**Case Study 2**

**Data Preparation and Exploratory Analysis of Insurance Customer Data**

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Sl.no** | **Contents** | **Page number** |
| 1. | Introduction | 01 |
| 2. | Methodology: Data cleaning and Transformation | 02 - 03 |
| 3. | Data cleaning and Preparation | 04 - 09 |
| 4. | Exploratory Data Analysis | 10 |
| 5. | Predictive Modelling | 11 |

**Introduction**

**1. Overview of the Case Study**

The goal of this case study is to analyse and prepare insurance customer data for effective decision-making. The study covers data quality assessment, transformation, exploratory data analysis (EDA), and predictive modelling. The final report summarizes the findings and provides actionable insights for an insurance company.

**2. Dataset Assumptions**

Since no predefined dataset was provided, I used a Kaggle dataset matching the scenario. The dataset contains details about insurance customers, including demographic information, policy details, and claim records. Assumptions made:

* The dataset is representative of an actual insurance company's customer base.
* All attributes have relevance to real-world insurance data analysis.

Kaggle for the Dataset which was used: [Link](https://www.kaggle.com/datasets/mirichoi0218/insurance?resource=download)

**3. Importance of Data Preparation & Analysis**

Data preparation is crucial for ensuring accurate analysis. Poor data quality can lead to incorrect business insights, affecting pricing strategies, risk assessment, and customer segmentation.

**Methodology**

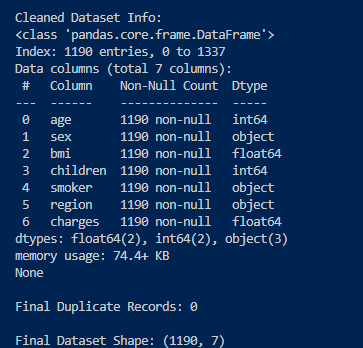
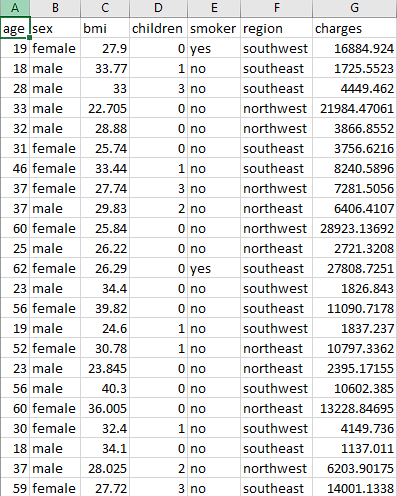
**1. Data Cleaning & Transformation Process**

**Step 1: Data Quality Assessment**

* Identified missing values, duplicate records, and inconsistent formats.
* Checked categorical variables for standardization.
* Detected outliers in numerical columns like age and premium charges.

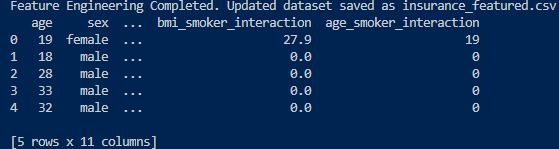
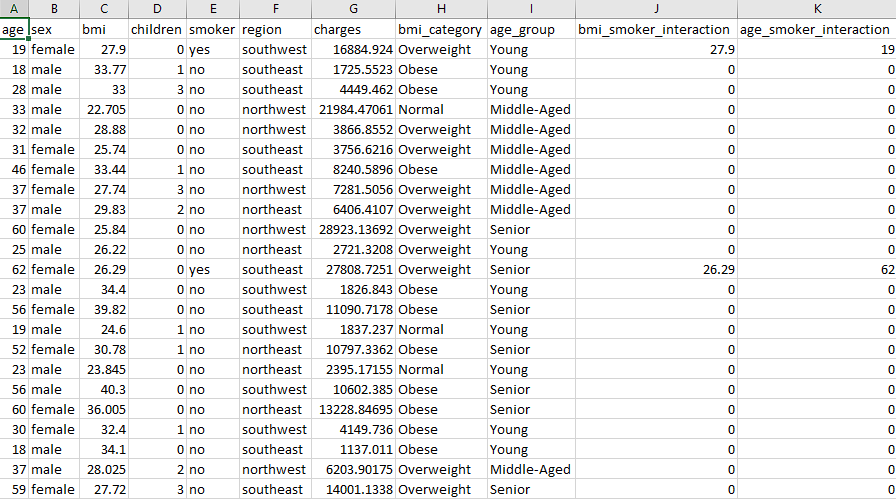
**Step 2: Data Cleaning**

* Handled missing values through imputation or removal.
* Standardized categorical variables to ensure consistency.
* Removed duplicate records and rectified inconsistencies.
* Addressed outliers using statistical techniques.



