Analysis of diabetes in Indian women

- NAME : SOUVIK ROY
- ROLL NO. : 193012-21-0464
- UNIVERSITY REG. NO. : 012-1111-1821-19
- COLLEGE : ASUTOSH COLLEGE

- TOPIC : PROJECT WORK
- PAPER : DSE-B-2
- DATE : 14/07/2022

Contents

- **Executive Summary of the Project:**

Nowdays diabetes is a very big problem in society, specially in women , because women have some more number of diabetic factors in their life(like pregnancies). So through my project I'm taking eight covariates and one response variable. Then I'm trying to see the how the observations, for 256 patients regarding to each covariate, are distributed, so here we take the help of one of the important statistical tools boxplot. Then I try to show the relationships between each of the nine variables(covariate and response) and also show the correlatin map. Then as the response(study) variable is a binary variable so I use binary logistic regression to analyse our data.

- **Introduction:**

Researchers have identified a bunch of districts in India that have the maximum prevalence for diabetes among women. At least 50 of the 640 districts studied have high prevalence of diabetes — greater than one in 10 — among women aged 35-49 years. Tamil Nadu, Kerala, Andhra Pradesh and Odisha have districts with the highest prevalence. The results were published in the Journal of Diabetes & Metabolic Disorders.

While Cuttack in Odisha has the highest prevalence of 20%, 14 districts in Tamil Nadu — the maximum among all States — have high prevalence, prompting the researchers to classify them as 'hotspots'.

In all, 254 districts have a "very high level" (greater than 10.7%) of diabetes burden, and 130 have a moderately high (8.7-10.6%) burden. The burden is higher in the southern and eastern parts of the country and lowest in central India.

The researchers sourced data from the National Family Health Survey-4 (2015-16) as it provides district-level health indicators for women. Demographic details of 2,35,056 women from 36 States/Union Territories were analysed for gleaning disease spread and analysing relationship among disease and socio-economic category, location, number of children, obesity and hypertension among others. This was also the first NHS survey to collected blood glucose levels in men and women thus helping determine diabetes.

- **Collected data:**

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
9	91	68	0	0	24.2	0.2	58	0
2	91	62	0	0	27.3	0.525	22	0
3	99	54	19	86	25.6	0.154	24	0
3	163	70	18	105	31.6	0.268	28	1
9	145	88	34	165	30.3	0.771	53	1
7	125	86	0	0	37.6	0.304	51	0
13	76	60	0	0	32.8	0.18	41	0
6	129	90	7	326	19.6	0.582	60	0
2	68	70	32	66	25	0.187	25	0
3	124	80	33	130	33.2	0.305	26	0
6	114	0	0	0	0	0.189	26	0
9	130	70	0	0	34.2	0.652	45	1
3	125	58	0	0	31.6	0.151	24	0
3	87	60	18	0	21.8	0.444	21	0
1	97	64	19	82	18.2	0.299	21	0
3	116	74	15	105	26.3	0.107	24	0
0	117	66	31	188	30.8	0.493	22	0
0	111	65	0	0	24.6	0.66	31	0
2	122	60	18	106	29.8	0.717	22	0
0	107	76	0	0	45.3	0.686	24	0
1	86	66	52	65	41.3	0.917	29	0
6	91	0	0	0	29.8	0.501	31	0
1	77	56	30	56	33.3	1.251	24	0
4	132	0	0	0	32.9	0.302	23	1
0	105	90	0	0	29.6	0.197	46	0
0	57	60	0	0	21.7	0.735	67	0
0	127	80	37	210	36.3	0.804	23	0
3	129	92	49	155	36.4	0.968	32	1
8	100	74	40	215	39.4	0.661	43	1
3	128	72	25	190	32.4	0.549	27	1
10	90	85	32	0	34.9	0.825	56	1
4	84	90	23	56	39.5	0.159	25	0
1	88	78	29	76	32	0.365	29	0
8	186	90	35	225	34.5	0.423	37	1

