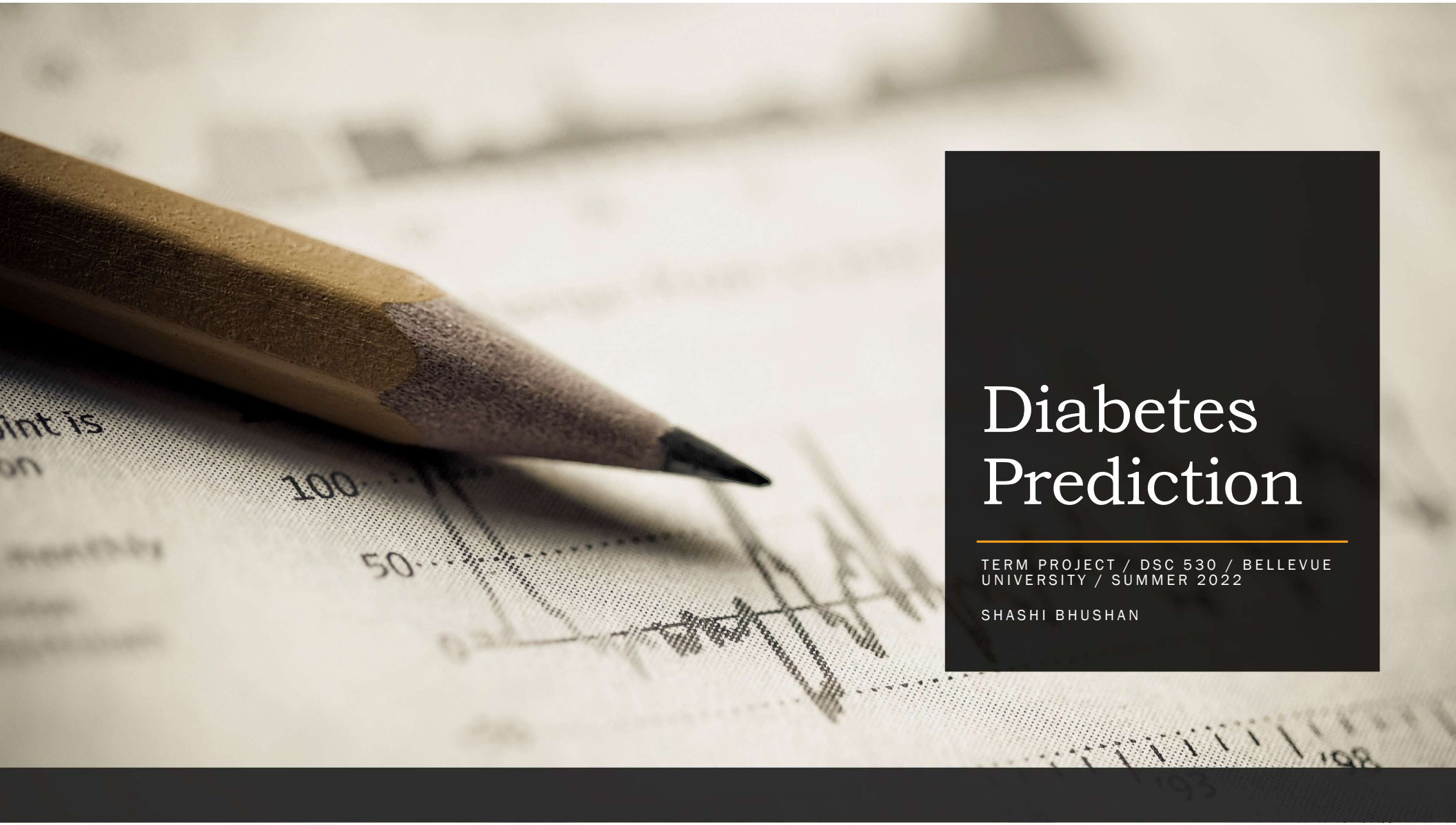


Diabetes Prediction

TERM PROJECT / DSC 530 / BELLEVUE UNIVERSITY / SUMMER 2022

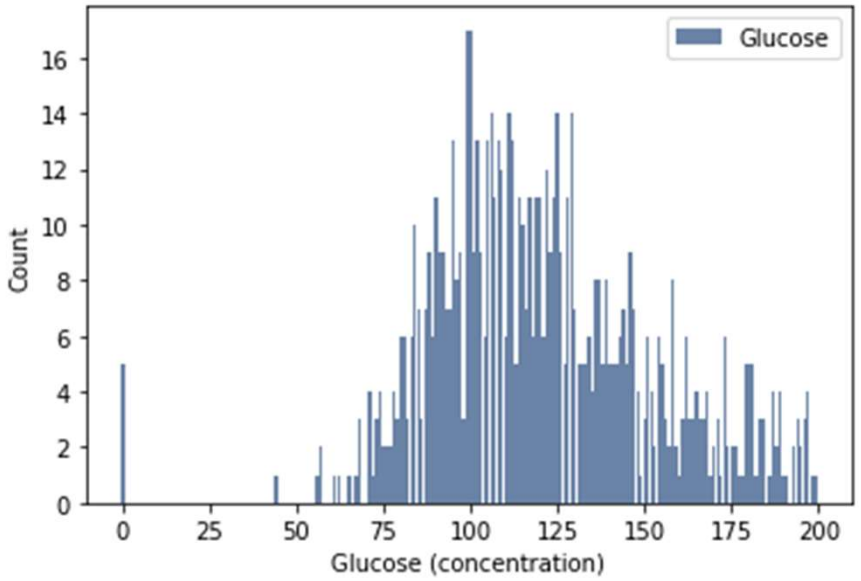
SHASHI BHUSHAN

Introduction

- Aims to identify diagnostic parameters that can be used to predict diabetes.
- Are other variables correlated?
- Utilizes data from the National Institute of Diabetes and Digestive and Kidney Diseases, India (Source: Pima Indians Diabetes Database | Kaggle).
  - Dataset included 8 predictor variables such as number of times pregnant, Glucose levels, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes pedigree function, and Age. Outcome variable was Boolean (0 for diabetes and 1 for non-diabetic).

Selected Variables

- Following variables were selected in the analysis based on generally associated diagnostic parameters with existing diabetic cases:
  - Glucose - Plasma glucose concentration at 2 hours in an oral glucose tolerance test.
  - BloodPressure - Diastolic blood pressure (mm Hg)
  - Insulin - 2-Hour serum insulin (mu U/ml)
  - BMI - Body mass index (weight in kg/(height in m)^2)
  - Age – Age in years




```
df.Glucose.describe()
```

```
count    768.000000
mean     120.894531
std       31.972618
min        0.000000
25%       99.000000
50%      117.000000
75%      140.250000
max      199.000000
Name: Glucose, dtype: float64
```

```
df.Glucose.mode()
```

```
0     99
1    100
Name: Glucose, dtype: int64
```

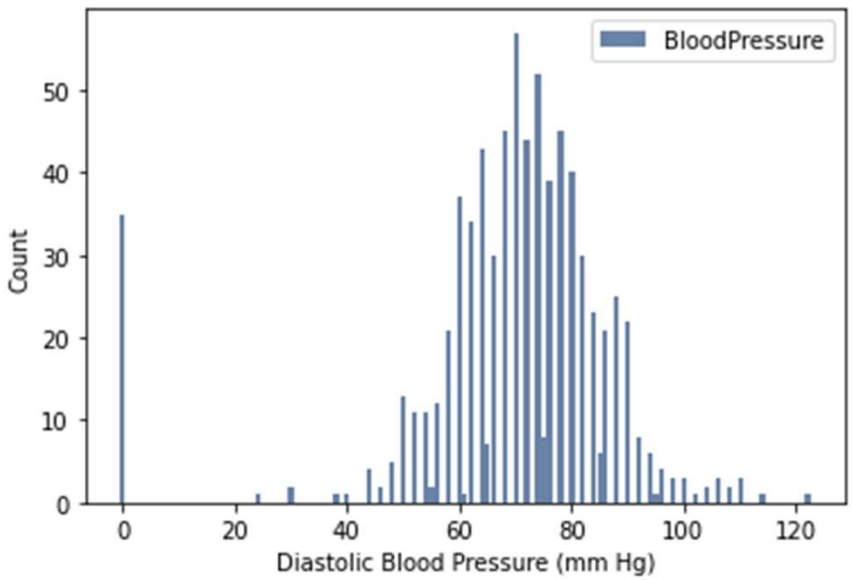
```
df.Glucose.tail()
```

```
763    101
764    122
765    121
766    126
767     93
Name: Glucose, dtype: int64
```

Glucose

Glucose levels below 70 are outliers as glucose levels are typically over 70 for general population.

