Group Members:

Daniyal Shahid (20L-2095)

Anzla Aslam(20L-2109)

Aqsa Arshad(20L-1166)

ASSIGNMENT NO 3

Exploratory Data Analysis Report

1)Customer ID

Summary Statistics:

Distinct Values: 2335

Distinct Values (%): 100.0%

Missing Values: 0

Missing Values (%): 0.0%

Memory Size: 18.4 KiB

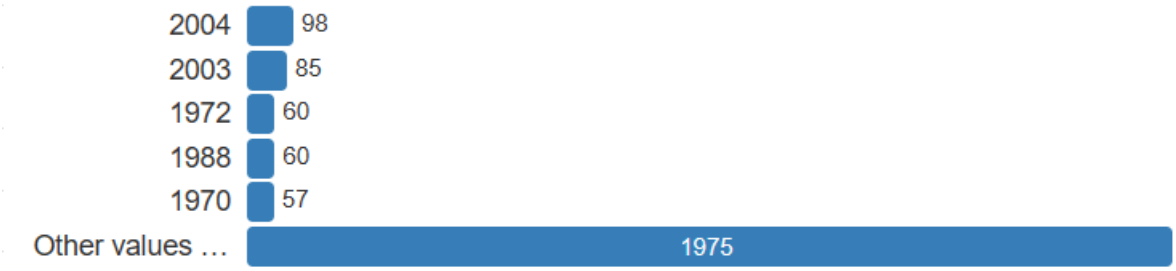
Insights and Observations:

- The Customer ID variable is categorical and represents unique customer identification numbers.
- There are no missing values in this variable.
- The data appears to follow a consistent format with the characters "I" and "d" followed by numeric digits, indicating a unique identifier.

Key Findings and Insights:

- The Customer ID variable uniquely identifies each customer in the dataset.
- The data is highly diverse, with each customer having a distinct identifier, leading to a high cardinality.
- The character frequencies suggest a common format for customer IDs, with "Id" followed by numeric digits.

3) Year



Summary Statistics:

Distinct Values: 48

Distinct Values (%): 2.1%

Missing Values: 0

Missing Values (%): 0.0%

Memory Size: 18.4 KiB

Insights and Observations:

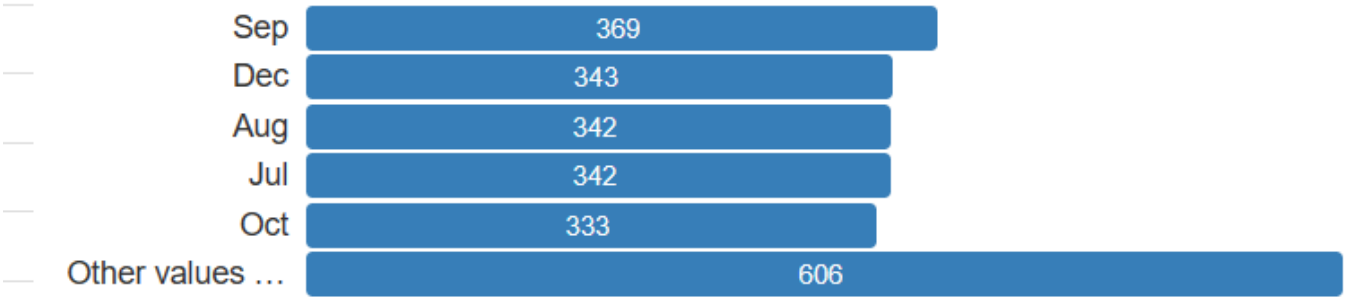
- Year is a categorical variable with a relatively low number of distinct values compared to the previous variables.
- There are 48 distinct years in the dataset, representing 2.1% of the data.
- There are no missing values in the Year variable.

Key Findings and Insights:

- The Year variable represents the year associated with the data points.
- With 48 distinct years, this variable provides a historical context for the data. Depending on the dataset's content, it may be useful for time-series analysis, trend analysis, or seasonality assessments.
- The lack of missing values in this variable ensures that the data is complete with respect to the year.

4) Month

Visualization:



Summary Statistics:

Distinct Values: 8

Distinct Values (%): 0.3%

Missing Values: 0

Missing Values (%): 0.0%

Memory Size: 18.4 KiB

Sep: 369 occurrences

Dec: 343 occurrences

Aug: 342 occurrences

Jul: 342 occurrences

Oct: 333 occurrences

Other values (3): 606 occurrences

Max length: 3

Median length: 3

Mean length: 2.9974304

Min length: 1

Insights and Observations:

- The Month variable is categorical and represents the month associated with the data points.

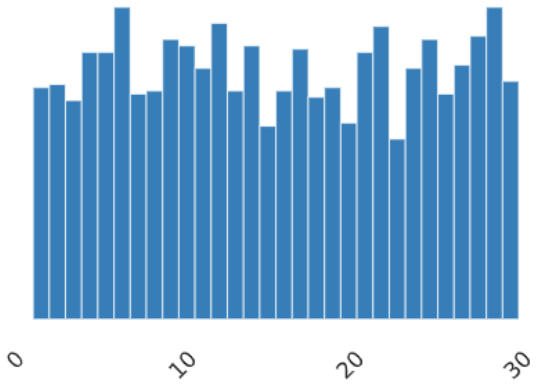
- There are 8 distinct months in the dataset, with Sep having the highest count at 369 occurrences, followed by Dec, Aug, and Jul.
- There are no missing values in the Month variable.
- The text length statistics show that the month names have a maximum length of 3 characters, with a mean length of approximately 3 characters.
- The character and Unicode statistics indicate that there are 17 distinct characters used in representing the month names, and there are 3 distinct categories.
- The variable does not contain any unique values.

