EXPLORATORY DATA AND STATISTICAL ANALYSIS ON DIABETES

Presented by:-

Saumya Achantani	18csu194
Saumya Gupta	18csu195
Rupali Taneja	18csu182

Under the supervision of:-

Ms. Shakshi Sharma

(Faculty in-charge)

NCU
THE NORTHCAP UNIVERSITY®
NAAC ACCREDITED
(Formerly ITM University, Gurugram)

TABLE OF CONTENTS

PROBLEM STATEMENT

To study and analyze the number of people suffering from diabetes, various causes of diabetes, which factors are majorly responsible for diabetes and relationship among these factors.

INTRODUCTION

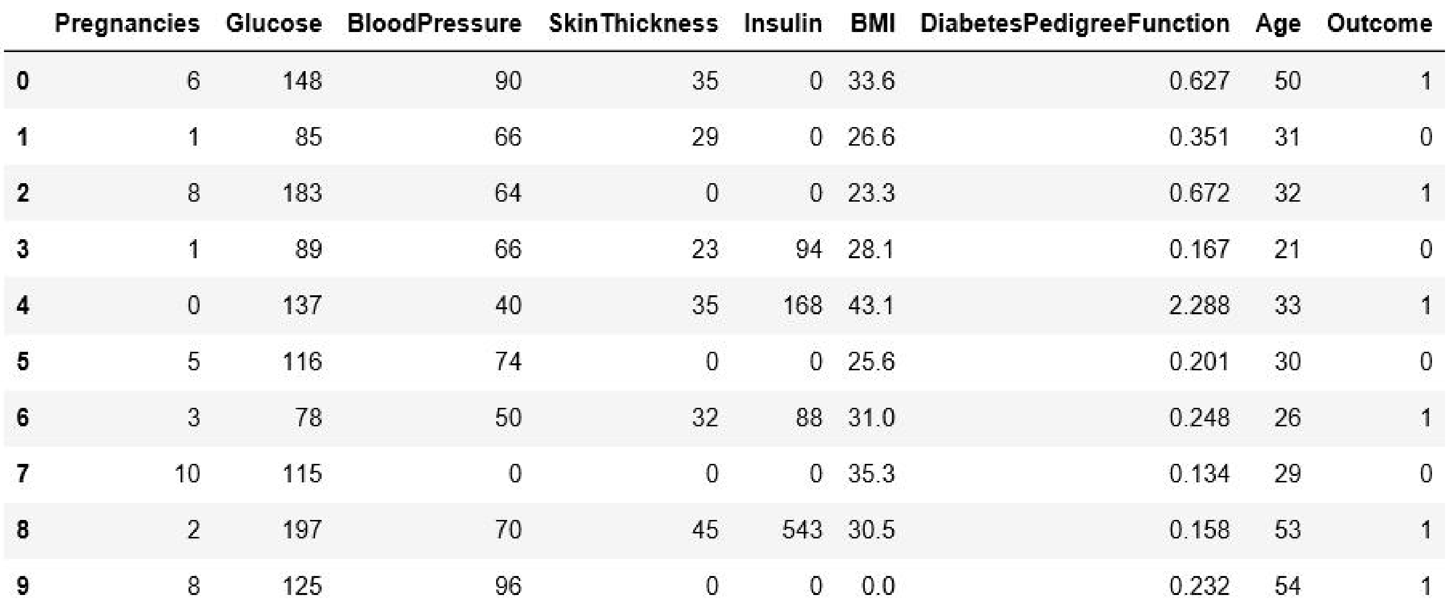
- The main objective is to raise awareness about diabetes by analyzing various factors responsible for the same.

- We will try to analyze this data with the help of data visualization, probability and statistics which helps in getting a visual picture of the factors and the extent to which they affect a diabetic person.

DATASET DESCRIPTION

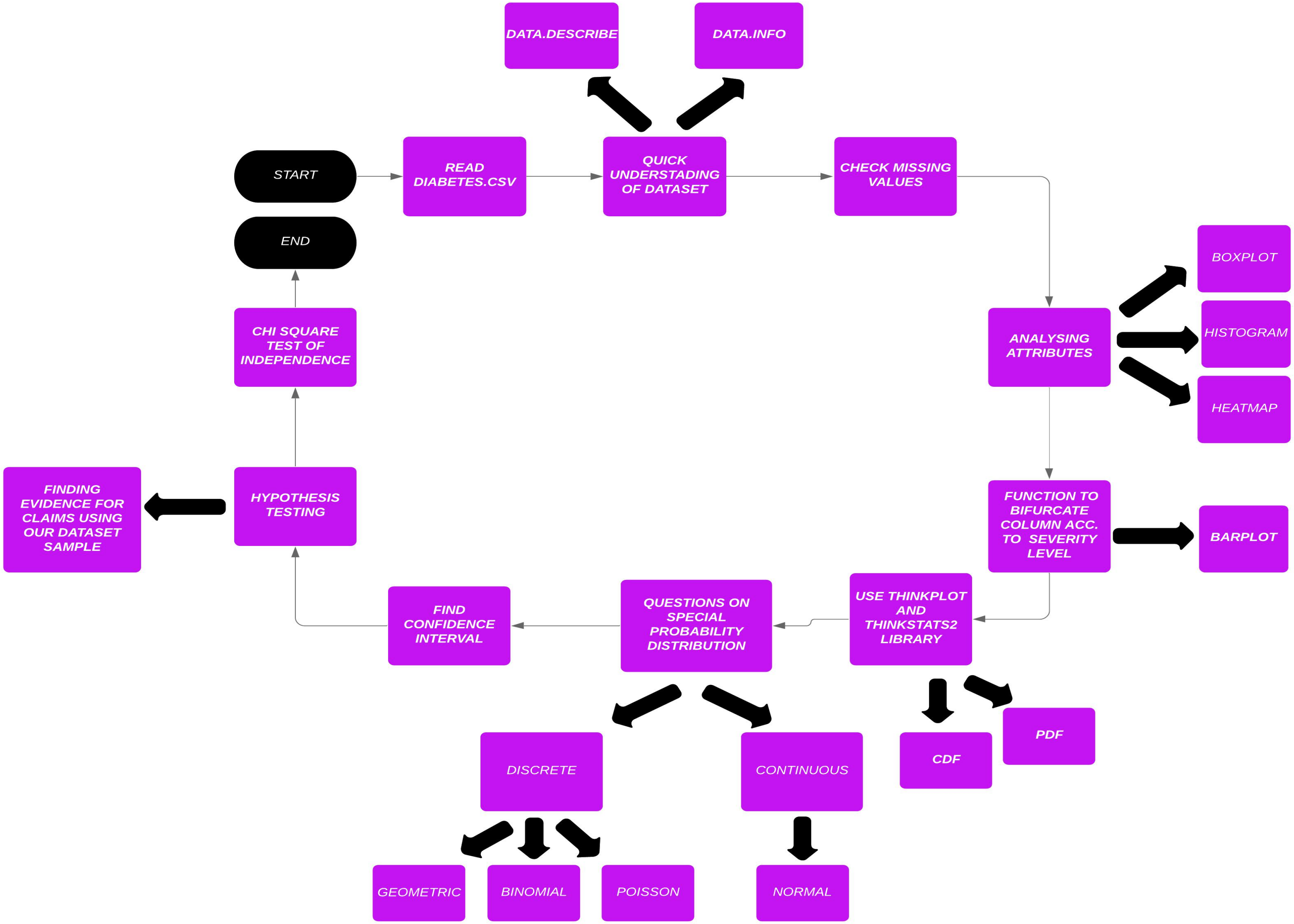
- ❖ Pima Indians Diabetes Dataset is originally from the National Institute of Diabetes Diseases.

The DATASET contains following attributes:-



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	90	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31.0	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0.0	0.232	54	1

