Medical Insurance Cost Analysis: A Deep Dive into Key Drivers

Prepared for: Stakeholders

Prepared by: [Soham Waghe]

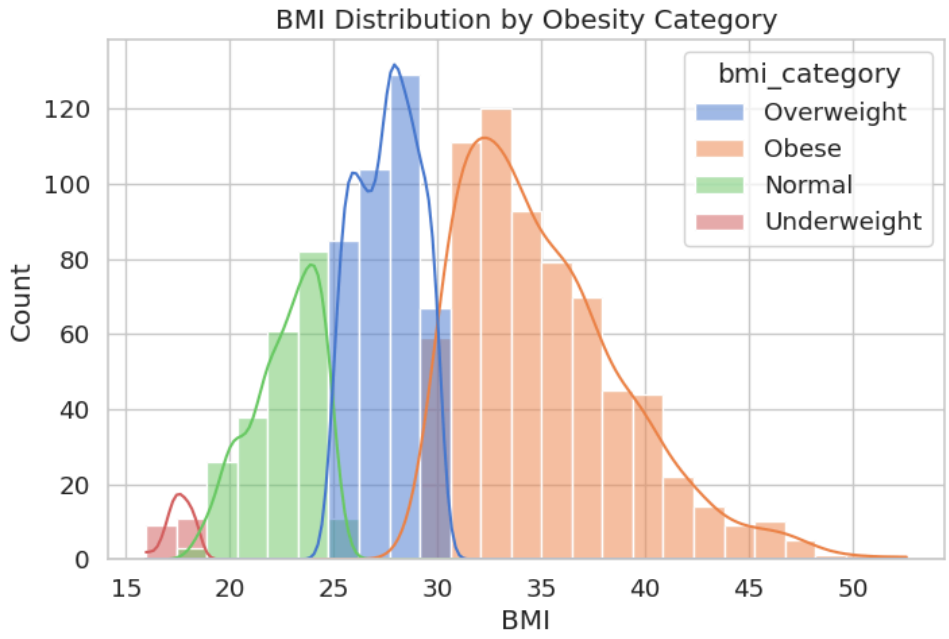
Date: September 25, 2025

---

1. Executive Summary

This report presents a comprehensive analysis of a medical insurance dataset to identify the primary factors influencing policyholder charges. The investigation reveals that **smoking status** is the single most significant driver of medical costs, with smokers incurring dramatically higher charges than non-smokers.

Furthermore, a combination of factors, particularly **high Body Mass Index (BMI)** and **age**, compound these costs. Our analysis indicates a clear positive correlation between higher BMI, older age, and increased insurance charges. An interactive Power BI dashboard has been developed to allow for dynamic exploration of these findings.



Key recommendations include a **review of premium pricing structures** to more accurately reflect the risk associated with smoking, the development of **targeted wellness programs** to promote smoking cessation and healthy weight management, and leveraging these insights for **risk-based marketing campaigns**.

---

2. Introduction

The primary objective of this project was to analyze medical insurance data to understand the relationships between policyholder attributes and their total medical charges. By identifying these key drivers, we aim to provide actionable insights that can inform pricing strategies, improve risk assessment models, and guide the development of customer-facing wellness initiatives. This report details the methodology, from data preparation to exploratory analysis and the final dashboard implementation.

---

3. Data Preparation and Cleaning

The initial dataset required several preparation steps to make it suitable for analysis. The process was as follows:

- **Handling Categorical Data:** Non-numerical data such as `sex`, `smoker` status, and `region` were converted into a numerical format. `Sex` and `smoker` were label-encoded (e.g., male/female becomes 1/0), while `region` was one-hot encoded to prevent implying any ordinal relationship.
- **Feature Engineering:** To uncover deeper insights, new features were engineered from the existing data:
    - A **`bmi_category`** was created to classify individuals as 'Underweight', 'Normal', 'Overweight', or 'Obese' based on their BMI. This simplifies the interpretation of BMI's impact.
    - An **`age_smoker` interaction feature** was created by multiplying a policyholder's age by their smoking status. This helps capture the compounding effect of these two critical variables on medical charges.
- **Outlier Review:** Initial exploration using boxplots identified the presence of outliers in the `bmi` and `charges` columns. These were retained for the analysis as they represent legitimate, albeit high, values that are crucial for understanding the full scope of costs.

