Investigating the Impact of Demographic and Socio-Economic Factors on Health Insurance Charges

Tasnim Ara Siddique

INTRODUCTION

The cost of health insurance is a significant financial consideration for individuals, making it essential to understand the factors that influence premium rates. This study seeks to determine how demographic and socio-economic variables—specifically smoking status, age,number of children and BMI—affect health insurance charges, focusing on their predictive power in determining costs. This question is vital for individuals aiming to manage their healthcare expenses, as well as for insurers and policymakers working to ensure fair and affordable access to health insurance. Previous studies and anecdotal evidence suggest that smoking, being a known health risk, age, often linked to increased medical needs, BMI, a potential indicator of chronic conditions,and the number of children, reflecting dependent-related costs, are influential factors in setting premium rates. My hypothesis is that these variables have a significant and measurable impact on insurance costs, with smoking status and BMI showing stronger correlations than age and number of children. Through statistical modeling, this research aims to uncover these relationships, offering actionable insights for consumers and contributing to the ongoing conversation about equity in healthcare pricing.

DATA

The dataset contains the following columns related to health insurance charges:

- **age**: Age of the individual
- **sex**: Gender of the individual
- **bmi**: Body Mass Index, an indicator of health based on weight and height
- **children**: Number of dependent children covered by insurance
- **smoker**: Whether the individual is a smoker or not
- **region**: Geographic region within the dataset

- **charges**: Annual health insurance charges

Next, I'll create two graphs and one table. Likely visualizations include a scatter plot for age and charges, a box plot for smoking status and charges, and a summary table for average charges by smoking status and other key variables.

