1. Brief description of the data, including its origin and quality issues.

Data Origin:

- **Source**: The PimaDiabetes dataset was originally collected by the USA's National Institute of Diabetes and Digestive and Kidney Diseases from the population of the Pima Indian tribe near Phoenix, Arizona[1].
- **Collection Method**: Each community resident over 5 years of age was asked to undergo a standardized examination every two years, which included an oral glucose tolerance test. Diabetes was diagnosed according to World Health Organization Criteria.
- **Variables**: Eight variables were chosen to form the basis of the datasets of diabetes within five years in Pima Indian women, then one additional dataset 'Outcome' that showing the results whether they have diabetes or not.
- **Time Frame**: The original data collection took place from 1965 to 1970. However, the specific time frame for the datasets provided in this assignment is not explicitly stated.

Data Quality:

- **Completeness**: No variables that having missing values.
- **Consistency**: There are no discrepancies in how the Pima Diabetes dataset was recorded.
- **Relevance**: All of the eight variables are having a high relevancy to the diabetes outcome. However, the objective of this study is to determine which variables being the most important causes to the diabetes cases.
- **Accuracy**: All of the variables is valid as their data types is already aligned to what the data is intended to, however the Glucose, BloodPressure, SkinThickness, Insulin, and BMI showed lot of 0 values (which not possible for that kind of data). This issue will be handled further.

2. Exploratory Data Analysis

Head of the data:

- Looking at the first five rows (head) to get a sense of what the data looks like.

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigree	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1

Figure 1. The first five rows of data in PimaDiabetes.csv

Dataset Info:

- Observing further the name of columns, count of non-missing rows, data types, and printing the length of datasets row.

```
Data columns (total 9 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Pregnancies       750 non-null    int64
 1   Glucose           750 non-null    int64
 2   BloodPressure     750 non-null    int64
 3   SkinThickness     750 non-null    int64
 4   Insulin           750 non-null    int64
 5   BMI               750 non-null    float64
 6   DiabetesPedigree  750 non-null    float64
 7   Age               750 non-null    int64
 8   Outcome           750 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 52.9 KB
Length of row is : 750
```

Figure 2. The information of column name, non-missing count, data type, and rows length from PimaDiabetes.csv

Summary Statistics:

- Using the describe function helps us see the statistics for each variable. This lets us notice if there are any outliers or if certain variables have a value of 0, like Glucose, which is not possible for a human. This gives us an idea of how the data is behaving and what kind of treatment we may need to make.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigree	Age
count	750.000000	750.000000	750.000000	750.000000	750.000000	750.000000	750.000000	750.000000
mean	3.844000	120.737333	68.982667	20.489333	80.378667	31.959067	0.473544	33.166667
std	3.370085	32.019671	19.508814	15.918828	115.019198	7.927399	0.332119	11.708872
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.244000	24.000000
50%	3.000000	117.000000	72.000000	23.000000	36.500000	32.000000	0.377000	29.000000
75%	6.000000	140.750000	80.000000	32.000000	129.750000	36.575000	0.628500	40.750000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000

Figure 3. Summary statistics of variables from PimaDiabetes.csv

Zero/Error Values Handling:

- The Glucose, BloodPressure, SkinThickness, Insulin, and BMI columns have some 0 values, likely due to errors. We need to fix this by imputing or removing data. We removed (*not dropping*) data points with zero values for Insulin and SkinThickness, and filled in the missing values for Glucose, BloodPressure, and BMI with averages. This led to the following statistics summary.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigree	Age
count	388.000000	388.000000	388.000000	388.000000	388.000000	388.000000	388.000000	388.000000
mean	3.288660	122.604651	70.561856	28.992268	155.371134	33.052196	0.526655	30.760309
std	3.205966	30.807249	12.478562	10.484020	117.993415	7.028717	0.350231	10.127700
min	0.000000	56.000000	24.000000	7.000000	14.000000	18.200000	0.085000	21.000000
25%	1.000000	99.000000	62.000000	21.000000	76.750000	28.375000	0.279750	23.000000
50%	2.000000	119.000000	70.000000	29.000000	125.500000	33.200000	0.452000	27.000000
75%	5.000000	143.000000	78.000000	36.000000	190.000000	37.025000	0.687000	36.000000
max	17.000000	198.000000	110.000000	63.000000	846.000000	67.100000	2.420000	81.000000

Figure 4. Summary statistics after data imputation and removal

Univariate Data Analysis:

- Combining Rug plot, Histogram, KDE, Mean, and Median in a single plot gives a quick overview of data density, distribution, and central tendencies. For instance, Glucose's KDE shows a balanced, unimodal distribution with high density around its mean and median.

