Analysis 1 -- cleaning data

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

In [2]:
```
# Read the data from 'Diabetes.csv' file into a pandas dataframe

df = pd.read_csv('../Data/Diabetes.csv')
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
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
```

In [3]:
```
# Get Statistics on the columns

df.describe()
```

Out[3]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

In [4]:
```
# Check for missing values

df.isnull()
```

Out[4]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...
763	False	False	False	False	False	False	False	False	False
764	False	False	False	False	False	False	False	False	False
765	False	False	False	False	False	False	False	False	False
766	False	False	False	False	False	False	False	False	False
767	False	False	False	False	False	False	False	False	False

768 rows × 9 columns

In [5]:
```
df.columns
```

Out[5]:
```
Index(['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',
       'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],
      dtype='object')
```

OBSERVATION: For columns such as 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', I see a value of 0.

In [6]:
```
# Find the records for which 'Glucose', 'BloodPressure', 'SkinThickness', 'BMI' are 0

bad_glu_df = df[df['Glucose'] == 0]
bad_glu_df
```

Out[6]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
75	1	0	48	20	0	24.7	0.140	22	0
182	1	0	74	20	23	27.7	0.299	21	0
342	1	0	68	35	0	32.0	0.389	22	0
349	5	0	80	32	0	41.0	0.346	37	1
502	6	0	68	41	0	39.0	0.727	41	1

In [8]:
```
bad_bp_df = df[df['BloodPressure'] == 0]
bad_bp_df
```

Out[8]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
7	10	115	0	0	0	35.3	0.134	29	0
15	7	100	0	0	0	30.0	0.484	32	1
49	7	105	0	0	0	0.0	0.305	24	0
60	2	84	0	0	0	0.0	0.304	21	0
78	0	131	0	0	0	43.2	0.270	26	1
81	2	74	0	0	0	0.0	0.102	22	0
172	2	87	0	23	0	28.9	0.773	25	0
193	11	135	0	0	0	52.3	0.578	40	1
222	7	119	0	0	0	25.2	0.209	37	0
261	3	141	0	0	0	30.0	0.761	27	1
266	0	138	0	0	0	36.3	0.933	25	1
269	2	146	0	0	0	27.5	0.240	28	1
300	0	167	0	0	0	32.3	0.839	30	1
332	1	180	0	0	0	43.3	0.282	41	1
336	0	117	0	0	0	33.8	0.932	44	0
347	3	116	0	0	0	23.5	0.187	23	0
357	13	129	0	30	0	39.9	0.569	44	1
426	0	94	0	0	0	0.0	0.256	25	0
430	2	99	0	0	0	22.2	0.108	23	0
435	0	141	0	0	0	42.4	0.205	29	1
453	2	119	0	0	0	19.6	0.832	72	0
468	8	120	0	0	0	30.0	0.183	38	1
484	0	145	0	0	0	44.2	0.630	31	1
494	3	80	0	0	0	0.0	0.174	22	0
522	6	114	0	0	0	0.0	0.189	26	0
533	6	91	0	0	0	29.8	0.501	31	0
535	4	132	0	0	0	32.9	0.302	23	1
589	0	73	0	0	0	21.1	0.342	25	0
601	6	96	0	0	0	23.7	0.190	28	0
604	4	183	0	0	0	28.4	0.212	36	1
619	0	119	0	0	0	32.4	0.141	24	1
643	4	90	0	0	0	28.0	0.610	31	0
697	0	99	0	0	0	25.0	0.253	22	0
703	2	129	0	0	0	38.5	0.304	41	0
706	10	115	0	0	0	0.0	0.261	30	1

In [10]:
```
bad_bp_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 35 entries, 7 to 706
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Pregnancies               35 non-null     int64  
 1   Glucose                   35 non-null     int64  
 2   BloodPressure             35 non-null     int64  
 3   SkinThickness             35 non-null     int64  
 4   Insulin                   35 non-null     int64  
 5   BMI                       35 non-null     float64
 6   DiabetesPedigreeFunction  35 non-null     float64
 7   Age                       35 non-null     int64  
 8   Outcome                   35 non-null     int64  
dtypes: float64(2), int64(7)
memory usage: 2.7 KB
```

In [11]:
```
bad_skin_df = df[df['SkinThickness'] == 0]
bad_skin_df
```

Out[11]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
2	8	183	64	0	0	23.3	0.672	32	1
5	5	116	74	0	0	25.6	0.201	30	0
7	10	115	0	0	0	35.3	0.134	29	0
9	8	125	96	0	0	0.0	0.232	54	1
10	4	110	92	0	0	37.6	0.191	30	0
...	...	...	...	...	...	...	...	...	...
757	0	123	72	0	0	36.3	0.258	52	1
758	1	106	76	0	0	37.5	0.197	26	0
759	6	190	92	0	0	35.5	0.278	66	1
762	9	89	62	0	0	22.5	0.142	33	0
766	1	126	60	0	0	30.1	0.349	47	1

227 rows × 9 columns

In [12]:
```
bad_skin_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 227 entries, 2 to 766
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Pregnancies               227 non-null    int64  
 1   Glucose                   227 non-null    int64  
 2   BloodPressure             227 non-null    int64  
 3   SkinThickness             227 non-null    int64  
 4   Insulin                   227 non-null    int64  
 5   BMI                       227 non-null    float64
 6   DiabetesPedigreeFunction  227 non-null    float64
 7   Age                       227 non-null    int64  
 8   Outcome                   227 non-null    int64  
dtypes: float64(2), int64(7)
memory usage: 17.7 KB
```

In [13]:
```
bad_bmi_df = df[df['BMI'] == 0]
bad_bmi_df
```

Out[13]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
9	8	125	96	0	0	0.0	0.232	54	1
49	7	105	0	0	0	0.0	0.305	24	0
60	2	84	0	0	0	0.0	0.304	21	0
81	2	74	0	0	0	0.0	0.102	22	0
145	0	102	75	23	0	0.0	0.572	21	0
371	0	118	64	23	89	0.0	1.731	21	0
426	0	94	0	0	0	0.0	0.256	25	0
494	3	80	0	0	0	0.0	0.174	22	0
522	6	114	0	0	0	0.0	0.189	26	0
684	5	136	82	0	0	0.0	0.640	69	0
706	10	115	0	0	0	0.0	0.261	30	1

In [14]:
```
bad_bmi_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 11 entries, 9 to 706
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Pregnancies               11 non-null     int64  
 1   Glucose                   11 non-null     int64  
 2   BloodPressure             11 non-null     int64  
 3   SkinThickness             11 non-null     int64  
 4   Insulin                   11 non-null     int64  
 5   BMI                       11 non-null     float64
 6   DiabetesPedigreeFunction  11 non-null     float64
 7   Age                       11 non-null     int64  
 8   Outcome                   11 non-null     int64  
dtypes: float64(2), int64(7)
memory usage: 880.0 bytes
```

In [17]:
```
# Deleting records with 'Glucose', 'BloodPressure', 'SkinThickness', 'BMI' are 0

clean_df = df[(df['Glucose'] > 0) & (df['BloodPressure'] > 0) & (df['SkinThickness'] > 0) & (df['BMI'] > 0)]
clean_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 532 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Pregnancies               532 non-null    int64  
 1   Glucose                   532 non-null    int64  
 2   BloodPressure             532 non-null    int64  
 3   SkinThickness             532 non-null    int64  
 4   Insulin                   532 non-null    int64  
 5   BMI                       532 non-null    float64
 6   DiabetesPedigreeFunction  532 non-null    float64
 7   Age                       532 non-null    int64  
 8   Outcome                   532 non-null    int64  
dtypes: float64(2), int64(7)
memory usage: 41.6 KB
```

In [18]:
```
# Send clean_df to a csv file called "clean_Diabetes.csv"

clean_df.to_csv("../data/clean_Diabetes.csv")
```

In [ ]: