Analysis of Patient Data from a Diabetes Screening Program. (Report of Prevalence, BMI, Age, and association with smoking.

Introduction:

For this project, we aim to investigate Diabetes prevalence and risk factors from those who attended a local screening programme. Our dataset contains information collected during clinic visits of 253680 participants over a period of one month. The data collected included the Patient's Body Mass Index (BMI), Age, Smoking status, Diabetes diagnosis, Cholesterol check, Cholesterol level(High or Low), If the patient had ever had stroke, Sex of the patient, Number of days in the past 30 days the patient says they have suffered with Mental health which includes stress depression and other factors.

Objective:

This project is aimed to summarize data collected, provide insights into diabetes prevalence, compare our findings with previous screening results from another region, to explore associations between some key risk factors such as smoking and diabetes, and to interpret the findings and provide possible explanations for observed trends.

Tabular summary of patients' population data:

Table 1.0 (Summary of Continuous (Numerical) Data);

Variable	Mean	SD (Standard, deviation)	Minimum	Maximum	Median
Age	63.19	17.22	18.00	93.00	66.39
BMI	28.80	5.93	12.00	54.44	27.00

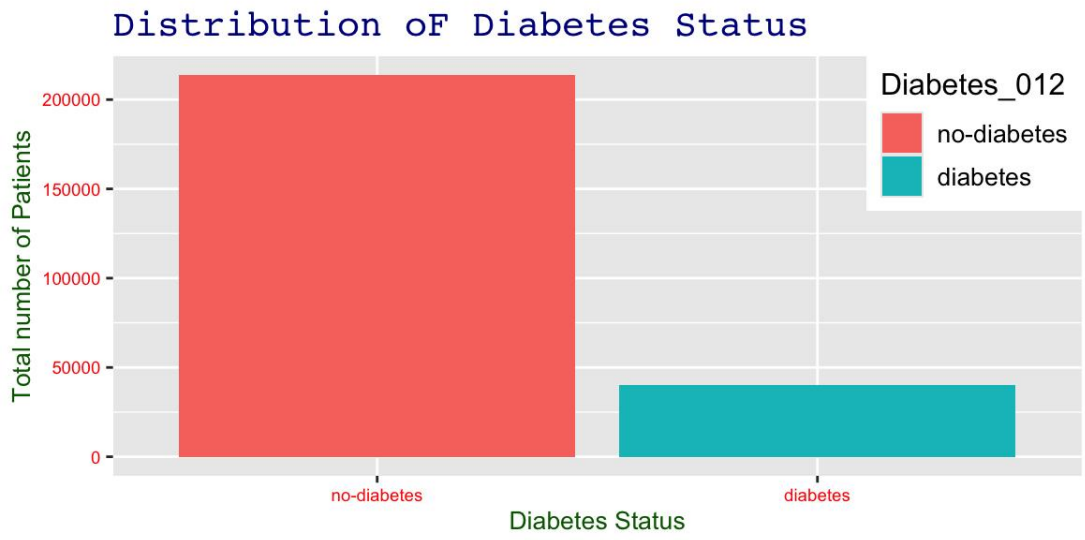
Table 1.1 (Frequency and Percentage Distribution of Categorical Table);

