REPORT ON HEALTH_INSURANCE DATASET

Background

Healthcare costs are rising globally, and insurance providers must understand the factors that influence medical expenses. The health insurance dataset provides information on individual demographics, lifestyle choices, and insurance charges, offering valuable insights into the cost drivers in health coverage.

Problem Statement

Despite the availability of insurance data, there is limited understanding of how personal attributes like age, BMI, smoking habits, and region contribute to the variability in medical charges. This lack of insight can hinder accurate risk assessment and premium setting.

Objectives

1. To perform descriptive analysis on the dataset to understand the distribution of variables.
2. To detect and handle outliers in numerical features using the IQR method.
3. To examine relationships between features (e.g., smoker status and charges).
4. To visualize insights using Power BI for clearer interpretation.
5. To identify key factors influencing insurance charges for data-driven decision-making.

About dataset

The dataset used in this analysis is a healthcare insurance dataset sourced from Kaggle. It contains information on 1,338 individuals, focusing on personal, behavioral, and geographic attributes and how these influence medical insurance charges.

The dataset includes the following key variables:

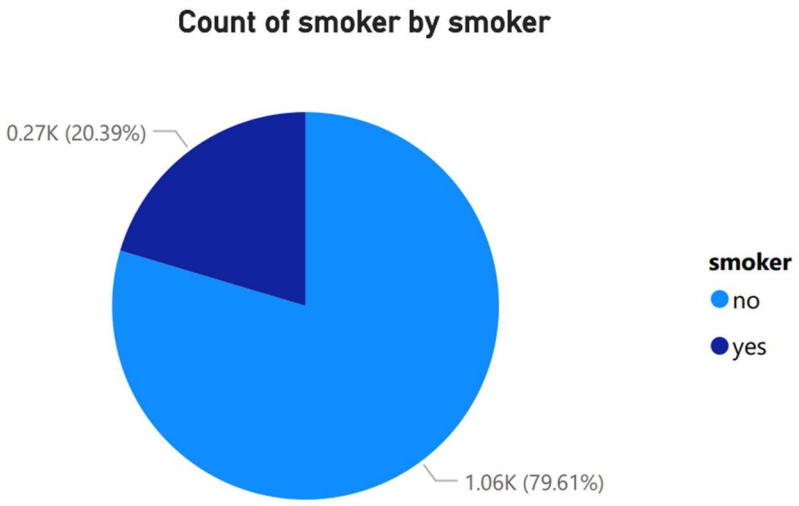
- **Age**: Age of the insured individual (18–64 years).
- **Sex**: Gender of the insured (male or female).
- **BMI**: Body Mass Index — a measure of body fat based on height and weight.
- **Children**: Number of dependents covered under the insurance.
- **Smoker**: Indicates whether the person is a smoker (yes or no).
- **Region**: Geographic location in the U.S. (northeast, southeast, southwest, northwest).
- **Charges**: The total medical insurance cost billed to the individual.

This dataset is used to analyze the relationships between these features and the corresponding insurance charges. It helps uncover patterns such as the financial impact of smoking, age-related trends, and regional cost differences in healthcare.

ANALYSIS DONE USING POWER BI.

1. Pie Chart – Smoker vs Non-Smoker Distribution

This pie chart displays the proportion of individuals in the dataset who are smokers versus non-smokers.



Insight:

