Name: Abed Tabbalat

Team Status: Individual

Milestone 2 – Data Selection & Project Proposal

- **What types of model or models do you plan to use and why?**

The plan is to run 3 different models and see which ones fits best for the dataset. The models I am choosing to run and find out which is the best model are:

1. Random Forest
2. Logistic Regression
3. Decision Tree

The reason I chose the models above is because the dataset shows variables for different people with an outcome variable of ones and zeros that determines if the person is diabetic or not. The additional variables that come with it are characteristics of the person that could be triggers on being diabetic. Once the model's accuracy is tested, a confusion matrix will be tested on the models for a better visual.

- **How do you plan to evaluate your results?**

Results will be evaluated by choosing the most accurate model and applying a confusion matrix to the model and then an ROC curve to determine how fit the model will be. Accuracy percentage is the key to have a successful model.

- **What do you hope to learn?**

I recently found out that a close family member has gotten diabetes. I have never paid attention to it as it is a highly common condition people go through. Once I learned that this could be a genetic passthrough, my curiosity has been risen to learn more about it. What environmental factors could cause it? What variables from the data collected highly impacts the outcome of diabetes.

In addition, learning which model best fits certain types of datasets based on what variables it contains is one thing I would want to be excellent with. The ability to look at a dataset and narrowing down which models to use. Of course, this comes with experience and repetition, and this is one of the technical goals that I have.

- **Assess any risks with your proposal.**

The main risk I currently see is the dataset is less than 1,000 rows which can impact the model results. The more data and information we have the better the outcome of the model will be. This risk may not exist depending on how the variables are correlated with each other. In addition, the models I have chosen could result in inaccurate results if overfitting happens which could result in choosing a different approach in predictions.

- **Identify a contingency plan if your original project plan does not work out.**

If the plan does not work out because the analysis shows that there isn’t enough data to support the model results, then the plan will be to find a larger dataset within the same topic. That said, having the same topic and possible similar variables should not impact the choice of the models that will be run. If the results are yet to be inaccurate due to the data not having enough features to predict. Finding a different dataset at that time will be a must and starting over. Even though that will be a stretch, a backup plan must be in place.

- **Include anything else you believe is important.**

The main question regarding diabetes would be, what variables are mostly impacting the outcome of having a positive diabetes diagnosis. To add, this topic or healthcare in general has never been my cup of tea, as my interest has always been in the P&C insurance industry. Sometimes life gets in the way, and we forget the importance of health that keeps us alive and well to be able to continue doing what we do. I take that for granted and there is nothing better than learning about it through datasets and predictions. Choosing diabetes in specific, as mentioned above, having a close family member diagnosed with it, while never paid attention to it except for “can’t have sugar” is not enough anymore. This will help me understand finding out what correlates with the outcome what I should monitor for myself and others to avoid getting it. I believe as common as it is, anyone would be interested in knowing what variables do correlate to be able to monitor themselves and live a better life.

Milestone 3 – Preliminary Analysis

- **Brief explanation about the data**

The data was downloaded as a CSV file from Kaggle, containing 768 rows and 9 columns. Each column is a variable that will be used for the modeling, those variables are the following

1. **Pregnancies:** Amount of time the patient been pregnant
2. **Glucose:** Patient glucose levels
3. **BloodPressure:** Patient blood pressure
4. **SkinThickness:** Patient skin thickness
5. **Insulin:** Patient insulin level
6. **BMI:** Patient Body Mass Index
7. **DiabetesPedigreeFunction:** Indicates the function which scores the likelihood of diabetes based on family history
8. **Age:** Patient age
9. **Outcome:** 1 for positive diabetes, 0 negative diabetes

- **Will I be able to answer the questions I want to answer with the data I have?**

The main question towards the dataset and the project is whether it is possible to predict the outcome based on the variables available. After some data exploration with the data, I believe predicting the outcome with the variables provided is possible with a chance of error. Finding out how accurate the model would be in predicting the outcome is key.

