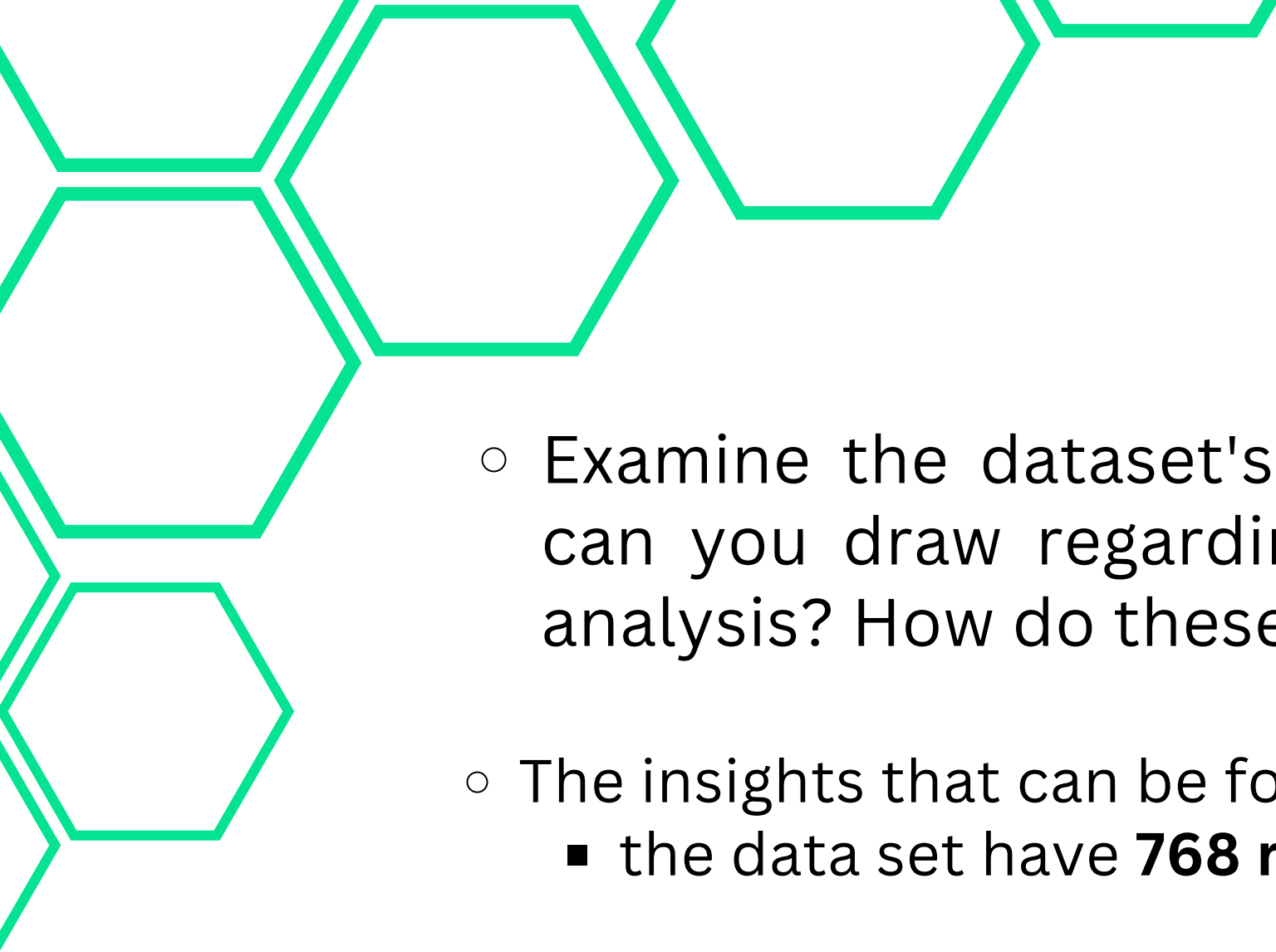

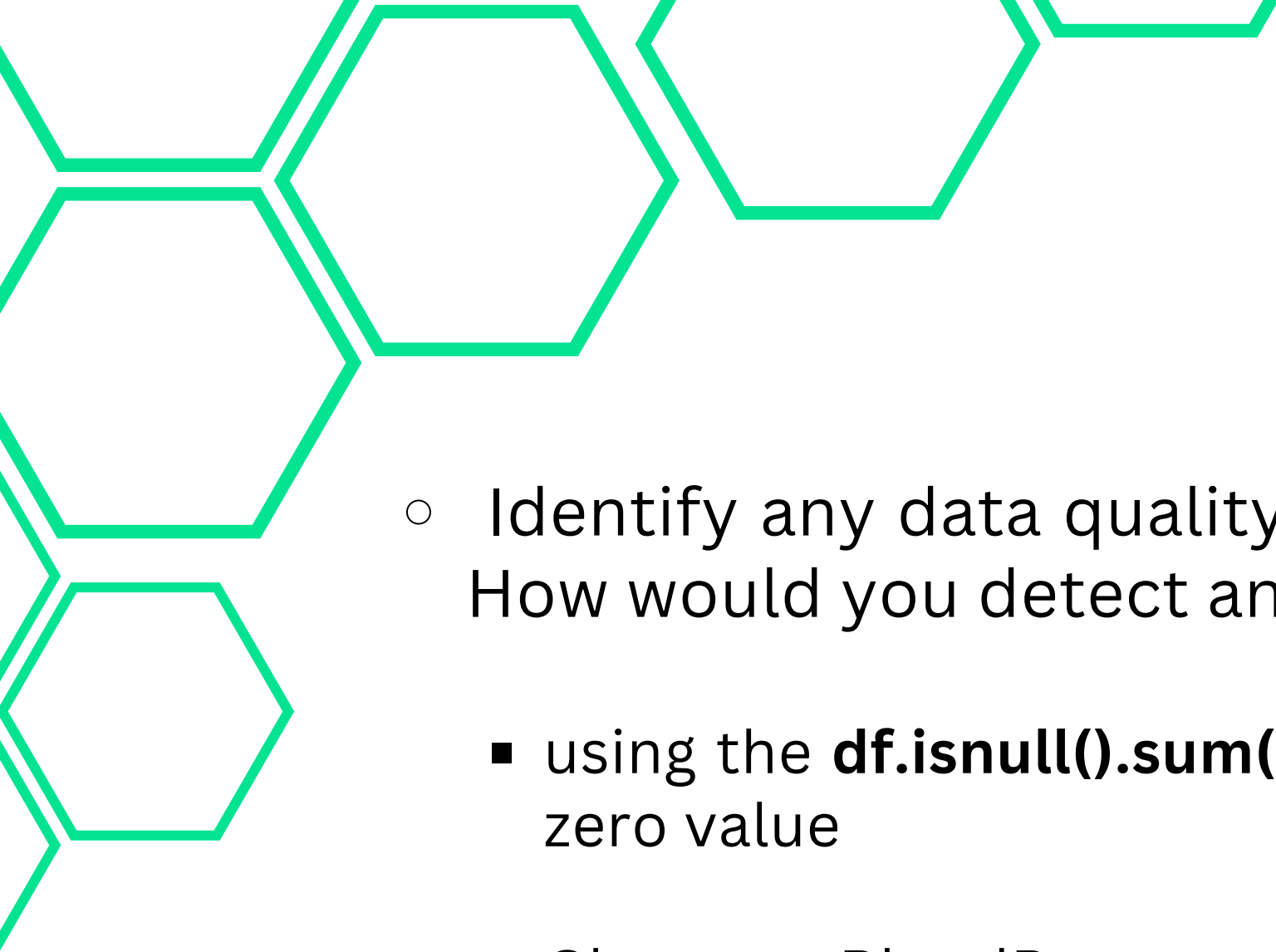





EXPLORATORY DATA ANALYSIS

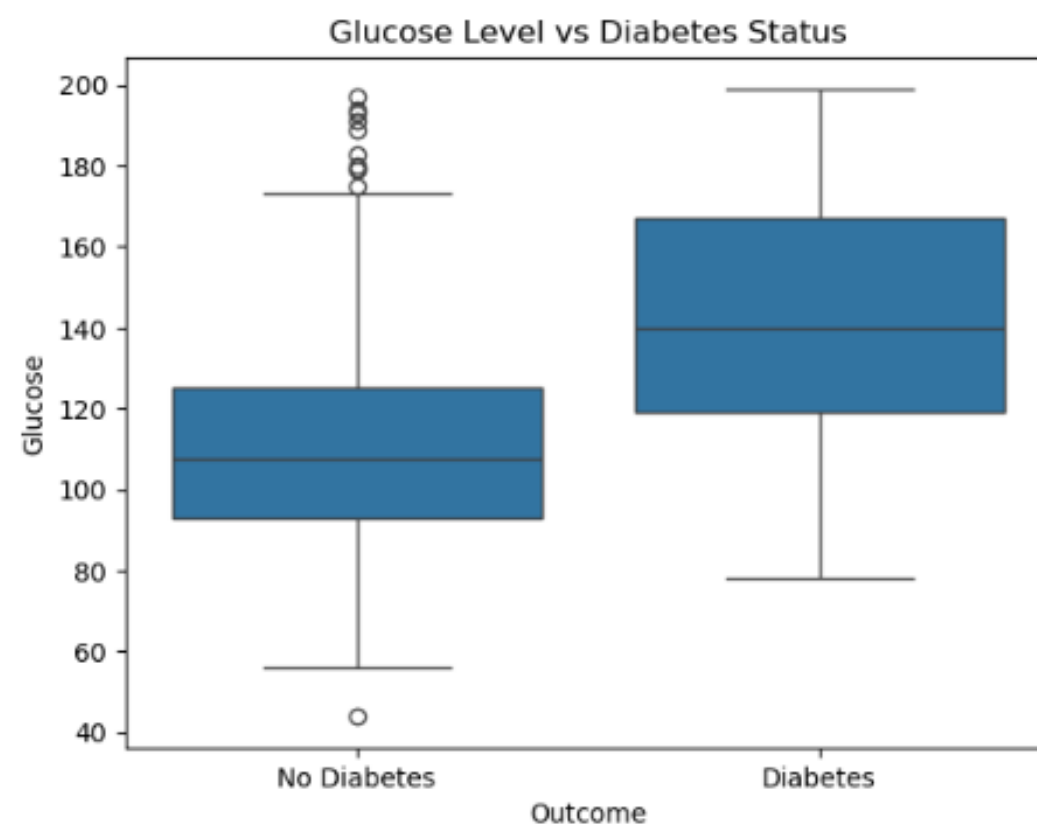
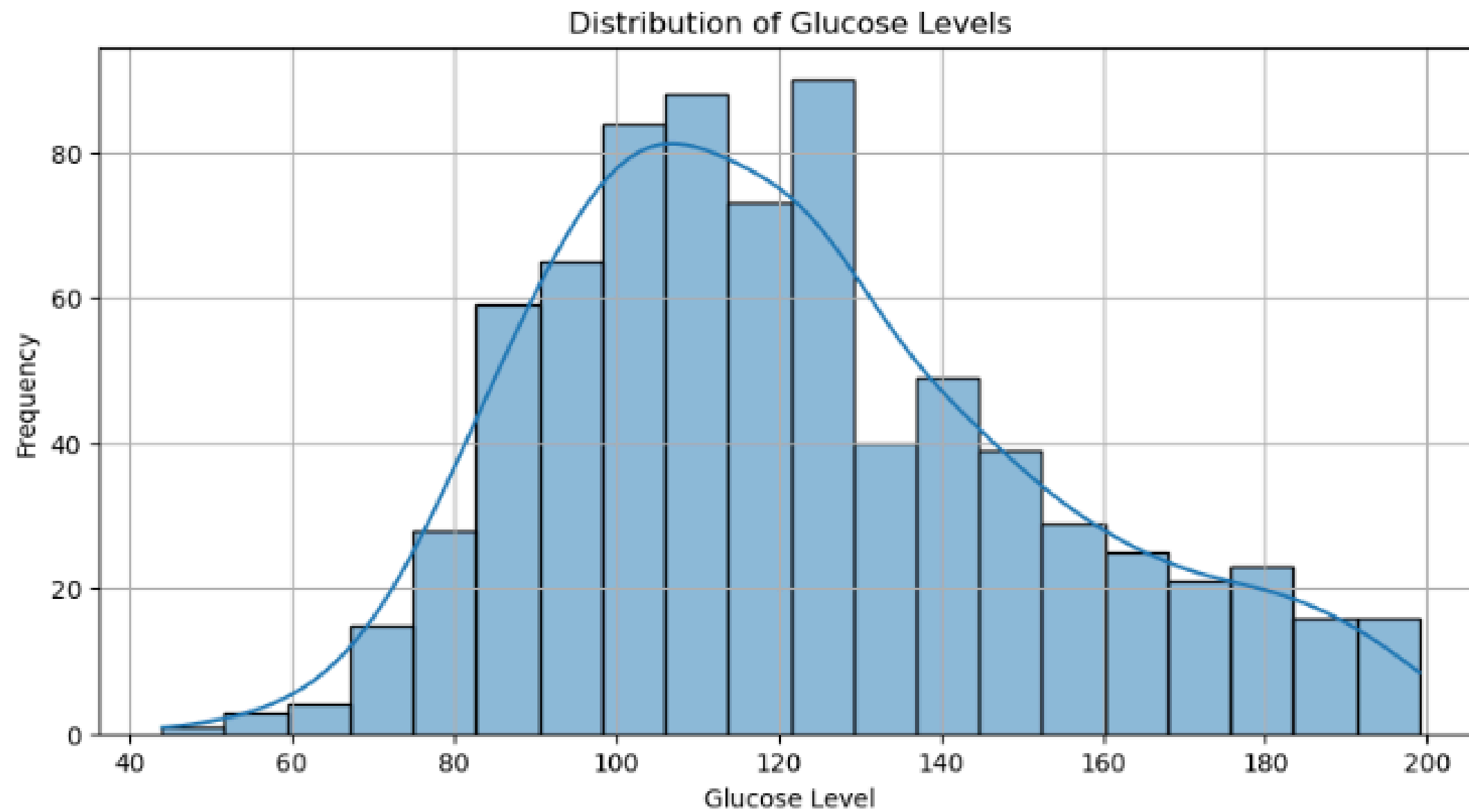