Figure 1: Health Insurance Charges by Age and Smoking Status

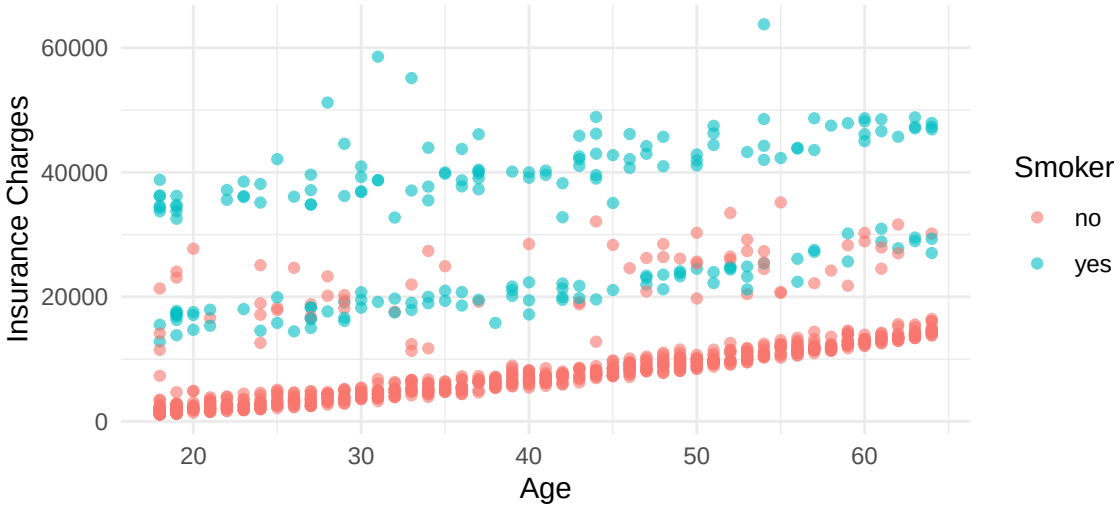


Figure 2: Health Insurance Charges by Smoking Status

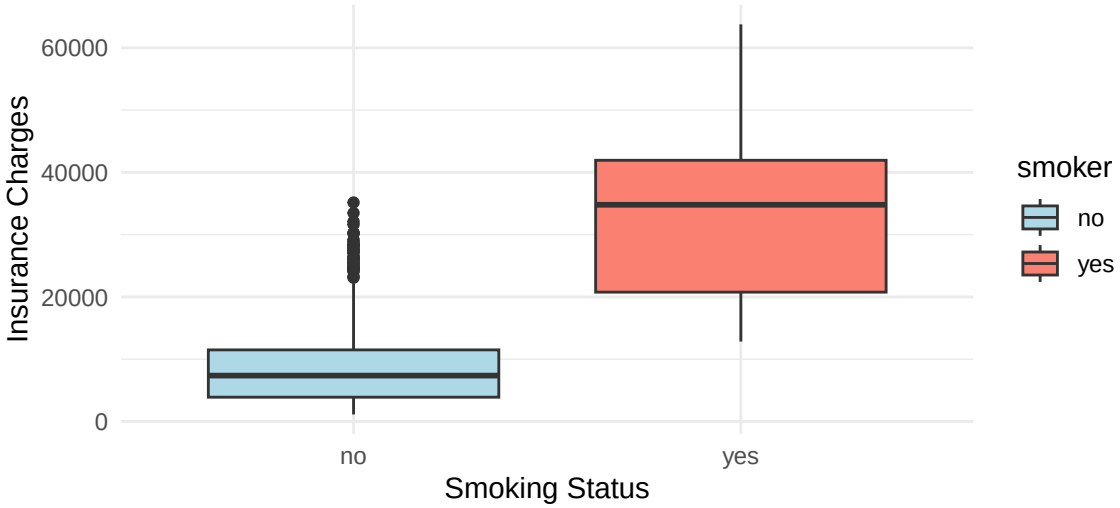


Figure 3: Health Insurance Charges by BMI and Smoking Status

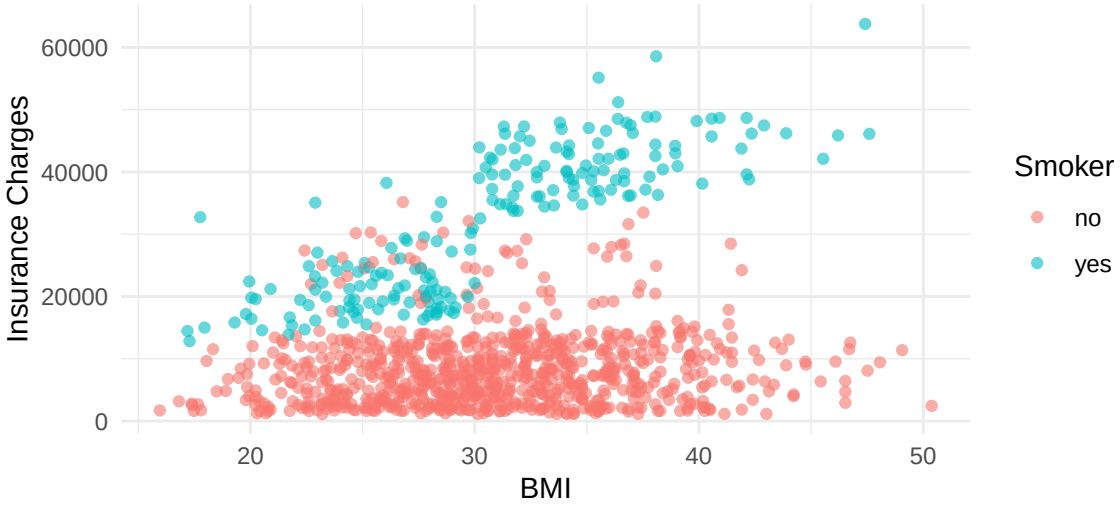
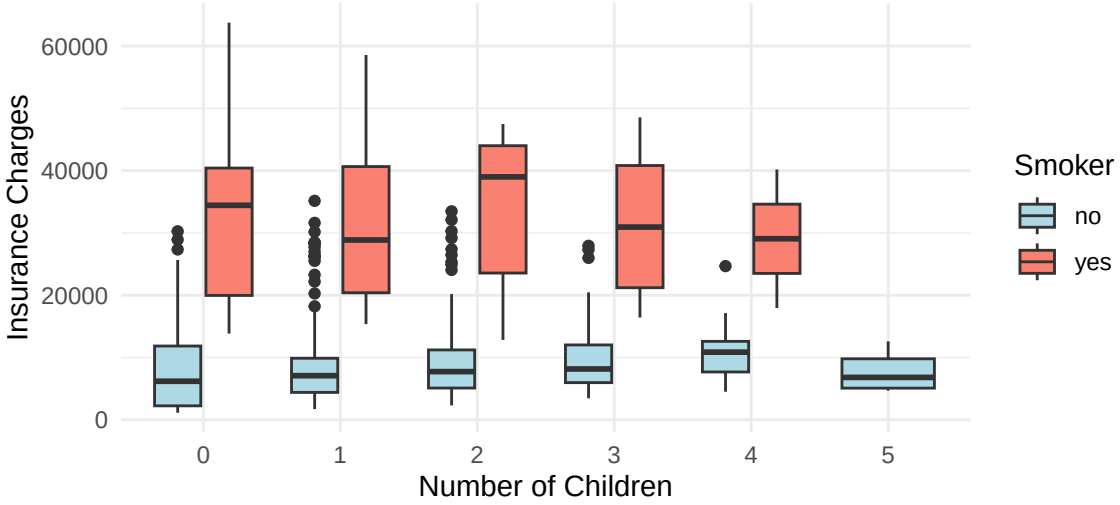