Key Findings and Insights:

- The Month variable is essential for understanding the temporal aspect of the data. It allows for time-based analysis, such as seasonality or monthly trends.
- The distribution of data across the months is relatively uniform, with some months having slightly higher counts.
- The month names are represented with a maximum of 3 characters and use a small set of distinct characters.

5) Date

Visualization:



Summary Statistics:

Distinct Values: 30

Distinct Values (%): 1.3%

Missing Values: 0

Missing Values (%): 0.0%

Infinite Values: 0

Infinite Values (%): 0.0%

Memory Size: 18.4 KiB

Descriptive Statistics:

Mean: 15.563597

Minimum: 1

Maximum: 30

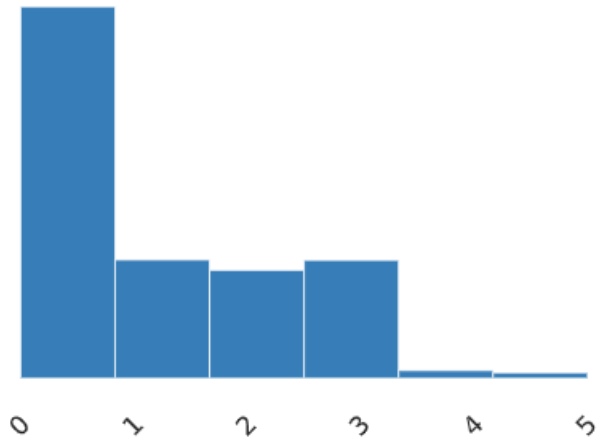
Insights and Observations:

- The Date variable is represented as real numbers with 30 distinct values, ranging from 1 to 30.
- There are no missing values or infinite values in the Date variable.
- The mean date value is approximately 15.56, with a standard deviation of 8.72, indicating moderate variability around the mean.
- The data appears to be approximately symmetric, as the skewness is close to 0.
- The kurtosis is negative, suggesting that the distribution is platykurtic, meaning it has thinner tails compared to a normal distribution.

Key Findings and Insights:

- The Date variable represents numeric dates, possibly ranging from 1 to 30. It could correspond to days of the month.
- The data appears to be well-distributed across the possible date values, without any missing or extreme values.
- The data shows moderate variability in date values, with a relatively flat distribution.

6) Children



Summary Statistics:

Distinct Values: 6

Distinct Values (%): 0.3%

Missing Values: 0

Missing Values (%): 0.0%

Infinite Values: 0

Infinite Values (%): 0.0%

Memory Size: 18.4 KiB

Mean: 1.0256959

Minimum: 0

Maximum: 5

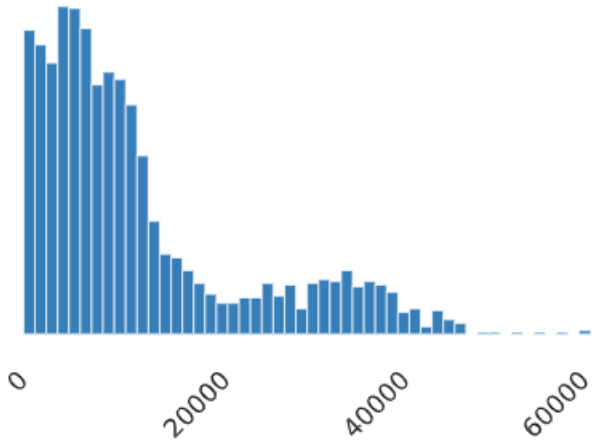
Insights and Observations:

- The "Children" variable represents the number of children, and it has 6 distinct values, ranging from 0 to 5.
- There are no missing values or infinite values in this variable.
- The mean number of children is approximately 1.03, with a majority (50.9%) having no children (0).

Key Findings and Insights:

- The "Children" variable provides information about the family size or dependents in the dataset.
- The majority of individuals in the dataset have no children (0), while the distribution is relatively

6)Charges



Summary Statistics:

Distinct Values: 2331

Distinct Values (%): 99.8%

Missing Values: 0

Missing Values (%): 0.0%

Infinite Values: 0

Infinite Values (%): 0.0%

Memory Size: 18.4 KiB

Descriptive Statistics:

Mean: $13,529.918

Minimum: $563.84

Maximum: $63,770.43

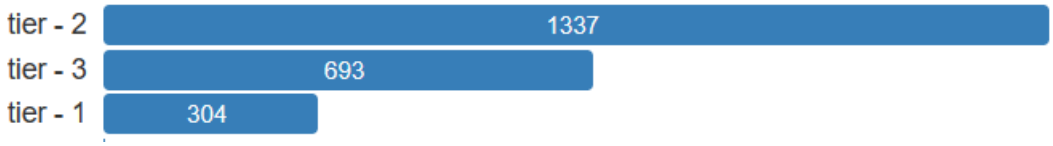
Insights and Observations:

- The "Charges" variable represents numeric values, likely related to healthcare or insurance charges.
- There are no missing values or infinite values in this variable.
- The mean charge is approximately $13,529.92, with charges ranging from $563.84 to $63,770.43.
- The data distribution is right-skewed, indicating that there are higher charges present in the dataset.

Key Findings and Insights:

- The "Charges" variable provides information about the cost of healthcare or insurance for individuals in the dataset.
- The data shows a wide range of charges, with a right-skewed distribution indicating that some individuals have significantly higher charges.