PROJECT DESIGN



START

READ DIABETES.CSV

QUICK UNDERSTADING OF DATASET

DATA.DESCRIBE

DATA.INFO

CHECK MISSING VALUES

ANALYSING ATTRIBUTES

BOXPLOT

HISTOGRAM

HEATMAP

FUNCTION TO BIFURCATE COLUMN ACC. TO SEVERITY LEVEL

BARPLOT

USE THINKPLOT AND THINKSTATS2 LIBRARY

CDF

PDF

QUESTIONS ON SPECIAL PROBABILITY DISTRIBUTION

DISCRETE

GEOMETRIC

BINOMIAL

POISSON

CONTINUOUS

NORMAL

FIND CONFIDENCE INTERVAL

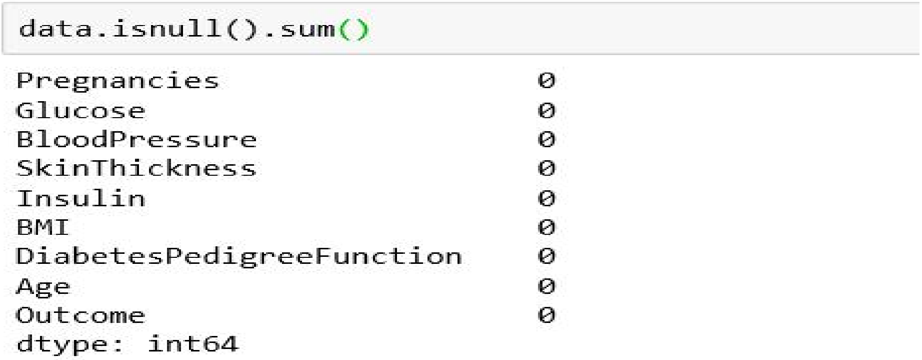
HYPOTHESIS TESTING

FINDING EVIDENCE FOR CLAIMS USING OUR DATASET SAMPLE

CHI SQUARE TEST OF INDEPENDENCE

END

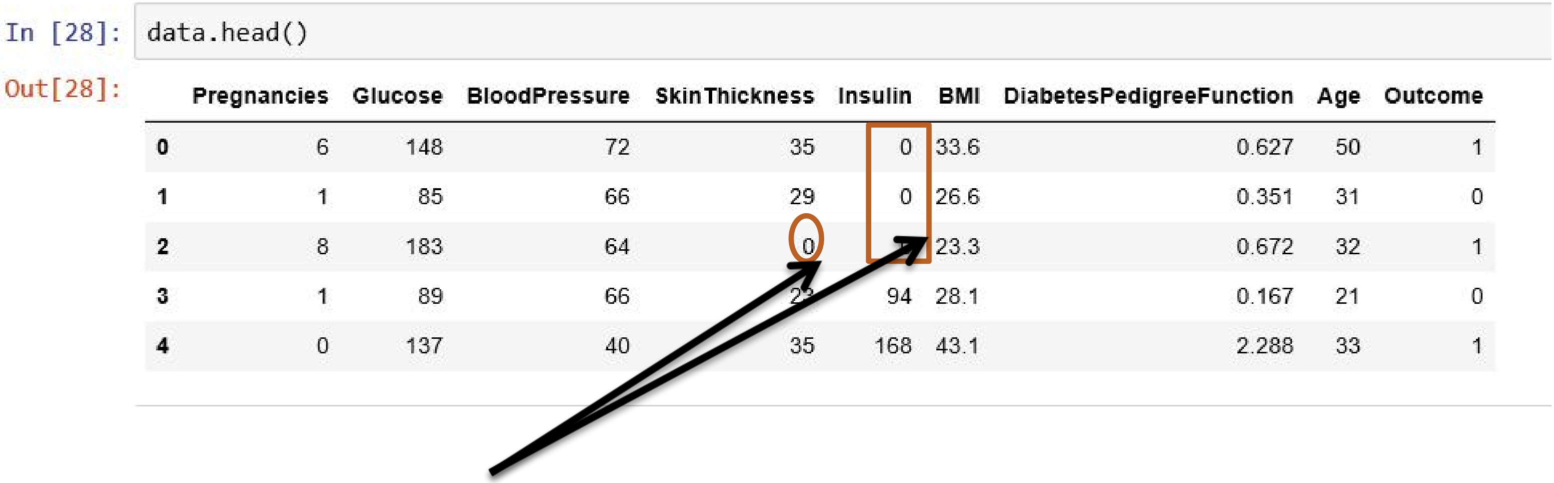
CHECKING FOR MISSING VALUES



```
data.isnull().sum()
```

```
Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

No NAN values?



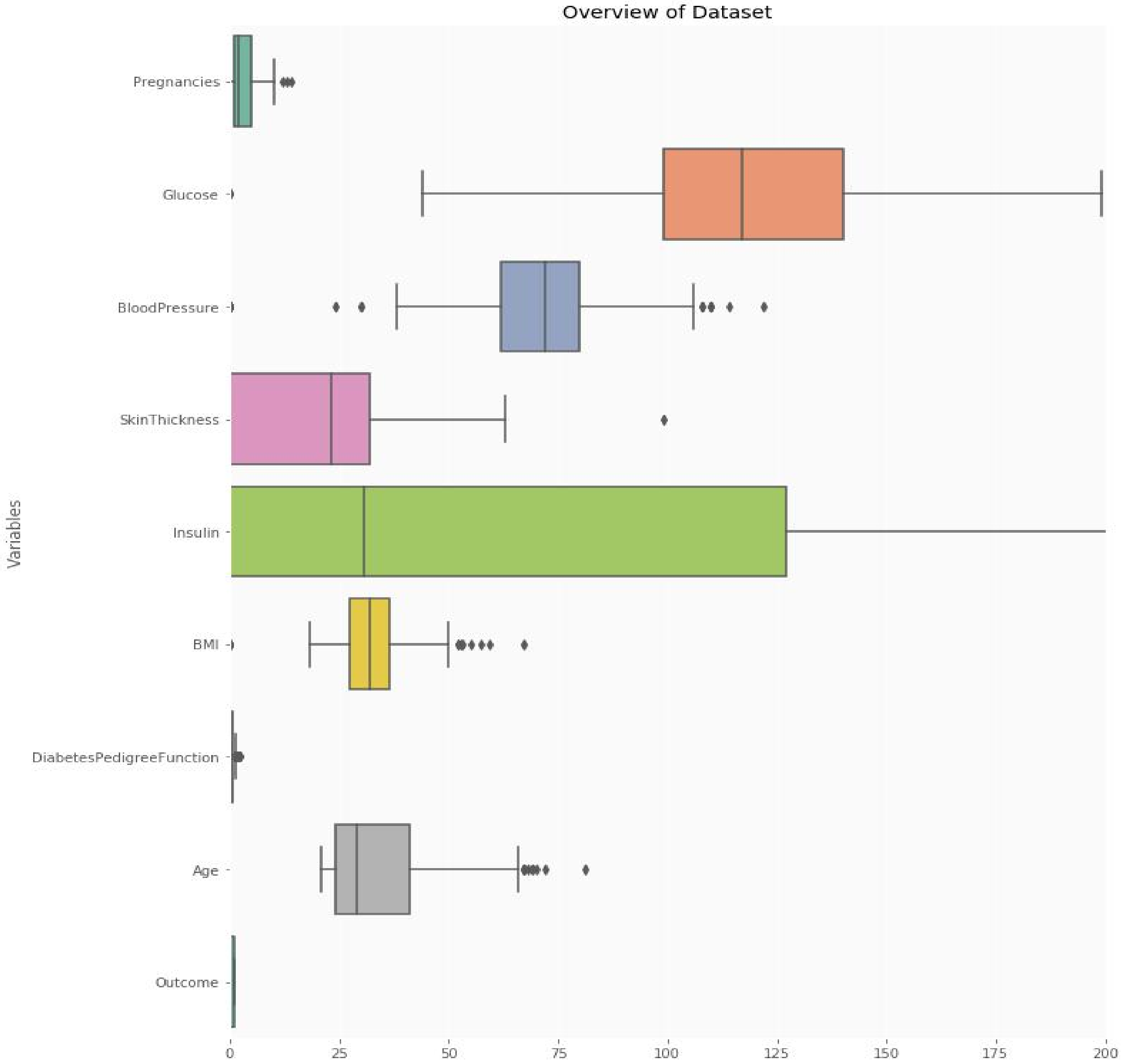
```
In [28]: data.head()
```

Out[28]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	[illegible]	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

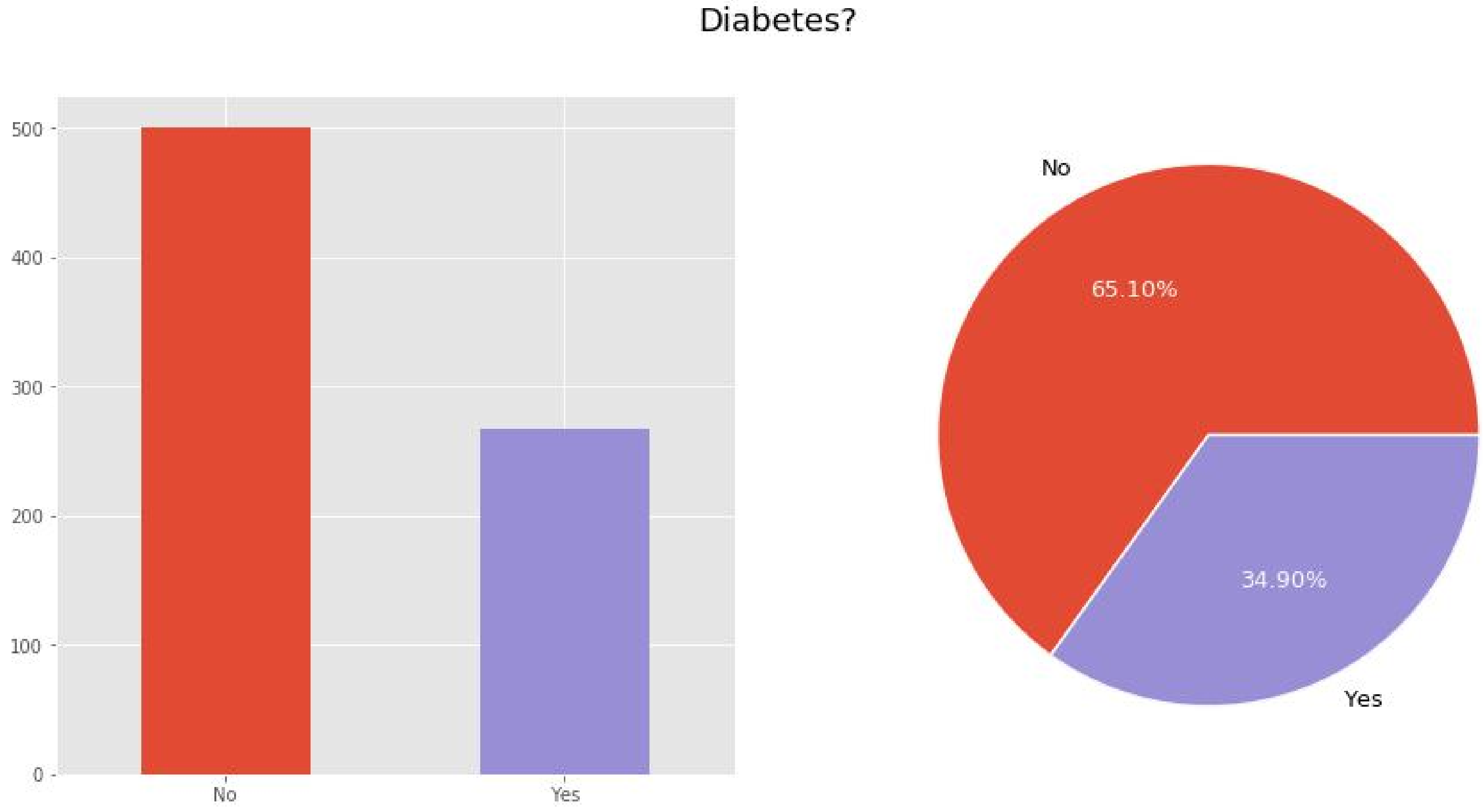
Attributes have to be replaced from 0 to NAN.
Then we have performed imputation..

Fig 1: Overview of Dataset



This plot shows the range and distribution of each attribute in dataset .

Fig2:Distribution of people with outcome of diabetes

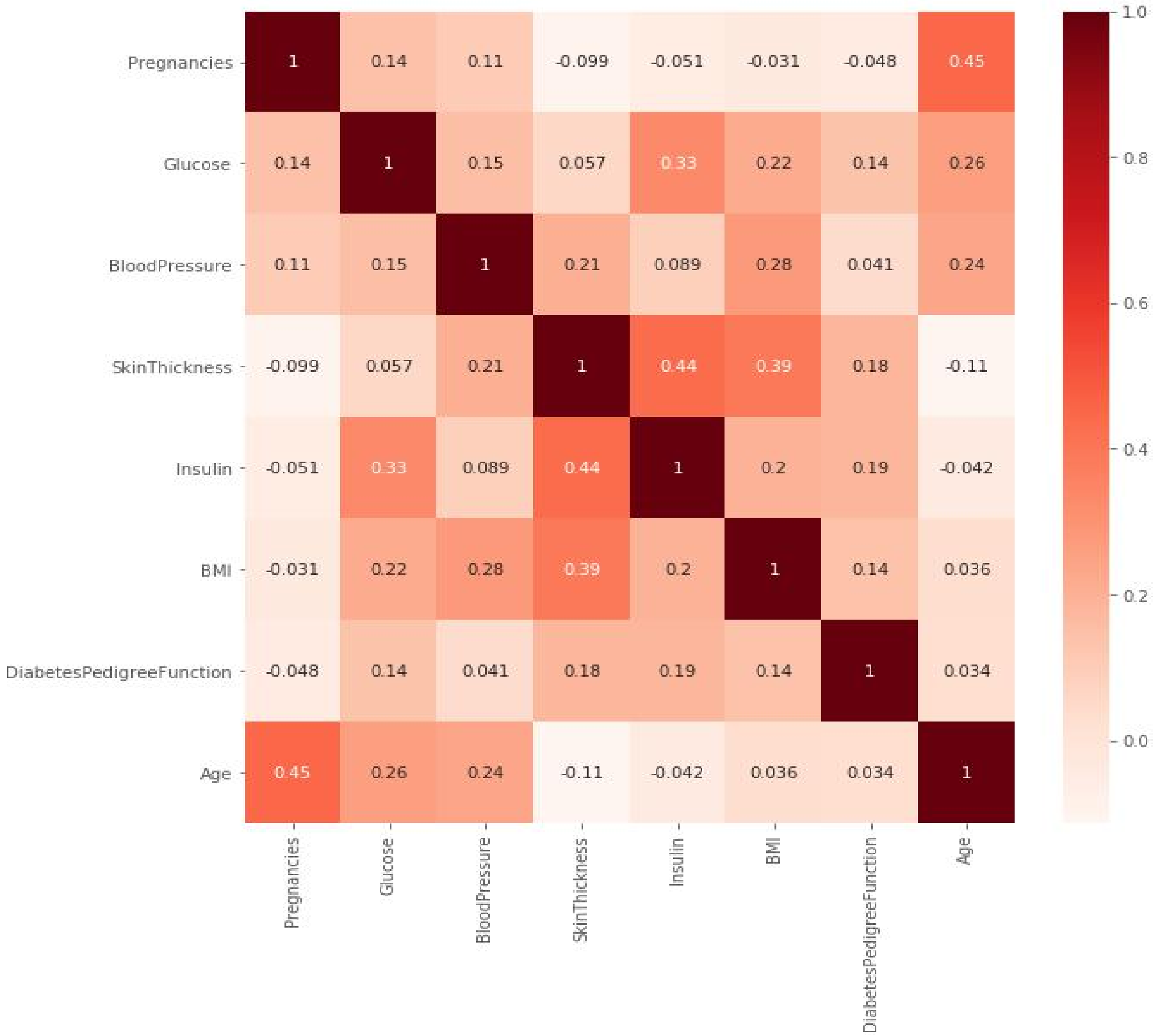


Barplot

Piechart

It can be *seen that the number of non-diabetics is almost twice the number of diabetic patients*

Fig 3: Heatmap



Examination of correlation among the variables.

Fig 4:Histogram of all variables

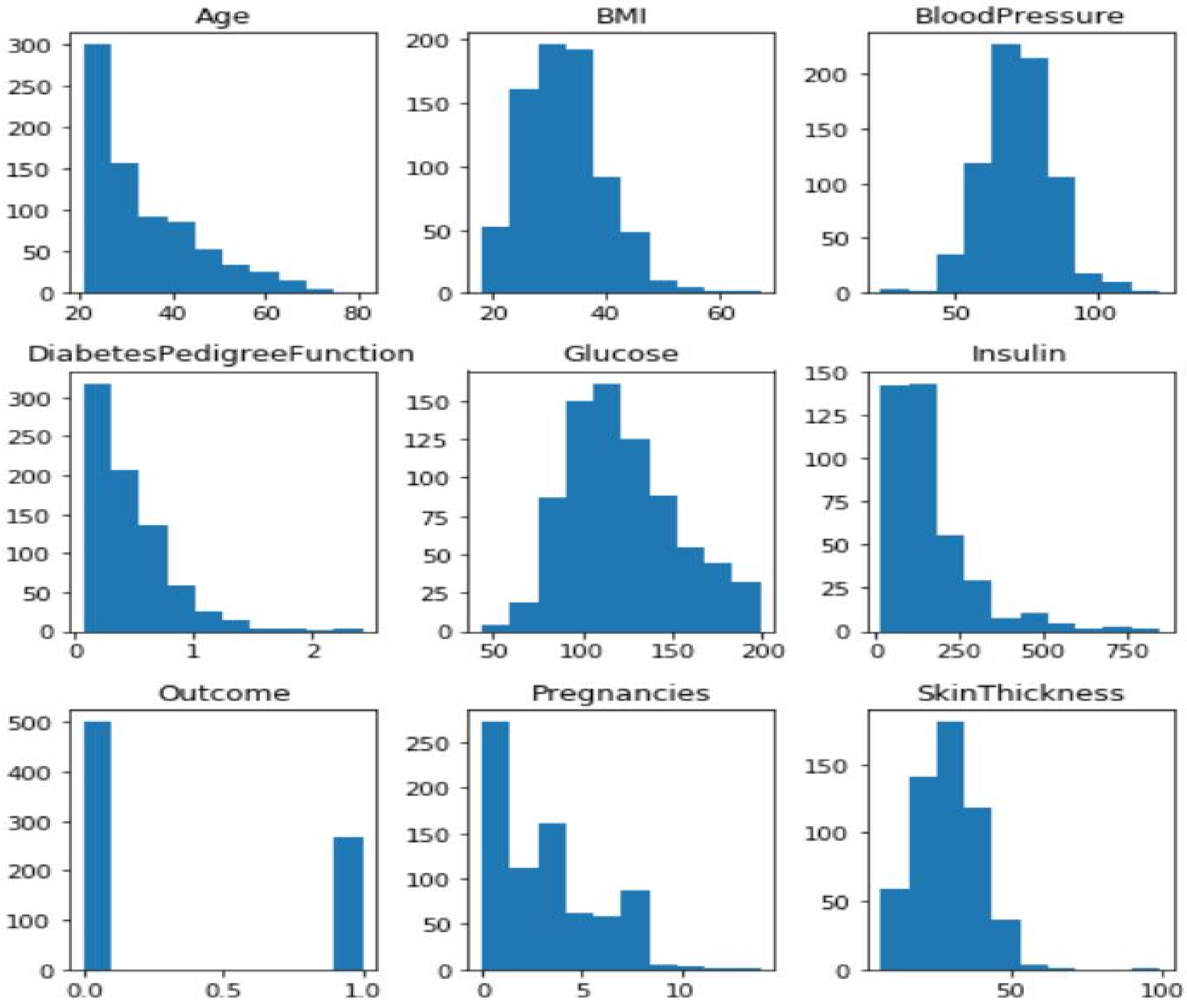
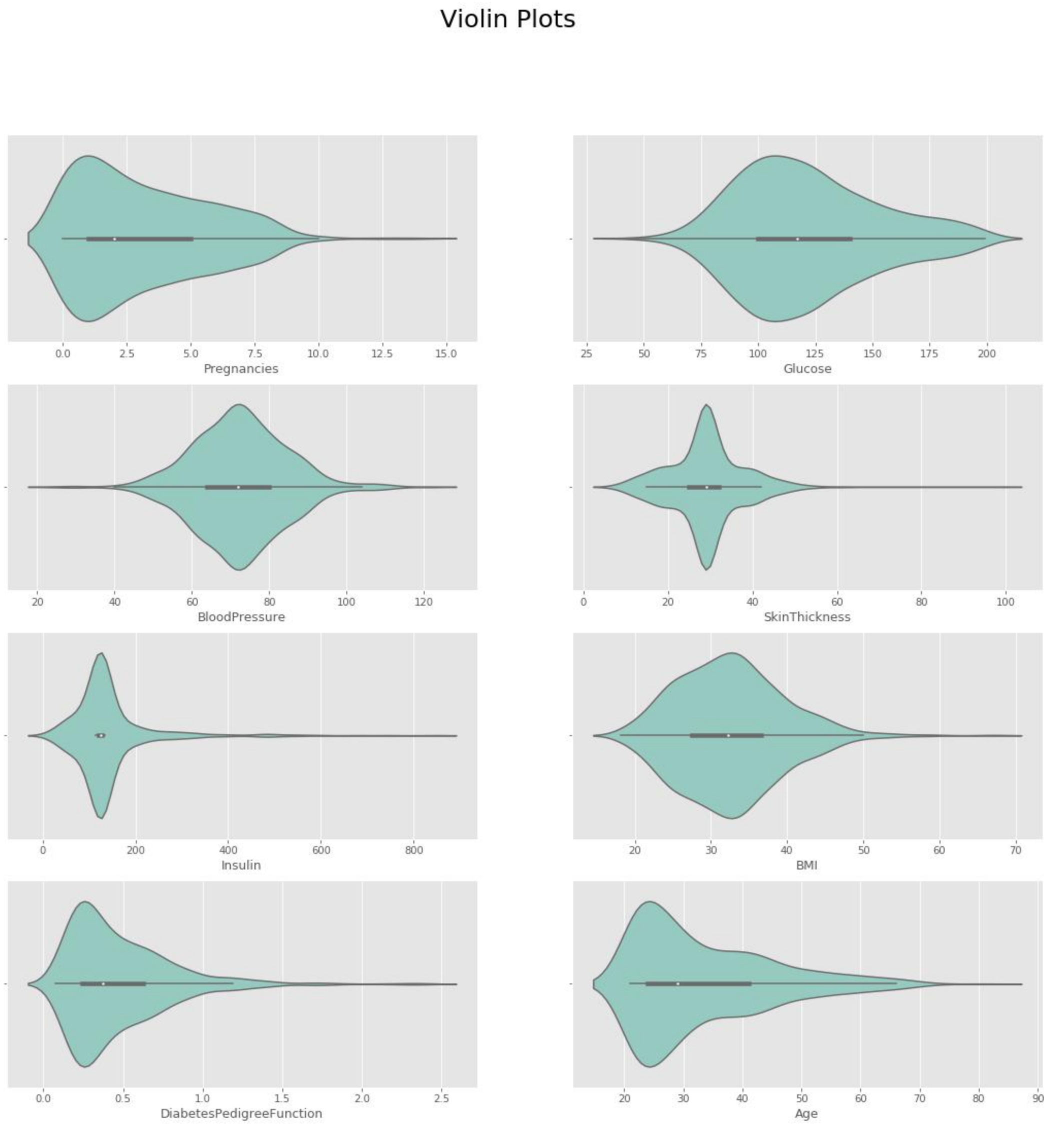


Fig. 5:Violin plot



A violin plot clearly shows the presence of different peaks, their position and relative amplitude.

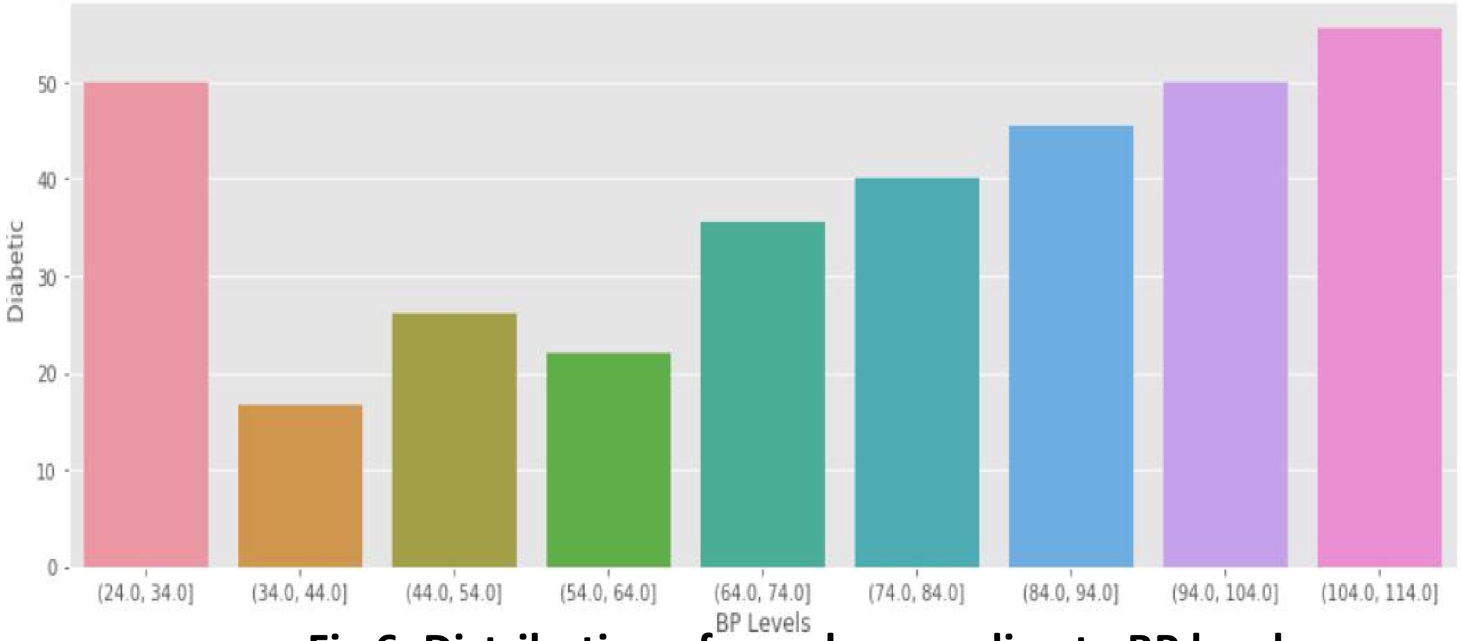


