1. Brief description of the data set 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.

2. Initial plan for data exploration

   Descriptive statistics and visualization were reported for data exploration.

   - Summary statistics e.g. counts, average, standard deviation, median, minimum and maximum.
   - Histogram and box plot for distribution and outlier visualization
   - Pairwise correlation
   - Contingency tables between categorical features

3. Actions taken for data cleaning and feature engineering

   - There are no missing data or duplicates in the data.
   - Feature engineering and encoding not needed for this EDA and hypothesis testing.

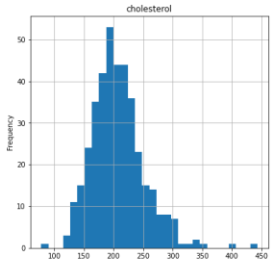
4. Key Findings and Insights, which synthesizes the results of Exploratory Data Analysis in an insightful and actionable manner

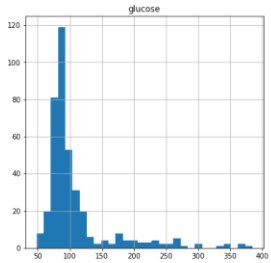
   - There are 390 observation rows from 390 different patients.
   - Following is the last five entries of the data:

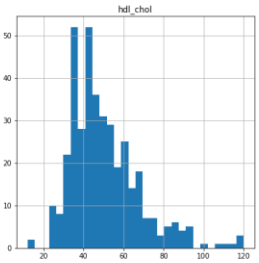
patient_number	cholesterol	glucose	hdl_chol	chol_hdl_ratio	age	gender	height	weight	bmi	systolic_bp	diastolic_bp	waist	hip	waist_hip_ratio	diabetes
386	227	105	44	5.2	83	female	59	125	25.2	150	90	35	40	0.88	No diabetes
387	226	279	52	4.3	84	female	60	192	37.5	144	88	41	48	0.85	Diabetes
388	301	90	118	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

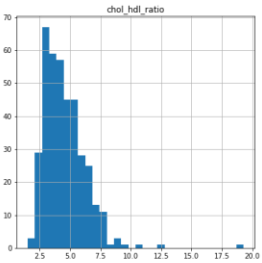
   - There are no duplicates in the data.
   - There are no missing values from the data.
   - Customer number was not considered for EDA because it adds no value to the analysis.

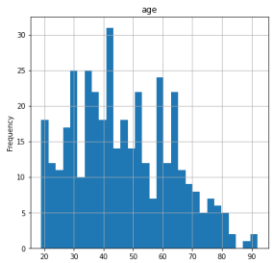
- Following is the histogram of all the numerical features. Glucose looks the most right skewed compared to the rest. There is an obvious spike at 140 for systolic blood pressure.

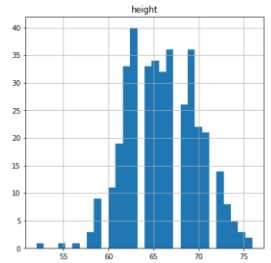


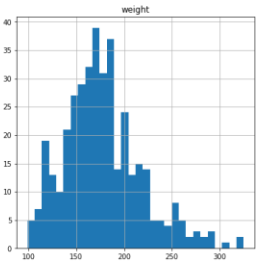


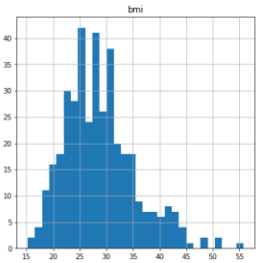


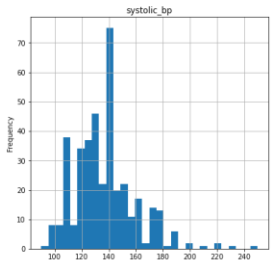


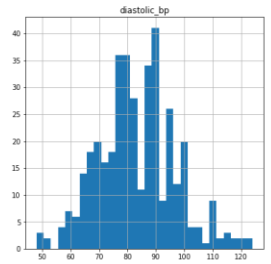


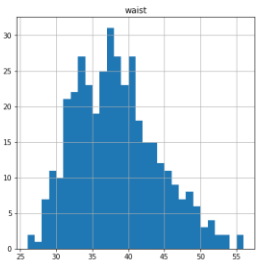


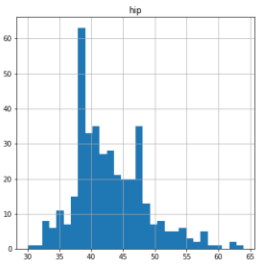


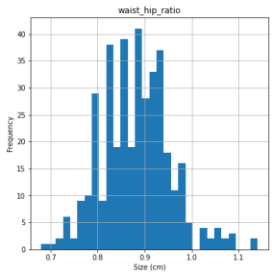




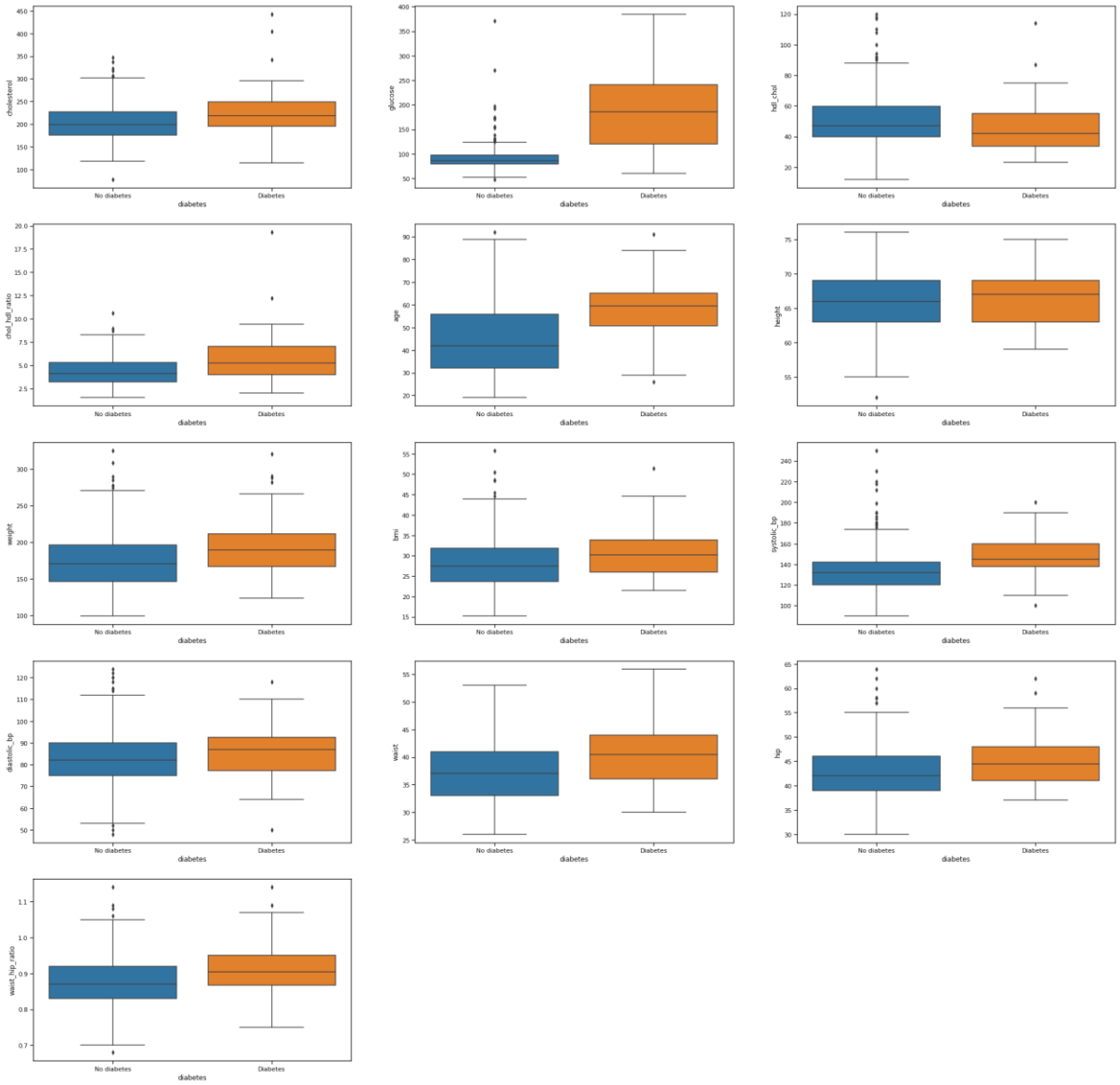




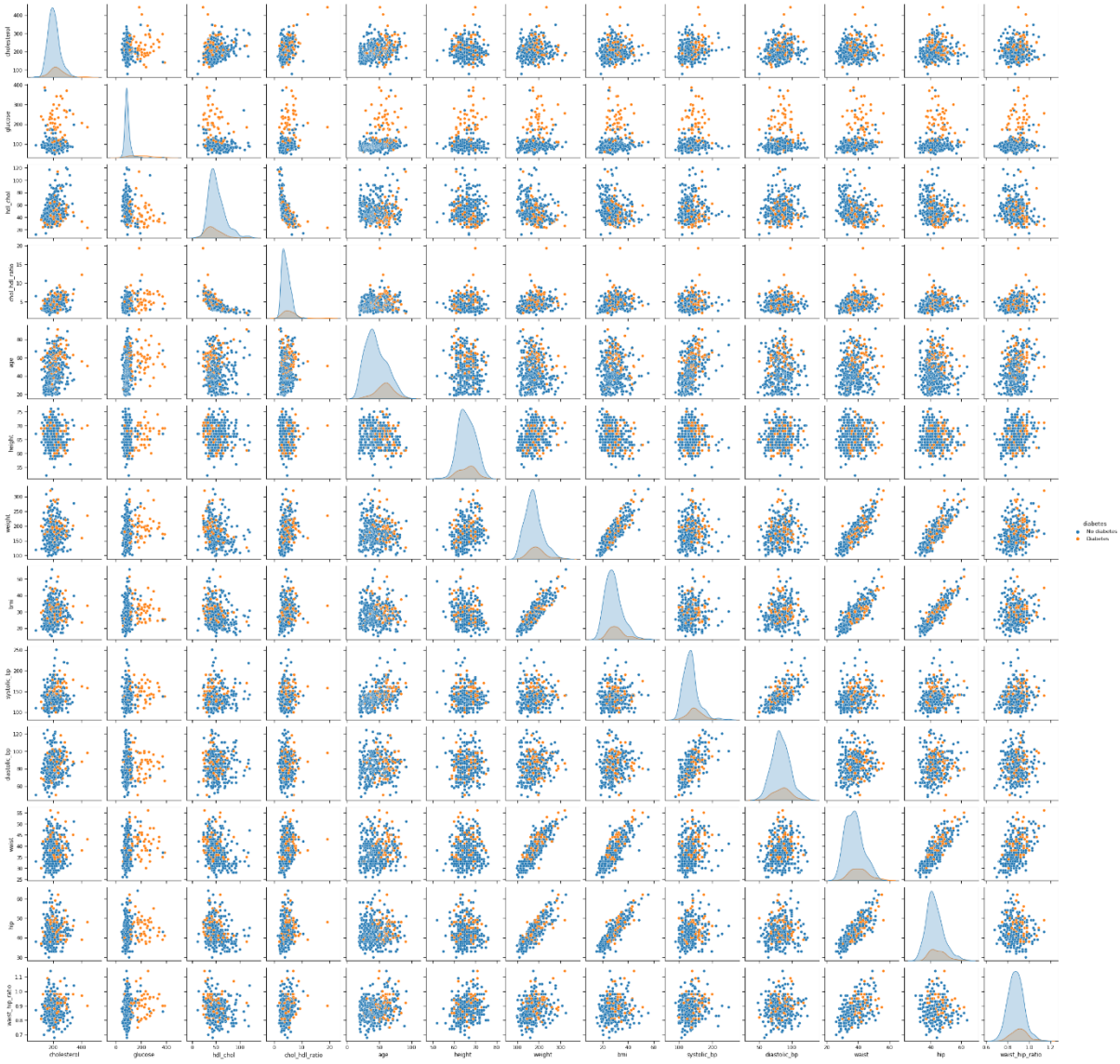




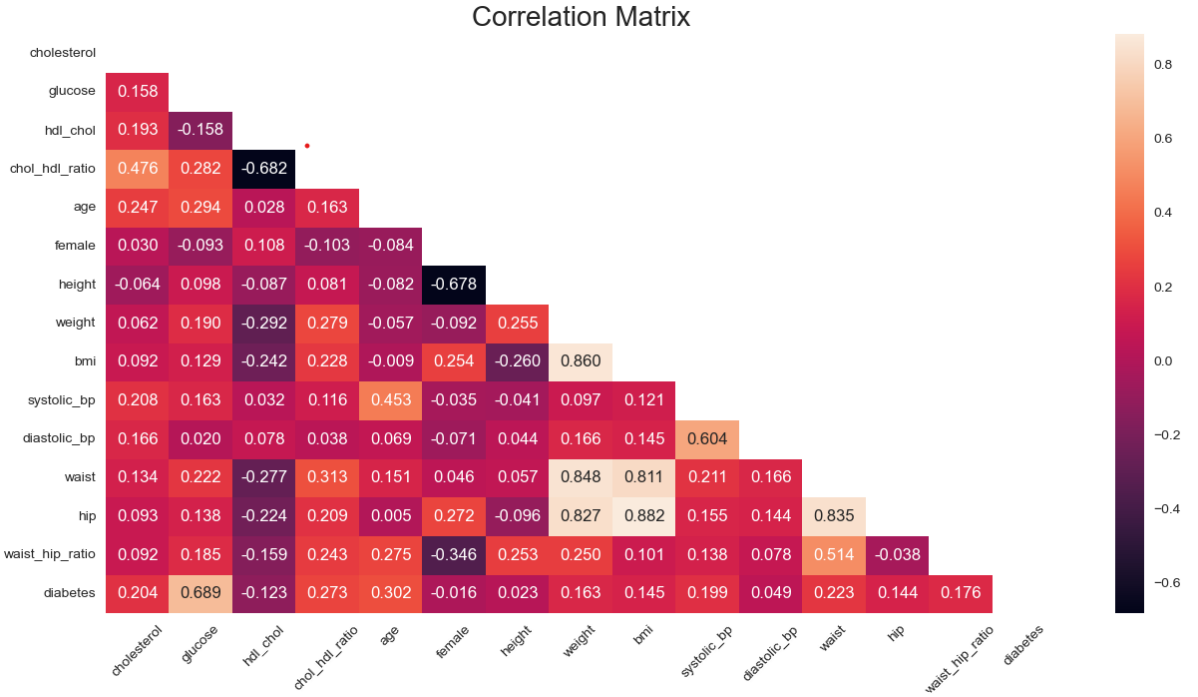
- The following are the boxplots of all numerical features by diabetes groups. The variation of glucose is the greatest for this groups compared to the other features. Outliers are most seen in non-diabetes group than diabetes group for most of this features.



- 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.



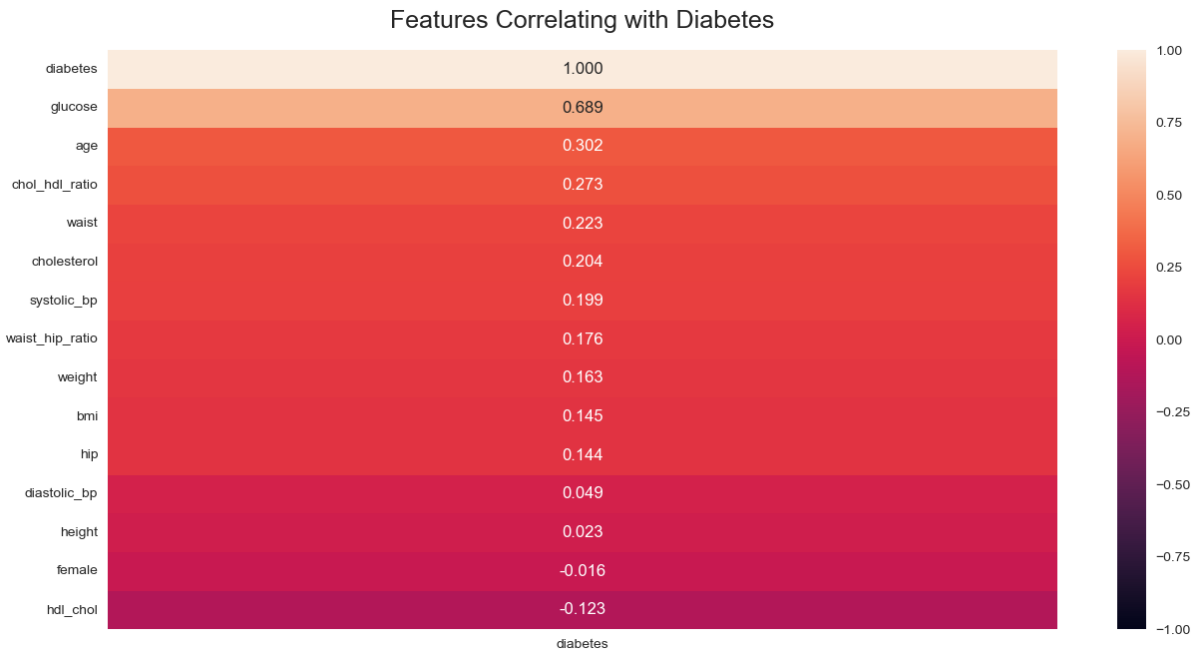
- The following is the correlation matrix of the data:



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 above shows the correlation coefficient of each feature on diabetes in descending order. Glucose is most correlated following by age with diabetes.

- 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

The proportion of either no diabetes or diabetes is only differed slightly between males and females. Overall, we have a larger proportion of females (58.46%) in the data.