A check on the frequency of occurrence produced 16 data points out of 768 below 70.




```
df.BloodPressure.describe()
```

```
count    768.000000
mean      69.105469
std       19.355807
min        0.000000
25%       62.000000
50%       72.000000
75%       80.000000
max      122.000000
Name: BloodPressure, dtype: float64
```

```
df.BloodPressure.mode()
```

```
0    70
Name: BloodPressure, dtype: int64
```

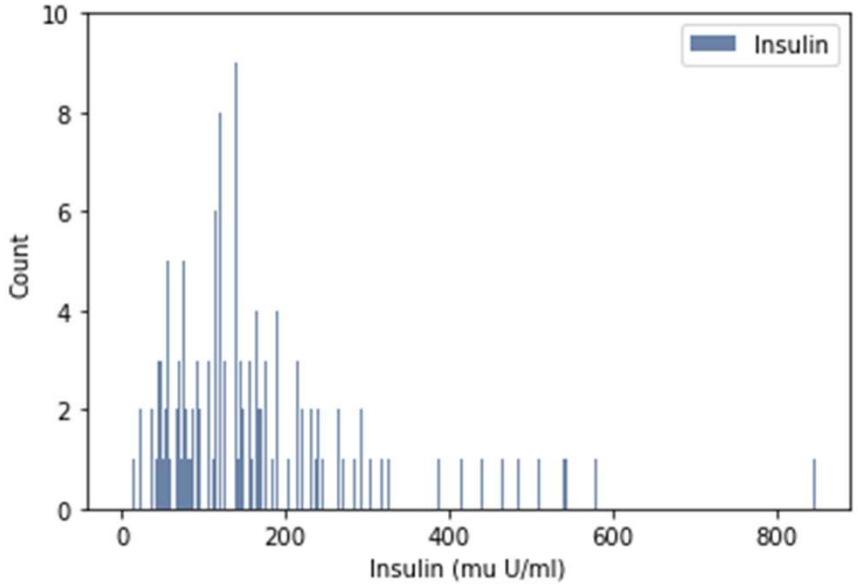
```
df.BloodPressure.tail()
```

```
763    76
764    70
765    72
766    60
767    70
Name: BloodPressure, dtype: int64
```

Blood Pressure

Blood Pressure below 50 and over 100 are outliers as these typically are considered too low or too high for diastolic levels.

Latest and smallest 20 numbers indicated single digit frequencies.




```
df.Insulin.describe()
```

```
count    768.000000
mean      79.799479
std      115.244002
min        0.000000
25%        0.000000
50%       30.500000
75%      127.250000
max      846.000000
Name: Insulin, dtype: float64
```

```
df.Insulin.mode()
```

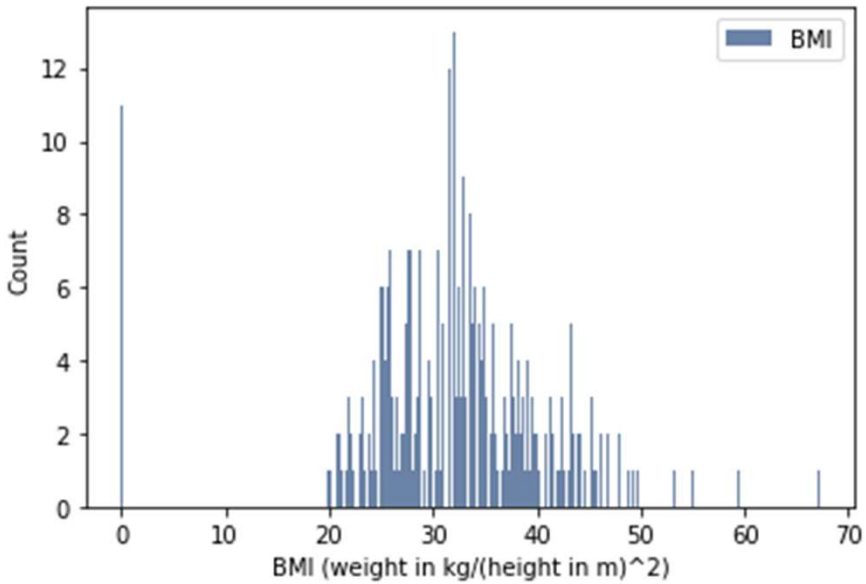
```
0    0
Name: Insulin, dtype: int64
```

```
df.Insulin.tail()
```

```
763    180
764      0
765    112
766      0
767      0
Name: Insulin, dtype: int64
```

Insulin

Insulin test results less than 50 and higher than 600 were considered outliers as these typically fall under too low and too high for physically active individuals.




```
df.BMI.describe()
```

```
count    768.000000
mean      31.992578
std        7.884160
min        0.000000
25%       27.300000
50%       32.000000
75%       36.600000
max       67.100000
Name: BMI, dtype: float64
```

```
df.BMI.mode()
```

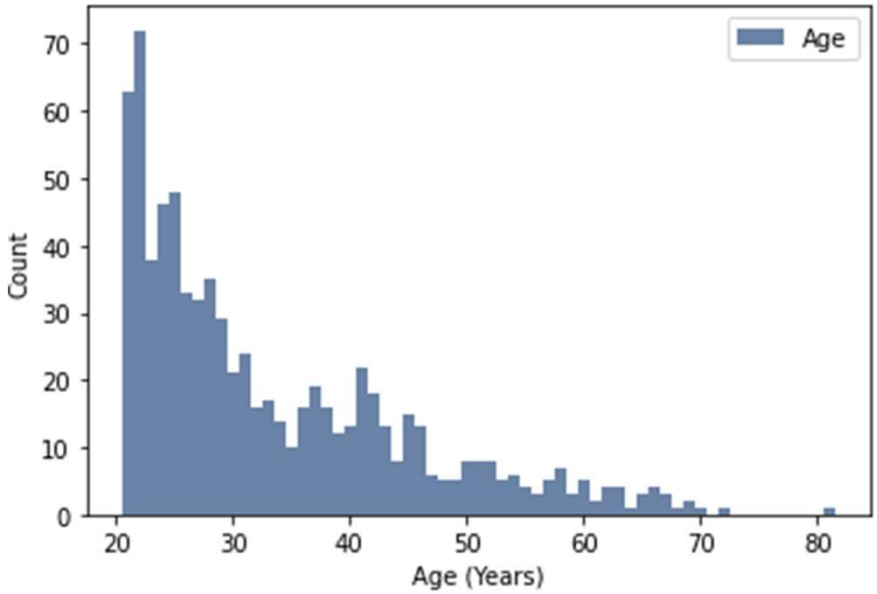
```
0    32.0
Name: BMI, dtype: float64
```

```
df.BMI.tail()
```

```
763    32.9
764    36.8
765    26.2
766    30.1
767    30.4
Name: BMI, dtype: float64
```

Histogram - BMI

BMI less than 18.2 are outliers and only zero showed as the value which is not a correct representation for BMI,




```
df.Age.describe()
```

```
count    768.000000
mean      33.240885
std       11.760232
min       21.000000
25%       24.000000
50%       29.000000
75%       41.000000
max       81.000000
Name: Age, dtype: float64
```

```
df.Age.mode()
```

```
0    22
Name: Age, dtype: int64
```

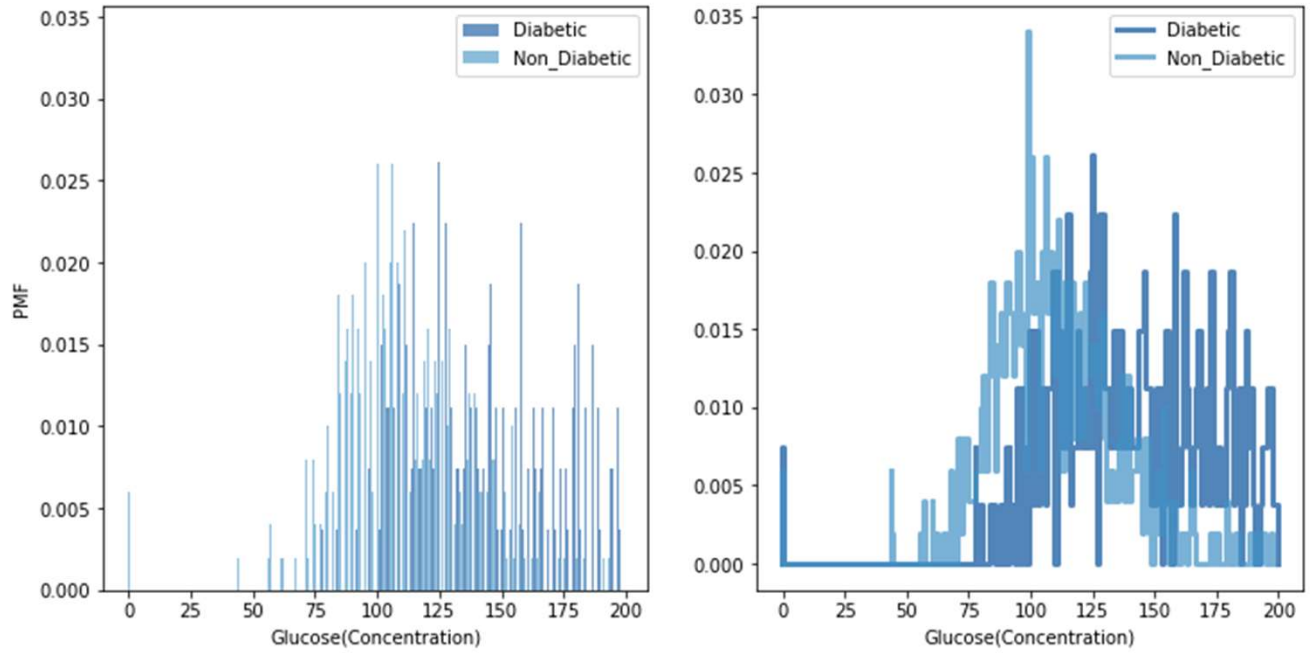
```
df.Age.tail()
```

```
763    63
764    27
765    30
766    47
767    23
Name: Age, dtype: int64
```

Histogram - Age

No outliers were observed as 20 to 80+ years can have diabetic people.

PMF - Glucose

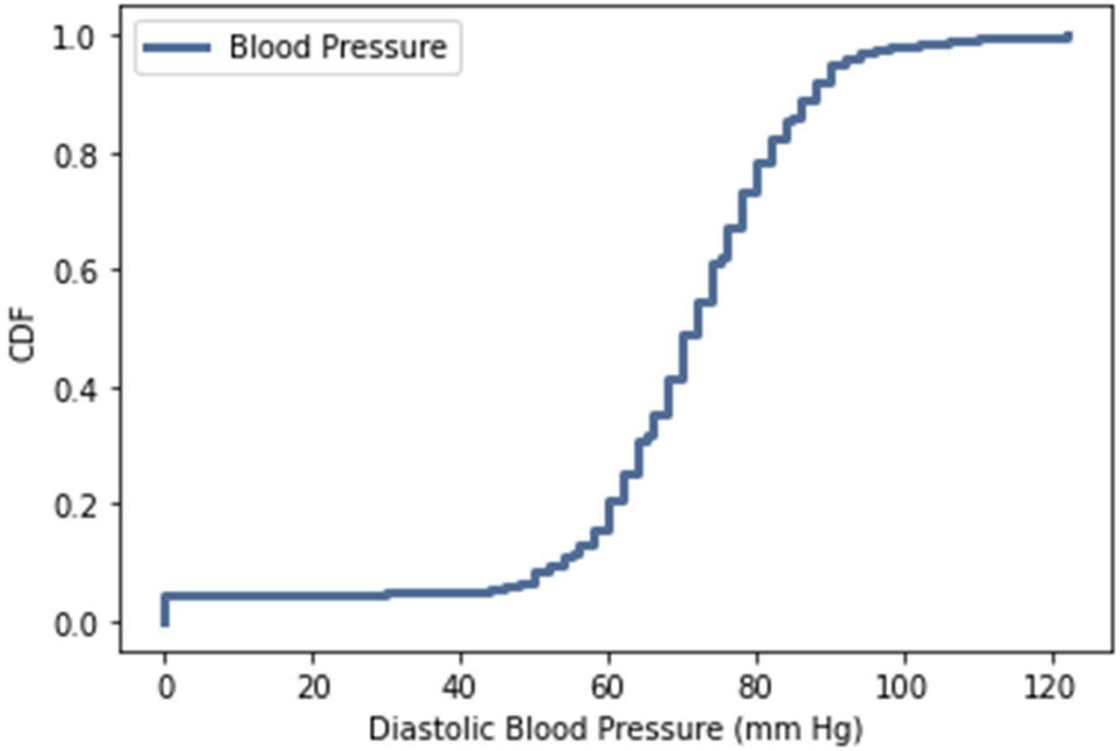

Diabetic
Non_Diabetic
PMF
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
25
50
75
100
125
150
175
200
Glucose(Concentration)
Diabetic
Non_Diabetic
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
25
50
75
100
125
150
175
200
Glucose(Concentration)

CDF – Blood Pressure

Very few people have their diastolic Blood Pressure below 50. Therefore, it is flat.

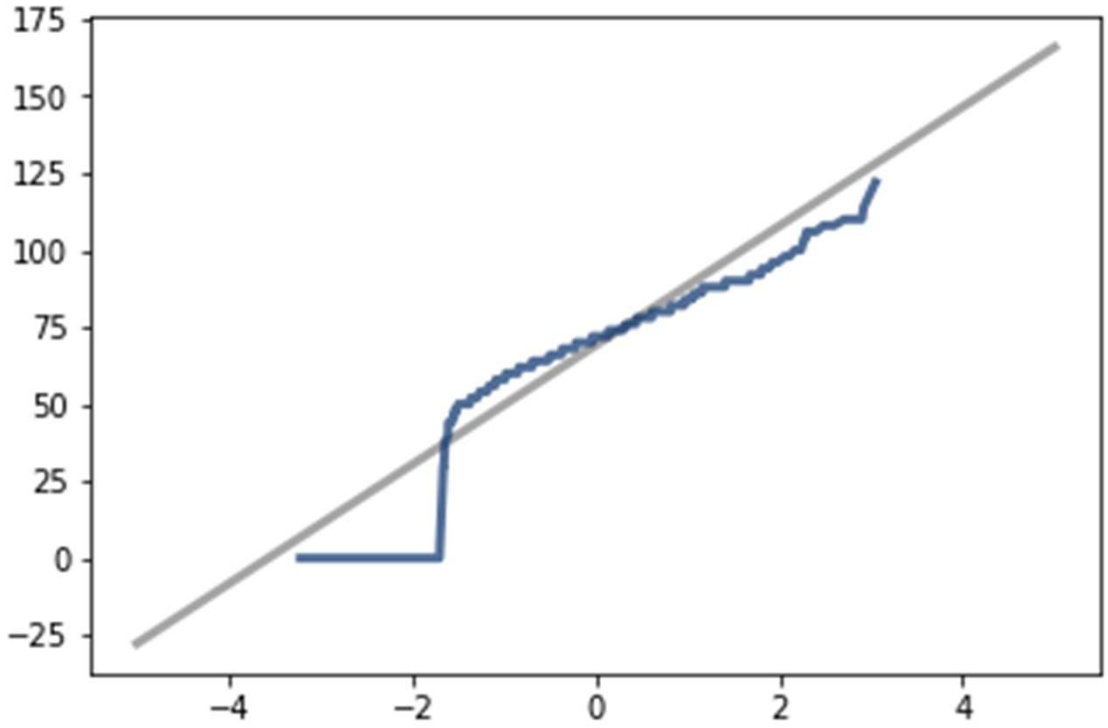
About 80% of the people have their diastolic Blood Pressure in the normal range, i.e., 80 mm Hg per recommended limit.

