1. Main objective of the analysis that specifies whether your model will be focused on prediction or interpretation.

- The objective of the analysis is to predict glucose given the diagnostic measures. It may help health care professionals to predict the glucose level of individuals for better glycaemic control and guide insulin dosing to minimise the risk of development of health complications especially diabetes due to elevated glucose level.

2. Brief description of the data set you chose and a summary of its attributes.

- The data was uploaded to Kaggle by Houcem Benmansourand. It's called "Predict diabetes based on diagnostic measure" version one and originally from the National Institute of Diabetes and Digestive and Kidney Diseases. It is accessible via https://www.kaggle.com/houcembenmansour/predict-diabetes-based-on-diagnostic-measures.
- The dataset contains 390 entries from 390 different patients and 16 data columns in CSV file format.
- The 16 data columns are the patient number, cholesterol, glucose, HDL cholesterol, cholesterol/HDL ratio, age, gender (male or female), height, weight, BMI, systolic blood pressure, diastolic blood pressure, waist size, hip size, waist/hip ratio and diabetes status (diabetes or no diabetes). All are numerical values except for gender and diabetes status which are dichotomous.

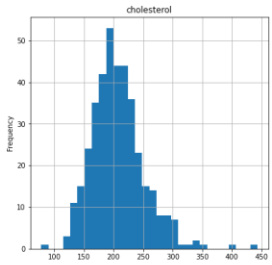
3. Brief summary of data exploration and actions taken for data cleaning and feature engineering.

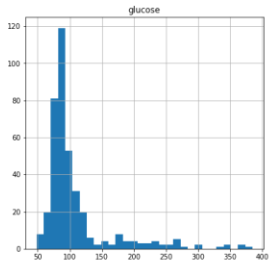
- Following is the last five entries of the data:

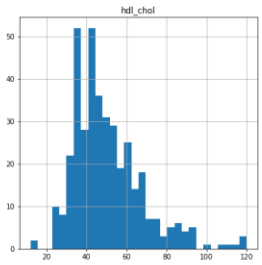
patient_number	cholesterol	glucose	hdl_chol	chol_hdl_ratio	age	gender	height	weight	bmi	systolic_bp	diastolic_bp	waist	hip	waist_hip_ratio	diabetes
386	227	105	44	5.2	83	female	59	125	25.2	150	90	35	40	0.88	No diabetes
387	226	279	52	4.3	84	female	60	192	37.5	144	88	41	48	0.85	Diabetes
388	301	90	118	2.6	89	female	61	115	21.7	218	90	31	41	0.76	No diabetes
389	232	184	114	2.0	91	female	61	127	24.0	170	82	35	38	0.92	Diabetes
390	165	94	69	2.4	92	female	62	217	39.7	160	82	51	51	1.00	No diabetes

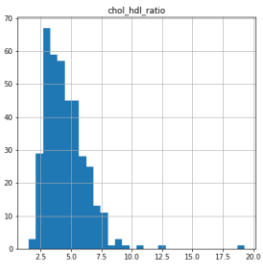
- There are neither duplicates nor missing values in the data.
- Patient number was not considered because it adds no value to the analysis.
- One-hot encoding was done for gender and diabetes.

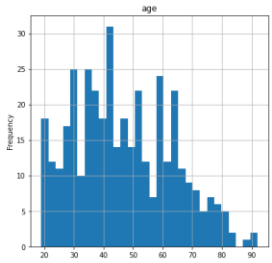
- Following is the histogram of all the numerical features. Glucose looks the most right skewed compared to the rest. There is an obvious spike at 140 for systolic blood pressure. Normality is