- 
- Examine the dataset's size, structure, and data types. What insights can you draw regarding its completeness, scope, and readiness for analysis? How do these datatypes influence preprocessing?
 - The insights that can be found are the following:
 - the data set have **768 rows/records**
 - **9 columns** - 'Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'
 - all data have **numerical in int and float value**
 - dataset is manageable and the data is **all numeric so it can be easy to do statistical analysis**
- 

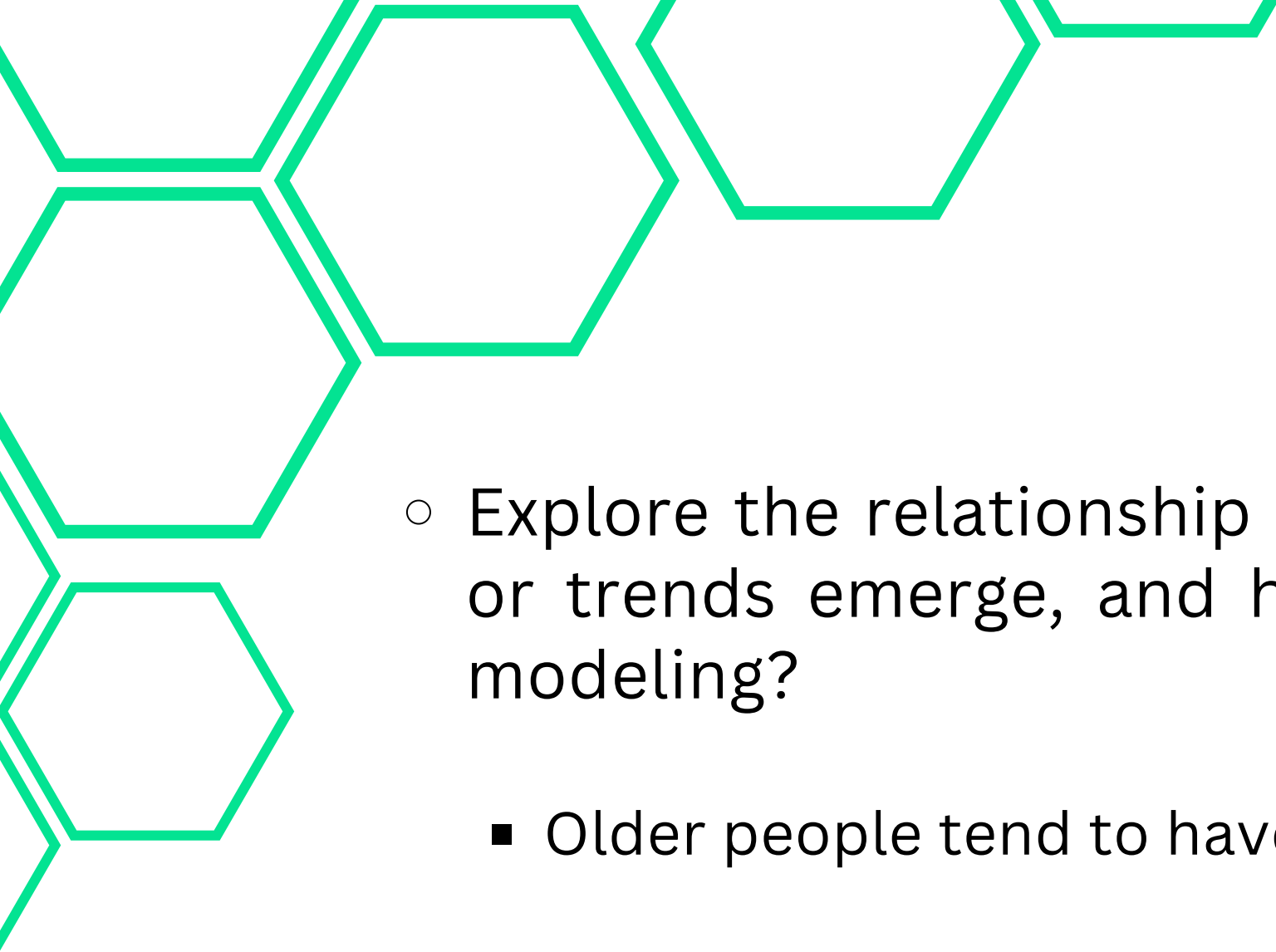
