Medical Costs and Demographics

By: Philip Kwan

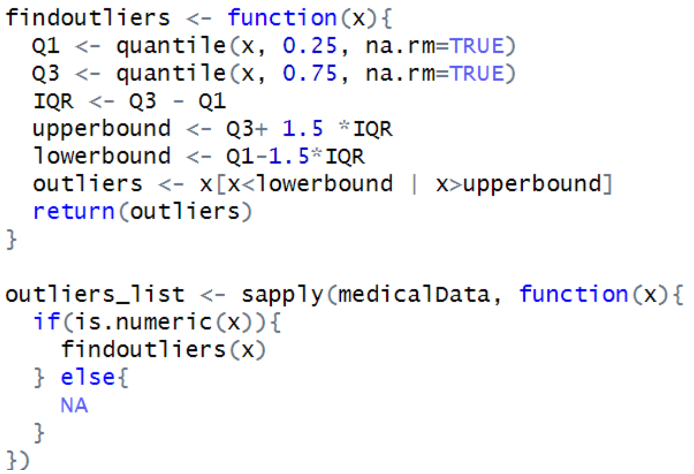
30th July 2024

Objective: To explore the relationship between Medical Costs and Demographic Factors. (Age, Number of Children, Smoking, Gender, Region, BMI)

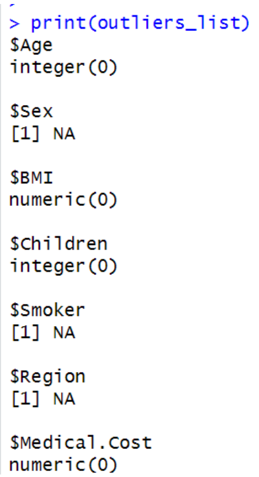
Methods: I ran analyses on an open source data file through the use of R Studio Software. This data file contained 10,000 data points which contained information on six demographic factors (Age, Number of Children, Smoking, Gender, Region, BMI) and their associated medical costs. In my analyses, I checked for clean data, constructed graphs (bar plots, scatter plots and boxplots), and ran hypothesis tests to determine relationships and identify cost distributions across these varying demographics. The demographics I focused on were smoker status, gender, and region and their roles in influencing healthcare costs.

Key Findings:

- Key Findings Part 1 Clean Data:
- Outlier and normality: The first step of my statistical procedure was to scan the data for **outliers** and determine normalcy for all variables before proceeding with analysis. I first created a function to find outliers across all the variables which upon running informed me there were no outliers in the data.



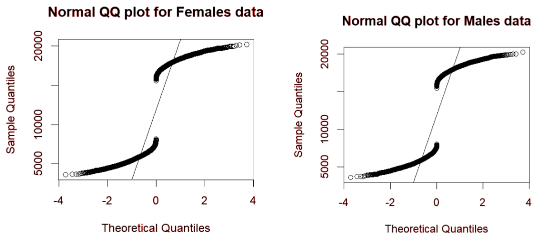
```
findoutliers <- function(x){
  Q1 <- quantile(x, 0.25, na.rm=TRUE)
  Q3 <- quantile(x, 0.75, na.rm=TRUE)
  IQR <- Q3 - Q1
  upperbound <- Q3+ 1.5 *IQR
  lowerbound <- Q1-1.5*IQR
  outliers <- x[x<lowerbound | x>upperbound]
  return(outliers)
}

outliers_list <- sapply(medicalData, function(x){
  if(is.numeric(x)){
    findoutliers(x)
  } else{
    NA
  }
})
```



```
> print(outliers_list)
$Age
integer(0)

$Sex
[1] NA

$BMI
numeric(0)

$Children
integer(0)

$Smoker
[1] NA

$Region
[1] NA

$Medical.Cost
numeric(0)
```

