

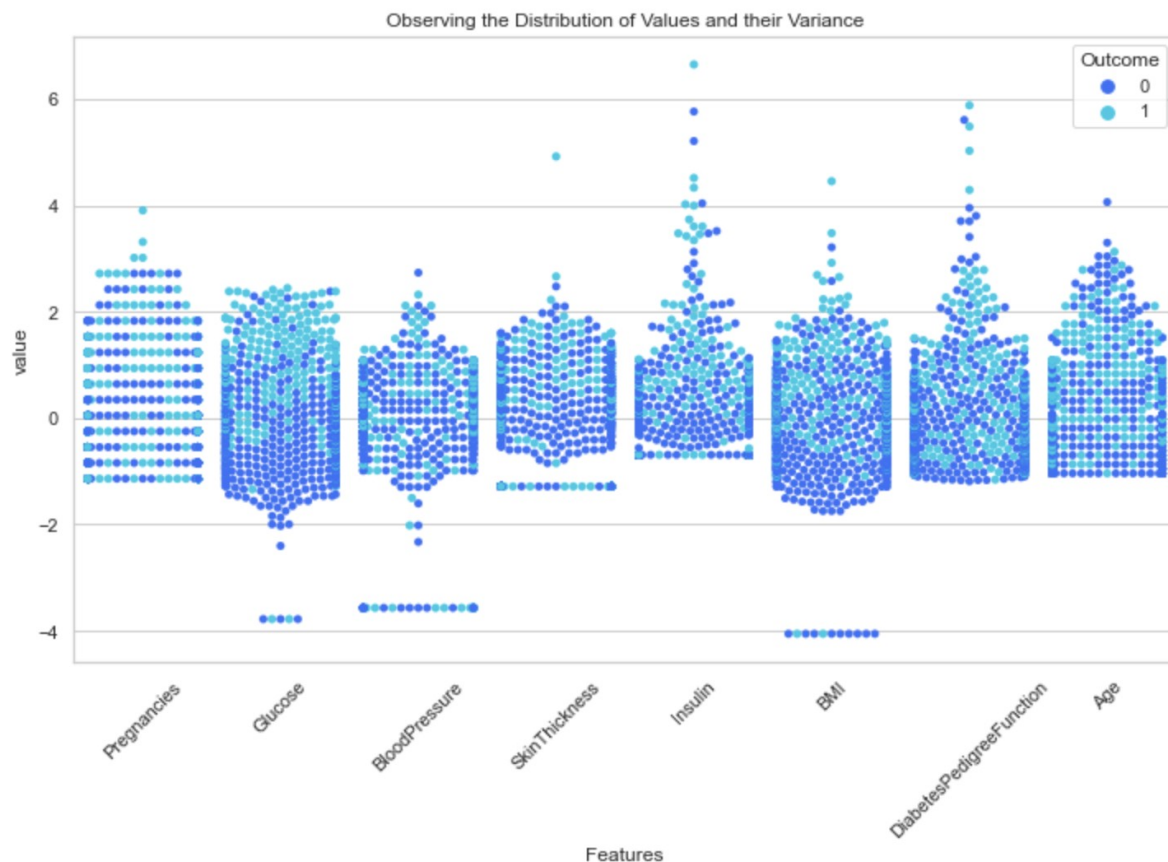
Experimental Design for Data Science

Aryan Chaurasia

Introduction

For the datasets, my approach was to first visualize and conduct some exploratory analysis. Then look for variables that can be used for further experiments. After identifying those variables, I conducted t-tests, ANOVA and regression. My research question was to explore what variables would be significant when it comes to diabetes.

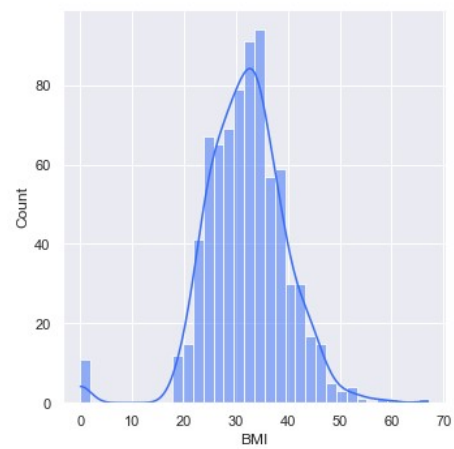
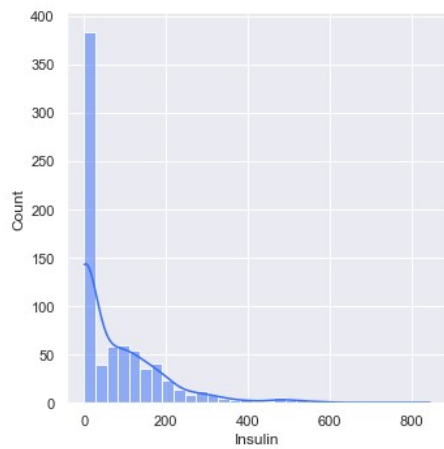
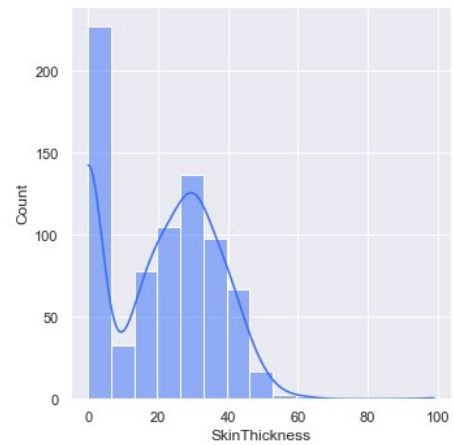
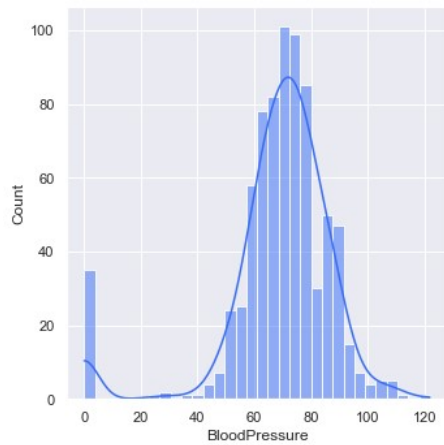
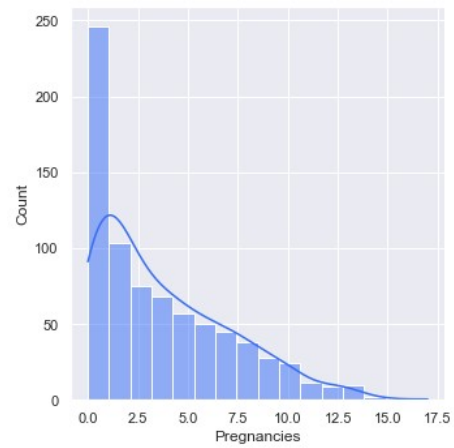
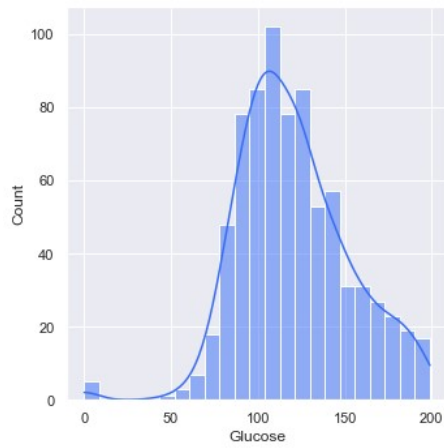
Visualizing distribution of values depending on the Outcome