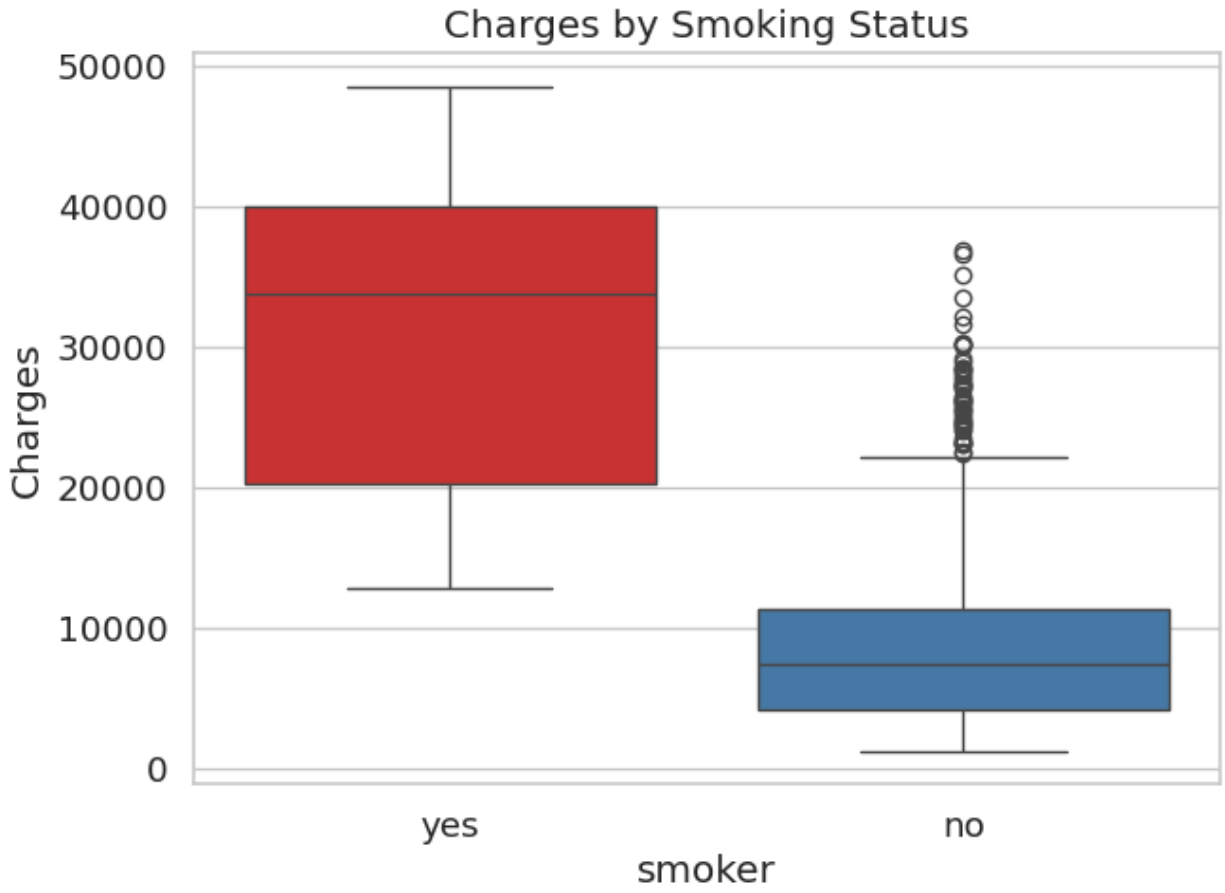
---

4. Exploratory Data Analysis (EDA) & Key Findings

EDA was performed to visualize patterns and relationships within the data. The following are the most significant findings.