Below shows the data types of the dataset:

```
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

The data is all numeric which means applying a regression model would be best as a quick observation. Going further with data exploration, below shows some exploratory analysis on each variable.

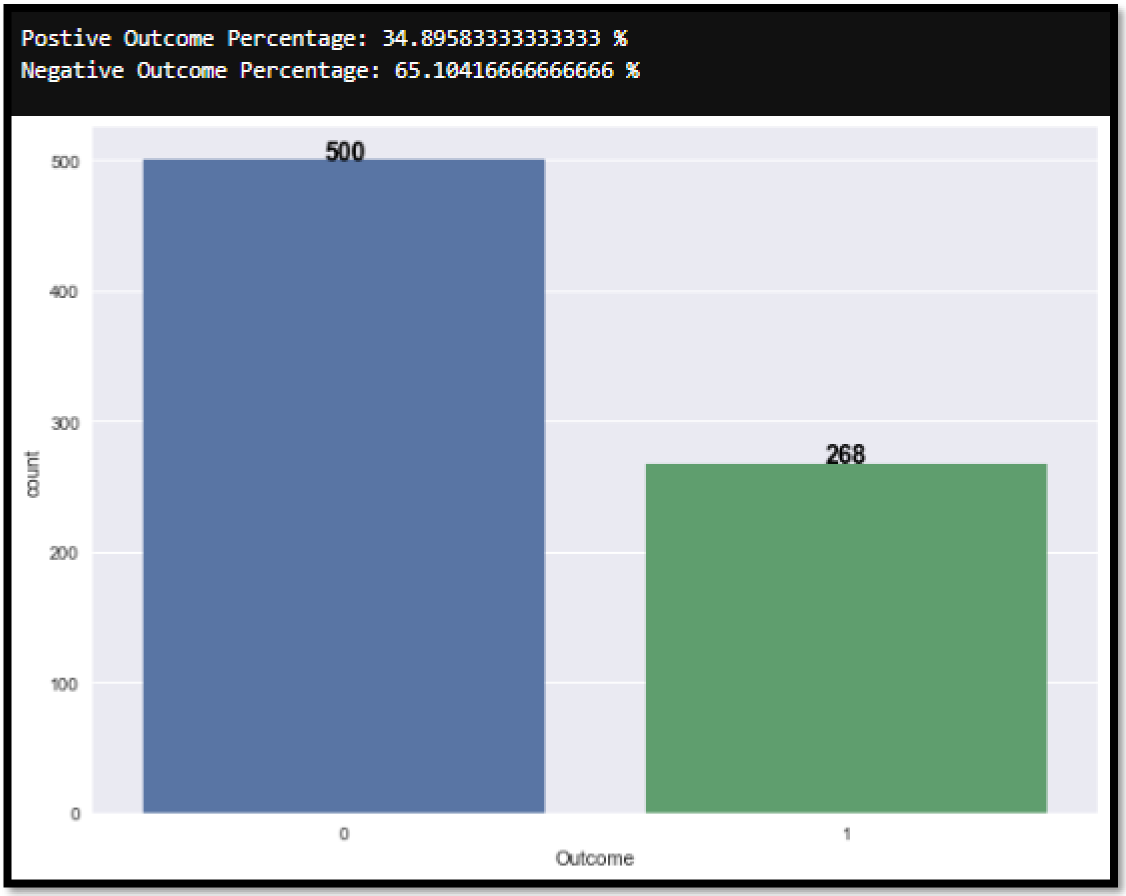
	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.0000	6.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.0000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.0000	80.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	23.0000	32.00000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	30.5000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.0000	36.60000	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	0.3725	0.62625	2.42
Age	768.0	33.240885	11.760232	21.000	24.00000	29.0000	41.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.00000	0.0000	1.00000	1.00

Checks for blanks also applied.