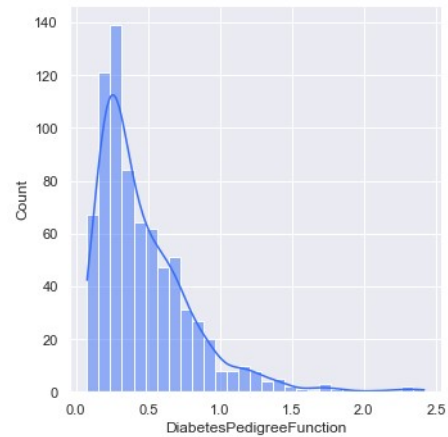
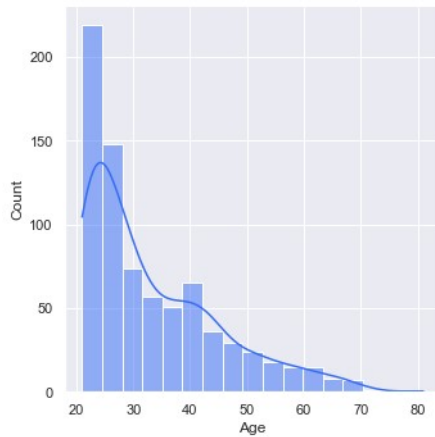


First, I generated some descriptive statistics where I grouped the dataset by Outcome and checked the difference by mean.

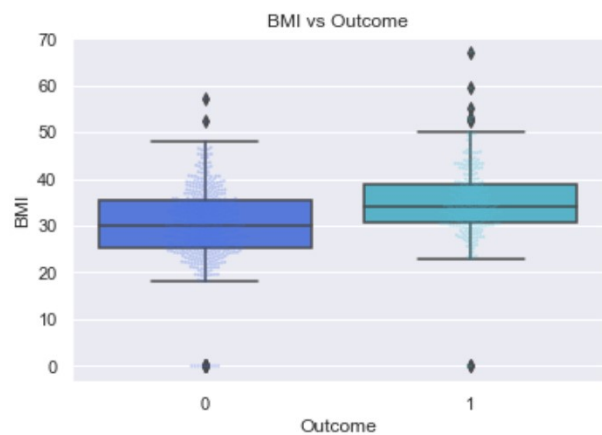
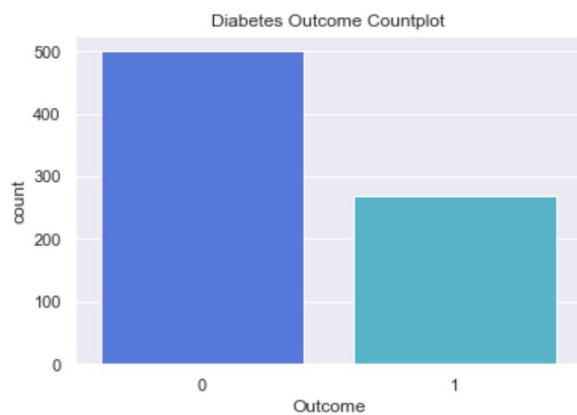
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
Outcome								
0	3.298000	109.980000	68.184000	19.664000	68.792000	30.304200	0.429734	31.190000
1	4.865672	141.257463	70.824627	22.164179	100.335821	35.142537	0.550500	37.067164

Creating Distribution plot for various features, this is necessary to get the basic idea of the data.

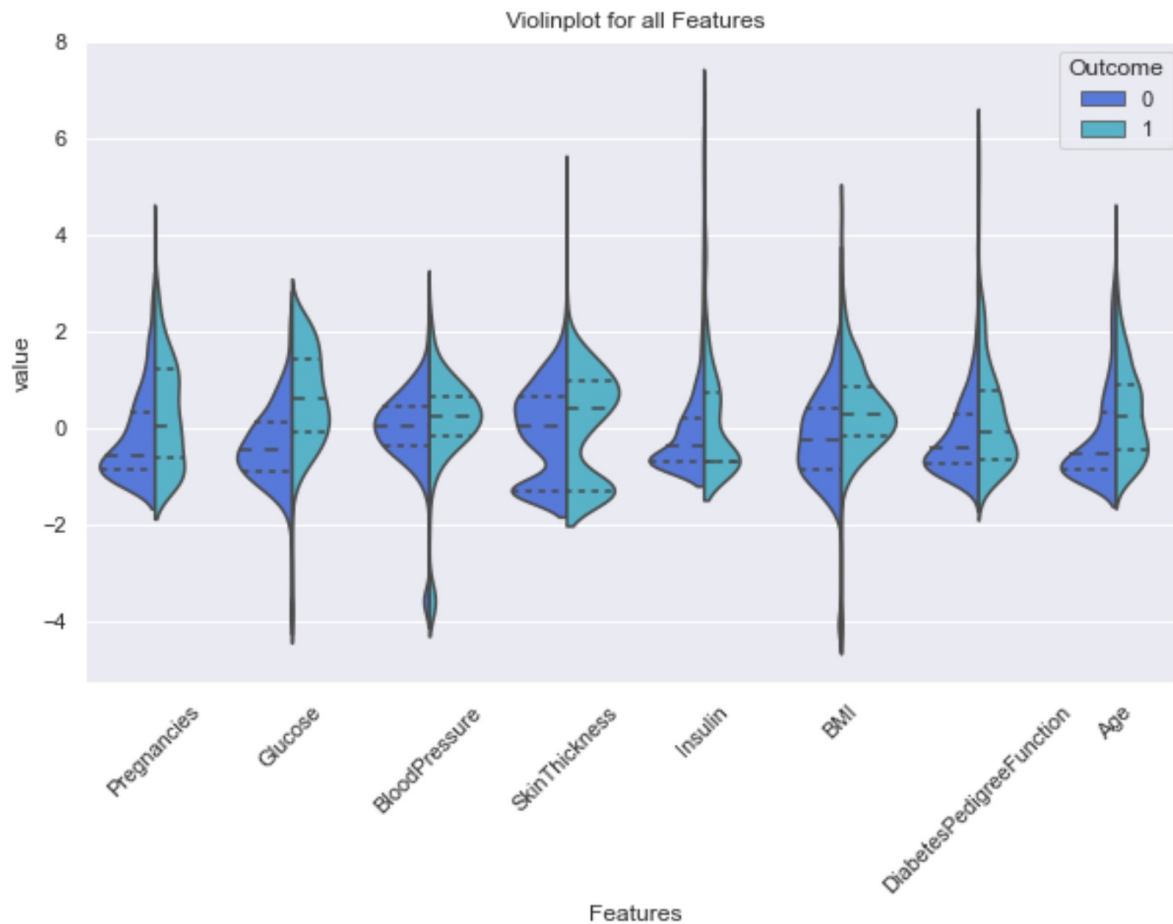




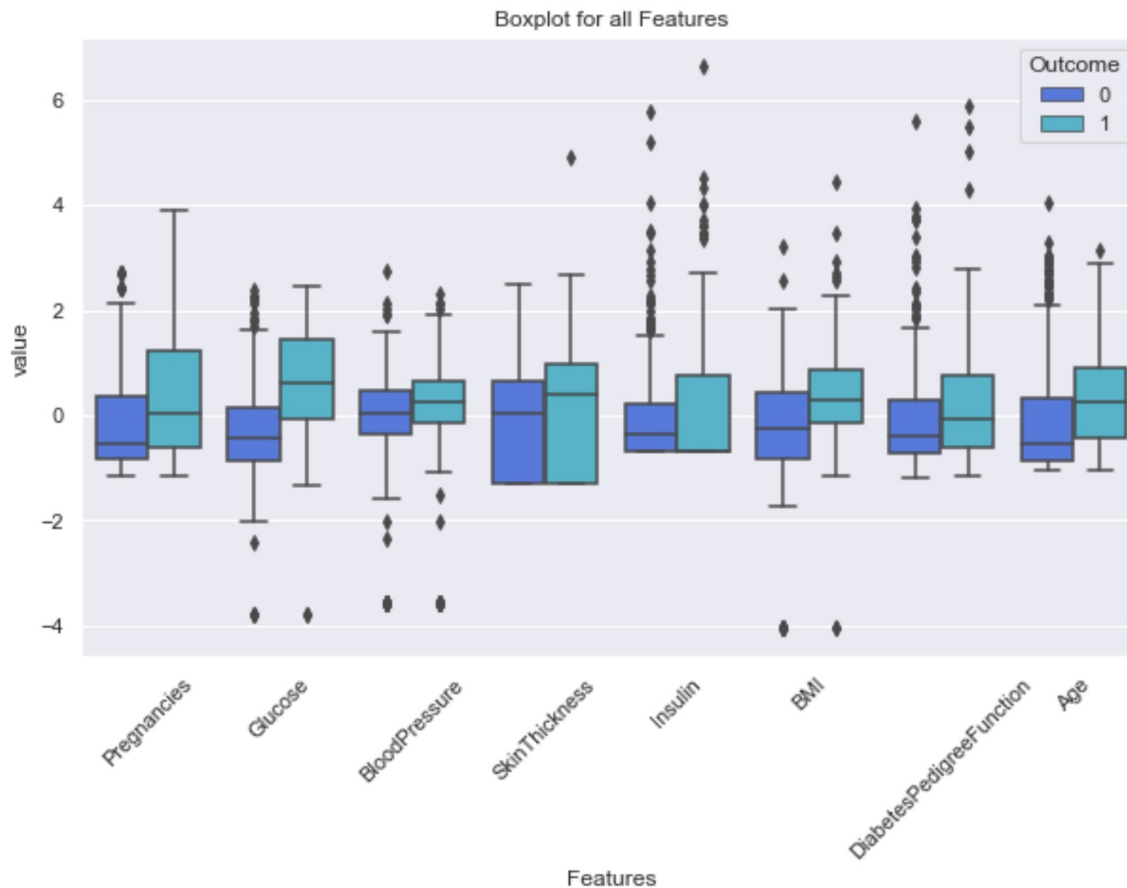
As seen in the distribution plot there are a lot of variances in the data to visualize it together, I standardized the data i.e. subtract the mean and divide by the standard deviation. This is done to put different variables in the same scale. (Frost, 2022)



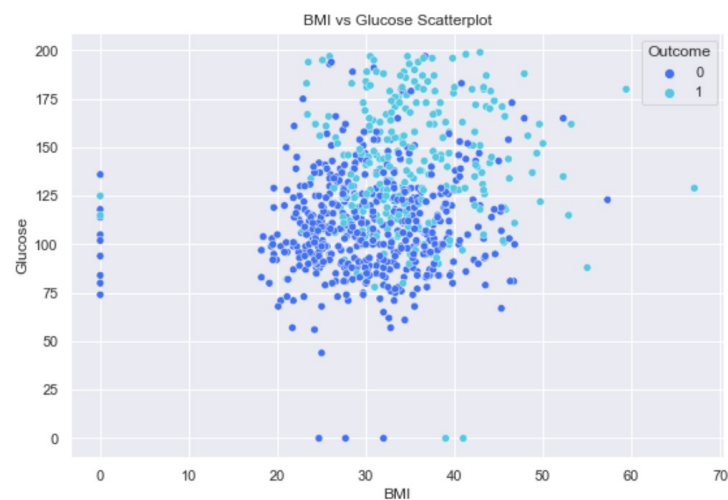
After standardizing the data, I generated a violin plot for all features depending on the outcome. This is important because it helps us visually see which variables have major changes depending on diabetic or non-diabetic. As seen below the difference between glucose and BMI is easily visible in terms of diabetes outcome. Hence, I would choose this for my Anova.



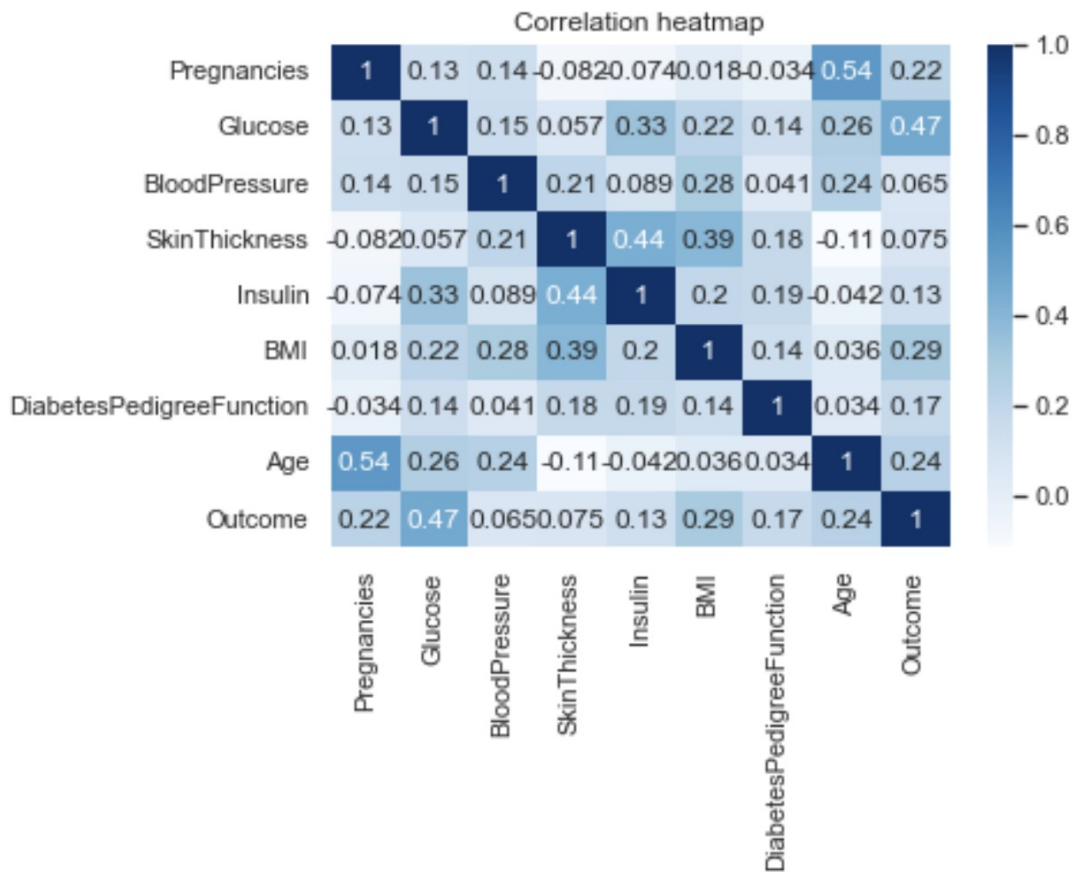
Similarly, a boxplot for all features depending on the outcome.



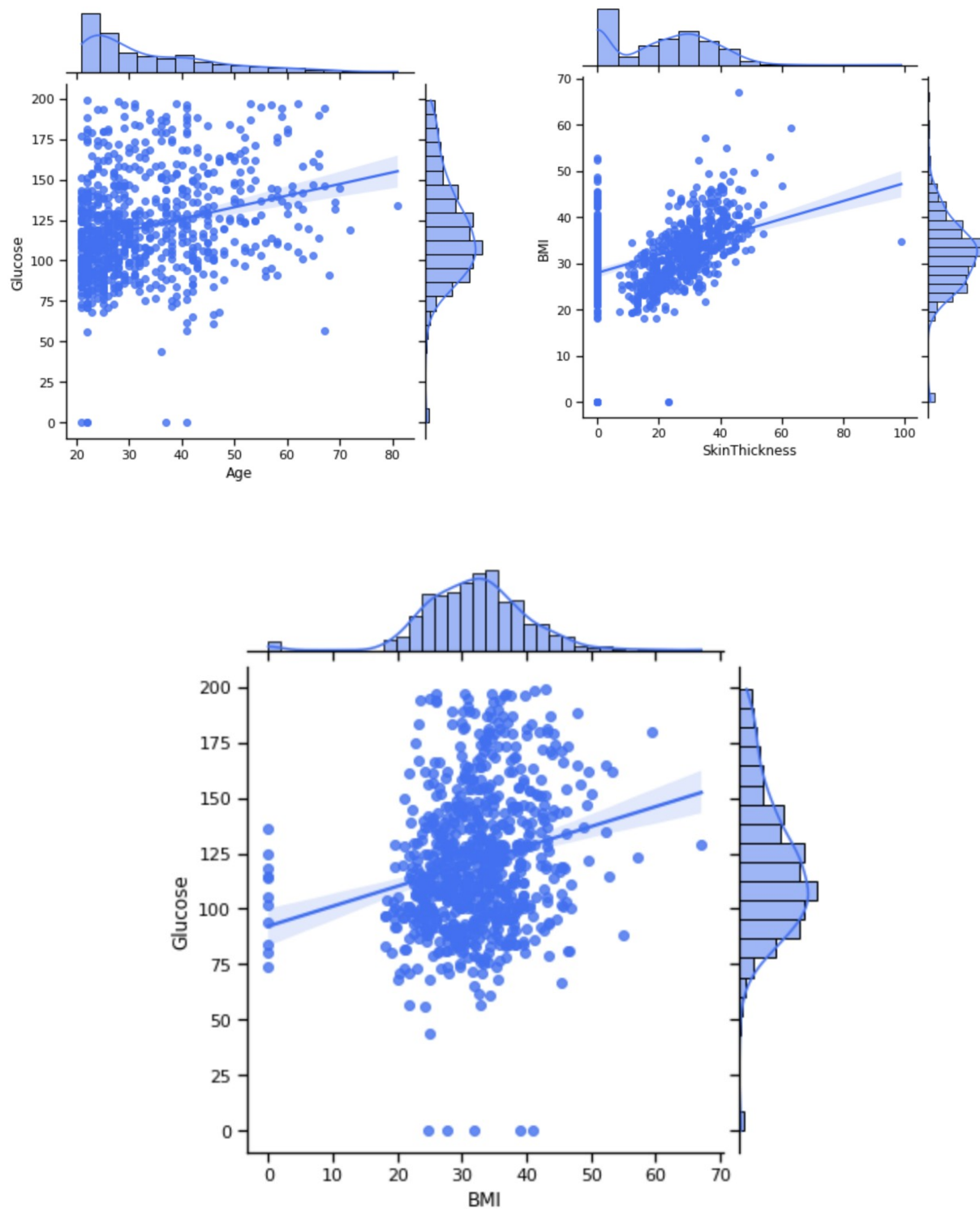
Some scatterplot generated to better visualize the data depending on the outcome. For example, below BMI for people with diabetes is generally higher as seen on scatterplot.



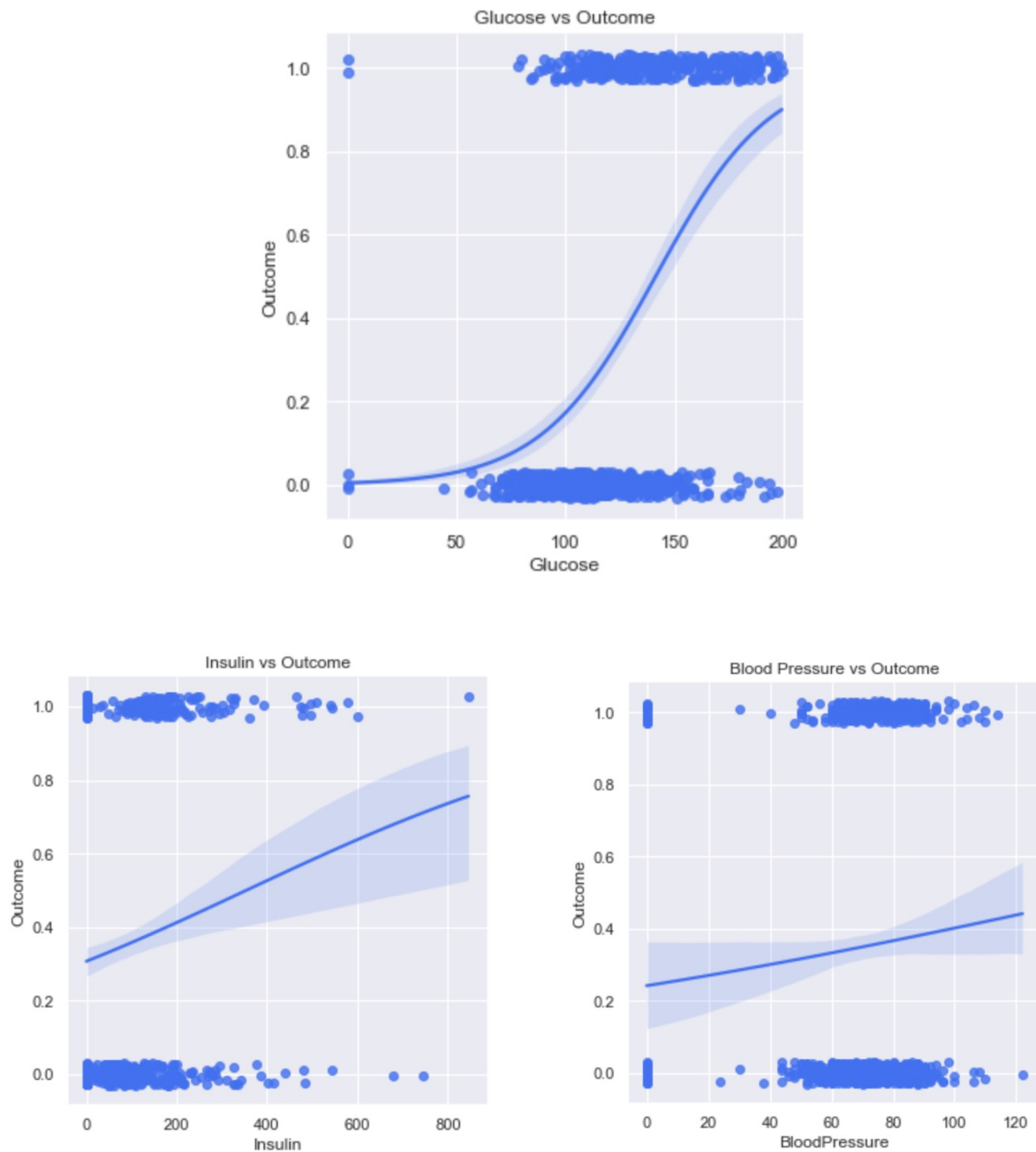
Then I also generated a correlation map to check which variables are correlated with each other. There is a high correlation between glucose and outcome, Insulin and glucose, Insulin and Skin Thickness etc.



Also, then I visualized linear relations between some of the features discovered in the correlation map just to explore further details and discover how these relation look visually.

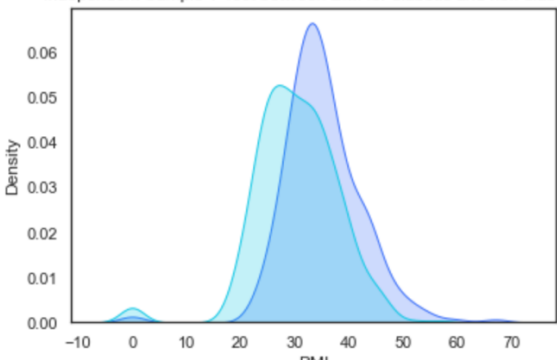
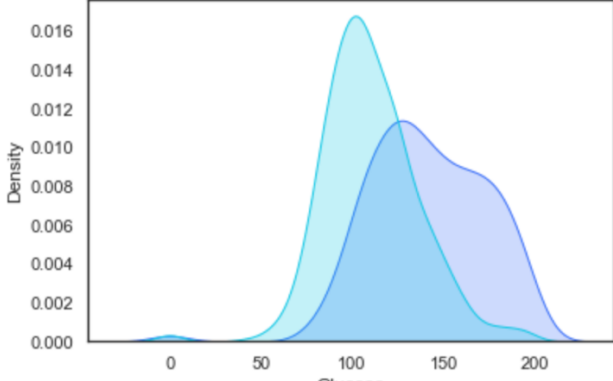
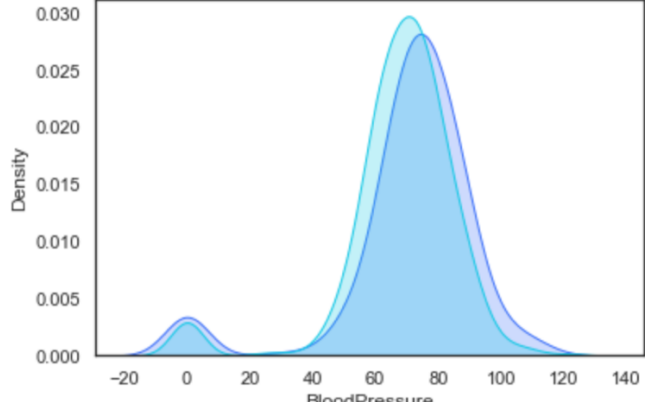


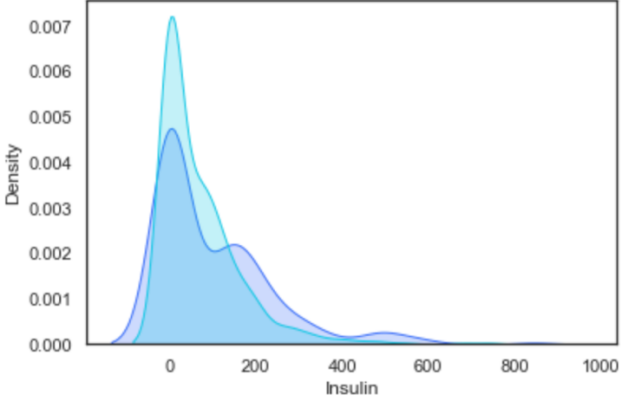
Later I also visualized some features logistically. I picked glucose as one of the main features which was directly proportional to diabetes and when visualized logistically the relation could easily be seen as compared to other graphs below. The glucose vs Outcome graph below clearly shows how the probability of diabetes increases when glucose increases. (Note I didn't conduct any logistic regression it's just visualization to explore the outcome.). Even in the violin plot shown before there is a clear difference in glucose depending on the outcome. Hence due to these visualizations, I chose glucose as one of the important features of my dataset.



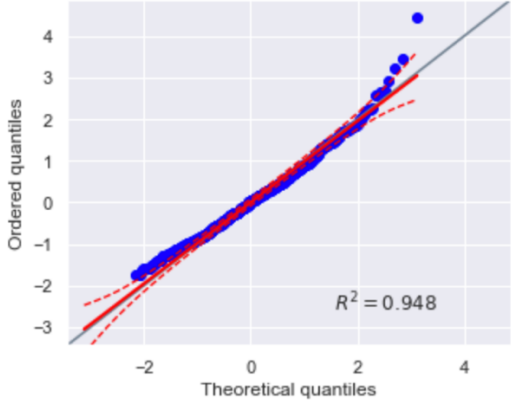
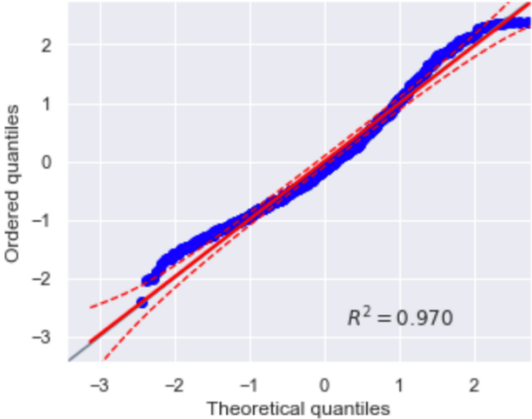
T-Tests

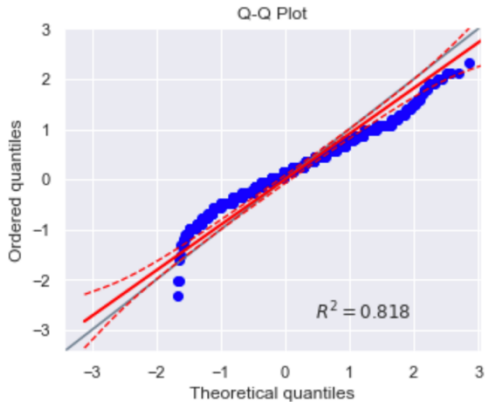
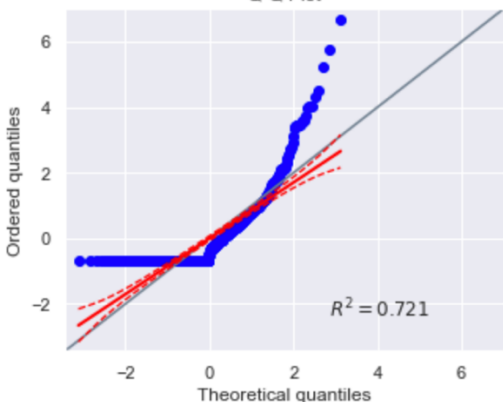
Conducted a few t-tests between features between diabetic and non-diabetic outcomes. Even though in violin plot difference was clearly visible wanted to do a t-test to check if there is any significant difference.

