**Healthcare Insurance**

Healthcare insurance is a critical area that deals with risk management, cost prediction, and operational efficiency in policy design. Insights derived from this project will not only help ABC Insurance optimize their premium structures but also educate policyholders about healthier lifestyle choices.

The objective of this project is to analyze an insurance dataset using Python and uncover insights related to premium charges and their influencing factors. This dataset includes information such as age, gender, BMI, number of children, smoking habits, region, and insurance charges.

By using Python libraries such as Pandas, Numpy, Matplotlib, and Seaborn, the insurance.csv dataset is loaded into a variable called data. This dataset contains details about individuals (like age, BMI, number of children, etc.) and their insurance charges.

* **Shape of Data**: The dataset has 1338 rows and 7 columns.
* **Data types**: Helps us identify whether columns are numerical (age, BMI, children, and charges) or categorical (sex, smoker, and region).
* **First and last rows**: Using .head() and .tail(), we can see the first and last 5 rows to understand how the data looks.
* **Missing Values**: No missing values were detected in the dataset.

Age ranges from 18 to 64 years. BMI (Body Mass Index) ranges from 15.96 to 53.13.

Charges have a wide range from $1,121.87 to $63,770.43. Most individuals have 0 to 2 children.

**Five-Point Summary:**

Charges has a wide range, indicating significant variability in insurance costs.

BMI distribution suggests the dataset includes individuals with varied body mass indices.

**Scatter Plots:**

**Age vs Charges**:

* As age increases, insurance charges tend to rise, likely due to higher health risks with aging.
* Smokers generally have higher charges than non-smokers in all age groups.

**BMI vs Charges**:

* Individuals with higher BMI tend to have higher charges, particularly smokers.
* For non-smokers, the correlation between BMI and charges is less pronounced.

**Children vs Charges**:

* No strong correlation between the number of children and charges is observed.

**Count Plots:**

**Sex Distribution**:

* The dataset has a relatively balanced number of males and females.
* Both genders show similar patterns in insurance charges.

**Smoker Distribution**:

* The number of non-smokers is significantly higher than smokers.
* Smokers incur much higher charges.

**Region Distribution**:

* The dataset is evenly distributed across the four regions (northeast, northwest, southeast, southwest).

**Pairplot:**

* The pairplot shows clear separations in charges based on smoker status.
* Smokers have distinct clusters with higher charges across all numerical features.

**Histograms:**

**BMI**:

Normally distributed with most individuals having BMI between 20–35.

**Age**:

Uniform distribution, with individuals across all age groups represented.

**Children**:

Most policyholders have 0–2 children, with fewer having 3 or more.

**Charges**:

Right-skewed distribution indicates that most policyholders incur lower charges, with a few outliers incurring very high charges.

**Boxplots:**

**BMI** : Few outliers are present, indicating extreme cases.

**Charges**: Many high outliers correspond to smokers with higher insurance costs.

Age does not show significant outliers, while `children` is discrete.

**Heatmap:** Moderate positive correlation between:

Age and charges (older individuals tend to have higher charges). BMI and charges (higher BMI leads to higher charges). Children shows little to no correlation with charges.

**Premium Charges for Smokers vs Non-Smokers**

**Age Trends**:

**Smokers**: Charges increase sharply with age. Smoking adds significant costs, reflecting higher health risks and claims.

**Non-Smokers**: Charges increase more gradually with age, suggesting lower health risks.

**Impact of Smoking**:

At all age groups, smokers incur consistently higher charges compared to non-smokers.

The gap between smokers and non-smokers widens as age increases.

**Insights:**

**Smoking is the Strongest Predictor**:

* Smokers incur significantly higher charges compared to non-smokers.

**BMI and Age Influence Premiums**:

* Higher BMI and older age groups are associated with increased charges.
* Encouraging healthy lifestyles (e.g., weight management programs) could help reduce premiums.

**Children and Region Have Minimal Impact**:

* Neither the number of children nor the region significantly affects insurance charges.

**Target Audience for Cost Optimization**:

* Smokers and individuals with high BMI are key targets for cost-reduction strategies. Tailored health plans could lower claims and improve customer outcomes.