Fig 6: Distribution of people according to BP level

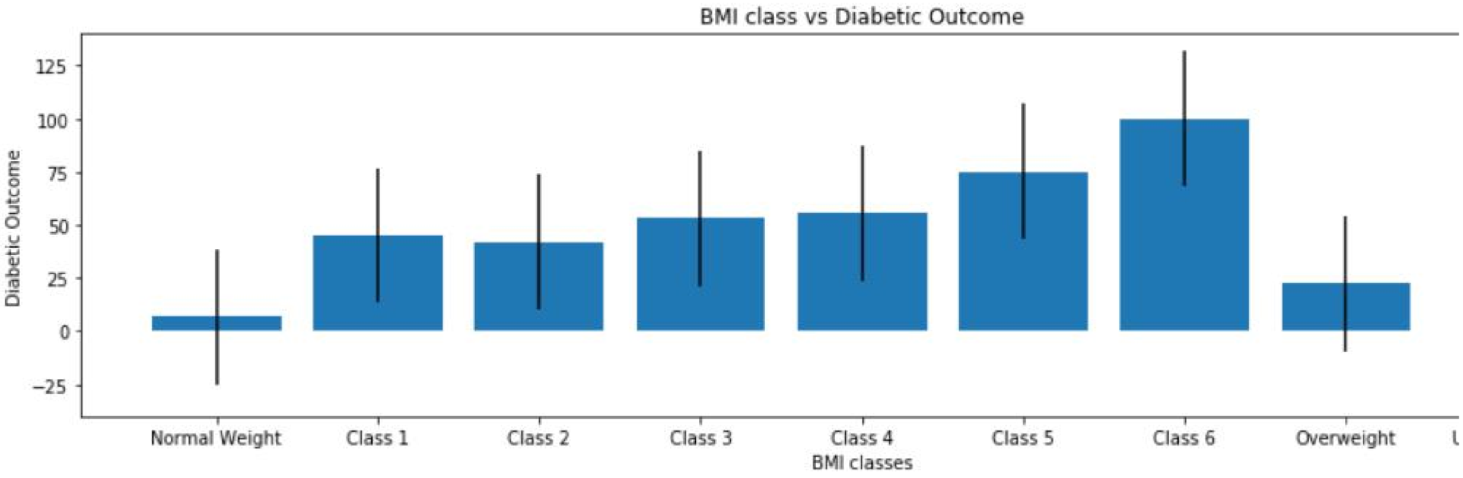
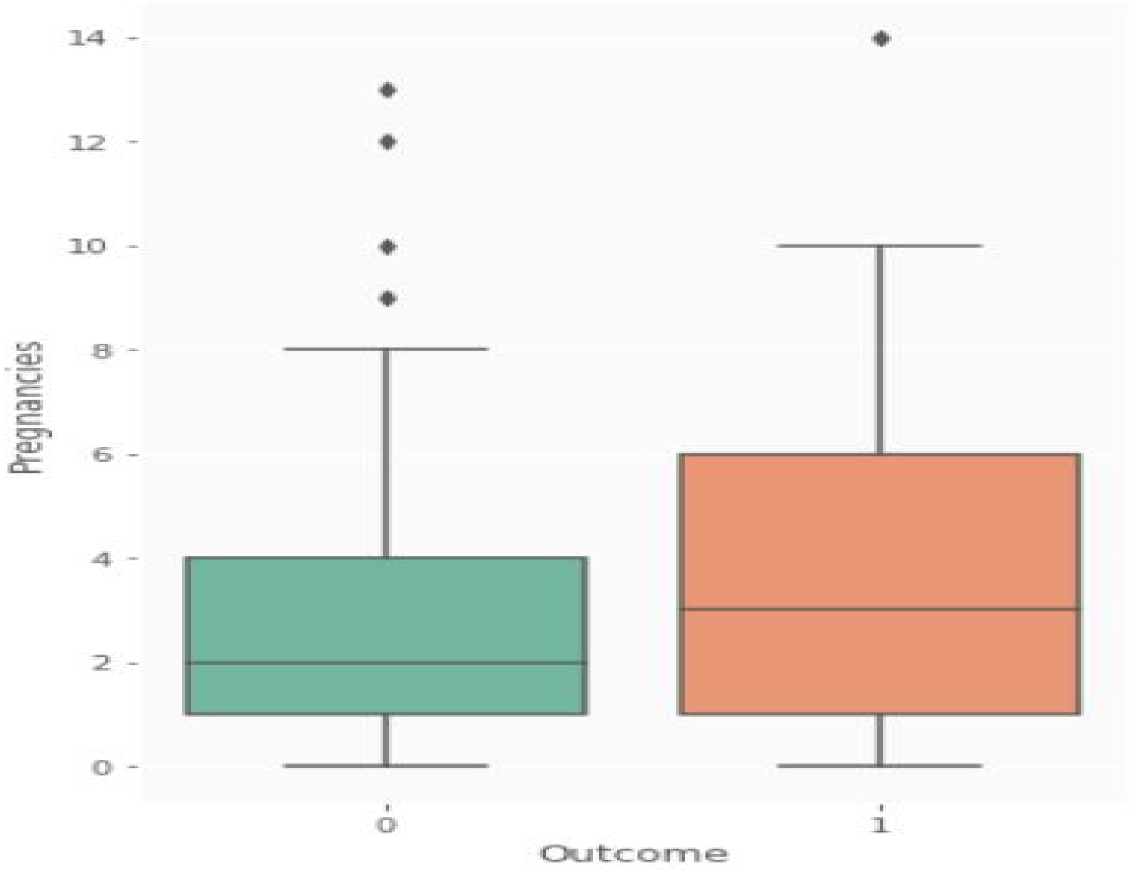


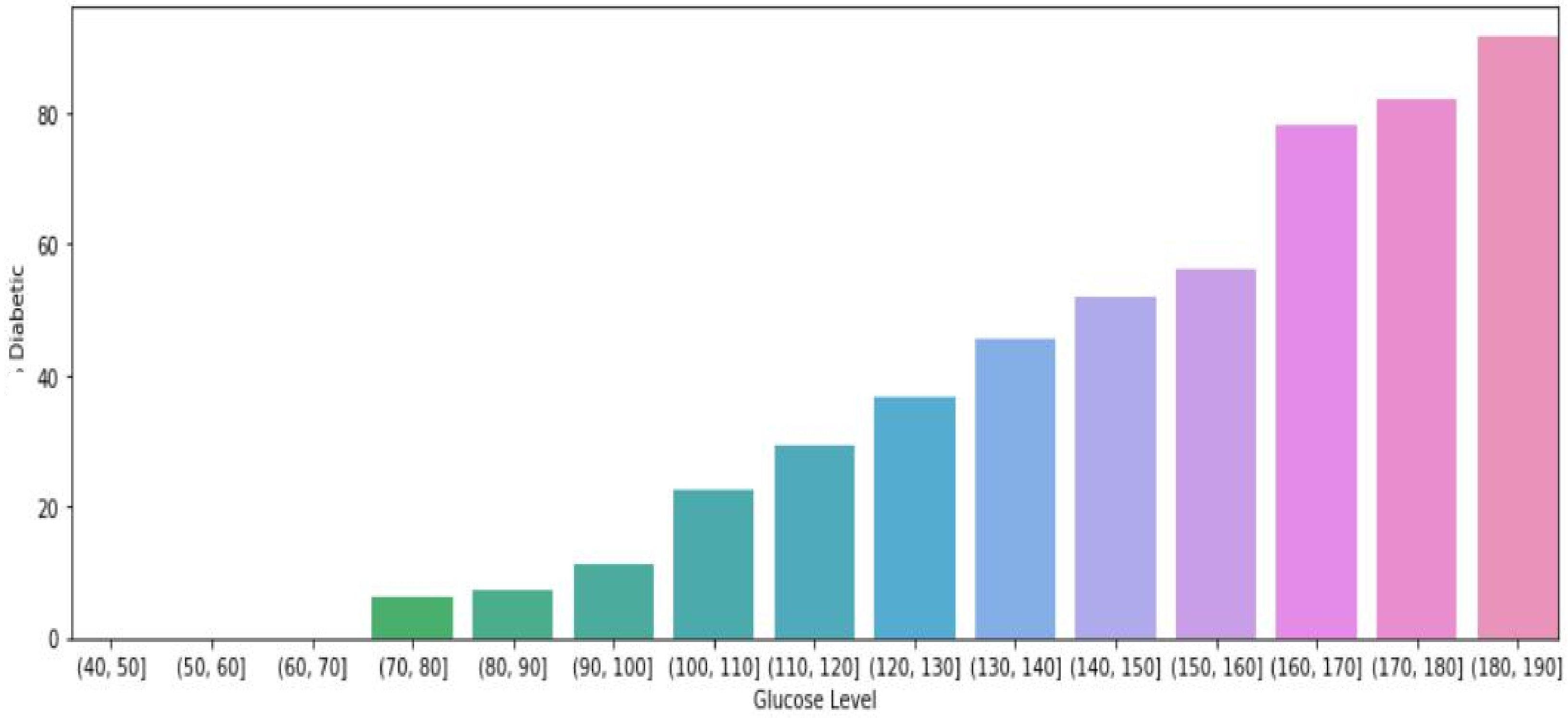
Fig 7:Body Mass Index effect on outcome

Fig 8:Relation between number of pregnancies and outcome



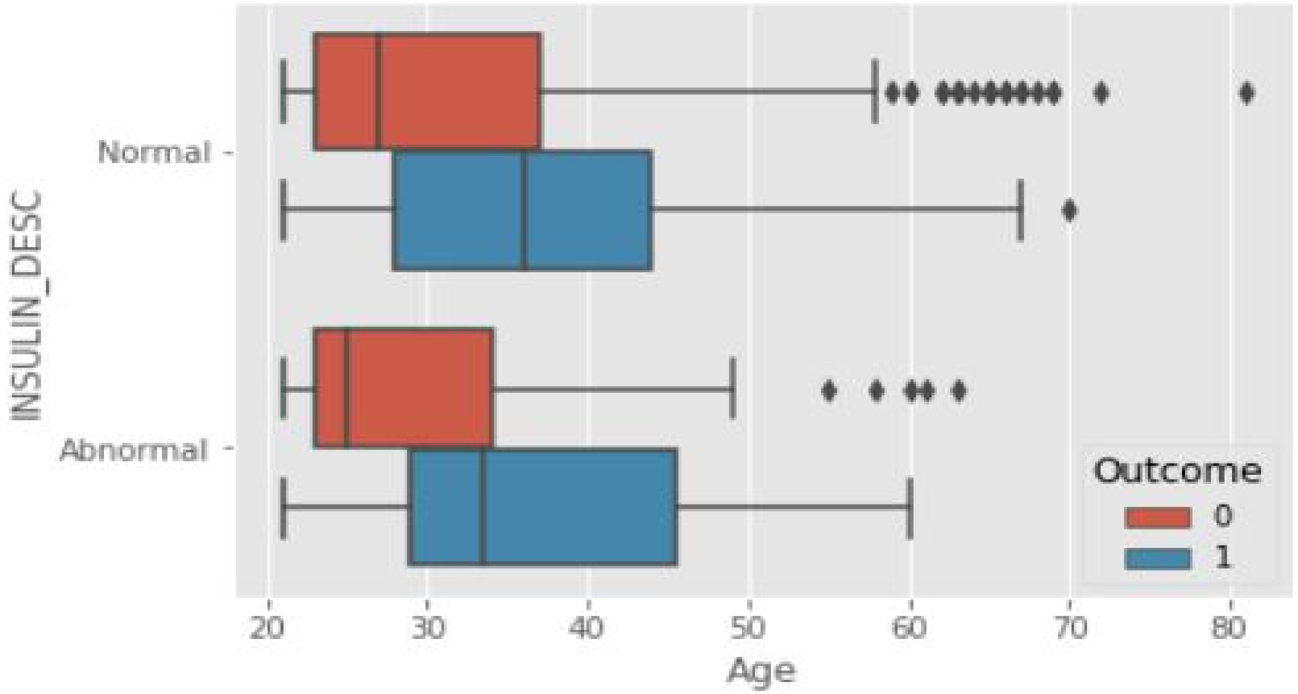
It can be seen that higher number of pregnancies increases the risk of diabetes significantly.

Fig 9:Glucose Level vs diabetic outcome



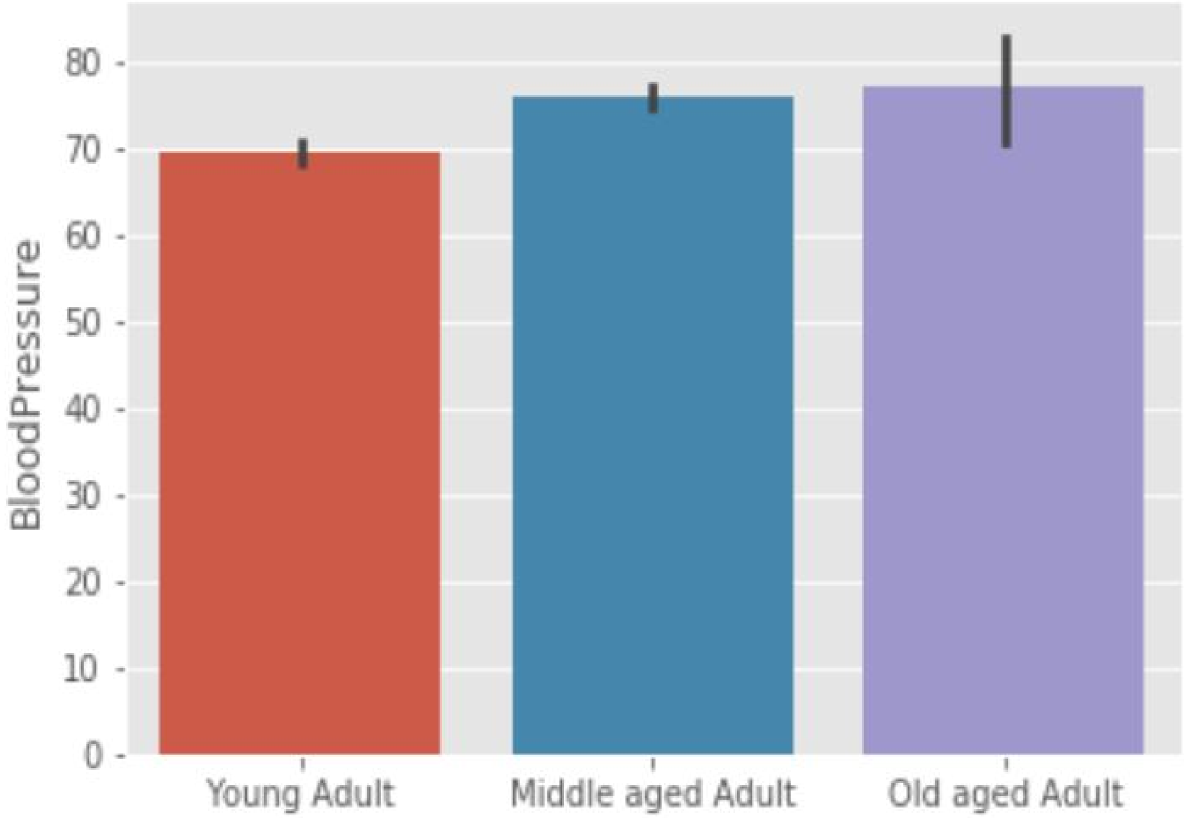
Most people having diabetes have glucose level in 180-190 level. Blood sugar spikes occur in people with diabetes because their body is unable to use insulin effectively.

Fig 10:Age vs Insulin_level



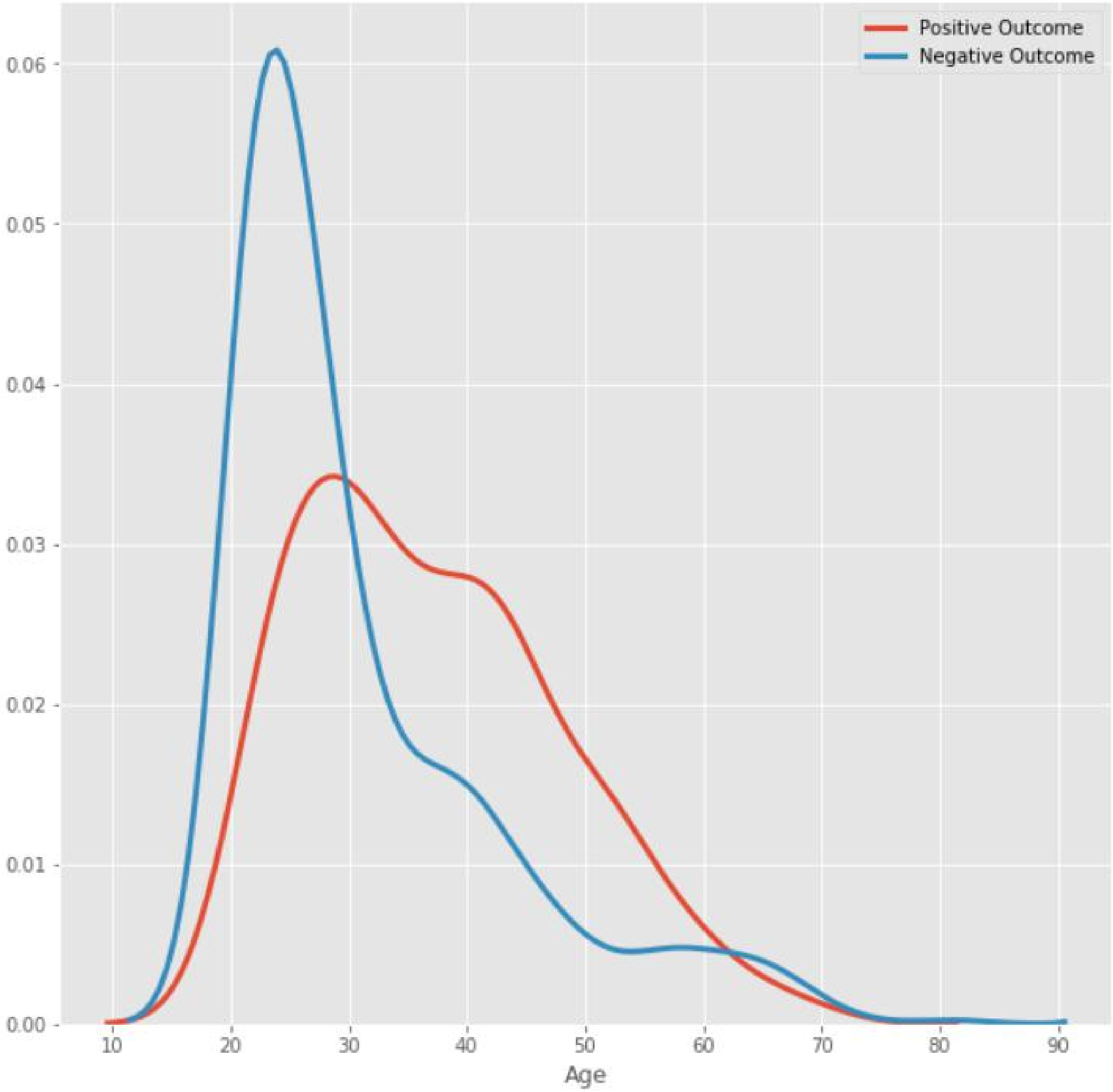
People having normal insulin levels having diabetes are within the age range from 25 and 45 whereas patients having abnormal insulin levels are more diabetic in the age range of late 20's to mid 40’s.

Fig 11: Relation between age and blood pressure.



Blood pressure increases with age mostly.

Fig 12: Distplot



It shows that Graph is right skewed

Mostly, diabetes starts to affect predominantly middle aged people, i.e., around 30.

CDF-Cumulative distribution function

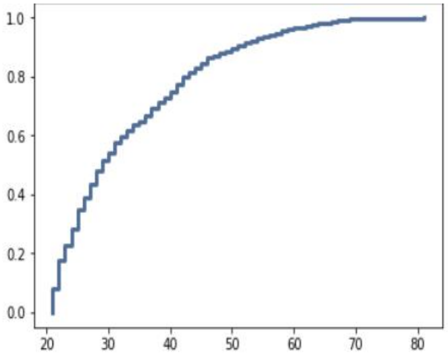


Fig 13.1:Age

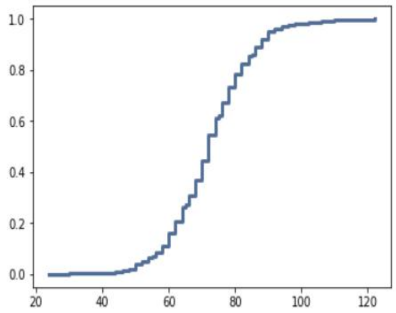


Fig 13.2:Blood Pressure

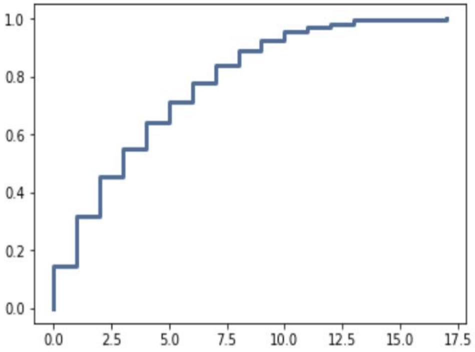


Fig 13.3:Pregnancies

PMF-Probability mass function

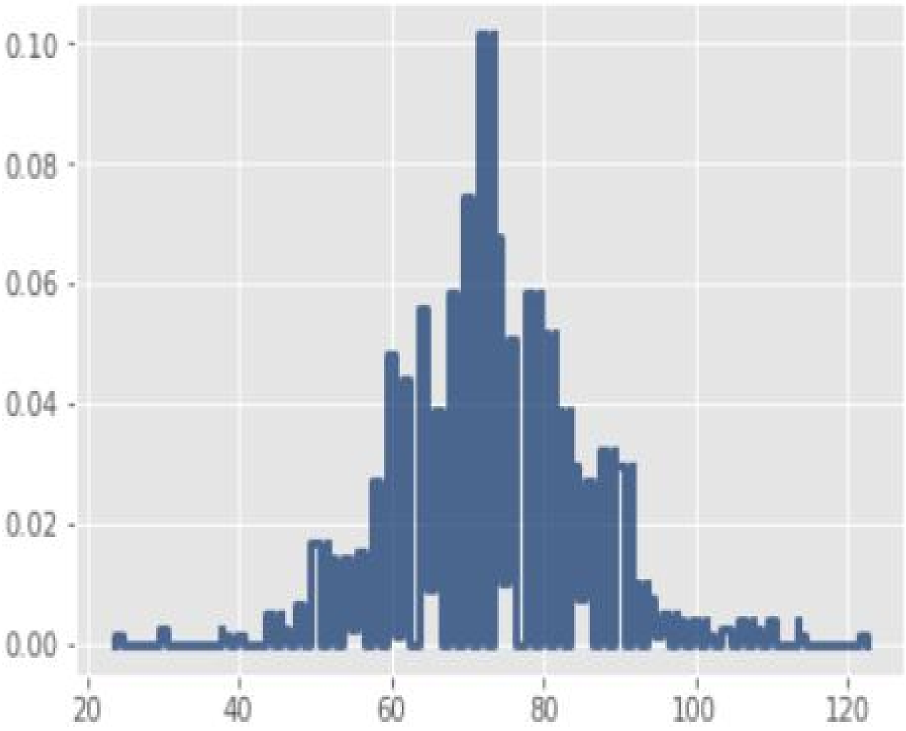


Fig 13.4 :Blood Pressure

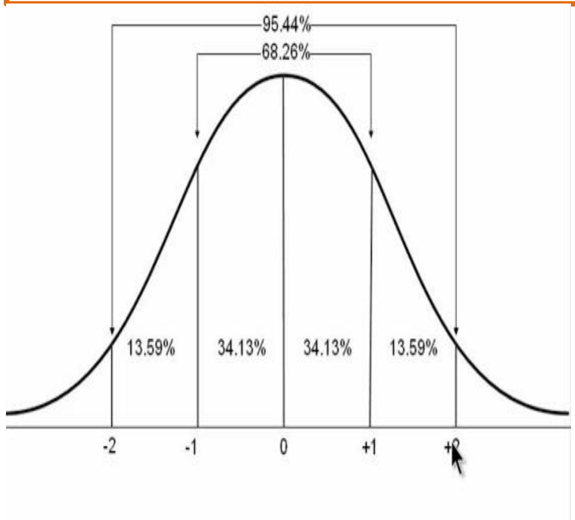
INTERPRETATION USING PROBABILITY DISTRIBUTIONS

NORMAL DISTRIBUTION