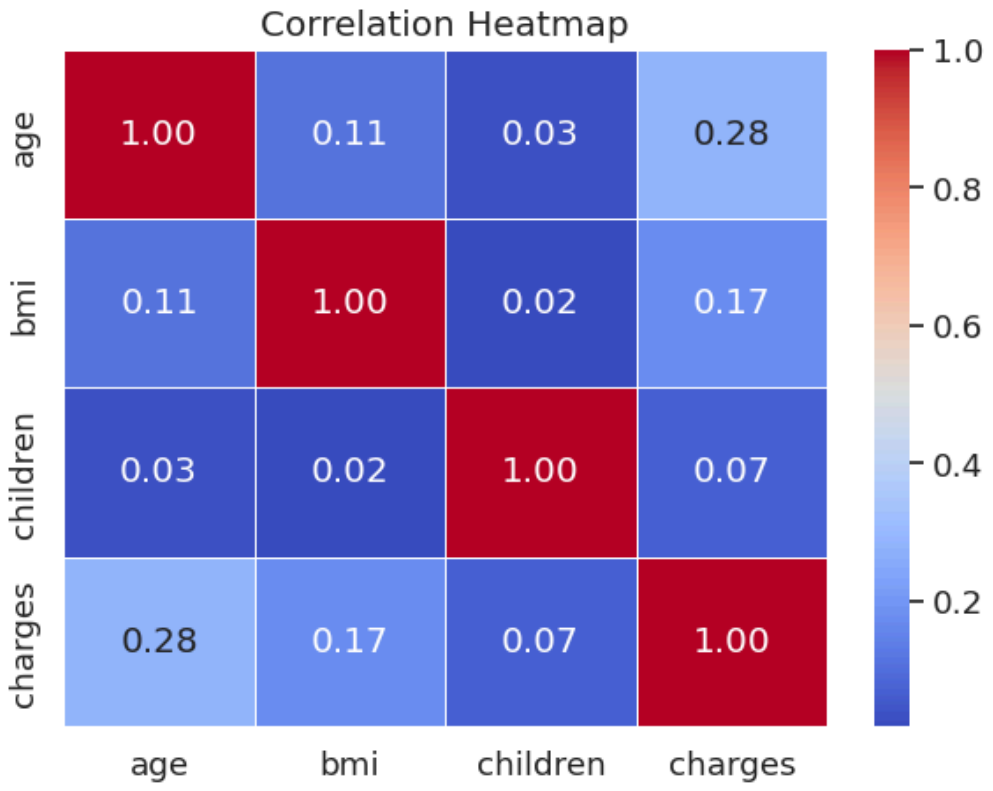
4.1 The Overwhelming Impact of Smoking

The most striking finding from the analysis is the profound impact of smoking on insurance charges. As shown in the boxplot below, the median charge for smokers is approximately **four times higher** than for non-smokers. The range of charges for smokers is also significantly wider, indicating much greater cost variability and risk.



4.2 Correlation of Key Factors

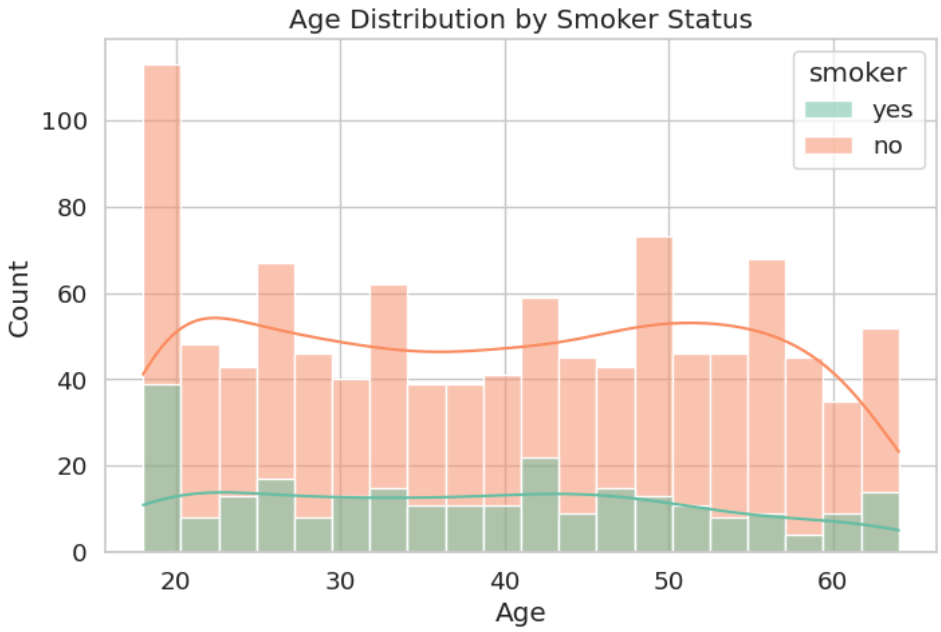
A correlation heatmap was generated to quantify the linear relationships between numerical variables.