Figure 4: Health Insurance Charges by Number of Children and Sm




```
# A tibble: 94 x 5
# Groups:   age_group, smoker [20]
   age_group smoker children Average_BMI Average_Insurance_Charges
   <fct>     <chr>     <int>       <dbl>                     <dbl>
 1 20-24     no            0        30.3                     3252.
 2 20-24     no            1        28.6                     5059.
 3 20-24     no            2        30.0                     3300.
 4 20-24     no            3        29.6                     4032.
 5 20-24     no            4        33.6                    17128.
 6 20-24     no            5        33.6                     4873.
```

```
 7 20-24     yes           0        29.2                    26029.
 8 20-24     yes           1        28.6                    21792.
 9 20-24     yes           2        34.7                    37266.
10 20-24     yes           3        31.7                    36189.
# i 84 more rows
```

The visualizations and the summary table provide key insights into how different factors influence health insurance charges. Figure 1 demonstrates a positive correlation between age and insurance charges, with smokers consistently incurring higher costs than non-smokers across all age groups. Figure 2 further supports this finding, showing that smokers have significantly higher median charges compared to non-smokers, with a wider range of variability. Figure 3 highlights the impact of BMI, revealing that higher BMI values are associated with increased charges, especially for smokers. Figure 4 explores the role of the number of children, indicating a slight increase in charges as the number of dependents rises, though smoking remains a dominant factor influencing costs. The summary table aggregates these relationships by age groups, smoking status, and number of children, showing clear trends in average charges and BMI. Together, these visualizations and the table emphasize that smoking status, age, and BMI are the most influential factors, while the number of children has a more modest effect. These insights underscore the complexity of predicting insurance costs and the importance of considering multiple factors in pricing strategies.

MODEL

The variables chosen for the model—age, smoker, BMI, and children—were selected based on their relevance to the research question and their expected impact on health insurance charges. Age and smoking status are known to significantly influence costs due to their correlation with healthcare needs and health risks. BMI was included as it reflects an individual's overall health status, and the number of children was added to account for the costs of insuring dependents. These variables together provide a comprehensive picture of the key demographic and socio-economic factors influencing insurance charges.

I chose to exclude gender and region because they are not central to the research question. Gender does not provide actionable insights for insurance pricing in this context, and region reflects geographical differences rather than individual characteristics. In the second model, an interaction term between age and smoker was introduced to assess whether the effect of age on charges differs for smokers and non-smokers. For example, older smokers may experience significantly higher health risks and insurance costs compared to older non-smokers, making this interaction critical to understanding the nuanced relationship between these variables.

