Data550 Final Dataset

Mia Yang

Final Project: Modeling Medical Insurance Charges Using Demographic and Lifestyle Factors

Besides the loading data code chunk, all other codes used to generate the table 1 and the boxplot will not be included in the report but will be in the .rmd file.

1. Introduction:

1.1 Dataset

My dataset was chosen from *kaggle* and it is a *Medical Insurance Cost dataset.*(Click on italic text to see the website and the data).

```
absolute_path_data_final <- here::here("raw_data", "insurance.csv")

final_data <- read.csv(absolute_path_data_final, header = TRUE)

head(final_data)
```

```
##   age    sex    bmi children smoker    region   charges
## 1  19 female 27.900        0    yes southwest 16884.924
## 2  18   male 33.770        1     no southeast  1725.552
## 3  28   male 33.000        3     no southeast  4449.462
## 4  33   male 22.705        0     no northwest 21984.471
## 5  32   male 28.880        0     no northwest  3866.855
## 6  31 female 25.740        0     no southeast  3756.622
```

1.2 Objective of this Project:

Health insurance costs are influenced by a range of demographic and lifestyle characteristics, such as age, sex, body mass index (BMI), smoking status, and geographic region. Understanding how these factors contribute to medical expenses can provide valuable insights for health policy, insurance providers, and individuals seeking affordable care.

1.3 Description of the Data:

The dataset used in this analysis, the Medical Insurance Cost Dataset (sourced from Kaggle), contains records for 1,338 individuals. Each record includes demographic variables (age, sex, region), lifestyle indicators (BMI, smoking status, number of children), and the corresponding insurance charges billed to the individual. This dataset is widely used for teaching statistical modeling and health economics because it offers a structured, real-world–inspired look at cost variation.

Characteristic	By Insurance Charges (High vs Low) Overall N = 1,338[1]	High N = 669[1]	Low N = 669[1]	p-value[2]
age	39 (27, 51)	51 (36, 57)	31 (23, 40)	<0.001
sex				>0.9
female	662 (49%)	332 (50%)	330 (49%)	
male	676 (51%)	337 (50%)	339 (51%)	
bmi	30.4 (26.3, 34.7)	31.1 (26.8, 35.5)	29.9 (25.8, 34.1)	0.001
children				0.016
0	574 (43%)	295 (44%)	279 (42%)	
1	324 (24%)	142 (21%)	182 (27%)	
2	240 (18%)	119 (18%)	121 (18%)	
3	157 (12%)	90 (13%)	67 (10%)	
4	25 (1.9%)	17 (2.5%)	8 (1.2%)	
5	18 (1.3%)	6 (0.9%)	12 (1.8%)	
region				0.2
northeast	324 (24%)	178 (27%)	146 (22%)	
northwest	325 (24%)	157 (23%)	168 (25%)	
southeast	364 (27%)	180 (27%)	184 (28%)	
southwest	325 (24%)	154 (23%)	171 (26%)	
smoker	274 (20%)	274 (41%)	0 (0%)	<0.001

[1] Median (Q1, Q3); n (%)
[2] Wilcoxon rank sum test; Pearson's Chi-squared test

2. Descriptive Analysis:

2.1 Table 1:

Characteristics of the 1338 data points from the insurance database are displayed in the table below.

Table 1 Analysis:

Table 1 summarizes the baseline characteristics of the study population (N = 1,338) stratified by insurance charges categorized as high versus low (above vs. below the median). Individuals in the high-charge group tended to be older, with a median age of 51 years compared to 31 years in the low-charge group ($p < 0.001$). BMI was also higher among those with high charges (median 31.1 vs. 29.9, $p = 0.001$). The distribution of sex was similar across groups (approximately half male and half female, $p > 0.9$). The number of children showed some differences, with the high-charge group slightly more likely to have three or more children ($p = 0.016$). Regional distribution did not differ significantly ($p = 0.2$). Notably, smoking status showed the strongest association with charges: 41% of the high-charge group were smokers, compared to none in the low-charge group ($p < 0.001$). These findings suggest that age, BMI, number of children, and smoking status are important predictors of higher insurance costs.