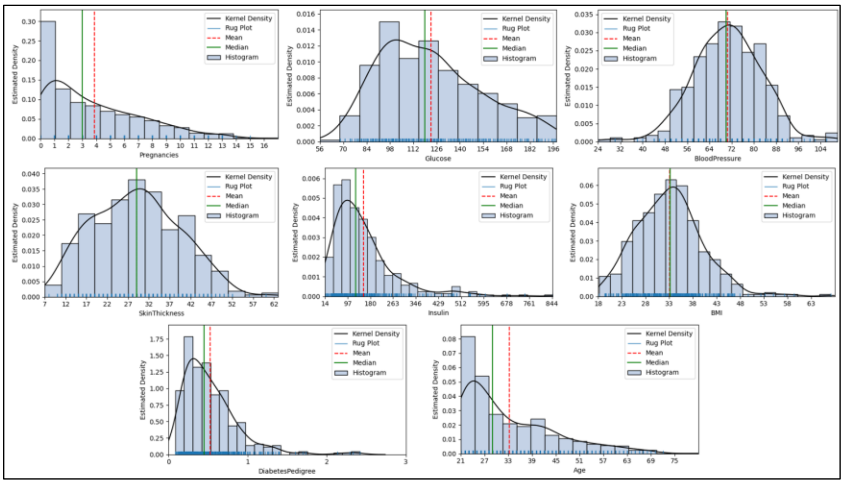

Figure 5. Combined plots of Rug plot (lossless), Histogram (lossy), Kernel Density Estimate (KDE, lossy), and vertical line of Mean and Median of each PimaDiabetes.csv (except Outcome) variables.

- Then comparing the variables value with ECDF (multiplies by 100%) that can be used to observe the distribution of data points of each variable from the lowest to the highest again their percentiles.

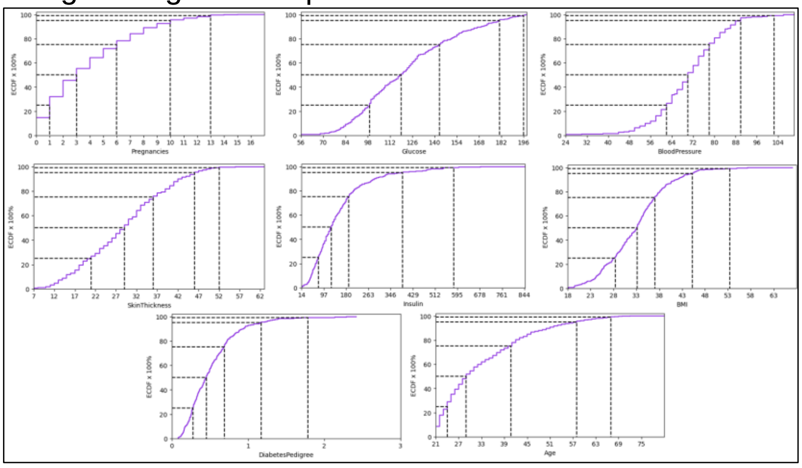

Figure 6. Combined plots of quantiles and ECDF multiplied by 100% of each PimaDiabetes.csv (except Outcome) variables.

- Additionally, comparing Gaussian and uniform kernels in univariate analysis helps us understand how they capture data distribution. When the lines are closely aligned, it indicates that the choice of kernels has little impact on the data within that range. This observation applies to all PimaDiabetes variables after data imputation and removal, as shown below.

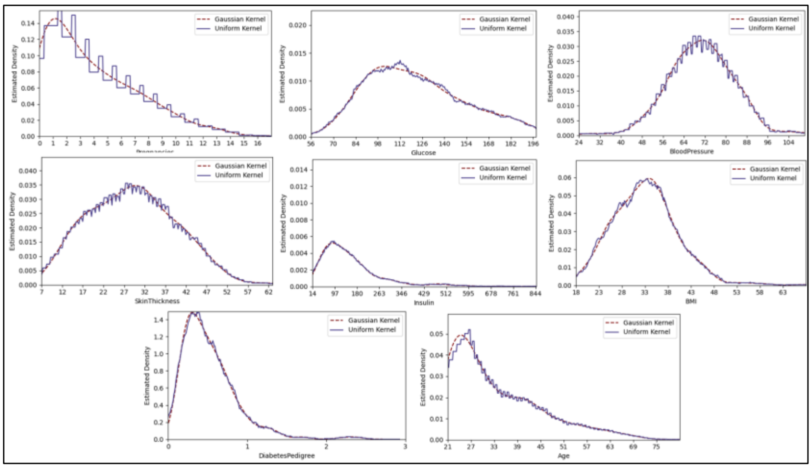

Figure 7. Combined density estimates using the uniform and gaussian kernels of each PimaDiabetes.csv (except Outcome) variables.

Multivariate Data Analysis:

- Start with the straightforward pair-plot to observe the relationship between each pair of variables.

