Report: Healthcare Insurance Analysis

Safayatul Islam
12th February, 2024
California State University Northridge

The purpose of this report is to analyze a healthcare insurance dataset containing 1,338 records and 7 attributes. The objective is to explore key trends, identify influential factors affecting insurance charges, and provide insights into data distribution and correlations.

Dataset Overview

The dataset consists of the following attributes:

- **Age**: Age of the insured individual.
- **Sex**: Gender of the individual (Male/Female).
- **BMI**: Body Mass Index, a measure of body fat.
- **Children**: Number of dependents covered under the insurance.
- **Smoker**: Indicates whether the individual is a smoker (Yes/No).
- **Region**: Geographic region where the policyholder resides.
- **Charges**: The medical insurance cost charged to the individual.

Missing Values Check

- No missing values were found in the dataset, ensuring the reliability of the data for analysis.

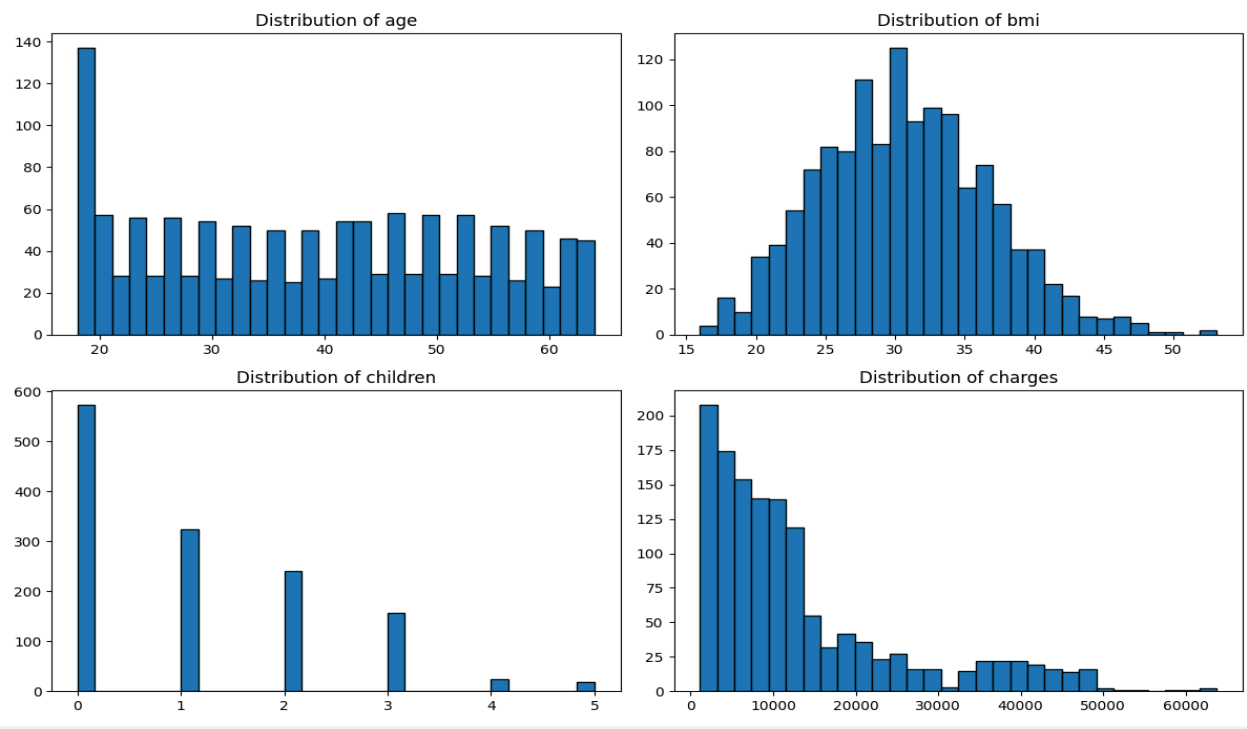
Summary Statistics

Summary Statistics

		age	bmi	children	charges
1	count	1338.0	1338.0	1338.0	1338.0
2	mean	39.207025411061 29	30.663396860986 55	1.0949177877429	13270.422265141 257
3	std	14.049960379216 154	6.09818691167901 4	1.20549273978191 4	12110.011236694 001
4	min	18.0	15.96	0.0	1121.8739
5	25%	27.0	26.29625	0.0	4740.28715
6	50%	39.0	30.4	1.0	9382.033
7	75%	51.0	34.69375	2.0	16639.912515
8	max	64.0	53.13	5.0	63770.42801