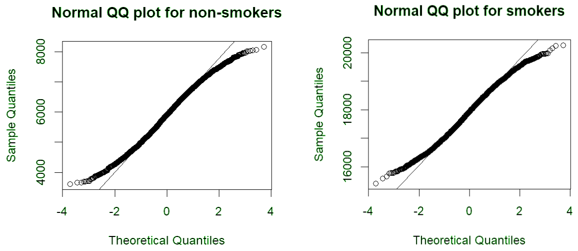
- I also created quantile-quantile plots (QQ-plots) to **determine** if data distributions for the relationships between demographic variables and medical costs contain **outliers** and is **normally distributed**. The QQ-plots showed that there are no outliers in the variables I am testing. The plots also presented the data as normally distributed for the Smoker Status variable. However, the QQ-plots corresponding to the variables Gender and Region, outlined in red and orange, displayed data that followed a skewed distribution.

- Quantile-Quantile plots for **Gender and Medical Cost:** Skewed distribution

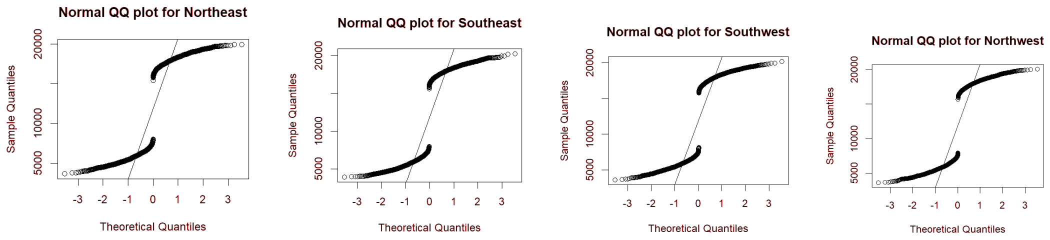


:

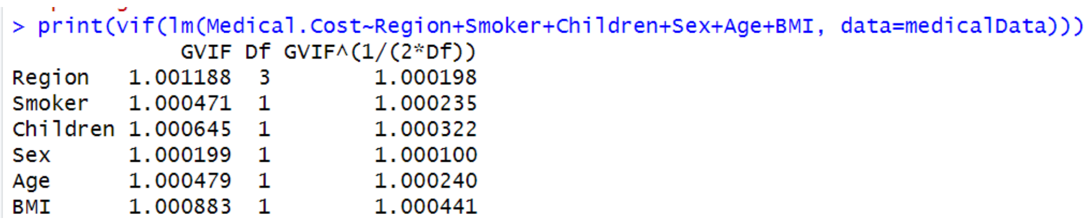
- Quantile Quantile plots for **Smoking Status and Medical Cost:** Normal Distribution



