- **Introduction**

This report will explain about Intractive visualization techniques that were used to create six charts for the challenging dataset about diabetes and provide a rigorous rationale for my design choices. It also shows what aspects of the data are chosen to convey the key features to the user effectively and which aspects of the data might be obscured or downplayed due to the selected visualization design. Moreover, I pose my initial question after that I investigate dataset and answer the question base on plot designed.

- **Loading and Summary of the Dataset**

Firstly, import the csv files for our analysis.

```
[ ]  1 d = pd.read_csv("/content/diabetes.csv")
     2
```

```
[ ]  1 print(f" the shape of dataset is: {d.shape}")
```

```
 the shape of dataset is: (768, 9)
```

```
     1 d.head()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.63	50	1
1	1	85	66	29	0	26.6	0.35	31	0
2	8	183	64	0	0	23.3	0.67	32	1
3	1	89	66	23	94	28.1	0.17	21	0
4	0	137	40	35	168	43.1	2.29	33	1

- **Summary of the dataset**

all variables are float and integer and we don't have any nominal variable.

```
1 d.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
```

- List of Features & Their Unique Values

```
1 d.nunique()
```

```
Pregnancies                 17
Glucose                    136
BloodPressure               47
SkinThickness               51
Insulin                    186
BMI                        248
DiabetesPedigreeFunction   517
Age                         52
Outcome                      2
dtype: int64
```

- **Data Preparation**

1. Data Cleaning

This is the initial and most common task in data preparation that is performed on raw data. Data cleansing is the process of examining, identifying, and mitigating errors in raw data.

In this part we check missing value of data. As the below code shows our dataset doesn't any missing value. But the missing values fill with 0 so we can fill them by mean since it can lead to the better performance in visualization.

```
1 d.isnull().sum()
```

```
Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

```
[5]  1 d[d==0]=np.nan
```

```
1 d.isnull().sum()
```

```
Pregnancies                 111
Glucose                       5
BloodPressure                35
SkinThickness               227
Insulin                     374
BMI                          11
DiabetesPedigreeFunction      0
Age                           0
Outcome                     500
dtype: int64
```

```
[7]  1 for i in d.columns:
     2   d[i]=d[i].fillna(d[i].mean())
```

2. Normalizing and Scaling the Dataset

```
from sklearn.preprocessing import MinMaxScaler

ss = MinMaxScaler()
dnorm = ss.fit_transform(d)
dnorm = pd.DataFrame(dnorm, columns=d.columns)
```

