**Student’s Name: Vishal Sharma**

**Roll Number: B20239**

**Mobile No: 9540140310**

**Branch: Data Science and Engineering**

Table 1 Mean, median, mode, minimum, maximum and standard deviation for all the attributes

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **S. No.** | **Attributes** | **Mean** | **Median** | **Mode** | **Min.** | **Max.** | **S.D.** |
| 1 | pregs | 3.845 | 3.00 | 1.00 | 0 | 17 | 3.369 |
| 2 | plas | 120.894 | 117.00 | 99.00 | 0 | 199 | 31.972 |
| 3 | pres (in mm Hg) | 69.105 | 72.00 | 70.00 | 0 | 122 | 19.355 |
| 4 | skin (in mm) | 20.536 | 23.00 | 0 | 0 | 99 | 15.952 |
| 5 | test (in mu U/mL) | 79.799 | 30.50 | 0 | 0 | 846 | 115.244 |
| 6 | BMI (in kg/m2) | 31.992 | 32.00 | 32.00 | 0 | 67.1 | 7.884 |
| 7 | pedi | 0.471 | 0.3725 | 0.254 | 0.078 | 2.42 | 0.331 |
| 8 | Age (in years) | 33.240 | 29.00 | 22.00 | 21.00 | 81 | 11.760 |

# Inferences:

1. From the above table, we observed that attribute pedi have a Standard Deviation of almost zero. It tells us that it’s spread is quite less.
2. For the attribute BMI, mean, mode and median are almost equal. It shows that the distribution of BMI is a nearly symmetrical distribution.

# a.

# Chart, scatter chart Description automatically generated

Figure 1 Scatter plot: Age (in years) vs. pregs

**Inferences:**

1. There are many patients who got pregnant almost 8 times before the age of 40 years.
2. As there is a dependency in the Age group of 20-30 years, but it is quite weak. Thus they are moderately dependent.

Chart, scatter chart

Description automatically generated

Figure 2 Scatter plot: Age (in years) vs. plas

**Inferences:**

1. From the above plot, there is a strong dependency between Age and Plasma-glucose concentration. Thus, they are strongly dependent.
2. There are many patients who have a plasma-glucose concentration between 75-150 in the age group of 20-30 years.

Chart, scatter chart

Description automatically generated

Figure 3 Scatter plot: Age (in years) vs. pres (in mm Hg)

**Inferences:**

1. From the above plot, we can infer that Diastolic Blood Pressure is weakly dependent on Age.
2. Almost everyone in all age groups have a Blood Pressure in between 50-96 mm Hg. But there are quite few whose Blood Pressure is quite high and some who have quite low.

Chart, scatter chart

Description automatically generated

Figure 4 Scatter plot: Age (in years) vs. skin (in mm)

**Inferences:**

1. From the above plot, the dependency of Triceps skin fold thickness is not too much with Age. Thus, it is weakly dependent.
2. People in the age group of 20-30 years have a thickness of 10-40 mm.

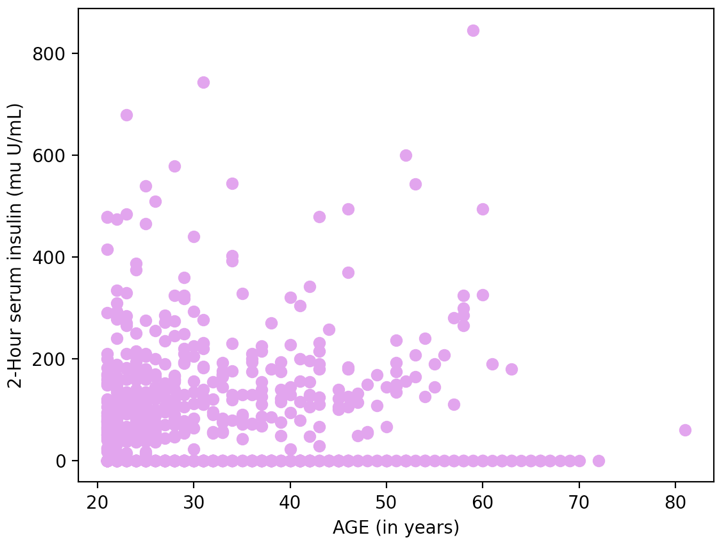


Figure 5 Scatter plot: Age (in years) vs. test (in mm U/mL)

**Inferences:**

1. The dependency between insulin intake and Age is very low. Thus, they are very weakly dependent on each other.
2. Most people in the age group of 20-30 years take a serum of insulin in the range of 0 – 200 mu U/mL

Chart, scatter chart

Description automatically generated

Figure 6 Scatter plot: Age (in years) vs. BMI (in kg/m2)

**Inferences:**

1. The dependency between BMI and Age is quite moderate, especially in the age group of 20-40 years. Thus, they are very weakly dependent on each other.
2. People in the age group of 20-30 years have a BMI in the range of 19-42 Kg/m^2.
3. Mostly the BMI lies in the range of 19-45 Kg/m^2 for all age groups.

Chart, scatter chart

Description automatically generated

Figure 7 Scatter plot: Age (in years) vs. pedi

**Inferences:**

1. The distribution is quite a weak dependency between Age and Diabetes pedigree function. Thus, they are very weakly dependent on each other.
2. The diabetes pedigree function is mostly in the range of 0 – 1 for almost every patient.

**b.**

Chart, scatter chart

Description automatically generated

Figure 8 Scatter plot: BMI (in kg/m2) vs. pregs

**Inferences:**

1. From the above plot, the dependency is quite weak between BMI and pregnancy, especially who have a BMI>50 or BMI<20 . Thus, they are weakly dependent on each other.
2. Mostly people who have a BMI in the range of 20-50 kg/m^2 were pregnant.

Chart, scatter chart

Description automatically generated

Figure 9 Scatter plot: BMI (in kg/m2) vs. plas

**Inferences:**

1. There is no such dependency between BMI and plasma glucose concentration. They are weakly dependent on each other.
2. The data is quite dense for patients who have a BMI between 20-40 Kg/m^2.

Chart, scatter chart

Description automatically generated

Figure 10 Scatter plot: BMI (in kg/m2) vs. pres (in mm Hg)

**Inferences:**

1. There is a weak dependency between BMI and Blood Pressure. They are weakly dependent on each other.
2. Most of the data is quite dense between BMI 20-40 kg/m^2.

Chart, scatter chart

Description automatically generated

Figure 11 Scatter plot: BMI (in kg/m2) vs. skin (in mm)

**Inferences:**

1. There is a dependency between BMI and Triceps thickness, but the slope of line of best fit is not 45 degrees. Thus they are moderately dependent on each other.
2. Most of the data is spread between BMI of 20-40 Kg/m^2 with a triceps thickness of 4mm -60mm.

Chart, scatter chart

Description automatically generated

Figure 12 Scatter plot: BMI (in kg/m2) vs. test (in mm U/mL)

**Inferences:**

1. As there is no such dependency observable from the above plot, thus they are weakly dependent on each other.
2. The insulin serum concentration is mainly between 0 – 200 mu U/mL.

Chart, scatter chart

Description automatically generated

Figure 13 Scatter plot: BMI (in kg/m2) vs. pedi

**Inferences:**

1. There is no such dependency observable from the above plot. Hus they are very weakly dependent on each other.
2. The Diabetes Pedigree Function is mainly in the range of 0-1.

Chart, scatter chart

Description automatically generated

Figure 14 Scatter plot: BMI (in kg/m2) vs. Age (in years)

**Inferences:**

1. There is no such dependency observable from the above plot. Hus they are very weakly dependent on each other.
2. BMI is mainly concentrated in 20-45 Kg/m^2 for the age group of 20 – 70 years.

# a.

Table 3 Correlation coefficient value computed between age and all other attributes

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Attributes** | **Correlation Coefficient Value** |
| 1 | pregs | 0.5443 |
| 2 | plas | 0.2635 |
| 3 | pres (in mm Hg) | 0.2395 |
| 4 | skin (in mm) | -0.1139 |
| 5 | test (in mu U/mL) | -0.0421 |
| 6 | BMI (in kg/m2) | 0.0362 |
| 7 | pedi | 0.0335 |

**Inferences:**

1. From the magnitude of correlation coefficient value, attributes pedi, BMI, skin and test have a coefficient near to zero. Thus they are almost independent with the attribute age. Attribute pregs in moderately correlated with Age, while plas and pres attributes are weakly corelated with Age.
2. From the sign of correlation coefficient value, skin and test are the attributes whose value decreases as the Age of a patient increases. While for the above attribute, the value increases as Age increases.
3. From the above scatter plots, we can show that the dependency of various attributes with Age. Though the attributes test, BMI and pedi have very small correlation coefficient, it is quite hard to observe.