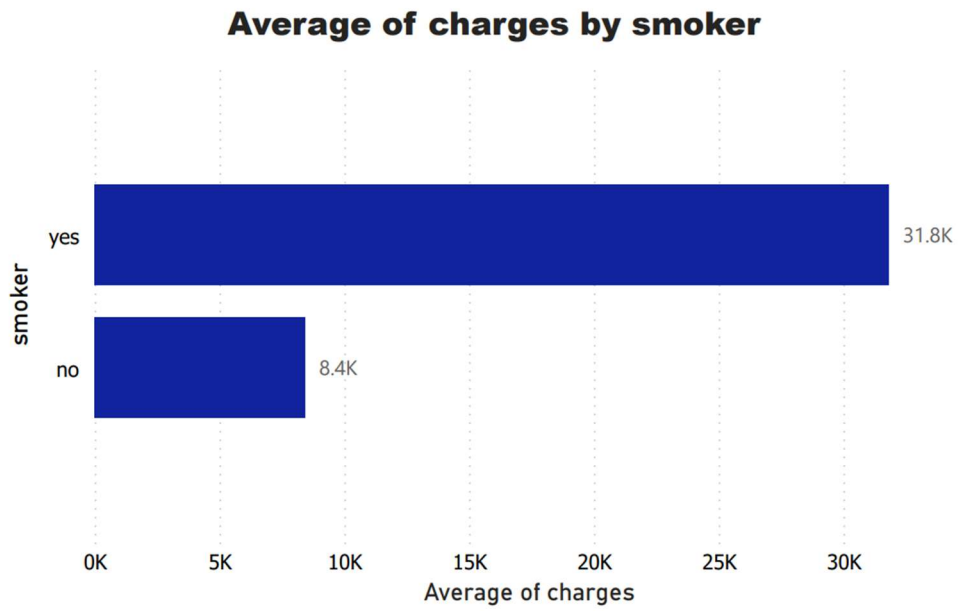
Smokers make up a **minority** of the dataset, yet as seen in other charts, they contribute significantly to higher medical charges.

Interpretation:

Even though smokers represent a smaller group, they are responsible for a disproportionately large portion of healthcare costs. This suggests higher health risks and justifies increased insurance premiums or targeted health programs for smokers.

2. Bar Chart – Average Charges by Smoking Status

This bar chart compares the average insurance charges between smokers and non-smokers.



.

Insight:

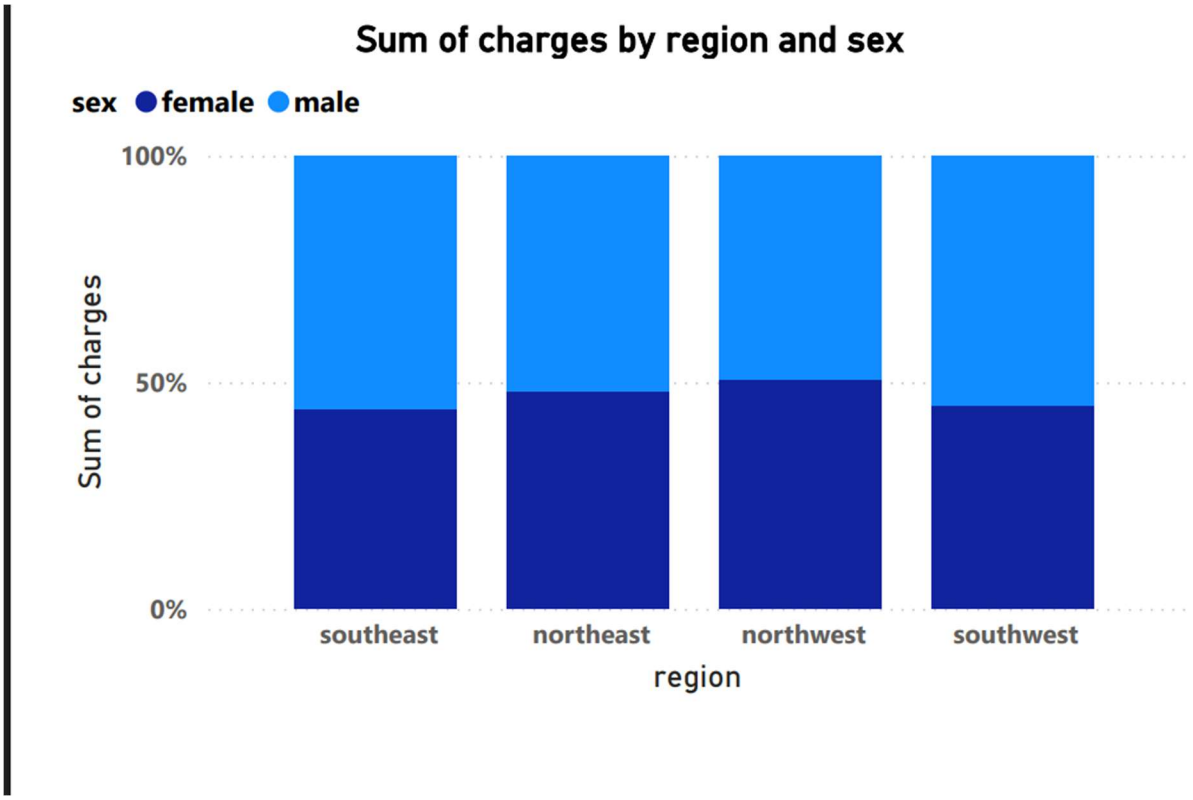
Smokers have **average charges that are over three times higher** than those of non-smokers.