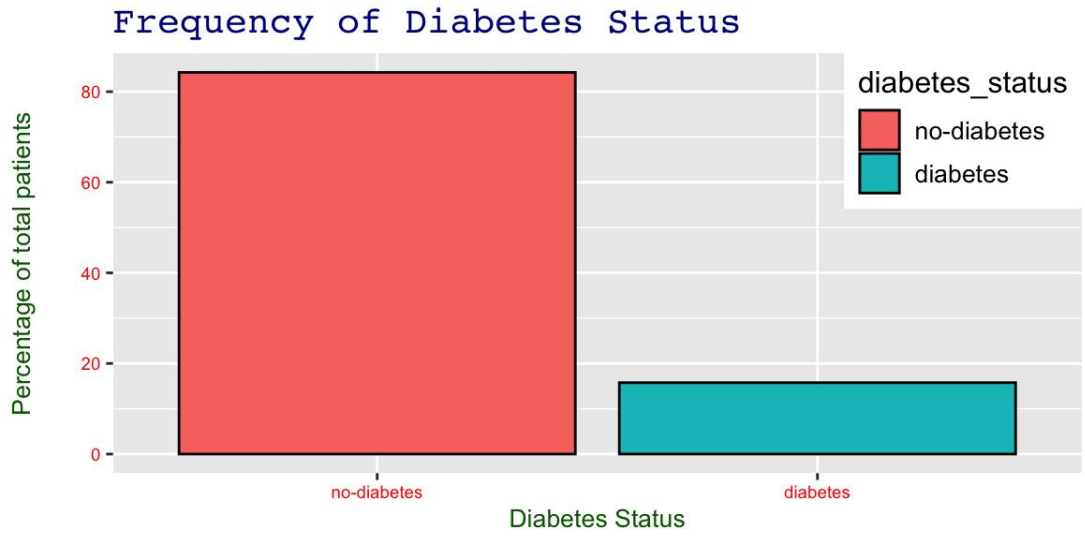
Variable	Category	Count	Percentage (%)
Diabetes	Diabetes	39977	15.76
Diabetes	No-Diabetes	213703	84.24
Smoking	Smoker	112423	44.32
Smoking	Non-Smoker	141257	55.68
HighCol	highcol	107591	42.41
HighCol	No-highcol	146089	57.59
CholCheck	checked	9470	3.73
CholCheck	Not-checked	244210	96.27
Stroke	yes	10292	4.06
Stroke	No-stroke	243388	95.94
Sex	male	111706	44.03
Sex	female	141974	55.97

Observation: From Tables 1.1 and 1.2, the tabular summary of each data set is being discussed leaving out the Mental Health which is not part of this study objective or have a direct impact to these analysis. From the tables 1.1 and 1.2, we can see the distributions of each likely factors of interest across all participants and the data summary which includes the statistical and the proportional observations are also well tabled out.

Graphical Summary of Data:
Extracting the two main categorical Data that affects the main aim of this report, which are the “Diabetes Status” and the “Smoking Status”, below are the graphical insights as derived from the data set, and the graph is presented using both Absolute counts (Total number of patients) and Relative frequency (Percentage).