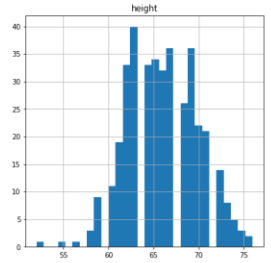


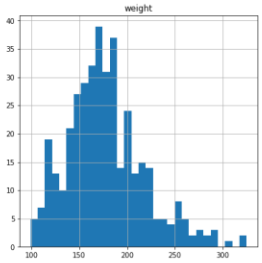


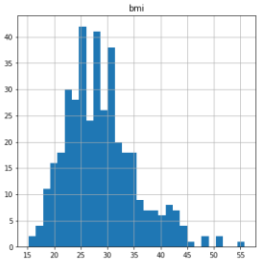


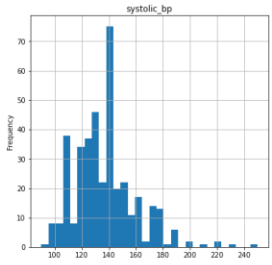


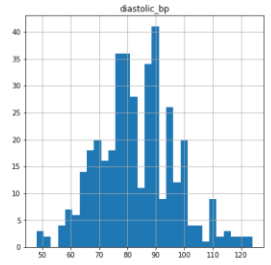


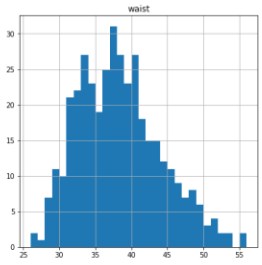


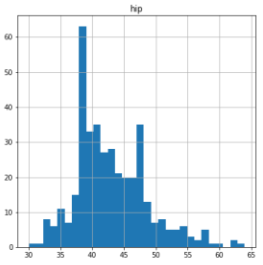


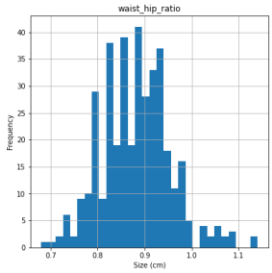




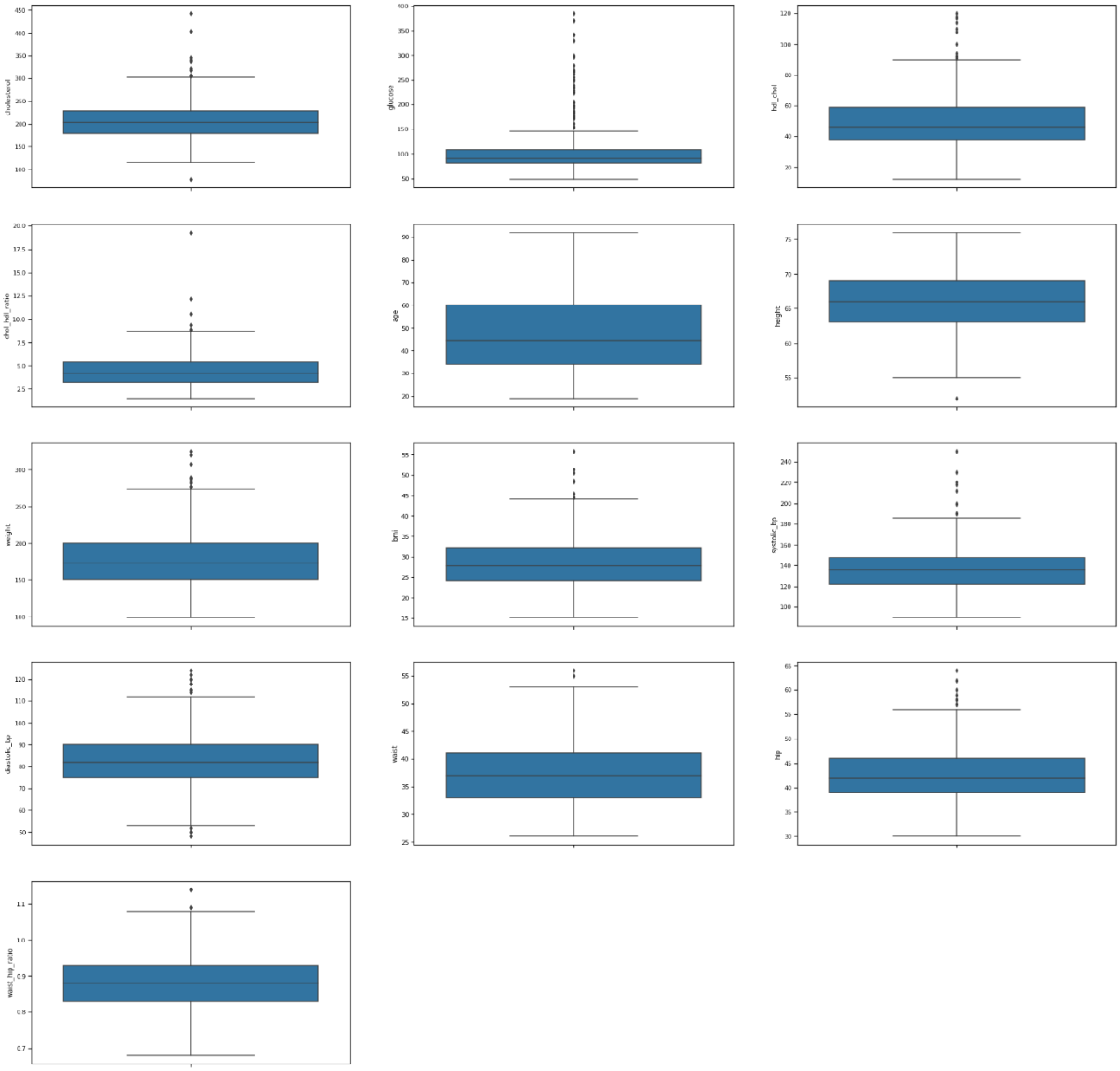




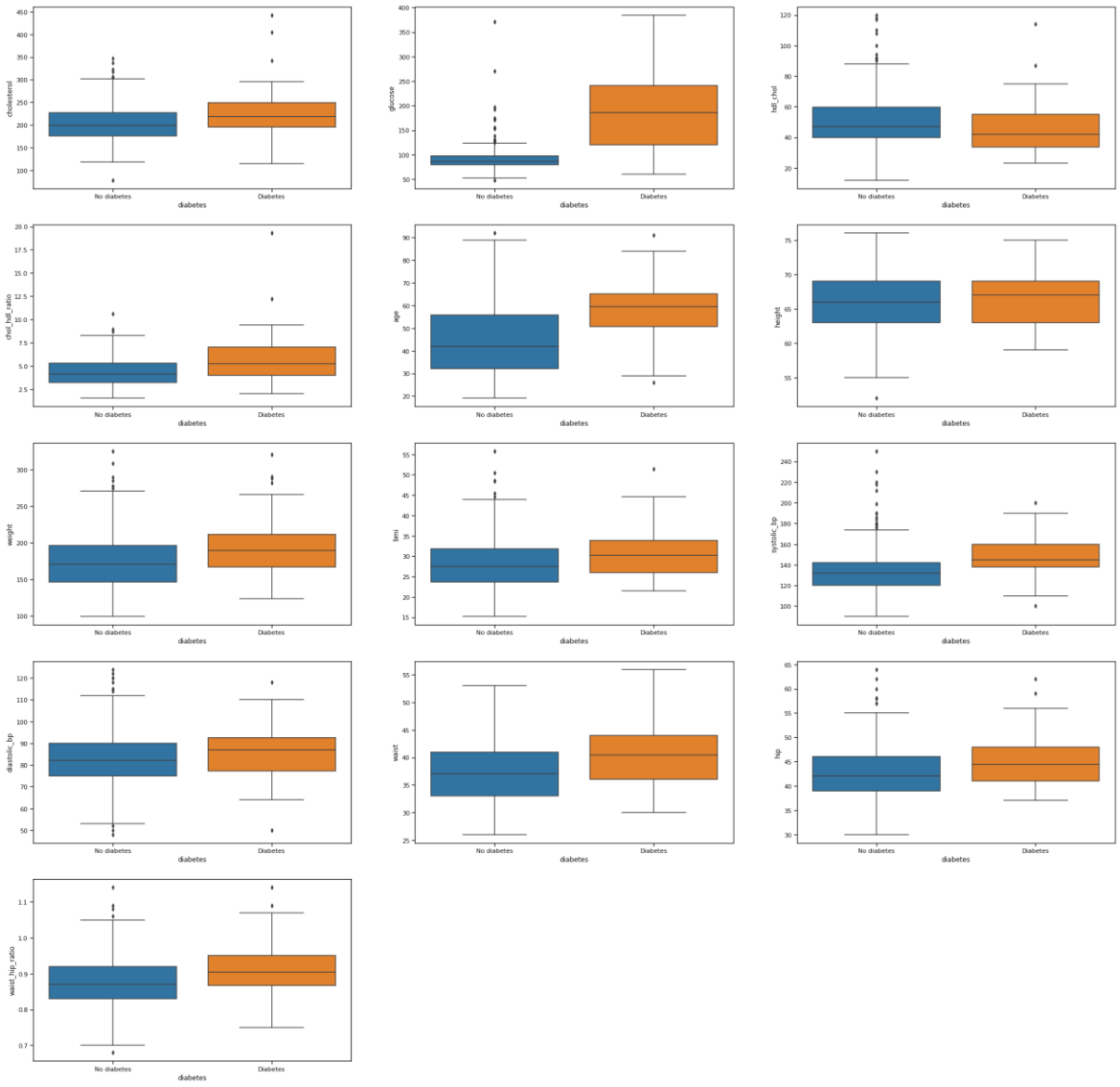




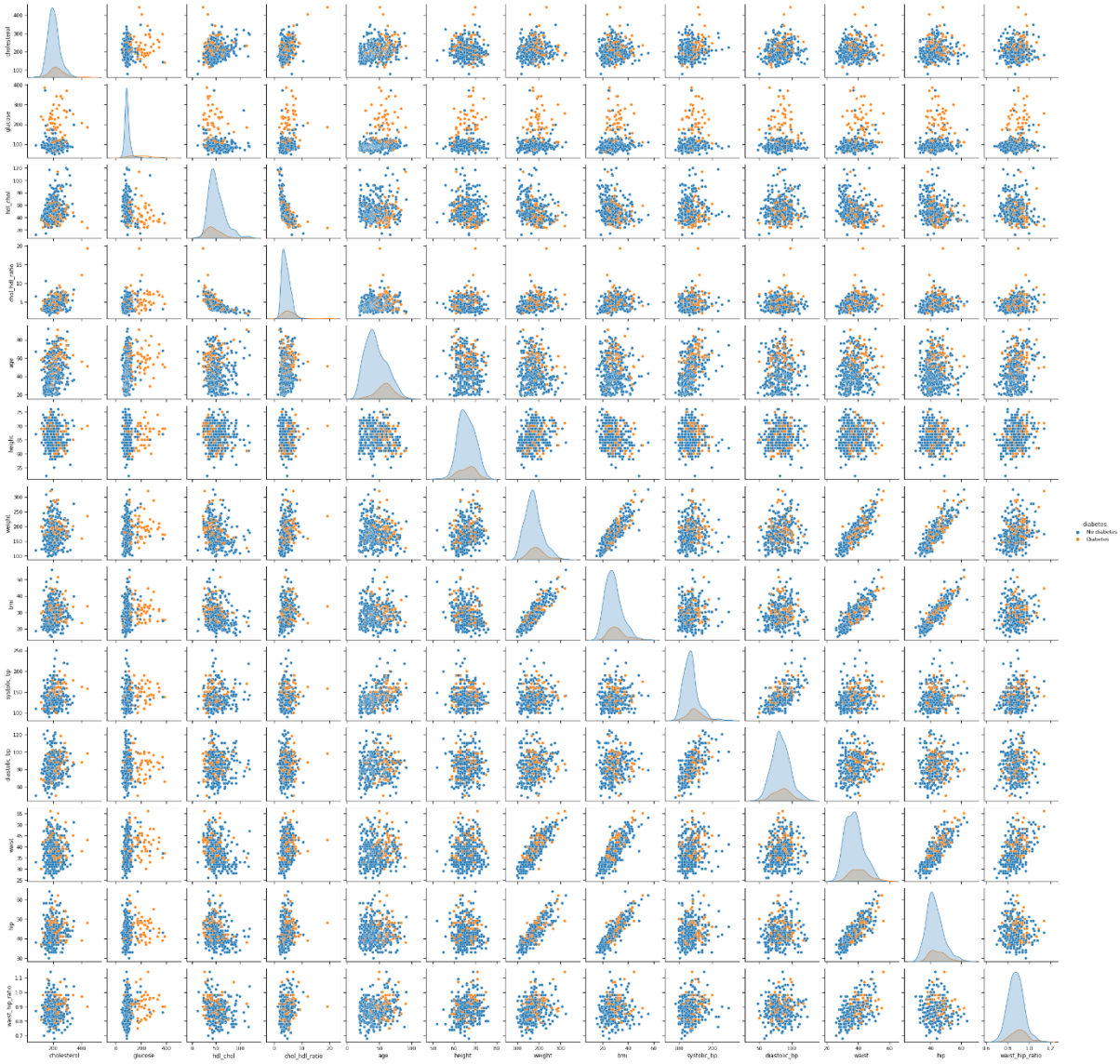
- The following are the boxplots of all numerical features. The variation of glucose is the greatest compared to the other features. There are more outliers in glucose than other features.



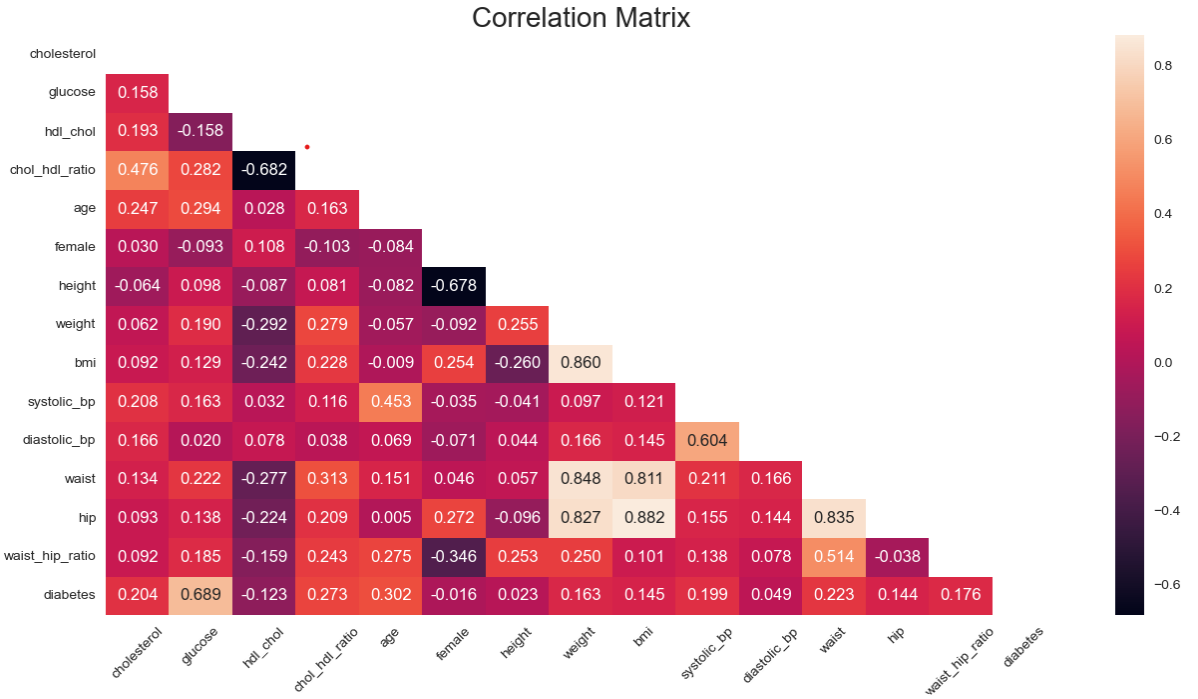
- The following are the boxplots of all numerical features by diabetes groups. The variation of glucose is the greatest for this groups compared to the other features. Outliers are most seen in non-diabetes group than diabetes group for most of this features.



- The following is the pairwise scatter plots with KDE. There are more non-diabetes than diabetes patients in the dataset hence the obvious spike in all KDE plots. The scatter plots provide a visual representation of the relationship between the features. The next correlation matrix gives the correlation coefficients between this features.



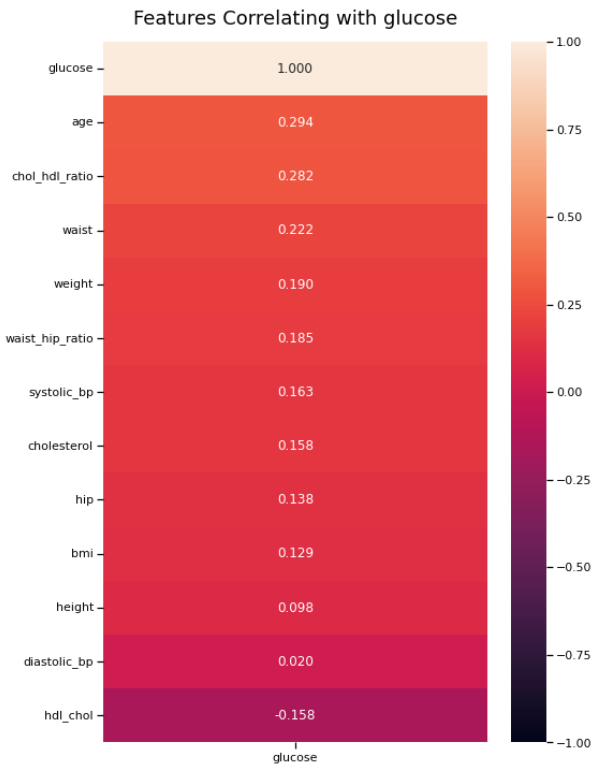
- The following is the correlation matrix of the data:



Together with the pairwise scatterplots visualization, high correlations are found between the following features:

- Glucose and diabetes (0.689)
- Cholesterol-to-HDL ratio and HDL cholesterol (-0.682)
- Female and height (-0.678)
- Weight with hip (0.827), waist (0.848) and BMI (0.86)
- BMI with hip (0.882) and waist (0.811)
- Systolic BP and diastolic BP (0.604)
- Waist and hip (0.835)

Looking at this correlations, it is reasonable to have anthropometric measurements (weight, hip, waist and BMI) correlating with each other. Glucose has been the standard biomarker for diabetes which as well shows here. The rest of the biomarkers are useful indicators for monitoring patients at risk for diabetes in healthcare.



The correlation coefficient of numerical features on glucose in descending order is shown. The magnitude of this correlations is not strong i.e. <0.5, where majority is positively correlated except for HDL cholesterol with glucose.

- The following table is the summary statistics by diabetes group. We can see that the average of glucose is much more elevated for diabetes patients (194.2) than patients without diabetes (91.6). On average, diabetes patients are older (58.4) and weighted heavier (192.8) with higher cholesterol (228.6).

Feature	Statistics	No diabetes	Diabetes	Total
cholesterol	**count**	330	60	390
	mean	203.35	228.60	207.23
	std	41.08	56.53	44.67
	min	78	115	78
	25%	175.25	195.75	179
	50%	199	219	203
	75%	226.75	249.75	229
	max	347	443	443
glucose	**count**	330	60	390
	mean	91.56	194.17	107.34
	std	26.87	77.44	53.80
	min	48	60	48
	25%	79	120	81
	50%	86.5	186	90
	75%	97	241.25	107.75
	max	371	385	385
HDL	**count**	330	60	390
	mean	51.17	45.28	50.27
	std	17.23	16.85	17.28
	min	12	23	12
	25%	40	33.75	38
	50%	47	42	46

	75%	59.75	55	59
	max	120	114	120
cholesterol/HDL ratio	**count**	330	60	390
	mean	4.32	5.64	4.52
	std	1.43	2.63	1.74
	min	1.5	2	1.5
	25%	3.2	4	3.2
	50%	4.1	5.2	4.2
	75%	5.3	7	5.4
	max	10.6	19.3	19.3
age	**count**	330	60	390
	mean	44.66	58.4	46.77
	std	16.11	13.12	16.44
	min	19	26	19
	25%	32	50.75	34
	50%	42	59.5	44.5
	75%	55.75	65.25	60
	max	92	91	92
height	**count**	330	60	390
	mean	65.91	66.17	65.95
	std	3.94	3.82	3.92
	min	52	59	52
	25%	63	63	63
	50%	66	67	66
	75%	69	69	69
	max	76	75	76
weight	**count**	330	60	390
	mean	174.60	192.83	177.41
	std	39.84	40.34	40.41
	min	99	123	99
	25%	146	166.5	150.25
	50%	170	189	173
	75%	195.75	211	200
	max	325	320	325
bmi	**count**	330	60	390
	mean	28.37	31.02	28.78
	std	6.58	6.29	6.60
	min	15.2	21.5	15.2
	25%	23.6	25.95	24.1
	50%	27.5	30.2	27.8
	75%	31.78	33.8	32.28
	max	55.8	51.4	55.8
systolic BP	**count**	330	60	390
	mean	135.2	147.77	137.13
	std	22.76	20.50	22.86
	min	90	100	90
	25%	120	138	122
	50%	132	145	136

