UNIVERSITY OF TEXAS AT EL PASO (UTEP)

ISAIAH THOMPSON OCANSEY

STAT 5329: Homework 3

In [1]:

```
pip install NumPy
```

```
Requirement already satisfied: NumPy in c:\users\thomo\anaconda3\lib\site-packages (1.2
0.3)
Note: you may need to restart the kernel to use updated packages.
```

In [2]:

```
import pandas as pd
import numpy as np
import statistics as st
import matplotlib.pyplot as plt
import seaborn as sns
```

In [14]:

```
df=pd.read_csv("diabetes.csv",encoding = "Latin-1")
df
```

Out[14]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	6	148	72	35	0	33.6	0.627	50
1	1	85	66	29	0	26.6	0.351	31
2	8	183	64	0	0	23.3	0.672	32
3	1	89	66	23	94	28.1	0.167	21
4	0	137	40	35	168	43.1	2.288	33
...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63
764	2	122	70	27	0	36.8	0.340	27
765	5	121	72	23	112	26.2	0.245	30
766	1	126	60	0	0	30.1	0.349	47
767	1	93	70	31	0	30.4	0.315	23

768 rows × 9 columns

The diabetes data named df has 9 variables and 768 observations.

Question(1) Calculate the mean, median,

standard deviation, IQR, and range for each variable

In [4]:

```
#mean of each variable
df.mean()
```

Out[4]:

```
Pregnancies                   3.845052
Glucose                     120.894531
BloodPressure                69.105469
SkinThickness                20.536458
Insulin                      79.799479
BMI                          31.992578
DiabetesPedigreeFunction      0.471876
Age                          33.240885
Outcome                       0.348958
dtype: float64
```

The mean for each variable in the dataset can be seen above. From the Output above, Glucose has the highest mean which means 768 sample of female patients of Pima Indian heritage has an average of 120.89 Glucose levels in their blood followed by an average of 79.79 insulin spike

In [5]:

```
#median of each variable
df.median()
```

Out[5]:

```
Pregnancies                   3.0000
Glucose                     117.0000
BloodPressure                72.0000
SkinThickness                23.0000
Insulin                      30.5000
BMI                          32.0000
DiabetesPedigreeFunction      0.3725
Age                          29.0000
Outcome                       0.0000
dtype: float64
```

The median for each variable in the dataset can be seen above. The variable Glucose has the highest median of 117.

In [6]:

```
#Standard deviation of each variable in the data set
df.std()
```

Out[6]:

```
Pregnancies                   3.369578
Glucose                      31.972618
BloodPressure                19.355807
SkinThickness                15.952218
Insulin                     115.244002
BMI                           7.884160
DiabetesPedigreeFunction      0.331329
Age                          11.760232
Outcome                       0.476951
dtype: float64
```

The standard deviation for each variable in the dataset can be seen above. Here, Insulin has the highest variability of 115.2 followed by Glucose.

In [7]:

```
#IQR of the variables
from scipy.stats import iqr
iqr(df)
```

Out[7]:

```
60.53775
```

The Interquartile Range (IQR) for each variable in the dataset can be seen above

In [8]:

```

# Rande of each variable
df.max() - df.min()
```

Out[8]:

```
Pregnancies                  17.000
Glucose                     199.000
BloodPressure               122.000
SkinThickness                99.000
Insulin                     846.000
BMI                          67.100
DiabetesPedigreeFunction      2.342
Age                          60.000
Outcome                       1.000
dtype: float64
```

The range for each variable in the dataset can be seen above

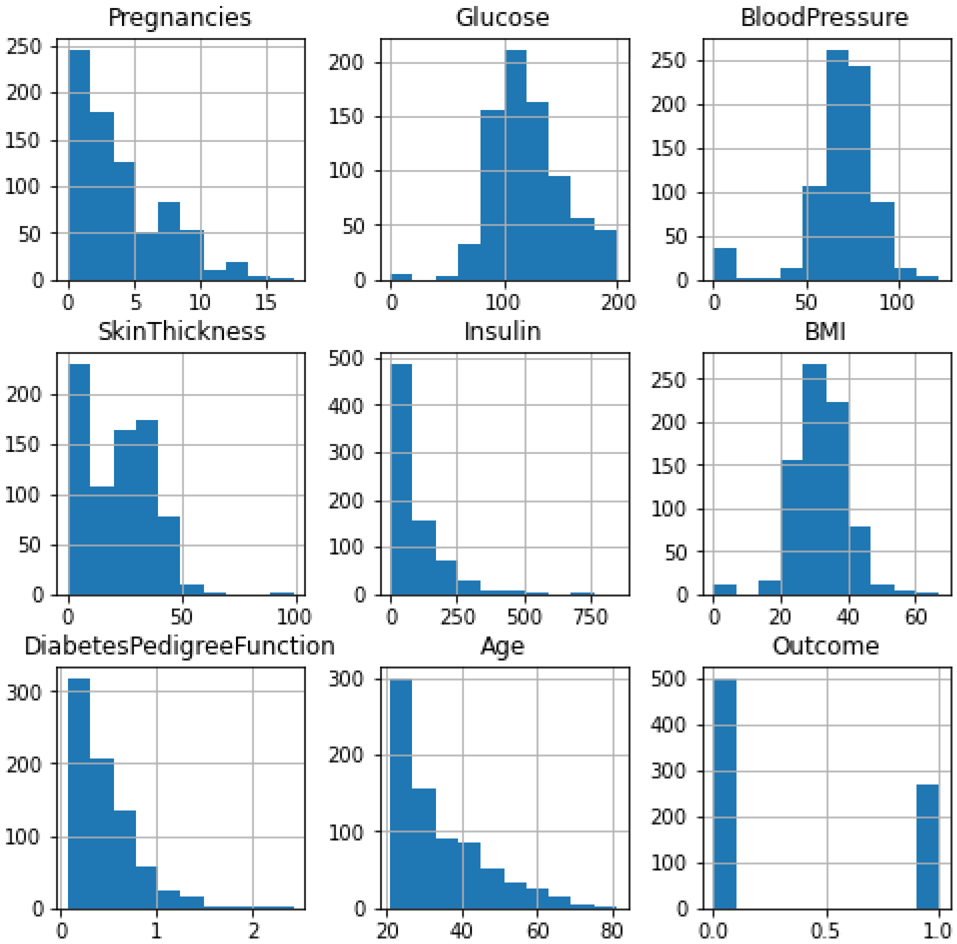
Question(2) For each variable, construct a histogram and comment on the shape of the distribution. Identify the variables that have similar histograms in terms of shape