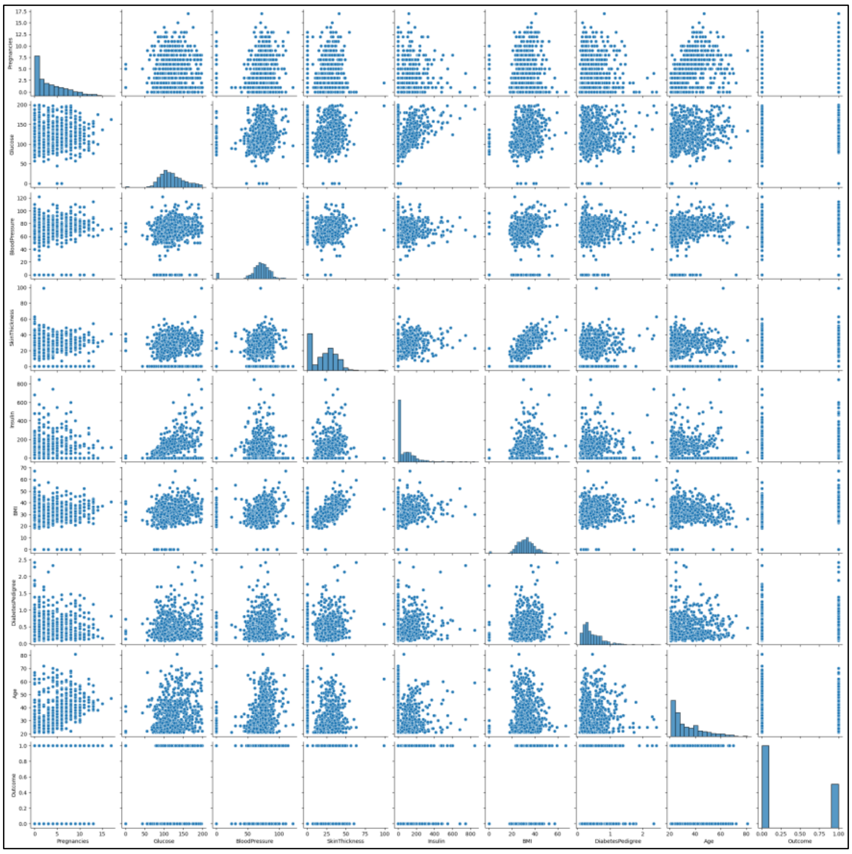

Figure 8. Pair plot between each variable of PimaDiabetes.csv

- Then, linear model between each variable is conducted to determine which variables can be used and most optimum as a predictor of diabetes 'Outcome' in the later phase. Use the python library called smf (statsmodel.formula.api) and apply to each 8 variables towards the 'Outcome'.

```
Summary for variable Pregnancies:
                            OLS Regression Results
==============================================================================
Dep. Variable:                Outcome   R-squared:                       0.053
Model:                            OLS   Adj. R-squared:                  0.051
Method:                 Least Squares   F-statistic:                     41.49
Date:                Fri, 03 Nov 2023   Prob (F-statistic):           2.12e-10
Time:                        09:59:19   Log-Likelihood:                -487.06
No. Observations:                 750   AIC:                             978.1
Df Residuals:                     748   BIC:                             987.4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.2221      0.026      8.643      0.000       0.172       0.273
Pregnancies    0.0324      0.005      6.441      0.000       0.023       0.042
==============================================================================
Omnibus:                   101072.938   Durbin-Watson:                   1.985
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              106.995
Skew:                           0.618   Prob(JB):                     5.84e-24
Kurtosis:                       1.623   Cond. No.                         7.93
==============================================================================
```

```
Summary for variable Glucose:
                            OLS Regression Results
==============================================================================
Dep. Variable:                Outcome   R-squared:                       0.212
Model:                            OLS   Adj. R-squared:                  0.211
Method:                 Least Squares   F-statistic:                     201.1
Date:                Fri, 03 Nov 2023   Prob (F-statistic):           1.33e-40
Time:                        09:59:19   Log-Likelihood:                -418.01
No. Observations:                 750   AIC:                             840.0
Df Residuals:                     748   BIC:                             849.3
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.4799      0.060     -7.959      0.000      -0.598      -0.362
Glucose        0.0068      0.000     14.181      0.000       0.006       0.008
==============================================================================
Omnibus:                       64.756   Durbin-Watson:                   1.988
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               54.437
Skew:                           0.578   Prob(JB):                     1.51e-12
Kurtosis:                       2.362   Cond. No.                         488.
==============================================================================
```

```
Summary for variable BloodPressure:
                            OLS Regression Results
==============================================================================
Dep. Variable:                Outcome   R-squared:                       0.004
Model:                            OLS   Adj. R-squared:                  0.002
Method:                 Least Squares   F-statistic:                     2.781
Date:                Fri, 03 Nov 2023   Prob (F-statistic):             0.0958
Time:                        09:59:19   Log-Likelihood:                -505.92
No. Observations:                 750   AIC:                             1016.
Df Residuals:                     748   BIC:                             1025.
Df Model:                           1
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
Intercept         0.2442      0.064      3.824      0.000       0.119       0.370
BloodPressure     0.0015      0.001      1.668      0.096      -0.000       0.003
==============================================================================
Omnibus:                     4741.482   Durbin-Watson:                   1.957
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              128.734
Skew:                           0.643   Prob(JB):                     1.11e-28
Kurtosis:                       1.430   Cond. No.                         264.
==============================================================================
```

```
Summary for variable SkinThickness:
                            OLS Regression Results
==============================================================================
Dep. Variable:                Outcome   R-squared:                       0.007
Model:                            OLS   Adj. R-squared:                  0.005
Method:                 Least Squares   F-statistic:                     5.089
Date:                Fri, 03 Nov 2023   Prob (F-statistic):             0.0244
Time:                        09:59:20   Log-Likelihood:                -504.76
No. Observations:                 750   AIC:                             1014.
Df Residuals:                     748   BIC:                             1023.
Df Model:                           1
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
Intercept         0.2963      0.028     10.477      0.000       0.241       0.352
SkinThickness     0.0025      0.001      2.256      0.024       0.000       0.005
==============================================================================
Omnibus:                     4861.300   Durbin-Watson:                   1.978
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              127.370
Skew:                           0.641   Prob(JB):                     2.20e-28
Kurtosis:                       1.441   Cond. No.                         42.3
==============================================================================
```

```
Summary for variable Insulin:
                            OLS Regression Results
==============================================================================
Dep. Variable:                Outcome   R-squared:                       0.017
Model:                            OLS   Adj. R-squared:                  0.016
Method:                 Least Squares   F-statistic:                     13.05
Date:                Fri, 03 Nov 2023   Prob (F-statistic):           0.000324
Time:                        09:59:20   Log-Likelihood:                -500.82
No. Observations:                 750   AIC:                             1006.
Df Residuals:                     748   BIC:                             1015.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.3031      0.021     14.398      0.000       0.262       0.344
Insulin        0.0005      0.000      3.612      0.000       0.000       0.001
==============================================================================
Omnibus:                     5516.829   Durbin-Watson:                   1.995
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              122.795
Skew:                           0.640   Prob(JB):                     2.16e-27
Kurtosis:                       1.486   Cond. No.                         171.
==============================================================================
```

```
Summary for variable BMI:
                            OLS Regression Results
==============================================================================
Dep. Variable:                Outcome   R-squared:                       0.084
Model:                            OLS   Adj. R-squared:                  0.083
Method:                 Least Squares   F-statistic:                     68.60
Date:                Fri, 03 Nov 2023   Prob (F-statistic):           5.56e-16
Time:                        09:59:20   Log-Likelihood:                -474.40
No. Observations:                 750   AIC:                             952.8
Df Residuals:                     748   BIC:                             962.0
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.2098      0.069     -3.031      0.003      -0.346      -0.074
BMI            0.0174      0.002      8.282      0.000       0.013       0.022
==============================================================================
Omnibus:                     4316.642   Durbin-Watson:                   1.947
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               96.791
Skew:                           0.564   Prob(JB):                     9.59e-22
Kurtosis:                       1.648   Cond. No.                         137.
==============================================================================
```

```
Summary for variable DiabetesPedigree:
                            OLS Regression Results
==============================================================================
Dep. Variable:                Outcome   R-squared:                       0.029
Model:                            OLS   Adj. R-squared:                  0.028
Method:                 Least Squares   F-statistic:                     22.45
Date:                Fri, 03 Nov 2023   Prob (F-statistic):           2.59e-06
Time:                        09:59:20   Log-Likelihood:                -496.22
No. Observations:                 750   AIC:                             996.4
Df Residuals:                     748   BIC:                             1006.
Df Model:                           1
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
Intercept            0.2308      0.030      7.725      0.000       0.172       0.289
DiabetesPedigree     0.2447      0.052      4.738      0.000       0.143       0.346
==============================================================================
Omnibus:                     6429.433   Durbin-Watson:                   1.964
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              116.899
Skew:                           0.626   Prob(JB):                     4.13e-26
Kurtosis:                       1.525   Cond. No.                         3.75
==============================================================================
```

```
Summary for variable Age:
                            OLS Regression Results
==============================================================================
Dep. Variable:                Outcome   R-squared:                       0.054
Model:                            OLS   Adj. R-squared:                  0.053
Method:                 Least Squares   F-statistic:                     42.90
Date:                Fri, 03 Nov 2023   Prob (F-statistic):           1.07e-10
Time:                        09:59:20   Log-Likelihood:                -486.39
No. Observations:                 750   AIC:                             976.8
Df Residuals:                     748   BIC:                             986.0
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.0325      0.051      0.639      0.523      -0.067       0.132
Age            0.0095      0.001      6.550      0.000       0.007       0.012
==============================================================================
Omnibus:                    13203.737   Durbin-Watson:                   1.961
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              101.253
Skew:                           0.566   Prob(JB):                     1.03e-22
Kurtosis:                       1.600   Cond. No.                         106.
==============================================================================
```