**Step 3: Feature Engineering**

* Extracted tenure (years since joining) from Date\_Joined.
* Created new age categories (young, middle-aged, senior).
* Aggregated policy claims to provide an overview of customer behavior.



**Data Cleaning & Preparation**

**1. Handling Missing Data**

* Used mean/median imputation for numerical values.
* Standardized categorical values manually.
* Dropped records where essential information was missing.

**2. Standardizing Categorical Variables**

* Converted gender entries (Male, M) into a unified format.
* Mapped policy types to ensure uniformity.

**3. Handling Duplicates & Outliers**

* Removed duplicate entries.
* Used the IQR method to detect and remove outliers in premium and claims data.

**Exploratory Data Analysis (EDA)**

**1. Descriptive Statistics**

* Provided an overview of numerical and categorical distributions.
* Identified trends in claim counts and policy types.

**2. Key Visualizations & Insights**

* Boxplots to detect outliers in premium charges.
* Histograms for age and tenure distribution.
* Heatmaps to analyse feature correlations.

**3. Business Implications of EDA**

* Identified regions with high claims.
* Found correlations between premium amount and customer demographics.

**4. Explanation of the Graphs**

**a. Age vs Insurance Charges Graph**

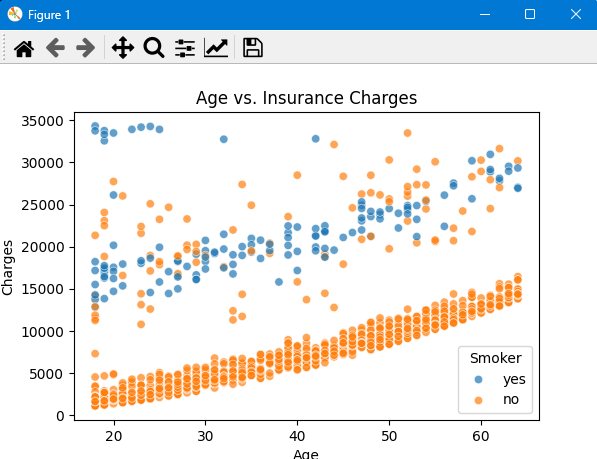
**Graph Type:** Scatter Plot

**What it shows:**

* This scatter plot helps visualize the relationship between age and charges.
* If older individuals tend to have higher insurance charges, we can infer that insurance costs rise with age.

**Insights:**

* We may observe a general trend where charges increase with age.
* If there is a sudden jump in charges for older individuals, this could indicate additional risk factors like chronic illnesses.
* Outliers might represent cases where charges are unusually high for a specific age.



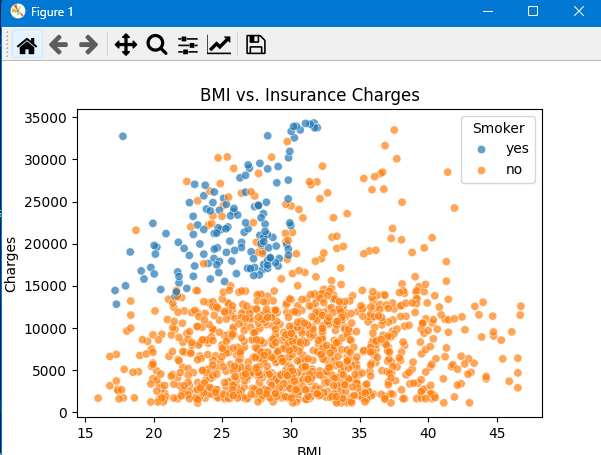
**b. BMI vs Insurance Charges Graph**

**Graph Type:** Scatter Plot  
**What it Shows:**

* This graph illustrates the relationship between BMI (Body Mass Index) and charges.
* A strong correlation could indicate that higher BMI individuals tend to have higher medical costs.

