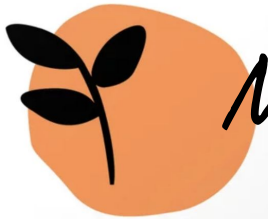


Mini Project 1

Yee Rin Lew
#2024-02-27-ds-pt-sg
4th May 2024

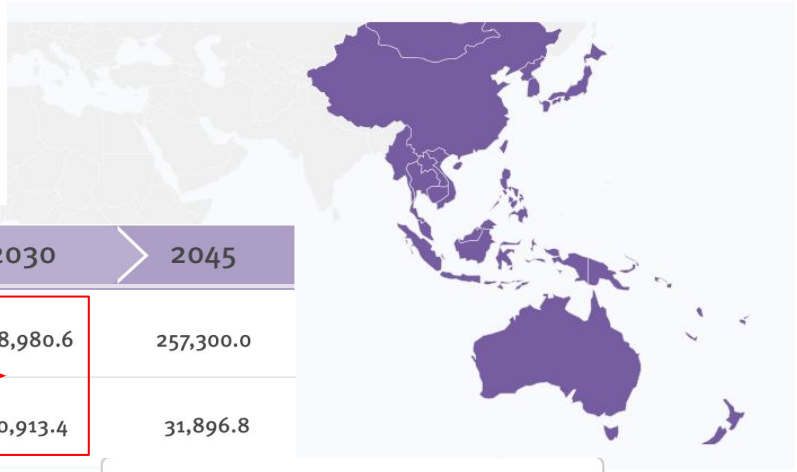
Agenda

- Background
- Objective
- Data Cleaning
- EDA
- Summary

Background

Western Pacific Region diabetes data

Roughly 163 million adults aged 20–79 years have diabetes in the IDF Western Pacific Region. This is the highest number of all IDF Regions and represents 35% of the world's total number of adults with diabetes in this age group.

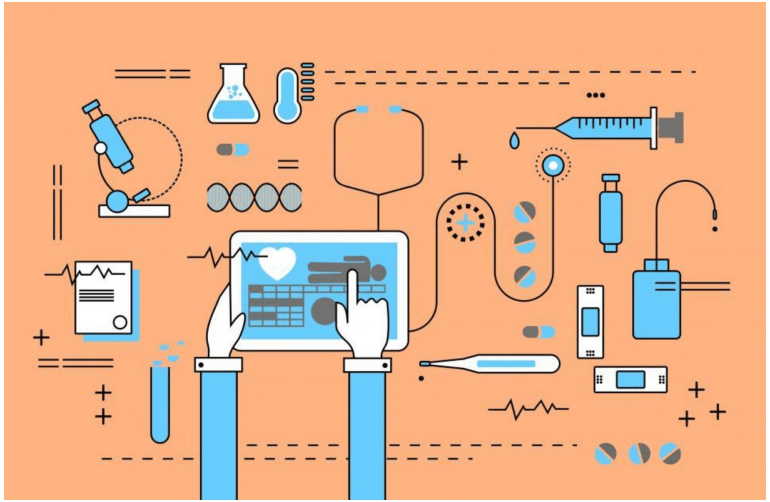


Diabetes-related health expenditure	2000	2011	2021	2030	2045
Total diabetes-related health expenditure, USD million	-	72,200.0	241,313.1	248,980.6	257,300.0
Diabetes-related health expenditure per person, USD	-	1,169.9	1,203.8	30,913.4	31,896.8

The highest number of deaths due to diabetes in 2019 occurred in the Western Pacific Region – well over 1 million.

World Diabetes Foundation. (n.d.). Western Pacific. In Diabetes Atlas. Retrieved from https://diabetesatlas.org/data/en/region/8/wp.html

Objective



Stakeholders:

1. Ministry of Health(MOH)
2. Healthcare Professional and Provider
3. Medical technological device manufacturers
4. Insurance companies

Problem Statements:

1. What are the risk factors that contribute to Diabetes Mellitus?
2. What are the relationships between these risk factor?

Solution:

Conducting a thorough analysis of the provided dataset to investigate whether there are statistically significant differences in various health parameters between diabetic and non-diabetic patients, identify key insights that can inform diagnostic criteria, treatment strategies, and preventive measures for diabetes management.

