Diabetes

James Patalan

DSC530 – Data Exploration and Analysis

08/12/23

Data Overview

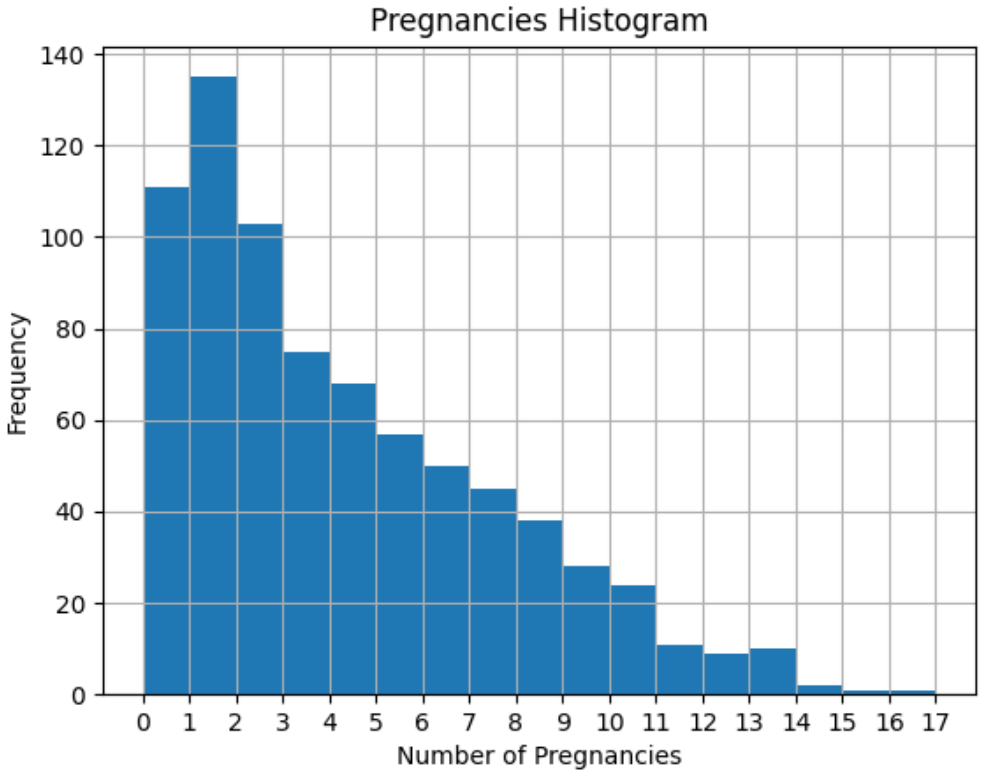
This dataset contains medical information related to Diabetes, from a group of 768 Pima Native American women.

My goal of the dataset is to build a predictive model that could determine whether a person has diabetes based on their medical history.

My hypothesis is that plasma glucose concentration will have the greatest effect on prediction.