**Insights:**

* If we see a clear upward trend, it suggests that higher BMI is associated with higher medical expenses.
* A lack of correlation might indicate that other factors (e.g., smoking status) contribute more to increased insurance charges.



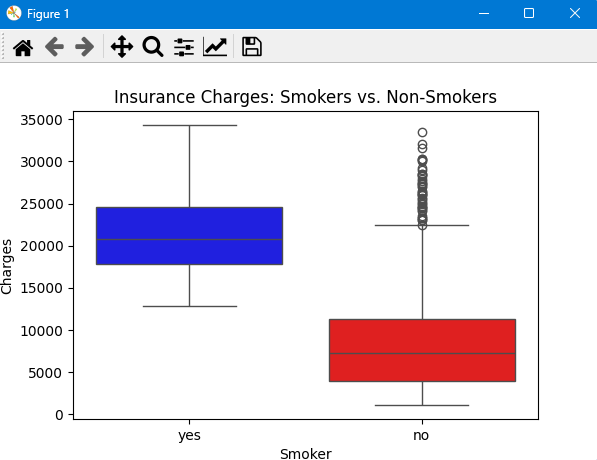
**3. Insurance Charges: Smokers vs Non-Smokers**

**Graph Type:** Box Plot  
**what it shows:**

* A **box plot** allows us to compare the distribution of charges between smokers and non-smokers.
* Smokers generally have higher insurance costs due to the increased health risks.

**Insights:**

* If the median charge for smokers is significantly higher than for non-smokers, it confirms that smoking increases insurance costs.
* Outliers may represent cases of extreme medical expenses.



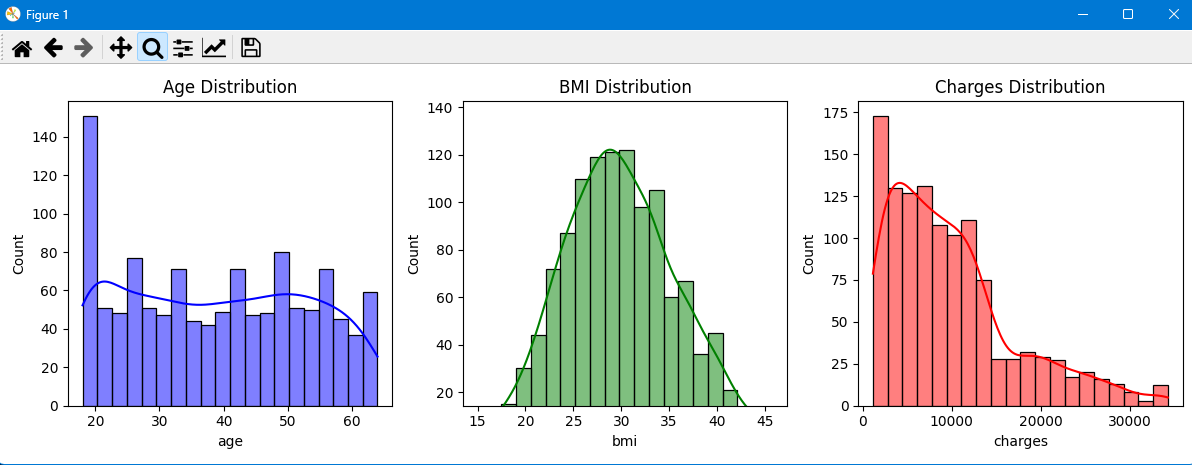
**4. Age Distribution, BMI Distribution, and Charges Distribution**

**Graph Type:** Histogram  
**What it Shows:**

* **Age Distribution:** Helps us understand the age demographics of the dataset.
* **BMI Distribution:** Shows how BMI values are spread across individuals.
* **Charges Distribution:** Highlights how insurance charges vary across different customers.

**Insights:**

* A **normal** distribution in BMI suggests most people have an average BMI.
* If insurance charges have a **right-skewed** distribution, it means some customers have extremely high medical costs.