Q1: According to the survey conducted by W.H.O, it was observed that blood pressure has a major impact on diabetic level. A sample of 768 people with or without diabetes was studied having mean of 69.1 mm hg and standard deviation 19.35 mm hg. Find:-

(a)the percentage of people having blood pressure 90 mm hg or higher?
Sol: 14%

(b) the percentage of people having normal blood pressure?
Sol: 71%



Diastolic blood pressure

- Less than 80 is normal,
- Between 90 or 120 is high,
- higher than 120 is critical

NORMAL DISTRIBUTION

Q4 :High level of BMI can contribute to high risk of diabetes. BMI level for women follows a normal distribution with mean of 31.99258 and standard deviation 7.88416. BMI levels above 30 indicates obesity and demand medical attention.

a)What percentage of women aged above 45 need medical attention?

Sol: 8.46%

b)What percentage of people are under normal range of BMI(between 18.5 and 24.9)?

Sol: 14.07%

BMI ranges

- below 18.5 –underweight
- between 18.5 and 24.9 – normal weight
- between 25 and 29.9 –overweight
- 30 and above- obesity

BINOMIAL DISTRIBUTION

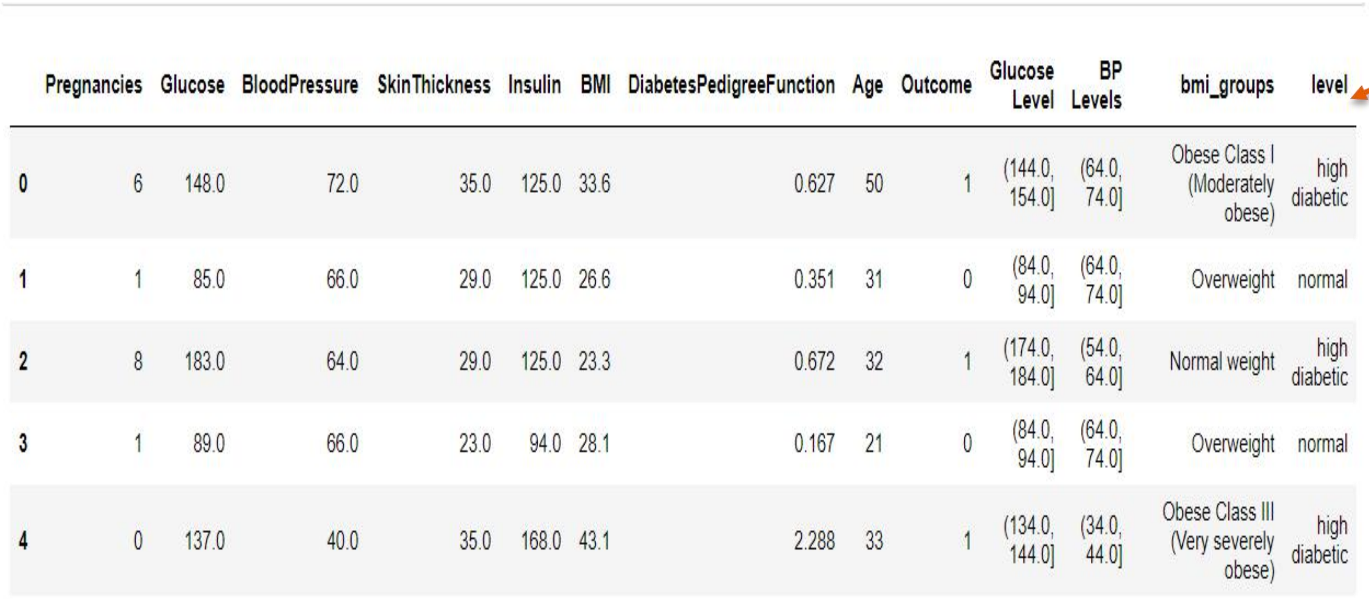
Q2. Analysis show that probability of people having high risk of diabetes is 0.24. Assume that data follows binomial distribution, what is the probability that more than 200 people out of total (768 people) are at high risk?

Sol: 0.26

GEOMETRIC DISTRIBUTION

Q3. The risk of people having diabetes due to inheritance is about 47%. Let X be the number of people you ask until one says he or she has diabetes. Then X is a discrete random variable with a geometric distribution as X~G(0.47).
What is the probability that you ask two people before one says he or she has diabetes?

Sol: 0.24



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome	Glucose Level	BP Levels	bmi_groups	level