-Quantile Quantile plots for **Region and Medical Cost:** Skewed Distribution



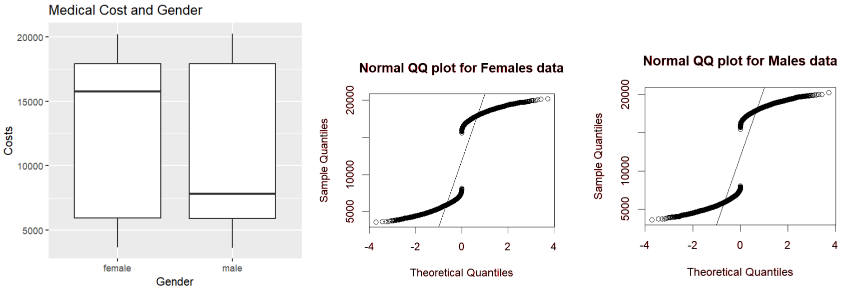
- Key Findings Part 2: Negligent Multicollinearity (VIF):
- Before running multiple hypothesis tests to analyze individual relationships between specific independent variables (Gender, Smoking Status, Region) and the dependent variable (Medical Cost), it is important to determine whether the independent variables are correlated with other independent variables in our dataset. To investigate this, I measured the Variance Inflation Factor (VIF) which 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. I found that each independent variable had a VIF value of less than 1.0004. This indicates there was negligible multicollinearity between the independent variables and that we could proceed with our tests.



```
> print(vif(lm(Medical.Cost~Region+Smoker+Children+Sex+Age+BMI, data=medicalData)))
             GVIF Df GVIF^(1/(2*Df))
Region   1.001188  3        1.000198
Smoker   1.000471  1        1.000235
Children 1.000645  1        1.000322
Sex      1.000199  1        1.000100
Age      1.000479  1        1.000240
BMI      1.000883  1        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.

- Part 2: Plots and Hypothesis tests
- Plot 1: Medical Cost and Gender:

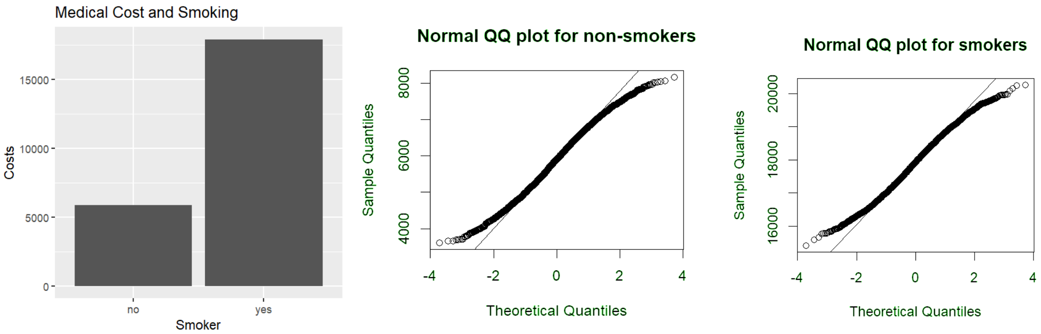


- This box plot shows a nearly identical range of healthcare cost for males [3617.09 to 20268.21] vs females [36532.5 to 20234.32] along with a similar first quartile 5895.767 (males) vs 5917.767 (females) and third quartile 17915.062 (males) 17953.912 (females). Interestingly, the boxplot also depicts a significant difference in median healthcare costs across genders, with females at $15781.390 and males at $7793.610.
- Quantile-Quantile plots were created for determining the normality of the data. These findings indicated that our data is skewed. I decided to proceed with a Mann-Whitney U-test, a non-parametric hypothesis test that compares median medical costs across genders. 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 similarly skewed 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 ratio and is thus at least ordinal.
- 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 our conclusion of our Mann-Whitney U-test, that the median healthcare cost for females is significantly higher than that of males as seen in the box plot. This presents as an *anomaly* that requires greater analysis. To further explore this *anomaly*, I ran a multiple linear regressions analysis to determine the

impact of all independent variables (potential confounding variables) on medical costs and compared it to a simple regression analysis of gender on medical cost. The results of this test indicated that both tests were statistically insignificant since they returned p-values greater than 0.05. This result led me to deduct that confounding variables had little influence on the skewed medical cost medians 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 healthcare cost is more likely to be influenced by other independent variables rather than Gender. The anomaly in median costs between genders is likely the result of random variation rather than systemic bias.

- Plot 5: Medical Cost and Smoking:

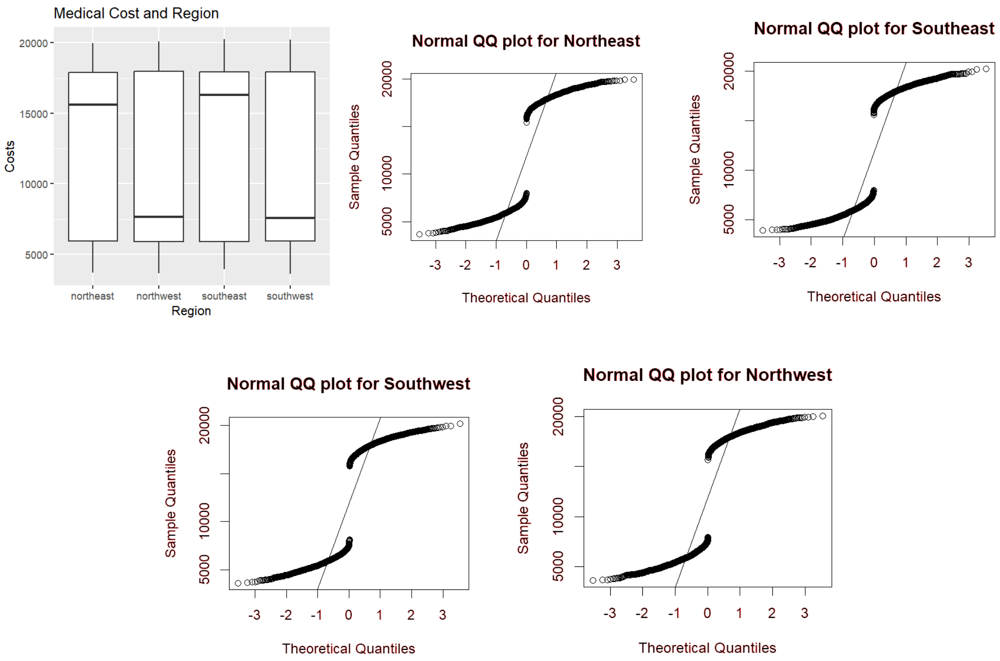


- This bar plot shows the mean healthcare costs for smokers (17922.253) are significantly higher than those for non-smokers (5894.855). To further analyze whether this difference is statistically significant, I decided to run a 2-sample t-test to compare mean medical costs across smokers and non-smokers. Listed are the **assumptions of the 2-sample t-test:**
    1) Normality: Each group (smokers and nonsmokers) and their associated medical cost appears to be 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 there was no bias in data acquisition.
    4) Homogeneous Variance: The data has approximately the same variance with smokers (726031.6) and non-smokers (721854.5) being nearly equal and the variance ratio between these two groups valued at less than four.
    5) Ratio Dependent variable: The data of each group can be measured on a ratio scale allowing for comparison with multiplication and division.
- **Hypothesis testing:**
    a. 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 test was that the probability value (p-value) is less than 2.2e^-16 meaning there is a miniscule probability of observing our data given that the

null hypothesis is true. This miniscule p-value means we reject the null hypothesis, indicating a significant difference between means for medical costs of smokers and non-smokers.

- Confidence interval: Another detail from the 2-sample t-test was the confidence interval for our plot telling us 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 a higher mean medical costs for smokers at $17922.253 compared to the mean cost of non-smokers $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:



- This box plot shows the median healthcare costs in the United States is 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. To further investigate this difference in medians for statistical significance, I ran a Kruskal-Wallis test which can test across more than two groups under non-normal conditions, as indicated by the QQ-plot. These are following **assumptions of the Kruskal-Wallis test**:
    - Ratio Dependent Variable: The dependent variable medical cost is measured on a ratio level allowing for accurate comparisons with multiplication and division.
    - Independence and Comparison: the independent variable region contains more than two independent groups that can be compared since nobody can live in more than one region simultaneously.

    - Homogeneous Distributions: All groups (northwest, northeast, southwest, southeast) are similar in skewness and shape as shown in the QQ-plots above.
- **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 that the null hypothesis is true is 0.92. This p-value is significantly greater than the common p-value of 0.05 which indicates that median medical costs do not significantly differ by region and we fail to reject the null hypothesis.
- **Anomaly and Further Analysis:**
- Contrary to this conclusion, the boxplot depicts varying median medical costs across regions which presents an anomaly. Investigating further, I tested for potential confounding variables that may skew the data and cause this anomaly. I ran a multiple regressions analysis to understand the impact of all independent variables (potential confounding variables) on medical costs and compared it with a simple regression analysis analyzing the impact of region on medical cost. Both tests yielded p-values significantly greater than 0.05, indicating no statistical significance. This finding suggests that confounding variables had little influence in skewing the medians and that region is not a significant factor in determining healthcare costs.
- 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 our differences 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:
    - My procedures entailed ensuring clean data, testing for VIF, developing visualizations, and analyzing hypothesis tests for the independent variables Gender, Smoking, and Region and the dependent variable Medical Cost. 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 likely has the smallest 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. People who smoke usually have higher medical costs than those who do not smoke.
        3) **Region**: Finally, median healthcare costs do not significantly vary by region in the US.
    - Therefore, smoking status was the most influential on medical costs while gender and region were the least 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 see how random variation and potential external variables play a role in skewing the medians (region and gender) would allow us to better understand this anomaly.

    - **Expanding the dataset:** I recommend that we perform a study that collects a larger sample of data which will lead to increased accuracy, lower margin of error, and greater representation in performing our study. A larger dataset leads to more accurate conclusions regarding the relationship between Gender, Smoking Status, Region, and healthcare costs.

    - **Further analysis omitted variables**: I initially intended to analyze the additional variables Age, BMI, and number of Children for their impact on medical costs. However, after completing my graphs, tests, and analysis, I found that the initial QQ plots I used for testing normality for these relationships were incorrectly formed. In creating new QQ plots that adjusted for my mistake, I found the relationships between these variables and medical costs to be skewed, meaning the normality assumption costs could not be met for these variables. This error unfortunately rendered the tests I conducted to analyze the relationships between the variables Age, BMI, and Number of Children with medical costs to be invalid.
    - Despite conducting hours of data analysis on these variables, I decided that they could not be included in my findings. Documentation on my flawed analysis is stored in a document which helps me analyze my errors to better prepare me for next data analysis.
    - **Recommendations for further analysis**: I would advise that these relationships would be analyzed and tested with the following:
        - <u>Age and BMI</u>: Both variables should be tested with the Spearman's Rank Order Correlation Test which is a non-parametric test that is used for finding correlation in scatter graphs with skewed data distributions.
        - <u>Number of Children</u>: A Kruskal Wallis Test should be used to test for differences across multiple independent groups (# of children) for skewed data.
        - <u>Additional tests</u>: If the data presents anomalies, it is important that we run additional tests to understand their causes and implications.

- ## Questions:
- I am a passionate undergraduate rising second year mathematics major at Washington and Lee university. In this project, I seeked to review and apply concepts learned in my probability course to a real world application. In the span of three weeks during my

summer vacation, I learned data analysis techniques and deepened my understanding of statistical methods. These are questions I will look to answer about the dataset:

  - **Non-parametric Tests**: How would non-parametric tests (Kruskal-Wallis and Mann-Whitney U-test) be performed without the use of statistical software such as Python or R? Is this a reasonable task for a statistician or data analyst at work?
  - **Ranking Influence on Medical Cost**: How would I rank the independent variables from most to least influential on medical cost? How would I quantify these rankings?
  - **Reducing Medical Costs**: If possible, how could I use my findings to reduce the pricing of medical costs for immutable factors?
  - **Survey Bias Determination**: Is there a way to analyze the data to determine whether the source was from a biased survey?
  - **Efficiency**: How can I improve my ability to efficiently analyze data with more timely and accurate findings?
- Overall, this project was a fruitful learning experience. I look forward to answering these questions and further developing my statistical analysis skills in the near future.

Acknowledgements:

- I acknowledge the use of Chat GPT for assistance in this project. Though the analysis and conclusions are my own, Chat GPT had a role in leading me through many aspects of my data analysis. I used Chat GPT as a data science learning tool, using this knowledge to apply to my first data science project.

Bibliography:

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