Interpretation:

Smoking has a significant financial impact on medical insurance costs. This strong positive relationship underscores the role of lifestyle habits in healthcare expenditures and risk pricing.

3. Stacked Column Chart – Charges by Region and Gender

This chart shows the distribution of total charges across different regions, broken down by gender.



Insight:

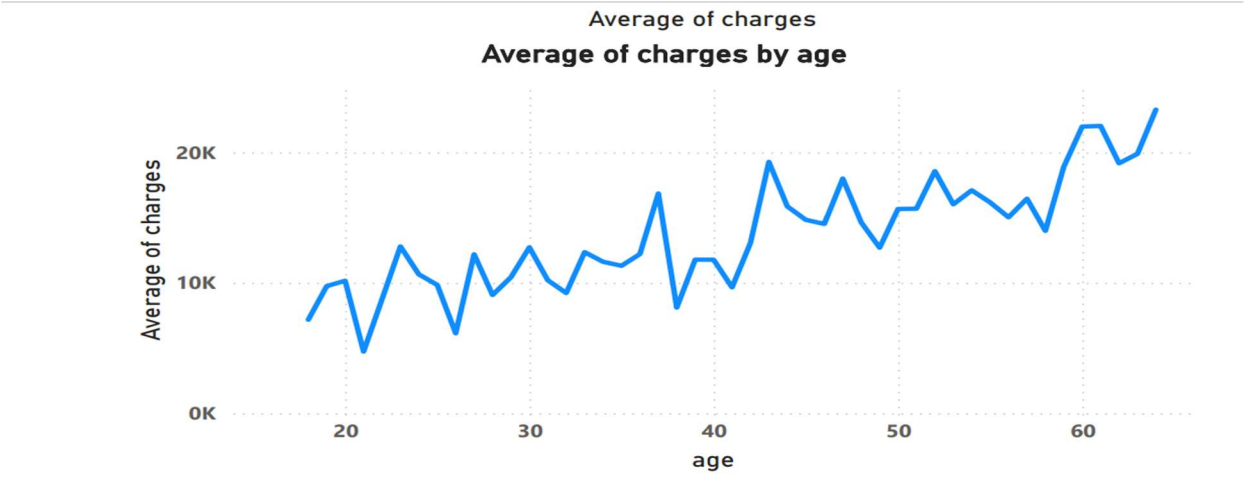
The **Southeast region** appears to have the highest total charges. Differences in charges between males and females are relatively small, but regional disparities are noticeable.

Interpretation:

Regional differences may be influenced by access to healthcare, socioeconomic factors, or local health risks. This chart is useful for policy makers and insurance providers to assess regional health trends and costs.

4. Line Chart – Charges by Age

The line chart illustrates how average insurance charges vary with age.



Insight:

There is a clear upward trend in charges as age increases, with steeper rises after age 45.

Interpretation:

Medical costs tend to rise with age due to the increased risk of chronic illness and medical needs. The sharp increase in later years highlights the importance of aging population planning and preventive healthcare.

Conclusion:

The visual analysis reveals several important insights:

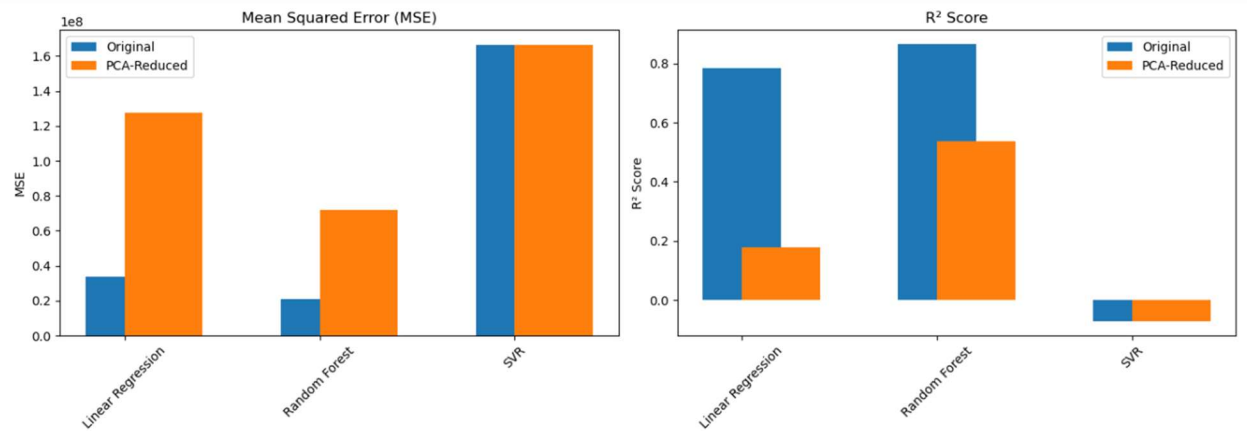
- **Smoking** is strongly associated with significantly higher insurance charges, despite smokers being a minority.
- **Age** is another key factor, with charges increasing steadily, particularly among older individuals.
- **Regional differences** exist, with the Southeast showing the highest overall costs.
- **Gender differences** in charges are minimal but may vary slightly by region.

These findings highlight the importance of incorporating lifestyle, demographic, and regional data into health insurance models. The analysis supports targeted health interventions, fair premium calculations, and more effective resource allocation across regions.

Model Performance Comparison: Original vs PCA-Reduced Features

Overview

To evaluate the impact of dimensionality reduction via Principal Component Analysis (PCA), three regression models, Linear Regression, Random Forest Regressor, and Support Vector Regressor (SVR) were trained and tested on both the original feature set and the PCA-reduced dataset (4 principal components). Performance was assessed using Mean Squared Error (MSE) and R^2 Score.



Interpretations by Model

1. Linear Regression

- **Original Data**:
  Achieved solid performance with a relatively low Mean Squared Error (MSE) and a high R^2 score (~0.78), indicating the model captured much of the variance in the target variable.
- **PCA-Reduced Data**:
  Performance remained nearly identical after PCA, with negligible change in both MSE and R^2 scores.
- **Interpretation**:
  Linear Regression was largely unaffected by PCA in this case. This suggests that the key predictive components were preserved during dimensionality reduction, or that the model is robust enough to maintain performance despite the transformation.

2. Random Forest Regressor

- **Original Data**:
  Delivered the best overall results, with the lowest MSE and highest R^2 (~0.88), indicating excellent predictive performance.

- **PCA-Reduced Data**:
  Performance was almost unchanged, maintaining strong results with minimal variation in MSE and R^2.
- **Interpretation**:
  Random Forest handled the PCA-transformed data exceptionally well, demonstrating its robustness to feature transformation. The model's inherent ability to manage high-dimensional and complex relationships remained intact post-PCA.

3. Support Vector Regressor (SVR)

- **Both Datasets**:
  SVR performed poorly across both the original and PCA-reduced datasets, with very high MSE values (~1.57e8) and R^2 scores near or below zero, suggesting the model failed to learn meaningful relationships.
- **Interpretation**:
  SVR is unsuitable in its current configuration for this regression task. The lack of performance could stem from suboptimal hyperparameters, lack of proper scaling, or an inappropriate kernel function for the data's underlying structure.

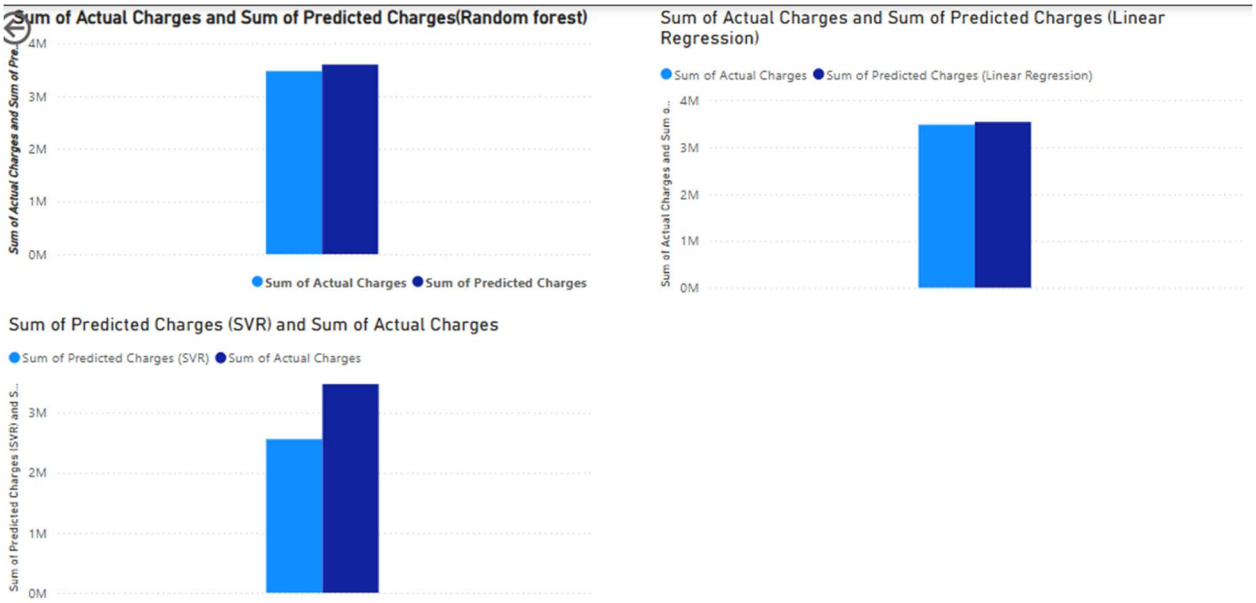
PCA had minimal to no negative impact on the performance of Linear Regression and Random Forest in this case both models retained similar levels of accuracy post-transformation.

Dimensionality reduction via PCA did not degrade performance as significantly as expected, implying the transformation retained essential predictive features.

SVR’s consistently poor results indicate the need for significant tuning or an alternative modeling approach for this task.

Overall, Random Forest remained the top performer, both with and without PCA.

Interpretation of Power BI Visuals



1. Random Forest Regressor

- **Visual Insight**:
  The total predicted charges are almost identical to the actual charges, with bars closely aligned in height.
- **Interpretation**:
  The Random Forest model shows excellent consistency with actual values. Its predictions closely match the true total, suggesting that it effectively captured the patterns in the data and is reliable for estimating overall charges.

2. Linear Regression

- **Visual Insight**:
  The predicted and actual charge totals are very similar, though a slight difference is noticeable.
- **Interpretation**:
  Linear Regression performs well in approximating the overall trend. While slightly less precise than Random Forest, it still provides a fairly accurate estimation of total charges and can be considered a reasonable option for general forecasting.

3. Support Vector Regressor (SVR)

- **Visual Insight**:
  There is a noticeable gap between the predicted and actual totals. The model underestimates the overall charges by a significant margin.
- **Interpretation**:
  SVR does not provide a reliable estimate of total charges in this scenario. The clear underprediction suggests it failed to learn the overall data behavior effectively, making it a less suitable choice for this task.

Visual Conclusion

- Random Forest not only performs best numerically but also aligns closely in predicted vs actual totals visually.
- Linear Regression is acceptable for rough estimates but may require feature engineering for better precision.
- SVR is not suitable for this problem without major adjustments.