Feature	T-Stat	P-Value
BMI Independent Sample T-Test between BMI for diabetic and non diabetic 	8.619	6.56623762470833e-17
Glucose Independent Sample T-Test between Glucose for diabetic and non diabetic 	13.751	2.6441613495403223e-36
Blood Pressure Independent Sample T-Test between Blood Pressure for diabetic and non diabetic 	1.713	0.087

<p style="text-align: center;">Insulin</p> <p style="text-align: center;">Independent Sample T-Test between Insulin for diabetic and non diabetic</p> 	3.300	0.001
--	-------	-------

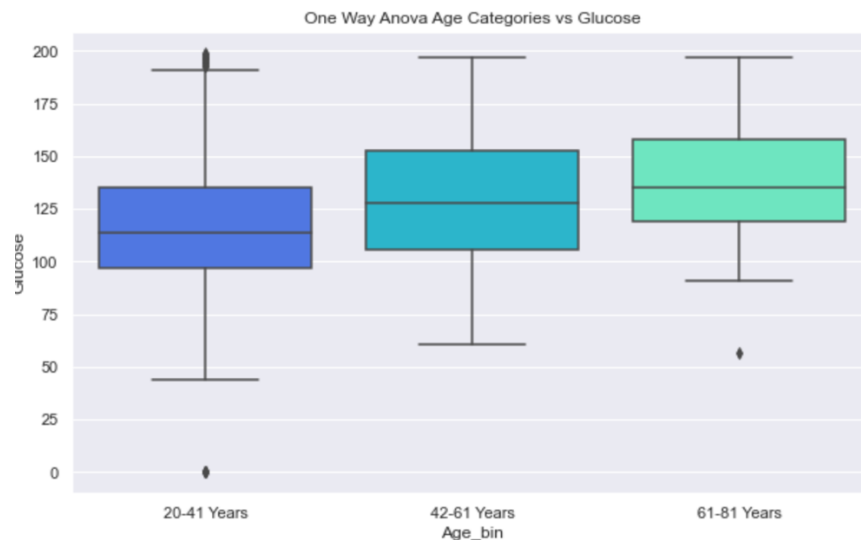
Testing normality with Q-Q plot and Shapiro-Wilk Test

Feature	Stat	P-Value
<p style="text-align: center;">BMI</p> <p style="text-align: center;">Q-Q Plot</p> 	0.949	1.8405621485603632e-15.3f
<p style="text-align: center;">Glucose</p> <p style="text-align: center;">Q-Q Plot</p> 	0.970	1.9867612763291298e-11.3f

<p data-bbox="444 197 662 231">Blood Pressure</p>  <p data-bbox="613 535 711 562">$R^2 = 0.818$</p> <p>The Q-Q Plot for Blood Pressure shows ordered quantiles on the y-axis (ranging from -3 to 3) and theoretical quantiles on the x-axis (ranging from -3 to 3). The data points (blue dots) closely follow the diagonal line, indicating a good fit to the normal distribution. The R^2 value is 0.818.</p>	<p data-bbox="927 197 1008 231">0.818</p>	<p data-bbox="1073 197 1408 268">1.5840069624449098e-28.3f</p>
<p data-bbox="505 663 602 697">Insulin</p>  <p data-bbox="613 1018 711 1045">$R^2 = 0.721$</p> <p>The Q-Q Plot for Insulin shows ordered quantiles on the y-axis (ranging from -2 to 6) and theoretical quantiles on the x-axis (ranging from -2 to 6). The data points (blue dots) show a significant deviation from the diagonal line, particularly at the higher end of the theoretical quantiles, indicating a poor fit to the normal distribution. The R^2 value is 0.721.</p>	<p data-bbox="927 663 1008 697">0.722</p>	<p data-bbox="1073 663 1408 735">7.915248149269491e-34.3f</p>

One-way ANNOVA

Divided age into three categories and took blood pressure as a continuous variable. Glucose is already seen as significantly different in terms of diabetic vs nondiabetic through visualization and t-test conducted before. Through Anova wanted to see the difference in terms of age. Had binned age in three categories.



	Sum_sq	df	F	PR(>F)
Age_bin	24748.894217	2.0	12.467085	0.000005
Residual	759315.562815	765.0	NaN	NaN