5	187	76	27	207	43.6	1.034	53	1
4	131	68	21	166	33.1	0.16	28	0
1	164	82	43	67	32.8	0.341	50	0
4	189	110	31	0	28.5	0.68	37	0
1	116	70	28	0	27.4	0.204	21	0
3	84	68	30	106	31.9	0.591	25	0
6	114	88	0	0	27.8	0.247	66	0
1	88	62	24	44	29.9	0.422	23	0
1	84	64	23	115	36.9	0.471	28	0
7	124	70	33	215	25.5	0.161	37	0
1	97	70	40	0	38.1	0.218	30	0
8	110	76	0	0	27.8	0.237	58	0
11	103	68	40	0	46.2	0.126	42	0
11	85	74	0	0	30.1	0.3	35	0
6	125	76	0	0	33.8	0.121	54	1
0	198	66	32	274	41.3	0.502	28	1
1	87	68	34	77	37.6	0.401	24	0
6	99	60	19	54	26.9	0.497	32	0
0	91	80	0	0	32.4	0.601	27	0
2	95	54	14	88	26.1	0.748	22	0
1	99	72	30	18	38.6	0.412	21	0
6	92	62	32	126	32	0.085	46	0
4	154	72	29	126	31.3	0.338	37	0
0	121	66	30	165	34.3	0.203	33	1
3	78	70	0	0	32.5	0.27	39	0
2	130	96	0	0	22.6	0.268	21	0
3	111	58	31	44	29.5	0.43	22	0
2	98	60	17	120	34.7	0.198	22	0
1	143	86	30	330	30.1	0.892	23	0
1	119	44	47	63	35.5	0.28	25	0
6	108	44	20	130	24	0.813	35	0
2	118	80	0	0	42.9	0.693	21	1
10	133	68	0	0	27	0.245	36	0
2	197	70	99	0	34.7	0.575	62	1
0	151	90	46	0	42.1	0.371	21	1
6	109	60	27	0	25	0.206	27	0
12	121	78	17	0	26.5	0.259	62	0
8	100	76	0	0	38.7	0.19	42	0
8	124	76	24	600	28.7	0.687	52	1
1	93	56	11	0	22.5	0.417	22	0
8	143	66	0	0	34.9	0.129	41	1
6	103	66	0	0	24.3	0.249	29	0
3	176	86	27	156	33.3	1.154	52	1
0	73	0	0	0	21.1	0.342	25	0
11	111	84	40	0	46.8	0.925	45	1
2	112	78	50	140	39.4	0.175	24	0

3	132	80	0	0	34.4	0.402	44	1
2	82	52	22	115	28.5	1.699	25	0
6	123	72	45	230	33.6	0.733	34	0
0	188	82	14	185	32	0.682	22	1
0	67	76	0	0	45.3	0.194	46	0
1	89	24	19	25	27.8	0.559	21	0
1	173	74	0	0	36.8	0.088	38	1
1	109	38	18	120	23.1	0.407	26	0
1	108	88	19	0	27.1	0.4	24	0
6	96	0	0	0	23.7	0.19	28	0
1	124	74	36	0	27.8	0.1	30	0
7	150	78	29	126	35.2	0.692	54	1
4	183	0	0	0	28.4	0.212	36	1
1	124	60	32	0	35.8	0.514	21	0
1	181	78	42	293	40	1.258	22	1
1	92	62	25	41	19.5	0.482	25	0
0	152	82	39	272	41.5	0.27	27	0
1	111	62	13	182	24	0.138	23	0
3	106	54	21	158	30.9	0.292	24	0
3	174	58	22	194	32.9	0.593	36	1
7	168	88	42	321	38.2	0.787	40	1
6	105	80	28	0	32.5	0.878	26	0
11	138	74	26	144	36.1	0.557	50	1
3	106	72	0	0	25.8	0.207	27	0
6	117	96	0	0	28.7	0.157	30	0
2	68	62	13	15	20.1	0.257	23	0
9	112	82	24	0	28.2	1.282	50	1
0	119	0	0	0	32.4	0.141	24	1
2	112	86	42	160	38.4	0.246	28	0
2	92	76	20	0	24.2	1.698	28	0
6	183	94	0	0	40.8	1.461	45	0
0	94	70	27	115	43.5	0.347	21	0
2	108	64	0	0	30.8	0.158	21	0
4	90	88	47	54	37.7	0.362	29	0
0	125	68	0	0	24.7	0.206	21	0
0	132	78	0	0	32.4	0.393	21	0
5	128	80	0	0	34.6	0.144	45	0
4	94	65	22	0	24.7	0.148	21	0
7	114	64	0	0	27.4	0.732	34	1
0	102	78	40	90	34.5	0.238	24	0
2	111	60	0	0	26.2	0.343	23	0
1	128	82	17	183	27.5	0.115	22	0
10	92	62	0	0	25.9	0.167	31	0
13	104	72	0	0	31.2	0.465	38	1
5	104	74	0	0	28.8	0.153	48	0
2	94	76	18	66	31.6	0.649	23	0