Therefore about 20% of the people are expected to have diabetes.



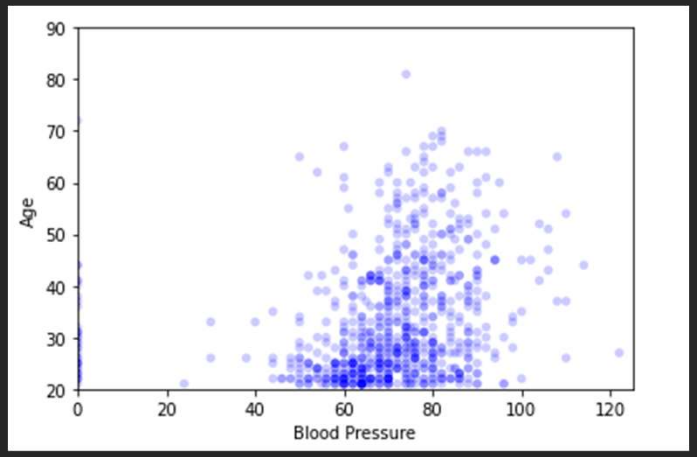
Normal Distribution – Blood Pressure

The normal probability plot shows data is almost normal as it matches the model seen as the gray line



Scatterplot – Blood Pressure / Age

Blood Pressure and Age are +vely correlated but strength of relationship is weak. It can be non-linearly correlated.




```
Cov(df.BloodPressure, df.Age)
```

```
54.45245869954427
```

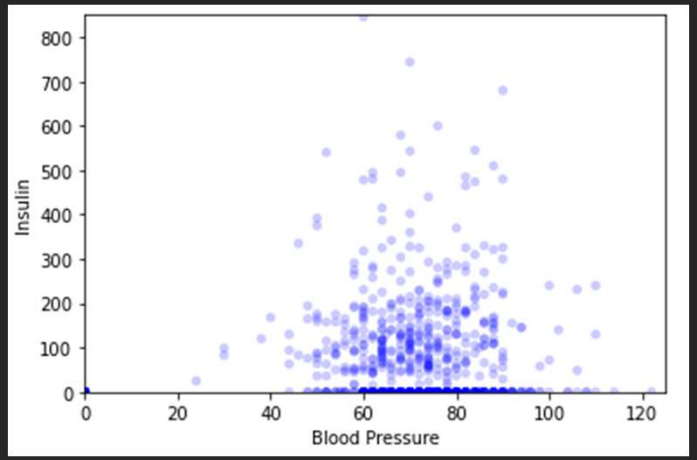
```
Corr(df.BloodPressure, df.Age)
```

```
0.23952794642136355
```

Scatterplot – Insulin / Blood Pressure

Blood Pressure and Insulin levels seem to be +vely correlated but strength of relationship is very less.

It can be non-linearly correlated.




```
Cov(df.BloodPressure, df.Insulin)
```

```
198.12010701497397
```

```
Corr(df.BloodPressure, df.Insulin)
```

```
0.08893337837319301
```

Hypothesis Test

- Difference in means for BMI in diabetic and non-diabetic people was tested.
- pvalue indicated that we expect to see a difference as big as the observed effect about 0% of the time.
- Result is statistically significant.

Regression Analysis

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-8.0932	0.692	-11.698	0.000	-9.449	-6.737
df.Glucose	0.0344	0.004	9.621	0.000	0.027	0.041
df.BloodPressure	-0.0122	0.005	-2.388	0.017	-0.022	-0.002
df.Insulin	-0.0010	0.001	-1.196	0.232	-0.003	0.001
df.BMI	0.0908	0.014	6.404	0.000	0.063	0.119
df.Age	0.0331	0.008	4.133	0.000	0.017	0.049

Logistical Regression was performed for predicting diabetes with glucose, blood pressure, insulin, BMI, and Age as predictor variables.

As evident from the pvalues Glucose, BloodPressure, BMI, and Age are statistically significant in predicting diabetes if we consider alpha to be 5%.

The End