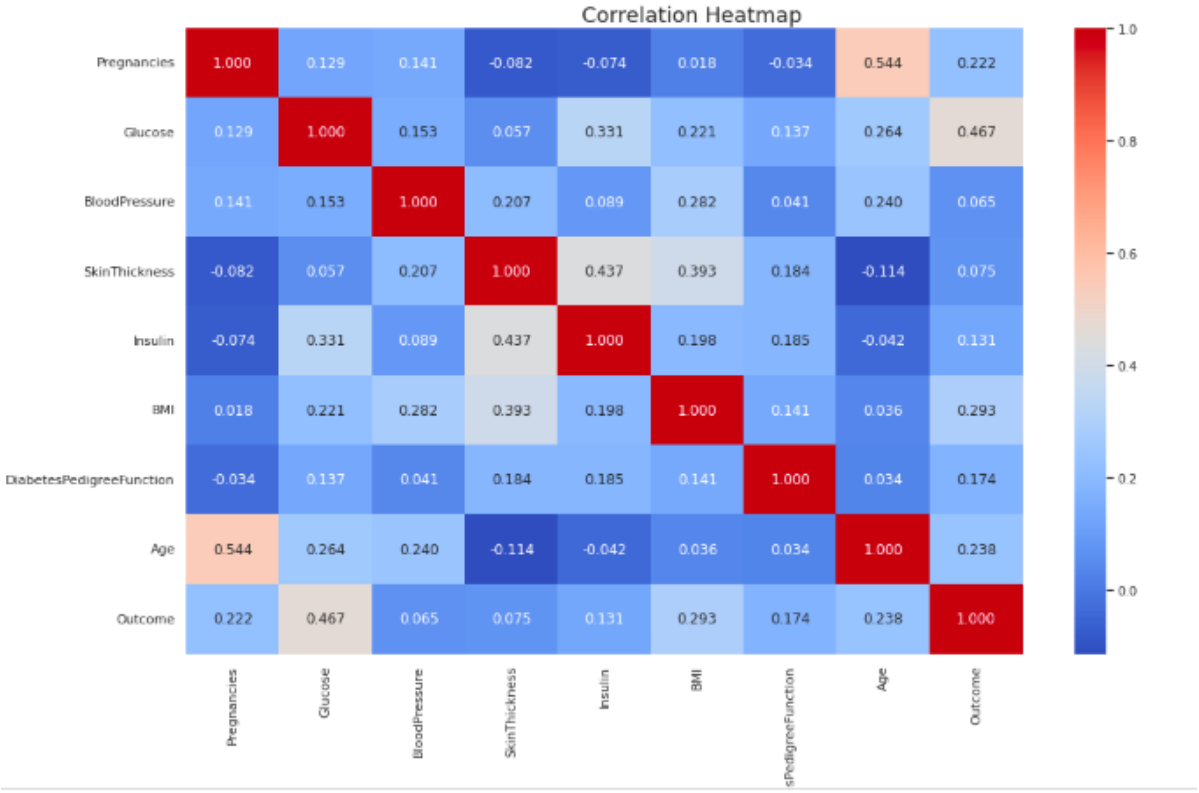
In this step the data has been normalize by minmaxscaler method.

- **Visualization**

Plot Selection 1 -Correlation Heatmap for selecting the suitable variables

I use plot a heatmap to investigate the correlation among our features. As it's illustrated, all the variables are related to each other, some of the features are highly correlated. Also some of them have a positive relationship and some have a negative relationship.

The below plots shows that Pregnancies, Glucose, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age and Blood Pressure have a positive relationship with diabetes(outcome). But the correlation of Blood Pressure and Skin Thickness are really low. So I investigate the effect of only Pregnancies, Glucose, Insulin, BMI, Diabetes Pedigree Function and Age variables on diabetes.

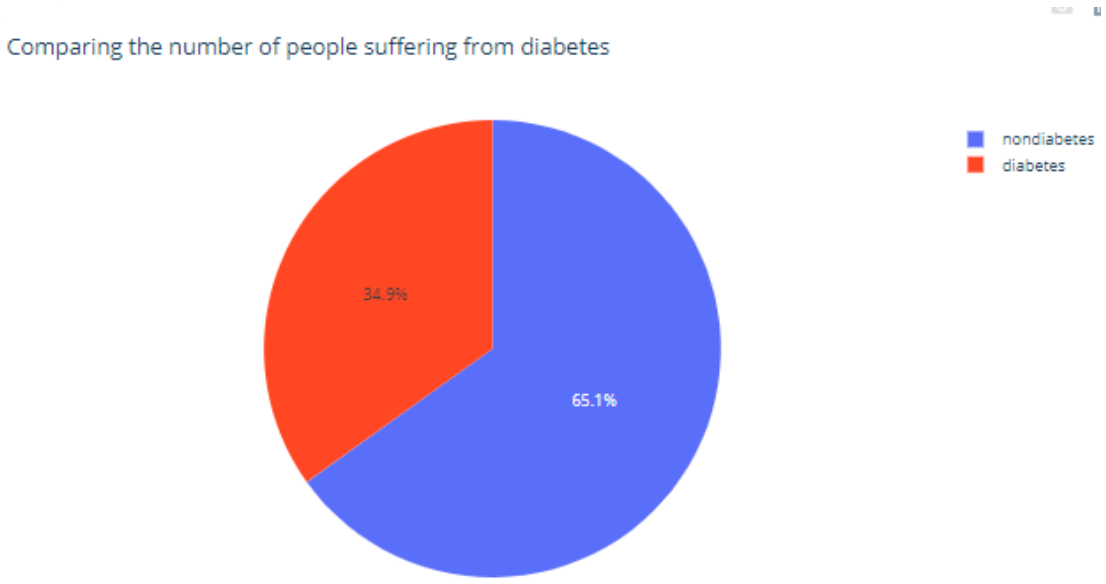


Plot Selection 2 - interactive pie Plot for diabetes:

what percent of people in dataset are diabetic?