7	97	76	32	91	40.9	0.871	32	1
1	100	74	12	46	19.5	0.149	28	0
0	102	86	17	105	29.3	0.695	27	0
4	128	70	0	0	34.3	0.303	24	0
6	147	80	0	0	29.5	0.178	50	1
4	90	0	0	0	28	0.61	31	0
3	103	72	30	152	27.6	0.73	27	0
2	157	74	35	440	39.4	0.134	30	0
1	167	74	17	144	23.4	0.447	33	1
0	179	50	36	159	37.8	0.455	22	1
11	136	84	35	130	28.3	0.26	42	1
0	107	60	25	0	26.4	0.133	23	0
1	91	54	25	100	25.2	0.234	23	0
1	117	60	23	106	33.8	0.466	27	0
5	123	74	40	77	34.1	0.269	28	0
2	120	54	0	0	26.8	0.455	27	0
1	106	70	28	135	34.2	0.142	22	0
2	155	52	27	540	38.7	0.24	25	1
2	101	58	35	90	21.8	0.155	22	0
1	120	80	48	200	38.9	1.162	41	0
11	127	106	0	0	39	0.19	51	0
3	80	82	31	70	34.2	1.292	27	1
10	162	84	0	0	27.7	0.182	54	0
1	199	76	43	0	42.9	1.394	22	1
8	167	106	46	231	37.6	0.165	43	1
9	145	80	46	130	37.9	0.637	40	1
6	115	60	39	0	33.7	0.245	40	1
1	112	80	45	132	34.8	0.217	24	0
4	145	82	18	0	32.5	0.235	70	1
10	111	70	27	0	27.5	0.141	40	1
6	98	58	33	190	34	0.43	43	0
9	154	78	30	100	30.9	0.164	45	0
6	165	68	26	168	33.6	0.631	49	0
1	99	58	10	0	25.4	0.551	21	0
10	68	106	23	49	35.5	0.285	47	0
3	123	100	35	240	57.3	0.88	22	0
8	91	82	0	0	35.6	0.587	68	0
6	195	70	0	0	30.9	0.328	31	1
9	156	86	0	0	24.8	0.23	53	1
0	93	60	0	0	35.3	0.263	25	0
3	121	52	0	0	36	0.127	25	1
2	101	58	17	265	24.2	0.614	23	0
2	56	56	28	45	24.2	0.332	22	0
0	162	76	36	0	49.6	0.364	26	1
0	95	64	39	105	44.6	0.366	22	0
4	125	80	0	0	32.3	0.536	27	1

5	136	82	0	0	0	0.64	69	0
2	129	74	26	205	33.2	0.591	25	0
3	130	64	0	0	23.1	0.314	22	0
1	107	50	19	0	28.3	0.181	29	0
1	140	74	26	180	24.1	0.828	23	0
1	144	82	46	180	46.1	0.335	46	1
8	107	80	0	0	24.6	0.856	34	0
13	158	114	0	0	42.3	0.257	44	1
2	121	70	32	95	39.1	0.886	23	0
7	129	68	49	125	38.5	0.439	43	1
2	90	60	0	0	23.5	0.191	25	0
7	142	90	24	480	30.4	0.128	43	1
3	169	74	19	125	29.9	0.268	31	1
0	99	0	0	0	25	0.253	22	0
4	127	88	11	155	34.5	0.598	28	0
4	118	70	0	0	44.5	0.904	26	0
2	122	76	27	200	35.9	0.483	26	0
6	125	78	31	0	27.6	0.565	49	1
1	168	88	29	0	35	0.905	52	1
2	129	0	0	0	38.5	0.304	41	0
4	110	76	20	100	28.4	0.118	27	0
6	80	80	36	0	39.8	0.177	28	0
10	115	0	0	0	0	0.261	30	1
2	127	46	21	335	34.4	0.176	22	0
9	164	78	0	0	32.8	0.148	45	1
2	93	64	32	160	38	0.674	23	1
3	158	64	13	387	31.2	0.295	24	0
5	126	78	27	22	29.6	0.439	40	0
10	129	62	36	0	41.2	0.441	38	1
0	134	58	20	291	26.4	0.352	21	0
3	102	74	0	0	29.5	0.121	32	0
7	187	50	33	392	33.9	0.826	34	1
3	173	78	39	185	33.8	0.97	31	1
10	94	72	18	0	23.1	0.595	56	0
1	108	60	46	178	35.5	0.415	24	0
5	97	76	27	0	35.6	0.378	52	1
4	83	86	19	0	29.3	0.317	34	0
1	114	66	36	200	38.1	0.289	21	0
1	149	68	29	127	29.3	0.349	42	1
5	117	86	30	105	39.1	0.251	42	0
1	111	94	0	0	32.8	0.265	45	0
4	112	78	40	0	39.4	0.236	38	0
1	116	78	29	180	36.1	0.496	25	0
0	141	84	26	0	32.4	0.433	22	0
2	175	88	0	0	22.9	0.326	22	0
2	92	52	0	0	30.1	0.141	22	0

