**Factors Affecting Health Insurance Charges for XYZ Insurance Company**

**Varsha Tomar**

**Introduction**

The aim of this paper is to investigate the factors that influence health insurance charges for XYZ Health Insurance Company, a major insurance provider in the USA. To achieve this, a dataset spanning one year was collected [1]. The data was thoroughly cleaned to identify and address any missing values. Categorical variables such as sex, smoker status, and region were converted into numerical values using dummy variables [2]. Subsequently, exploratory data analysis techniques were employed, including the creation of various graphs [3]. Finally, the dataset was split into 80% training and 20% testing data to assess the suitability of a linear regression model [4] in fitting the dataset. The key drivers of insurance charges were picked by developing a predictive model for potential customers.

**Method**

The dataset was first cleaned, ensuring that there were no missing values. Categorical variables such as sex, smoker status, and region were converted into numerical values using dummy variables. Exploratory data analysis techniques were applied, including the examination of correlation matrices and the creation of various graphs such as bar plots, scatter plots, and box plots.

**Results and Discussion**

The initial step in the analysis involved cleaning the dataset and ensuring that there were no missing values [5]. Any missing values were replaced with either the mean or median values, depending on the nature of the missingness [5]. As a result, the dataset was complete and ready for further analysis.

Next, the categorical variables of sex, smoker status, and region were converted into numerical values through the creation of dummy variables. This transformation allowed for the inclusion of these variables in the subsequent analyses.

Exploratory data analysis techniques were employed to gain insights into the relationships between various factors and insurance charges. The correlation matrix revealed that smoking status exhibited the strongest correlation with insurance charges (Fig. 1). Specifically, individuals classified as "smoker\_yes" displayed a positive correlation with insurance charges, while those classified as "smoker\_no" exhibited a negative correlation. Age, BMI, and the Southeast region also demonstrated significant effects on insurance charges, in that order.

To further explore the impact of smoking status on insurance charges, a bar plot was generated. The plot clearly illustrated that smokers had insurance charges approximately 3.8 times higher than non-smokers (Fig. 2a). This finding aligned with the observations from the correlation matrix, emphasizing the substantial influence of smoking status on insurance charges.

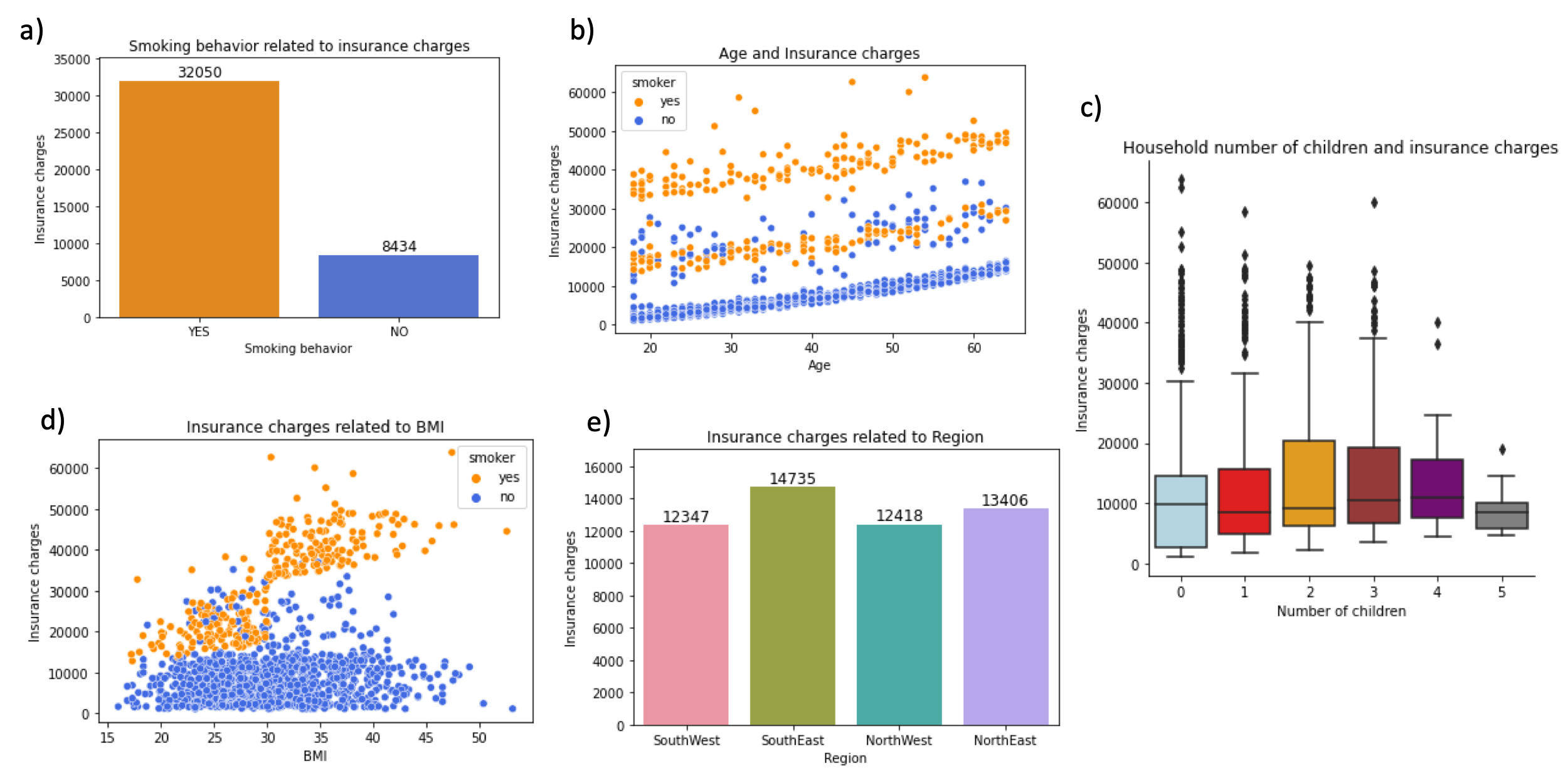
A screenshot of a game

Description automatically generated with medium confidence

**Figure 1.** Correlation matrix depicting the heat map to provide an understanding of relationship among different variables.

Additionally, scatter plots were created to examine the impact of smoking status and age on insurance charges (Fig. 2b). The plots indicated that smoking status had a more substantial effect on insurance charges compared to age. Although there was a slight increase in charges with age, the difference in charges between smokers and non-smokers remained significant across all age groups.

The analysis also investigated the relationship between the number of children and insurance charges using a box plot (Fig. 2c). The plot showed no apparent relationship between these two variables. However, the presence of numerous outliers within the dataset suggested potential anomalies or unique cases that required further examination.



**Figure 2.** Exploratory data analysis. a) Smoking status vs insurance charges, b) Relation of age and insurance cost, c) Relation between number of children and insurance charges, d) Relation between BMI and insurance charges, and e) Insurance charges with region.

Furthermore, the relationship between insurance charges and BMI was explored through scatter plots (Fig. 2d). The plots revealed that there was no significant relationship between charges and BMI when the smoking status was "no." However, when the smoking status was "yes," BMI appeared to have a linear relationship with charges, suggesting that BMI played a more important role in determining insurance charges among smokers.

Lastly, the impact of region on insurance charges was assessed using a bar plot (Fig. 2e). Although the differences were relatively minor, the Southeast region had the highest insurance charges compared to other regions.

A linear regression model was successfully constructed to predict insurance charges based on various features, including age, BMI, sex, smoking status, the number of children, and region. The categorical variables were transformed into numerical values through the creation of dummy variables, and the dataset was split into training and testing sets. The linear regression model achieved a reasonable R-squared value of 0.799 on the test set, indicating its ability to predict charges within 20% of the actual values. This model can be utilized by insurance companies to predict charges for potential customers and identify the key drivers of insurance costs (Fig. 3). Further research can be conducted to enhance the model's performance and compare it with alternative machine learning algorithms.

A screen shot of a graph

Description automatically generated with medium confidence

**Figure 3.** Comparison between real insurance price and prediction price that were calculated using the fitting of linear regression model.

**Ethical Concerns and Risks**

Some of the ethical concerns and risks associated with this project includes:

Privacy: One of the biggest ethical concerns is privacy. Personal data is collected for data analysis, and there is a risk that this data could be misused or shared without consent. It is essential to ensure that the data collected is anonymized and that the privacy of individuals is protected.

Bias: Data analysis can be biased if the data used for analysis is not representative of the population. If the data is biased, it can lead to incorrect conclusions and unfair treatment of certain groups. It is essential to ensure that the data is representative and unbiased.

Discrimination: The analysis could reveal factors that may be used to discriminate against certain groups. For example, if the analysis reveals that people with certain medical conditions have higher health insurance charges, this information could be used to discriminate against them. It is essential to ensure that the analysis does not result in discrimination against certain groups.

Misuse of results: The results of the analysis could be misused for financial gain, such as charging higher premiums to certain groups or denying coverage altogether. It is important to ensure that the results of the analysis are used ethically and not to discriminate against certain groups or individuals.

Informed consent: Individuals should be informed about the use of their data for analysis, and their consent should be obtained before their data is used. They should also be informed about how their data will be used and what risks are associated with the analysis.

**Conclusions**

A linear regression model was successfully developed to predict insurance charges based on various features, including age, BMI, sex, smoking status, number of children, and region. Initially, the categorical variables were transformed into numerical values by creating a new data frame containing dummy variables. Subsequently, the data frame was divided into training and testing sets. By using the 'charges' column as the target variable, the linear regression model was fitted, resulting in a satisfactory R-squared value of 0.799 on the test set. This indicates that the model's predictions will deviate from the actual values by no more than 20%, as depicted in the 'Real vs Prediction' charges comparison graph. Insurance companies can employ this model to forecast charges for prospective customers and identify the primary factors driving insurance costs. Future research should focus on enhancing the model's performance and exploring alternative machine learning algorithms for comparative analysis.

**Reference**

1. Kaggle US Health Insurance Dataset <https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset?resource=download>.
2. Dummy variable (statistics) <https://en.wikipedia.org/wiki/Dummy_variable_(statistics)#:~:text=In%20regression%20analysis%2C%20a%20dummy,expected%20to%20shift%20the%20outcome>.
3. Peng, R. (2016). Exploratory Data Analysis with R. Lulu.com.
4. Abbott, D. (2014). Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst (1st Edition). Wiley.
5. McGregor (2022). Practical Python data wrangling and data quality: Getting started with reading, cleaning, and analyzing data (1st Edition). O’Reilly Media.