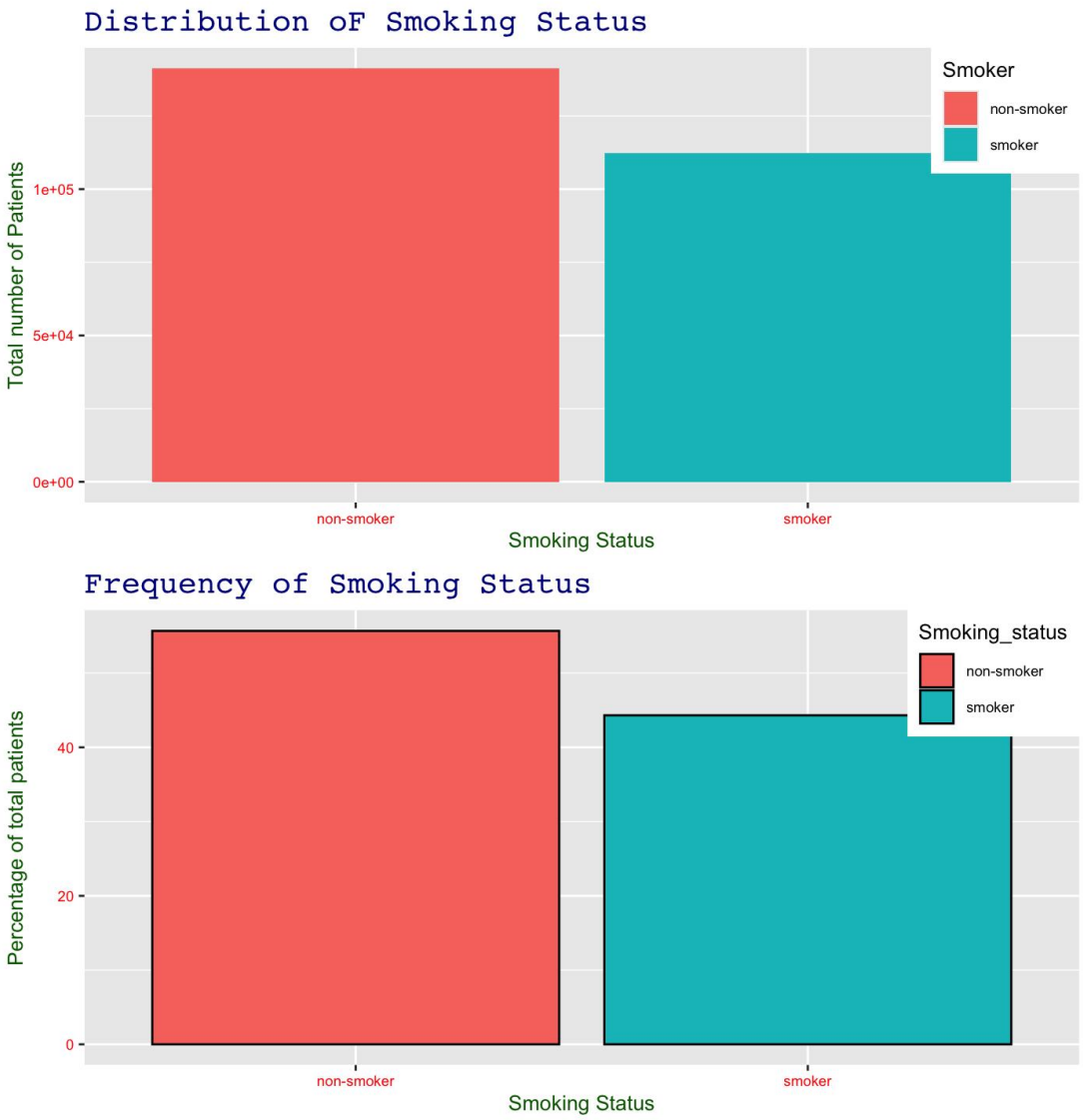
1. Absolute Counts and Proportional Frequency of Diabetes Status.