0	6	148.0	72.0	35.0	125.0	33.6	0.627	50	1	(144.0, 154.0]	(64.0, 74.0]	Obese Class I (Moderately obese)	high diabetic
1	1	85.0	66.0	29.0	125.0	26.6	0.351	31	0	(84.0, 94.0]	(64.0, 74.0]	Overweight	normal
2	8	183.0	64.0	29.0	125.0	23.3	0.672	32	1	(174.0, 184.0]	(54.0, 64.0]	Normal weight	high diabetic
3	1	89.0	66.0	23.0	94.0	28.1	0.167	21	0	(84.0, 94.0]	(64.0, 74.0]	Overweight	normal
4	0	137.0	40.0	35.0	168.0	43.1	2.288	33	1	(134.0, 144.0]	(34.0, 44.0]	Obese Class III (Very severely obese)	high diabetic

Making
a new
column

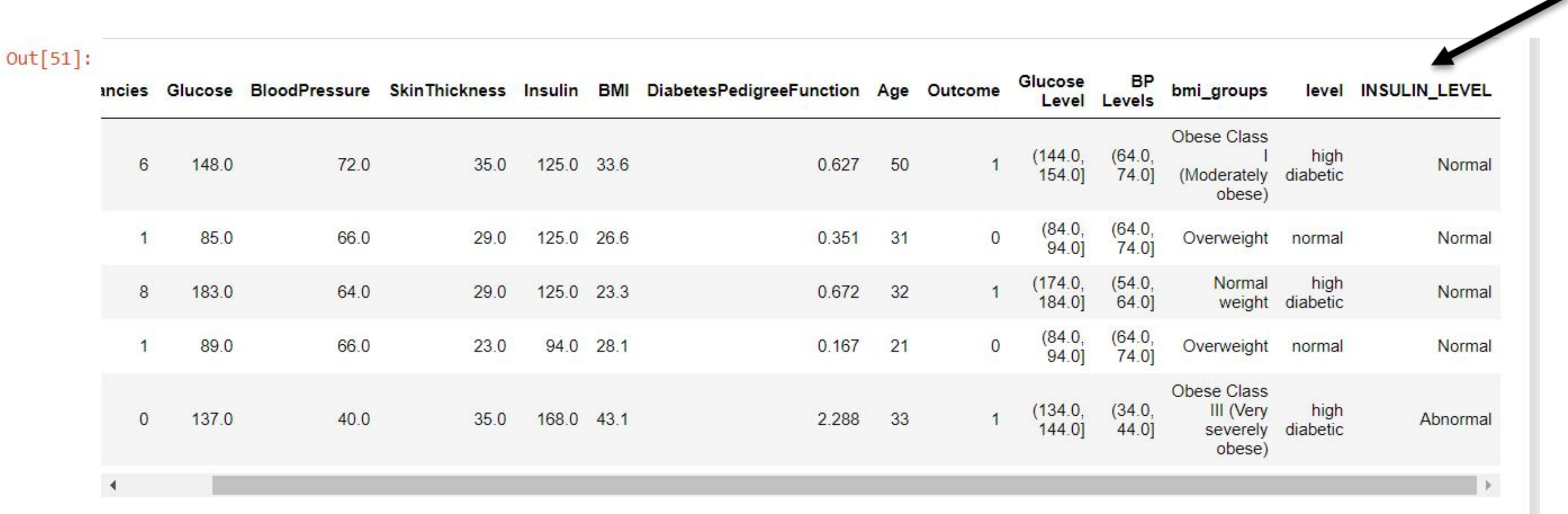
Q 5. A research shows that glucose have a major effect on diabetes level. A sample of 768 people with or without diabetes is studied. What is the probability of that people with high glucose level have diabetes

(a)Find the probability of people having high diabetes ??
Sol: 0.375

Diabetes level(glucose)

- <= 100 'normal'
- <=126 'prediabetic'
- else 'high diabetic'

new column



Out[51]:

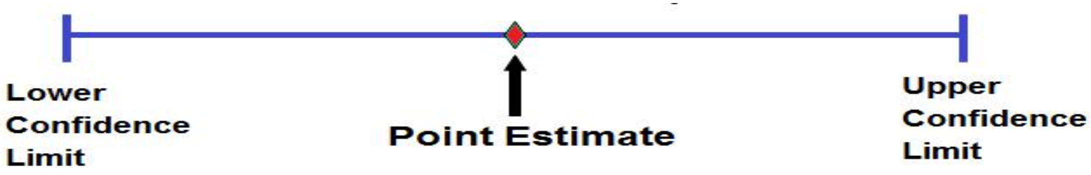
ancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome	Glucose Level	BP Levels	bmi_groups	level	INSULIN_LEVEL
6	148.0	72.0	35.0	125.0	33.6	0.627	50	1	(144.0, 154.0]	(64.0, 74.0]	Obese Class I (Moderately obese)	high diabetic	Normal
1	85.0	66.0	29.0	125.0	26.6	0.351	31	0	(84.0, 94.0]	(64.0, 74.0]	Overweight	normal	Normal
8	183.0	64.0	29.0	125.0	23.3	0.672	32	1	(174.0, 184.0]	(54.0, 64.0]	Normal weight	high diabetic	Normal
1	89.0	66.0	23.0	94.0	28.1	0.167	21	0	(84.0, 94.0]	(64.0, 74.0]	Overweight	normal	Normal
0	137.0	40.0	35.0	168.0	43.1	2.288	33	1	(134.0, 144.0]	(34.0, 44.0]	Obese Class III (Very severely obese)	high diabetic	Abnormal

Q 6) . A research shows that insulin has an effect on diabetes level. A sample of 768 people with or without diabetes is studied.)