**b.**

Table 4 Correlation coefficient value computed between BMI and all other attributes

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Attributes** | **Correlation Coefficient Value** |
| 1 | pregs | 0.0176 |
| 2 | plas | 0.2210 |
| 3 | pres (in mm Hg) | 0.2818 |
| 4 | skin (in mm) | 0.3925 |
| 5 | test (in mu U/mL) | 0.1978 |
| 6 | pedi | 0.1406 |
| 7 | Age (yrs) | 0.0362 |

**Inferences:**

1. From the magnitude of correlation coefficient value, attributes pedi, pregs and Age have a coefficient near to zero. Thus, they are almost independent with the attribute age. Attribute skin is moderately correlated with BMI, while plas and pres attributes are weakly corelated with BMI.
2. From the sign of correlation coefficient value, all the attributes have a positive correlation with BMI. Thus, as BMI increases, value of each attribute increases.
3. From the above scatter plots, we can show that the dependency of various attributes with BMI. Though the attributes test, Age has a very small correlation coefficient, it is quite hard to observe.

# a.

Chart, histogram

Description automatically generated

Figure 15 Histogram depiction of attribute pregs

**Inferences:**

1. From the above histogram, it can be interpreted that most people were pregnant in the range of 0-2, with a frequency of almost 250. It gradually decreases as the no of pregnancy increases.
2. As the first bin have the maximum height, it can be interpreted that the mode for this distribution lies in the range of 0 – 2.

Chart, histogram

Description automatically generated

Figure 16 Histogram depiction of attribute skin

**Inferences:**

1. The first bin has the highest frequency. Thus, most people have a Triceps skin fold thickness in the range of 0-10mm. It first decreases, then an increase in the thickness then ultimately decreases and approaches zero.
2. As the first bin is the highest, it can be interpreted that the mode of this distribution lies in the range of 0-10mm.

Chart, histogram

Description automatically generated

Figure 17 Histogram depiction of attribute pregs for class 0

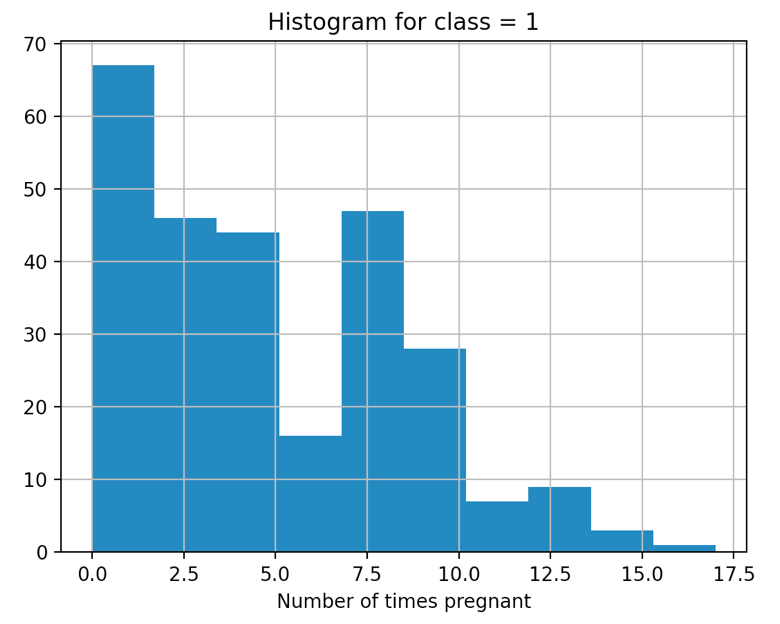


Figure 18 Histogram depiction of attribute pregs for class 1

**Inferences:**

1. From the histogram, it can be inferred that maximum number of people were pregnant in the range of 0-2 times.
2. Also, the mode of the above histogram lies in the range of 0-2 as it has the maximum frequency. But the bins of class 0 in range 0-2 have more frequency that class 1 for the same range.

Chart, box and whisker chart

Description automatically generated

Figure 19 Boxplot for attribute pregs

**Inferences:**

1. The outliers lie above the range of 14-17.4 (approximately), while Q3+(1.5\*IQR) is at 13.
2. The Inter Quartile Range (IQR) is 6-1 = 5 (Q3 – Q1)
3. As the median is not at the center, attribute pregs have a positive spread.
4. The minimum value, maximum value, and median are 0, 17, 3 respectively.
5. The above distribution is skewed towards left.