```
------- Blanks count -------
Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

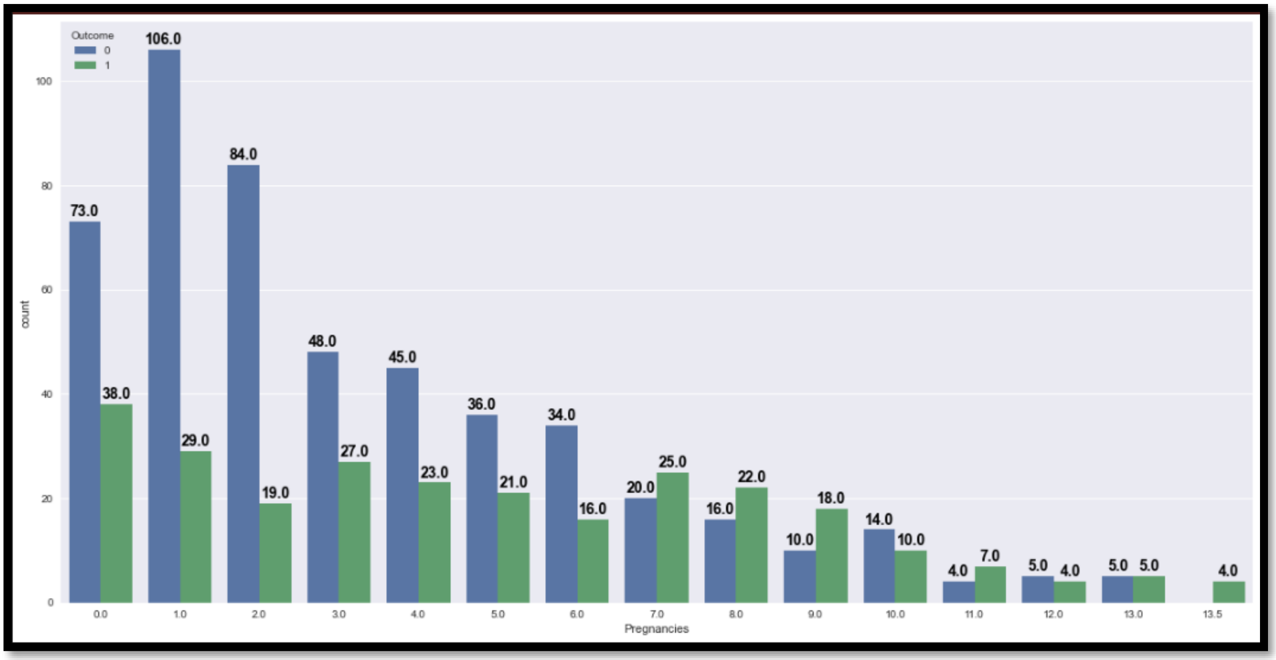
- **What visualizations are especially useful for explaining my data?**

Many visualizations came in handy to understand the data better. The target variable is the Outcome column which determines if the patient is positive with diabetes or not. Therefore, the first exploratory visualization is to compare the split between positive and negative outcomes to diabetes on the dataset.



35% (268 out of 768) of the patients have diabetes and 65% (500 out of 768) as seen in the bar chart above.

If we focus on the pool with positive outcome, we can check the weighted percentages of pregnancies.



	Pregnancies	Outcome	Outcome_Pct
0	0.0	38	14.179104
1	1.0	29	10.820896
2	2.0	19	7.089552
3	3.0	27	10.074627
4	4.0	23	8.582090
5	5.0	21	7.835821
6	6.0	16	5.970149
7	7.0	25	9.328358
8	8.0	22	8.208955
9	9.0	18	6.716418
10	10.0	10	3.731343
11	11.0	7	2.611940
12	12.0	4	1.492537
13	13.0	5	1.865672
14	13.5	4	1.492537