(b)Find the probability of people having high insulin level ??
Sol: 0.174

Q7. The probability of people having high insulin level is is 0.174. Assume that data follows binomial distribution, what is the probability that more than 200 people out of total (768 people) have high insulin level?

Sol: 0.26

CONFIDENCE INTERVAL

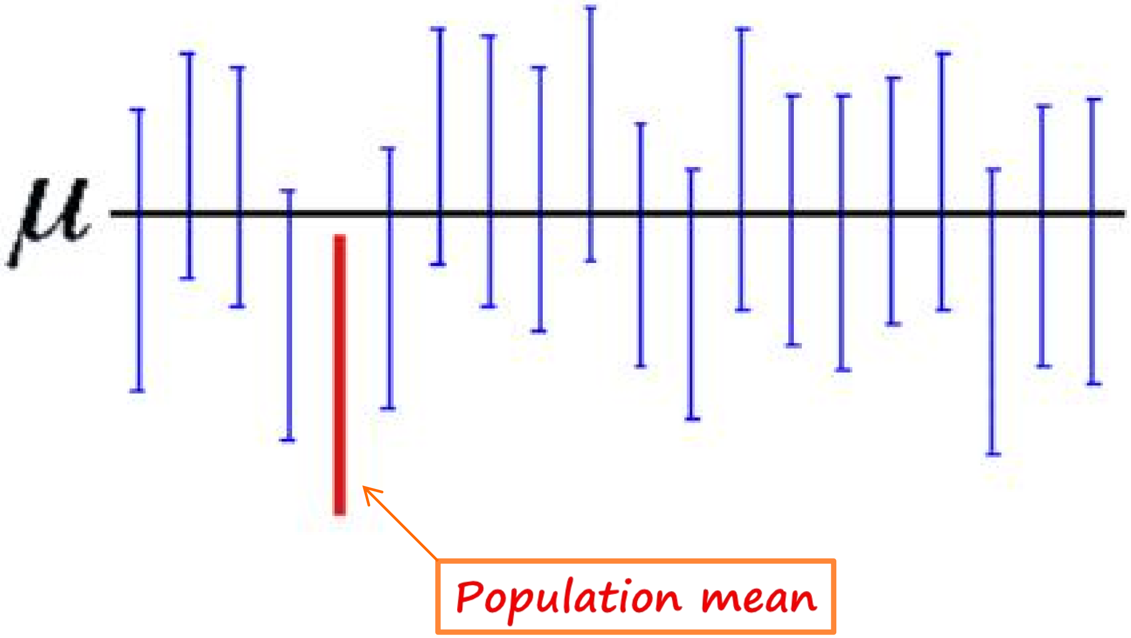


Lower Confidence Limit

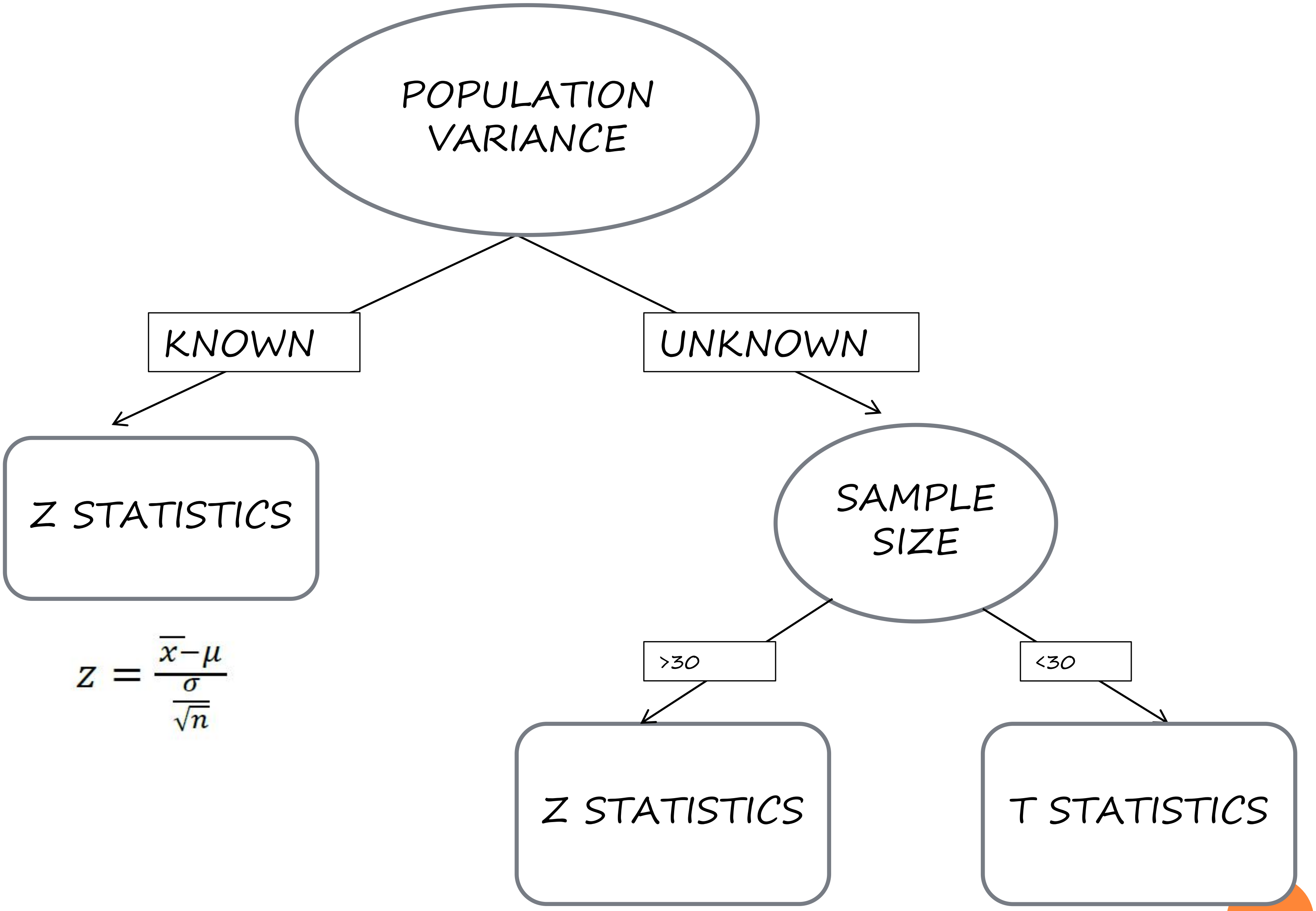
Point Estimate

Upper Confidence Limit

- SO WHAT DOES A 95% CONFIDENCE LEVEL MEAN?



Population mean



POPULATION VARIANCE

KNOWN

Z STATISTICS

$$z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

UNKNOWN

SAMPLE SIZE

>30

Z STATISTICS

$$z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

<30

T STATISTICS

$$t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$$

CASE 1

Q8.A random sample of 20 females was selected as a part of study on diabetes and their BMI was measured. The average BMI level for the sample was found to be 32 and standard deviation of 4.45. Assuming BMI to be normally distributed construct 97% C.I for mean BMI.

Sol: (24.70 , 34.27)

CASE 2

Q9. A random sample of 40 females was selected as a part of study on diabetes and their glucose level was measured. The average glucose for the sample was found to be 128 and standard deviation of 29 . Assuming glucose to be normally distributed in the sample, construct 95% ci for mean BMI.
NOTE:-Z stats will be used since N>30 EVEN if population variance is unknown

Sol: (100.2,120.29)

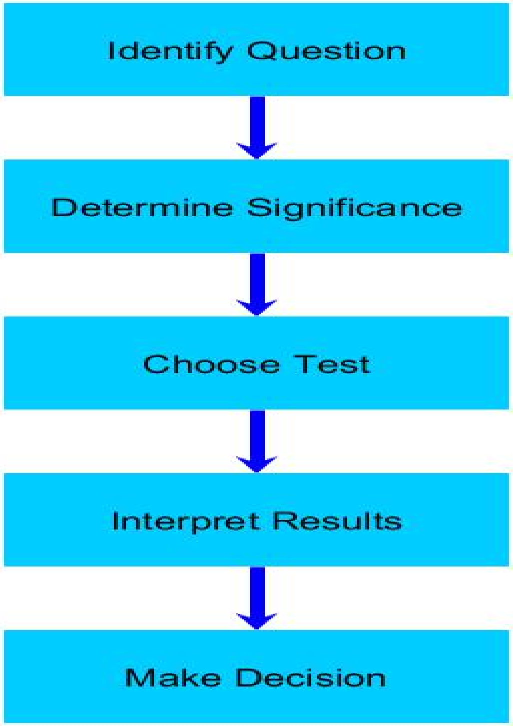
CASE 3

Q 10.A random sample of 20 females having diabetes was selected and their insulin level was measured. The population mean equals 80 mg/dL and the standard deviation is 56.The average insulin for the sample was found to be 53 mg/dL and standard deviation of 110.Construct a 90% ci for mean insulin.