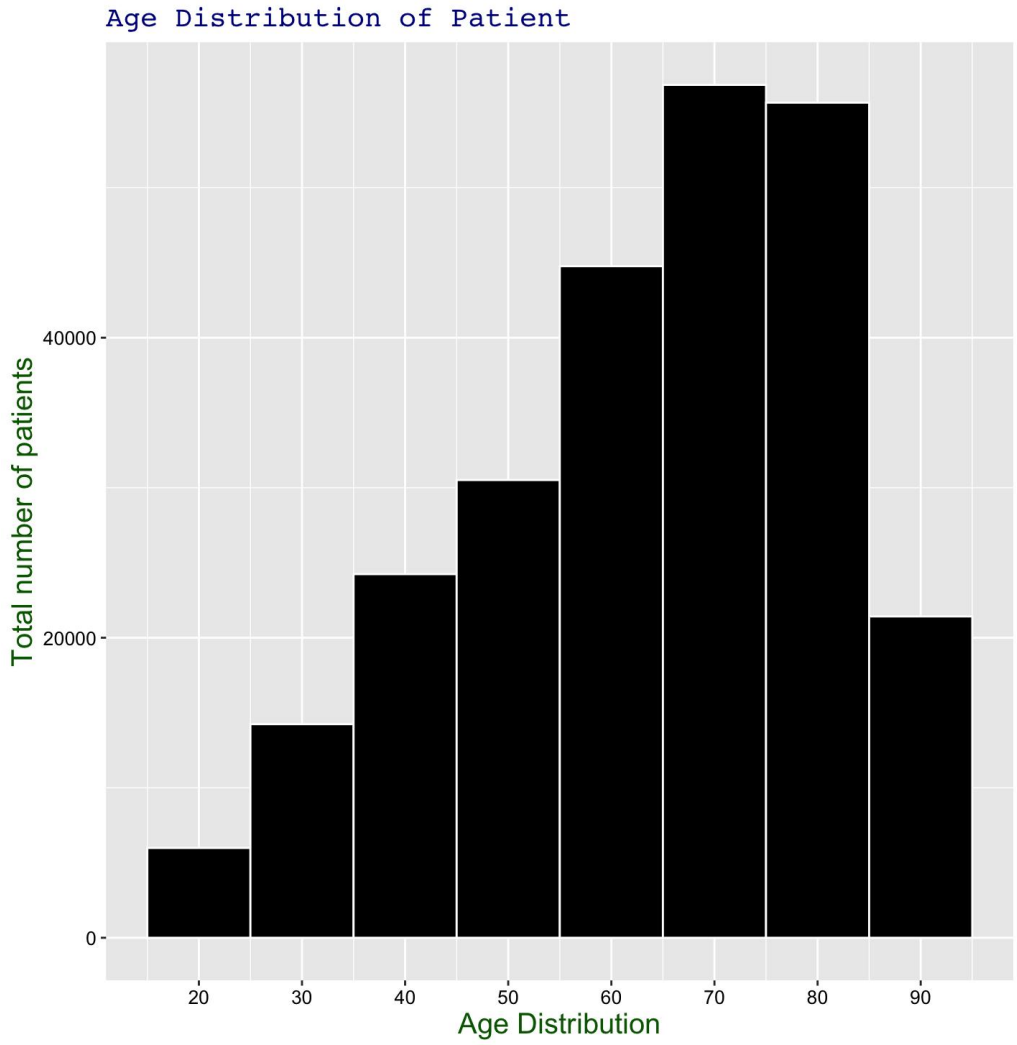
Observation:
From both the Absolute Counts and the frequency graph, it is evident that the plot confirms the severe imbalance as first suggested from the table 1.1, where we see the vast difference between the Diabetes and no-Diabetes category when proportioned in to percentage and we got (15.76% and 84.24% respectively).
Based on this insight we can say the larger part of our population data are not Diabetic as the data is highly skewed.

