**Project Report**

**US Health Insurance**

**Vishwasree Samala**

**INTRODUCTION:**

Numerous research findings indicate that smoking increases health care expenses, which could potentially represent 6–8% of the US national health care budget. (Fishman et al., 2003, p. 733-734). In line with the latest Surgeon General report, the percentage of adult smokers who smoke cigarettes is at a record-low level of 14%. Smoking-related diseases affect 16 million Americans (Kilroy, 2022). Smokers use more medical resources and spend more money on healthcare than non-smokers do because of the health risks associated with smoking (Izumi et al., 2001, p. 616). The amount that people with commercial insurance spend on medical care varies greatly depending on where they live. (Johnson & Biniek, 2021, p. 548).

**AIM:**

We aim to determine if there's a statistically significant difference in mean medical charges between smokers and non-smokers and to assess regional disparities in medical charges across the four specified US regions.

**RESEARCH QUESTIONS:**

1. Does being a smoker significantly influence medical charges compared to non-smokers?
2. Is there a significant difference in medical charges across different regions (northeast, southeast, southwest, northwest)?

**DATA:**

The "US Health Insurance Dataset" is obtained from Kaggle. With 1338 rows and 7 columns, this data offers insights into various aspects of insurance, demographics, and related factors, making it a valuable resource for research and analysis within the healthcare industry (US Health Insurance Dataset, n.d.).

Link to the dataset: <https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset/data>

The following are the variables:

**Continuous Numerical:** Age, BMI, Charges

**Discrete Numerical:** Children

**Categorical - Nominal:** Gender

**Categorical - Ordinal:** Smoker, Region

**For Research Question 1:**

Null hypothesis (H0): There is no significant difference in mean medical charges between smokers and non-smokers.

Alternate hypothesis (H1): There is a significant difference in mean medical charges between smokers and non-smokers.

**For Research Question 2:**

Null hypothesis (H0): There is no significant difference in mean medical charges across the four specified US regions.

Alternate hypothesis (H1): There is a significant difference in mean medical charges across at least one pair of regions.

**METHODS:**

**Data Transformation - Encoding Smoker Status:**

We converted the categorical distinction of 'smoker' and 'non-smoker' into a binary numeric representation. 'Smoker' entries are coded as '1', while 'non-smoker' entries are coded as '0'.

A screenshot of a computer

Description automatically generated

Fig 1: Code showing cleaned dataset with encoded smoker status.

**Outlier Detection and Data Cleansing:**

By calculating quartiles (Q1 and Q3) and the Interquartile Range (IQR), we established boundaries to identify outliers and removed them from the dataset.

A screenshot of a computer

Description automatically generated

Fig 2: Code for identifying outliers and data without outliers.

**For Research Question 1:**

**Descriptive Statistics:**

We computed summary statistics separately for smokers and non-smokers.

A close-up of a line

Description automatically generated

Fig 3: Code showing Descriptive Statistics

For non-smokers, the central tendency (mean) of medical charges is approximately $8,441. Smokers, on the other hand, exhibit a higher mean charge of around $32,050. These statistics indicate notable differences with smokers having higher median and mean charges compared to non-smokers.

**Exploratory Data Analysis (EDA):**

We produced a visual representation of medical charges categorized by smoking status using a box plot and histogram. The blue box represents non-smokers, while the orange box represents smokers.

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

Fig 4: Boxplot of charges by smoking status Fig 5: Histogram of chargers for smokers and non–smokers.

From the box plot, we can conclude that charges are high for smokers when compared to non-smokers.

**Statistical tests:**

We conducted a two-sample Welch's t-test to compare average medical charges.

**A screenshot of a computer program

Description automatically generated**

Fig 6: Code showing Welch Two Sample t-test and the result.

The analysis revealed a statistically significant difference in mean charges between the two groups (p-value < 2.2e-16). Smokers exhibited notably higher mean charges (approximately $22,014) compared to non-smokers (approximately $8,362). This suggests that there is a substantial and statistically significant difference in mean medical charges between smokers and non-smokers.

A screenshot of a computer

Description automatically generated

Fig 7: Figure showing ANOVA code and the results.

The highly significant p-value (<2e−16) indicates a strong association between smoking status and medical charges. The F-value of 680.6 further supports this, suggesting a substantial difference in charges between smokers and non-smokers. The ANOVA analysis demonstrates that smoking status significantly influences medical charges within the dataset.

**Correlation Analysis:**

**A close up of a computer code

Description automatically generated**

Fig 8: Figure showing correlation analysis code and result

The Pearson correlation coefficient of 0.602 signifies a moderate positive relationship between smoking status and medical charges suggesting higher charges are associated with being a smoker.

**Regression Analysis:**

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

Fig 9: Linear Regression Analysis Fig 10. Figure showing summary model.

The regression analysis highlights a statistically significant relationship between smoking status and medical charges (p < 0.05). Smokers tend to have higher medical charges compared to non-smokers.

The analysis revealed that smoking status significantly influences medical charges. Smokers, on average, incur approximately $13652.2 higher charges compared to non-smokers, as indicated by the model's coefficients. Both coefficients (smoker and intercept) hold high significance (p < 0.001), affirming a robust relationship between smoking status and charges. The model explains around 36.27% of the variability in charges based on smoking status.

**For Research Question 2:**

A screenshot of a computer

Description automatically generated

Fig 11: Converting categorical region column into numerical.

The code converts the categorical 'region' column in the 'insurance' dataset into numerical values (1, 2, 3, 4) to represent the different regions ('northeast', 'northwest', 'southeast', 'southwest') for analytical purposes.

**Exploratory Data Analysis:**

A box plot illustrating how medical charges ('Charges') vary across different regions, identified numerically as 1, 2, 3, and 4. The plot showcases these variations using distinct colors ('blue' and 'orange') for better comparison between regions.

**A screenshot of a computer

Description automatically generated**

Fig 12: Box plot of charges by region.

A screenshot of a computer

Description automatically generated

Fig 13: Histogram

The code generates four histograms, each displaying the distribution of medical charges for different numerical regions (1, 2, 3, 4) in the 'insurance' dataset using distinct colors (orange, blue, green, red) for easy comparison of charge distributions across regions.

**Statistical Tests:**

**A close-up of a test

Description automatically generated A screenshot of a computer code

Description automatically generated**

**A close-up of a computer code

Description automatically generated A screenshot of a computer

Description automatically generated**

Fig: 14: Figure showing Welch Two Sample t-test

No significant difference in means between northeast and northwest (p = 0.5755), northeast and southeast (p = 0.053), northeast and southwest (p = 0.002465), and northwest and southeast (p = 0.182). Significant differences between northwest and southwest (p = 0.01794), northeast and southwest (p = 0.002465), and southeast and northeast (p = 0.053). Substantial differences between southwest and northeast (p = 0.002465), southwest and northwest (p = 0.01794), and southwest and southeast (p = 0.3257), regarding mean charges.

A screenshot of a computer

Description automatically generated

Fig 14: ANOVA test

The ANOVA results indicate a significant association (p = 0.00153) between region\_numeric and charges, suggesting variations in charges across the identified regions.

**Correlation Analysis:**

A screenshot of a computer

Description automatically generated

Fig 15: Figure showing correlation analysis code and result.

The correlation coefficient of approximately -0.091 indicates a weak negative linear relationship between 'charges' and 'region\_numeric'.

**Regression Analysis:**

**A screenshot of a computer

Description automatically generated**

Fig 16: Scatterplot showing regression analysis.

A screenshot of a computer

Description automatically generated

Fig 17: Linear Regression Model

The analysis shows a significant negative association (p = 0.00153) between 'region\_numeric' and 'charges'. Each unit increase in 'region\_numeric' correlates with an estimated $596.8 decrease in charges, but 'region\_numeric' explains only a small variance in 'charges' (Adjusted R-squared: 0.0075).

**Welch's Two Sample t-test:**

**Rationale:** To compare mean charges between groups. (smokers vs. non-smokers, different regions).

**Limitation:** Assumes normality and independence; sensitive to outliers.

**ANOVA (Analysis of Variance):**

**Rationale:** To examine differences in mean charges among multiple groups (regions).

**Limitation:** Assumes equal variances and normality; results are affected by outliers.

**Linear Regression (lm):**

**Rationale:** To understand the relationship between 'region\_numeric' and 'charges'.

**Limitation:** Assumes linearity, homoscedasticity, and independence of residuals.

**CONCLUSION:**

**For Research Question 1**: The null hypothesis is rejected (p-value < 0.001). Smoking significantly influences medical charges, with smokers incurring notably higher costs compared to non-smokers.

**For Research Question 2:** The null hypothesis is not rejected for most regions (p-values > 0.05), indicating no significant differences in charges between these regions. However, comparisons involving the southwest region resulted in rejecting the null hypothesis (p-values < 0.05). This suggests there are significant differences in charges when comparing the southwest region to the

northeast, northwest, and southeast regions.

**References**

Fishman, P. A., Khan, Z. M., Thompson, E. E., & Curry, S. J. (2003). Health care costs among smokers, former smokers, and never smokers in an HMO. *Health Services Research, 38*(2), 733–749. https://doi.org/10.1111/1475-6773.00142

Izumi, Y., Tsuji, I., Ohkubo, T., Kuwahara, A., Nishino, Y., & Hisamichi, S. (2001, June 1). Impact of smoking habit on medical care use and its costs: a prospective observation of National Health Insurance beneficiaries in Japan. *International Journal Of Epidemiology*. https://doi.org/10.1093/ije/30.3.616

Johnson, W. C., & Biniek, J. F. (2021). Sources of Geographic Variation in Health Care Spending Among Individuals With Employer Sponsored Insurance. *Medical Care Research And Review : MCRR, 78*(5), 548–560. https://doi.org/10.1177/1077558720926095

Kilroy, A. (2022, August 16). Life Insurance For Smokers. *Forbes Advisor.* https://www.forbes.com/advisor/life-insurance/smokers/

US Health Insurance Dataset. (n.d.). Www.kaggle.com. https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset/data