	75%	142	160	148
	max	250	200	250
diastolic BP	**count**	330	60	390
	mean	83.01	84.85	83.29
	std	13.60	12.91	13.50
	min	48	50	48
	25%	75	77.25	75
	50%	82	87	82
	75%	90	92.5	90
	max	124	118	124
waist	**count**	330	60	390
	mean	37.32	40.88	37.87
	std	5.60	5.75	5.76
	min	26	30	26
	25%	33	36	33
	50%	37	40.5	37
	75%	41	44	41
	max	53	56	56
hip	**count**	330	60	390
	mean	42.65	44.9	43.00
	std	5.64	5.476297	5.66
	min	30	37	30
	25%	39	41	39
	50%	42	44.5	42
	75%	46	48	46
	max	64	62	64
waist/hip ratio	**count**	330	60	390
	mean	0.88	0.91	0.88
	std	0.07	0.08	0.07
	min	0.68	0.75	0.68
	25%	0.83	0.87	0.83
	50%	0.87	0.91	0.88
	75%	0.92	0.95	0.93
	max	1.14	1.14	1.14

Males have slightly higher glucose (113.29) on average than females (103.11).
There are more females than males in the data.

	gender	**Female**	**Male**	**Total**
glucose	**count**	228	162	390
	mean	103.11	113.29	107.34
	std	44.07	64.75	53.80
	min	52	48	48
	25%	81	81	81
	50%	90	89.5	90
	75%	105.25	112	107.75
	max	299	385	385

4. Summary of training at least three linear regression models which should be variations that cover using a simple linear regression as a baseline, adding polynomial effects, and using a regularization regression. Preferably, all use the same training and test splits, or the same cross-validation method.

- Considering all the features and target i.e. glucose, a train-test split with 80% and 20% train and test set respectively was used before starting training the linear regression models, 2nd degree polynomial regression, Ridge and Lasso by fit_transforming the training set, then transforming the test set.

  The following are the performance metrics from the four linear regressions on the data.

Linear regression	R^2	RMSE
Linear regression	0.6457	33.5723
Lasso regression	0.6333	34.1552
Ridge regression	0.6434	33.6818
Polynomial regression (2nd degree polynomial)	-0.4846	68.7234

5. A paragraph explaining which of your regressions you recommend as a final model that best fits your needs in terms of accuracy and explainability.

- Linear regression, Lasso and Ridge regression are very similar to each other in terms of explaining the variability of glucose given the features aaccording to their performance metrics of both R^2 and Root Mean Square Error (RMSE). However, linear regression shows as the best regression with the lowest RMSE (33.5723) and greatest R^2 (64.57%). 2nd degree polynomial regression is the worst due to the negative R^2 and largest RMSE.

6. Summary Key Findings and Insights, which walks your reader through the main drivers of your model and insights from your data derived from your linear regression model.

- Elevated glucose level can lead to many health problems and affect our quality of life if treatment is delayed. Hence, glucose prediction will help healthcare professionals to provide timely diagnosis and treatment due to elevated glucose level to patients. Moreover, it may also help in developing health strategy and raise public awareness about glucose monitoring.

  Among all this four regressions, linear regression model is the best and it is able to explain about 65% of the glucose variability given the diagnostic measures in glucose prediction.

7. Suggestions for next steps in analyzing this data, which may include suggesting revisiting this model adding specific data features to achieve a better explanation or a better prediction.

- Perform feature engineering on numerical BMI - underweight, normal, overweight, obese for example based on healthcare standard to study its impact on glucose
- Assess the impact of outliers and collinearity from this features
- Investigate if any interaction between features on glucose
- Gather more data and different features e.g. family medical history, laboratory, diet or lifestyle covering all other aspects of patients to introduce more information to better study glucose prediction
- To assess again this regression models by incorporating the previous stated points and tuning the hyperparameters given the data on glucose prediction

References

1. https://www.kaggle.com/houcembenmansour/predict-diabetes-based-on-diagnostic-measures
2. https://www.coursera.org/learn/ibm-exploratory-data-analysis-for-machine-learning?specialization=ibm-machine-learning
3. https://www.coursera.org/learn/supervised-machine-learning-regression
4. https://scikit-learn.org/stable/supervised_learning.html#supervised-learning
5. https://matplotlib.org/stable/index.html