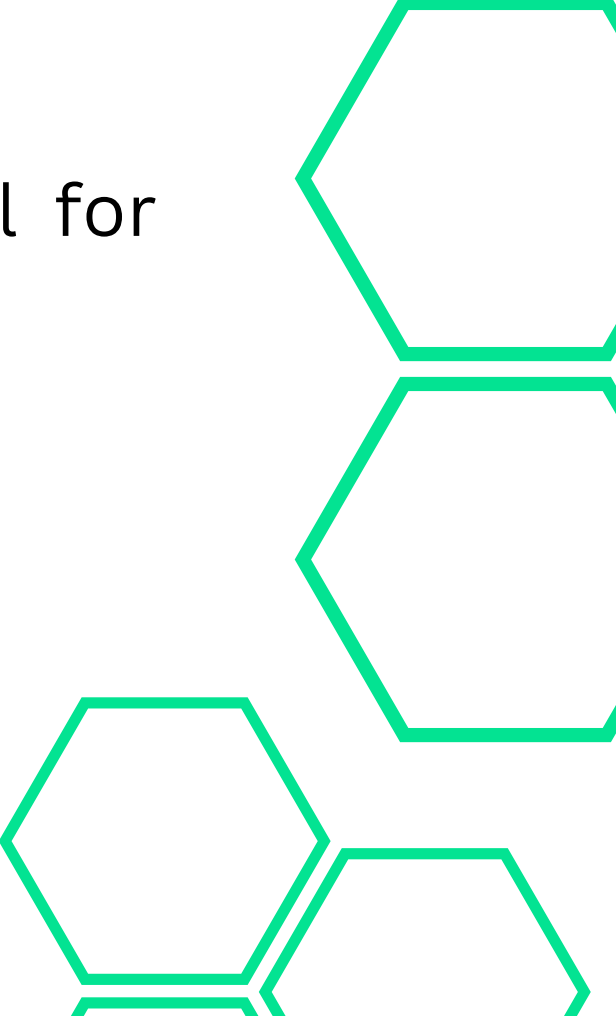
- 
- Identify any data quality issues, such as missing or inconsistent values. How would you detect and handle these issues in this dataset?
 - using the **df.isnull().sum()** but in this case it doesn't contain null values, but a zero value
 - Glucose , BloodPressure , SkinThickness , Insulin , and BMI have zero values that don't make sense medically.
 - replaced the zero values with NAN using:
 - **df[cols_with_zeros].replace(0, np.nan)**
 - **df.fillna(df.mean(), inplace=True)** to fill the null values with the mean of each column
- 

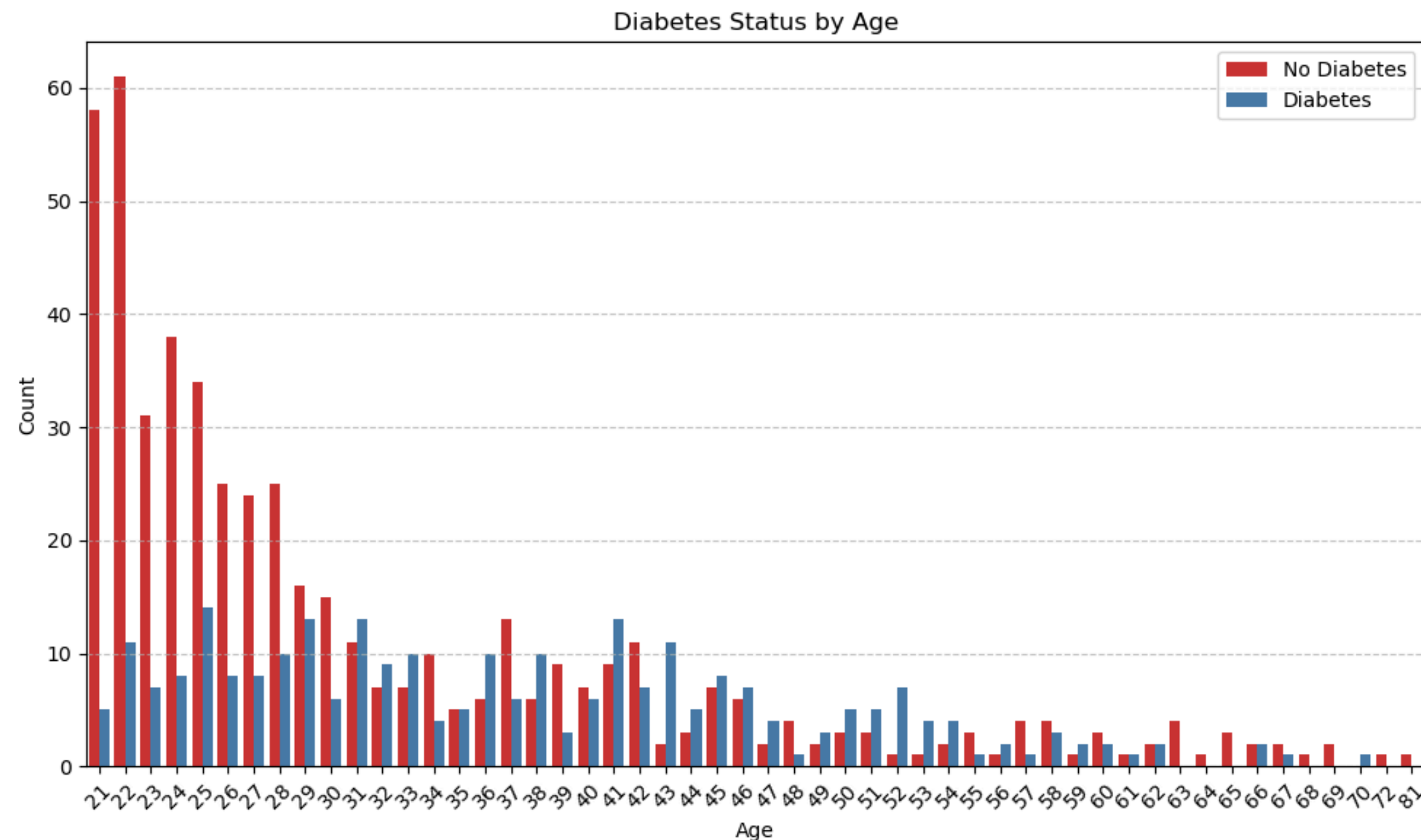
- Generate and interpret summary statistics for all features/input variables. What can you infer from their statistical properties?
 - Pregnancies : average women have about **3 pregnancies**.
 - Glucose : average of **121** (higher than normal that can be a risk to have diabetes)
 - BMI : average of **32** (classified as in obese range)

	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	121.686763	72.405184	29.153420	155.548223	32.457464	0.471876	33.240885	0.348958
std	3.369578	30.435949	12.096346	8.790942	85.021108	6.875151	0.331329	11.760232	0.476951
min	0.000000	44.000000	24.000000	7.000000	14.000000	18.200000	0.078000	21.000000	0.000000
25%	1.000000	99.750000	64.000000	25.000000	121.500000	27.500000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.202592	29.153420	155.548223	32.400000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	155.548223	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