3	130	78	23	79	28.4	0.323	34	1
8	120	86	0	0	28.4	0.259	22	1
2	174	88	37	120	44.5	0.646	24	1
2	106	56	27	165	29	0.426	22	0
2	105	75	0	0	23.3	0.56	53	0
4	95	60	32	0	35.4	0.284	28	0
0	126	86	27	120	27.4	0.515	21	0
8	65	72	23	0	32	0.6	42	0
2	99	60	17	160	36.6	0.453	21	0
1	102	74	0	0	39.5	0.293	42	1
11	120	80	37	150	42.3	0.785	48	1
3	102	44	20	94	30.8	0.4	26	0
1	109	58	18	116	28.5	0.219	22	0
9	140	94	0	0	32.7	0.734	45	1
13	153	88	37	140	40.6	1.174	39	0
12	100	84	33	105	30	0.488	46	0
1	147	94	41	0	49.3	0.358	27	1
1	81	74	41	57	46.3	1.096	32	0
3	187	70	22	200	36.4	0.408	36	1
6	162	62	0	0	24.3	0.178	50	1
4	136	70	0	0	31.2	1.182	22	1
1	121	78	39	74	39	0.261	28	0
3	108	62	24	0	26	0.223	25	0
0	181	88	44	510	43.3	0.222	26	1
8	154	78	32	0	32.4	0.443	45	1
1	128	88	39	110	36.5	1.057	37	1
7	137	90	41	0	32	0.391	39	0
0	123	72	0	0	36.3	0.258	52	1
1	106	76	0	0	37.5	0.197	26	0
6	190	92	0	0	35.5	0.278	66	1
2	88	58	26	16	28.4	0.766	22	0
9	170	74	31	0	44	0.403	43	1
9	89	62	0	0	22.5	0.142	33	0
10	101	76	48	180	32.9	0.171	63	0
2	122	70	27	0	36.8	0.34	27	0
5	121	72	23	112	26.2	0.245	30	0
1	126	60	0	0	30.1	0.349	47	1
1	93	70	31	0	30.4	0.315	23	0

- ***METHODOLOGY:***

Brief description of the process of analysis-

I. Correlation coefficient:

Correlation coefficients are used to measure how strong a relationship is between two variables. There are several types of correlation coefficient, but the most popular is Pearson's. Pearson's correlation (also called Pearson's R) is a correlation coefficient commonly used in linear regression. Formula of computing correlation coefficient is-

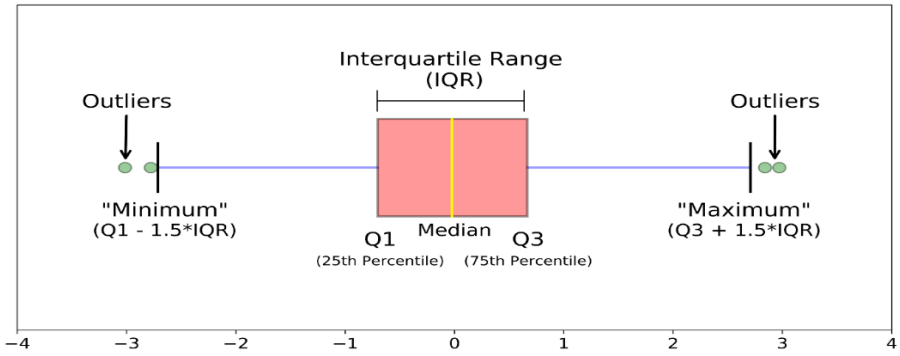
$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

II. Correlation matrix:

A correlation matrix is simply a table which displays the correlation coefficients for different variables. The matrix depicts the correlation between all the possible pairs of values in a table. It is a powerful tool to summarize a large dataset and to identify and visualize patterns in the given data. Example of a correlation matrix-