Process, Workflow, & Tools

Data Source	National Institute of Diabetes and Digestive and Kidney Diseases kaggle Data Shape:(768, 9)
Data Cleaning	  pandas
EDA	matplotlib pandas   seaborn 
Reporting	jupyter 

Data Cleaning

Null Value

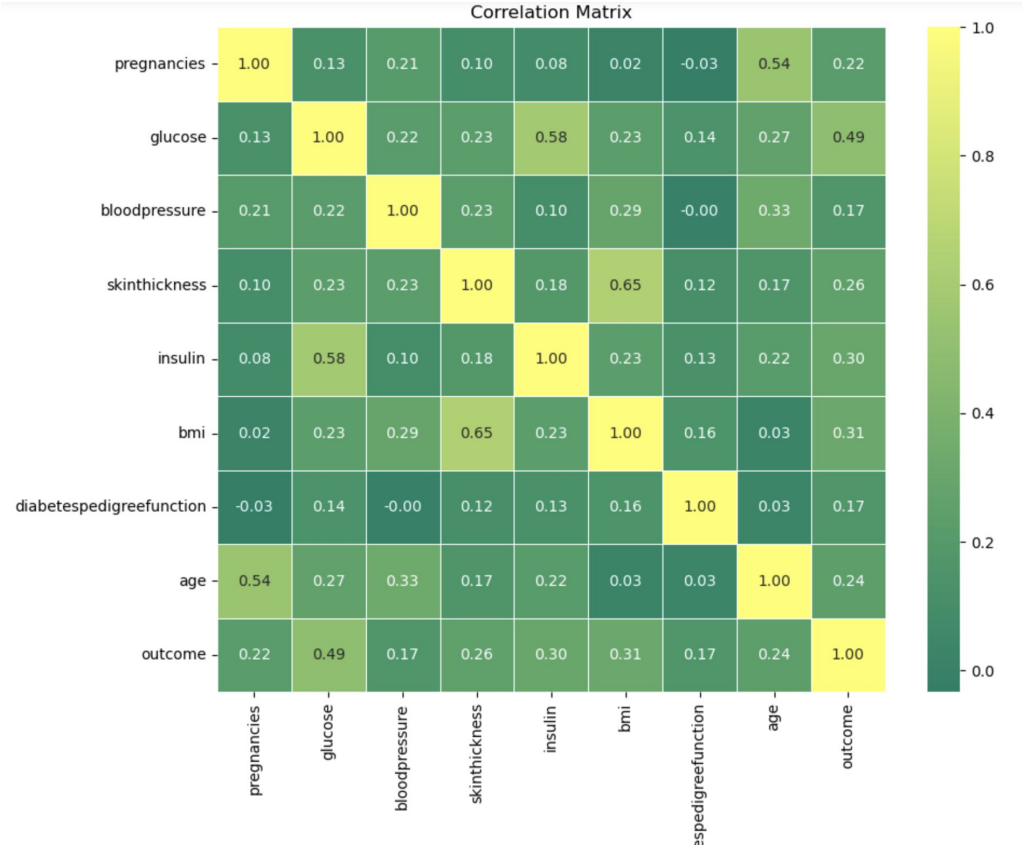
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 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
```

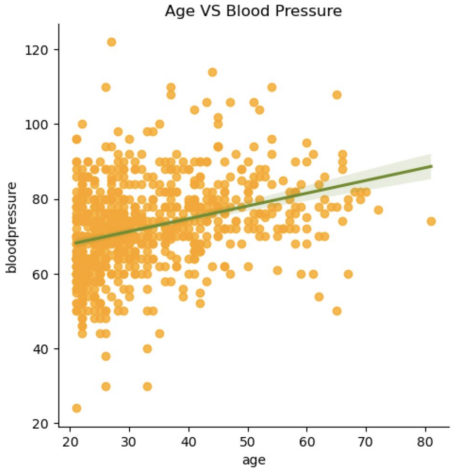
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	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

```
# real null values counts
df.isnull().sum()
```

```
pregnancies                 0
glucose                     5
bloodpressure              35
skinthickness             227
insulin                   374
bmi                        11
diabetespedigreefunction    0
age                         0
outcome                     0
dtype: int64
```

Null Value in ['bloodpressure']



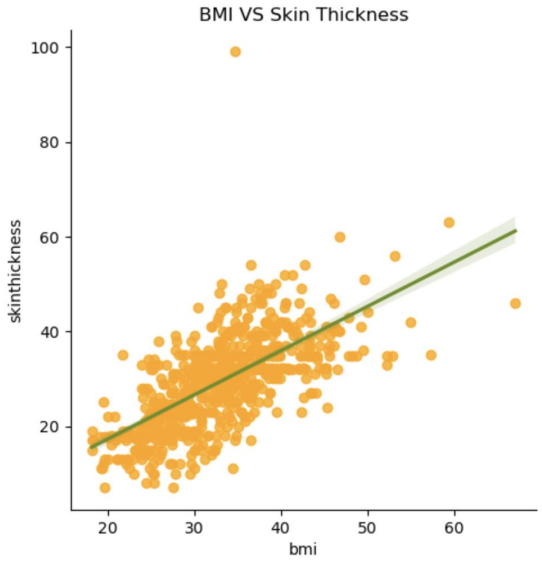



