**PREDICTING HEALTH INSURANCE PREMIUMS IN THE U.S.**

Team – 7

DATS6103: Summary Report

Prof. Dr. Ning Rui

**I. Introduction**

Insurance is a safety net for people and organizations in the event of unforeseen calamities or disasters, making it an essential component of contemporary financial planning. Diverse insurance categories address different demands, and people frequently select multiple plans to guarantee complete protection for their unique situation. Insurance is broadly classified into Personal, Asset, Specialty, Financial Protection, Liability, Business, and Cyber. Health insurance is a type of coverage that helps individuals manage and mitigate the financial risks associated with medical expenses. It functions as a legal contract between an insurance provider and a policyholder, whereby the provider agrees to pay regular premiums in return for the insurer agreeing to pay a percentage of the policyholder's medical expenses. A variety of medical services, including doctor visits, hospital stays, surgeries, prescription drugs, preventive care, and other medically required treatments, can be covered by health insurance.

Debugging some interesting facts about Health Insurance in the United States. While the majority of Americans have health insurance, a notable portion still grapples with medical debt. In 2022, the U.S. Census Bureau noted an increase in health insurance coverage, with 92.1% of the population having coverage, up from 91.7% in 2021. Despite this, approximately 8.4% or 27.6 million American adults faced periods without healthcare coverage in 2022. Racial and ethnic disparities were evident, with Hispanic and Black working-age adults less likely to have healthcare coverage than White/Non-Hispanic or Asian adults. A worrisome revelation indicates that nearly 25% of adults acknowledged skipping medication doses, cutting pills, or not filling prescriptions in the past year due to cost concerns. Moreover, about 41% of adults reported having outstanding medical or dental bill debt. Dental services ranked as the most frequently delayed healthcare type due to costs (35%), followed by vision services (25%) and doctor's visits (24%). These statistics underscore persistent challenges in achieving comprehensive and accessible healthcare for all Americans.

Evidently, the pivotal determinant shaping individuals' decisions regarding health insurance is the cost, a variable that is neither rigid nor malleable but contingent upon diverse factors including age, gender, medical history, exercise routines, and smoking habits. The insurance cost for each individual exhibits variation based on these nuanced factors. In our project, a comprehensive analysis of insurance data from a diverse range of individuals was conducted. Leveraging various sophisticated machine learning models, we endeavored to fine-tune and optimize the costs for individuals based on specific and discerning criteria.

We employed diverse visualization techniques to illustrate distinct factors and conducted statistical tests to assess the significance of variables in relation to the cost. Subsequently, we utilized machine learning models for further analysis.

**II. SMART Questions**

We formulated several SMART goals to guide our research and endeavored to address them through our comprehensive analysis. The following are the questions we established:

* To what extent is "region" a useful variable for estimating insurance costs? Does the data show any regional trends that affect premiums? Else what specific factors affect premiums?
* In comparison to non-smokers, how much does being a "smoker" add to the rise in

insurance costs?

* How much does age impact insurance premiums, and is this impact consistent across

different regions?

* How can insurance companies use the data on smoking habits and exercise frequency

to devise strategies for premium adjustments?

* How relevant is gender in determining insurance premiums, and is there a

gender-based disparity in premiums?

* How can we provide individuals with real-time estimates of their health insurance

premiums based on their unique characteristics beforehand?

**III. Literature Review**

As international students embarking on our academic journey in the United States, this project is rooted in our collective experiences upon arrival. The initial phase posed significant challenges, particularly in selecting an appropriate insurance provider. Our primary focus centered on two critical aspects: cost and coverage. As students, our financial resources were inherently constrained, magnifying the importance of securing comprehensive coverage. Given the exorbitant costs associated with emergencies in the United States, the selection of an insurance provider became a crucial and intricate decision-making process.