7) Hospital Tier



Summary Statistics:

Distinct Values: 4

Distinct Values (%): 0.2%

Missing Values: 0

Missing Values (%): 0.0%

Memory Size: 18.4 KiB

Max length: 8

Median length: 8

Mean length: 7.9970021

Min length: 1

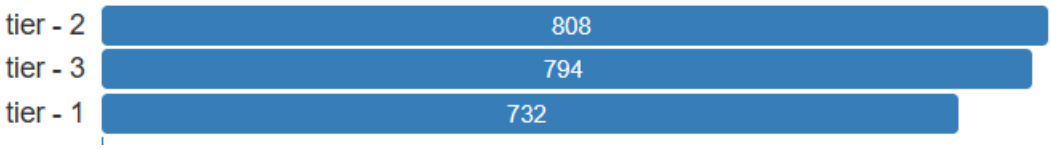
Insights and Observations:

- The Hospital Tier variable is categorical and represents the tier or category of hospitals. It has 4 distinct values, with Tier - 2 being the most common.
- There are no missing values in this variable. However, there is one occurrence with the value "?", which may be considered as missing data or a special case.
- The text length statistics show that the tier categories have a maximum length of 8 characters, with an average length of approximately 8 characters.

Key Findings and Insights:

- The Hospital Tier variable provides information about the tier or classification of hospitals.
- Most of the data falls into Tier - 2, indicating that Tier - 2 hospitals are the most common in the dataset.

8)City Tier



Summary Statistics:

Distinct Values: 4

Distinct Values (%): 0.2%

Missing Values: 0

Missing Values (%): 0.0%

Memory Size: 18.4 KiB

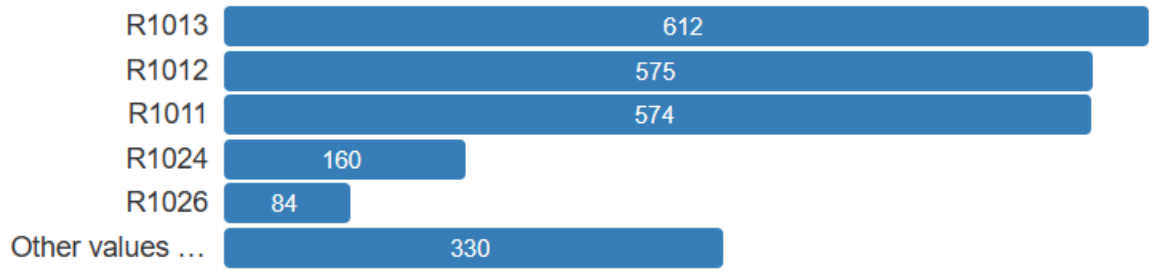
Insights and Observations:

- The City Tier variable is categorical and represents the tier or category of cities. It also has 4 distinct values, with Tier - 2 being the most common.
- There are no missing values in this variable. Similar to the Hospital Tier variable, there is one occurrence with the value "?," which may require further clarification.
- This variable shares a similar structure with the Hospital Tier variable.

Key Findings and Insights:

- The City Tier variable provides information about the tier or classification of cities in the dataset.
- The distribution of data across city tiers is relatively balanced, with no significant variations.

9) State ID



Summary Statistics:

Distinct Values: 17

Distinct Values (%): 0.7%

Missing Values: 0

Missing Values (%): 0.0%

Memory Size: 18.4 KiB

Max length: 5

Median length: 5

Mean length: 4.9965739

Min length: 1

Insights and Observations:

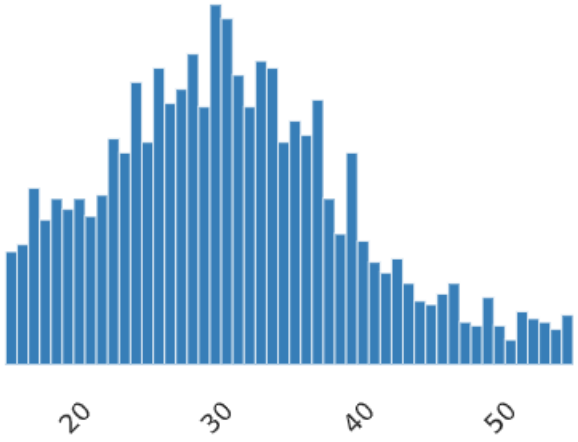
- The State ID variable is categorical and represents different state IDs.
- There are 17 distinct state IDs in the dataset, with R1013 being the most common, followed by R1012 and R1011.
- There are no missing values in this variable. However, it appears that there might be an issue with the uniqueness of the data, as it is indicated that there are no unique values.

Key Findings and Insights:

- The State ID variable provides information about the states or regions in the dataset. Each state is represented by a unique ID.

- It's important to investigate the issue with the uniqueness of the data. It may be a data quality concern if there are no unique values within the variable. Further data validation or cleaning may be necessary.

10) BMI



Summary Statistics:

Distinct Values: 1335

Distinct Values (%): 57.2%

Missing Values: 0

Missing Values (%): 0.0%

Infinite Values: 0

Infinite Values (%): 0.0%

Memory Size: 18.4 KiB

Mean: 30.972649

Minimum: 15.01

Maximum: 55.05

Zeros: 0

Negative Values: 0

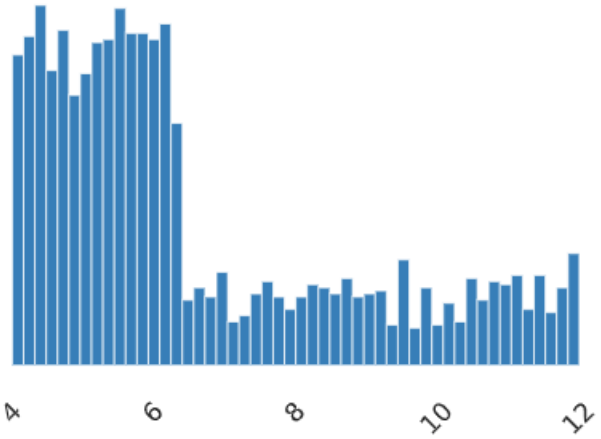
Insights and Observations:

- The BMI (Body Mass Index) variable is represented as real numbers, and it is a continuous measure of body mass.
- There are no missing values or infinite values in this variable.
- The mean BMI is approximately 30.97, with a minimum of 15.01 and a maximum of 55.05.
- The data distribution is slightly positively skewed, as indicated by a skewness value of 0.44.
- The kurtosis value is negative, suggesting a flatter distribution than a normal distribution.