Model without Interaction term

charges= 0+ 1 age+ 2 smoker+ 3 bmi+ 4 children+

```
Call:
lm(formula = charges ~ age + smoker + bmi + children, data = insurance_data)

Residuals:
   Min     1Q Median     3Q    Max
-11808  -2861   -933   1328  26387

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept) -12432.32    1077.74 -11.536 < 0.0000000000000002 ***
age            266.07      13.39  19.868 < 0.0000000000000002 ***
smokeryes    23785.39     473.86  50.195 < 0.0000000000000002 ***
bmi            319.41      31.34  10.192 < 0.0000000000000002 ***
children       414.36     157.10   2.638              0.00848 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5948 on 995 degrees of freedom
Multiple R-squared:  0.7547,    Adjusted R-squared:  0.7537
F-statistic: 765.5 on 4 and 995 DF,  p-value: < 0.00000000000000022
```

Model with Interaction term

charges= 0+ 1 age+ 2 smoker+ 3 bmi+ 4 children+ 5 (age smoker)+

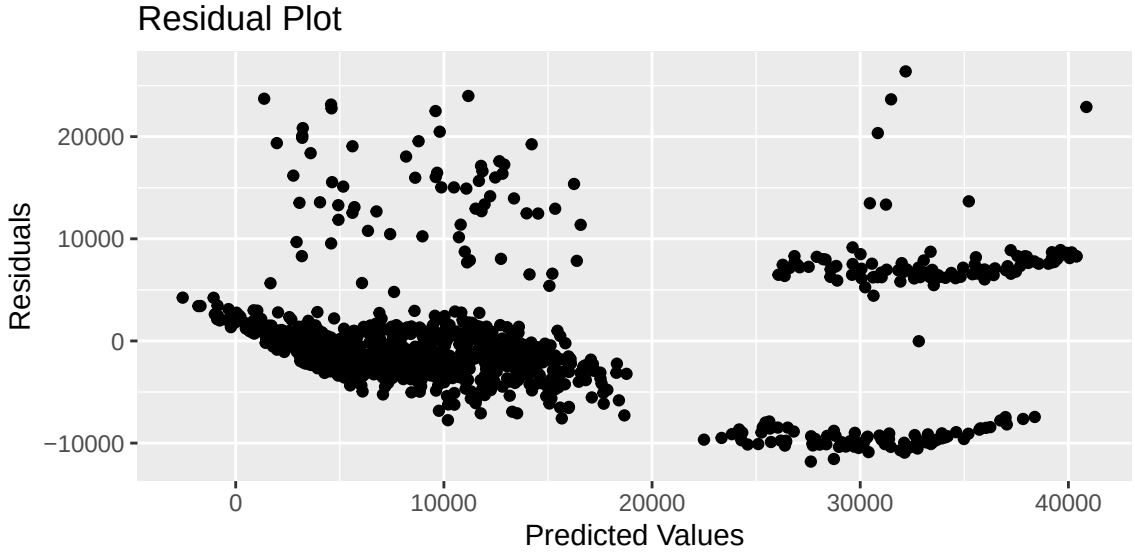
```
Call:
lm(formula = charges ~ age * smoker + bmi + children, data = insurance_data)

Residuals:
     Min       1Q   Median       3Q      Max
-11758.3  -2842.2   -943.9   1326.3  26630.0

