**Introduction**

The healthcare landscape is intricate, influenced by a myriad of factors that interplay to shape medical insurance costs. In this analysis, we delve into a comprehensive dataset comprising personal attributes, geographic factors, and the consequential impact on medical insurance charges for 1338 US citizens. Understanding these intricate relationships not only contributes to the existing knowledge base but also holds practical significance in healthcare planning and policy-making.

The dataset encapsulates information on age, gender, BMI (Body Mass Index), family size, smoking habits, and the geographical region of the insured individuals. Each variable brings a unique perspective to the intricate web of factors influencing medical insurance charges. As we embark on this exploration, we aim to unravel the nuanced relationships within the dataset, seeking answers to pertinent questions regarding the influence of these variables on healthcare expenses.

**Variable Distributions and Summary Statistics**

Visualizations, including histograms and boxplots, highlighted the distribution and summary statistics of variables. Numeric and categorical variables were examined independently, revealing key insights into the dataset's characteristics.

**Summary Statistics of the Variables**

Table 1

| Age | Sex | BMI | Children | Smoker | Region | Charges |
| --- | --- | --- | --- | --- | --- | --- |
| Min. | 18.00 | 15.96 | 0.000 |  |  | 1122 |
| 1st Quartile. | 27.00 | 26.30 | 0.000 |  |  | 4740 |
| Median | 39.00 | 30.40 | 1.000 |  |  | 9382 |
| Mean | 39.21 | 30.66 | 1.095 |  |  | 13270 |
| 3rd Quartile. | 51.00 | 34.69 | 2.000 |  |  | 16640 |
| Max. | 64.00 | 53.13 | 5.000 |  |  | 63770 |
| Length |  |  |  | 1338 | 1338 |  |

The age distribution is roughly symmetric with a slight rightward skew. The mean age is approximately 39.21 years, with a median of 39 years. The standard deviation is around 14.04 years, indicating moderate variability.

BMI follows a roughly normal distribution with a slight rightward skew. The mean BMI is about 30.66, with a median of 30.40. The standard deviation is approximately 6.10, suggesting moderate variability.

The distribution of the number of children is right-skewed, indicating a higher frequency of individuals with fewer children. The mean number of children is approximately 1.09, with a median of 1. The standard deviation is around 1.21, suggesting moderate variability.

The charges variable is highly right-skewed, indicating a concentration of lower-cost cases. The mean insurance charges are approximately $13,270.42, with a median of $9,382.03. The standard deviation is around $12,110.01, indicating a wide spread of charges.

The smoking status variable is binary, with "yes" and "no" categories. The dataset has a significant imbalance, with approximately 20.48% of individuals being smokers.

The region variable is categorical, with four regions: northeast, northwest, southeast, and southwest. The dataset shows a relatively balanced representation of individuals across the four regions.

The gender variable is binary, with "male" and "female" categories. The dataset has a relatively balanced distribution between genders, with approximately 50.52% males.

|  |  |
| --- | --- |
| A graph of a number of people  Description automatically generated  *Figure 1: Age Distribution* | A graph of a number of red bars  Description automatically generated with medium confidence  *Figure 2: Charge Distribution* |
| A green and black graph  Description automatically generated  *Figure 3: BMI distribution* | A graph showing different colored rectangular shapes  Description automatically generated  *Figure 4: Region Distribution* |
| A graph with a bar graph  Description automatically generated  *Figure 5: Children Distribution* | **A graph showing a number of smokers  Description automatically generated**  *Figure 6: Smoker Distribution* |
| **A blue and orange rectangular shapes  Description automatically generated**  *Figure 7: Sex Distribution* |  |

Table 2: **Correlation test results for all variables including charge variable**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Test Statistic (t) | Degrees of Freedom (df) | Correlation Coefficient (r) | 95% Confidence Interval | p-value | Conclusion |
| Age and Age | ∞ | 1336 | 0 |  | 1 | There is a perfect positive correlation between age and itself, which is expected |
| Age and BMI | 4.0181 | 1336 | 0.1093 | [0.0560, 0.1619] | 6.19e-05 | There is a statistically significant positive correlation (0.1093) between age and BMI. |
| Age and Children | 1.5537 | 1336 | 0.0425 | [-0.0111, 0.0958] | 0.1205 | There is no significant correlation between age and the number of children. |
| Age and Charges | 11.4531 | 1336 | 0.2990 | [0.2494, 0.3470] | 4.89e-29 | There is a strong positive correlation (0.2990) between age and insurance charges. |
| BMI and Age | 4.0181 | 1336 | 0.1093 | [0.0560, 0.1619] | 6.19e-05 | There is a statistically significant positive correlation (0.1093) between BMI and age. |
| BMI and BMI | ∞ | 1336 | 1 |  | 0 | There is a perfect positive correlation between BMI and itself, which is expected |
| BMI and Children | 0.4664 | 1336 | 0.0128 | [-0.0409, 0.0663] | 0.6410 | There is no significant correlation between BMI and the number of children. |
| BMI and Charges | 7.3966 | 1336 | 0.1983 | [0.1463, 0.2493] | 2.46e-13 | There is a moderate positive correlation (0.1983) between BMI and insurance charges |
| Children and Age | 1.5537 | 1336 | 0.0425 | [-0.0111, 0.0958] | 0.1205 | There is no significant correlation between the number of children and age |
| Children and BMI | 0.4664 | 1336 | 0.0128 | [-0.0409, 0.0663] | 0.6410 | There is no significant correlation between the number of children and BMI |
| Children and Children | ∞ | 1336 | 1 |  | 0 | There is a perfect positive correlation between the number of children and itself, which is expected |
| Children and Charges | 2.4912 | 1336 | 0.0680 | [0.0145, 0.1211] | 0.0129 | There is a statistically significant positive correlation (0.0680) between the number of children and insurance charges |
| Charges and Age | 11.4531 | 1336 | 0.2990 | [0.2494, 0.3470] | 4.89e-29 | There is a strong positive correlation (0.2990) between insurance charges and age |
| Charges and BMI | 7.3966 | 1336 | 0.1983 | [0.1463, 0.2493] | 2.46e-13 | There is a moderate positive correlation (0.1983) between insurance charges and BMI |
| Charges and Children | 2.4912 | 1336 | 0.0680 | [0.0145, 0.1211] | 0.0129 | There is a statistically significant positive correlation (0.0680) between insurance charges and the number of children |
| Charges and Charges | ∞ | 1336 | 1 |  | 0 | There is a perfect positive correlation between insurance charges and itself, which is expected |

Table 3: **Chi-Square Test Results**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | X-squared | Test | df | p-value | Conclusion |
| **Smoker and Region** | 7.3435 | Pearson's Chi-squared test | 3 | 0.06172 | There is no significant association between smoking status and region |
| **Smoker and Sex** | 7.3929 | Pearson's Chi-squared test with Yates' continuity correction | 1 | 0.006548 | There is a significant association between smoking status and gender. |
| **Region and Smoker** | 7.3435 | Pearson's Chi-squared test | 3 | 0.06172 | There is no significant association between region and smoking status. |
| **Region and Sex** | 0.43514 | Pearson's Chi-squared test | 3 | 0.9329 | There is no significant association between region and gender. |
| **Sex and Smoker** | 7.3929 | Pearson's Chi-squared test with Yates' continuity correction | 3 | 0.006548 | There is a significant association between gender and smoking status. |
| **Sex and Region** | 0.43514 | Pearson's Chi-squared test | 3 | 0.9329 | There is no significant association between gender and region. |

|  |  |
| --- | --- |
| A collage of black dots  Description automatically generated  *Figure 8: Scatter plot visualization of correlation* | A screen shot of a graph  Description automatically generated  *Figure 9: Visualization of the correlation matrix* |

**The result from the linear regression model**

**A screenshot of a computer

Description automatically generated**

*Figure 8: Linear Regression result*

The linear regression model is represented as follows:

charges=−11927.17+257.19×age−128.16×sexmale+336.91×bmi+390.98×children1+1635.78×children2+964.34×children3+2947.37×children4+1116.04×children5+23836.41×smokeryes−380.04×regionnorthwest−1033.14×regionsoutheast−952.89×regionsouthwest

**Coefficients:**

The intercept is -11927.17, indicating the estimated charges when all other predictors are zero.

Each additional year of age is associated with an increase of $257.19 in charges.

The coefficient for males is -128.16, suggesting a slightly lower charge for males, but it is not statistically significant.

An increase of one unit in BMI is associated with an increase of $336.91 in charges.

The presence of children has varying effects on charges, with statistically significant increases for children2, children3, and children4.

Smoking is a substantial predictor, with smokers incurring significantly higher charges ( $23836.41 on average).

Regions southeast and southwest have negative coefficients, indicating lower charges compared to the baseline region (northeast). Only the effect of the southeast region is statistically significant.

**Model Performance**:

The model explains approximately 75.19% of the variance in insurance charges. The adjusted R-squared, accounting for the number of predictors, is 74.97%. The F-statistic of 334.7 with a very low p-value (< 2.2e-16) indicates that the overall model is statistically significant.

The residuals exhibit a range from -11689.4 to 30042.7, indicating some variability not explained by the model.

**Generalized Additive Model (GAM)**

We also used a Generalized Additive Model (GAM) to predict health insurance charges. Our model includes various factors like age, gender, BMI (Body Mass Index), number of children, smoking status, and region.

**Impact of Different Factors**:

* **Age**: The relationship between age and charges is not linear. As people age, their insurance charges tend to increase, but not at a constant rate.
* **BMI**: BMI also has a non-linear relationship with charges. Generally, higher BMI is associated with higher insurance charges.
* **Smoking**: Smokers tend to have significantly higher insurance charges compared to non-smokers.
* **Number of Children**: Having more children is associated with higher insurance charges.
* **Gender and Region**: Gender and certain regions don't seem to have a significant impact on insurance charges.

**Model Performance:**

Our model can explain around 75.8% of the variability in insurance charges, which suggests it's quite effective in predicting charges. The model's predictive performance, measured by Generalized Cross Validation (GCV), is reasonably good.

Our GAM model provides valuable insights into how different factors influence health insurance charges. Smoking, age, BMI, and the number of children all play significant roles. However, gender and certain regional factors may not affect charges as much.

**Analysis of Differences in Predictor Variables Based on CHARGE-split**

We looked into our data and split it into two groups: one with "High" insurance charges and the other with "Low" charges as CHARGE-split. Our goal was to check if certain factors like age, BMI, children, and charges are different between these groups.

**Hypotheses**:

* **Null Hypothesis (H0)**: There is no significant difference in the means of predictor variables between the "High" and "Low" charge groups.
* **Alternative Hypothesis (Ha)**: There is a significant difference in the means of predictor variables between the "High" and "Low" charge groups.

**Age**: People in the "High" charge group were, on average, older than those in the "Low" charge group. Age seems to play a role in insurance charges. Older individuals might have higher costs. (See *Figure 5 )*

**BMI (Body Mass Index)**: People in the "High" charge group had slightly higher average BMI than those in the "Low" charge group. BMI might influence insurance charges, with higher values linked to higher costs. ( *see Figure 6* )

**Children**: The number of children didn't show a significant difference between the "High" and "Low" charge groups. Having children might not impact insurance charges much in our dataset.

**Charges:** People in the "High" charge group had significantly higher average insurance charges than those in the "Low" charge group. Insurance charges can vary a lot, and certain factors might lead to significantly higher costs.

|  |  |
| --- | --- |
| A screenshot of a graph  Description automatically generated*Figure 10: Boxplot of age by CHARGE-split* | A screenshot of a graph  Description automatically generated *Figure 11: Boxplot of BMI by CHARGE-split* |

**Assumptions:**

* **Independence of Observations:** Each observation within a geographic region for the interval predictor variables is assumed to be independent.
* **Normal Distribution (for ANOVA)**: ANOVA assumes that the residuals (deviations from the group means) are normally distributed. We also consider using the Kruskal-Wallis test.
* **Homogeneity of Variances (for ANOVA)**: ANOVA assumes that the variances of the interval predictor variables are equal across all geographic regions.

**Hypotheses:**

1. **Null Hypothesis (H0):** There is no significant difference in the central tendencies (means) of BMI and Age across different geographic regions.

* H0 for ANOVA: μ\_BMI\_southwest = μ\_BMI\_southeast = μ\_BMI\_northwest = μ\_BMI\_northeast (No significant difference in BMI means across regions)
* H0 for Kruskal-Wallis: The distributions of BMI are the same across regions.
* H0 for ANOVA: μ\_Age\_southwest = μ\_Age\_southeast = μ\_Age\_northwest = μ\_Age\_northeast (No significant difference in Age means across regions)
* H0 for Kruskal-Wallis: The distributions of Age are the same across regions.

1. **Alternative Hypothesis (Ha):** There is a significant difference in the central tendencies (means) of BMI and Age across different geographic regions.

* Ha for ANOVA: At least one region has a different mean BMI or Age compared to the others.
* Ha for Kruskal-Wallis: The distributions of BMI or Age are different across regions.

**ANOVA Results for BMI with Respect to Region:**

A close-up of numbers

Description automatically generated

*Figure 12: Summary result from ANOVA test for BMI against Region*

The main effect of 'Region' is statistically significant (F(3, 1334) = 39.49, p < 0.001), indicating that there are significant differences in mean BMI across at least two geographic regions.

**ANOVA Results for Age with Respect to Region:**

A black numbers on a white background

Description automatically generated

*Figure 13: Summary result from ANOVA test for Age against Region*

The results indicate that there is no statistically significant difference in the mean age across different geographic regions.

**Post-hoc Analysis Results for BMI with Respect to Region**:

A table with numbers and symbols

Description automatically generated

*Figure 14: Summary result from ANOVA test for BMI against Region*

The comparison between the 'northeast' and 'southeast' regions shows a significant difference in mean BMI (p < 0.001).

Similar significant differences are observed between 'northeast' and 'southwest' regions, as well as between 'northwest' and 'southeast' regions.

The 'southeast' and 'southwest' regions also significantly differ in mean BMI.

**Post-hoc Analysis Results for Age with Respect to Region:**

A table with numbers and symbols

Description automatically generated

*Figure 14: Summary result from post-hoc  test for Age against Region*

The post-hoc analysis does not reveal any significant differences in mean age between the specified regions. All adjusted p-values are above the commonly used significance threshold of 0.05, indicating no evidence to reject the null hypothesis of equal mean ages across regions.

These results suggest that, based on the available data, the mean age does not differ significantly among the specified geographic regions.

**Chi-square Test Results for Sex with Respect to Region**:

A black text on a white background

Description automatically generated

*Figure 14: Summary result from Chi-square test for sex against Region*

The chi-squared (X-squared) statistic is 0.43514, the degrees of freedom (df)  is 3 and the p-value is 0.9329. Indicating that there is no significant association between the variables 'sex' and 'region' (X-squared = 0.43514, df = 3, p = 0.9329). This implies that, based on the available data, the distribution of sexes does not differ significantly across the specified geographic regions. The non-significant p-value suggests that any observed differences in the distribution of sexes across regions are likely due to random variation, and there is no strong evidence to reject the null hypothesis of independence between 'sex' and 'region.'

**Chi-square Test Results for Smoker with Respect to Region:**

A black text on a white background

Description automatically generated

*Figure 15: Summary result from Chi-square test for Smoker against Region*

The chi-squared (X-squared) statistic is 7.3435 with degrees of freedom (df) = 3 and p-value of 0.06172. This indicates a marginally significant association between the variables 'smoker' and 'region' (X-squared = 7.3435, df = 3, p = 0.06172). While the p-value is slightly above the conventional significance threshold of 0.05, there may be some evidence to suggest a non-random association between smoking status and geographic region.

The Kruskal-Wallis rank sum test was conducted to assess the differences in BMI and age across different geographic regions.

* For BMI, the test yielded a highly significant result (Kruskal-Wallis chi-squared = 94.689, df = 3, p-value < 2.2e-16), indicating that there are significant differences in BMI across regions.
* However, for age, the test result was not significant (Kruskal-Wallis chi-squared = 0.41382, df = 3, p-value = 0.9374), suggesting that there are no significant differences in age across regions.
* These findings suggest that geographic region may play a role in influencing BMI but not age among the individuals in the dataset.

**Conclusion**

In conclusion, our analysis revealed several important insights regarding the relationships between predictor variables and health insurance charges. We found strong positive correlations between age, BMI, and the number of children with insurance charges. Additionally, smoking status emerged as a significant predictor, with smokers having substantially higher charges. Region and gender showed less pronounced associations with charges. Furthermore, our chi-square tests indicated significant associations between smoking status and gender, while other demographic factors showed no significant relationships. Overall, these findings contribute to a better understanding of the factors influencing health insurance charges, which is crucial for informed decision-making in healthcare policy and practice.

**References**

1. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

This reference provides information about the R programming language, which was used for data analysis and visualization.

1. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer.

This book by Hadley Wickham serves as a comprehensive guide to creating visualizations using the ggplot2 package in R.

1. Zeileis, A., Hothorn, T. (2002). Diagnostic checking in regression relationships. R News, 2(3), 7-10.

This source discusses diagnostic checks and tests for regression models, which may be relevant for assessing assumptions in regression analyses.

1. Crawley, M. J. (2012). The R Book. John Wiley & Sons.

The R Book by Michael J. Crawley is a valuable resource for various statistical analyses using R, including regression modeling and hypothesis testing.

1. Fox, J., Weisberg, S. (2018). An R Companion to Applied Regression. Sage Publications.

This book provides additional insights and guidance on regression analysis using R.