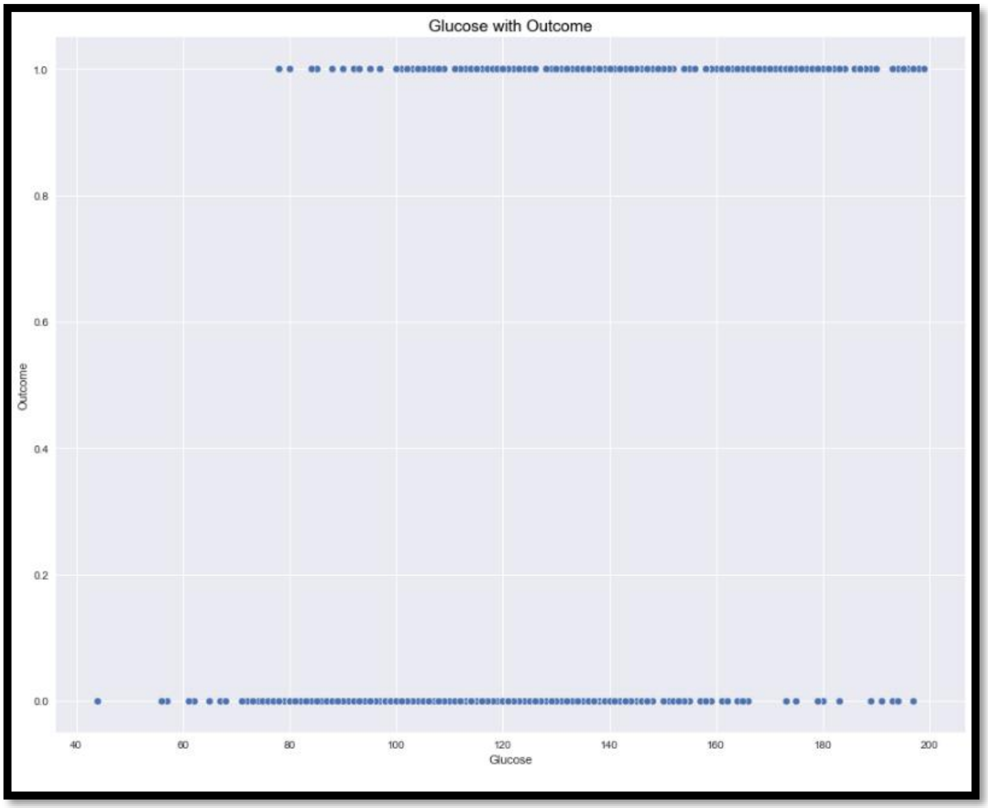
The bar chart shows the outcome for diabetes for each patient's pregnancy count and we can determine that for positive outcome, 14% (38 out of 268) is the highest count of positive outcome with no pregnancies.

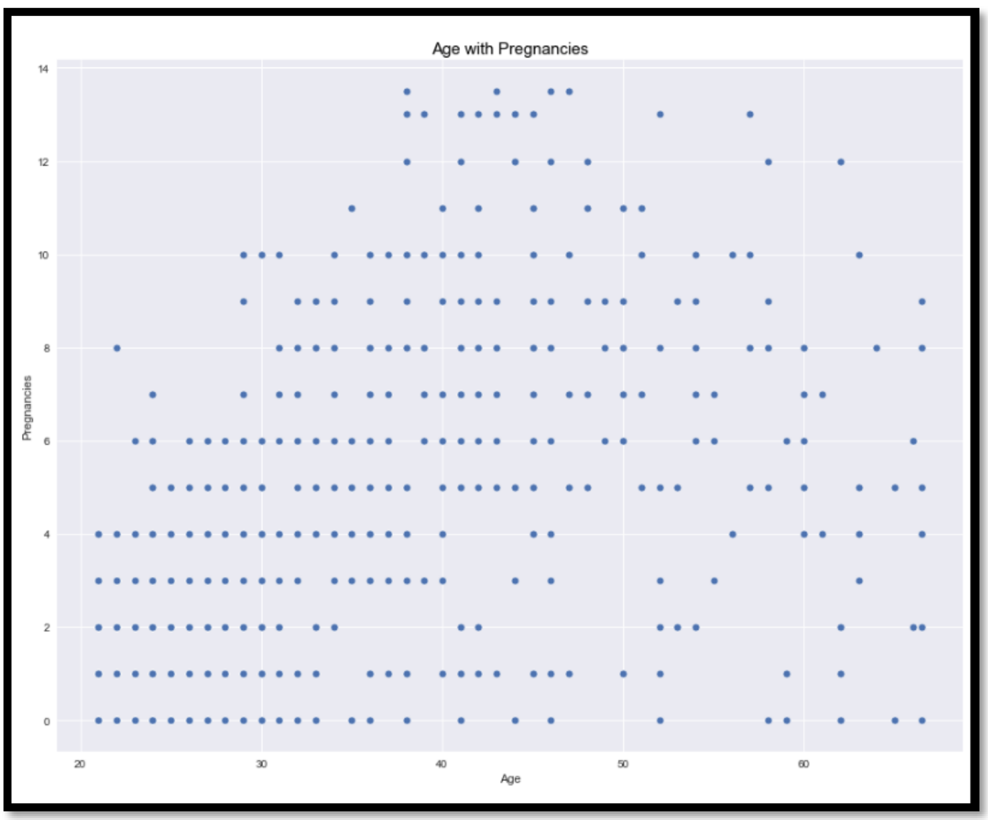
Next step performed was a correlation matrix to determine which variables are highly correlated. This could help with the modeling process and the accuracy of the model.