Figure 9. Results of linear model using all eight variables as independent variable and ‘Outcome’ as dependent variable.

- Using the linear model mentioned earlier, some variables have significant coefficient estimates. However, the low r2 value suggests a weak linear relationship with the 'Outcome' variable. It's important to also examine the correlation of all variables with 'Outcome'. Here are the results and visualization:

```
Correlation of Glucose with Outcome: 0.46030993500130307
Correlation of BMI with Outcome: 0.289831696615122
Correlation of Age with Outcome: 0.23289168318538497
Correlation of Pregnancies with Outcome: 0.2292346741958749
Correlation of DiabetesPedigree with Outcome: 0.17068833814035475
Correlation of Insulin with Outcome: 0.13092845050409388
Correlation of SkinThickness with Outcome: 0.08220532372704004
Correlation of BloodPressure with Outcome: 0.060860342401823815
```

Figure 10. Eight variables from PimaDiabetes.csv correlations with ‘Outcome’

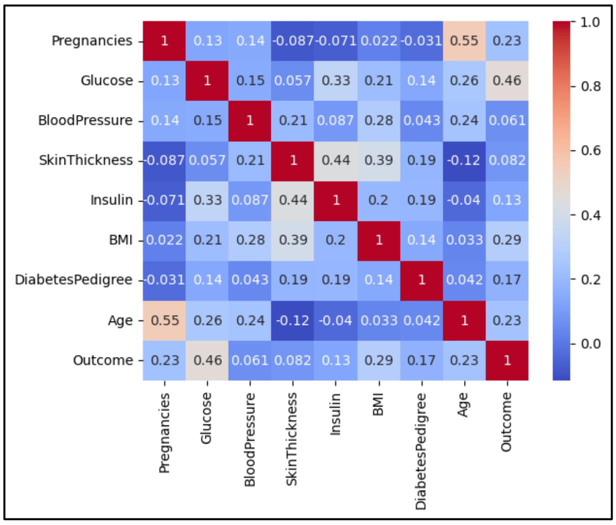

Figure 11. Visualization of correlations between each variable to 'Outcome'

3. Adding 'SevenOrMorePregnancies' column, fit the regression model and answer the probabilities of two conditions.

Adding 'SevenOrMorePregnancies' column:

- Adding 'SevenOrMorePregnancies' column and answer whether they have 7 pregnancies or not (coded '1' for 7 pregnancies or more, '0' otherwise).