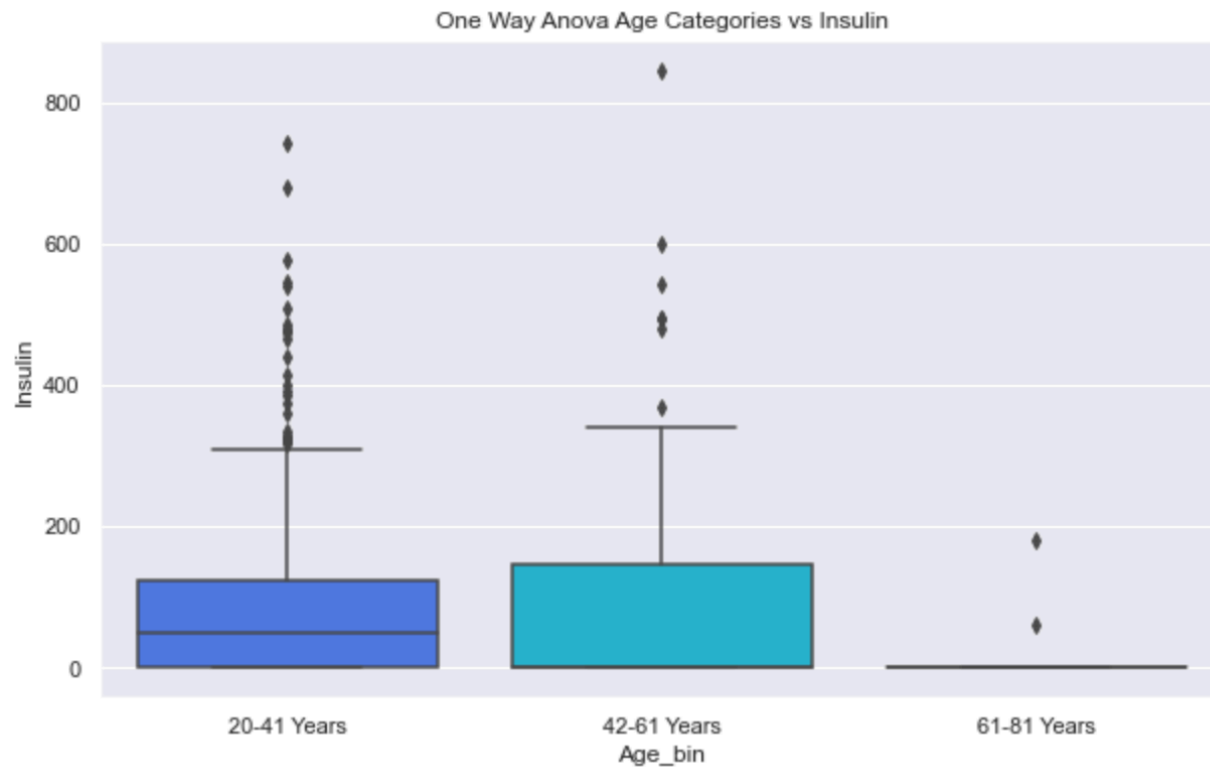
Post hoc test

	coef	std err	t	P> t	Conf. Int. Low	Conf. Int. Upp.	pvalue-hs	reject-hs
42-61 Years-20-41 Years	12.602212	2.901303	4.343639	0.000016	6.906752	18.297672	0.000048	True
61-81 Years-20-41 Years	18.272416	6.431808	2.840945	0.004618	5.646327	30.898505	0.009214	True
61-81 Years-42-61 Years	5.670204	6.815787	0.831922	0.405712	-7.709661	19.050069	0.405712	False

Result Interpretation

A significant difference was found as the age group increases there is a significant difference 42-61 Years old have difference with 20-41 years old group and similarly, 61-81 years old group are also significantly different from than 20-41 Years old. However, there isn't a significant difference between 42-61 years and 61-81 years.

Another one-way ANNOVA between age bins and Insulin



	Sum_sq	df	F	PR(>F)
Age_bin	1.344487e+05	2.0	5.115948	0.006207
Residual	1.005222e+07	765.0	NaN	NaN

Post Hoc test

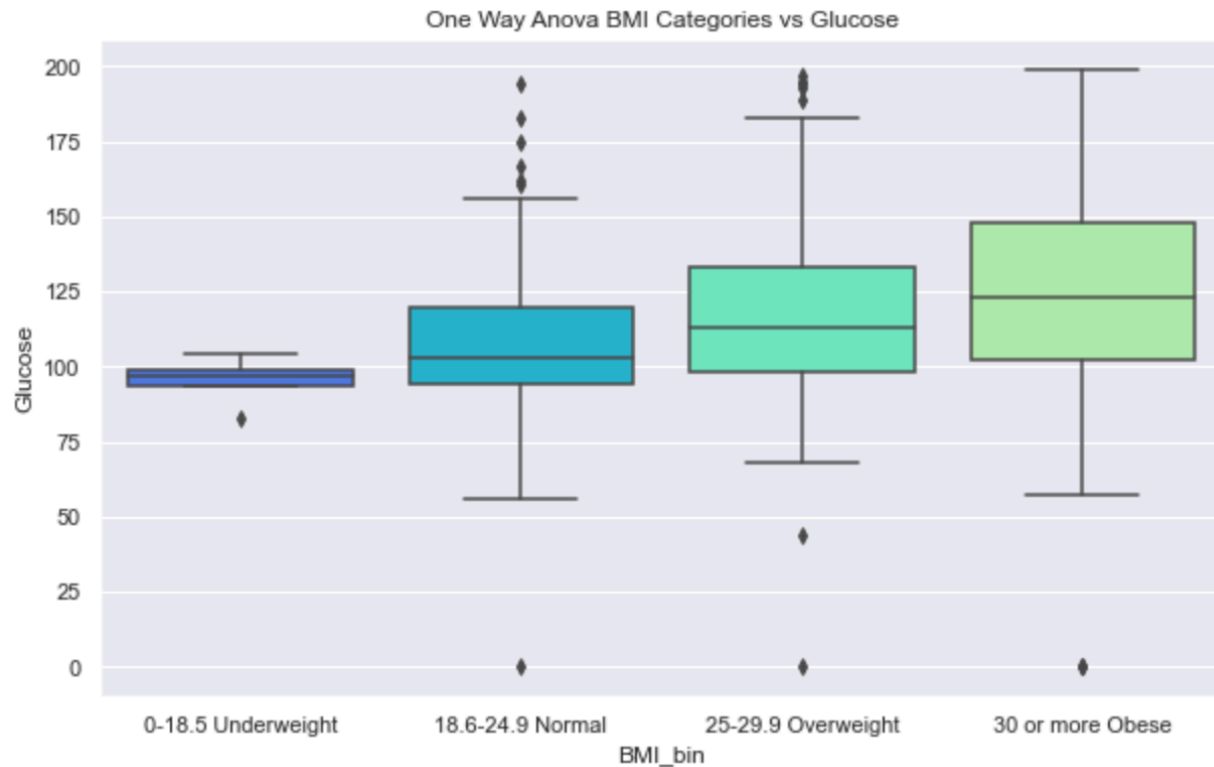
	coef	std err	t	P> t	Conf. Int. Low	Conf. Int. Upp.	pvalue-hs	reject-hs
42-61 Years-20-41 Years	7.761916	10.556326	0.735286	0.462391	-12.960890	28.484722	0.462391	False
61-81 Years-20-41 Years	-71.025839	23.401994	-3.035034	0.002487	-116.965587	-25.086091	0.004967	True
61-81 Years-42-61 Years	-78.787755	24.799090	-3.177042	0.001548	-127.470101	-30.105409	0.004636	True

Result Interpretation

In terms of Insulin and age there is a significant difference between 61-81 years and 20-41 years. And 61-81 years and 42-61 years. However, there isn't significant difference between 20-41 years and 42-61 years.

Another one-way ANNOVA between BMI bins and Glucose

Other than age as categorical values I also tried with **BMI as categorical values**, because BMI is directly related to health and can easily be caricaturized. The category is chosen as per CDCP. (CDCP, 2021)