2.2 Table 2: Summary of Insurance Charges by Region

region	mean_charges	median_charges	sd_charges	min_charges	max_charges	n
southeast	14735.41	9294.13	13971.10	1121.87	63770.43	364

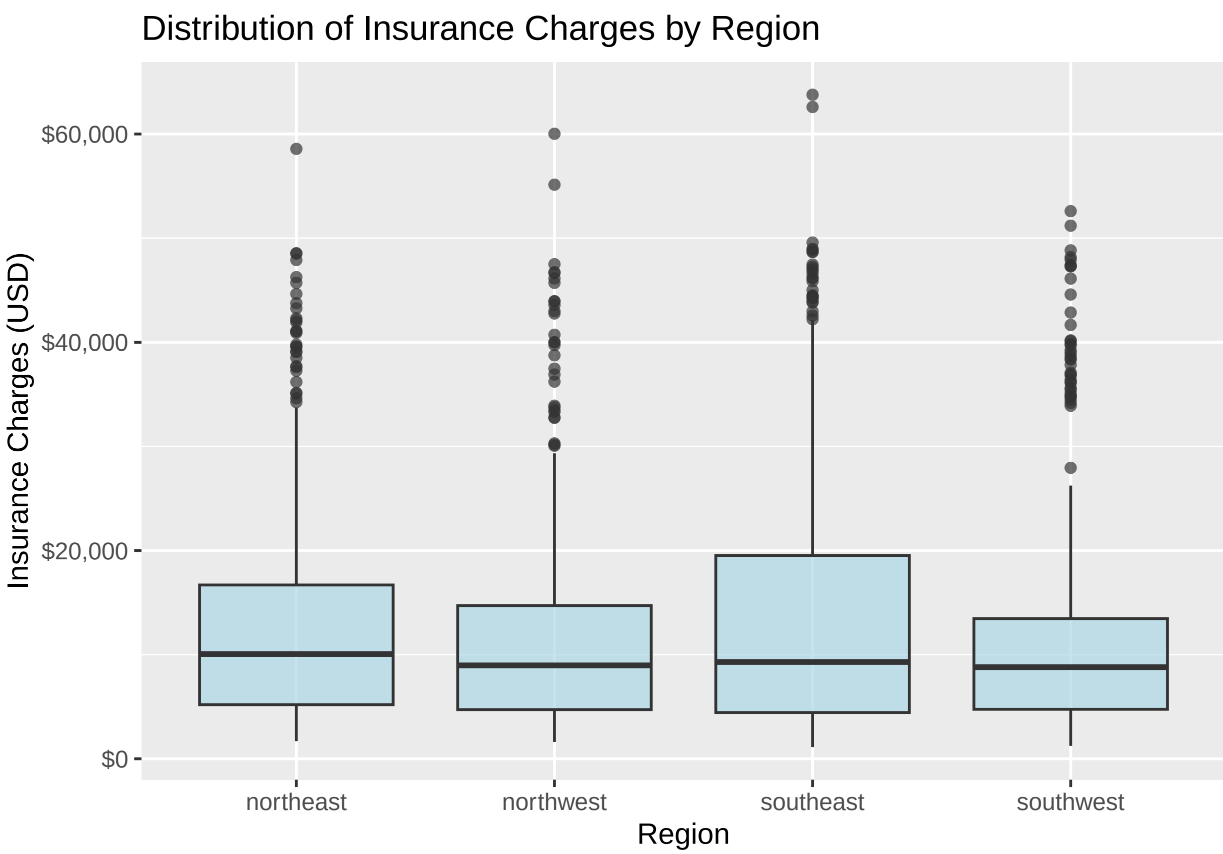
region	mean__charges	median__charges	sd__charges	min__charges	max__charges	n
northeast	13406.38	10057.65	11255.80	1694.80	58571.07	324
northwest	12417.58	8965.80	11072.28	1621.34	60021.40	325
southwest	12346.94	8798.59	11557.18	1241.57	52590.83	325

Table 2 Analysis:

Average insurance charges vary notably by region, with the southeast having the highest mean costs and the northwest and southwest showing lower averages. The wide variation within regions suggests differences in population health, lifestyle, and access to care, emphasizing the need to consider regional factors when evaluating medical expenses.