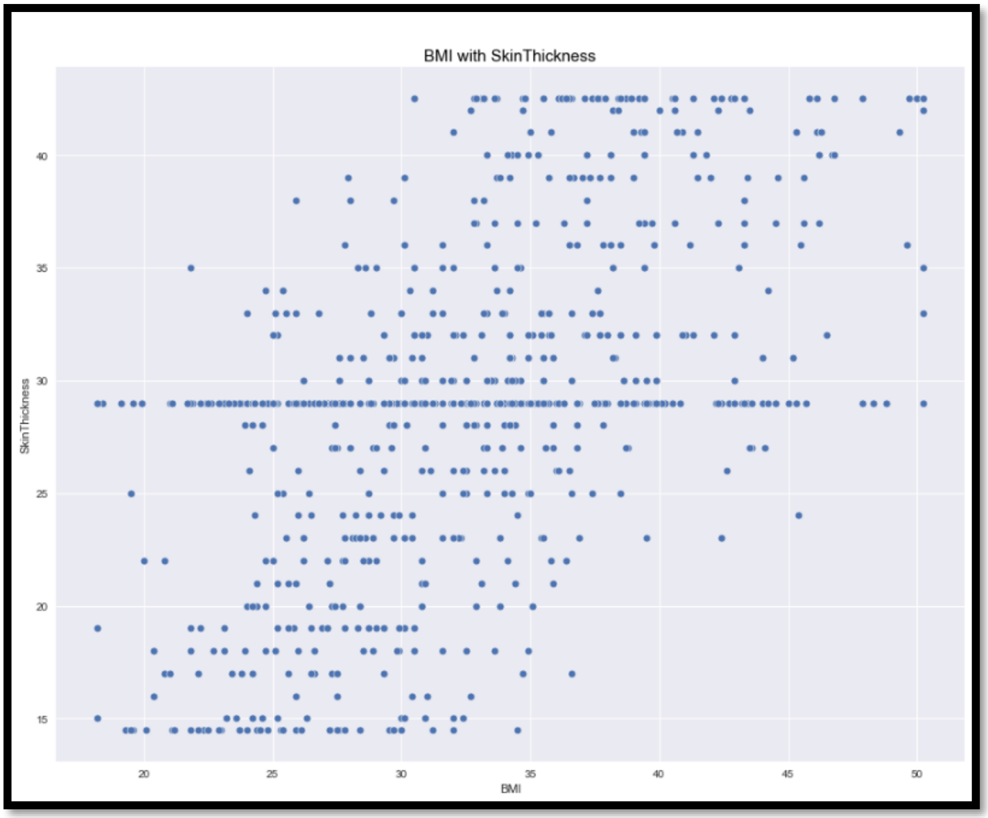
Heatmap shows a high positive correlation between the following:

- Glucose and Outcome
- Pregnancies and Age
- SkinThickness and BMI

Now that we know which variables are highly correlated, we can visualize these relationships with a scatter plot.

