Report:

» In this data-driven analysis project, we explore a large insurance dataset to uncover the key drivers of insurance costs. By conducting descriptive analyses, correlation analysis, and regression modelling, we aim to identify the factors that strongly influence insurance charges. Additionally, we develop predictive models to estimate charges for new patients. The dataset includes information such as age, sex, body mass index (BMI), number of children, smoking status, region, and insurance charges. We pre-process the data by encoding categorical variables and categorizing BMI into different weight categories. Our analysis reveals insights into the distribution of charges, the prevalence of smoking among patients, the demographic composition of the dataset, and the relationship between the number of children and charges. The findings from this study can provide valuable insights for insurance companies in understanding the factors that contribute to insurance costs.

» **Introduction:**

Insurance costs play a crucial role in the healthcare industry, impacting individuals, families, and insurance providers. Understanding the drivers behind these costs is essential for insurance companies to make informed decisions, develop accurate pricing models, and assess risk. In this project, we delve into a comprehensive insurance dataset to identify the key factors influencing insurance charges. By utilizing descriptive analyses, correlation analysis, and regression modeling techniques, we aim to uncover insights that will help insurance companies better understand the drivers of insurance costs.

» **Data Pre-processing:**

Our analysis begins with data pre-processing steps to ensure the dataset's readiness for analysis. We import necessary libraries such as NumPy, Pandas, Matplotlib, and Seaborn. After loading the insurance dataset, we examine its structure and content. The dataset contains information about various attributes, including age, sex, BMI, number of children, smoking status, region, and insurance charges. To facilitate our analysis, we handle missing values, check for duplicates, and convert categorical variables into numerical codes. Additionally, we categorize BMI into different weight categories to gain insights into the relationship between weight and insurance charges.

» **Descriptive Analyses:**

To gain a comprehensive understanding of the data, we perform descriptive analyses. We explore the distribution of charges through visualization, examining any potential outliers that may impact our analysis. Furthermore, we investigate the

distribution of sex, smoking status, and regions among the dataset's individuals. These visualizations provide valuable insights into the demographic composition of the dataset.

» **Correlation Analysis and Regression Modelling:**

To identify the factors that strongly influence insurance charges, we conduct correlation analysis and regression modelling. By calculating the correlation matrix, we assess the relationships between variables. We visualize the correlation matrix using a heat-map, which highlights the strength and direction of these relationships. Subsequently, we develop regression models to estimate charges based on various independent variables. These models enable us to quantify the impact of each factor on insurance costs and identify the most significant drivers.

» **Predictive Modelling:**

In addition to analysing the existing data, we develop predictive models to estimate charges for new patients. These models leverage the relationships and insights gained from our analysis to provide accurate cost estimates. By applying these models to new patient data, insurance companies can make more informed pricing decisions and improve their risk assessment capabilities.

» **Conclusion:**

Through this data-driven analysis, we uncover the key drivers of insurance costs using a large insurance dataset. Our findings provide insights into the distribution of charges, the prevalence of smoking among patients, the demographic composition of the dataset, and the relationship between the number of children and charges. Insurance companies can leverage these insights to optimize their pricing models, assess risks more effectively, and make informed decisions. By understanding the factors that contribute to insurance costs, insurance providers can better meet the needs of their customers while maintaining a sustainable and profitable business.