We have also referred:

Kaushik, K., Bhardwaj, A., Dwivedi, A. D., & Singh, R. (2022). Machine Learning-Based Regression Framework to Predict Health Insurance Premiums. *International Journal of Environmental Research and Public Health*, *19*(13), 7898. <https://doi.org/10.3390/ijerph19137898>

**IV. Description of Data**

Our data was obtained from Kaggle and is formatted as a comma-separated values (CSV) file. It comprises 1 million records distributed across 12 columns. Each row in the dataset represents a distinct record, while each column corresponds to a different variable. The variables encompass age, sex, BMI, number of children, smoker status, region, income, education, occupation, and type of insurance plan. Notably, the "charges" column denotes the actual premium amount charged.

+----+------------------------+---------+

**| # | Column | Dtype** |

+----+------------------------+---------+

| 1 | age | int64 |

| 2 | gender | object |

| 3 | bmi | float64 |

| 4 | children | int64 |

| 5 | smoker | object |

| 6 | region | object |

| 7 | medical\_history | object |

| 8 | family\_medical\_history | object |

| 9 | exercise\_frequency | object |

| 10 | occupation | object |

| 11 | coverage\_level | object |

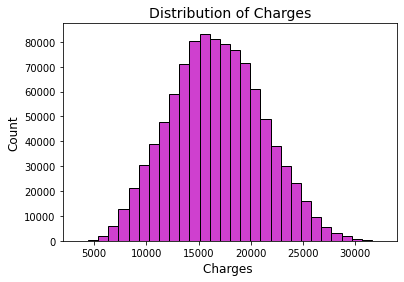
| 12 | charges | float64 |

+----+------------------------+---------+

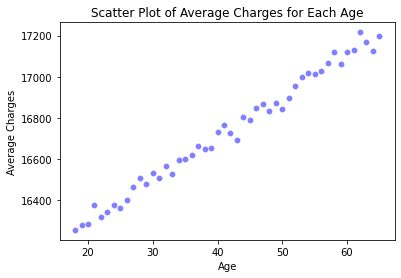
**V. Data Preprocessing**

Our data preprocessing involved systematically identifying and removing duplicates, addressing missing values with a thoughtful imputation strategy, and transforming categorical variables through efficient encoding techniques. These streamlined steps laid the groundwork for a refined dataset ready for subsequent analysis and modelling.

**VI. Exploratory Data Analysis**

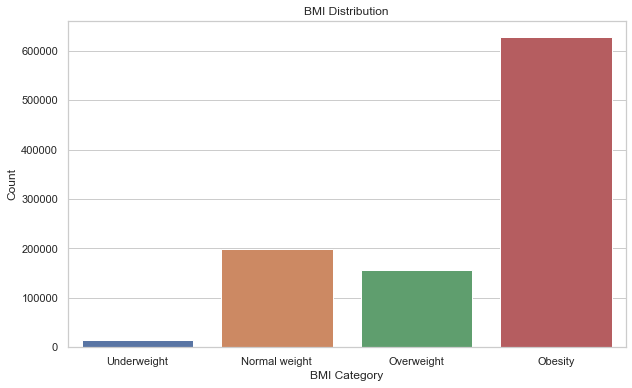
****

The histogram depicts the charges distribution, showcasing a range spanning from a minimum of 5000 to a maximum of 30000 and exhibiting characteristics reminiscent of a normal distribution. Moving forward, our analysis will focus on understanding how these charges are applied, considering a variety of influencing factors.



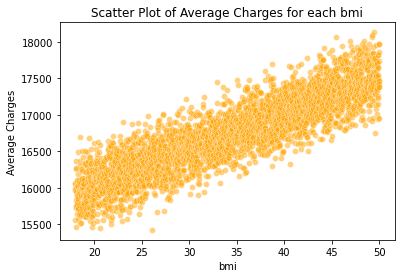
The scatter plot illustrates the correlation between age and mean charges. It indicates that as age rises, so do the charges, resembling a straightforward linear relationship between age and mean charges.