```
age_class  outcome
20-39      0          69.397403
           1          72.482993
40-59      0          76.702703
           1          79.145833
60+        0          77.181818
           1          80.888889
Name: bloodpressure, dtype: float64
```

Null Value in ['bmi', 'insulin', 'glucose']

```
df['X'] = df['X'].fillna(df.groupby('outcome')['X'].transform('median'))
```

Null Value in ['skinthickness']

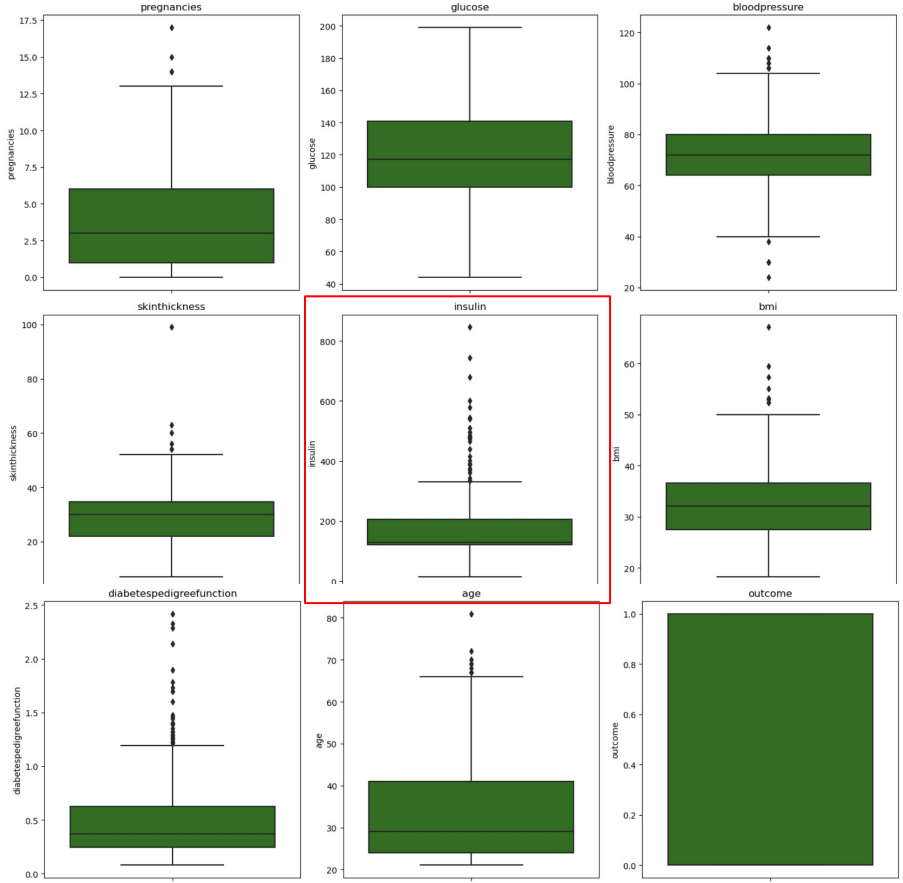


National Center for Biotechnology Information. (n.d.). *StatPearls [Internet]*.. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK541070/

- Underweight - BMI under 18.5 kg/m^2
- Normal weight - BMI greater than or equal to 18.5 to 24.9 kg/m^2
- Overweight – BMI greater than or equal to 25 to 29.9 kg/m^2
- Obesity – BMI greater than or equal to 30 kg/m^2