According the below pie plot 65% of people have diabetes and 45 do not have diabetes. This plot is interactive so we can zoom it.

The interactive pie plot blend well to analysing these variables, since we can filter and move the plot around and hover over any point to see its exact values We can filter our plot by selecting the desired legend so that the plot is displayed only based on diabetic and non-diabetic.



For design this plot I use the value_counts()for Outcome variable to calculate the number of people who have diabetes or not. The name of pie plot is non-diabetic and diabetic and the value of it is the number of them.

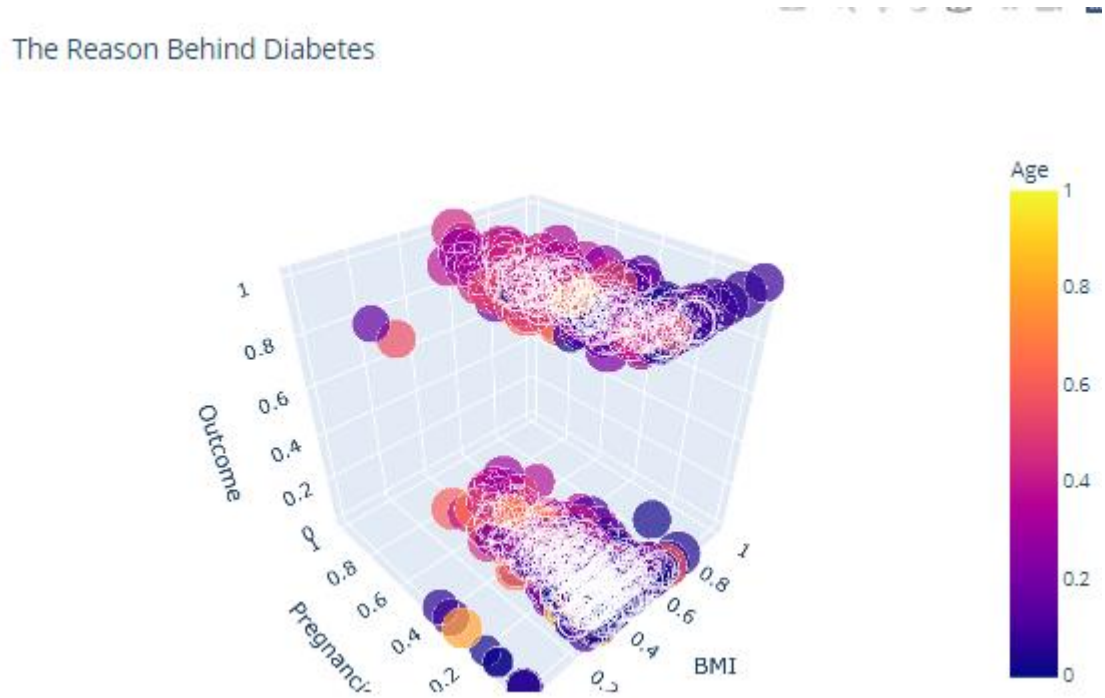
Plot Selection 3 - interactive 3D Scatter Plot (with colour and size)

how is the relationship between BMI, Pregnancies, Age, Glucose and Diabetes?