Chart, box and whisker chart

Description automatically generated

Figure 20 Boxplot for attribute plas

**Inferences:**

1. The outlier at the bottom is approximately zero, which is smaller than the minimum range which is Q1-1.5\*IQR = 44.
2. The range Q3 – Q1 = 140.25-99 = 41.25 (approximately).
3. As the median is not at the center, it is positively spread and has a range of 44-199 with some values at zero.
4. The above distribution is skewed towards left.
5. The minimum, maximum and median is given by 0, 199, 117 respectively with Q1 at 99 and Q3 at 140.25.

Chart, box and whisker chart

Description automatically generated

Figure 21 Boxplot for attribute pres(in mm Hg)

**Inferences:**

1. The outlier at the bottom is at zero, which is smaller than the minimum range which is Q1-1.5\*IQR = 38 and the outlier at the top is at 122, which is greater than Q3+1.5\*IQR = 106
2. The range Q3 – Q1 = 80 - 62 = 18.
3. As the median is approximately at the center, it is symmetrically spread (almost) and has a range of 38-106 with some values at zero and some at 122.
4. The above distribution is skewed towards center.
5. The minimum, maximum and median is given by 0, 122, 72 respectively with Q1 at 62 and Q3 at 80.

Chart, box and whisker chart

Description automatically generated

Figure 22 Boxplot for attribute skin(in mm)

**Inferences:**

1. The outlier at the top is approximately at 99, which is greater than the maximum range which is Q3+1.5\*IQR = 63.
2. The range Q3 – Q1 =32 - 0 = 32 (approximately).
3. As the median is not at the center, it is negatively spread and has a range of 0-99 with some values at range greater than 63.
4. The above distribution is skewed towards right.
5. The minimum, maximum and median is given by 0, 99, 23 respectively with Q1 at 0 and Q3 at 32.

Chart, box and whisker chart

Description automatically generated

Figure 23 Boxplot for attribute test (mu U/mL)

**Inferences:**

1. The outlier at the top is approximately at 846, which is greater than the maximum range which is Q3+1.5\*IQR = 318.
2. The range Q3 – Q1 = 127 - 0 = 127 (approximately).
3. As the median is not at the center, it is positively spread and has a range of 0-846 with some values at range greater than 318.
4. The above distribution is skewed towards left.
5. The minimum, maximum and median is given by 0, 846, 31.5 respectively with Q1 at 0 and Q3 at 127.

Chart, box and whisker chart

Description automatically generated

Figure 24 Boxplot for attribute BMI (in kg/m2)

**Inferences:**

1. The outlier at the bottom is approximately zero, which is smaller than the minimum range which is Q1-1.5\*IQR = 18.2 and another outlier at the top with value 67.1 which is greater than Q3+1.5\*IQR = 50.
2. The range Q3 – Q1 =36.6 – 27.3 = 9.3 (approximately).
3. As the median is at the center, it is symmetrically spread and has a range of 0-67.1.
4. The above distribution is skewed towards center.
5. The minimum, maximum and median is given by 0, 67.1, 32 respectively with Q1 at 27.3 and Q3 at 36.6.

Chart, box and whisker chart

Description automatically generated

Figure 25 Boxplot for attribute pedi

**Inferences:**

1. The outlier at the top is approximately at 2.42, which is greater than the maximum range which is Q3+1.5\*IQR = 1.19.
2. The range Q3 – Q1 = 0.625 – 0.243 = 0.382 (approximately).
3. As the median is not at the center, it is positively spread and has a range of 0.243-2.42.
4. The above distribution is skewed towards left.
5. The minimum, maximum and median is given by 0.243, 2.42, 0.372 respectively with Q1 at 0.243 and Q3 at 0.625.

Chart, box and whisker chart

Description automatically generated

Figure 26 Boxplot for attribute Age (in years)

**Inferences:**

1. The outlier at the top is at 81, which is greater than the maximum range which is Q1+1.5\*IQR = 66.
2. The range Q3 – Q1 = 41 – 24 = 17.
3. As the median is not at the center, it is positively spread and has a range of 21-81.
4. The above distribution is skewed towards right.
5. The minimum, maximum and median is given by 21, 81, 29 respectively with Q1 at 21 and Q3 at 41.