```
outcome  bmi_class
0        underweight    17.000000
         healthy        17.689655
         overweight     22.830000
         obese          32.283582
1        underweight          NaN
         healthy        15.000000
         overweight     24.666667
         obese          34.728477
Name: skinthickness, dtype: float64
```

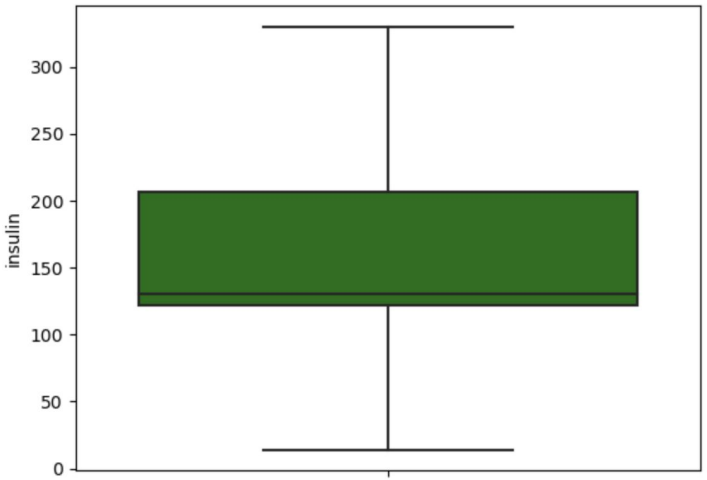
Outliers



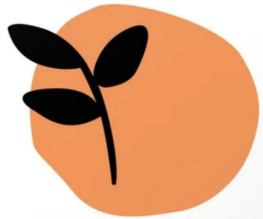

```python
# Replace 'insulin' upper outliers with median

Q1 = df.insulin.quantile(0.25)
Q3 = df.insulin.quantile(0.75)
IQR = Q3-Q1
upper = Q3+1.5*IQR
lower = Q1-1.5*IQR

median_value = df['insulin'].median()
insulin_outlier = df[df['insulin']>upper]
df.loc[insulin_outlier.index, 'insulin'] = median_value
```

```python
sns.boxplot(y = df['insulin'], color = 'green')
```

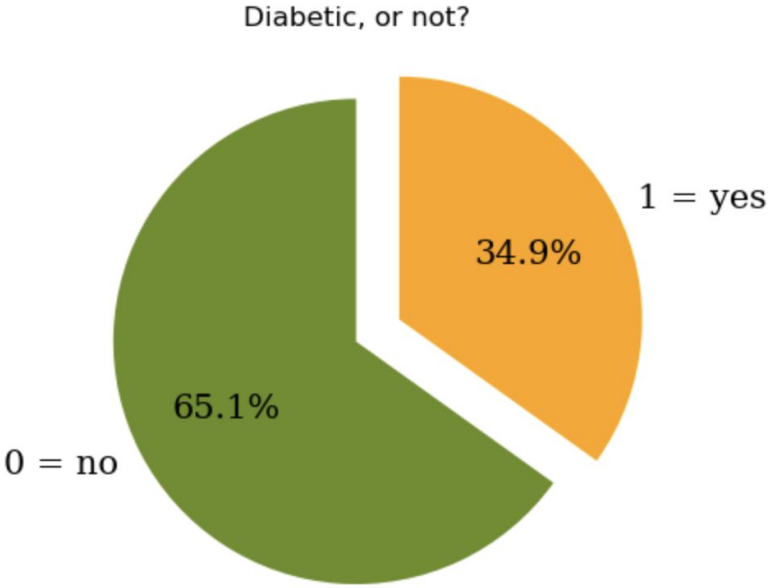
```
<Axes: ylabel='insulin'>
```

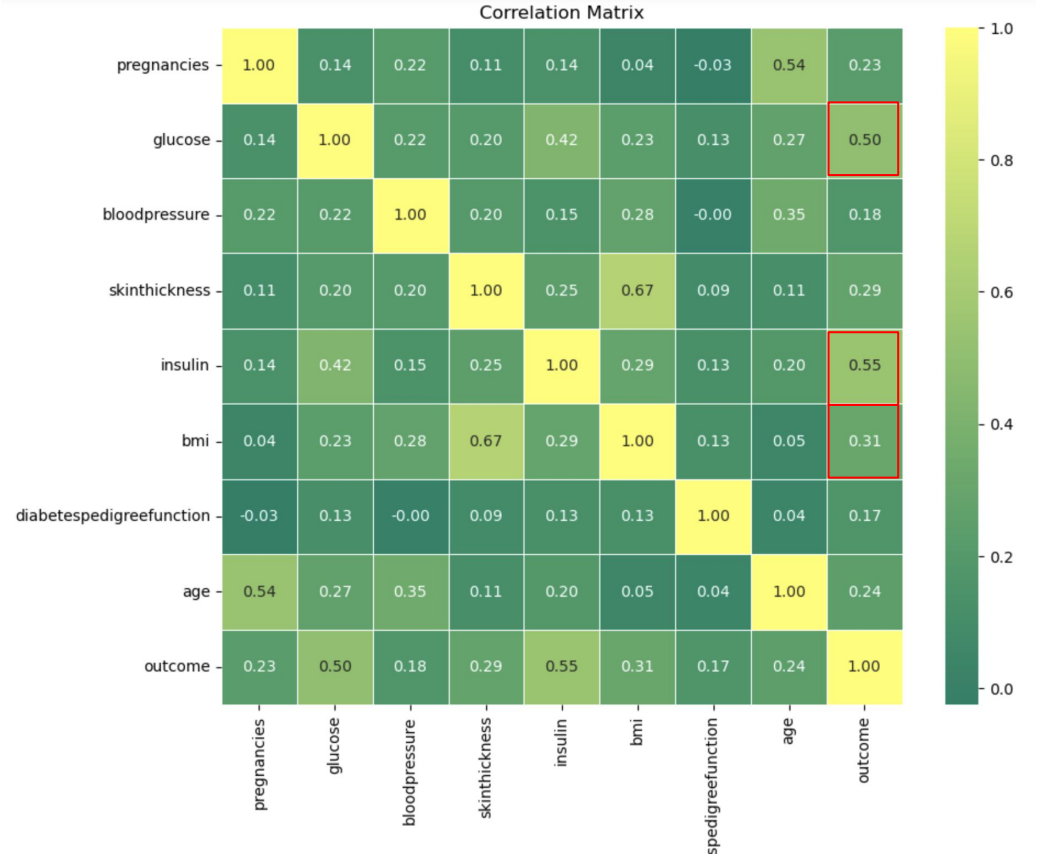


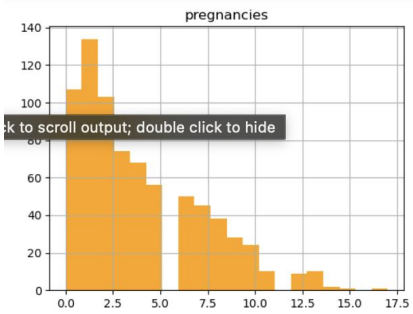
*Outliers for 'bmi' only 8 of them, decided to just drop

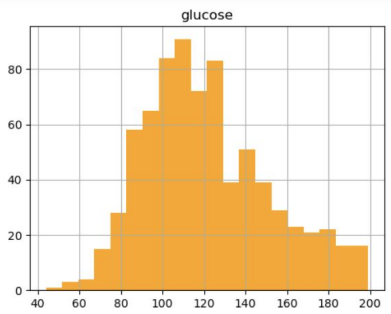


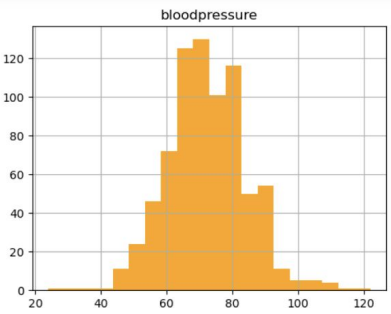
EDA

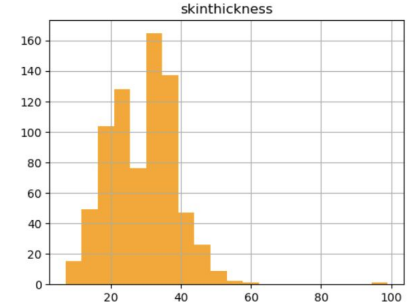

Diabetic, or not?
1 = yes
34.9%
65.1%
0 = no

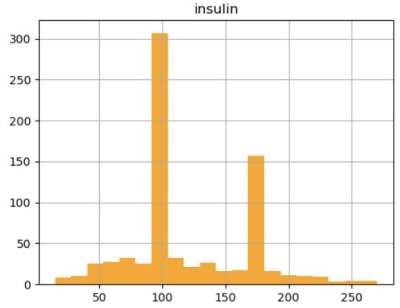

Correlation Matrix
pregnancies 1.00 0.14 0.22 0.11 0.14 0.04 -0.03 0.54 0.23
glucose 0.14 1.00 0.22 0.20 0.42 0.23 0.13 0.27 0.50
bloodpressure 0.22 0.22 1.00 0.20 0.15 0.28 -0.00 0.35 0.18
skinthickness 0.11 0.20 0.20 1.00 0.25 0.67 0.09 0.11 0.29
insulin 0.14 0.42 0.15 0.25 1.00 0.29 0.13 0.20 0.55
bmi 0.04 0.23 0.28 0.67 0.29 1.00 0.13 0.05 0.31
diabetespedigreefunction -0.03 0.13 -0.00 0.09 0.13 0.13 1.00 0.04 0.17
age 0.54 0.27 0.35 0.11 0.20 0.05 0.04 1.00 0.24
outcome 0.23 0.50 0.18 0.29 0.55 0.31 0.17 0.24 1.00