For the answering the second question, I decided to make an interactive 3D scatter plot (with colour and size as extra dimensions) for third visualization technique. This plot shows the relation between BMI, Pregnancies, Age, Glucose and diabetes. I considered BMI as x axe, Pregnancies as y axe, Diabetes as a z axe, Age as a colour and Glucose as a size. The interactivity and the 5D nature of the plot blend well to analysing these variables, as the user can freely move the plot around and hover over any point to see its exact values. The colour change from blue to yellow makes the data visually distinct and easy to parse.

- When we zoom plot can observe that the Number of times pregnant and BMI in diabetic people is more than non-diabetic ones. And by increasing these variables, the number of diabetic person boost. Therefore, the pregnant and BMI has a positive relationship with diabetes.

- the number of blue point in the non-diabetic persons (ponies which is in the bottom of picture) is more than diabetic people, since the younger people have diabetes less than older people. So, the age has positive relationship with diabetes.
- the number of large points in the diabetic points are more than the number of large point in the non-diabetic points. so, Plasma glucose have the positive relationship with diabetes.

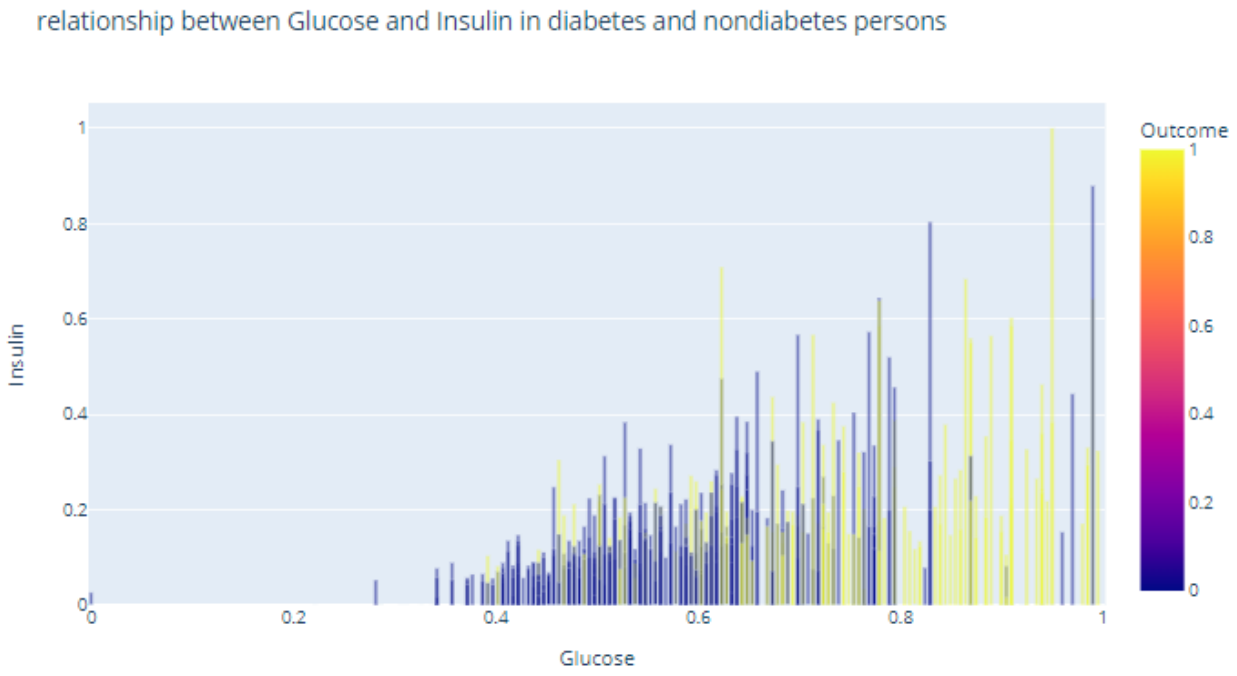


Plot Selection 4 - interactive Bar Plot

do a change in Glucose and Insulin variables have significant effect on Diabetes?

In order to answer the third question, I design an interactive bar plot with colour for forth visualization technique. This plot shows the effect of Glucose and Insulin variables on diabetes. I considered Glucose as x axe, Insulin as y axe, Diabetes as colour. The interactivity and the bar plot blend well to analysing these variables, as the user can freely move and zoom on it.