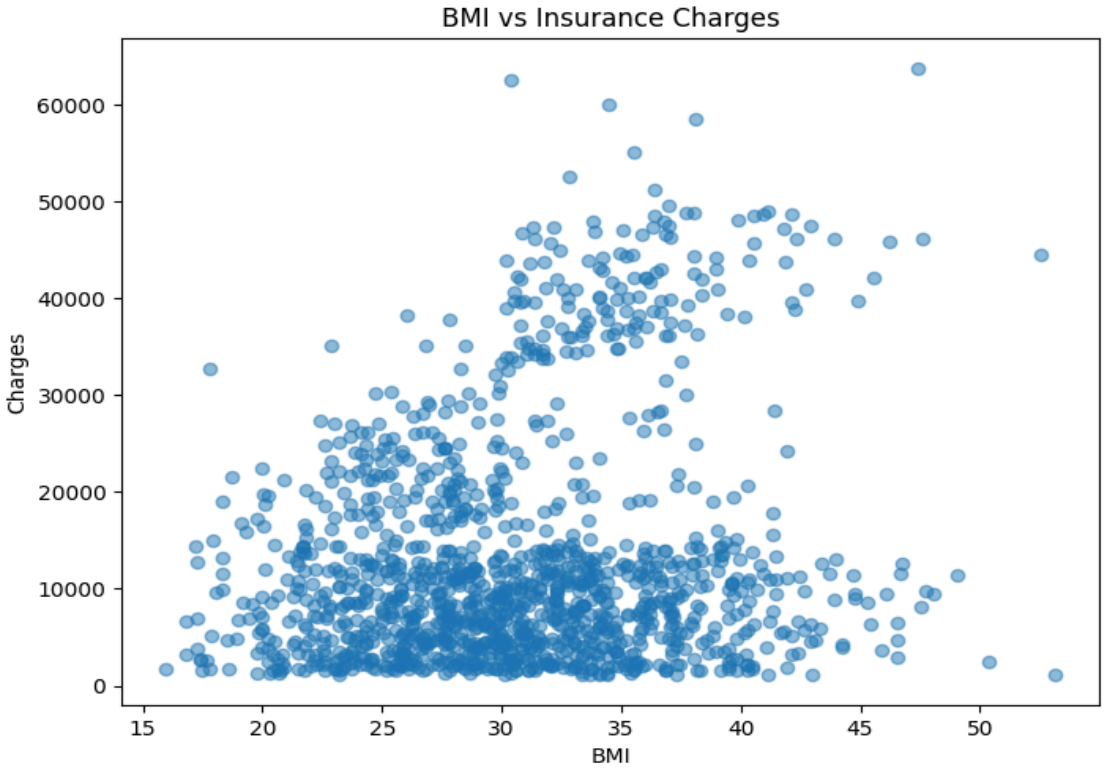
- **Age**: Ranges from **18 to 64 years**, with an average of **39.2 years**.
- **BMI**: Varies between **15.96 and 53.13**, with a mean of **30.66**.
- **Children**: Number of dependents varies from 0 to 5, averaging 1.09 per individual.
- **Charges**: Insurance costs range from $1,121.87 to $63,770.42, with an average charge of $13,270.42.

Data Distribution Analysis.