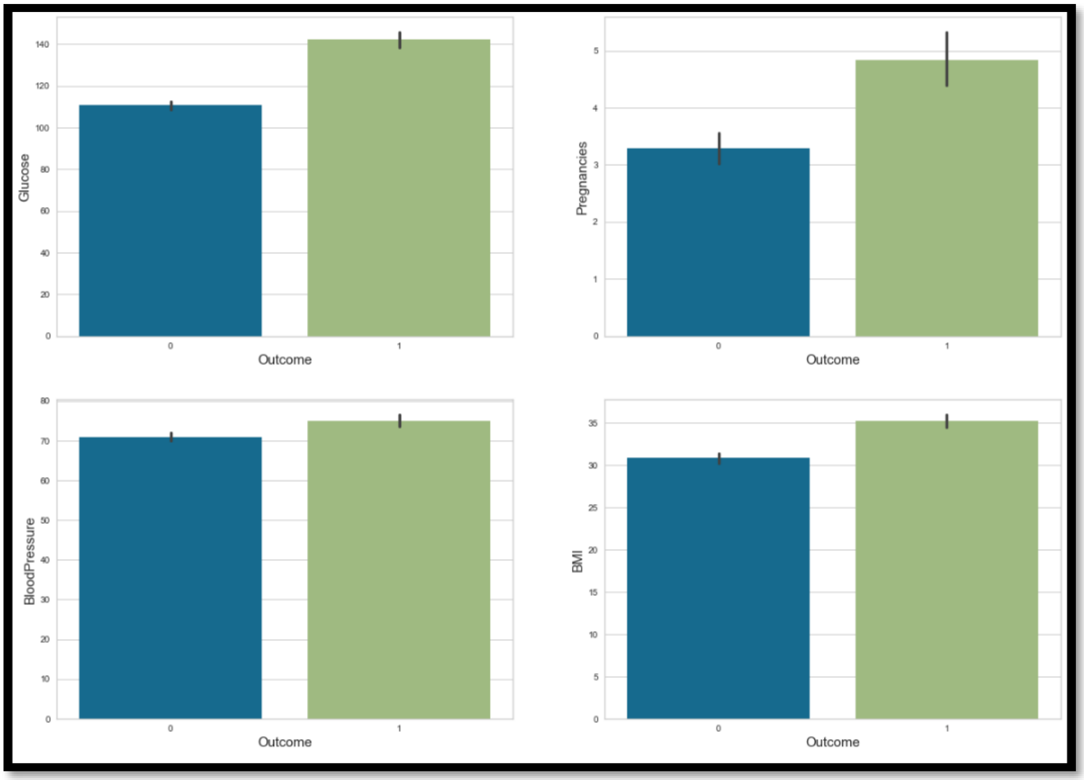

Glucose with Outcome
1.0
0.8
0.6
0.4
0.2
0.0
Outcome
40
60
80
100
120
140
160
180
200
Glucose