**Body Mass Index (BMI):**

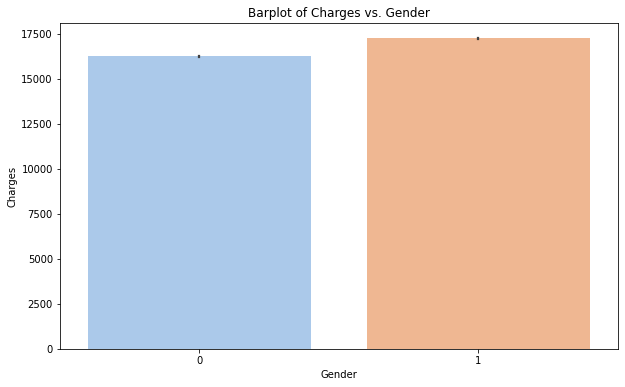
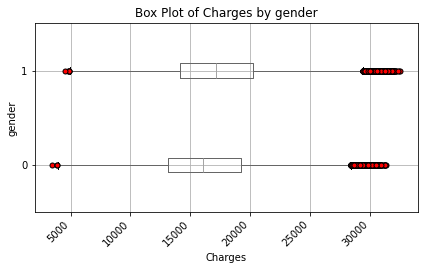
Body Mass Index (BMI) quantifies the amount of body fat relative to an individual's height and weight. BMI is classified on the following basis

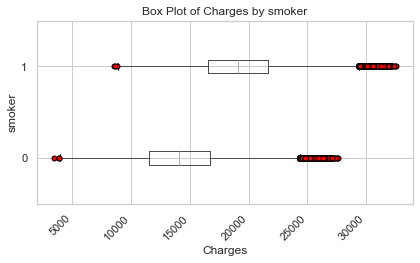
|  |  |
| --- | --- |
| **BMI** | **Category** |
| < 18.5 | Under Weight |
| 18.50 - 24.90 | Normal Weight |
| 25.00 - 29.90 | Over Weight |
| >=30 | Obesity |

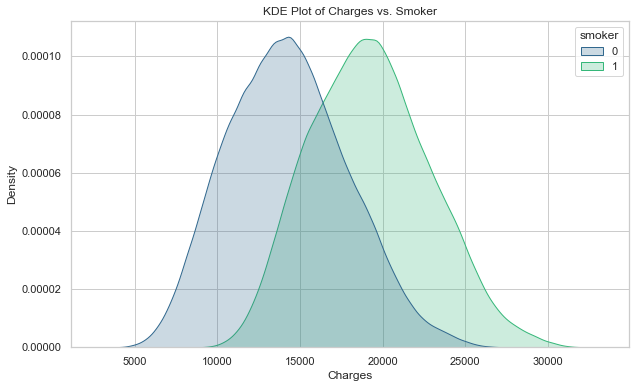
The bar plot above illustrates the distribution of BMI categories among individuals in the dataset. A significant majority is dealing with obesity. In line with 2021 statistics, a noteworthy 40% of Americans are currently grappling with obesity, and this issue is escalating at an alarming rate, posing a concerning trend for humanity.



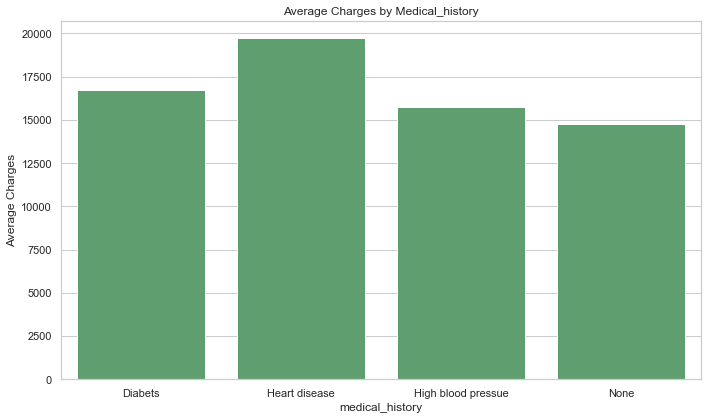
The scatter plot visually explores the relationship between BMI (Body Mass Index) and average charges. The scatter plot suggests a nearly linear correlation between BMI and charges. Specifically, as BMI increases, there is a noticeable upward trend in average charges, indicating that individuals with higher BMI tend to incur higher medical costs.