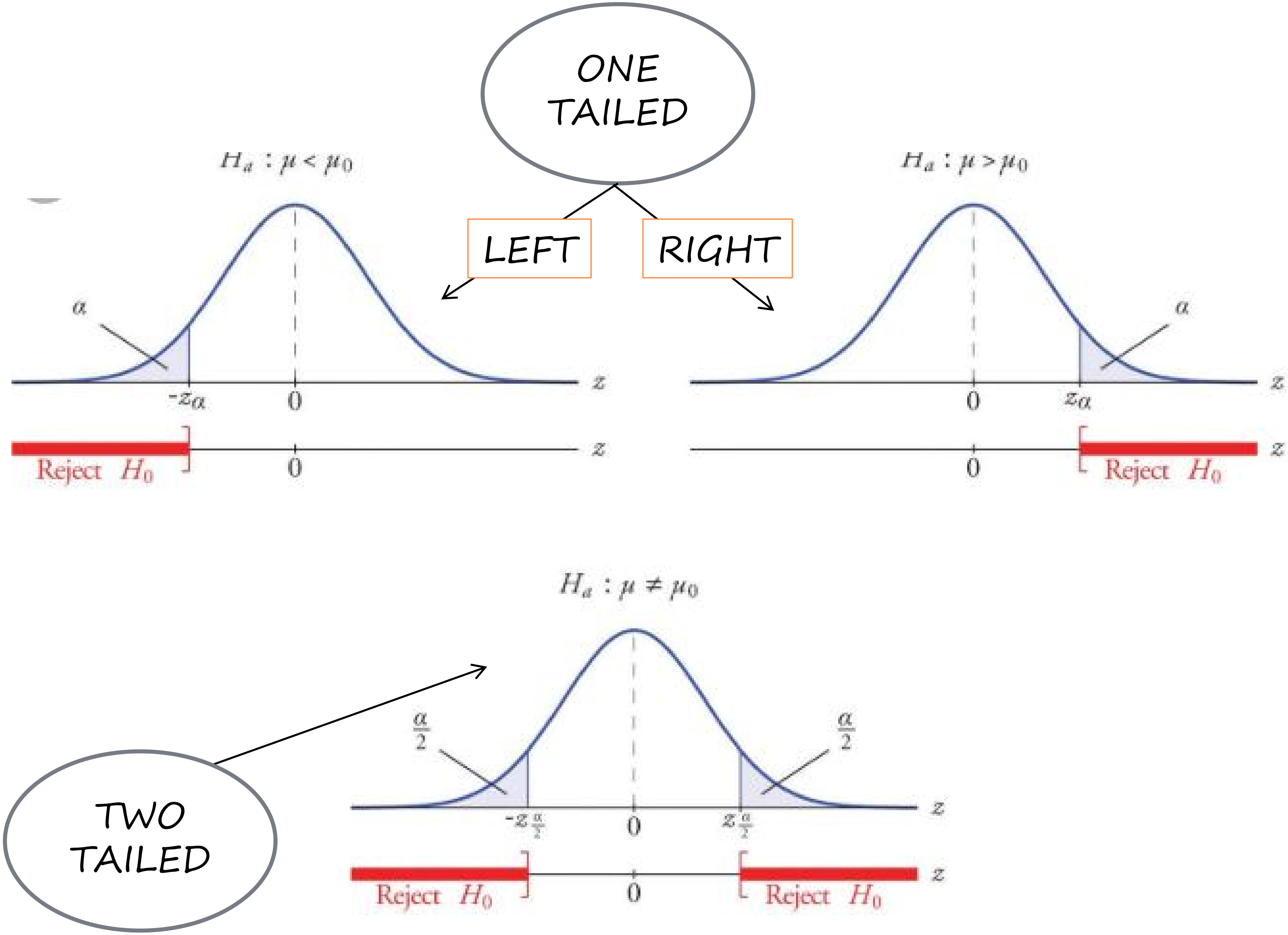
Sol: (71.01,112.08)

HYPOTHESIS TESTING

- *Hypothesis testing is the use of statistics to determine the probability that a given claim is true.*
- *Statistical analysts test a hypothesis by measuring and examining a random sample of the population being analyzed.*



Rejection region



ONE TAILED

LEFT

RIGHT

TWO TAILED

Q 11. A researcher claims that high BMI has a negative effect on diabetes level. A sample of 50 people having diabetes have a mean BMI level of 33.3kg/m ^{2} and standard deviation of 6.31 kg/m ^{2}.Test the hypothesis that BMI has effect on diabetes. Take significance level to be 0.05.

SOLUTION:

1.**Assumptions**: the population is approximately normally distributed

2.Deciding Hypothesis:

Ho : BMI has affect on diabetes

Ha : Null hypothesis is false

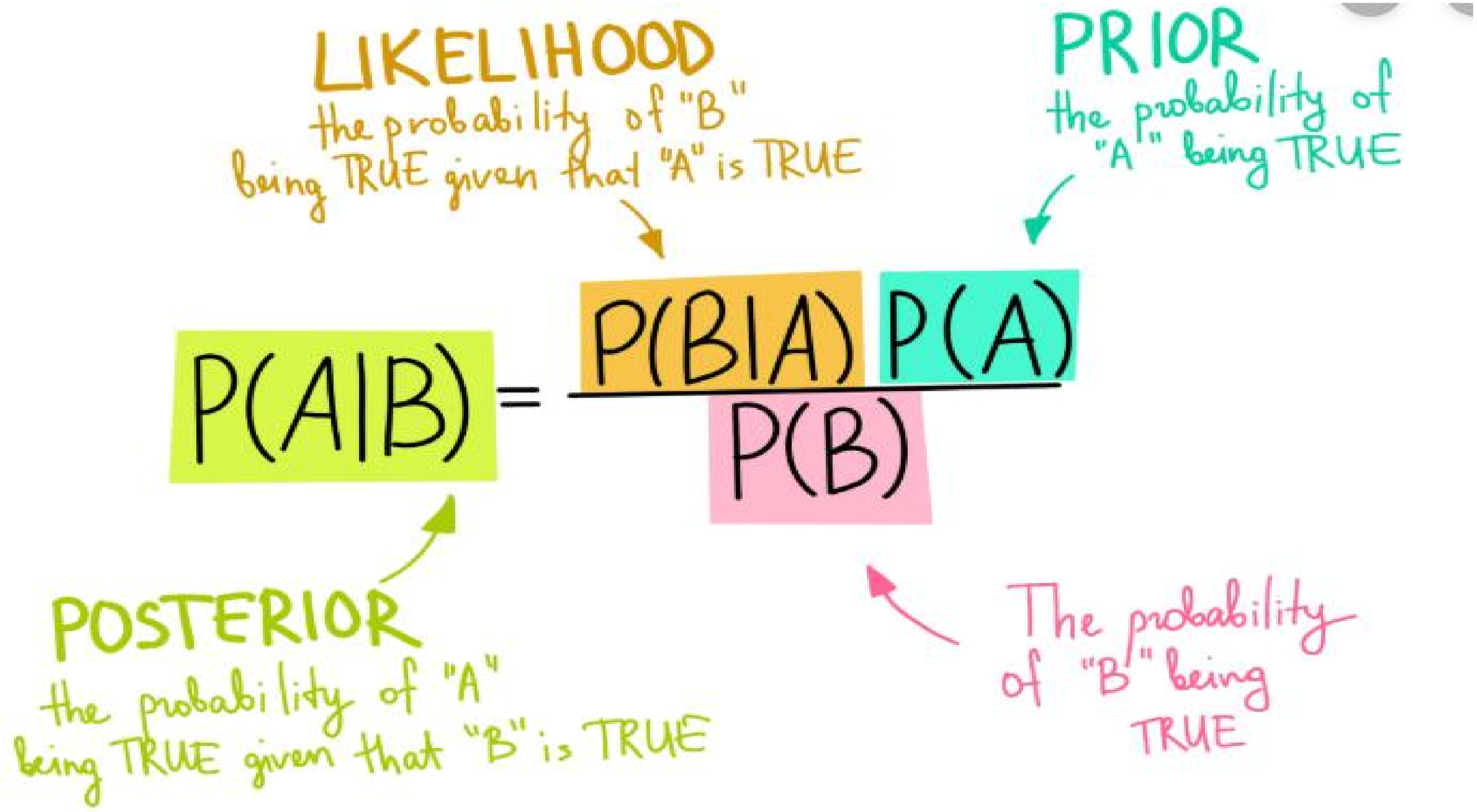
3.Statistic used: Since n>30, therefore, we will use z- statistics

4.Decision rule: If z-score > 1.64 (critical region), we will reject null hypothesis

5.**Decision**: **We fail to reject Ho** , since z score is in the rejection region.

We conclude that, **BMI has affect on diabetes of a patient**

BAYES THEOREM


LIKELIHOOD
the probability of "B" being TRUE given that "A" is TRUE
PRIOR
the probability of "A" being TRUE
P(A|B) = P(B|A) P(A) / P(B)
POSTERIOR
the probability of "A" being TRUE given that "B" is TRUE
The probability of "B" being TRUE

Q 12. According to the survey, 25% of population have high glucose level. 50% of population have high glucose level when they have diabetes. About 27% of people in a given year will have diabetes. What is the probability that a person will have diabetes if it is given that he/she has high glucose level?

