Medical Costs and Demographics

By: Philip Kwan

30th July 2024

Objective: To explore the relationship between Medical Costs and Demographic Factors, which include Age, Number of Children, Smoking, Gender, Region, BMI, using various statistical methods.

Methods: I ran analyses on an open-source dataset that contained 10,000 data points across six demographic factors (Age, Number of Children, Smoking, Gender, Region, BMI) and their associated medical costs. This paper focuses on Smoker Status, Gender, and Region and their roles in influencing healthcare costs. My analysis entailed data cleaning, exploratory data analysis, and hypothesis testing to identify significant relationships. Additionally, multiple regression analysis was used to identify confounding relationships and verify robust findings.

Key Findings: Upon initial analyses, I found no outliers in the dataset and the variables presented as independent. After creating visualizations and running hypothesis tests, I found that smoking was the only statistically significant factor influencing healthcare costs. At the same time, gender and region were less influential on healthcare costs. Through further analysis, I discovered that the cause of differing results in visualizations compared to hypothesis tests was random variation and subgroups within the data rather than confounding variables.

Detailed Analysis:

Detailed Analysis Part 1 Initial Screenings:

- Outlier and Normality:
- The first step of my statistical procedure was to scan the data for **outliers** before proceeding with analysis to ensure the robustness of future findings. I first created a series of functions to find outlier values within each independent variable's dataset values. The function I developed "outliers_list" took the data contained in each variable and applied another function "findoutliers" which tested each data point in that specific variable to see whether it fell outside the range of our upper bound and lower bound. Data points that fell outside of our upper bound or lower bound were added to our outliers list.
- Upon running this outlier detection function and printing its results, I found no outliers within the dataset's independent variables. This initial outlier scan yielded promising results regarding the robustness of future findings. To expand my analysis, I tested for outliers of each independent variable concerning medical cost before analyzing these relationships. I completed this procedure by developing and analyzing QQ plots of our relations.
- These quantile-quantile plots (QQ-plots) showed that the data distributions for the relationships between demographic variables of Gender, Smoking Status, and Region and medical costs contained no outliers.
- Also, the plots presented the relation between Smoker Status and medical cost as the only normal data distribution. The QQ plots corresponding to the variables Gender and Region, outlined in red and orange, followed skewed distributions.

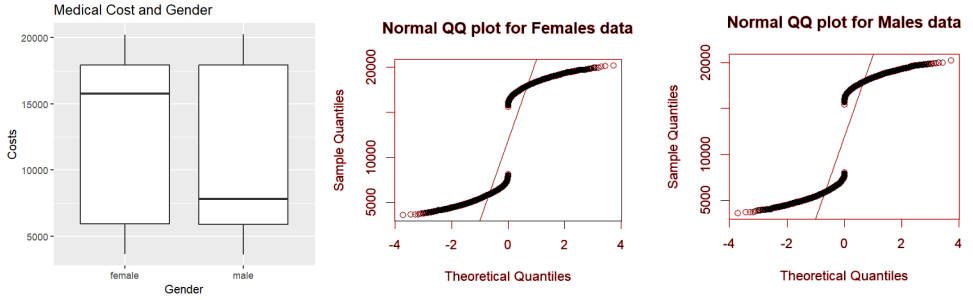
- Negligent Multicollinearity (VIF):
- Before running multiple hypothesis tests to analyze relationships between individual independent variables (Gender, Smoking Status, Region) and the dependent variable (Medical Cost), it is important to determine whether the independent variables are correlated with each other to ensure robustness of findings. To investigate this, I measured the Variance Inflation Factor (VIF) for a multiple linear regression containing all independent variables. A VIF measures inflation in the variance of the regression coefficient (effect on the independent variable) as a result of correlations among independent variables (multicollinearity). A VIF value greater than 10 indicates high multicollinearity and instability in estimating an independent variable's impact. In my analysis of the VIF, I found that each independent variable had a VIF value of less than 1.00044. These VIF values indicate negligible multicollinearity between the independent variables, meaning we could proceed with testing relationships between our independent variables and medical cost.

Variable	Region	Smoker	Children	Gender	Age	BMI
Adjusted General VIF	1.000198	1.000235	1.000322	1.000100	1.000240	1.000441

- **Summary**: I therefore concluded that the data appears to contain no outliers, the independent variables are not correlated, and two variables contain skewed data. Given this information, I was then enabled to proceed with creating statistical graphs and conducting hypothesis tests accordingly.

Detailed Analysis Part 2: Plots and Hypothesis tests

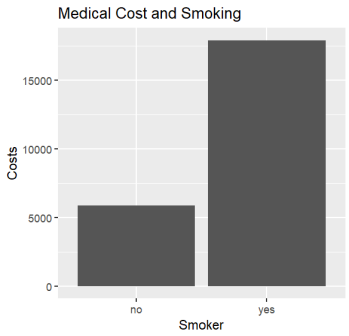
- Plot 1: Medical Cost and Gender:

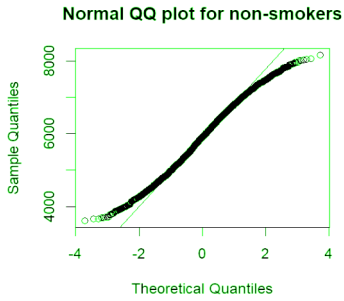


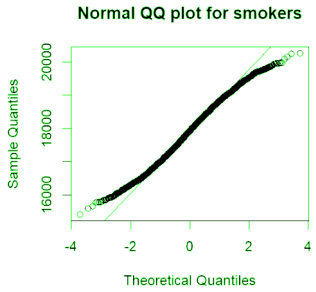
Measure	Min	1st Q.	Median	3rd Q.	Max	Std. Dev.	Mean
Males	3617.090	5895.767	7793.610	17915.062	20268.210	6070.258	11879.896
Females	3653.500	5917.767	15781.390	17953.912	20234.320	6077.988	11917.711

- The boxplot and the table above notably show the median medical cost is significantly higher for females than males. Also, the QQ plots depict the relations between gender groups (males and females) and medical cost as skewed right.
- To investigate this difference in means, I ran a Mann-Whitney U-test which compared median medical costs across the skewed distributions for males and females. Here are the **assumptions for the Mann-Whitney U-test**:
    1) Random samples: The Samples are randomly selected, assuming there was no bias in data acquisition.
    2) Homogeneous Distribution Shape: The QQ-plots indicate that both groups are skewed right and appear similar in distribution meaning this assumption is fulfilled.
    3) Independence: The male and female groups are separate with no overlaps.
    4) Ordinal dependent variable: The dependent variable medical cost is measured on a ratio scale and thus has the properties of an ordinal variable.
- **Hypothesis Testing**:
    - The null hypothesis of our Mann-Whitney U-test is that median medical costs do not differ across gender.
- In conducting our test, we find that the probability value (p-value) calculated was 0.6634 in our test meaning there is a 66.34% chance of observing our data given that the medians do not differ. Since this p-value is higher than the common p-level of 0.05, we fail to reject the null hypothesis and conclude that median healthcare costs do not significantly differ across genders.
- **Anomaly and further analysis:** Despite the conclusion of our Mann-Whitney U-test, the box plot presents the median healthcare cost for females as significantly higher than that of males. This discrepancy presents as an anomaly that requires greater analysis.
- To further explore this anomaly, I analyzed how confounding variables could skew the medians across genders. This analysis involved comparing the outputs of a multiple linear regression test, encompassing all independent variables including potential confounding variables, to a simple linear regression test on solely the variable gender.
- I printed a summary of these two linear regressions, and found that both tests output p-values greater than 0.05 for gender and were not statistically significant. For instance, the p-value for the subgroup “Male” is 0.403 in the multiple linear regressions test containing all exogenous variables and 0.756 in the simple linear regression of gender.
- We observe that confounding variables had little influence on skewing the median medical costs across genders. Therefore, the anomaly that the median healthcare price differs across genders is likely caused by random variation or subgroups that skew the medians rather than Confounding variables or systematic differences between regions.
- **Conclusion**: Overall, our findings prove that medical costs are more likely to be influenced by independent variables other than Gender. The discrepancy of medians between our hypothesis test and our boxplot is likely the result of random variation rather than systemic bias.

- Plot 5: Medical Cost and Smoking:





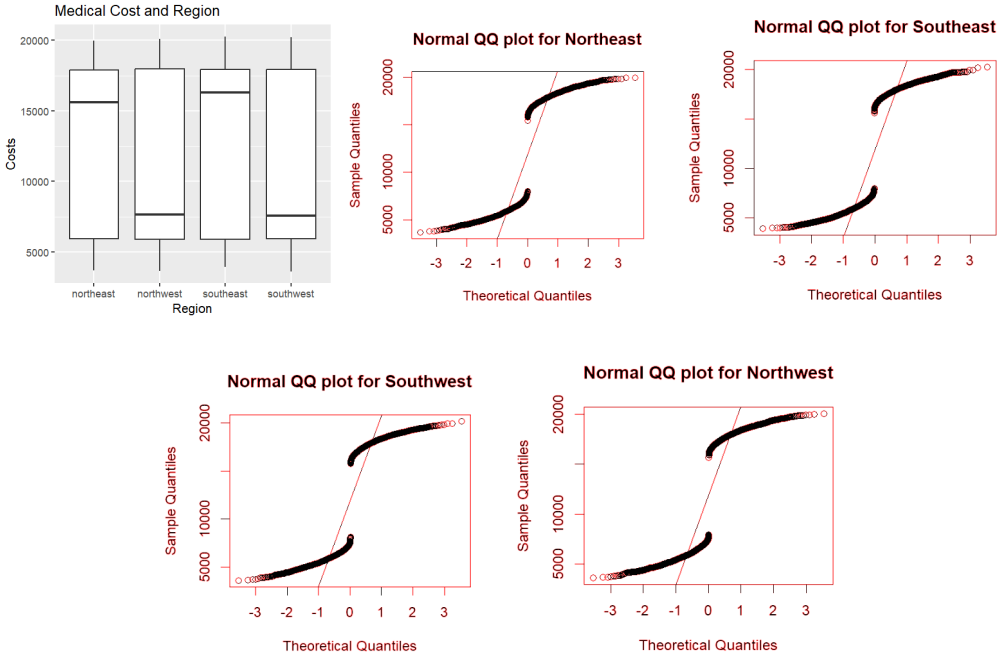


Measure	Min	1st Q.	Median	3rd Q.	Max	Std. Dev.	Mean
Non-smoker	3617.090	5258.260	5910.735	6531.472	8166.970	849.6202	5894.855
Smoker	15421.730	17292.030	17934.305	18548.783	20268.210	852.0749	17922.253

- This bar plot notably presents the mean healthcare costs for smokers (17922.253) as significantly higher than those for non-smokers (5894.855). Also, these QQ plots for the relations between smoking status and medical costs present normal distributions.
- To further analyze whether the difference in mean healthcare costs is statistically significant, I ran a 2-sample t-test which compared mean medical costs across smokers and non-smokers. Listed are the **assumptions of the 2-sample t-test:**
    1) Normality: Each group (smokers and nonsmokers) and their associated medical costs appear normally distributed in the QQ plots.
    2) Independence: The data of smokers and non-smokers are independent and individuals cannot belong to both groups.
    3) Random Samples: The samples are randomly selected, assuming no bias in data acquisition.
    4) Homogeneous Variance: The data has a similar variance for smokers at 726031.6 and non-smokers at 721854.5 and the ratio of these variances is valued at less than four.
    5) Ratio Dependent variable: The dependent variable medical cost can be measured on a ratio scale.
- **Hypothesis testing:**
    - The null hypothesis of the 2-sample t-test was that there is no difference in means of smokers vs non-smokers.
- P-value: The first finding of the t-test was the p-value being less than 2.2e^-16 meaning there is a minuscule probability of observing our data given a true null hypothesis. This miniscule p-value means we reject the null hypothesis, indicating there is a significant difference between means for medical costs of smokers and non-smokers.
- Confidence interval: Another finding from the t-test, was the confidence interval, indicating we are 95% confident that the difference between means is between -12060.75 and -11994.04. This interval does not contain zero, meaning we can be confident there is a difference of means for both groups.
- Mean Comparison: Finally, the 2-pair t-test displayed higher mean medical costs for smokers at $17922.253 compared to non-smokers at $5894.855.

- The 2-sample t-test confirms our finding in the bar plot that the mean healthcare cost for smokers is significantly higher compared to non-smokers. Our analysis proves that smoking is associated with higher medical costs

- Plot 6: Medical Cost and Region:



Measure	Southwest	Northwest	Northeast	Southeast
Min	3617.090	3653.500	3690.860	3920.110
First Quartile	5917.340	5888.240	5918.140	5912.640
Median	7554.800	7654.895	15601.560	16282.070
Third Quartile	17922.340	17957.822	17905.225	17946.155
Max	20234.320	20077.310	19979.130	20268.210
Std. Dev.	6071.614	6086.49	6062.989	6076.742
Mean	11823.973	11858.047	11914.446	12000.430

- The boxplot and table above notably show the median healthcare costs in the United States as highest in the southeast (16282.070) followed by the northeast (15601.560), northwest (7654.895), then southwest (7554.800) which has the lowest median healthcare costs. Also, QQ plots display the relations between all regions and medical costs as skewed right.

- To further investigate these differences in medians for statistical significance, I ran a Kruskal-Wallis test to compare median medical costs for the skewed distributions representing regions. These are the following **assumptions of the Kruskal-Wallis test**:
    - Ratio Dependent Variable: The dependent variable medical cost is measured on a ratio scale, allowing for accurate comparisons with multiplication and division.
    - Independence and Comparison: the independent variable region contains more than two comparable, independent groups.
    - Homogeneous Distributions: All groups (northwest, northeast, southwest, southeast) are similar in skewness and shape as shown in the QQ plots above.
    - Ordinal Data: The dependent variable medical cost is measured on a ratio scale and is thus at least ordinal.
- **Hypothesis Testing:**
    - The null hypothesis of this test is that median medical costs do not differ by region.
- The test found the probability value (p-value) that our data exists given a true null hypothesis is 0.92. This p-value is significantly greater than the common p-value of 0.05, meaning we fail to reject the null hypothesis and median medical costs do not significantly differ by region.
- **Anomaly and Further Analysis:**
- Contrary to this conclusion, the boxplot depicts median medical costs that vary across regions. Investigating further, I tested for potential confounding variables that may skew the medians and cause this anomaly.
- To test for confounding variables, I again compared the results of a multiple linear regression containing all independent variables to a simple linear regression on the variable region.
- Both linear regressions yielded p-values significantly greater than 0.05, indicating no statistical significance of confounders. This finding suggests that confounding variables had little influence in skewing the medians.
- Therefore, the anomaly that median healthcare costs vary across regions in our dataset is likely the result of random variation or the presence of subgroups within regions. Confounding variables or systematic differences between regions are unlikely to be the cause for the difference in medians across regions displayed in our boxplot. Our findings prove that healthcare cost is more likely to be influenced by other independent variables rather than region.

Conclusion:

- Summary: This study emphasizes smoking as an influential factor in determining medical costs. The statistical methods used in my analyses provided greater identification of factors that influence medical costs.
    - Through my procedures, I concluded the following;
        1) **Gender**: Median healthcare costs do not significantly differ across genders. A person's gender is a demographic that has an insignificant impact on medical cost in relation to the other variables analyzed.

2) **Smoking**: The median healthcare costs for smokers is greater than the median healthcare costs for non-smokers. Smoking is associated with higher healthcare costs than non-smoking.
3) **Region**: Finally, median healthcare costs do not significantly vary by region in the US.

- Therefore, smoking status was the most influential demographic on medical costs while gender and region were less influential on medical costs among the variables measured.

Recommendations:

- **Anomaly analysis:** I recommend further analysis of the anomalies presented in the relationships between medical cost and both region and gender. Further investigation into the differing medians in these groups would allow us to create more accurate conclusions about these relationships. Running additional tests to assess the impact of random variation and potential external variables on skewing the medians for region and gender would allow us to better understand these anomalies.

- **Expanding the dataset:** The data set was limited by its small sample size compared to the population of the US and the small number of demographic variables. I recommend that we perform further research on larger datasets that more accurately represent the population of the United States.

- **Recommendations for further analysis**: I advise that the relationships below would be analyzed with the following hypothesis tests:
    - Age and BMI: Both variables should be tested with the Spearman's Rank Order Correlation Test which is a non-parametric test used for finding correlation in scatter graphs with skewed data distributions.
    - Number of Children: A Kruskal Wallis Test should be used to test for differences of means across multiple independent groups (# of children) for skewed distributions.
    - Additional tests: If the data presents anomalies, it is important that we run additional tests to understand their causes and implications.

Acknowledgements:

- I acknowledge the use of Chat GPT for assistance in this project. I used Chat GPT for help understanding R code and statistical tests. The analysis and conclusions are my own.

Bibliography:

- Ali, Waqar. (2024). Medical Costs [medical_costs.csv]. Kaggle. https://www.kaggle.com/datasets/waqi786/medical-costs/data