	No diabetes		**Diabetes**		**Total**	
	n	%	n	%	n	**%**
Male	136	83.95	26	16.05	162	41.54
Female	194	85.09	34	14.91	228	58.46
Total	330	84.62	60	15.38	390	100

5. Formulating at least 3 hypotheses about this data

   i. Null hypothesis: The mean glucose for diabetes and non-diabetes populations are equal.
   Alternative hypothesis: The mean glucose for diabetes and non-diabetes populations are different.

   ii. Null hypothesis: The mean of BMI for diabetes and non-diabetes populations are equal.
   Alternative hypothesis: The mean of BMI for diabetes is greater than non-diabetes population.

   iii. Null hypothesis: There is no relation between gender and diabetes.
   Alternative hypothesis: There is relation between gender and diabetes.

6. Conducting a formal significance test for one of the hypotheses and discuss the results

   Hypothesis (i),

Feature	Statistics	No diabetes	Diabetes	Total
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

	T	dof	alternative	p-val	CI95%	cohen-d	BF10	power
T-test	10.154	61.605	two-sided	0.0	[82.41, 122.82]	2.629	2.56e+18	1.0

The mean glucose for diabetes group is 194.17. Whereas, it is 91.56 for non-diabetes group. Welch *t*-test from pingouin package was used to test the difference due to unequal variances from the Levene test i.e. p-value<0.05 which clearly depicted in the boxplot on page 3. It shows that the difference of 102.61 in glucose from both groups was significant with 95% confidence that the true means difference lying between 82.41 and 122.82, suggesting that a genuine difference in both the groups. Hence, we can reject the null hypothesis of no difference.

Hypothesis (ii),

Feature	Statistics	No diabetes	Diabetes	Total
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

	T	dof	alternative	p-val	CI95%	cohen-d	BF10	power
T-test	2.892	388	greater	0.002	[1.14, inf]	0.406	15.051	0.893

Variances are equal between both groups according to Levene test i.e. p-value>0.05. 2-sample *t*-test from pingouin package was used to test the difference of both groups without correcting for unequal variances. The p-value of 0.002 of the test is less than the significance level alpha of 0.05. This means that the average BMI for diabetes group is greater than that of non-diabetes group. Hence, we can reject the null hypothesis.

Hypothesis (iii),

```
Observed counts:

diabetes  Diabetes  No diabetes
gender
female          34          194
male            26          136

Expected counts:

diabetes   Diabetes  No diabetes
gender
female    35.076923   192.923077
male      24.923077   137.076923

                 test  lambda   chi2  dof   pval  cramer  power
0             pearson   1.000  0.094  1.0  0.759   0.016  0.061
1       cressie-read    0.667  0.094  1.0  0.759   0.016  0.061
2      log-likelihood   0.000  0.094  1.0  0.759   0.016  0.061
3       freeman-tukey  -0.500  0.094  1.0  0.760   0.015  0.061
4  mod-log-likelihood  -1.000  0.093  1.0  0.760   0.015  0.061
5             neyman   -2.000  0.093  1.0  0.760   0.015  0.061
```

The expected counts are all more than 5 from the 2x2 contingency tables between gender and diabetes. Therefore, Chi-square test was used to test the association between both features. Since the p-values are all >0.05, we cannot reject the null hypothesis. Gender does not qualify as a good predictor for diabetes on this dataset.

7. Suggestions for next steps in analyzing this data

   - Perform feature engineering on numerical BMI - underweight, normal, overweight, obese for example based on healthcare standard to study its impact on diabetes
   - Assess the impact of outliers from this features
   - Collect more data and different features e.g. family medical history, laboratory, diet or lifestyle etc to study diabetes
   - Study the relationship of this features on diabetes prediction by building a classification machine learning model
   - To assess different classification models on which algorithm performs better given the data

8. A paragraph that summarizes the quality of this data set and a request for additional data if needed

   The quality of this data is decent. The data contains neither missing nor duplicate data. Data types are well reflected the features themselves. However, the measurement unit of this features is lacking which makes correct data and result interpretation a challenge to some degree. The data is not large. More data and different features covering all other aspects of the patients to introduce more information to better study diabetes are required.

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://pingouin-stats.org/generated/pingouin.ttest.html
4. https://www.marsja.se/how-to-perform-a-two-sample-t-test-with-python-3-different-methods/
5. https://matplotlib.org/stable/index.html
6. https://support.minitab.com/en-us/minitab/18/Assistant_Two_Sample_t.pdf
7. http://gureckislab.org/courses/fall19/labincp/chapters/11/00-ttest.html
8. https://www.saem.org/about-saem/academies-interest-groups-affiliates2/cdem/for-students/cdem-voice/educational-research-column/educational-research-column-choosing-wisely-chi-square-vs.-fisher-s-exact