Diabetes case study

This case study is about analyzing the dataset that contains details about diabetes. This kaggle dataset is originally from National institute of Diabetes and Digestive and Kidney diseases. This dataset has details like patient number,age,weight,bmi,cholesterol,pressure and whether the patient has diabetes or not.

Business Task:

To identify the different parameters from the given varaiables that contribute to diabetes.

Importing the packages

In order to analyze, clean and visualize the data, I am importing numpy,pandas,matplotlib and seaborn packages.

In [14]:

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

%matplotlib inline
```

Importing the dataset

Importing the diabetes dataset into a dataframe and previewing first 5 lines of dataset.

In [15]:

```
diab_df = pd.read_csv('C:/Users/senth/Downloads/diabetes.csv')
```

Data Cleaning

In [16]:

```
diab_df.head()
```

Out[16]:

	patient_number	cholesterol	glucose	hdl_chol	chol_hdl_ratio	age	gender	height	weight	bmi
0	1	193	77	49	3,9	19	female	61	119	22,5
1	2	146	79	41	3,6	19	female	60	135	26,4
2	3	217	75	54	4	20	female	67	187	29,3
3	4	226	97	70	3,2	20	female	64	114	19,6
4	5	164	91	67	2,4	20	female	70	141	20,2

In [17]:

```
diab_df.dtypes
```

Out[17]:

```
patient_number      int64
cholesterol         int64
glucose             int64
hdl_chol            int64
chol_hdl_ratio     object
age                 int64
gender             object
height              int64
weight              int64
bmi                object
systolic_bp         int64
diastolic_bp        int64
waist               int64
hip                 int64
waist_hip_ratio    object
diabetes           object
dtype: object
```

By observing the data types of the different columns we can see that chol_hdl_ratio,bmi,waist_hip_ratio are objects when that should be numeric. Lets convert the datatype of those columns.

In [18]:

```
diab_df['chol_hdl_ratio'] = diab_df['chol_hdl_ratio'].str.replace(',','.')
diab_df['bmi'] = diab_df['bmi'].str.replace(',','.')
diab_df['waist_hip_ratio'] = diab_df['waist_hip_ratio'].str.replace(',','.')
```

In [19]:

```
diab_df[['chol_hdl_ratio','bmi','waist_hip_ratio']] = diab_df[['chol_hdl_ratio','bmi','
```

In [20]:

```
diab_df.dtypes
```

Out[20]:

```
patient_number       int64
cholesterol          int64
glucose              int64
hdl_chol             int64
chol_hdl_ratio     float64
age                  int64
gender              object
height               int64
weight               int64
bmi                float64
systolic_bp          int64
diastolic_bp         int64
waist                int64
hip                  int64
waist_hip_ratio    float64
diabetes            object
dtype: object
```

In [21]:

```
ProfileReport(diab_df)
```

Report generated with pandas-profiling.

Out[21]:

By using pandas profiling , i observed if there are any missing values or duplicate values in any of the columns. There are no missing or duplicate values in the dataset. Some statistic summaries are available from pandas profiling which shows relation between different variables.

Analyzing and Visualizing the data

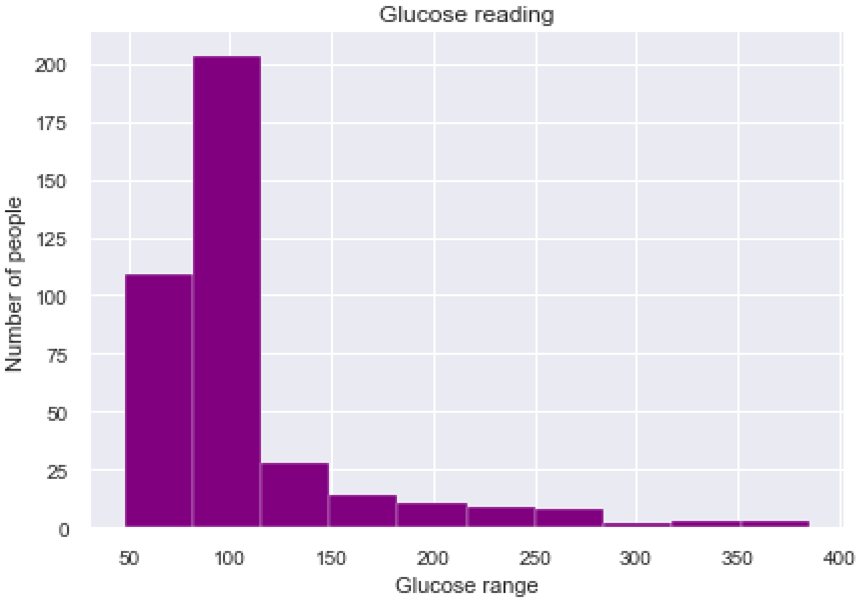
I want to find the range of glucose reading in the patients .

In [22]:

```
fig,ax = plt.subplots(nrows = 1, ncols = 1)
plt.hist(diab_df['glucose'],color='purple')
fig.tight_layout()
plt.title('Glucose reading')
plt.xlabel('Glucose range')
plt.ylabel('Number of people')
plt.figure(figsize=(10,10))
```

Out[22]:

```
<Figure size 720x720 with 0 Axes>
```



```
<Figure size 720x720 with 0 Axes>
```

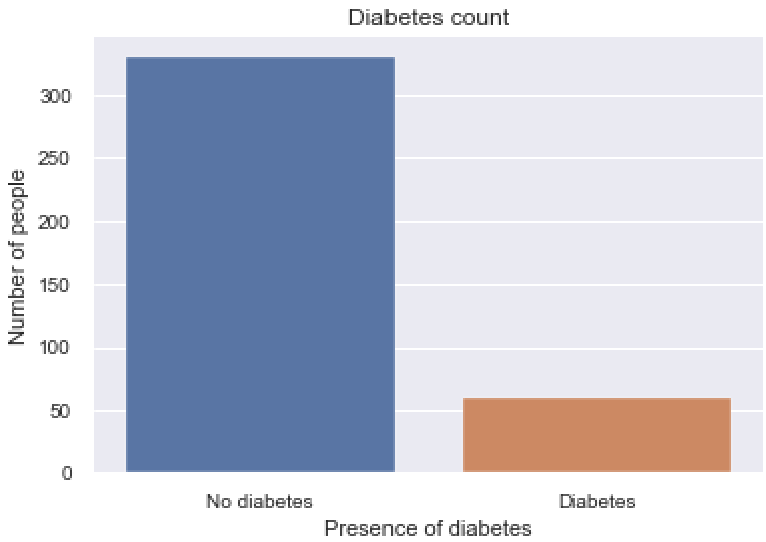
From this glucose range histogram, I find that most of the patients glucose reading is in the normal range of less than 140 mg/dl. So I want to find the number of people with diabetes.

In [23]:

```
sns.set_style('darkgrid')
sns.set_palette('deep')
sns.countplot(data=diab_df,x='diabetes').set(title='Diabetes count',xlabel='Presence of
```

Out[23]:

```
[Text(0.5, 1.0, 'Diabetes count'),
 Text(0.5, 0, 'Presence of diabetes'),
 Text(0, 0.5, 'Number of people')]
```



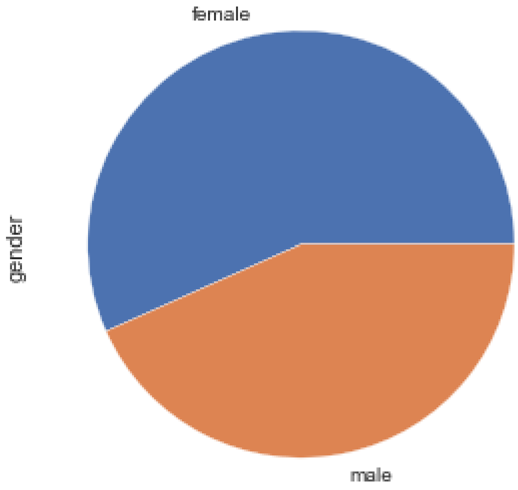
0f the 390 patients , 60 patients are diabetic. I filtered the dataframe to create a new dataframe containing only diabetic patient details.

In [24]:

```
diab= diab_df[diab_df['diabetes'] == 'Diabetes']
```

In [25]:

```
fig = diab['gender'].value_counts().plot.pie().get_figure()
fig.tight_layout()
```



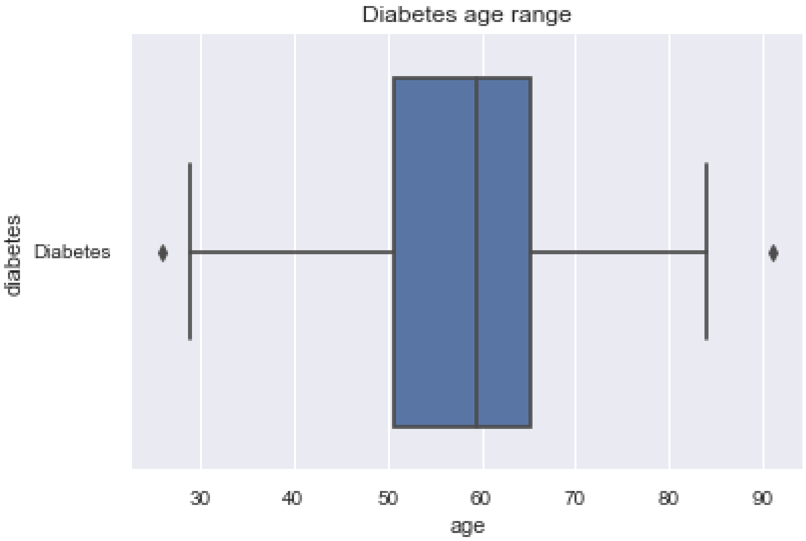
From this pie chart ,we can see that the number of female diabetic patients is more than the male diabetic patients. Next I want to find whether diabetes is widespread in older people or young people.

In [26]:

```
sns.set_style('darkgrid')
sns.boxplot(x='age',y='diabetes',data = diab).set(title='Diabetes age range')
```

Out[26]:

```
[Text(0.5, 1.0, 'Diabetes age range')]
```



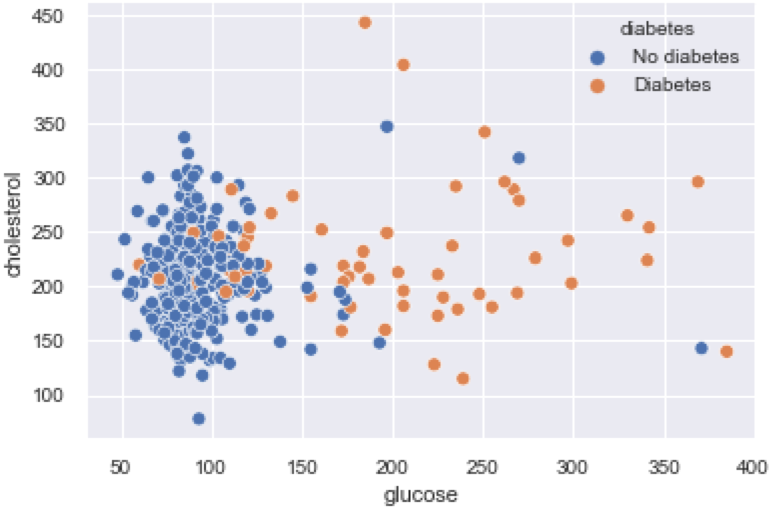
From the above boxplot, it is clear that old people between the age 50 to 65 are diabetic with some outlier patients having diabetes before the age of 30 and after the age of 90.

In [121…

```
sns.scatterplot(x=diab_df['glucose'],y=diab_df['cholesterol'],hue=diab_df['diabetes'])
```

Out[121…

```
<AxesSubplot:xlabel='glucose', ylabel='cholesterol'>
```



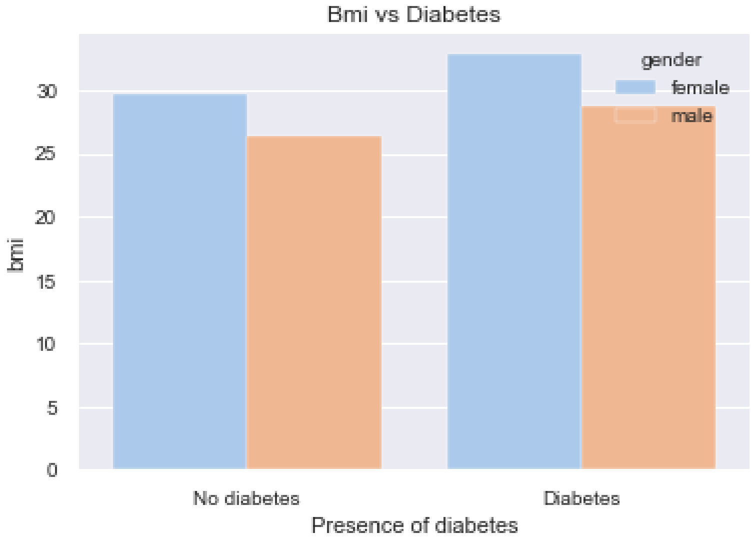
From an earlier heatmap I observed there is no correlation between cholesterol and glucose.I visualized using scatterplot and confirmed there is no correlation between cholesterol and glucose.

In [28]:

```
sns.set_style('darkgrid')
sns.set_palette('pastel')
sns.barplot(x='diabetes',y='bmi',hue='gender', data= diab_df,ci=None).set(title='Bmi vs
```

Out[28]:

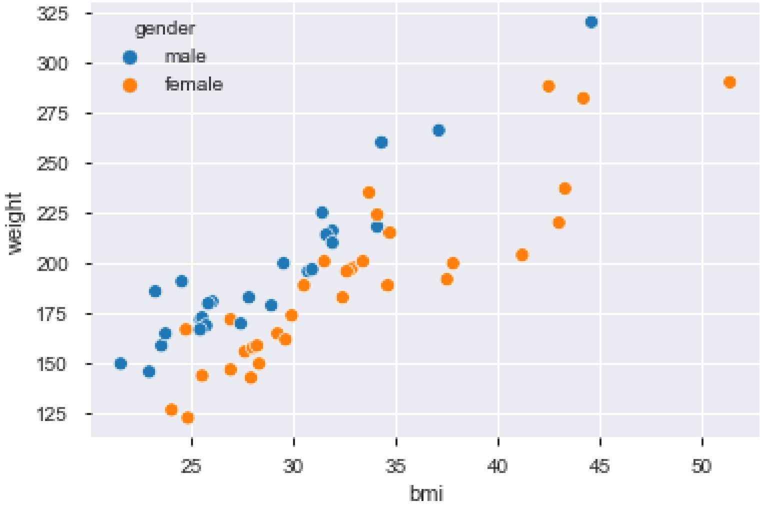
```
[Text(0.5, 1.0, 'Bmi vs Diabetes'), Text(0.5, 0, 'Presence of diabetes')]
```



When looking at the bmi of diabetic and non-diabetic patients, we can clearly see that the bmi of diabetic patients are more than the bmi of non-diabetic patients. So I visualized the relation between bmi and weight of the diabetic patients.

```
In [14]: sns.scatterplot(y='weight',x='bmi',hue='gender',data=diab)
```

```
Out[14]: <AxesSubplot:xlabel='bmi', ylabel='weight'>
```



From the above scatter plot we can see that the bmi increases as the weight of the patient increases . We have a positive linear correlation between the 2 variables.

Recommendations:

- Since most diabetic patients are old people between the age 50 to 65, it is essential that the middle age people take care of their health by keeping the glucose levels in check in order to avoid becoming diabetic.
- The diabetic patients bmi is higher than the non-diabetic patients. Bmi increases as weight increases. So people whose bmi is more than the recommended level of 25, can take measures to reduce the weight that keeps the glucose level normal.