| Description |  | Excel Function | Result |
| --- | --- | --- | --- |
| Count |  | =COUNT (I2:A769) | 768 |
| Minimum |  | =MIN (I2:I769) | 21 |
| Maximum |  | =MAX (I2:I769) | 81 |
| Mean |  | =AVERAGE (I2:I769) | 33.24 |
| Median |  | =MEDIAN (I2:I769) | 29 |
| First Quartile (Q1) |  | =QUARTILE.INC (I2:I769, 1) | 24 |
| Third Quartile (Q3) |  | =QUARTILE.INC (I2:I769, 3) | 41 |
| Interquartile Range (IQR) |  | =Q3 - Q1 | 17 |
| Standard Deviation |  | =STDEV.P(I2:I769) | 11.76 |
| Mean Deviation |  | =AVEDEV (I2:I769) | 9.85 |
|  | | | |

**PLR week task 4 Data Exploration with Excel Function**

**Description of the data**

The patients in this dataset have ages ranging from 21 to 81 years. The average age is approximately 33 years, but the median age is lower at 29 years, suggesting the data is right skewed with a concentration of younger patients and a tail of older ones. The standard deviation of about 11.76 years indicates a moderate spread in age."

**Ratio of Diabetic vs. Non-Diabetic Patients**

* Count of non-diabetic (0): =COUNTIF (J2:J769, 0) -> **500**
* Count of Diabetic (1): =COUNTIF (J2:J769, 1) -> **268**
* Ratio (Diabetic: Non-Diabetic): 268: 500, which simplifies to approximately 1: 1.87. This means for every 1 diabetic patient, there are about 1.87 non-diabetic patients.

**Appropriate Measure of Central Tendency**

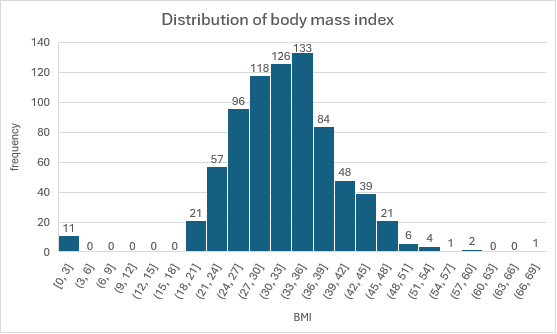
* **Age:** **Median**. The age data is skewed (as seen by the difference between mean and median). The median is less influenced by the few very old ages and better represents the "typical" patient.
* **Body Mass (mass):** **Mean**. BMI is typically a well-distributed, continuous variable where the mean is a meaningful measure of central body mass, assuming no extreme outliers.
* **Plasma (plas):** **Mean**. This is a key clinical measurement, and the mean plasma glucose level is a standard indicator.
* **Blood Pressure (pres):** **Mean**. Like plasma, blood pressure is a standard clinical metric where the average is most commonly used and informative.

**3.Data Visualization with Charts**

**Body Mass Index (BMI) - Histogram and Line Chart**

1. **Histogram:**

**Purpose**: To show the distribution of BMI values.



**Observations from Your Data:**

1. **Shape**: The distribution will be roughly bell-shaped (normal distribution)
2. **Center**: The histogram will peak around BMI 30-35
3. **Spread**: Most values will fall between 20 and 45 BMI

**Diabetics vs. Non-Diabetic - Pie Chart**

**Purpose:** To show the composition of the dataset.

* **Observation:** The chart will visually confirm that non-diabetic patients make up a larger portion (about 65%) of the dataset.

**Relationship between Skin, Age, and Class - Scatter Plot**

**Further Exploration & Observed Trends**

* **Area Chart:** You could plot the running total of patients by age to see the cumulative distribution.
* **Trendline:** As mentioned in the scatter plot, adding a trendline can help visualize if there is a positive or negative correlation between variables like plas (glucose).

**4.Final Description of the Dataset**

This dataset contains medical records from 768 female Pima Indian patients for diabetes prediction. The typical patient is around 33 years old (mean), though the median age of 29 better represents the population due to right-skewed distribution favoring younger patients with fewer older outliers. Clinical variables such as plasma glucose concentration, BMI, and blood pressure are appropriately summarized using mean values as standard continuous metrics.

The dataset shows class imbalance, with non-diabetic cases (class=0) outnumbering diabetic cases (class=1) approximately 2:1. Analysis indicates that higher plasma glucose levels strongly correlate with diabetes occurrence, while age also shows positive association with increased diabetes prevalence. Notable outliers exist in age and insulin test values, requiring verification for potential data errors or genuine extreme cases.