2. Absolute Counts and Percentage distribution of Smoking Status.



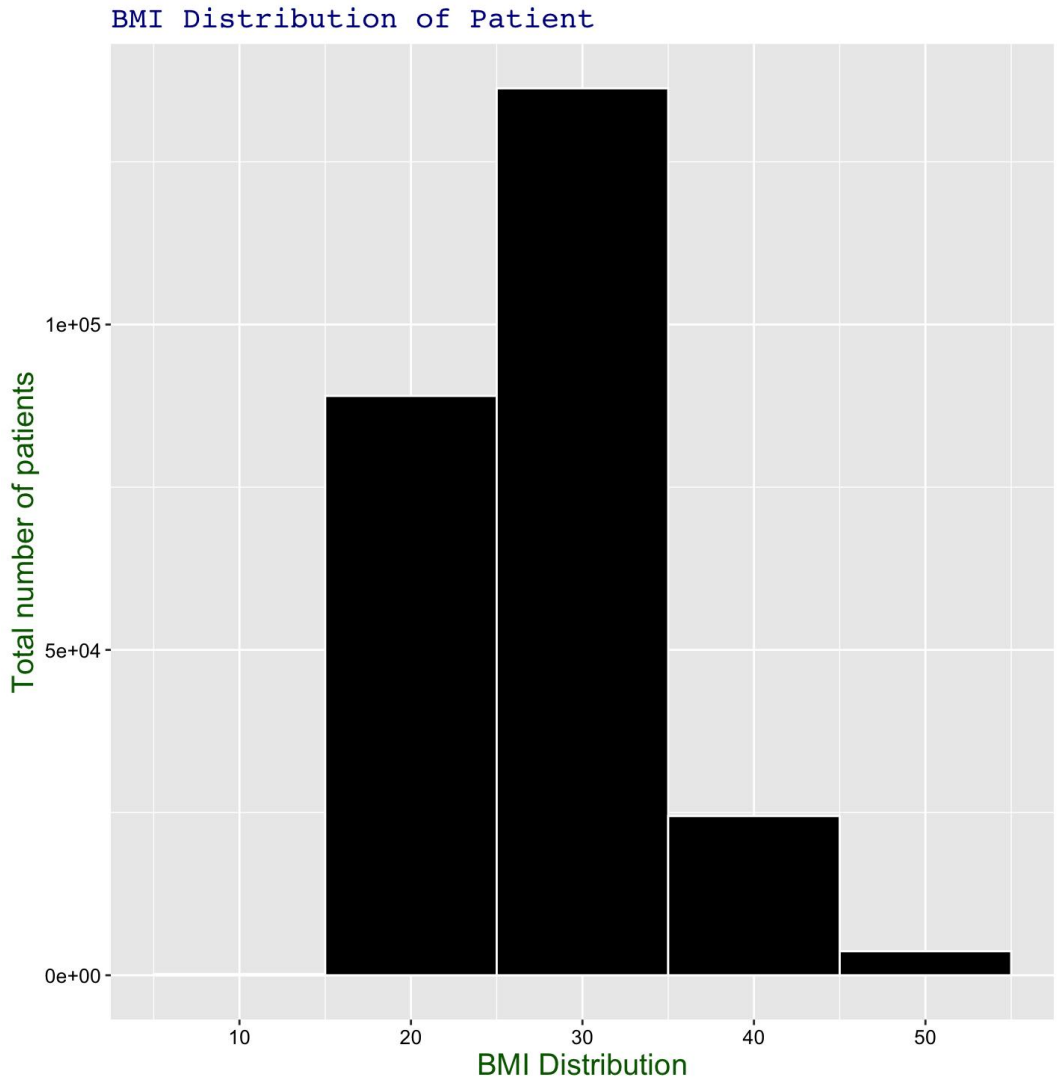
Observation:
Both the raw count and the percentage graph shows that the number of non-smokers are greater than the number and proportion of smokers.
The data is skewed towards the non-smoker category which is a strong indication that the patient population is predominantly consisted of individuals who are non-smokers. This further more support our tabular summary in table 1.1, which shows that the non-smoker to smoker percentage is 55.68% to 44.32%.

3. Age Distribution of patient.
This graph is from a continuous data distribution of the Patients' Age, hence the histogram below best gives us the summary of the Age distribution data as seen in tabel1.0.



Observation:
From the Age distribution graph, it can be observed that the majority of patients fall within the middle age range. The distribution shows a gradual increase in the number of patients from younger ages up to around 50 to 60 years after which there is a slight declination, which indicates that most patients in the population are adults, with fewer patients in the younger group (below 30 years) and older group above (60 years).

4. Body Mass Index (BMI) distribution.
This graph is from the patients' BMI data from the data set, and the histogram below describes the summary of the patients' BMI across the data set population.



Observation:

The BMI Distribution graph shows that the majority of patients have BMI values concentrated in the lower to mid-range categories. A large proportion of the patients fall between a BMI of 20 and 30 which indicates that most individual falls within the normal to overweight range. Also few patients has BMI values above 30 which indicates that obesity is less common among the patient population data. So although a few patients are underweight or obese, the overall population's BMI is moderately distributed around the healthy levels.

Comparison of data with previous study.

1. Summary: Comparison of Diabetes Prevalence.
This section compares the prevalence of diabetes in the currently observed data set with the 14% rate report from the previous year screening program.

This analysis is aimed to see if there is a significant contrast between this data set and the previous year finding using a one-sample proportion test.

From the data-set, a total of 39977 (15.76%) were identified as having diabetes out of 253680 participants. To compare with with the 14% last year reported diabetes rate, a one-sample proportion test was carried out.

Result:
Out of a total of 253,600 individuals, 39,977 were identified as having diabetes, giving an observed proportion ($\hat{p}$) of 0.1576 or (15.76%).