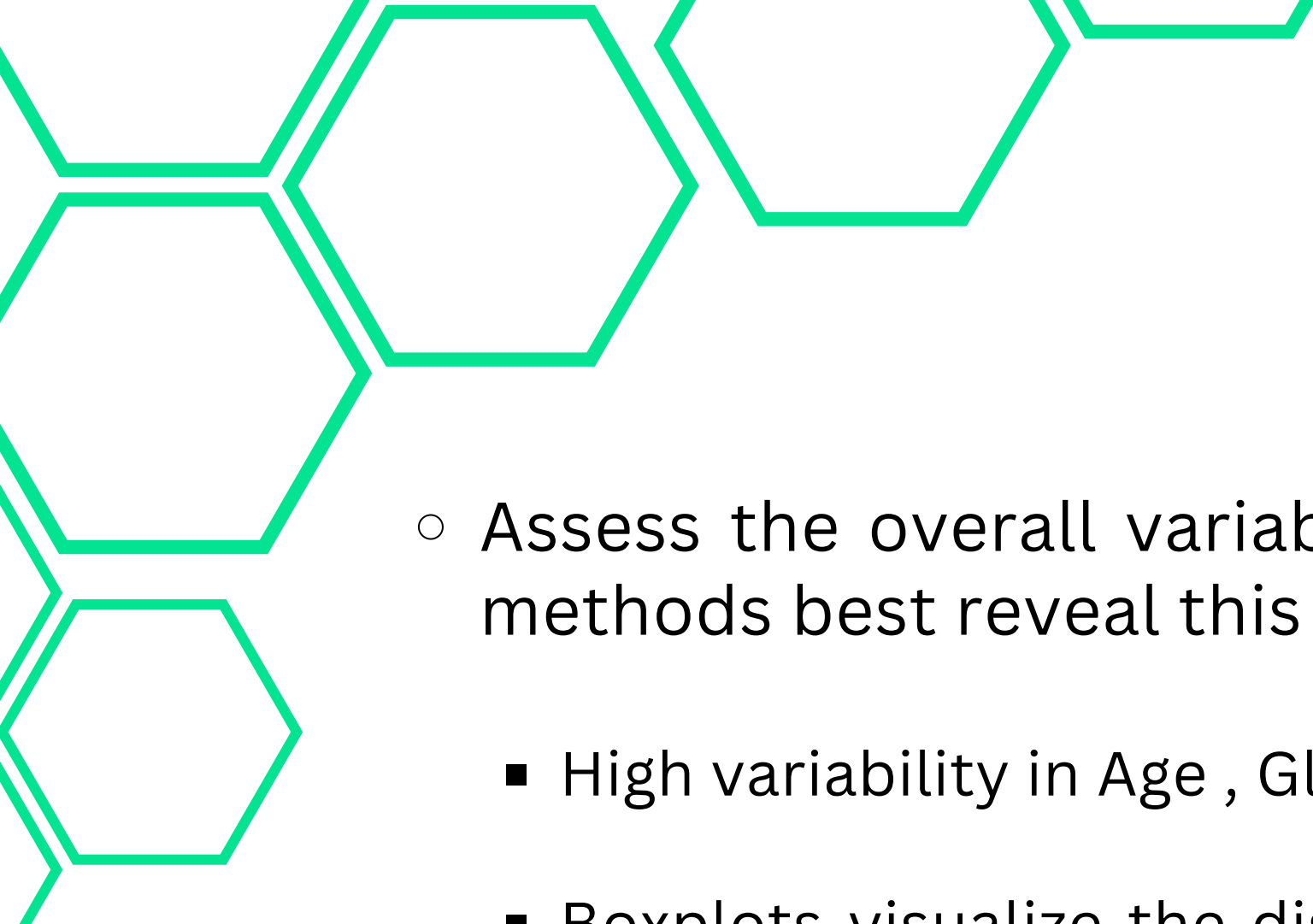

- Examine the distribution of a key variable related to diabetes risk. What insights can be drawn from the distribution using visual or statistical methods?
- Glucose - based on the graph show most people have moderate glucose levels, but some have very high levels.