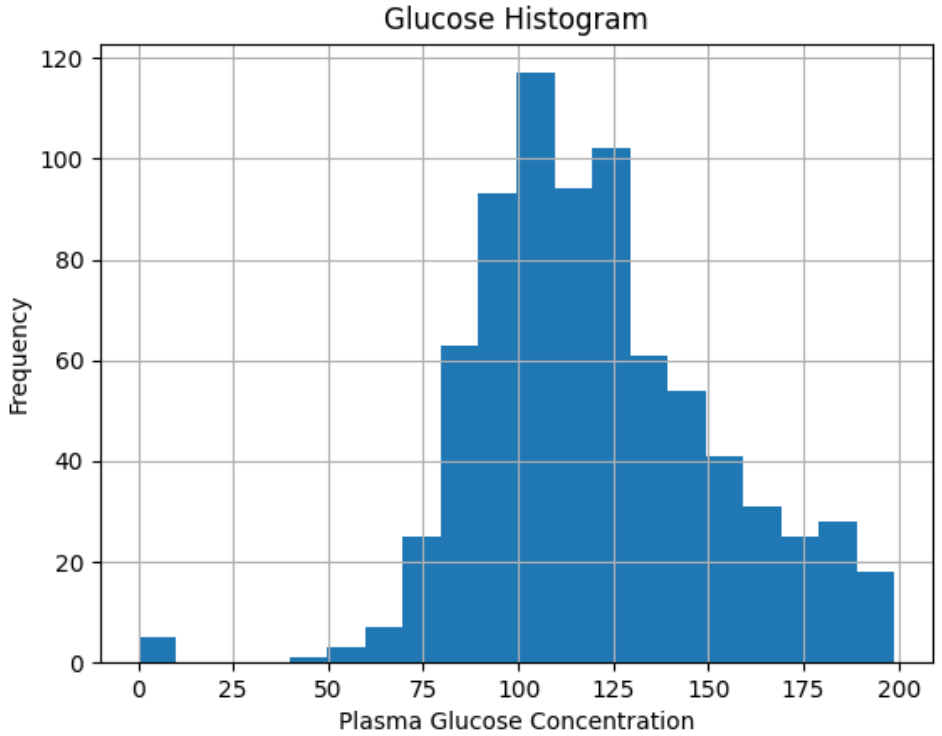
Variable 1: Pregnancies

- The first variable I chose is the number of times the woman has been pregnant
- I chose this variable because pregnancy greatly affects a woman's body and changes it
- The average number of pregnancies is 3.85, and the mode is 1 pregnancy, however there are some extreme outliers, the greatest being 17 pregnancies, that’s a lot of babies



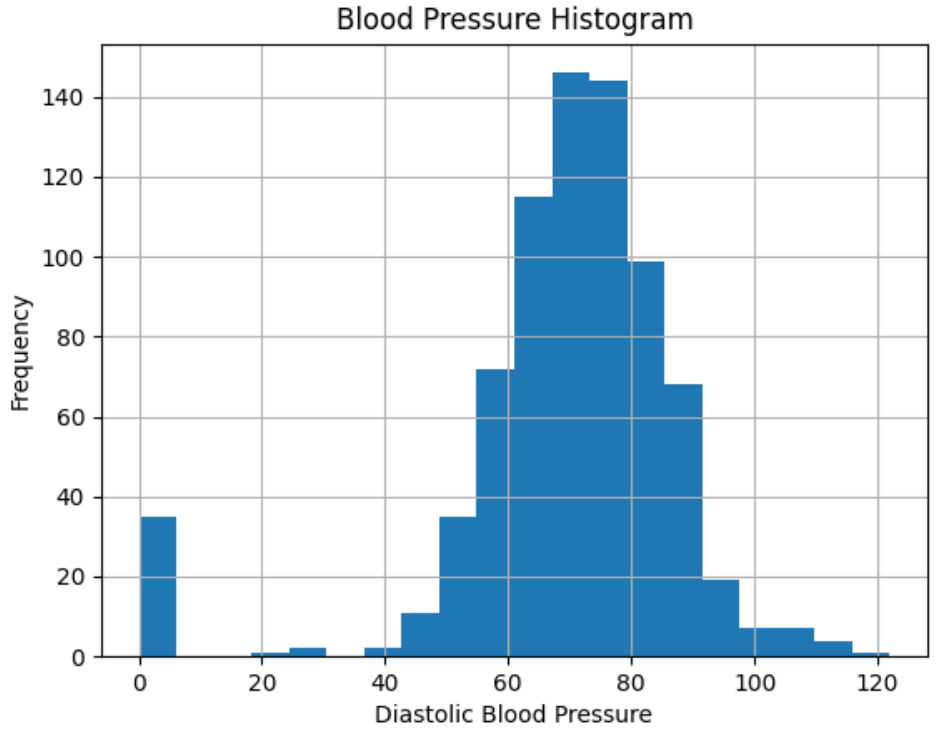
Variable 2: Glucose

- This variable is the woman's level of plasma glucose concentration (aka blood sugar)
- The reason I chose this variable is because diabetes is a condition that effects a person's blood sugar levels
- The average glucose level is 120, and the two modes are 99 and 100
- Normal glucose levels are expected between 70-100mg/dL
- Higher blood sugar levels as seen in the histogram bring the average up



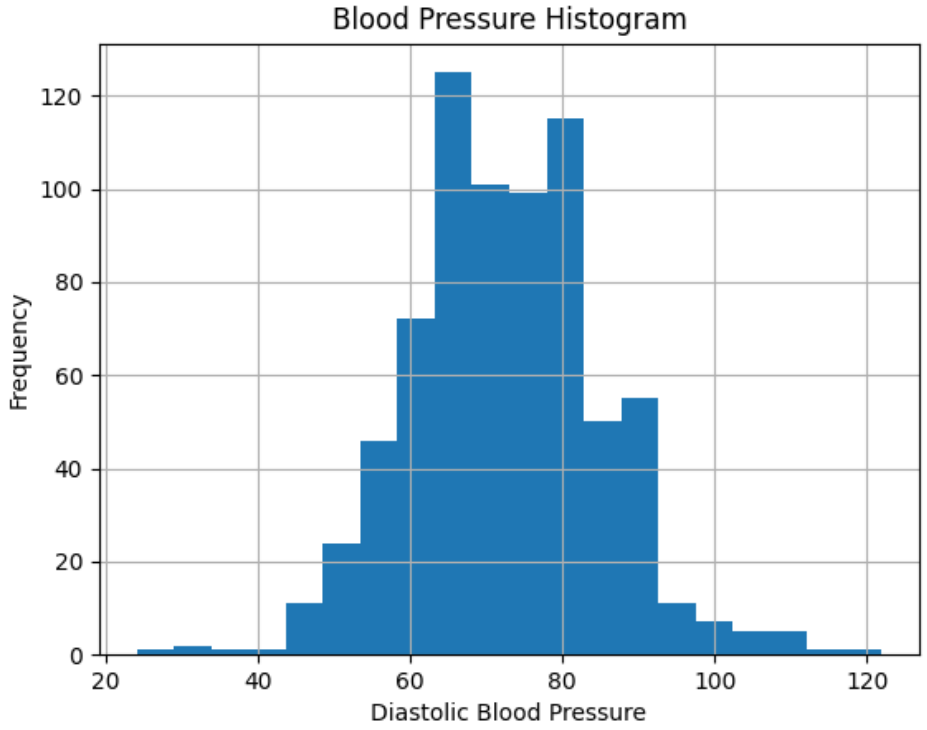
Variable 3: Blood Pressure

- The next chosen variable is Diastolic Blood Pressure reading
- This variable was chosen because high blood pressure can lead to many health-related complications
- Normal blood pressure for women ranges from 110-139 mm Hg depending on age
- In this dataset, there are several strange instances where a blood pressure of 0 is given, this greatly brings down the average resulting in a mean of 69.12



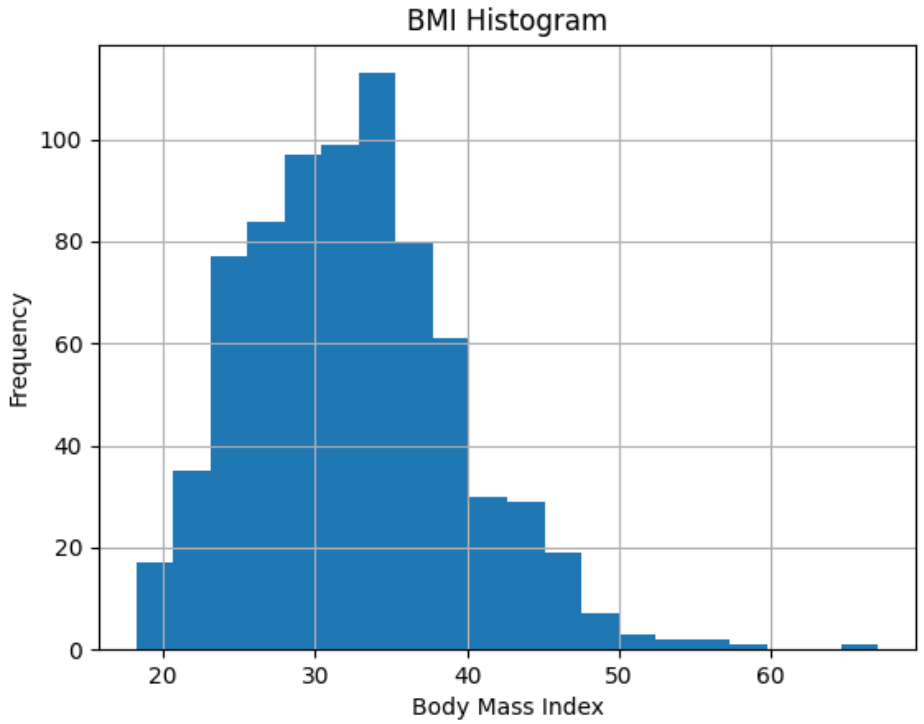
Variable 3: Blood Pressure Cont.