3. Visualization

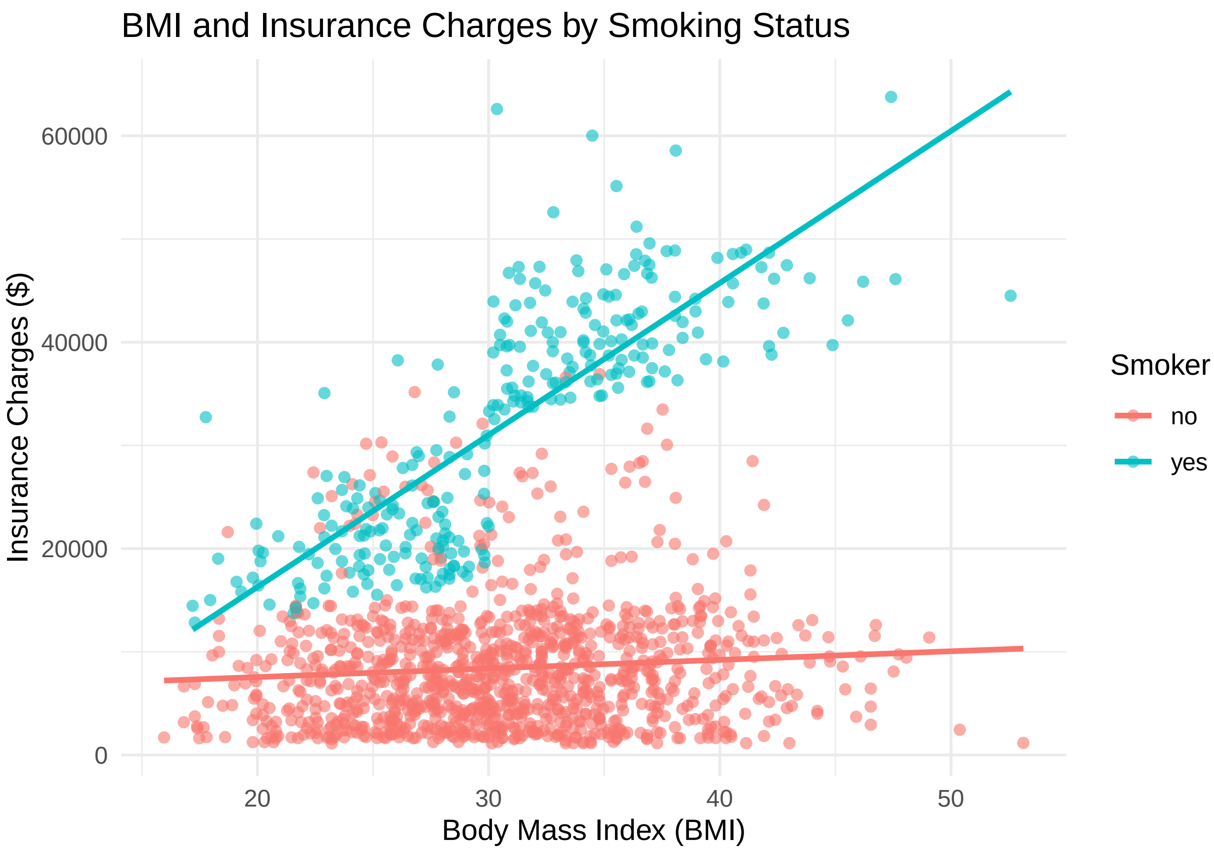
3.1 Boxplot of the Charges by Region



Boxplot by region Analysis:

The boxplot was created to examine whether geographic region is associated with differences in insurance charges. Region is a categorical factor in the dataset, and plotting charges across the four regions (northeast, northwest, southeast, and southwest) allows for a visual comparison of central tendencies, spread, and outliers. From the plot, we can see that the overall distributions of charges are broadly similar across regions, with medians clustering around comparable levels. However, the southeast region shows a slightly higher median and wider interquartile range compared to other regions, suggesting greater variability in charges. All regions exhibit numerous high-cost outliers, reflecting individuals with extremely high insurance expenses. This visualization suggests that while region may play some role in cost variability, other factors—such as smoking status, BMI, or age—are likely stronger predictors of differences in insurance charges.

3.2 Scatter Plot: BMI vs. Charges, Colored by Smoking Status



Scatter plot:

The scatter plot shows that insurance charges increase with higher BMI, particularly among smokers. Smokers consistently have higher medical costs than non-smokers across all BMI levels, with the steep upward slope indicating that obesity amplifies expenses for this group. In contrast, non-smokers show relatively stable charges regardless of BMI. Overall, smoking and high BMI appear to interact to substantially raise healthcare costs.