The null hypothesis stated that the true population proportion of diabetes (p) is equal to 0.14 (14%), while the alternative hypothesis suggest that it differs from 0.14, which makes the test a two-tailed test., The test produced a Chi-squared (χ^2) value of 655.13, with 1 degree of freedom, and the p-value less than $2.2 * 10^{-16}$. Since the P-value is far less than the generally accepted significance level of 0.05, we reject the null hypothesis. Therefore this indicates that the observed diabetes prevalence of (15.76%) from our current data-set is significantly different from the 14%

reported last year. Therefore this test rejects the hypothesis that the average BMI for our current data set is 14%.

Conclusion:
In summary the analysis demonstrates a notable increase in diabetes prevalence from 14% to 15.76% within the both years of observation, which indicates a possible upward trend that might warrant an advanced public health indication.

2. Summary: Comparison of Mean BMI with 28 kg/m²
This analysis aimed to determine whether the mean Body Mass Index (BMI) of the individuals in the current data set differs from the reference value of 28kg/m² that was reported from the last year screening programme.
A one-sample t-test was conducted to assess if the average BMI in the sample differs statistically from the previously observed mean.

Result:
The result indicated a sample mean of BMI of 28kg/m² with a confidence interval of 28.20 to 28.24. The test produced a t-value of 18.71 with 253,679 degree of freedom and a p-value less than 0.001.
The p-value shows a significant difference from the reference value. Although the difference from observation is small, but the large sample size makes it significant,which suggest that the average BMI in our current data set is slightly higher than 28kg/m². Thus this test rejects the hypothesis that the average BMI of our current dataset is 28kg/m²

Conclusion:
From the test result, there is an indication of a slight upward shift in the average BMI among participants compared to last year's screening population. Although the difference is small, it might however suggest a gradual increase in body weight trends over the observed period of time which might have a public health and life style interventions.

3. **Summary: Comparison of Mean Age with 55 Years**

The aim of this analysis was to determine whether the mean age of individuals in the current dataset is different from the 55 years value which was gotten from last year's screening programme.

A one-sample t-test was carried out to assess if the average age in the present sample shows any significant relationship or deviation from this benchmark.

Result:

The results indicated a mean age of 63.19 years as also supported by table 1, with a 95% confidence interval of 63.12 to 63.26. The test produced a t-value of 239.53, with 253,679 degrees of freedom, and a p-value less than 0.001, indicating a statistically significant difference between the current mean age and the reference value gotten from last year. This finding rejects the Hypothesis that the average age of participants in the current dataset is 55 years, but rather shows that there is a significantly higher than 55 years.

Conclusion

The analysis demonstrates that the mean age of the population under study has increased notably compared to last year's average of 55 years. This suggests that the current screening group consists of an older population, which could influence the observed rates of health conditions such as diabetes and elevated BMI.

4. **Summary: Association Between Smoking and Diabetes**

This test was aimed to investigate if there is a significant association between the smoking status and the prevalence of diabetes in the study population.

A Chi-squared test of independence was performed to assess the relationship between smoking and diabetes.

Result:

The Result indicated there is a statistically significant association between the two variables ($\chi^2 = 999.41$, $df = 1$, $p < 2.2 \times 10^{-16}$),

(Where χ^2 is the Chi-squared statistic, df is the degree of freedom, and the p signifies the p-value). Strongly suggesting that smoking status and diabetes prevalence are not independent.

Conclusion:
This result suggests that smoking status is not independent of diabetes status in the study population. Which rejects the null hypothesis that suggest that they were. In other words, the likelihood of having diabetes differs significantly between smokers and non-smokers in the observed population.

This Indicates that Individuals who smoke have a different probability of having diabetes compared to non-smokers, highlighting the potential role of smoking as a factor associated with diabetes.

General Conclusion :
The analysis reveals a diabetes prevalence 0f 15.76%, significantly higher than the 14% reported in the previous screening. The mean BMI (28.8kg/m^2) and age (63.19 years) were also higher than the reference value of 28kg/m^2 and 55 years, indicating an older and slightly heavier population. Also we found a significant association between Smoking and Diabetes, showing that smokers and non-smokers differ in their likelihood of having diabetes.
Overall, these findings suggests a rising trend in diabetes and related risk factors, emphasizing the need for continued public health awareness and lifestyle interventions.