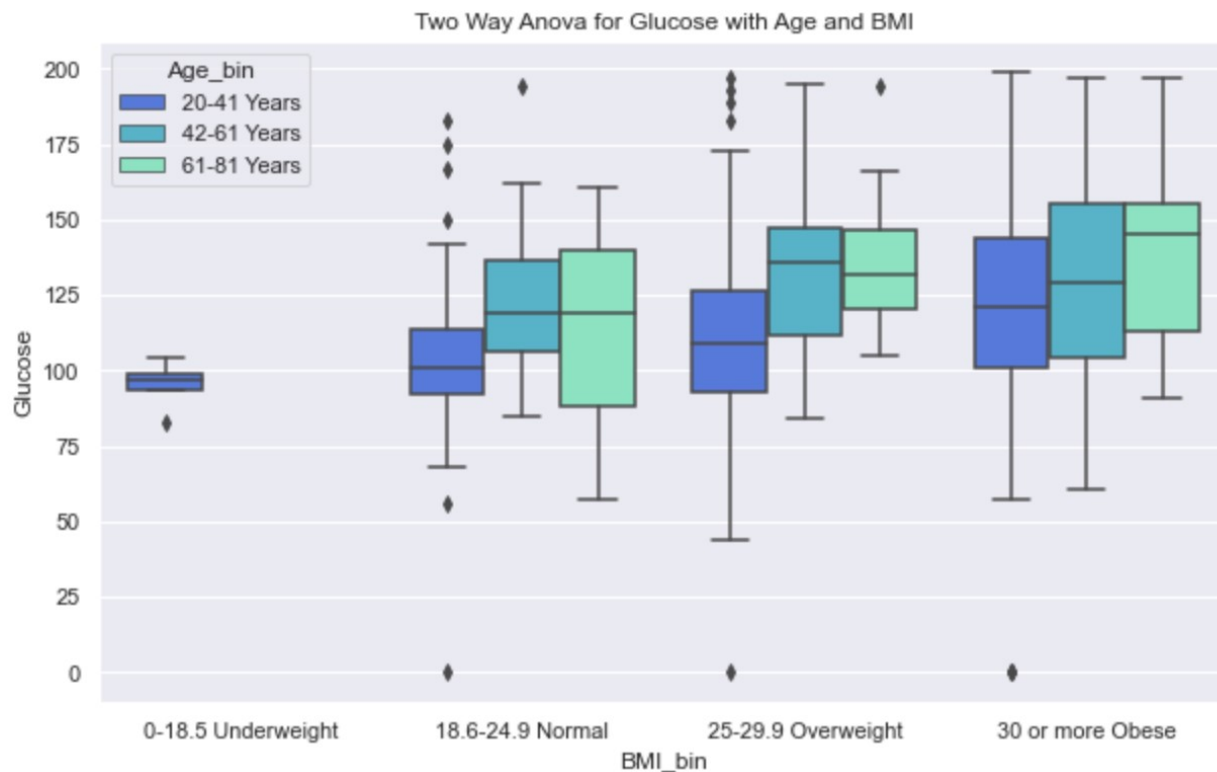
	Sum_sq	df	F	PR(>F)
BMI_bin	35674.853359	3.0	12.076956	9.981209e-08
Residual	741444.132110	753.0	NaN	NaN

Post hoc test

	coef	std err	t	P> t	Conf. Int. Low	Conf. Int. Upp.	pvalue-hs	reject-hs
18.6-24.9 Normal-0-18.5 Underweight	12.730392	15.994276	0.795934	4.263211e-01	-18.668281	44.129065	0.426321	False
25-29.9 Overweight-0-18.5 Underweight	21.096369	15.863929	1.329833	1.839761e-01	-10.046418	52.239156	0.334105	False
30 or more Obese-0-18.5 Underweight	30.764831	15.755936	1.952587	5.123875e-02	-0.165953	61.695614	0.145975	False
25-29.9 Overweight-18.6-24.9 Normal	8.365977	3.892856	2.149059	3.194783e-02	0.723836	16.008117	0.121797	False
30 or more Obese-18.6-24.9 Normal	18.034438	3.426309	5.263517	1.844946e-07	11.308184	24.760693	0.000001	True
30 or more Obese-25-29.9 Overweight	9.668462	2.754449	3.510126	4.745342e-04	4.261150	15.075774	0.002370	True

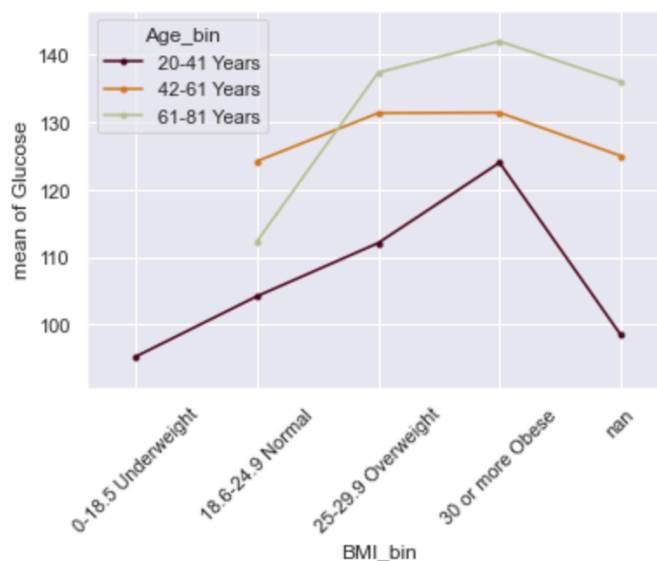
Two Way Anova

For two-way Anova, I chose BMI bins and age for categorical values and glucose for continuous values. This is important because BMI is an important health feature, along with age and if used with glucose a good analysis can be done to check variance. I also thought to use Insulin however in the dataset insulin has a lot of zero's which doesn't make it desirable.



	sum_sq	df	F	PR(>F)
C(BMI_bin)	3.567485e+04	3.0	1.241663e+01	6.238334e-08
C(Age_bin)	5.267262e-10	2.0	2.749905e-13	1.000000e+00
C(BMI_bin):C(Age_bin)	2.602934e+04	6.0	4.529753e+00	1.601104e-04
Residual	7.154148e+05	747.0	NaN	NaN

Interaction plot



Post Hoc Test

	group1	group2	Diff	Lower	Upper	q-value	p-value
0	30 or more Obese	25-29.9 Overweight	9.668462	2.240274	17.096649	5.033392	0.003622
1	30 or more Obese	18.6-24.9 Normal	18.034438	8.794380	27.274497	7.547691	0.001000
2	30 or more Obese	nan	21.742103	-4.068377	47.552584	3.257554	0.144753
3	30 or more Obese	0-18.5 Underweight	30.764831	-11.725719	73.255380	2.799938	0.276919
4	25-29.9 Overweight	18.6-24.9 Normal	8.365977	-2.132262	18.864215	3.081672	0.188651
5	25-29.9 Overweight	nan	12.073641	-14.213527	38.360810	1.776154	0.692696
6	25-29.9 Overweight	0-18.5 Underweight	21.096369	-21.685416	63.878154	1.906931	0.640290
7	18.6-24.9 Normal	nan	3.707665	-23.147799	30.563129	0.533893	0.900000
8	18.6-24.9 Normal	0-18.5 Underweight	12.730392	-30.402911	55.863695	1.141341	0.900000
9	nan	0-18.5 Underweight	9.022727	-40.386625	58.432079	0.706179	0.900000

Result Interpretation

A significant difference is observed between people with a BMI of more than 30 and people who are overweight and also a significant difference is found between Obese bin and people with normal weight.

Linear Regression

For linear regression I tried to predict BMI using various coefficients. These coefficients were Glucose, Blood Pressure, Insulin, Skin thickness, and age. I didn't choose diabetic function and pregnancies because their correlation was pretty low. Skin thickness had higher correlation with BMI.

The results were as follows:

	Actual	Predicted
661	42.9	41.556800
122	33.6	33.677324
113	34.0	25.972100
14	25.8	33.798867
529	24.6	27.923337
...
476	33.7	36.289348
482	27.8	30.031686
230	44.0	31.171224
527	26.3	31.335614
380	30.8	33.589928

Mean Absolute Error: 5.220901455625269

Mean Squared Error: 49.52461501490933

Root Mean Squared Error: 7.037372735254921

REFERENCE

Centers for Disease Control and Prevention. (2021, August 27). *About adult BMI*. Centers for Disease Control and Prevention. Retrieved March 28, 2022, from https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html

Frost, J. (n.d.). *Standardization- Statistics By Jim*. Statistics By Jim. Retrieved March 28, 2022, from <https://statisticsbyjim.com/glossary/standardization/>