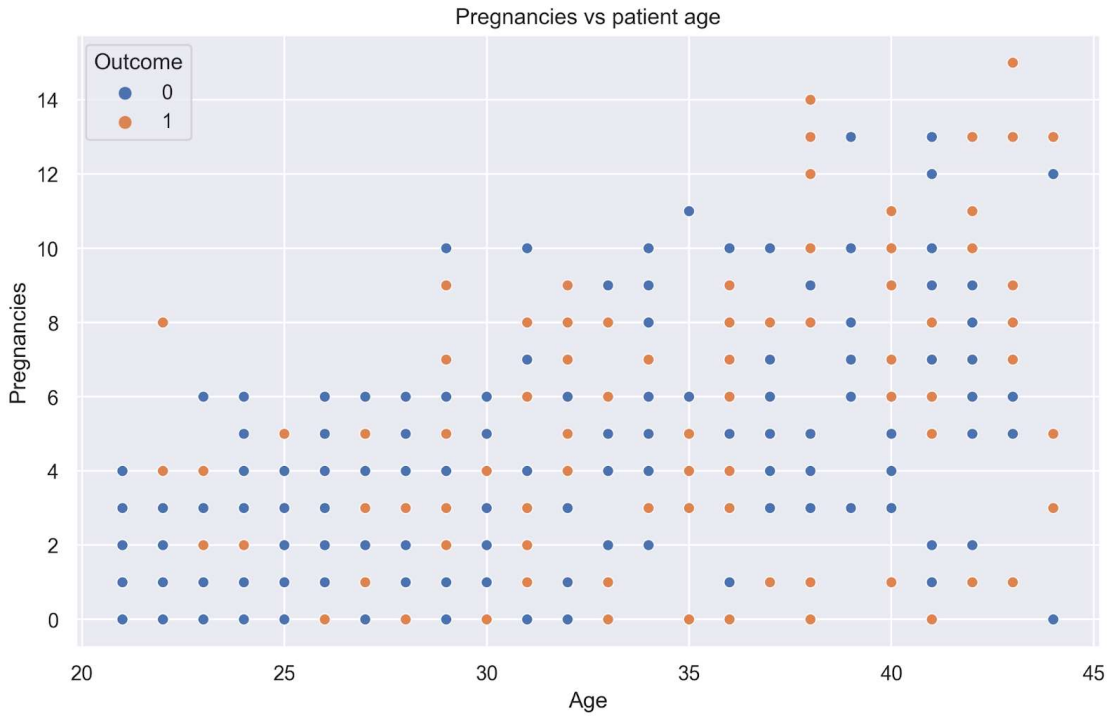
Age and number pregnancies

Earlier, it was shown that diabetics on average had more pregnancies. Is this due to a lurking variable such as age?



The box plot shows that the number of pregnancies increases as the age range increases. However, the number of pregnancies seem similar between 35-44 and 45+. Similar to diabetic outcome shown earlier, it is possible that most pregnancies occur before 45 years old.

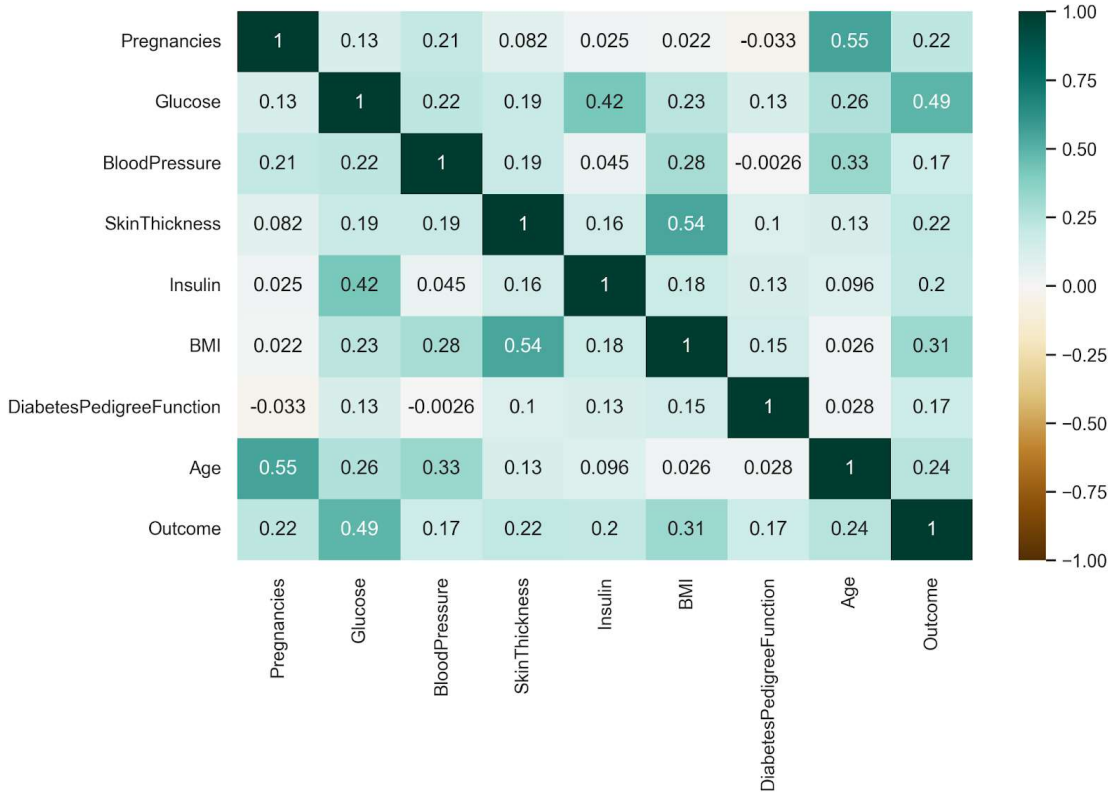
Below there is a scatter graph for pregnancies vs. age for patients between 18 and 45.



The scatter graph shows a correlation between age and number of pregnancies (r = 0.654). Number of pregnancies is associated with age, and age is associated with diabetes. Diabetic outcome is shown on the graph too to better visualise the relationships.

Correlations

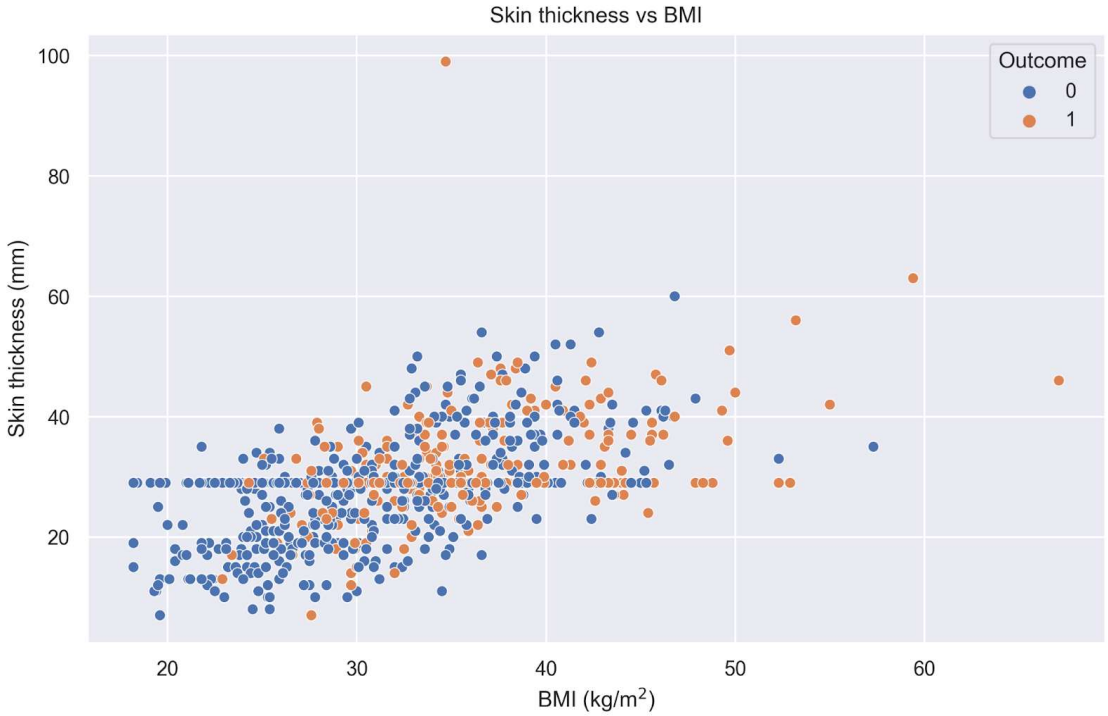
There were not many strong correlations found in the dataset. A summary is shown below:



The other correlations are between insulin and glucose, and skin thickness and BMI. Both insulin and skin thickness were missing a lot of data, so the conclusions cannot be as strong.

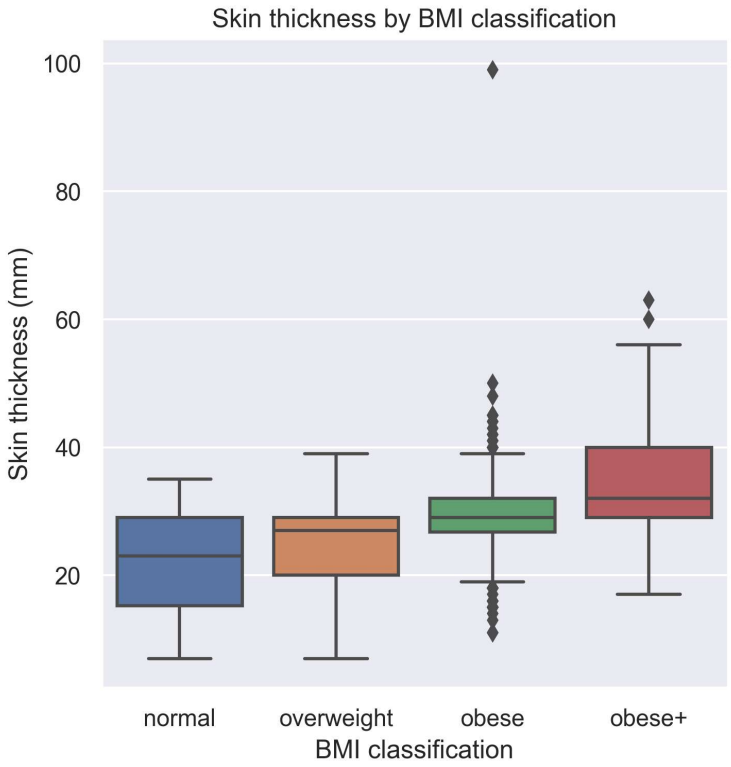
Skin thickness and BMI

Finally, the relationship between skin thickness and BMI can be examined. Skin thickness is the triceps skinfold thickness. It is measured with callipers and used to estimate body fat percentage. Normal thickness for females is 23mm.



The graph shows a positive correlation between BMI and skin thickness. Patients with higher BMIs tend to have thicker tricep skinfolds, indicating higher body fat. The distribution of positive outcomes for diabetes can also be seen. Orange dots dominate both the higher values for BMI and skin thickness.

The relationship between skin thickness and BMI can also be shown with a boxplot of BMI classifications:



It is easier to see the pattern on the box plot. As the BMI classification increases, so does the median skin thickness. Additionally, the general trend is that the interquartile range decreases as the BMI classification increases, until the 'obese+' classification which has the second largest. However the obese category has a lot of outliers, including an extreme outlier at 100mm.