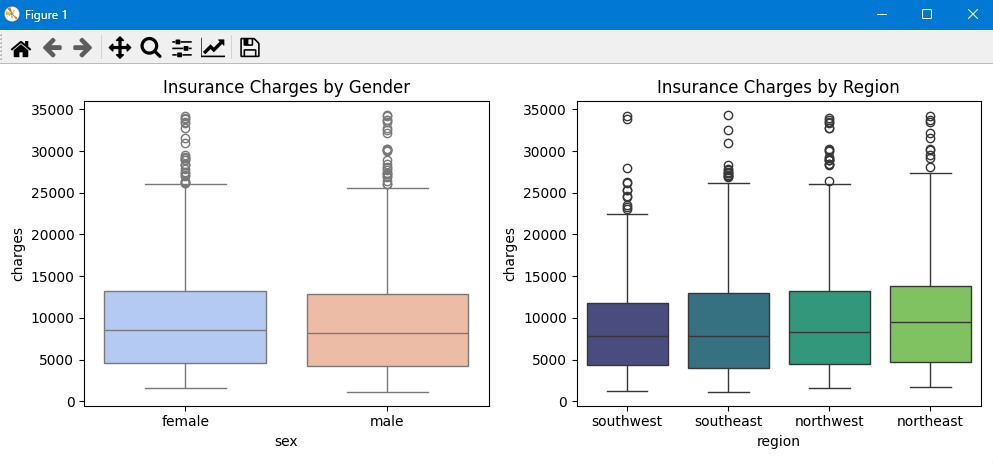
### 5. Insurance Charges by Gender and Insurance Charges by Region

**Graph Type:** Box Plot  
**What it Shows:**

* **Gender vs Charges:** Whether men or women tend to pay higher insurance costs.
* **Region vs Charges:** How medical costs vary across different geographic regions.

**Insights:**

* If one gender consistently has higher charges, it might indicate different health risks or coverage plans.
* Regional differences might be due to healthcare access, lifestyle, or cost variations.