- As can be seen in the below plot by increasing Glucose the height of bar boost which mean is that Insulin increase.
- More importantly, we can see more height yellow bar are in the right side of the plot which shows that diabetic person has more Glucose and Insulin in their blood so any change in Glucose and Insulin variables have significant effect on Diabetes.



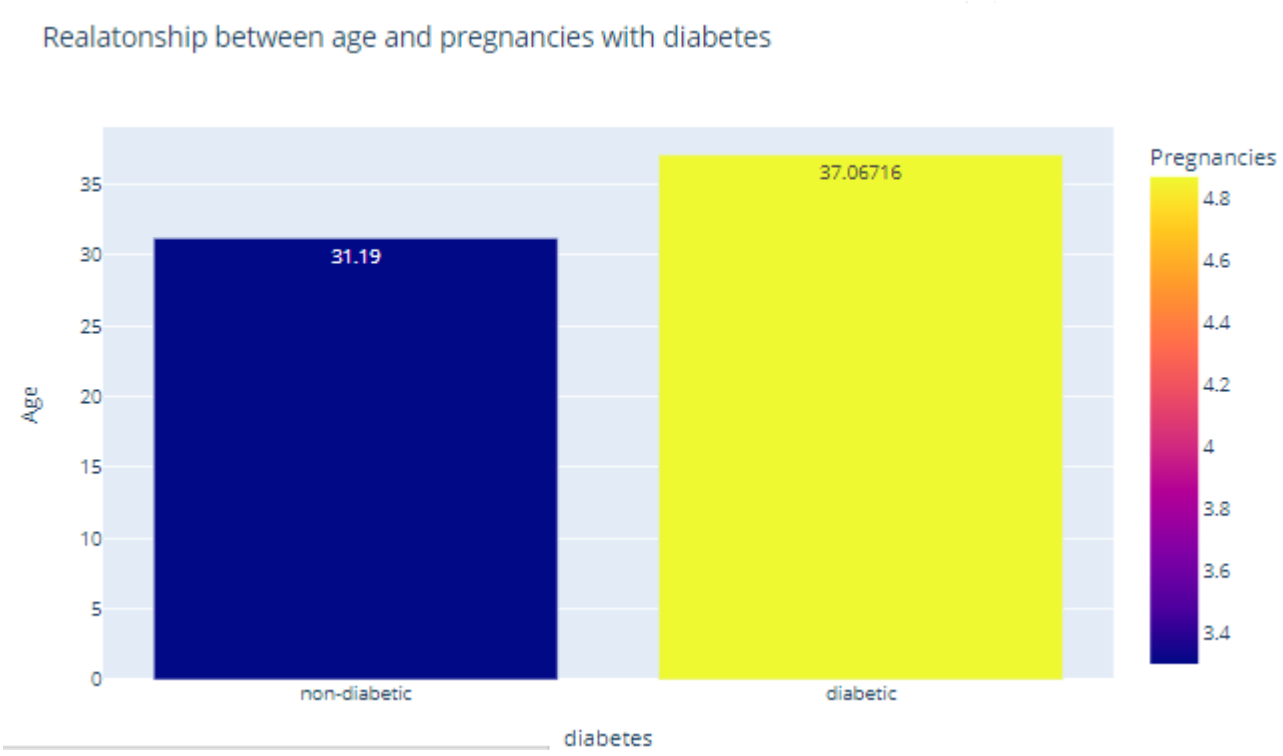
Plot 5 - interactive bar Plot

What is the average age and duration of pregnancy in diabetic patients?

for answering the forth question, I design an interactive bar plot with colour for fifth visualization technique. This plot shows the average age and pregnancy in diabetic. I

considered diabetes as x axe, Age as y axe, and pregnancy as colour. The interactivity and the bar plot blend well to analysing these variables, as the user can freely move and zoom on it.