	Always to vote in elections	Never to try to evade taxes	Always to obey laws	Keep watch on action of govt	Active in social/ political associations	Understand others' points of view	Choose products for politics/ ethics/ envir.	Help worse off people in America	Help worse off people in rest of World
Always to vote in elections	1.00	.94	.94	.94	.92	.92	.89	.93	.88
Never to try to evade taxes	.94	1.00	.97	.95	.90	.94	.91	.95	.89
Always to obey laws	.94	.97	1.00	.96	.91	.94	.91	.96	.90
Keep watch on action of govt	.94	.95	.96	1.00	.93	.95	.91	.95	.89
Active in social/political associations	.92	.90	.91	.93	1.00	.92	.88	.91	.87
Understand others' points of view	.92	.94	.94	.95	.92	1.00	.91	.94	.89
Choose products for politics/ethics/envi	.89	.91	.91	.91	.88	.91	1.00	.91	.86
Help worse off people in America	.93	.95	.96	.95	.91	.94	.91	1.00	.93
Help worse off people in rest of World	.88	.89	.90	.89	.87	.89	.86	.93	1.00

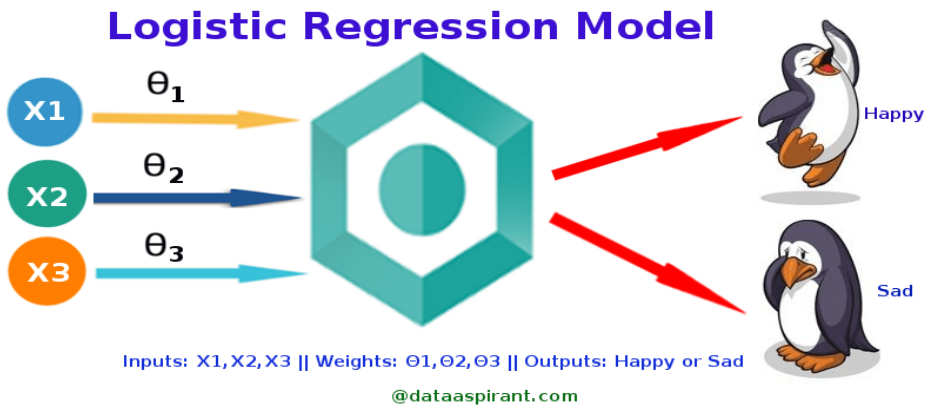
III. Boxplot:



The image above is a boxplot. A boxplot is a standardized way of displaying the distribution of data based on a five number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum"). It can tell you about your outliers and what their values are. It can also tell you if your data is symmetrical, how tightly your data is grouped, and if and how your data is skewed.

IV. Binary logistic regression:

Binary logistic regression (LR) is a regression model where the target variable is binary, that is, it can take only two values, 0 or 1. It is the most utilized regression model in readmission prediction, given that the output is modelled as readmitted (1) or not readmitted (0).



V. Hosmer-Lameshow goodness of fit :

The Hosmer-Lemeshow test (HL test) is a goodness of fit test for logistic regression, especially for risk prediction models. A goodness of fit test tells you how well your data fits the model. Specifically, the HL test calculates if the observed event rates match the expected event rates in population subgroups.

The test is only used for binary response variables (a variable with two outcomes like alive or dead, yes or no).

VI. Outliers:

In statistics, an **outlier** is a data point that differs significantly from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set. An outlier can cause serious problems in statistical analyses.

Outliers can occur by chance in any distribution, but they often indicate either measurement error or that the population has a heavy-tailed distribution. In the former case one wishes to discard them or use statistics that are robust to outliers, while in the latter case they indicate that the distribution has high skewness and that one should be very cautious in using tools or intuitions that assume a normal distribution. A frequent cause of outliers is a mixture of two distributions, which may be two distinct sub-populations, or may indicate 'correct trial' versus 'measurement error; this is modeled by a mixture model.

In most larger samplings of data, some data points will be further away from the sample mean than what is deemed reasonable. This can be due to incidental systematic error or flaws in the theory that generated an assumed family of probability distributions, or it may be that some observations are far from the center of the data. Outlier points can therefore indicate faulty data, erroneous procedures, or areas where a certain theory might not be valid. However, in large samples, a small number of outliers is to be expected (and not due to any anomalous condition).

VII. Mean:

In statistics, the mean is one of the measures of central tendency, apart from the mode and median. Mean is nothing but the average of the given set of values. It denotes the equal distribution of values for a given data set. The mean, median and mode are the three commonly used measures of central tendency. To calculate the mean, we need to add the total values given in a datasheet and divide the sum by the total number of values. The Median is the middle value of a given data when all the values are arranged in ascending order. Whereas mode is the number in the list, which is repeated a maximum number of times.