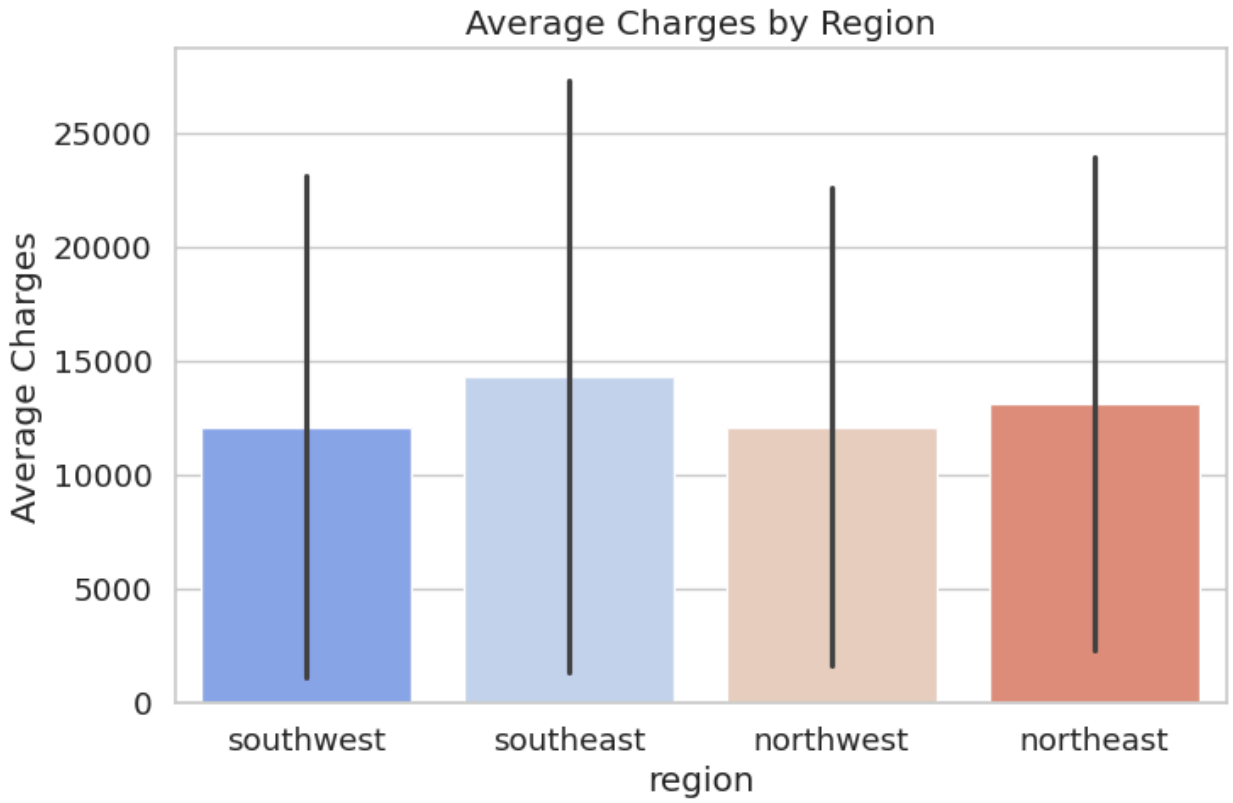
The heatmap indicates a moderate positive correlation between **`age`** **and** **`charges`** **(0.28)** and a weaker positive correlation between **`bmi`** **and** **`charges`** **(0.17)**. While these individual correlations are not overwhelmingly strong, their combined effect, especially when paired with smoking status, is substantial.

4.3 Demographic and Regional Analysis

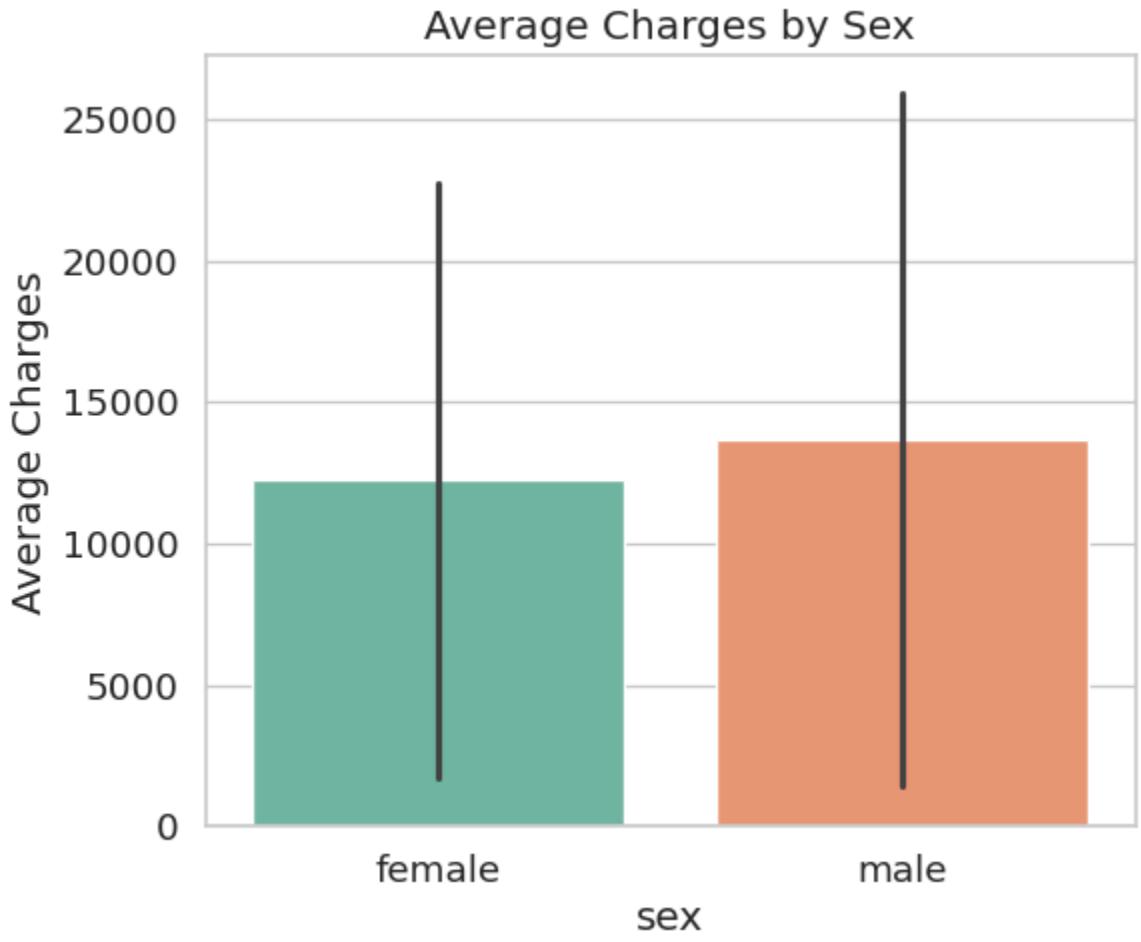
- **Age Distribution:** The analysis of age distribution shows that while non-smokers are spread fairly evenly across all age groups, smokers are slightly more concentrated in the younger-to-middle age brackets.



- **Regional Charges:** On average, the **southeast region** exhibits the highest insurance charges. However, the high variance (indicated by the error bars) suggests that other factors, like the higher prevalence of smokers or higher BMI in this region, are likely the underlying cause rather than geography alone.