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigree	Age	Outcome	SevenOrMorePregnancies
6	148	72	35	0	33.6	0.627	50	1	0
1	85	66	29	0	26.6	0.351	31	0	0
8	183	64	0	0	23.3	0.672	32	1	1
1	89	66	23	94	28.1	0.167	21	0	0
0	137	40	35	168	43.1	2.288	33	1	0

Figure 12. The first five rows of dataframe after added 'SevenOrMorePregnancies' column

Deciding model to be used:

- Since both 'Outcome' and 'SevenOrMorePregnancies' are categorical with binary values (either 1 or 0), using a classification regression type is appropriate. In this case, we'll use a Logistic Regression model. Here is a summary of the statistics for fitting 'SevenOrMorePregnancies' to 'Outcome'.

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                Outcome   No. Observations:                  750
Model:                          Logit   Df Residuals:                      748
Method:                           MLE   Df Model:                            1
Date:                Thu, 02 Nov 2023   Pseudo R-squ.:                 0.04469
Time:                        16:31:15   Log-Likelihood:                -462.39
converged:                       True   LL-Null:                       -484.02
Covariance Type:            nonrobust   LLR p-value:                 4.791e-11
==========================================================================================
                             coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------
Intercept                 -0.9199      0.092    -10.052      0.000      -1.099      -0.741
SevenOrMorePregnancies     1.1898      0.182      6.529      0.000       0.833       1.547
==========================================================================================
```

Figure 13. Summary statistics of logistic regression model fitting of 'SevenOrMorePregnancies' to the 'Outcome'

Probability gets diabetes given have six or fewer children:

- Create a dataframe subset with only women who have been pregnant <=6 times. Then, further subset using the 'SevenOrMorePregnancies' column. Finally, apply the previously created model to this subset dataframe, resulting in a 28.5% probability of getting diabetes given six or fewer children.

```
# Calculate the probabilities of getting diabetes given six or fewer pregnancies
df_six_or_fewer = df_original[df_original['Pregnancies'] <= 6]
X_six_or_fewer = df_six_or_fewer[['SevenOrMorePregnancies']]
prob_six_or_fewer = model.predict(X_six_or_fewer)
print(f'Probability of getting diabetes given six or fewer pregnancies: {prob_six_or_fewer.mean()}')

Probability of getting diabetes given six or fewer pregnancies: 0.2849829351535837
```

Probability gets diabetes given have seven or more children:

- Create a dataframe subset where only consist of women who pregnant >=7. Then, subset again using the ‘SevenOrMorePregnancies’ column, finally fit the model that created before to the subset dataframe, resulting the probability of getting diabetes given seven or more children is 56.7%.

```
# Calculate the probabilities of getting diabetes given seven or more pregnancies
df_seven_or_more = df_original[df_original['Pregnancies'] >= 7]
X_seven_or_more = df_seven_or_more[['SevenOrMorePregnancies']]
prob_seven_or_more = model.predict(X_seven_or_more)
print(f'Probability of getting diabetes given seven or more pregnancies: {prob_seven_or_more.mean()}')
✓ 0.0s

Probability of getting diabetes given seven or more pregnancies: 0.5670731707317076
```

4. Using the data in PimaDiabetes.csv, fit the chosen models to ‘ToPredict.csv’

Model Comparison, Evaluation, and Selection:

- During the exploratory data analysis (EDA) phase, 'Glucose' was observed to have the highest correlation with 'Outcome', followed by 'BMI'. Additionally, Pearson’s correlation indicated that 'Glucose' also has a high correlation with 'Insulin' and 'Age'. Therefore, three models will be tested and evaluated before choosing the final one. The first model uses ['Glucose', 'BMI'] as independent variables, the second uses ['Glucose', 'Insulin'], and the third uses ['Glucose', 'Age']. The performance metrics for each model are as follows:

```
[‘Glucose’, ‘BMI’]
Accuracy Model-1        : 0.7866666666666666
Recall Model-1          : 0.6521739130434783
Precision Model-1       : 0.6521739130434783
F1 Score Model-1        : 0.6521739130434783

[‘Glucose’, ‘Insulin’]
Accuracy Model-2        : 0.7333333333333333
Recall Model-2          : 0.4782608695652174
Precision Model-2       : 0.5789473684210527
F1 Score Model-2        : 0.5238095238095238

[‘Glucose’, ‘Age’]
Accuracy Model-3        : 0.76
Recall Model-3          : 0.4782608695652174
Precision Model-3       : 0.6470588235294118
F1 Score Model-3        : 0.55
```

Figure 14. Performance evaluation of (Model-1) [‘Glucose’, ‘BMI’], (Model-2) [‘Glucose’, ‘Insulin’], and (Model-3) [‘Glucose’, ‘Age’]

- Model-1 with ['Glucose', 'BMI'] is chosen because all the performance metrics evaluations indicate the highest values. This means that Model-1 performed better by correctly predicting a large number of outcomes and achieving the highest precision.

Applying Model-1 to predict the Outcome of 'ToPredict.csv' dataset:

- Applying the Model-1 to ToPredict dataset will be resulting the Outcome value of its datasets, below is the the first five row (the total rows of datasets is also only five rows) after the model was applied:

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigree	Age	Outcome
4	136	70	0	0	31.2	1.182	22	0
1	121	78	39	74	39.0	0.261	28	0
3	108	62	24	0	26.0	0.223	25	0
0	181	88	44	510	43.3	0.222	26	1
8	154	78	32	0	32.4	0.443	45	1

Figure 15. ToPredict dataset after the model was applied with the result is 'Outcome'

Repeat the process as on Pima data by adding 'SevenOrMorePregnancies':

- Adding 'SevenOrMorePregnancies' column and answer whether they have 7 pregnancies or not (coded '1' for 7 pregnancies or more, '0' otherwise).

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigree	Age	Outcome	SevenOrMorePregnancies
4	136	70	0	0	31.2	1.182	22	0	0
1	121	78	39	74	39.0	0.261	28	0	0
3	108	62	24	0	26.0	0.223	25	0	0
0	181	88	44	510	43.3	0.222	26	1	0
8	154	78	32	0	32.4	0.443	45	1	1

Figure 16. The first five rows of ToPredict dataset after added 'SevenOrMorePregnancies' column whether they have 7 pregnancies or not (coded '1' for 7 pregnancies or more, '0' otherwise)

Probability gets diabetes given have six or fewer children:

- Applying the same concept on PimaDiabetes, resulting the probability of getting diabetes given six or fewer children using ToPredict dataset is 24.9%.

```
# Calculate the probabilities of getting diabetes given six or fewer children
df_six_or_fewer = df_to_predict[df_to_predict['Pregnancies'] <= 6]
X_six_or_fewer = df_six_or_fewer[['SevenOrMorePregnancies']]
prob_six_or_fewer = model4.predict(X_six_or_fewer)
print(f'Probability of getting diabetes given six or fewer pregnancies: {prob_six_or_fewer.mean()}')
✓ 0.0s
Probability of getting diabetes given six or fewer pregnancies: 0.24999999979479076
```

Probability gets diabetes given have seven or more children:

- Applying the same concept on PimaDiabetes, resulting the probability of getting diabetes given seven or more children using ToPredict dataset is 99.9%.

```
# Calculate the probabilities of getting diabetes given seven or more children
df_seven_or_more = df_to_predict[df_to_predict['Pregnancies'] >= 7]
X_seven_or_more = df_seven_or_more[['SevenOrMorePregnancies']]
prob_seven_or_more = model4.predict(X_seven_or_more)
print(f'Probability of getting diabetes given seven or more pregnancies: {prob_seven_or_more.mean()}')
```

✓ 0.0s

```
Probability of getting diabetes given seven or more pregnancies: 0.9999999997522737
```

5. Python Notebook

Attached at Appendix

Reference

[1]Smith, J. W., Everhart, J., Dickson, W., Knowler, W., & Johannes, R. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the annual symposium on computer application in medical care (pp. 261–265).