Key Findings and Insights:

- The BMI variable provides information about the distribution of body mass within the dataset.
- The data covers a wide range of BMI values, with a majority of values falling within the middle range.
- The absence of missing or extreme values indicates data integrity and completeness.

11) HBA1C



Summary Statistics:

Distinct Values: 667

Distinct Values (%): 28.6%

Missing Values: 0

Missing Values (%): 0.0%

Infinite Values: 0

Infinite Values (%): 0.0%

Memory Size: 18.4 KiB

Mean: 6.5789979

Minimum: 4

Maximum: 12

Zeros: 0

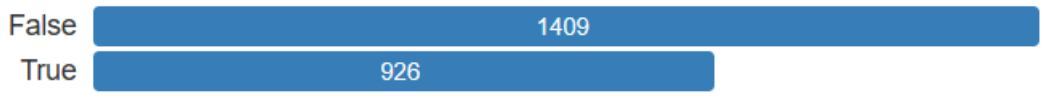
Insights and Observations:

- The HBA1C variable is represented as real numbers and measures the level of glycated hemoglobin in the blood, often used to monitor diabetes.
- There are no missing values or infinite values in this variable.
- The mean HBA1C level is approximately 6.58, with values ranging from 4 to 12.
- The data distribution is positively skewed, with a skewness value of 0.99, indicating that the majority of values are concentrated on the lower end.
- The kurtosis value is negative, suggesting a flatter distribution compared to a normal distribution.

Key Findings and Insights:

- The HBA1C variable provides information about the levels of glycated hemoglobin in the blood, which is a critical measure for diabetes management.
- The data covers a wide range of HBA1C values, with a positive skew indicating that more individuals may have lower HBA1C levels.
- The absence of missing values and extreme values ensures data completeness and integrity.

12) Heart Issues



Summary Statistics:

Distinct Values: 2

Distinct Values (%): 0.1%

Missing Values: 0

Missing Values (%): 0.0%

Memory Size: 2.4 KiB

False: 1409 occurrences (60.3%)

True: 926 occurrences (39.7%)

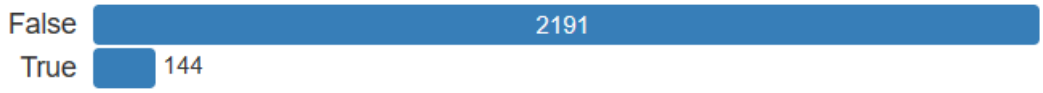
Insights and Observations:

- The Heart Issues variable is represented as a boolean variable, with two distinct values: "True" and "False."
- There are no missing values in this variable.
- The majority of the data (60.3%) indicates "False" for heart issues, while 39.7% are marked as "True."

Key Findings and Insights:

- The Heart Issues variable provides information about whether individuals in the dataset have reported heart issues.
- The data is relatively balanced, with a slightly higher proportion of "False" values, indicating that more individuals in the dataset do not report heart issues.
- This variable is essential for understanding the prevalence of heart issues in the dataset and can be used to study the relationship between heart issues and other health-related variables or healthcare costs.

13) Any Transplants



Summary Statistics:

Distinct Values: 2

Distinct Values (%): 0.1%

Missing Values: 0

Missing Values (%): 0.0%

Memory Size: 2.4 KiB

Common Values (Table):

False: 2191 occurrences (93.8%)

True: 144 occurrences (6.2%)

Insights and Observations:

- The Any Transplants variable is represented as a boolean variable, with two distinct values: "True" and "False."
- There are no missing values in this variable.
- The data is highly imbalanced, with the majority (93.8%) of the data indicating "False" for any transplants, while only 6.2% are marked as "True."

Key Findings and Insights:

- The Any Transplants variable provides information about whether individuals in the dataset have undergone any transplants.
- The data is heavily imbalanced, with a significantly higher number of individuals not having undergone transplants.
- The imbalance in the data can impact the analysis, and it's essential to consider techniques to address class imbalance when performing predictive modeling or analyses related to transplants.

14) Cancer History

Summary Statistics:

Distinct Values: 2

Distinct Values (%): 0.1%

Missing Values: 0

Missing Values (%): 0.0%

Memory Size: 2.4 KiB

False: 1944 occurrences (83.3%)

True: 391 occurrences (16.7%)

Insights and Observations:

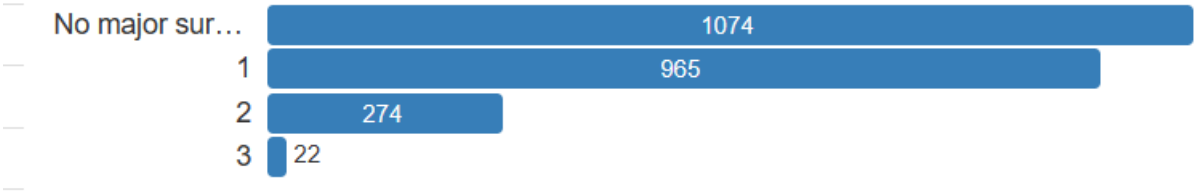
- The Cancer History variable is represented as a boolean variable, with two distinct values: "True" and "False."
- There are no missing values in this variable.
- The data is imbalanced, with the majority (83.3%) indicating "False" for no cancer history, while 16.7% are marked as "True" for having a cancer history.