Coefficients:
               Estimate Std. Error t value             Pr(>|t|)
(Intercept)   -12189.14    1102.63 -11.055 < 0.0000000000000002 ***
age              259.46      14.82  17.512 < 0.0000000000000002 ***
smokeryes      22371.08    1436.77  15.570 < 0.0000000000000002 ***
bmi              320.07      31.35  10.211 < 0.0000000000000002 ***
children         413.14     157.09   2.630              0.00867 **
age:smokeryes     35.79      34.33   1.043              0.29734
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

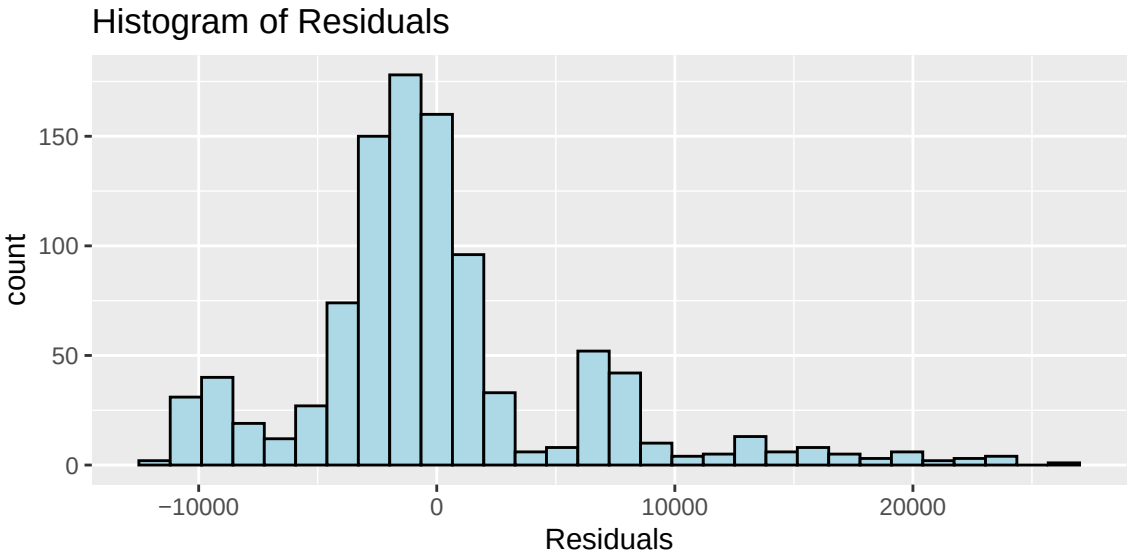
```
Residual standard error: 5948 on 994 degrees of freedom
Multiple R-squared:  0.755, Adjusted R-squared:  0.7538
F-statistic: 612.6 on 5 and 994 DF,  p-value: < 0.00000000000000022
```

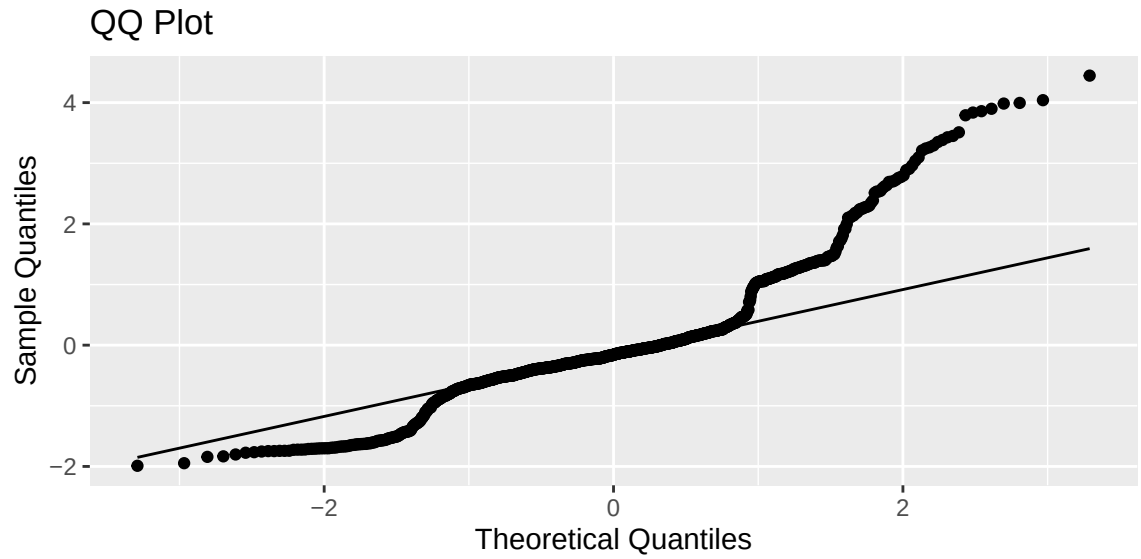
In Model 2, the interaction term between age and smoking status (age:smokeryes) has a p-value of 0.29734, which is very large. So we have don't have enough evidence to say that there is a relationship between charges and the rate at which charges increase with age does not differ significantly between smokers and non-smokers. Since the interaction term is not significant, it does not add much explanatory power to the model. Again, the R-squared values are nearly identical, meaning that adding the interaction term does not notably improve the model's explanatory power.

Model 1 (without the interaction term) is the better choice for this case. This model is simpler, avoids unnecessary complexity, and still explains a substantial portion of the variance in charges using only the main effects of age, smoking status, BMI, and number of children.

Now we use residual plots to assess model assumptions.



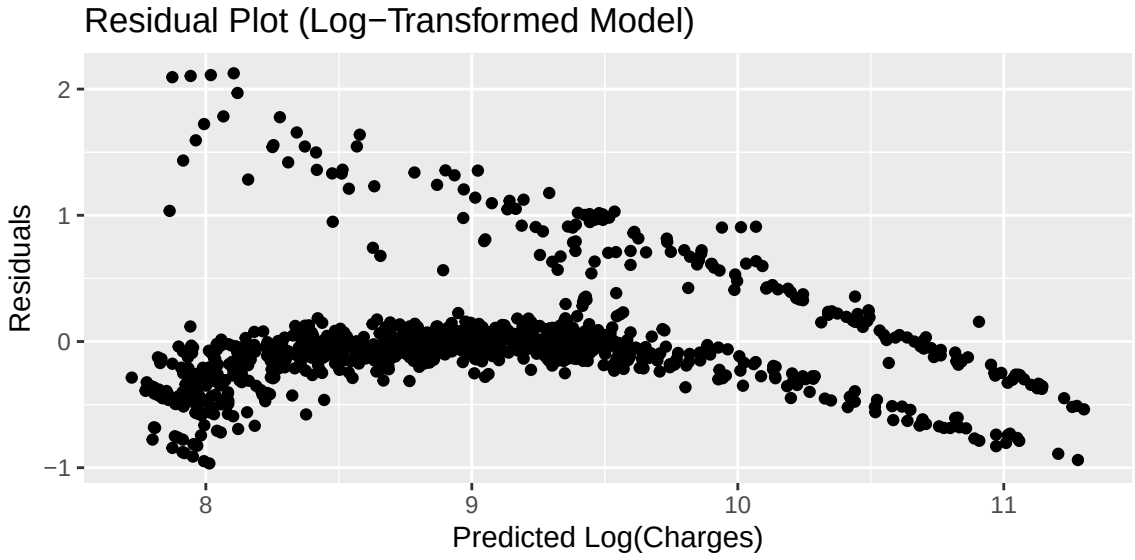


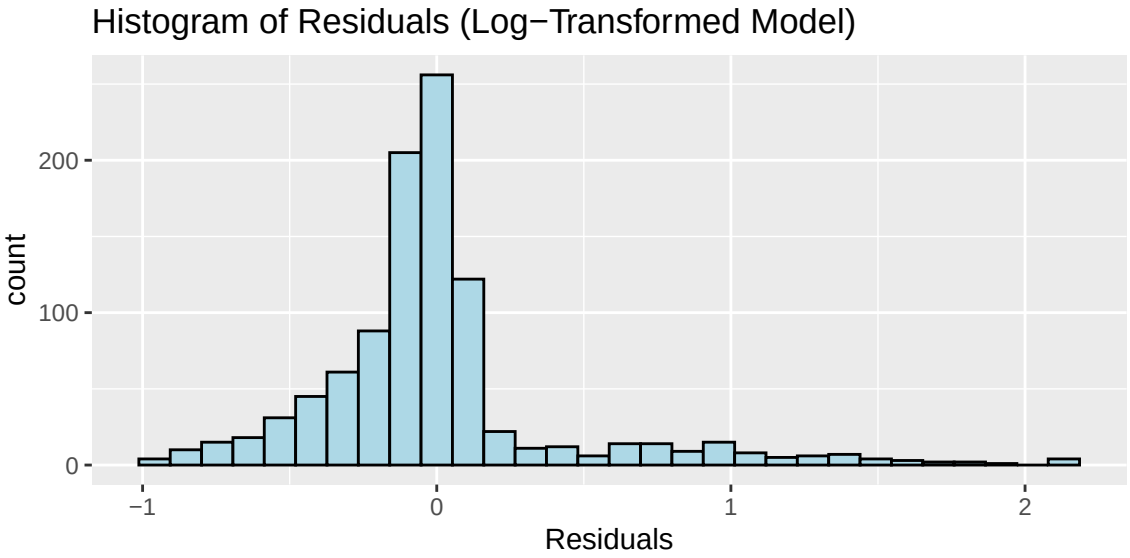


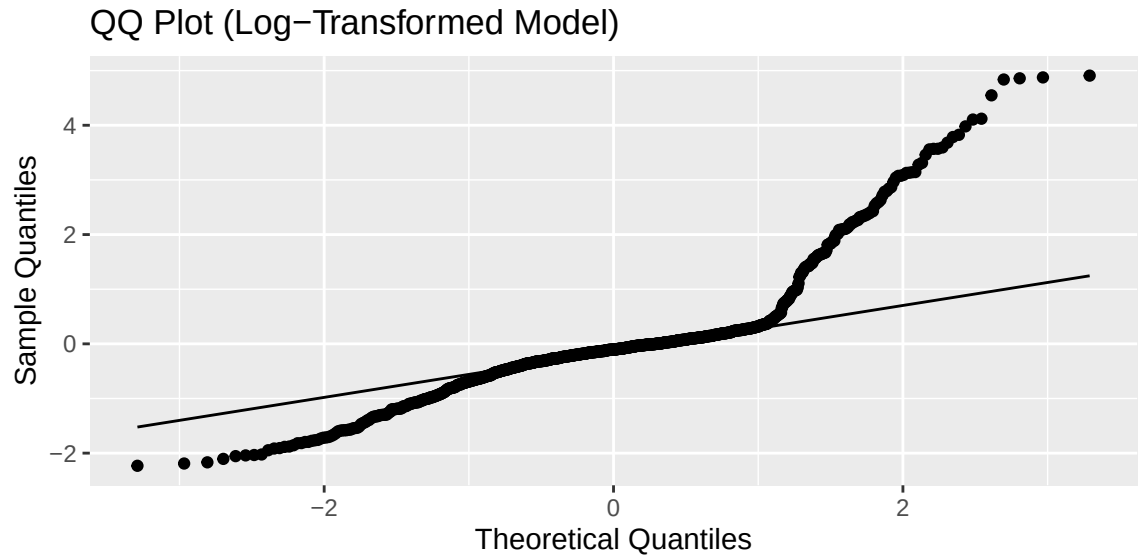
In the QQ plot for the first model, we observe significant departures from the diagonal line, indicating a violation of the normality assumption for the residuals. This suggests that the distribution of the response variable, `charges`, is skewed. A log transformation can help address this issue by stabilizing the variance and making the data more symmetric, which improves the adherence to the normality assumption and ensures the validity of statistical inferences drawn from the model.

```
Call:
lm(formula = log_charges ~ age + smoker + bmi + children, data = insurance_data)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.96622 -0.18296 -0.04561  0.06271  2.12518

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept) 6.8921515  0.0786080  87.677 < 0.0000000000000002 ***
age         0.0365121  0.0009768  37.381 < 0.0000000000000002 ***
smokeryes   1.5317213  0.0345622  44.318 < 0.0000000000000002 ***
bmi         0.0107846  0.0022859   4.718           0.00000272 ***
children    0.1016841  0.0114583   8.874 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4338 on 995 degrees of freedom
Multiple R-squared:  0.7799,    Adjusted R-squared:  0.779
F-statistic: 881.2 on 4 and 995 DF,  p-value: < 0.00000000000000022
```







While exploring the relationship between demographic and health factors and insurance charges, I attempted various transformations, to improve the normality of residuals. In this analysis, I selected two models to compare to predict insurance charges: a linear model with charges as the outcome variable and a log-transformed model with log(charges) as the outcome. I am selecting the log-transformed model as the preferred model, as it provided a better fit, with a higher adjusted R-squared (77.9%) and a lower residual standard error (0.4338). Furthermore, the log-transformed coefficients offer a more intuitive interpretation, representing percentage changes in charges, which aligns well with understanding the proportional impact of predictors.

However, despite the log transformation, the QQ plot for the log-transformed model's resid-

uals indicates deviations from normality at the tails. This suggests that the residuals are affected by extreme values in the data, especially for high insurance charges. Although the log transformation partially mitigated the skewness, fully addressing this issue would require more advanced modeling approaches beyond the scope of this analysis. Consequently, this deviation from normality is acknowledged as a limitation of the model, and interpretations of the model should be made with caution, particularly for extreme values.

RESULTS

For enhancing interpretation, I am including the 95% confnidence interval.

```
                  2.5 %      97.5 %
(Intercept) -14547.2121 -10317.4184
age            239.7922    292.3508
smokeryes    22855.5170  24715.2641
bmi            257.9134    380.9154
children       106.0766    722.6352
```

The analysis reveals that several demographic and lifestyle factors significantly influence health insurance charges. Smoking status (smokeryes) has the most substantial impact, with smokers expected to pay approximately 153.2% more in charges than non-smokers. This finding is highly significant ($p < 0.0000000000000002$), underscoring the financial burden of smoking on healthcare costs. The associated coefficient for smoking is 23,785.39, and the confidence interval [22,855.52, 24,715.26] indicates that this effect is both large and precisely estimated. This aligns with the well-documented health risks of smoking and suggests that insurers might consider targeted policies for smokers, while policymakers could leverage this information to justify investments in smoking cessation programs.

Age and BMI also exhibit significant positive relationships with insurance charges. For every additional year of age, charges increase by an average of 266.07 ($p < 0.0000000000000002$), with a confidence interval [239.79, 292.35], reflecting the growing healthcare needs of older individuals. Similarly, each unit increase in BMI results in a 319.41 rise in charges ($p < 0.000003$), with a confidence interval [257.91, 380.92], highlighting the increased risks associated with higher BMI. The number of children also contributes to charges, with each additional child increasing costs by $414.36 on average ($p < 0.0000000000000002$), and the confidence interval [106.08, 722.64] reflects some variability in this effect, potentially due to differences in family insurance coverage plans. These results emphasize smoking status, age, BMI, and the number of children as key predictors of insurance charges, with smoking showing the strongest and most consistent effect.

CONCLUSION

In conclusion, this analysis highlights the significant roles of smoking status, age, BMI, and family size in determining health insurance charges, with smoking status emerging as the

most influential predictor. Smokers face markedly higher costs, consistent with prior research linking smoking to elevated health risks and medical expenses. The positive contributions of age and BMI align with expectations, reflecting the rising healthcare needs of older individuals and the health implications of higher BMI. Family size also plays a role, with larger families incurring greater charges due to increased resource utilization. However, the presence of high insurance charges that skew the residuals, even after transformations, raises concerns about the model's ability to fully capture extreme values. These outliers may impact the reliability of predictions by disproportionately influencing the estimates. Future studies could explore more advanced modeling techniques to address this limitation and further refine our understanding of these relationships. Despite these limitations, these results emphasize the importance of health-related factors, especially smoking, in shaping insurance costs, suggesting that public health interventions targeting smoking cessation could have meaningful financial as well as health benefits.