pregnancies
glucose
bloodpressure
skinthickness
insulin
bmi
spedigreefunction
age
outcome
1.0
0.8
0.6
0.4
0.2
0.0

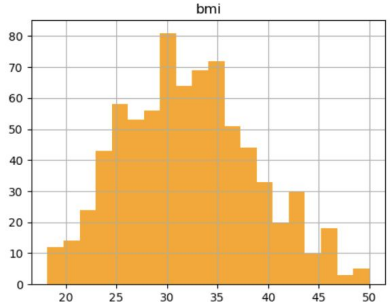

pregnancies
ck to scroll output; double click to hide

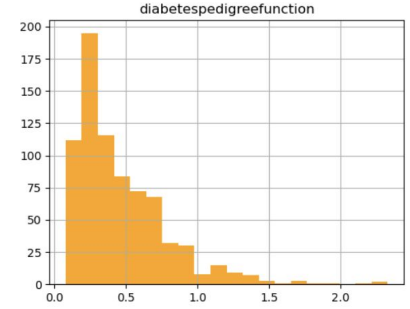

glucose

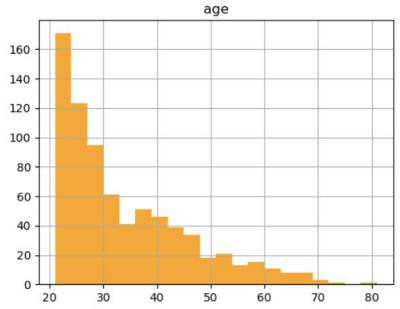

bloodpressure

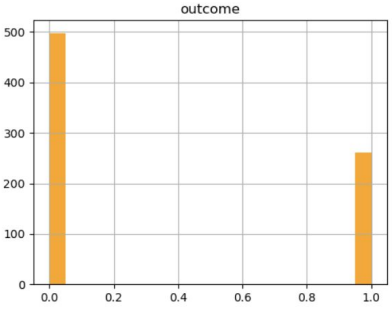

skinthickness


insulin

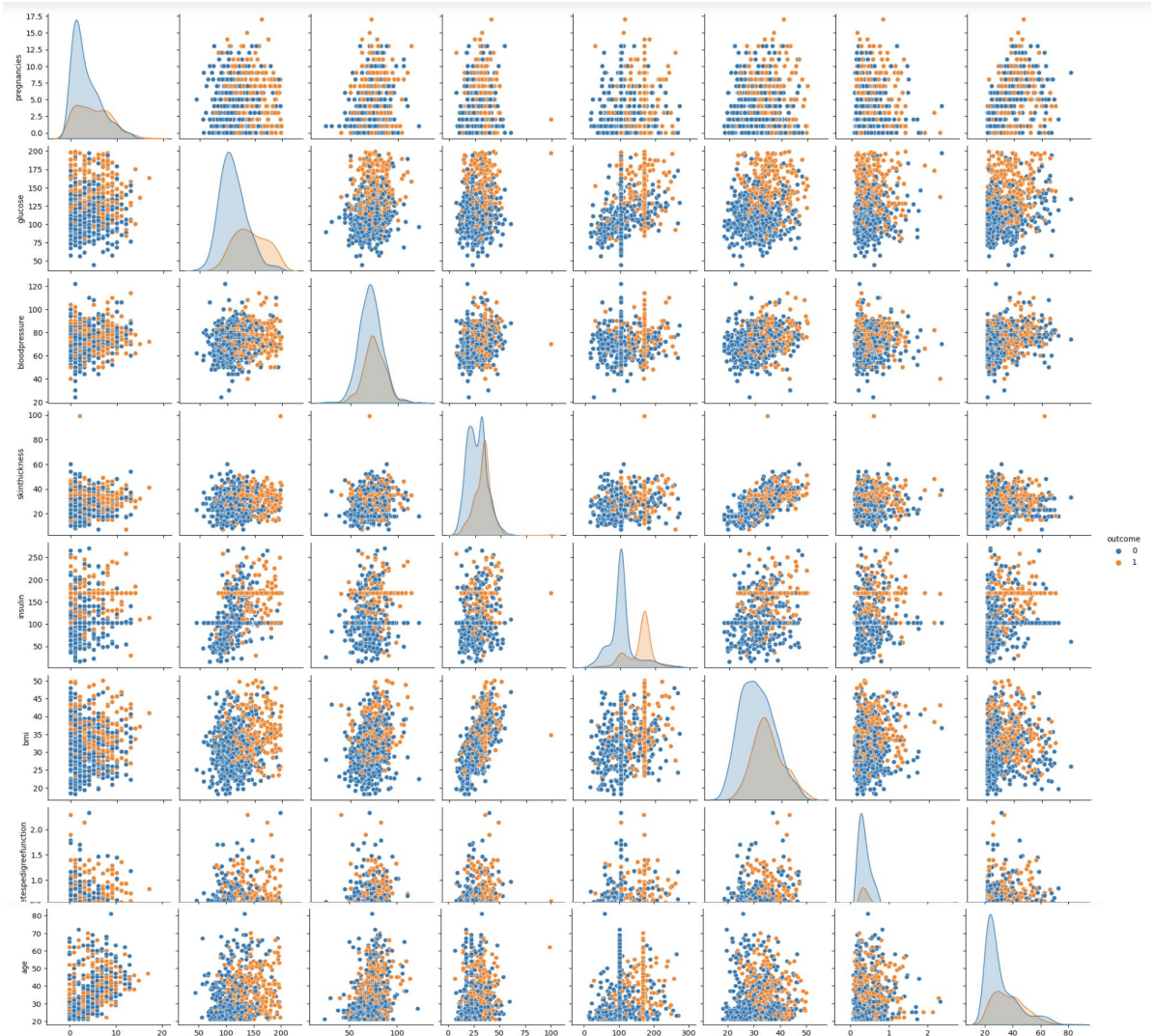

bmi


diabetespedigreefunction


age


outcome


Distribution

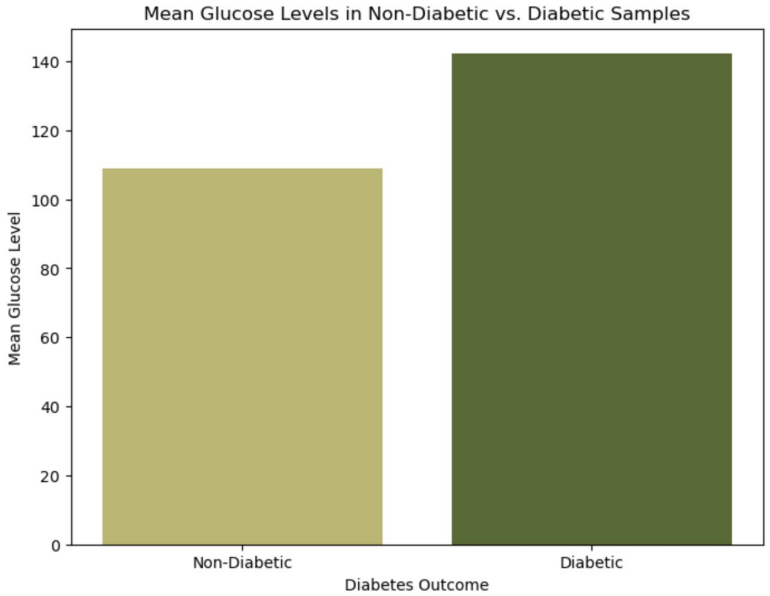

pregnancies
glucose
bloodpressure
skinthickness
insulin
bmi
diabetespedigreefunction
age
outcome
0
1

Pairplot

Question 1: Is there significant difference in mean glucose levels between diabetic & non diabetic patients?

Null Hypothesis (H0): There is no difference in the mean glucose level between diabetic and non-diabetic patient.

Alternative Hypothesis (HA): There is a difference in the mean glucose level between diabetic and non-diabetic patient.



t-test

```
#statistic
N = 200
a = nondm['glucose']
b = dm['glucose']

#set Alpha
alpha = 0.05
```

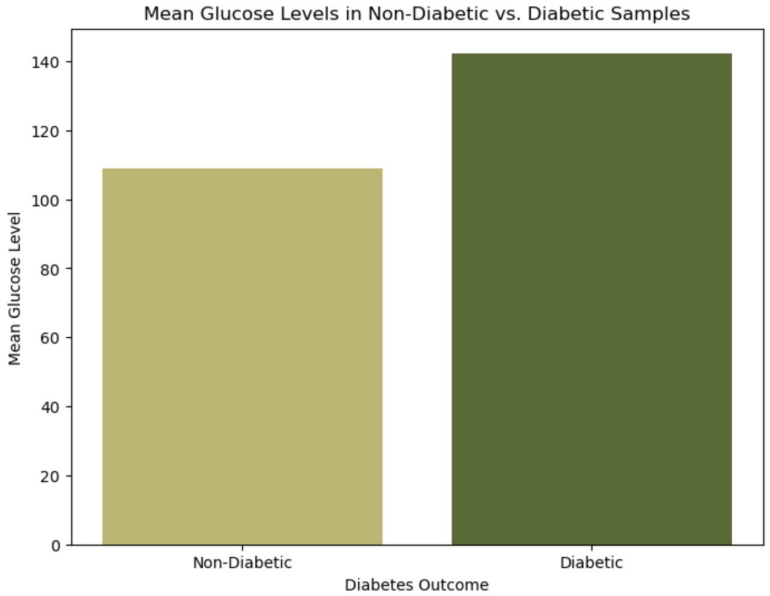
```
t, p = stats.ttest_ind(a,b)
print("t = " + str(t))
print("p = " + str(p))
```

```
t = -12.264208140098336
p = 1.494339909886565e-29
```

Question 1: Is there significant difference in mean glucose levels between diabetic & non diabetic patients?

~~Null Hypothesis (H0): There is no difference in the mean glucose level between diabetic and non-diabetic patient.~~

Alternative Hypothesis (HA): There is a difference in the mean glucose level between diabetic and non-diabetic patient.



t-test

```
1 #statistic
2 N = 200
3 a = nondm['glucose']
4 b = dm['glucose']
5
6 #set Alpha
7 alpha = 0.05
```

```
1 t, p = stats.ttest_ind(a,b)
2 print("t = " + str(t))
3 print("p = " + str(p))
```

```
t = -12.264208140098336
p = 1.494339909886565e-29
```

Question 2: Is there association between BMI classes and diabetes status?

Null Hypothesis (H0): There is no association between BMI class and diabetes status.

Alternative Hypothesis (HA): There is association between BMI class and diabetes status.

```
Chi-square statistic: 74.9084804674606
P-value: 3.790718167139003e-16
Degrees of freedom: 3
Expected frequencies:
[[  2.60416667   1.39583333]
 [ 66.40625     35.59375   ]
 [116.53645833  62.46354167]
 [314.453125   168.546875  ]]
```

Question 2: Is there association between BMI classes and diabetes status?

~~Null Hypothesis (H0): There is no association between BMI class and diabetes status.~~

Alternative Hypothesis (HA): There is association between BMI class and diabetes status.

```
Chi-square statistic: 74.9084804674606
P-value: 3.790718167139003e-16
Degrees of freedom: 3
Expected frequencies:
[[  2.60416667   1.39583333]
 [ 66.40625     35.59375   ]
 [116.53645833  62.46354167]
 [314.453125   168.546875  ]]
```

alpha = 0.05

Question 3: Is there a significant difference in terms of insulin levels between diabetic and non-diabetic patients?

Null Hypothesis (H0): There is no significant difference in terms of insulin levels between diabetic and non-diabetic patients

Alternative Hypothesis (HA): There is a significant difference in terms of insulin level between diabetic and non-diabetic patients

```
#statistic
N = 200
a1= nondm['insulin']
b1 = dm['insulin']

#set Alpha
alpha = 0.05
```

alpha = 0.05

```
#t-testing
t2, p2 = stats.ttest_ind(a1,b1)
print("t = " + str(t2))
print("p = " + str(p2))
```

```
t = -13.39852499060977
p = 4.7632087173222495e-34
```

Question 3: Is there a significant difference in terms of insulin levels between diabetic and non-diabetic patients?

~~Null Hypothesis (H0): There is no significant difference in terms of insulin levels between diabetic and non-diabetic patients~~

Alternative Hypothesis (HA): There is a significant difference in terms of insulin level between diabetic and non-diabetic patients

```
#statistic
N = 200
a1= nondm['insulin']
b1 = dm['insulin']

#set Alpha
alpha = 0.05
```

alpha = 0.05

```
#t-testing
t2, p2 = stats.ttest_ind(a1,b1)
print("t = " + str(t2))
print("p = " + str(p2))
```

```
t = -13.39852499060977
p = 4.7632087173222495e-34
```

Summary

Based on the observed differences in glucose levels, BMI, and insulin levels between diabetic and non-diabetic patients, it is reasonable to consider these factors as risk factors which can inform diagnostic criteria, treatment strategies, and preventive measures for diabetes management, while allocate medical resource accordingly and appropriately.

Limitation

Diabetes is a complex and multifactorial disease influenced by a combination of genetic, lifestyle, and environmental factors, so a comprehensive approach to risk assessment and prevention is warranted.

Reference

National Center for Biotechnology Information. (n.d.). *StatPearls [Internet]*.. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK541070/

World Diabetes Foundation. (n.d.). Western Pacific. In *Diabetes Atlas*. Retrieved from https://diabetesatlas.org/data/en/region/8/wp.html

Thank You!