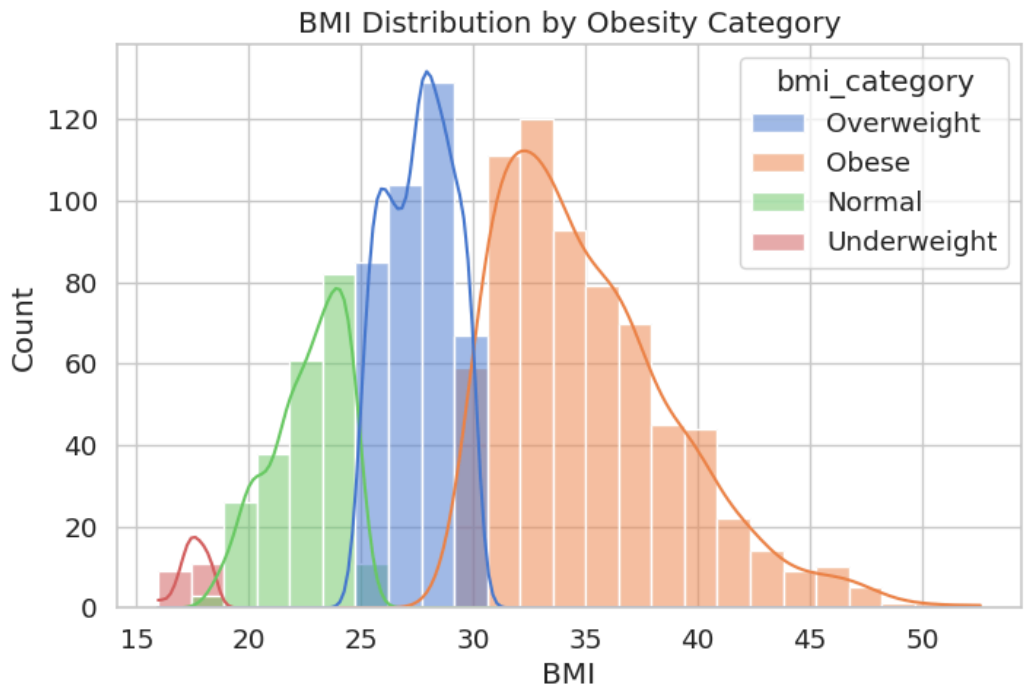


- **Charges by Sex:** While males show slightly higher average charges than females, the difference is not statistically significant due to the high variability in charges within each group.



4.4 The Role of Body Mass Index (BMI)

BMI is another critical factor influencing medical costs. The data shows a clear trend where higher BMI categories are associated with higher charges. The distribution plot below illustrates how policyholders are segmented into the different BMI categories.



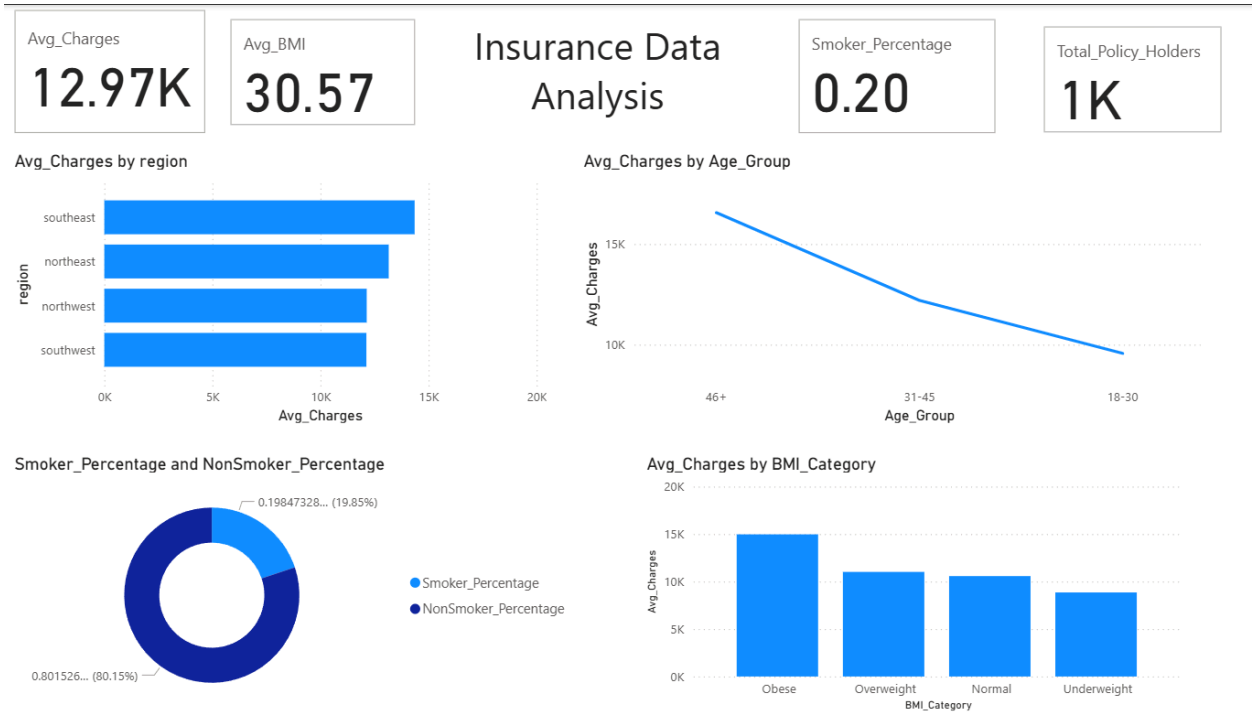
---

5. Interactive Dashboard Overview

To provide stakeholders with a tool for self-service analysis, an interactive Power BI dashboard was created. The dashboard consolidates the key findings into a user-friendly interface.