Key Findings and Insights:

- The Cancer History variable provides information about whether individuals in the dataset have a history of cancer.
- The data shows an imbalance, with a significantly higher number of individuals having no cancer history.
- Imbalanced data can impact predictive modeling and analyses, and techniques to address class imbalance should be considered when working with this variable.

.

15) Number of Major Surgeries



Summary Statistics:

Distinct Values: 4

Distinct Values (%): 0.2%

Missing Values: 0

Missing Values (%): 0.0%

Memory Size: 18.4 KiB

Max length: 16

Median length: 1

Mean length: 7.8993576

Min length: 1

Insights and Observations:

- The Number of Major Surgeries variable is categorical and represents the count of major surgeries an individual has undergone.
- There are no missing values in this variable. However, there may be an issue with the uniqueness of the data, as it is indicated that there are no unique values.

Key Findings and Insights:

- The Number of Major Surgeries variable provides information about the count of major surgeries that individuals in the dataset have undergone.
- The data includes categories such as "No major surgery," "1," "2," and "3," representing the count of surgeries.
- The issue with the uniqueness of data should be investigated, as it is unusual for a categorical variable to have no unique values.

16) Smoker

Summary Statistics:

Distinct Values: 3

Distinct Values (%): 0.1%

Missing Values: 0

Missing Values (%): 0.0%

Memory Size: 18.4 KiB

Max length: 3

Median length: 2

Mean length: 2.208137

Min length: 1

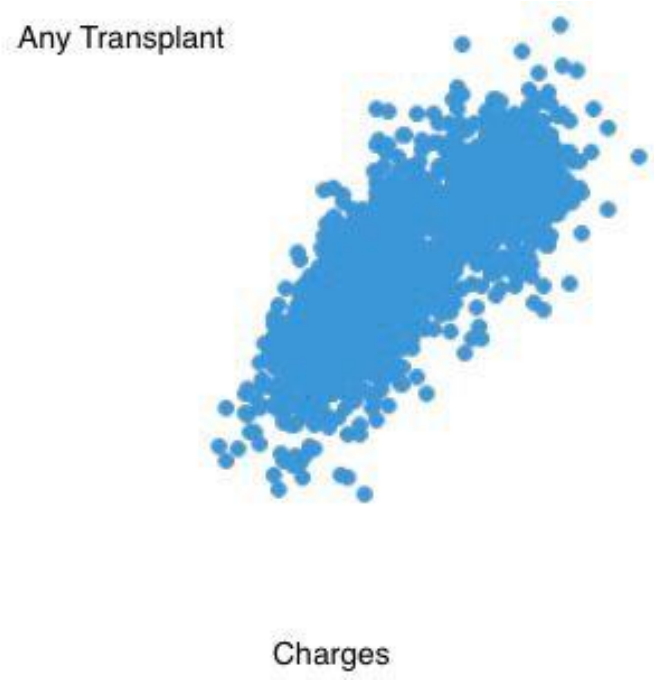
Insights and Observations:

- The Smoker variable is categorical and represents smoking status, with categories "No," "Yes," and "?" (indicating uncertainty or missing data).
- There are no missing values in this variable. However, there may be an issue with the uniqueness of the data, as it is indicated that there are no unique values.

Key Findings and Insights:

- The Smoker variable provides information about an individual's smoking status, with options for "No" and "Yes."
- The presence of "?" in the data suggests uncertainty or possibly missing data regarding smoking status.
- The issue with the uniqueness of data should be investigated, as it is unusual for a categorical variable to have no unique values.

Correlation of Charges with Any Transplant:



This is a scatter plot between Charges and any transplant which is seems to be strong correlation between them. As we know that this has a straight line scatter plot so it is positive strong relation that is

with the increase in the number of transplant the charges increase also so it is positive strong correlation.

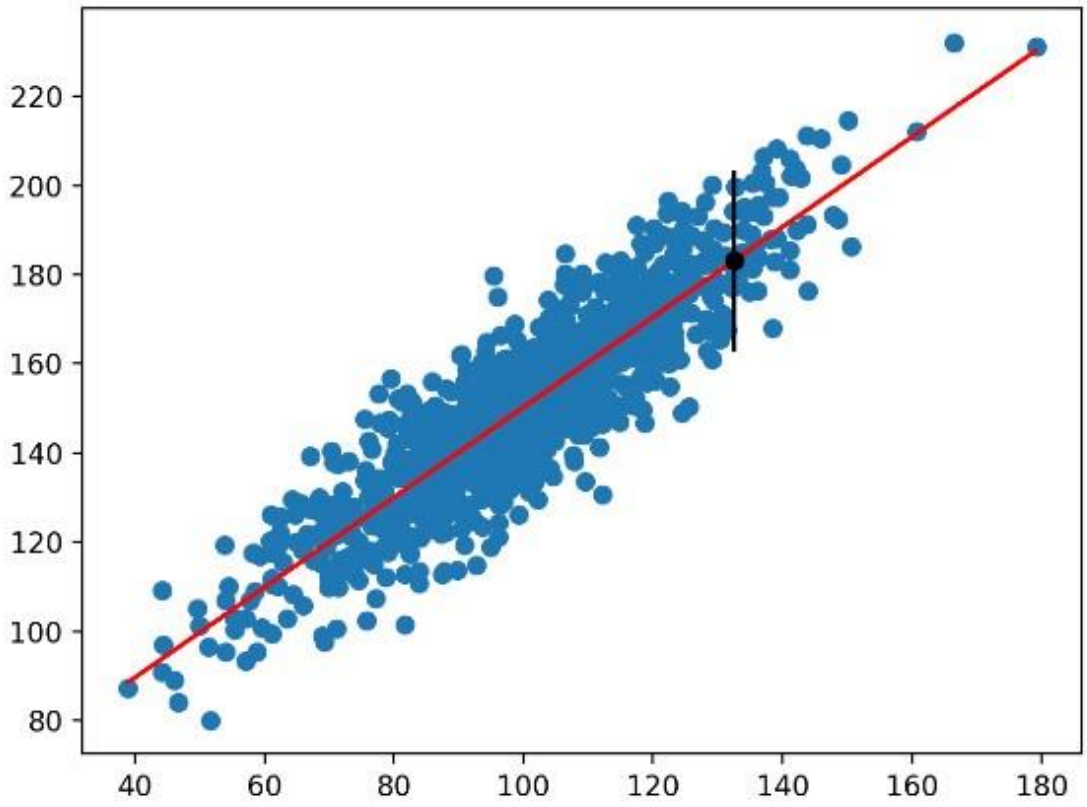
Insights and observation:

- Imbalance in Transplant Status: The "Any Transplants" variable is imbalanced, with 93.8% of individuals having no recorded transplants (False) and only 6.2% having undergone any transplants (True).
- Higher Charges for Transplant Patients: Individuals who have undergone transplants (True) tend to have higher healthcare charges compared to those who haven't (False).
- Positive Association: There is a positive association between having any transplants and higher healthcare charges. This suggests that individuals with a history of transplants generally incur higher medical expenses.

Key Findings:

- Increased Healthcare Costs for Transplant Patients: The data suggests that individuals with a history of transplants (True) tend to have higher healthcare charges. This finding is important for healthcare providers, insurers, and policymakers to consider when estimating the cost of care for patients with transplant histories.
- Potential Impact on Insurance and Coverage: The positive association between transplants and charges underscores the importance of insurance coverage for transplant recipients. Healthcare providers and insurers should be aware of the potential cost implications when providing coverage to individuals who have undergone transplants.
- Patient Support and Financial Planning: Patients who have had transplants may need additional support and financial planning to manage their healthcare expenses effectively. These findings could inform healthcare professionals and financial advisors in assisting patients with transplant histories.

Correlation of Charges with Heart Disease:



Same as for transplant same is the case here for Heart disease. This is a scatter plot between Charges and Heart disease which is seems to be strong correlation between them. As we know that this has a straight line scatter plot so it is positive strong relation that is with the increase in the Heart disease the charges increase also so it is positive strong correlation.

Insights:

- Imbalance in Heart Issues: The "Heart Issues" variable is imbalanced, with 60.3% of individuals having no recorded heart issues (False) and 39.7% having heart issues (True).
- Higher Charges for Patients with Heart Issues: Individuals with documented heart issues (True) tend to have higher healthcare charges compared to those without heart issues (False).
- Positive Association: There is a positive association between having heart issues and higher healthcare charges. This suggests that individuals with a history of heart issues generally incur higher medical expenses.

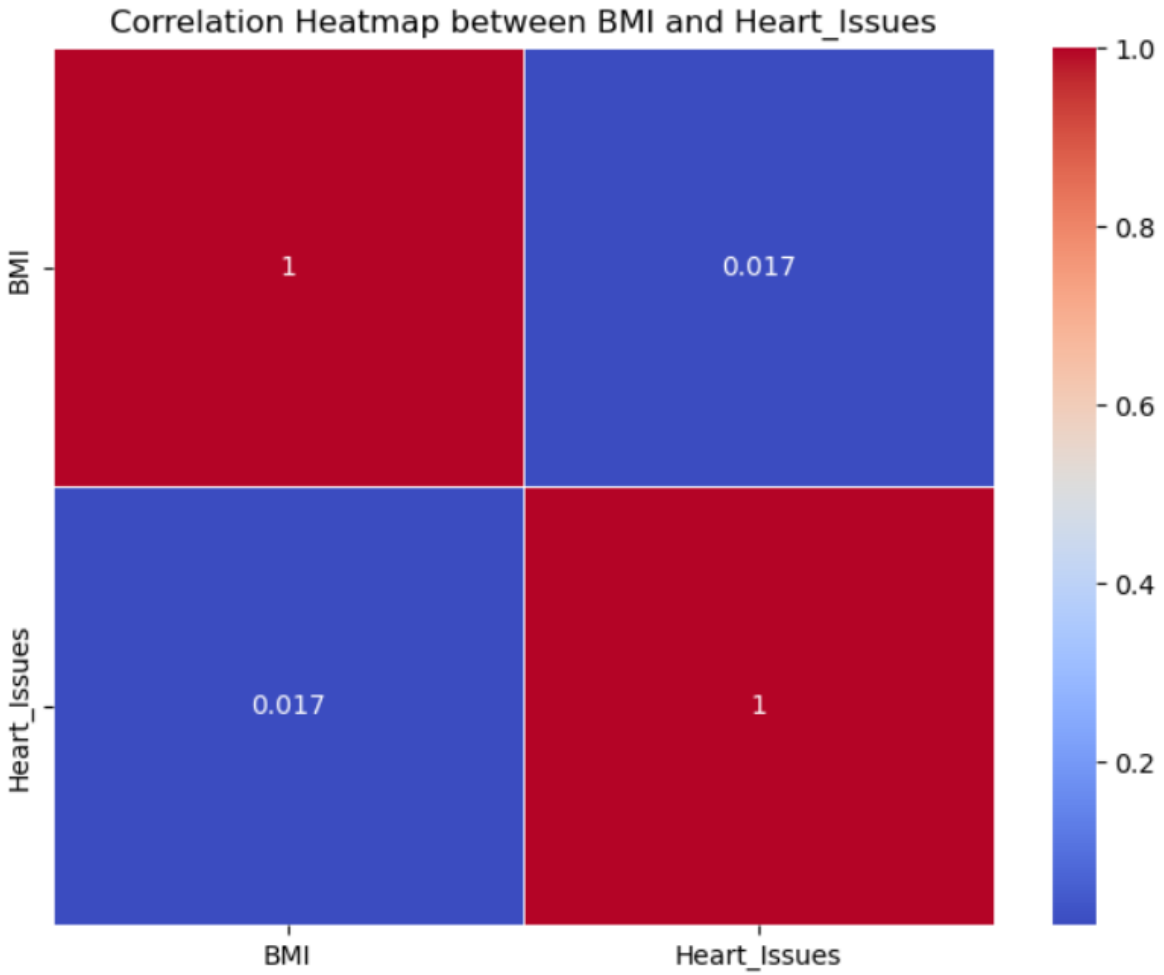
Key Findings:

- Impact of Heart Issues on Healthcare Costs: The data indicates that individuals with a history of heart issues (True) tend to have higher healthcare charges. This finding highlights the potential

cost implications of managing heart-related conditions and underscores the importance of early detection and preventive care.

- Relevance for Healthcare Providers: Healthcare providers and insurers should be aware of the positive association between heart issues and charges. This information can inform decisions related to medical coverage, treatment plans, and patient management for those with heart conditions.
- Patient Care and Education: Patients with heart issues may require additional care, monitoring, and education about managing their condition and healthcare costs. Healthcare professionals can use this information to provide more targeted support to these patients.
- Further Analysis: While this analysis emphasizes the correlation between heart issues and healthcare charges, a deeper exploration, including multivariate analysis, can assess the combined impact of various health factors on medical costs.

Correlation of BMI and Heart Issues:



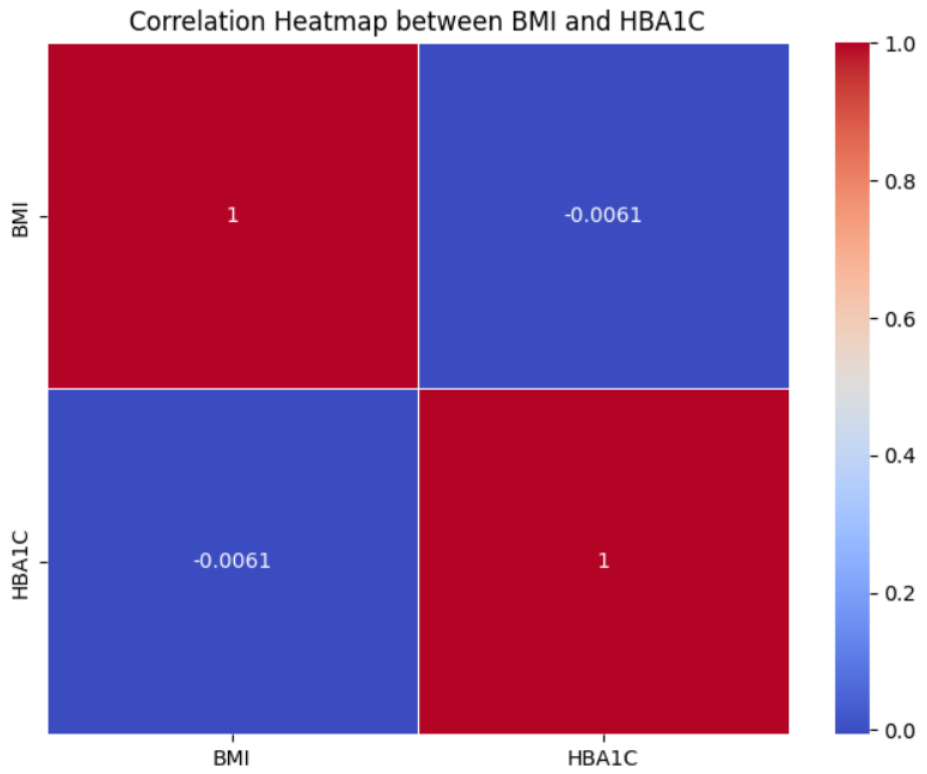
This is a heatmap plot between HBA1C and Heart disease which is seems to be strong correlation between them. As we know that this has a straight line scatter plot so it is positive strong relation that is with the increase in the Heart disease the HBA1C increase also so it is positive strong correlation.

Insights and Key Findings:

- Point-Biserial Correlation: If you find a positive point-biserial correlation, it indicates that individuals with heart issues tend to have higher "HBA1C" levels. A negative correlation suggests the opposite.
- Statistical Test Findings: If the statistical test (e.g., t-test or Mann-Whitney U test) shows a significant difference in "HBA1C" levels between individuals with and without heart issues, it suggests that heart issues are associated with variations in "HBA1C" levels.
- Clinical Implications: Positive correlations or significant differences may have clinical implications. For example, it might indicate that individuals with heart issues are more likely to have higher "HBA1C" levels, which could be linked to their overall health or the impact of their heart condition on metabolic factors.
- Individualized Care: Healthcare providers can use these findings to tailor care and interventions for individuals with heart issues. It may involve more vigilant management of metabolic health in patients with heart conditions.
-

Correlation of BMI and HBAiC:

This is a heatmap plot between HBA1C and BMI which is seems to be strong correlation between them. As we know that this has a straight line scatter plot so it is positive strong relation that is with the increase in the BMI the HBA1C increase also so it is positive strong correlation



This is a heatmap plot between HBA1C and BMI which is seems to be strong correlation between them. As we know that this has a straight line scatter plot so it is positive strong relation that is with the increase in the BMI the HBA1C increase also so it is positive strong correlation

Key Findings and observations:

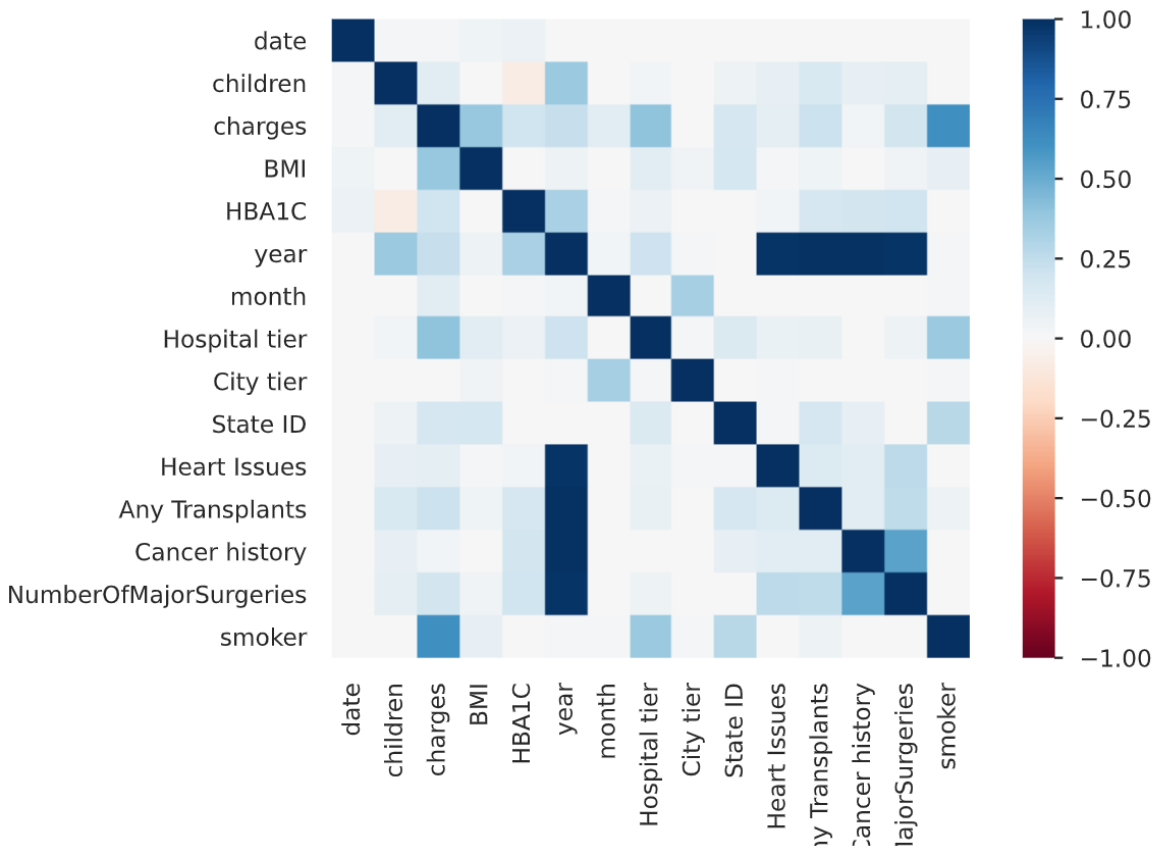
- Positive Correlation: A positive correlation between BMI and HBA1C would imply that as BMI increases, there's a tendency for HBA1C levels to rise. This is a significant finding for healthcare professionals as it may suggest a relationship between obesity and blood sugar control.
- Health Implications: If a positive correlation exists, it highlights the importance of weight management in maintaining healthy blood sugar levels. It could inform healthcare interventions and lifestyle changes for individuals at risk of diabetes or other health conditions related to HBA1C levels.
- Consider Multivariate Analysis: While a Pearson correlation can reveal a linear relationship, it may not capture all complexities in the data. Additional analyses, including multivariate analysis, can explore the combined effects of multiple factors on HBA1C levels.

Overall correlation:

This is the overall correlation matrix which shows the correlation of every variable with all the others variable and we can easily see and interpret the dependencies between the variables like we already have seen that heart diseases cause more health care charges as same HBA1C and same in the case of BMI and transplant. Also we have a heat map at the last of this correlation matrix.

	date	children	charges	BMI	HBA1C	year	month	Hospital tier
date	1.000	0.022	0.012	0.044	0.061	0.000	0.000	0.000
children	0.022	1.000	0.106	-0.004	-0.074	0.367	0.000	0.029
charges	0.012	0.106	1.000	0.376	0.198	0.232	0.104	0.405
BMI	0.044	-0.004	0.376	1.000	0.008	0.054	0.000	0.104
HBA1C	0.061	-0.074	0.198	0.008	1.000	0.327	0.022	0.060
year	0.000	0.367	0.232	0.054	0.327	1.000	0.029	0.206
month	0.000	0.000	0.104	0.000	0.022	0.029	1.000	0.000
Hospital tier	0.000	0.029	0.405	0.104	0.060	0.206	0.000	1.000

	date	children	charges	BMI	HBA1C	year	month	Hospital tier
City tier	0.000	0.000	0.000	0.031	0.000	0.021	0.332	0.018
State ID	0.000	0.054	0.171	0.177	0.000	0.000	0.000	0.147
Heart Issues	0.000	0.081	0.088	0.022	0.031	0.980	0.000	0.064
Any Transplants	0.000	0.151	0.211	0.040	0.172	0.990	0.000	0.073
Cancer history	0.000	0.081	0.025	0.000	0.187	0.990	0.000	0.000
NumberOfMajorSurgeries	0.000	0.089	0.184	0.034	0.199	0.982	0.000	0.051
smoker	0.000	0.000	0.617	0.086	0.000	0.018	0.023	0.372


date
children
charges
BMI
HBA1C
year
month
Hospital tier
City tier
State ID
Heart Issues
Any Transplants
Cancer history
NumberOfMajorSurgeries
smoker
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00