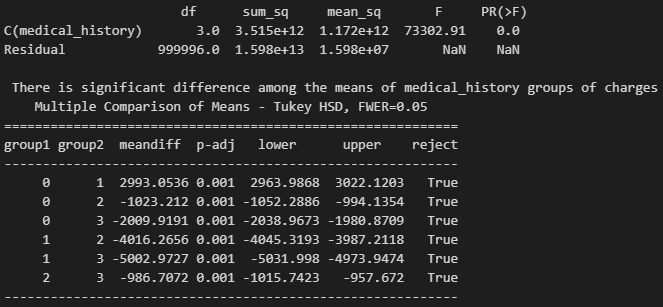
The above plots highlight the pricing contrast between male’s (1) and female’s (0). On average, males tend to incur higher charges than females. Notably, the dataset's maximum charge is attributed to a male, while females are associated with the minimum charges. In our investigation, a t-test comparing charges based on gender revealed an extremely low p-value (0.0), providing strong evidence that a significant difference exists in charges between genders. Importantly, our analysis failed to reject the null hypothesis.



The KDE and Box plots distinctly show that individuals who smoke generally face higher charges compared to non-smokers. To reinforce this observation, we conducted a T-test, resulting in an exceptionally low p-value (p < 0.05). This confirms a significant difference in charges between smokers and non-smokers, and notably, we did not reject the null hypothesis.

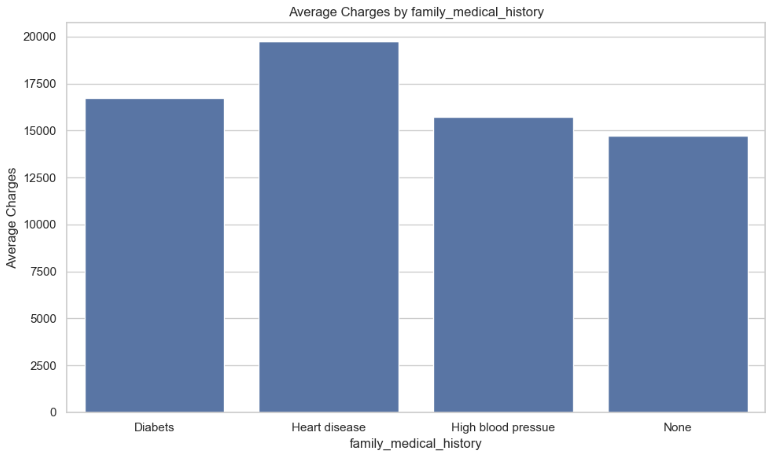


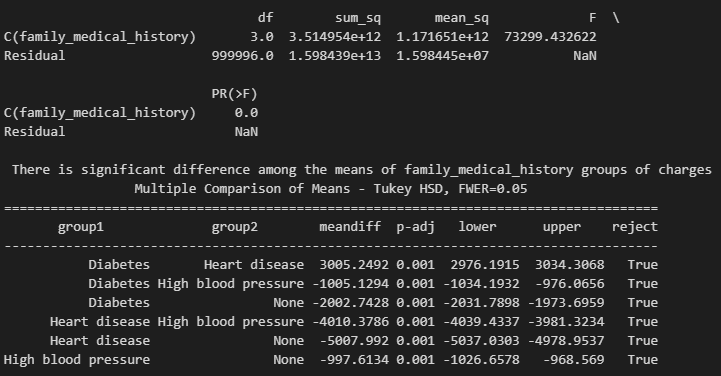
The plot above delineates the connection between charges and an individual's medical history. It distinctly shows that individuals with a history of "Heart Disease" incur notably higher charges, while the distinctions for Diabetes and High Blood Pressure are comparatively marginal.



The groups are labelled as follows: Diabetes (0), Heart disease (1), High blood pressure (2), and None (3). The presented analysis is an ANOVA table, indicating a significant difference among the means of the medical history groups concerning charges. The Tukey HSD post hoc test reveals specific pairwise mean differences and associated confidence intervals. Notably, all group comparisons demonstrate statistically significant differences (p < 0.05), suggesting that medical history has a discernible impact on charges.

The chart below depicts the relationship between charges and an individual's family medical history. Heart disease incurs the highest charges, followed by diabetes and high blood pressure. Conversely, the category "None" signifies the absence of the mentioned diseases, reflecting comparatively lower charges.





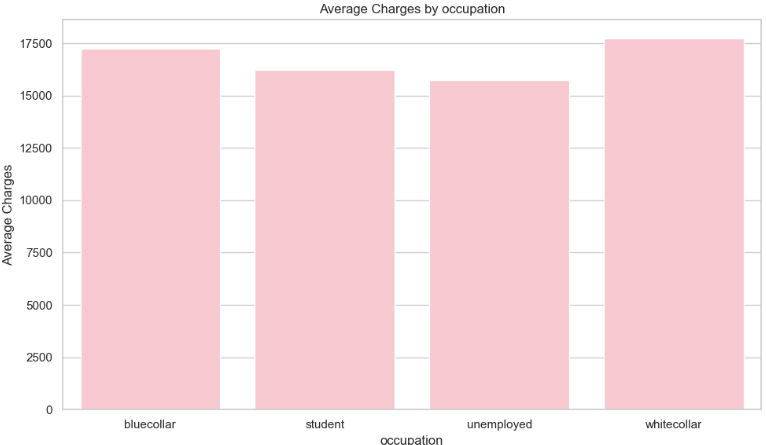
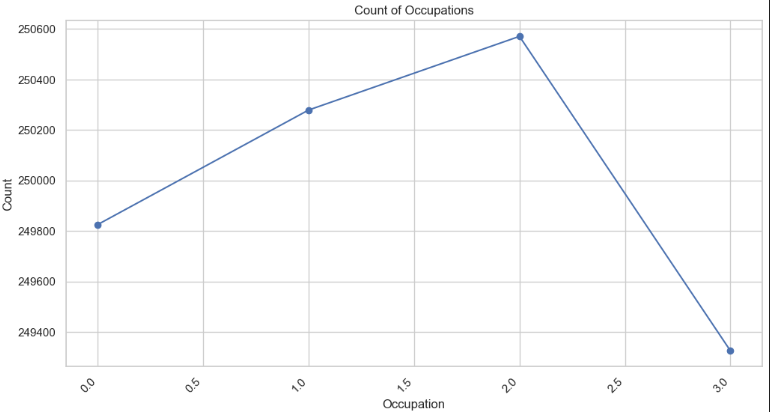
A screen shot of a computer

Description automatically generatedANOVA results highlight significant differences in mean charges among family medical history groups (Diabetes, Heart disease, High blood pressure, None). Tukey HSD post hoc tests confirm pairwise distinctions, all with p < 0.05, underscoring the significant impact of family medical history on charges.

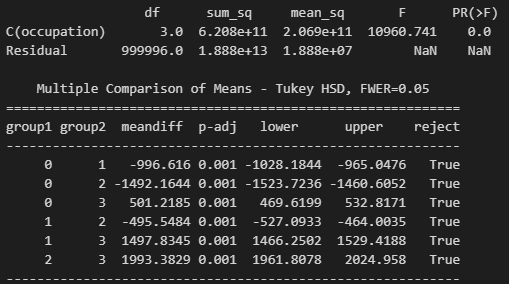
A graph of blue rectangular objects

Description automatically generated with medium confidence

The charges plot across regions reveals slight variations, with the North East exhibiting comparatively higher charges. An ANOVA test confirms significantly different mean insurance charges among the four regions, supported by an extremely low p-value. Subsequent Tukey HSD analysis identifies specific pairwise differences in mean charges.



The plots above and box plot below clearly indicate higher charges for individuals in white-collar occupations, followed by those in blue-collar and student roles. Unemployed individuals, while paying less in charges, constitute a higher count in the dataset.

A diagram of a box plot

Description automatically generated

The ANOVA test underscores a substantial difference in mean charges across diverse occupations. Additionally, the Tukey HSD test affirms statistical significance in all pairwise differences. Specifically, the smallest difference is observed between Group 1 (students) and Group 2 (unemployed), with an average charge difference of $495.5484, indicating the varying impact of occupation on charges.