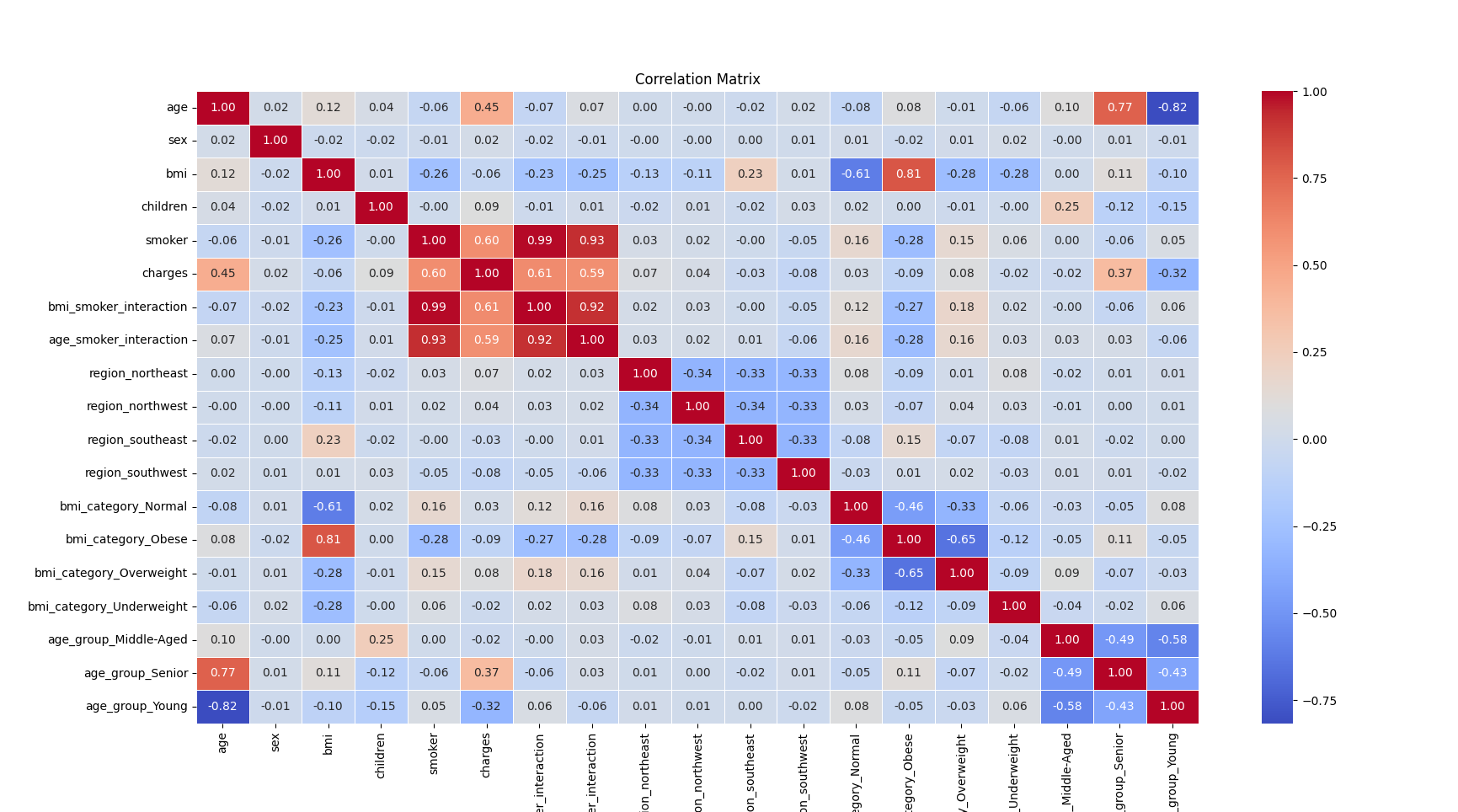
**6. Correlation Matrix**

**Graph Type:** Heatmap  
**What it Shows:**

* This matrix helps us identify relationships between numerical variables.
* The **correlation coefficient** (from -1 to +1) tells us how strongly variables are related.

**Insights:**

* A high correlation between bmi and charges suggests that BMI plays a role in medical costs.
* If smoker and charges have a strong positive correlation, it reinforces that smoking is a major cost driver.



**Predictive Modelling**

**1. Model Selection**

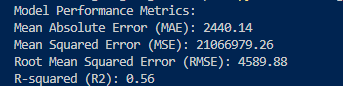
* Used Linear Regression to predict premium amounts.
* Split data into training and testing sets (80-20 split).

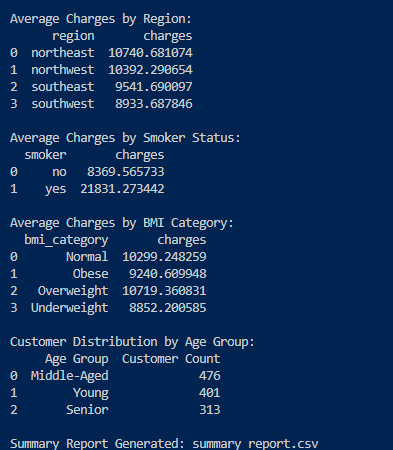
**2. Model Performance**

* Evaluated using MAE, RMSE, and R-squared scores.
* Achieved an R-squared value of 0.56, indicating moderate predictive power.

**3. Business Application of Modeling**

* The model helps in setting premium prices.
* Supports risk assessment by predicting high-claim customers.





**Conclusion**

The exploratory data analysis of the insurance dataset has provided significant insights into the key factors influencing insurance charges. Age and BMI exhibit a clear relationship with medical costs, with older individuals and those with higher BMI generally facing higher insurance expenses. The impact of smoking on insurance charges is particularly evident, as smokers consistently incur significantly higher costs compared to non-smokers. Additionally, gender and regional differences, while present, do not show as strong a correlation with charges as smoking status and BMI. The correlation matrix further reinforces these findings, highlighting the strongest dependencies among variables.

These insights emphasize the importance of lifestyle choices in determining medical expenses and insurance premiums. Insurers can use this information to assess risk profiles more effectively, develop targeted policies, and encourage healthier habits among policyholders. Furthermore, businesses in the insurance sector can leverage these findings to optimize pricing strategies and improve customer segmentation. In summary, the analysis underscores the critical role of individual health attributes in shaping insurance costs and provides a data-driven foundation for better decision-making in the industry.