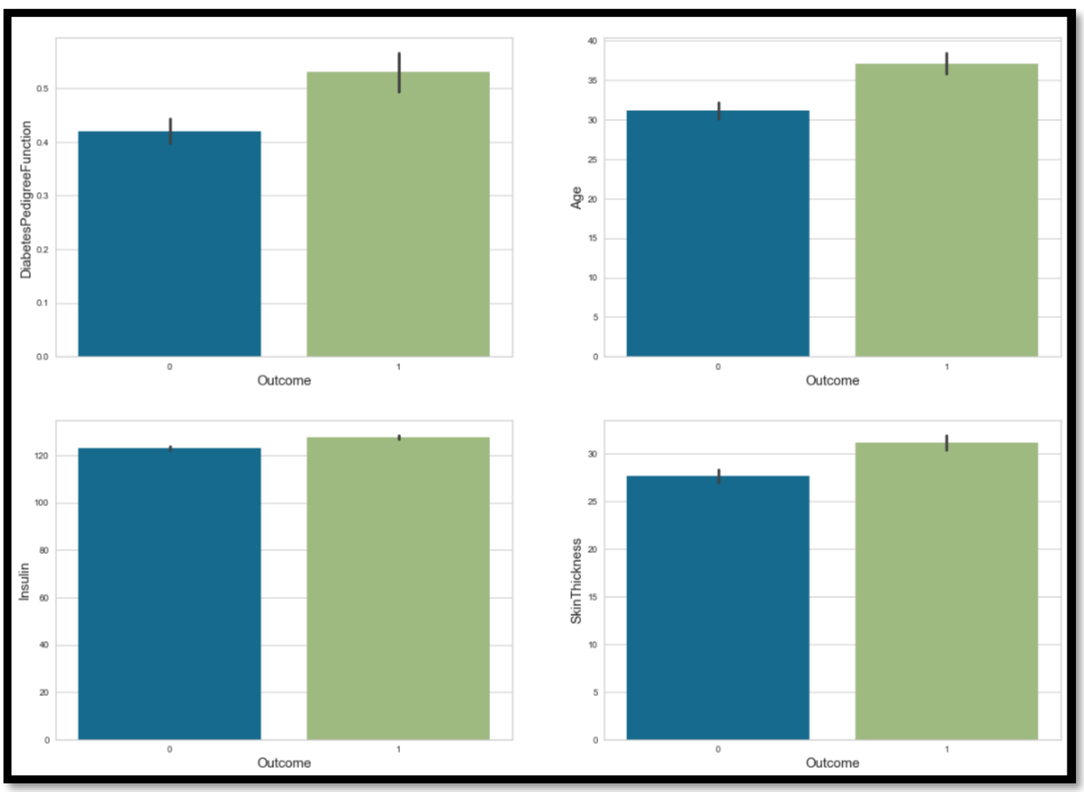

Age with Pregnancies
14
12
10
8
6
4
2
0
Pregnancies
20
30
40
50
60
Age



Based on general knowledge the results make sense. The higher the glucose levels are the more likely a person is diabetic. The older the patient is the more likely the patient has been through pregnancies. The higher the body mass index, skin thickness would more parallel to it.

Final area to visualize would be running bar charts of the target variable against all other variables.


Glucose
Outcome
Pregnancies
Outcome
BloodPressure
Outcome
BMI
Outcome


DiabetesPedigreeFunction
Outcome
Age
Outcome
Insulin
Outcome
SkinThickness
Outcome

The bar charts clearly state that patients with positive outcome have higher readings in all the variables than patients with negative outcome.

- **Do I need to adjust the data and/or driving questions?**

We can determine the following:

- There are `9` columns and `768` rows in the dataset
- All the variables are numeric
- There are no blanks within the dataset
- There are no duplicates within the dataset

The min column in the data description table shows `0` in the variables are unexpected, further investigation required.

The columns that require investigation are:

- Glucose
- BloodPressure
- SkinThickness
- Insulin
- BMI

Once the preliminary analysis has been complete, there was one thing that did not make sense at a first glance which is the **min** value for some of the variables is zero. The variables mentioned in the screenshot above would be impossible for a living human being to have 0 readings, which concludes that zeros have been used as fillers for missing values.

Once the zeros have been identified, they were replaced with **NaNs** to give a realistic look at the data. The last exercise to have the data ready for the rest of the analysis and modeling was determining the skewness of those variables through a histogram to determine if we are replacing the NaN values with mean, or median. The results indicated that all the variables will be replaced with median.

- **Do I need to adjust my model/evaluation choices?**

Until this point, there is still no model of choice, as the first step in the modeling process is to determine which model is the best fit. However, after exploring the data, realizing all the variables are numeric, the best choice would be applying a logistic regression to the model. Regression is recommended to use when prediction is reliant on numeric variables.

- **Are my original expectations still reasonable?**

After analyzing the data, and understanding what each variable contributes, my original expectations remain reasonable. The variables that are highly correlated make sense and would be vital in the modeling process. The original expectations are the variables are positively correlated, and therefore, the higher the readings the more likely for the patient to be positive in diabetes. The EDA performed compliments the original expectations.