In [9]:

```

fig = plt.figure(figsize = (8,8))
ax = fig.gca()
df.hist(ax=ax)
plt.show()
```

```
C:\Users\thomo\AppData\Local\Temp/ipykernel_110820/3175089013.py:3: UserWarning: To outp
ut multiple subplots, the figure containing the passed axes is being cleared
  df.hist(ax=ax)
```

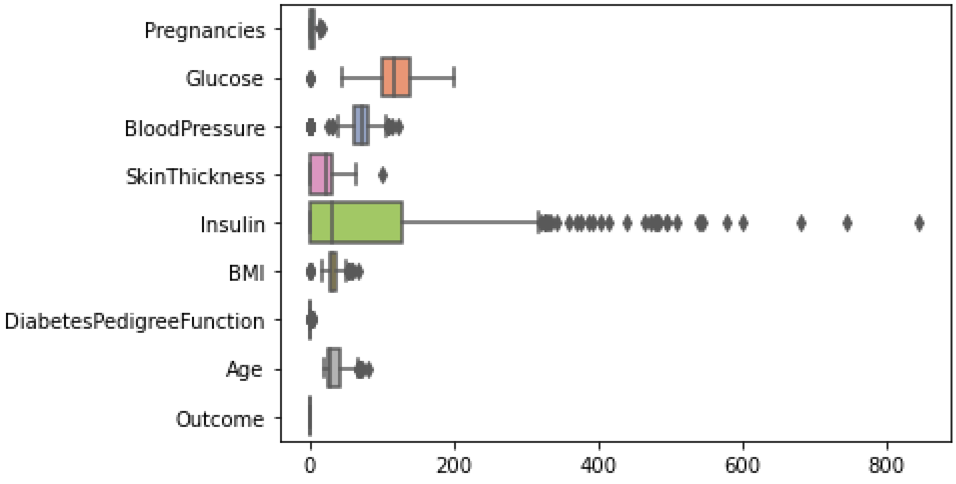


It can be observed from the histograms above that Glucose,Blood Presure and BMI are fairly normally distributed whiles Pregnancies, Skin Thickness, Insulin,Diabetes Pedigree Function and Age are skewed. We can also observe that Pregnancy, Age, Diabetes Pedigree Function and Insulin looks similar in terms of the shape of their histograms. Also, Glucose, BMI and Blood Predsure looks similar in terms of the shape of their histograms.

Question(3) Construct box-and-whisker plots for all variables

In [10]:

```
ax = sns.boxplot(data=df, orient="h", palette="Set2")
```

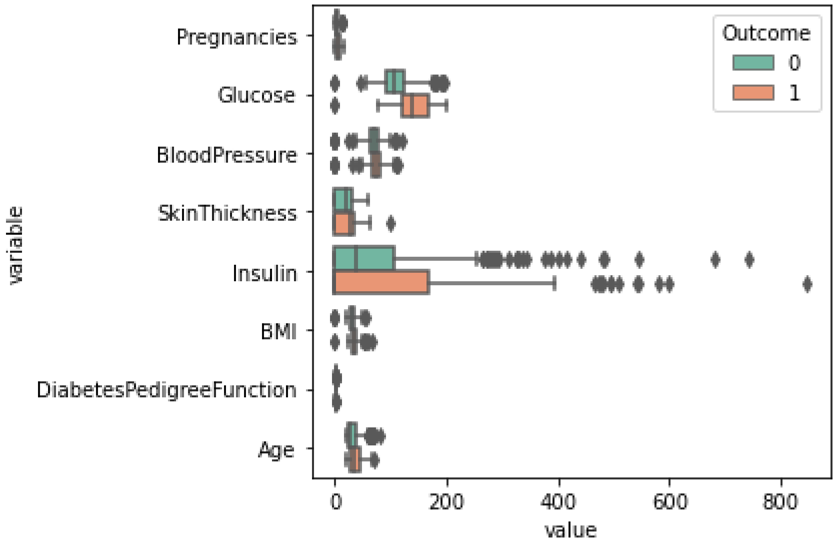


We can observe from the box-and-whisker plot above that the variable Insulin has outliers.

Question(4) Construct side-by-side boxplots for all the variables for the Outcome=1 and Outcome=0 groups. Identify the variables that have boxplots with different shapes between the two groups? What does it mean?

In [11]:

```
df_long = df.melt(id_vars=['Outcome'])
ax = sns.boxplot(data=df_long, x="value", y="variable", orient="h", palette="Set2", hue
plt.tight_layout()
plt.show()
```



It can be observed from the above boxplot that the variables; Insulin and Glucose have different shapes between the two groups and this means that; at least 21 years old female patients of Pima Indian heritage with high insulin and Glucose levels are at risk of being diabetic whiles those with low insulin and glucose levels are at no risk of being diabetic. This explanation is given on the premise that the response variable,Outcome=1 means diabetic and 0 therwise.

In [ ]: