Health Insurance Data Analysis

Kalyanapu Raja(rkalyana)

EE 590 Basics of Data Analysis and Machine Learning

Introduction

This report presents a comprehensive analysis of a health insurance dataset aimed at uncovering insights into various factors affecting insurance costs. The dataset includes several variables such as Age, Gender, BMI, Number of Children, Smoking Status, Region, and Insurance Expenses.

Data Preparation

The initial step involved loading the data from a CSV file into a Pandas Data Frame. Preliminary data inspection revealed a well-structured dataset with a mix of categorical and numerical variables.

Exploratory Data Analysis

Key findings from the exploratory analysis include:

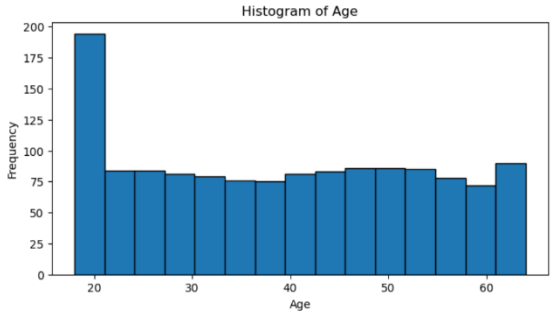
Visualizations:

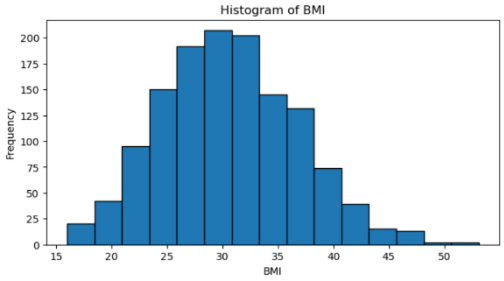
Histograms for numerical variables (Age, BMI, Children, Expenses) revealed their distribution characteristics.

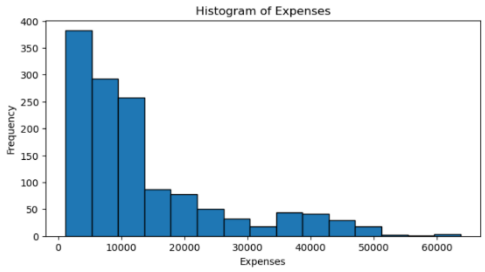
Bar charts for categorical variables (Gender, Smoker, Region) showed the count distribution among different categories.

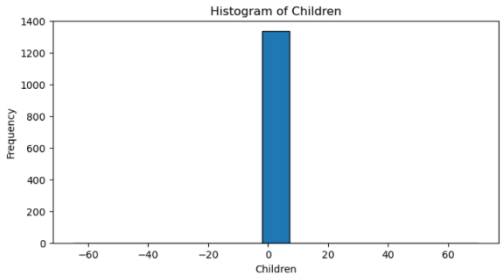
Insights:

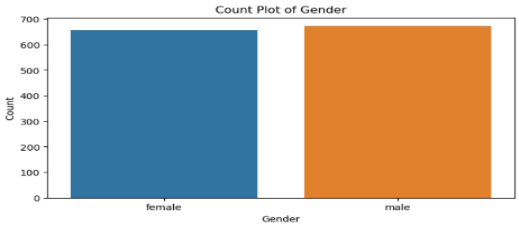
Age and BMI appeared normally distributed. Most insured individuals are non-smokers. Insurance costs (Expenses) exhibit a right-skewed distribution, indicating variability in insurance charges.

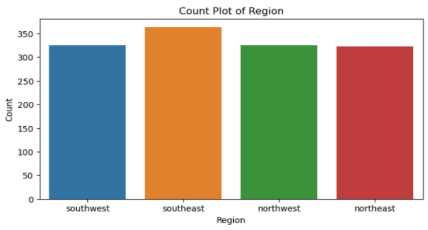












Descriptive Statistics and Variability Measures

Descriptive statistics provided an overview of central tendencies and dispersion in numerical data:

Age had a mean around [64.000] years, with a standard deviation of [14.04].

BMI indicated a mean value of [53.1000], suggesting an average toward the higher side of the BMI scale.

Quartile Analysis

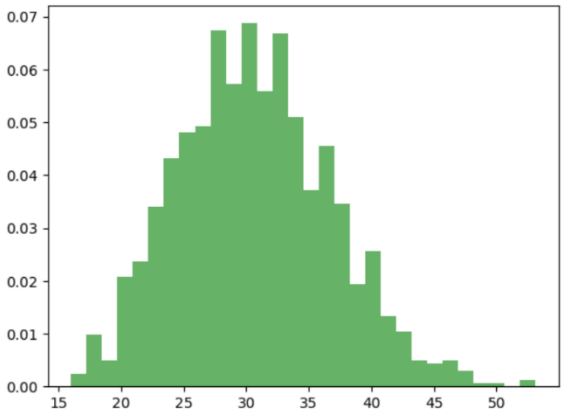
Quartile analysis for numerical variables highlighted:

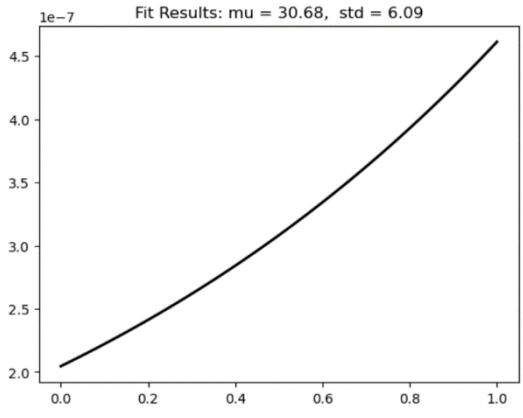
The median age of the population is 39.00.

The third quartile for BMI indicates that 75% of individuals have a BMI less than 34.650.

Probability Distribution Modelling

Probability distribution models, particularly the normal distribution, were fitted to columns like BMI. The distribution fit for BMI closely followed the theoretical normal distribution curve.





Feature Scaling

Feature scaling was performed using standardization techniques, which brought all numerical variables to a common scale, facilitating more efficient modeling.

Linear Regression Modeling

A linear regression model was developed with 'Expenses' as the target variable. The model aimed to understand the impact of various factors on insurance costs:

Model Evaluation

The linear regression model was evaluated using metrics like Mean Squared Error (MSE) and R-squared (R^2):

```
Mean Squared Error: 35405134.87719506
R^2 Score: 0.7608041536118314
```

Conclusion

The analysis revealed significant insights into health insurance data, highlighting key factors that influence insurance costs. This understanding could be pivotal for insurance companies in policy making and risk assessment.