- It is my educated guess that the zeros that are appearing here are meant to be NULL values
- I have written new code that will change zeros to NULL values in all columns where a value of 0 would not make sense, Pregnancies will not be altered as never being pregnant is possible, unlike having no blood pressure
- New blood pressure mean: 72.41
- New glucose mean: 121.69



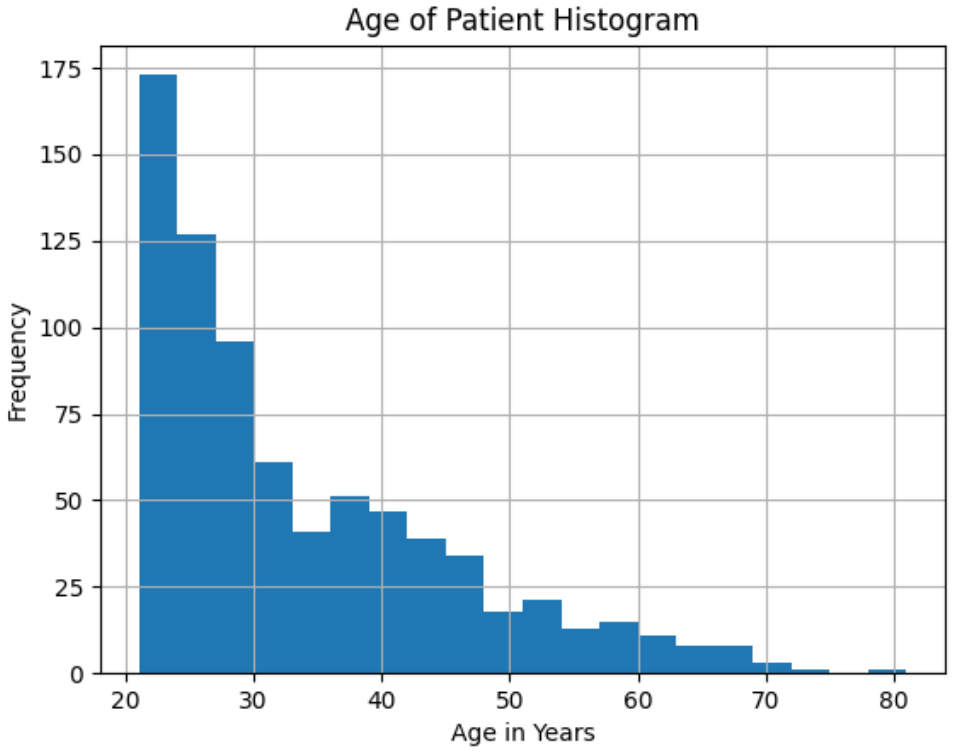
Variable 4: BMI

- The fourth chosen variable is Body Mass Index. BMI is a person's weight, divided by the square of their height, and it is used to broadly categorize if someone is Underweight, Normal, Overweight, etc.
- This variable was selected because being overweight increases risk for many health issues, including diabetes
- Both the average and mode for the sample was a BMI of 32
- According to the CDC, a BMI of 30 or higher falls into the obese range



Variable 5: Age

- My final chosen variable is age
- This variable was chosen because older adults are at a higher risk of development of type 2 diabetes due to increasing insulin resistance and pancreatic islet that come with age
- The average age of the group is 33
- Most patients are younger, with a mode of 22, and only a few elderly patient outliers