As can be seen in bar plot the age mean of the diabetic people is more than non-diabetic ones. 37 years for diabetic and 31 years for non-diabetic people. Also the diabetic people have the higher amount of pregnancy period than non-diabetic since diabetic is yellow.



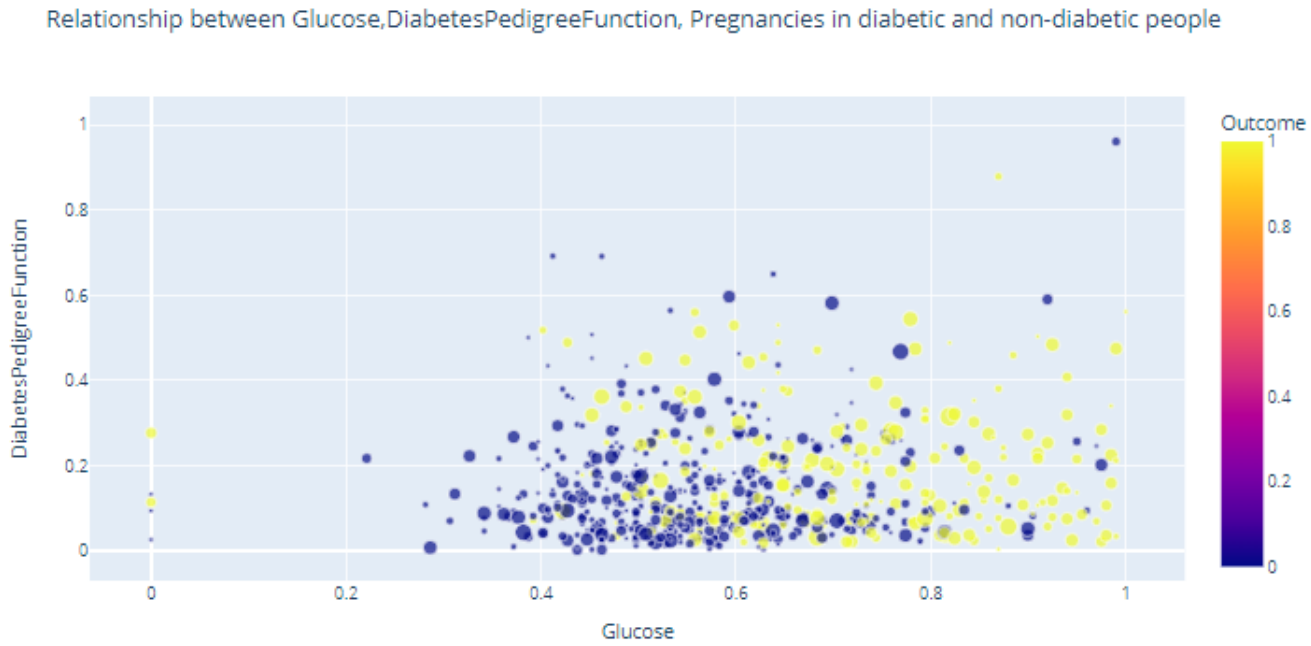
<u>Plot Selection 6 - Interactive 2-d scatter Plot with color and size :</u>

what do you notice from investigating difference of Glucose, Diabetes Pedigree Function, Pregnancies in diabetic and non-diabetic people?

I draw to make an interactive 2D scatter plot (with colour and size as extra dimensions) For answering the fifth question. This plot shows the relationship between Glucose, Diabetes Pedigree Function, Pregnancies in diabetic and non-diabetic people. I considered Glucose as x axe, Diabetes Pedigree Function as y axe, Diabetes as a colour and Pregnancies as a size. The

interactivity and the 4D nature of the plot blend well to analysing these variables, as the user can freely move the plot around and hover over any point to see its exact values.

- With the first look at the plot we sound that diabetic points with yellow colour have higher amount of Glucose and Diabetes Function. Also, as yellow points have the larger size diabetic person have more Pregnancies times. So all of them have the positive relationship with Diabetes.



----------END----------