VIII. Median:

The median of a set of data is the middlemost number or center value in the set. The median is also the number that is halfway into the set.To find the median, the data should be arranged, first, in order of least to greatest or greatest to the least value. A median is a number that is separated by the higher half of a data sample, a population or a probability distribution, from the lower half. The median is different for different types of distribution.

For example, the median of 3, 3, 5, 9, 11 is 5. If there is an even number of observations, then there is no single middle value; the median is then usually defined to be the mean of the two middle values: so the median of 3, 5, 7, 9 is (5+7)/2 = 6.

IX. Mode:

A mode is defined as the value that has a higher frequency in a given set of values. It is the value that appears the most number of times.

Example: In the given set of data: 2, 4, 5, 5, 6, 7, the mode of the data set is 5 since it has appeared in the set twice.

Statistics deals with the presentation, collection and analysis of data and information for a particular purpose. We use tables, graphs, pie charts, bar graphs, pictorial representation, etc. After the proper organization of the data, it must be further analyzed to infer helpful information.

For this purpose, frequently in statistics, we tend to represent a set of data by a representative value that roughly defines the entire data collection. This representative value is known as the measure of central tendency. By the name itself, it suggests that it is a value around which the data is centred. These measures of central tendency allow us to create a statistical summary of the vast, organized data. One such measure of central tendency is the mode of data.

- ***Computation:***

At first we are to import our csv file named as “project.csv” saved in our system. The required code is-

```
>data=read.csv("C:\\Users\\souvi\\Desktop\\project.csv",header=T)
```

Correlation matrix and Correlation map:

In the data we have total 8 covariates(variable on which the study variable is dependent) and one variables study variable named as Pregnencies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age and Outcome(study variable). Here we're finding the correlation matrix and correlation map involving all these 9 variables, using the software "R-studio".

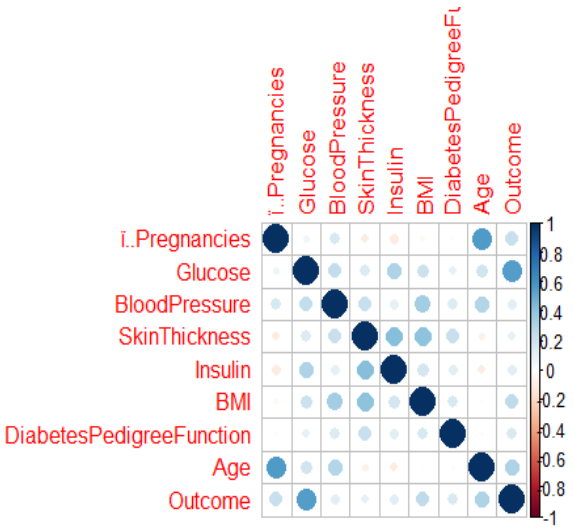
Through the outputs, i.e. correlation matrix and correlation map, we will be able to understand the interdependencies among the 9 variables those are present in our data.

❖ Required R codes and outputs:

```
>install.packeges("corrplot")
>library(corrplot)
>cor(data)
```

output

```
> corrplot(cor(data))
```

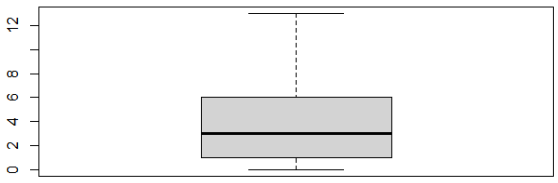


Boxplots:

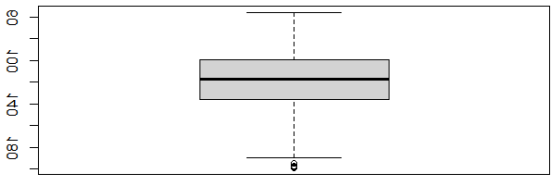
Now we're trying to see that how the 256 observations in each of the 8 covariates are distributed, using a five number summary. So finding the boxplots for each covariate, i.e. boxplots for each of the columns(except the last one) of our data, can be the solution.

❖ Required R codes and outputs:

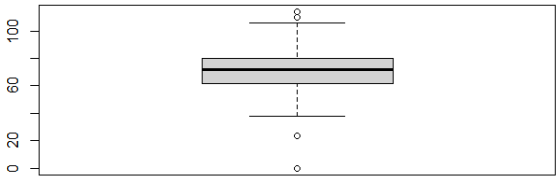
>boxplot(data$ï..Pregnancies)



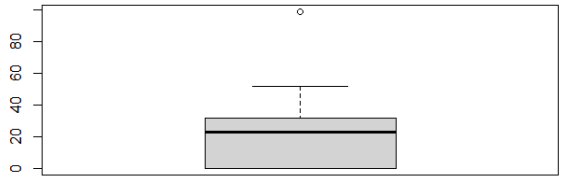
>boxplot(data$Glucose)



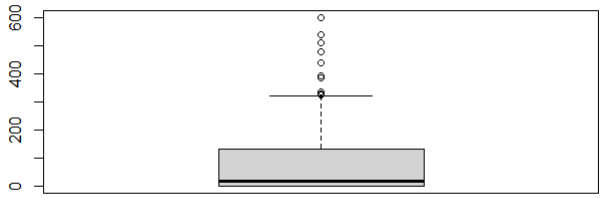
>boxplot(data$BloodPressure)



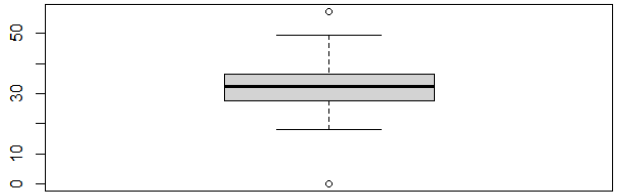
>boxplot(data$SkinThickness)



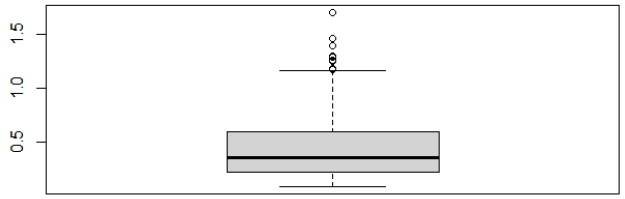
>boxplot(data$Insulin)



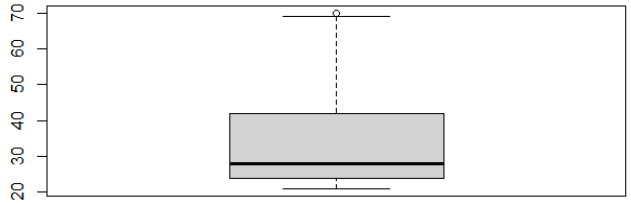
>boxplot(data$BMI)



> boxplot(data$DiabetesPedigreeFunction)



> boxplot(data$Age)



Finding Mean, Median, Mode of the data(required R codes and output):

```
>summary(data)
 Pregnancies       Glucose
Min.   : 0.000   Min.   : 56.0
1st Qu.: 1.000   1st Qu.: 99.0
Median : 3.000   Median :117.0
Mean   : 3.887   Mean   :120.8
3rd Qu.: 6.000   3rd Qu.:136.0
Max.   :13.000   Max.   :199.0
BloodPressure    SkinThickness
Min.   :  0.00   Min.   : 0.0
1st Qu.: 62.00   1st Qu.: 0.0
Median : 72.00   Median :23.0
Mean   : 69.65   Mean   :20.3
3rd Qu.: 80.00   3rd Qu.:32.0
Max.   :114.00   Max.   :99.0
  Insulin          BMI
Min.   :  0.00   Min.   : 0.00
1st Qu.:  0.00   1st Qu.:27.57
Median : 17.00   Median :32.40
Mean   : 79.23   Mean   :32.04
3rd Qu.:130.00   3rd Qu.:36.40
Max.   :600.00   Max.   :57.30
DiabetesPedigreeFunction
Min.   :0.0850
1st Qu.:0.2188
Median :0.3490
Mean   :0.4450
3rd Qu.:0.5958
Max.   :1.6990
    Age          Outcome
Min.   :21.00   Min.   :0.0000
1st Qu.:24.00   1st Qu.:0.0000
Median :28.00   Median :0.0000
Mean   :33.45   Mean   :0.3242
3rd Qu.:42.00   3rd Qu.:1.0000
Max.   :70.00   Max.   :1.0000
   group
Min.   :0.0000
1st Qu.:1.0000
Median :1.0000
Mean   :0.8789
3rd Qu.:1.0000
Max.   :1.0000
```

Binary Logistic Regression(with required R codes and ouput):

Now we're doing the binary logistic regression on our data, where the response variable is outcome.

####Generate groups based on pregnancies(group1=0,group2=1-5,group3=>5)

```
> data$group[data$Pregnancies==0]=0
> data$group[data$Pregnancies>0]=1
```

####Create factors

```
> data2<-within(data,{group<-factor(group)})
```

####multivariable logistic regression model

```
> logit1<-
glm(group~Age+BMI+Glucose+SkinThickness,data=data2,family="binomial"(link="logit"))
> summary(logit1)

Call:
glm(formula = group ~ Age + BMI + Glucose + SkinThickness, family = binomial(link = "logit"),
    data = data2)

Deviance Residuals:
    Min       1Q   Median
-3.2406   0.2617   0.4226
     3Q      Max
 0.5377   1.2168

Coefficients:
             Estimate Std. Error
(Intercept)    2.138147   1.246460
Age            0.074254   0.024655
BMI           -0.076532   0.031680
Glucose       -0.003628   0.006913
SkinThickness  0.027648   0.013867
             z value Pr(>|z|)
(Intercept)     1.715   0.0863 .
Age             3.012   0.0026 **
BMI            -2.416   0.0157 *
Glucose        -0.525   0.5997
SkinThickness   1.994   0.0462 *
---
Signif. codes:
```

0 '***' 0.001 '**' 0.01 '*'
0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 188.98 on 255 degrees of freedom
Residual deviance: 171.30 on 251 degrees of freedom
AIC: 181.3

Number of Fisher Scoring iterations: 6

- ***Analysis:***
  - **Mean, Median, Mode of the variables:**

measurements	pregnencies	glucose	bloodpressure	skinthickness	insulin	bmi	dpf	age
mean	3.887	120.8	69.65	20.3	79.23	32.04	0.4450	33.45
median	3	117.0	72	23	17	32.40	0.3490	28
mode	13	199	114	99	600	57.30	1.6990	70

  - **From the boxplots we can see that-**
    A. There's no outlier in the observations of the variable Pregnencies.
    B. There are outliers in the observations of other variables.

- ✓ ***####HL GOF test:***

>library("performance")

```
>performance_hosmer(logit1,n_bins=10)

Hosmer-Lameshow Goodness-of-fit Test

Chi-squared:  20.848

df: 8

p-value: 0.008
```

- ***Conclusion:***

  We can see that the there are outliers in the observations, except tha variable pregnencies. Because of that mean can't be taken as a measure of central tendency for those variables(all covariates except pregnencies). For those we can take median as a measure of central tendency.

  Here, using Hosmer-Lameshow Goodness of fit test, we can see that the the p-value for the adjusted model is (0.008)>0.001. So we can say that the model fits the data well.

ACKNOWLEDGEMENT:

I am indebted to number of person for helping me in the preparation of this project.
Firstly, Dr. Apurba Roy , Vice- Principal ,Asutosh College, university of Calcutta. Without whose help I couldn't have been a part of this prestigious college.
I owe a deep debt of gratitude to my supervisor Dr. Dhiman Dutta for necessary guidance,
for this presentation of this dissertation ,valuable comments and suggestions. I am extremely grateful to him for me the necessary stimulus, support and valuable time.
I am greatly indebted to Dr. Parthasarathi Bera , Dr. Shirsendu Mukhopadhay, Dr. Sankha Bhattacharya and Ondrila Bose (Faculty members) often took pains and stood by me in adverse circumstances. Without her encouragement and inspiration it was not possible for me to complete this project.
Finally my earnest thanks go to my friends who were always beside me when I needed them without any excuses and made these three years worthwhile.
This project is not only a mere project. It is the memories spend with the whole department which has created a mutual understanding among us. There are many emotions related to this piece of work, especially respect and duty towards teachers and vice versa; educational attachment with my friends; social attachment with my college.

…….
Souvik Roy
Student, Department of Statistics

DECLARATION

I Souvik Roy , a student of B.Sc sem -6, Statistics Honors, of university of Kolkata, Registration no. 012-1111-1821-19 and Roll no. 193012-21-0464 hereby declare that I have done this piece of project work entitled as " Analysis of diabetes in Indian women" under the supervision of Dr. Dhiman Dutta(Assistant professorDepartment and HOD of Statistics , Asutosh college) as a part of B.Sc. Sem-6 examination according to the syllabus paper DSE T5.

I further declare that the piece of project work has not been published elsewhere for any degree or diploma or taken from any published project.

……

Signature