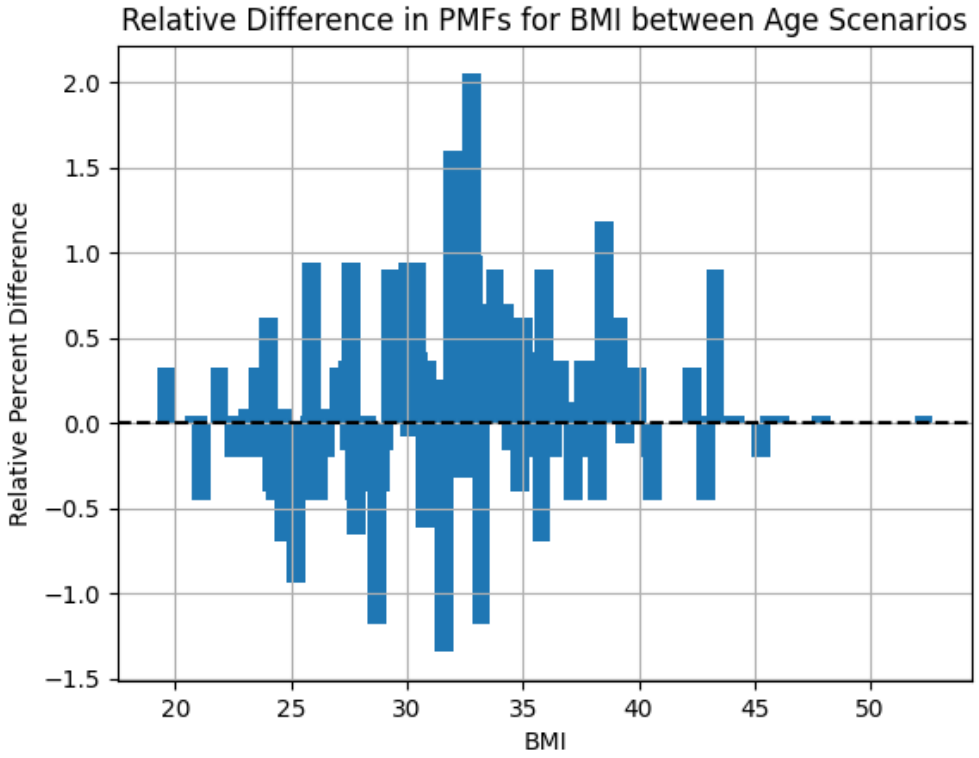


Primary Analysis

PMF, CDF, Analytical Distribution & Scatter Plots

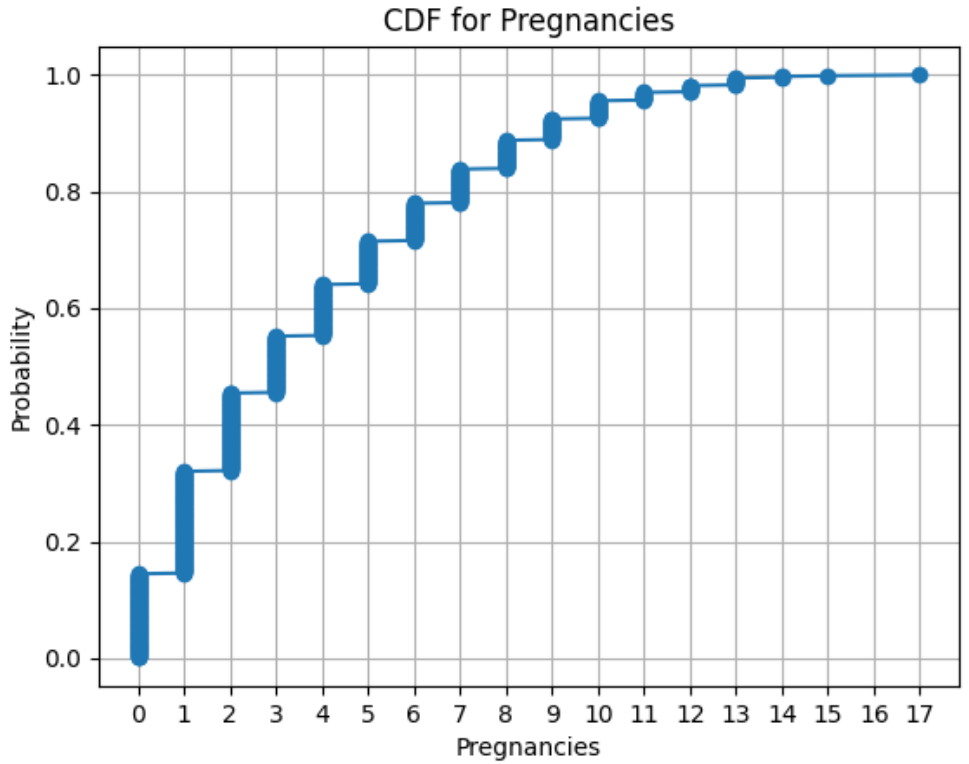
Probability Mass Function

- In examining the variable distribution, it was revealed that most of the patients were younger
- This graph shows the difference in percentage points between BMI for patients older than 31, compared to all patients younger than 30
- My hypothesis was that there would be a clear division between younger and older patients as young people are generally thought to be healthier and more active
- However, the distribution is evenly spread, showing no relationship between age and BMI