5.1 High-Level KPI Dashboard

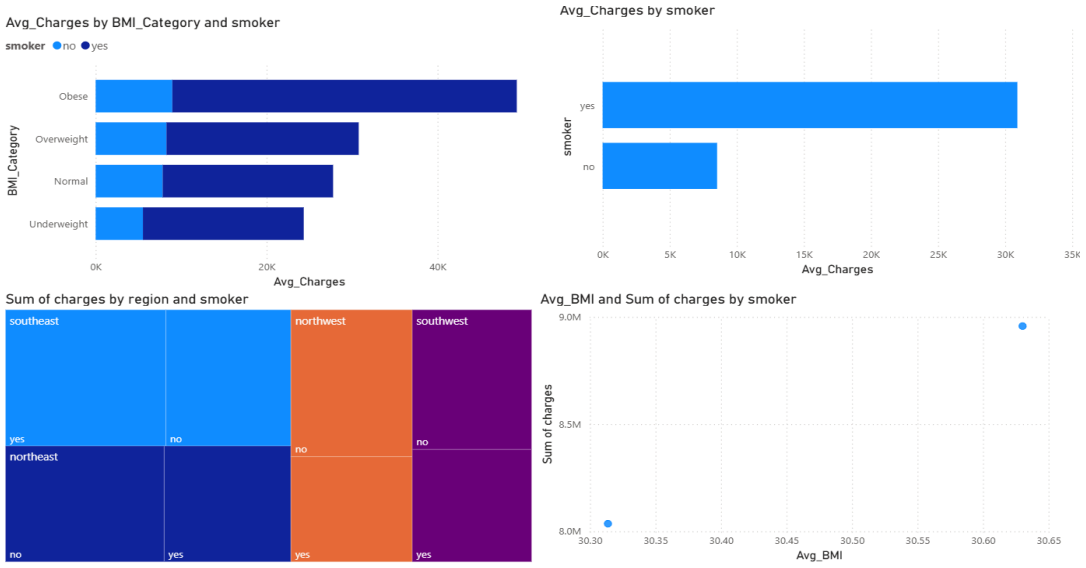
The main dashboard presents key performance indicators (KPIs) at a glance, including overall average charges, average BMI, the percentage of smokers, and the total number of policyholders. It also features visualizations that summarize average charges by region, age group, and BMI category.



5.2 Deep-Dive Analysis Dashboard

A second dashboard page allows for a more granular analysis, focusing on the combined effects of multiple factors. Key visuals include:

- **Charges by BMI & Smoker Status:** This chart powerfully illustrates that obese smokers face exponentially higher costs than any other group.
- **Regional Charges by Smoker Status:** A treemap visualizes the total sum of charges, highlighting that smokers in the southeast region are a major cost center.



---

6. Conclusion & Recommendations

This analysis concludes that policyholder lifestyle and health metrics are far more influential on medical costs than simple demographics like region or sex. **Smoking is the single most powerful predictor of high insurance charges.** This effect is significantly amplified when combined with a high BMI.

Based on these findings, we recommend the following actions:

1. **Re-evaluate Premium Models:** The current pricing structure should be reviewed to ensure it adequately reflects the high risk associated with smokers. Implementing more significant premium differentials for smokers could lead to a more balanced and profitable risk pool.
2. **Launch Targeted Wellness Initiatives:** Develop and promote wellness programs aimed at smoking cessation and weight management. Offering incentives, such as premium reductions for program completion, could encourage healthier lifestyles and reduce long-term costs.
3. **Data-Driven Marketing:** Utilize these insights to create targeted marketing campaigns. Emphasize the significant cost savings available to non-smokers and individuals with a healthy BMI, potentially attracting lower-risk applicants.