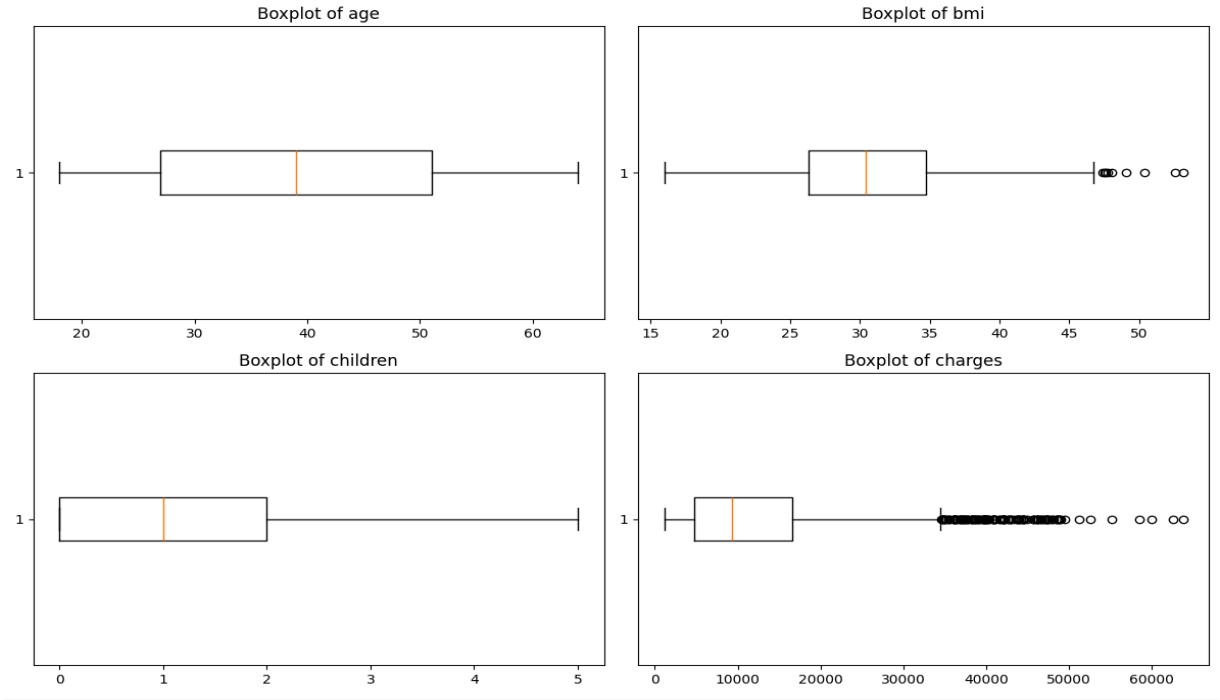
- The **age distribution** appears relatively uniform across different age groups.
- The **BMI distribution** is slightly right-skewed, with some individuals having extremely high BMI values.
- The **number of children** covered under insurance policies is concentrated between 0-2.
- **Insurance charges** are highly skewed, with a small subset of individuals incurring significantly high costs.

Correlation Analysis



- **Insurance charges strongly correlate with smoking status and BMI**, indicating that these factors have a significant impact on costs.
- **Age and BMI show a moderate correlation with charges**, suggesting that older individuals and those with higher BMI tend to incur higher costs.

Key Insights



2. **Higher BMI is associated with increased insurance charges**: Individuals with obesity tend to incur higher healthcare expenses.
3. **Outliers in insurance costs**: Some individuals have extremely high insurance charges, likely due to lifestyle choices or chronic health conditions.

This analysis demonstrates that lifestyle factors, such as smoking and obesity (high BMI), have a significant impact on insurance costs. The presence of outliers in charges suggests that certain individuals are incurring substantially higher medical expenses, likely due to riskier health conditions.

Future work could involve predictive modeling to estimate insurance costs based on personal health metrics and demographics.

Recommendations & Next Steps

- Develop a predictive model to estimate insurance costs based on age, BMI, and smoking status.
- Investigate regional disparities in insurance costs to determine location-based pricing trends.
- Analyze outliers further to understand extreme variations in insurance charges.

This Exploratory Data Analysis (EDA) provides critical insights that could assist insurance companies in refining risk assessment strategies and pricing models.