SOL: P(A) : person has diabetes =0.27

P(B) : person have high glucose level = 0.25

P(B/A) : person have high glucose level given that he have diabetes= 0.50

P(A/B):**probability that a person will have diabetes if it is given that he/she has high glucose level**

therefore, through calculation, P(A/B)= 0.593

CHI SQUARE TEST

Chi-square test for independence is used to determine whether there is a significant relationship between two variables.

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

χ^2 = the test statistic $\quad \sum$ = the sum of

O = Observed frequencies E = Expected frequencies

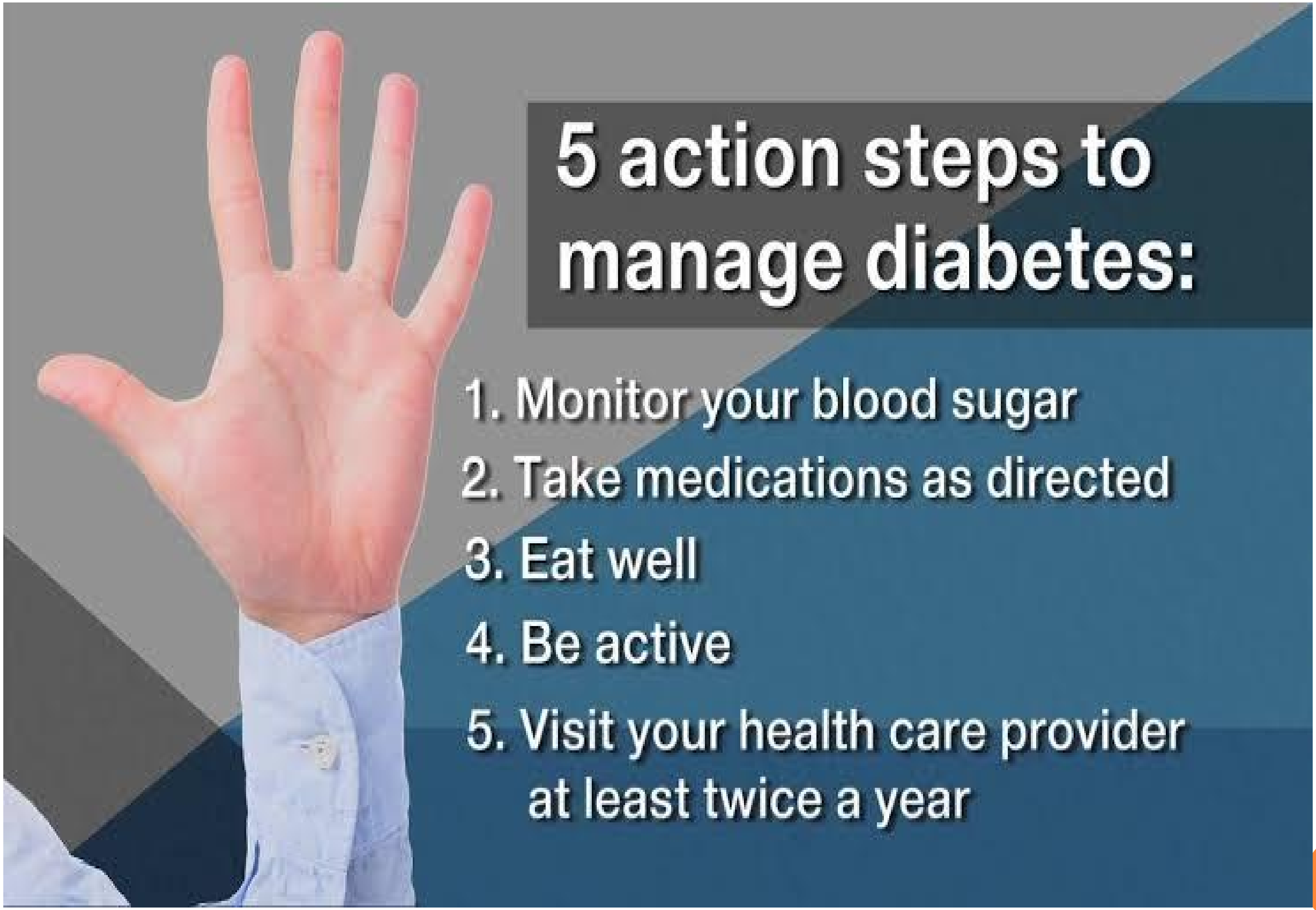
Q 13. A random sample of _women is surveyed and the relationship between skin thickening and bp is studied. The data that resulted from the survey is summarised. Is there enough evidence to say that they are dependent at 5% level of significance?

Sol: p value: 0.273
alpha: 0.050
therefore, failed to reject null hypothesis
INTERPRETATION:-

CONCLUSION

- In health related field , we have analyzed the serious growing problem of diabetes and the factors that influence it.

- There exists a dependency between the factors and some factors high correlation among them.



5 action steps to manage diabetes:

1. Monitor your blood sugar
2. Take medications as directed
3. Eat well
4. Be active
5. Visit your health care provider at least twice a year

REFERENCES

1. https://www.kaggle.com/uciml/pima-indians-diabetes-database
2. https://www.kaggle.com/devisangeetha/which-factor-causes-diabetes
3. https://www.journals.elsevier.com/diabetes-research-and-clinical-practice

THANK YOU