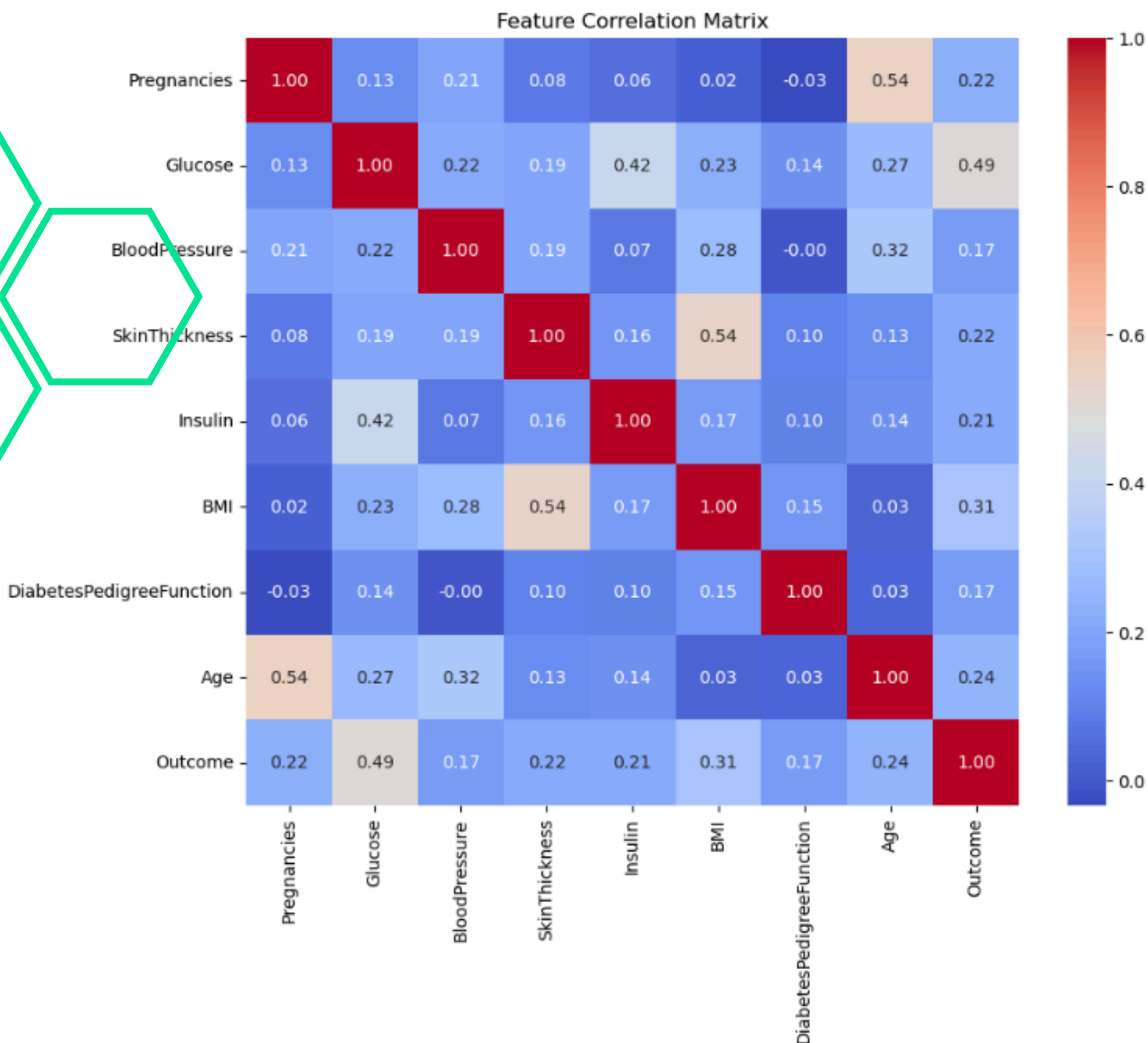
- 
- Explore the relationship between two or more variables. What patterns or trends emerge, and how might these insights influence analysis or modeling?
 - Older people tend to have more pregnancies.
 - **Higher BMI + higher glucose** = both **increase diabetes risk.**
 - With this relationship it help us choose features that are most useful for predicting diabetes.
- 



Analyze the relationship between age and diabetes status. What does the data suggest about this relationship?

- for the relationship between age and diabetes status:
 - people who have diabetes are generally older than those who don't.
- this conclude **age is a factor in predicting diabetes : older = higher risk.**

- 
- Assess the overall variability in the dataset. Which statistical or visual methods best reveal this variability?
 - High variability in Age , Glucose , and BMI
 - Boxplots visualize the distribution and spread of a dataset, especially when comparing data between different groups
- 



- Explore correlations or dependencies among features. How might these influence feature selection or modeling?
- **Glucose and Outcome** (diabetes yes/no) are moderately correlated – good for prediction.
- **BMI and Age** are somewhat related.



THANK YOU

