Question1: What is the correlation between features in the dataset. and what two features have very strong correlation with the independent variable? Justify with reason

When two sets of data are strongly linked together we say they have a High Correlation .The word correlation (Co means together hence it is together relation) is Positive when the values increase together and correlation is Negative when one value decreases as the other increase. example : Set of Icecream sales vs Set of ice cream temperature.

Pandas(loc and iloc) and Numpy are the 2 features have very strong correlation with independent variable.

```
!pip install pandas
```

```
import pandas as pd
import numpy as np
```

```
data = {'Naveen K': pd.Series([50,40,30,20]),
        'Manoj BV ': pd.Series([60,50,30,25])}
```

+ Code + Text

```
df = pd.DataFrame(data)
```

```
df
```

	Naveen K	Manoj BV
0	50	60
1	40	50
2	30	30
3	20	25

```
item = {'Naveen K': pd.Series([50,40,30,20], index=['English', 'Maths', 'Kannada', 'Science']
        'Manoj BV': pd.Series([60,50,30,25], index=['English', 'Maths', 'Kannada', 'Science'
```

```
cart = pd.DataFrame(item)
cart
```

	Naveen K	Manoj BV
English	50	60
Maths	40	50

```
cart.iloc[[1,2,3]]
```

	Naveen K	Manoj BV
Maths	40	50
Kannada	30	30
Science	20	25

```
cart.loc[['Kannada','Maths']]
```

	Naveen K	Manoj BV
Kannada	30	30
Maths	40	50

Question 2: Which feature has more outliers. Explain with a visualization.

Pandas/Data frames and some more are the features which have more outliers.

```
import matplotlib.pyplot as plt
import numpy as np
```

```
x=np.arange(0,10)
y=np.arange(10,20)
```

```
y
```

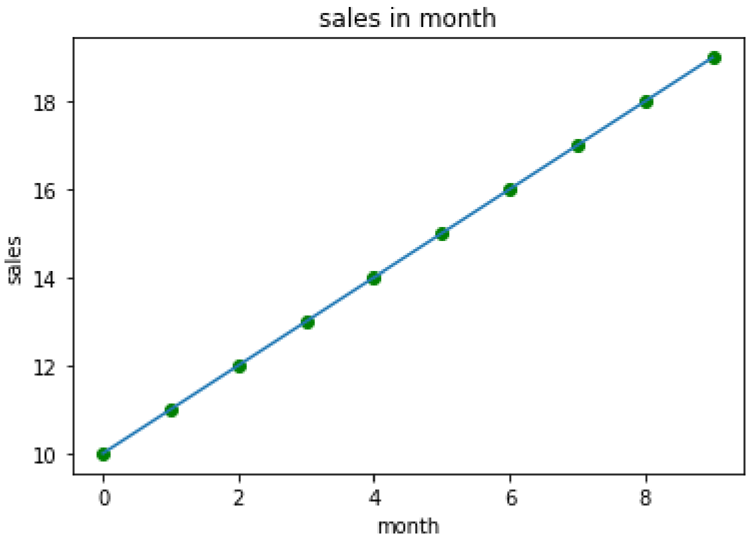
```
     array([10, 11, 12, 13, 14, 15, 16, 17, 18, 19])
```

```
x
```

```
     array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9])
```

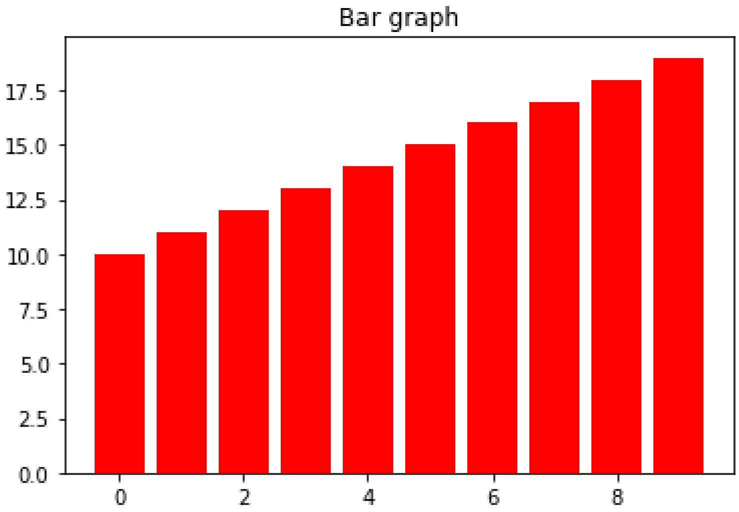
```
##plotting using matplotlib

##plt scatter
plt.scatter(x,y,c='r')
plt.scatter(x,y,c='g')
plt.xlabel('month')
plt.ylabel('sales')
plt.title('sales in month')
plt.savefig('Test.png')
plt.plot(x,y)
```

```
    [<matplotlib.lines.Line2D at 0x7f864da34250>]
```



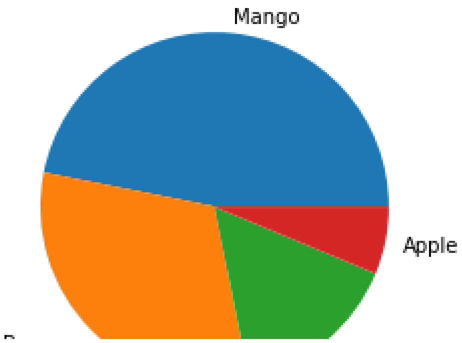
```
plt.bar(x, y, color = 'r')
plt.title('Bar graph')
```

```
    Text(0.5, 1.0, 'Bar graph')
```



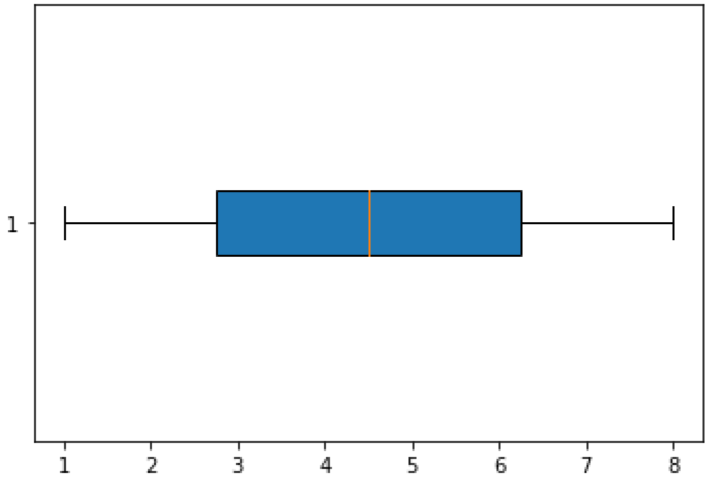
```
data = 'Mango', 'Banana', 'orange', 'Apple'
sizes = [150, 100, 50, 20]
plt.pie(sizes, labels=data)
```

```
([<matplotlib.patches.Wedge at 0x7f864d6e5e50>,
  <matplotlib.patches.Wedge at 0x7f864d6c7e90>,
  <matplotlib.patches.Wedge at 0x7f864d6f3250>,
  <matplotlib.patches.Wedge at 0x7f864d6f3850>],
 [Text(0.10781885436251686, 1.0947031993394165, 'Mango'),
  Text(-0.7778174593052025, -0.7778174593052023, 'Banana'),
  Text(0.6978326125800102, -0.8503114986990107, 'orange'),
  Text(1.0788638084435533, -0.21459935421774162, 'Apple')])
```



```
data = np.array([1,2,3,4,5,6,7,8])
plt.boxplot(data,vert=False,patch_artist=True)
```

```
{'boxes': [<matplotlib.patches.PathPatch at 0x7f864db4af50>],
 'caps': [<matplotlib.lines.Line2D at 0x7f864d9f5d50>,
  <matplotlib.lines.Line2D at 0x7f864db46790>],
 'fliers': [<matplotlib.lines.Line2D at 0x7f864db69390>],
 'means': [],
 'medians': [<matplotlib.lines.Line2D at 0x7f864db46090>],
 'whiskers': [<matplotlib.lines.Line2D at 0x7f864d9f5790>,
  <matplotlib.lines.Line2D at 0x7f864d9f5150>]}
```



Question3: In which Age group Majority of people have diabetes? Make a visualization to validate your finding.

Link : https://raw.githubusercontent.com/plotly/datasets/master/diabetes.csv

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

```
df = pd.read_csv('https://raw.githubusercontent.com/plotly/datasets/master/diabetes.csv')
```

```
df.head(3)
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFu
0	6	148	72	35	0	33.6	
1	1	85	66	29	0	26.6	
2	8	183	64	0	0	23.3	

```
df.tail()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigree
763	10	101	76	48	180	32.9	
764	2	122	70	27	0	36.8	
765	5	121	72	23	112	26.2	
766	1	126	60	0	0	30.1	
767	1	93	70	31	0	30.4	

```
df.ndim
```

```
2
```

```
df.shape
```

```
(768, 9)
```

```
df.columns
```

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

```
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
```

```
df.isnull().sum().sum()
```

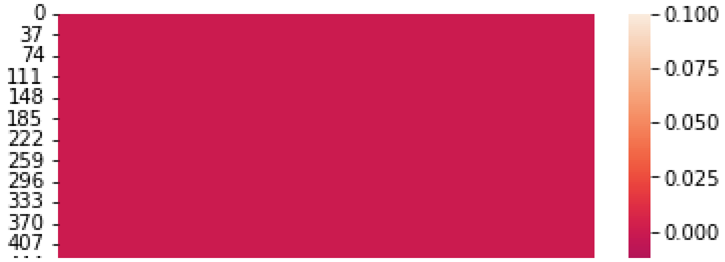
```
0
```

```
df.describe()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	Dia
count	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	
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	

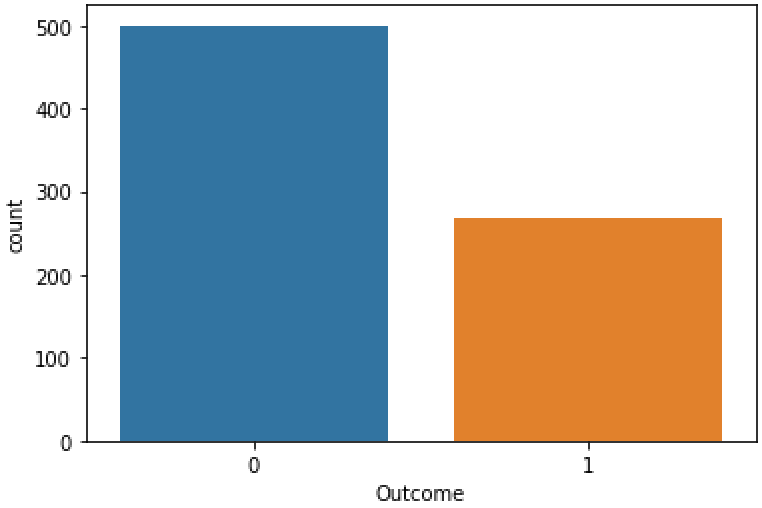
```
sns.heatmap(df.isnull())
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4a126ab850>
```



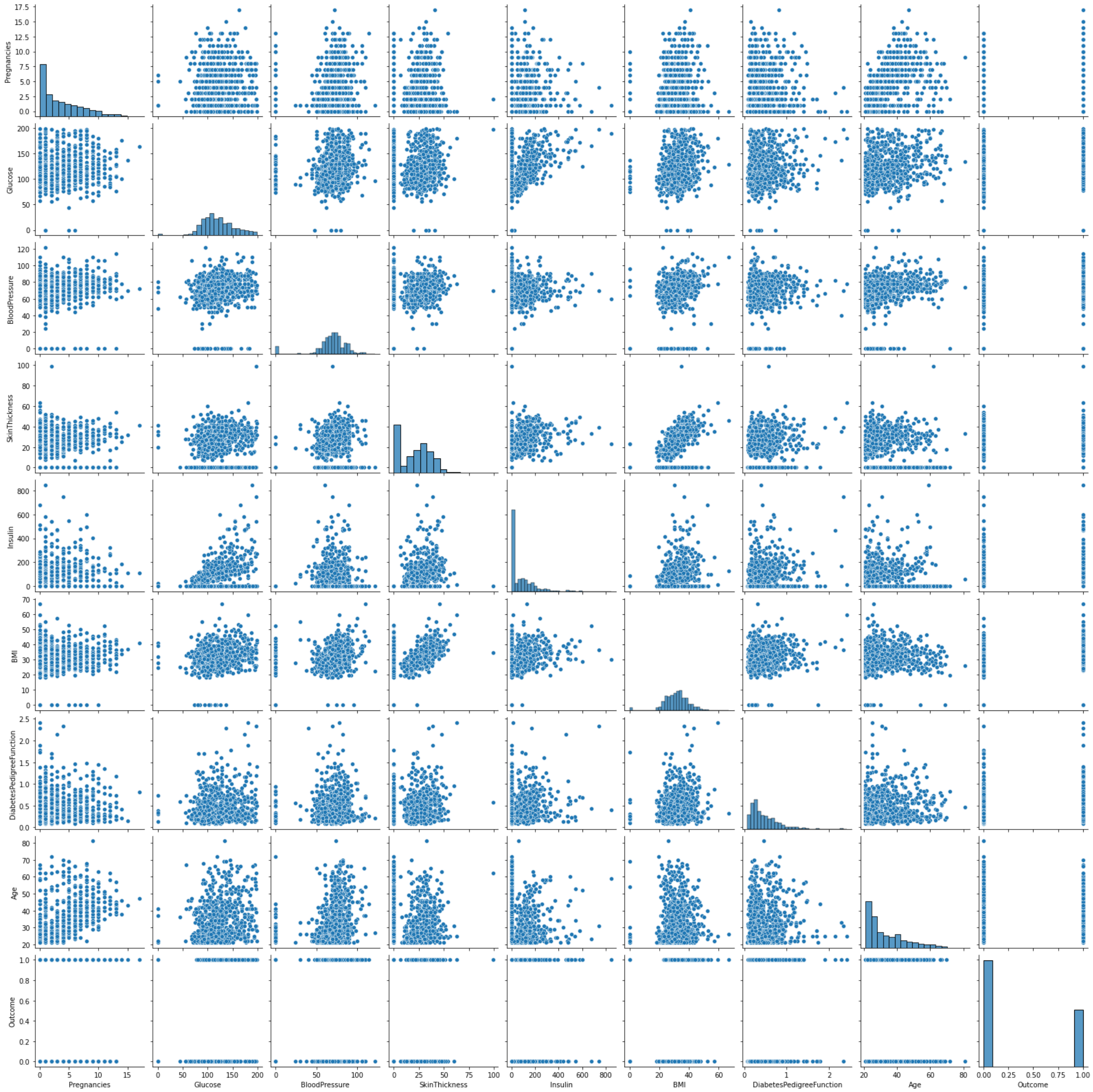
```
sns.countplot('Outcome',data=df)
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass th
  FutureWarning
<matplotlib.axes._subplots.AxesSubplot at 0x7f4a125a18d0>
```



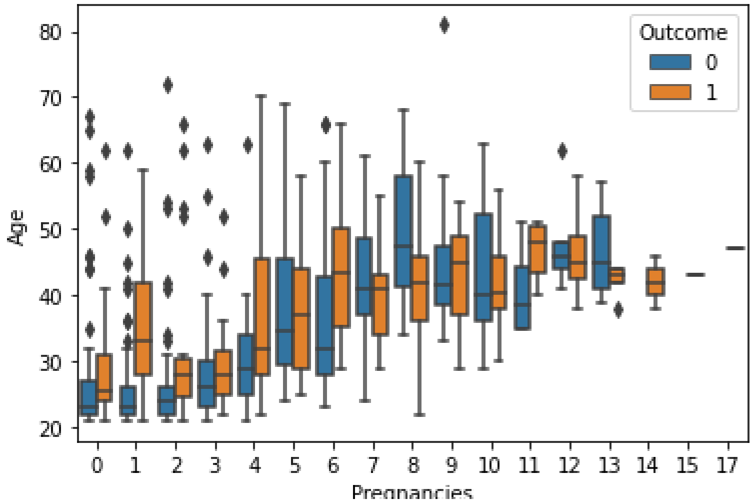
```
sns.pairplot(df)
```

```
<seaborn.axisgrid.PairGrid at 0x7f4a1255abd0>
```



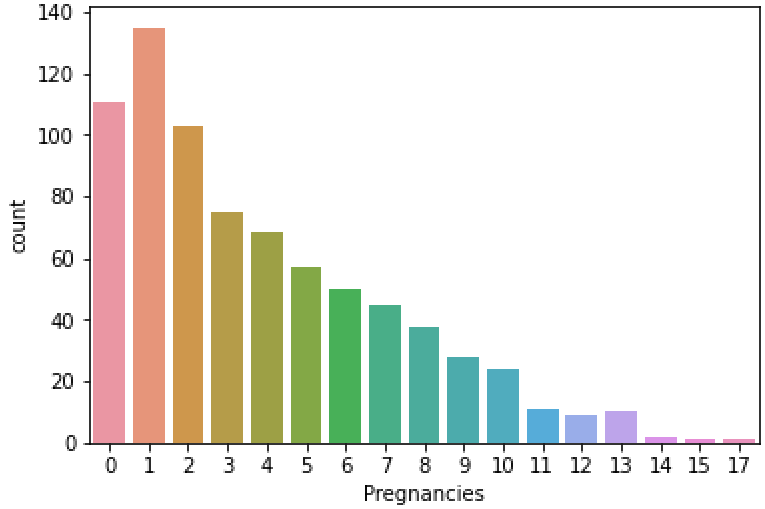
```
sns.boxplot(x="Pregnancies",y="Age",data=df,hue="Outcome")
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4a1030a3d0>
```



```
sns.countplot(x="Pregnancies",data=df)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4a1007e850>
```

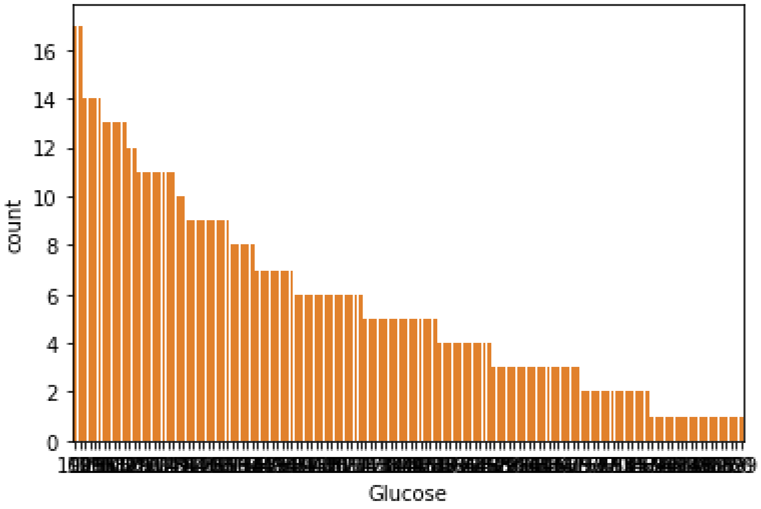


```
 df['Glucose'].value_counts()
```

```
100    17
99     17
129    14
125    14
111    14
       ..
177     1
172     1
169     1
160     1
199     1
Name: Glucose, Length: 136, dtype: int64
```

```
base_color = sns.color_palette()[1]
gen_order = df['Glucose'].value_counts().index
sns.countplot(data = df, x = 'Glucose', color = base_color,
sns.countplot(data = df, x = 'Glucose', color = base_color,
              order = gen_order)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4a0fec8450>
```


count
Glucose