**PROJECT DOCUMENTATION**

**DATE:** 20/12/2023

**PROJECT TITLE:** Health Insurance Analysis

**OBJECTIVE:** The goal of this data analysis project is to gain insights into the factors influencing health insurance charges. We aim to identify patterns, correlations, and trends that can help in understanding the key drivers of health insurance costs.

**INTRODUCTION:**

In an era where healthcare costs continue to be a significant concern, understanding the underlying factors that contribute to these expenses is crucial for informed decision-making and resource allocation. This data analysis project aims to shed light on the intricacies of healthcare charges by leveraging the power of Power BI and delving into a comprehensive dataset. The dataset encompasses diverse demographic and lifestyle factors, including age, sex, BMI, number of children, smoking status, region, and healthcare charges.

**DATASET DESCRIPTION:**

Age: age of the primary beneficiary

Sex: insurance contractor gender, female, male

BMI: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9

Children: Number of children covered by health insurance / Number of dependents

Smoker: Smoking

Region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.

Charges: Individual medical costs billed by health insurance

**TOOLS**

1. **Microsoft Excel:** Data Exploration, Data Cleaning and Preparation, Data Visualization, and Basic Analysis.
2. **Microsoft Power Query:** Data Transformation, and Data Integration.
3. **Microsoft Word:** Documentation and Reporting.
4. **Microsoft Power BI:** Data Visualization and dashboard.
5. **Microsoft PowerPoint:** Presentation of Findings.
6. **Microsoft Teams:** Collaboration, and Sharing Reports.

**METHODOLOGY**

* Data Exploration
* Data Cleaning and Preparation
* Data Visualization and Dashboard
* Insights and Recommendations
* Conclusion

**DATA EXPLORATION:**

* Dataset was loaded into PowerBI.
* Examine the number of rows and columns in the dataset to understand its size.

Number of rows: 1339

Number of columns: 7

* Review the data types of each column (e.g., numeric, categorical) to ensure they are assigned correctly.
* Calculate basic summary statistics (mean, median, standard deviation, min, max) for numerical columns to understand the central tendency and variability of the data.

Summary statistics of age column:

* 1. Mean,
  2. Median,
  3. Mode,
  4. Range,
  5. Q1,
  6. Q2,
  7. Q3,
  8. Q4,
  9. Standard deviation,
  10. Variance,
  11. Histogram,
  12. Standardization with Z-test,
  13. Hypothesis testing (Null, alternative, and T-test),

Summary statistics of BMI column:

* 1. Mean,
  2. Median,
  3. Mode,
  4. Range,
  5. Box plot,
  6. Q1,
  7. Q2,
  8. Q3,
  9. Q4,
  10. Standard deviation,
  11. Variance,
  12. Histogram,
  13. Standardization with Z-test,
  14. Hypothesis testing (Null, alternative, and T-test),

Summary statistics of children column:

* 1. Mean,
  2. Median,
  3. Mode,
  4. Range,
  5. Box plot,
  6. Q1,
  7. Q2,
  8. Q3,
  9. Q4,
  10. Standard deviation,
  11. Variance,
  12. Histogram,
  13. Standardization with Z-test,
  14. Hypothesis testing (Null, alternative, and T-test),

Summary statistics of charges column:

* 1. Mean,
  2. Median,
  3. Mode,
  4. Range,
  5. Box plot,
  6. Q1,
  7. Q2,
  8. Q3,
  9. Q4,
  10. Standard deviation,
  11. Variance,
  12. Histogram,
  13. Standardization with Z-test,
  14. Hypothesis testing (Null, alternative, and T-test),
* Count the unique values in categorical columns to understand the diversity of the data and identify potential issues.

Count of unique values in sex column:

Male:

Female:

Count of unique values in the smoker column:

Smokers:

Non-smokers:

Count of unique values in the region column:

Southwest:

Southeast:

Northwest:

Northwest:

* Create **histograms or box plots** to visualize the distribution of numerical variables. This helps identify skewness or outliers.
* Use visualizations or statistical methods to detect **outliers** in numerical data. Decide whether to keep, transform, or remove outliers based on the context of the analysis.

Histograms and box plots of Age

Histograms and box plots of BMI

Histograms and box plots of children

Histograms and box plots of charges

* Use **scatter plots, correlation matrices, or pair plots** to investigate relationships between numerical variables.

Relationships between Age and BMI:

Relationships between Age and children:

Relationships between Age and charges:

Relationships between BMI and children:

Relationships between BMI and charges:

Relationships between children and charges:

**DATA CLEANING:**

* No missing or null values

1. **Demographic Analysis:**
   * Explore the distribution of **charges** based on different categories (e.g., male vs. female, smoker vs. non-smoker, region).
   * Use visualizations like scatter plots or correlation matrices.
   * Analyze the distribution of charges based on demographics (age, sex, region).
   * Explore the impact of having children on healthcare charges.
2. **Smoking Status Analysis:**
   * Compare healthcare charges between smokers and non-smokers.
   * Use visualizations to highlight differences in charges.
3. **BMI Analysis:**
   * Explore the relationship between BMI and healthcare charges.
   * Group individuals into **BMI categories** and analyze charge distributions.
4. **Region-wise Analysis:**
   * Compare healthcare charges across different regions.
   * Identify any regional trends or variations.
5. **Multivariate Analysis:**
   * Conduct multivariate analysis considering multiple factors simultaneously.
   * Use visualizations or statistical models to identify the combined impact of variables on healthcare charges.
6. **Dashboard Creation:**
   * Design a Power BI dashboard to present key findings.
   * Include interactive visualizations for stakeholders to explore data dynamically.
7. **Insights and Recommendations:**
   * Summarize key insights gained from the analysis.
   * Provide recommendations based on the identified patterns and trends.

**DATA VISUALIZATION AND DASHBOARD**

**DATA INSIGHTS AND RECOMMENDATIONS**