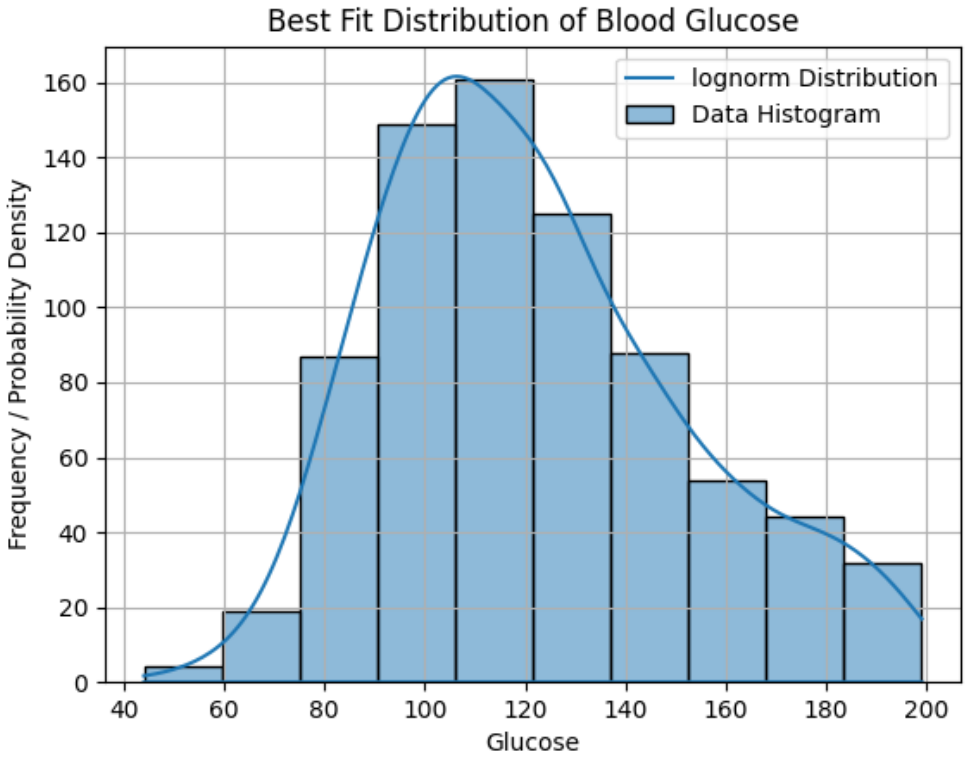
Cumulative Distribution Function

- Looking at the CDF for the number of pregnancies each women has had, one thing stands out, the women in this data set have had a lot of babies
- While nearly all patients have had less than 12 pregnancies, less than 20% have never been pregnant
- This tells me that pregnancies may not be affecting diabetes because more than 80% of patients have been pregnant

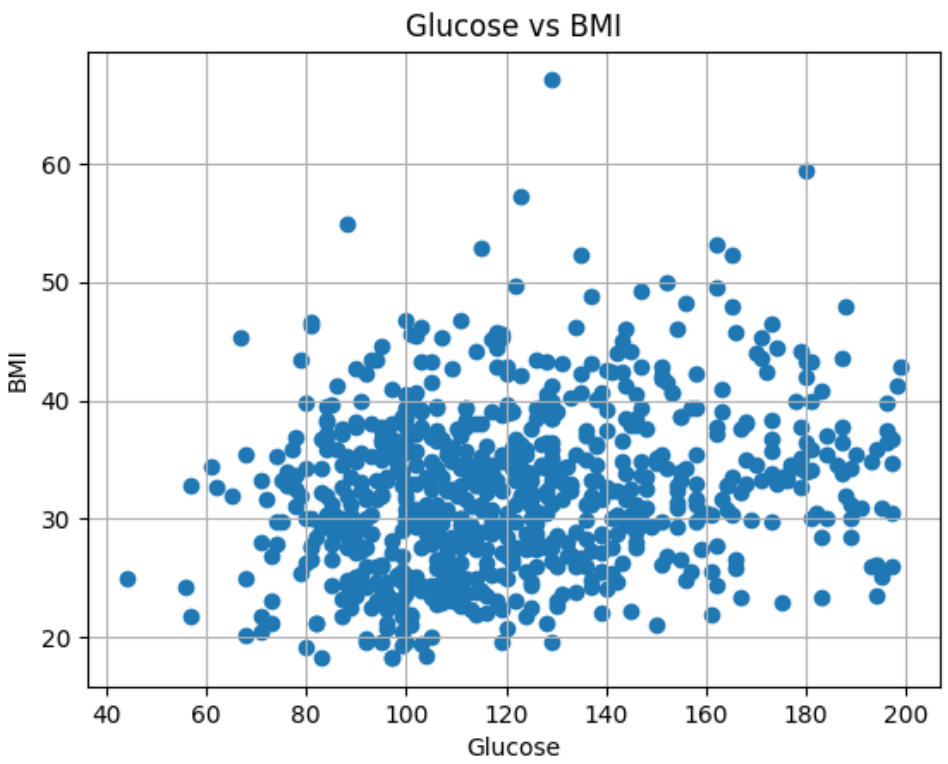


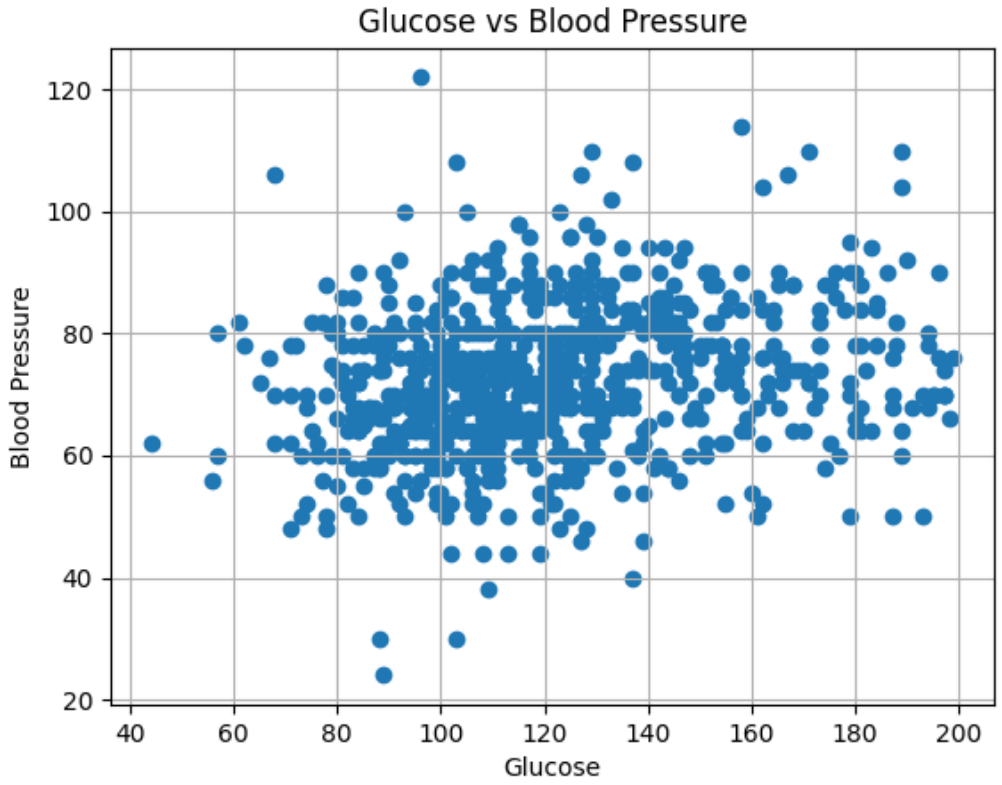
Analytical Distribution

- The code tested normal, exponential, and log normal distribution and picked log normal as the best fit for this data
- Applying the log function to the x-axis will change the distribution to appear normal
- This tells us that the data is positively skewed, and if we are to build a machine learning model with the Glucose variable, we should apply the log function to that column



Scatter Plots





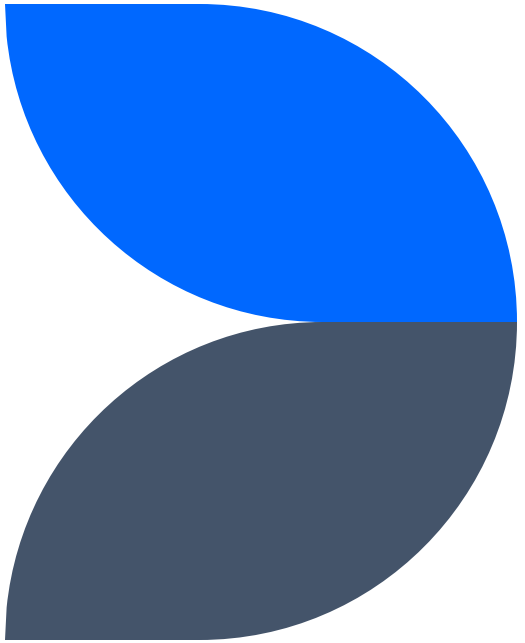
Scatter Plots Cont.

- The covariance matrix shows how the three tested variables change together
- The covariance between Glucose and Blood Pressure indicates that when Glucose levels increase, Blood Pressure tends to increase as well, since the covariance is not extremely high it only suggests a moderate association
- Glucose and BMI also suggests that there may be a relationship between blood sugar and BMI, however the covariance is not very high

	Glucose	Blood Pressure	BMI
Glucose	932.425376	84.811985	49.355133
Blood Pressure	84.811985	153.317842	24.644988
BMI	49.355133	24.644988	47.955463

Hypothesis Testing

Logistical Regression & Regression Analysis



Classical Hypothesis Testing

- The logistical regression results indicate that a significant relationship exists between Plasma Glucose levels, and likelihood of having diabetes
- The p-values are or are near to 0, which indicates the coefficients are significant, confidence intervals further suggests significance
- However, as shown by the Pseudo R-squared, the model only explains approximately 20% of the dependent variable, whether the person has diabetes

Logit Regression Results

Dep. Variable:	Outcome	No. Observations:	763
Model:	Logit	Df Residuals:	761
Method:	MLE	Df Model:	1
Date:	Sat, 12 Aug 2023	Pseudo R-squ.:	0.2028
Time:	12:54:56	Log-Likelihood:	-393.28
converged:	True	LL-Null:	-493.35
Covariance Type:	nonrobust	LLR p-value:	1.949e-45

	coef	std err	z	P>|z|	[0.025	0.975]
const	-5.7151	0.438	-13.045	0.000	-6.574	-4.856
Glucose	0.0406	0.003	12.014	0.000	0.034	0.047

Regression Analysis

- This regression analysis was performed on the explanatory variables Glucose, Blood Pressure, and BMI
- Even with multiple values, the R-squared indicates that the model can only explain 27.7% of variation
- This means that there are other untested variables that may have a higher significance in predicting diabetes, and further testing is required

OLS Regression Results

Dep. Variable:	Outcome	R-squared:	0.277
Model:	OLS	Adj. R-squared:	0.274
Method:	Least Squares	F-statistic:	91.84
Date:	Sat, 12 Aug 2023	Prob (F-statistic):	2.49e-50
Time:	13:19:25	Log-Likelihood:	-371.07
No. Observations:	724	AIC:	750.1
Df Residuals:	720	BIC:	768.5
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.9622	0.107	-9.019	0.000	-1.172	-0.753
Glucose	0.0068	0.001	13.388	0.000	0.006	0.008
BloodPressure	0.0004	0.001	0.336	0.737	-0.002	0.003
BMI	0.0136	0.002	5.881	0.000	0.009	0.018

Omnibus:	49.385	Durbin-Watson:	1.976
Prob(Omnibus):	0.000	Jarque-Bera (JB):	35.700
Skew:	0.438	Prob(JB):	1.77e-08
Kurtosis:	2.355	Cond. No.	1.05e+03

Data Source

https://www.kaggle.com/datasets/ashishkumarjayswal/diabetes-dataset

Thank you

James Patalan