Health Insurance Analysis – Advanced Report

This report provides an end■to■end analysis of a health insurance dataset using R for data cleaning and statistical exploration, and Power BI for interactive visual analytics. The goal of the project is to uncover the key drivers of insurance charges, understand risk patterns, and present insights useful for decision■making in healthcare and insurance industries.

1. Dataset Overview

The dataset contains customer attributes such as age, sex, BMI, number of children, region, smoking status, and medical charges. A cleaned version of the dataset was created (insurance_clean.csv) using R, where additional categorical fields such as age_group and bmi_category were added to enhance analysis.

2. Data Cleaning & Feature Engineering (R)

Using R, the following steps were performed: • Checked structure and summary statistics. • Verified missing values (none present). • Created engineered fields: – age_group (Young, Middle■aged, Senior) – bmi_category (Underweight, Normal, Overweight, Obese) • Exported the cleaned dataset for Power BI use. A full R script is included in the GitHub repository.

3. Exploratory Data Analysis (R Visualization)

Several visualizations were created in R: • Average charges by smoking status showed smokers have significantly higher medical charges. • Regional boxplots revealed variation in cost distribution across regions. • BMI vs Charges scatter plot indicated a positive relationship, especially among smokers. • Correlation heatmap identified strong relationships between charges, age, and BMI.

4. Power BI Dashboard Development

A Power BI dashboard was built using: • Bar charts to show charge distribution. • Donut charts for categorical comparisons. • Scatter charts to illustrate feature relationships. • A custom Risk Score measure created using DAX. The dashboard provides interactive filtering and drill■down insights.

5. Key Insights

From both R and Power BI analysis: • Smoking is the strongest driver of high insurance charges. • BMI is positively associated with charges, especially in higher BMI categories. • Older individuals (Senior age group) tend to incur higher costs. • The Southeast region shows higher variability in charges. • Combining age_group, smoker, and BMI provides a strong foundation for risk■scoring models.

6. Conclusion

The project demonstrates practical skills in: • Data cleaning and transformation using R. • Feature engineering for analytics. • Visualization using ggplot2. • Dashboard development using Power BI. • Interpretive insight generation suitable for business stakeholders. This makes the project portfolio■ready and showcases capabilities in real■world analytics.

7. Appendix

Additional supporting code, raw data, and high■resolution dashboard screenshots are included in the GitHub repository.