Project 2

Leyla Muminova

2025-11-18

Introduction and my motivation I chose a dataset from Kaggle titled *"Medical Insurance Cost Prediction"*, which includes different factors that influence the price of medical insurance. For this project, I decided to look at the question: **"How does the cost of medical insurance vary across regions for smokers compared to non-smokers?"** Coming from another country and still getting used to it, the medical insurance in the US is something interesting to observe and learn about. We usually assume that smokers pay more on average; however, I would like to analyze whether these differences are based on geographic region. By studying this question, I aim to determine whether the differences in charges are primarily driven by personal habits (such as smoking) or by broader regional economic and health system factors. As a student majoring in economics, I strongly believe that this question contributes to a broader conversation about inequality, price fairness, and other factors that shape insurance costs for individuals.

```
# Reading and importing a csv file

library(tidyverse)
```

Data Wrangling and Visualization

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 4.0.1     v purrr   1.1.0
## v tibble  3.3.0     v dplyr   1.1.4
## v tidyr   1.3.1     v stringr 1.5.2
## v readr   2.1.5     v forcats 1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
insurance <- read_csv("medical_insurance.csv")
```

```
## Rows: 100000 Columns: 54
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (10): sex, region, urban_rural, education, marital_status, employment_st...
## dbl (44): person_id, age, income, household_size, dependents, bmi, visits_la...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(insurance)
```

```
## # A tibble: 6 x 54
##   person_id   age sex    region  urban_rural income education    marital_status
##       <dbl> <dbl> <chr>  <chr>   <chr>        <dbl> <chr>        <chr>
## 1     75722    52 Female North   Suburban     22700 Doctorate    Married
## 2     80185    79 Female North   Urban        12800 No HS        Married
```

```
## 3     19865    68 Male   North   Rural        40700 HS           Married
## 4     76700    15 Male   North   Suburban     15600 Some College Married
## 5     92992    53 Male   Central Suburban     89600 Doctorate    Married
## 6     76435    63 Female North   Rural       305000 HS           Single
## # i 46 more variables: employment_status <chr>, household_size <dbl>,
## #   dependents <dbl>, bmi <dbl>, smoker <chr>, alcohol_freq <chr>,
## #   visits_last_year <dbl>, hospitalizations_last_3yrs <dbl>,
## #   days_hospitalized_last_3yrs <dbl>, medication_count <dbl>,
## #   systolic_bp <dbl>, diastolic_bp <dbl>, ldl <dbl>, hba1c <dbl>,
## #   plan_type <chr>, network_tier <chr>, deductible <dbl>, copay <dbl>,
## #   policy_term_years <dbl>, policy_changes_last_2yrs <dbl>, ...
```

```
class(insurance$smoker) # character
```

```
## [1] "character"
```

```
class(insurance$region) # character
```

```
## [1] "character"
```

```
class(insurance$sex) # character
```

```
## [1] "character"
```

```
insurance <- insurance %>%
  mutate(
    smoker = as.factor(smoker),
    region = as.factor(region),
    sex = as.factor(sex)
  )
```

After reading the CSV file, I decided to check the class of variables I am planning to work with, and after discovering that they are character strings, I decided to change them into factors, as it will make it easier to work since those columns are categories instead of text.

```
regional_smoker_summary <- insurance %>%
  group_by(region, smoker) %>%
  summarise(mean_charges = mean(annual_medical_cost))
```

```
## `summarise()` has grouped output by 'region'. You can override using the
## `.groups` argument.
```

```
regional_smoker_summary
```

```
## # A tibble: 15 x 3
## # Groups:   region [5]
##    region  smoker  mean_charges
##    <fct>   <fct>          <dbl>
##  1 Central Current        4194.
##  2 Central Former         3050.
##  3 Central Never          2729.
##  4 East    Current        4215.
##  5 East    Former         3141.
##  6 East    Never          2727.
##  7 North   Current        4273.
##  8 North   Former         3174.
##  9 North   Never          2732.
## 10 South   Current        4441.
## 11 South   Former         3238.
```

```
## 12 South   Never          2761.
## 13 West    Current        4263.
## 14 West    Former         3129.
## 15 West    Never          2772.
```

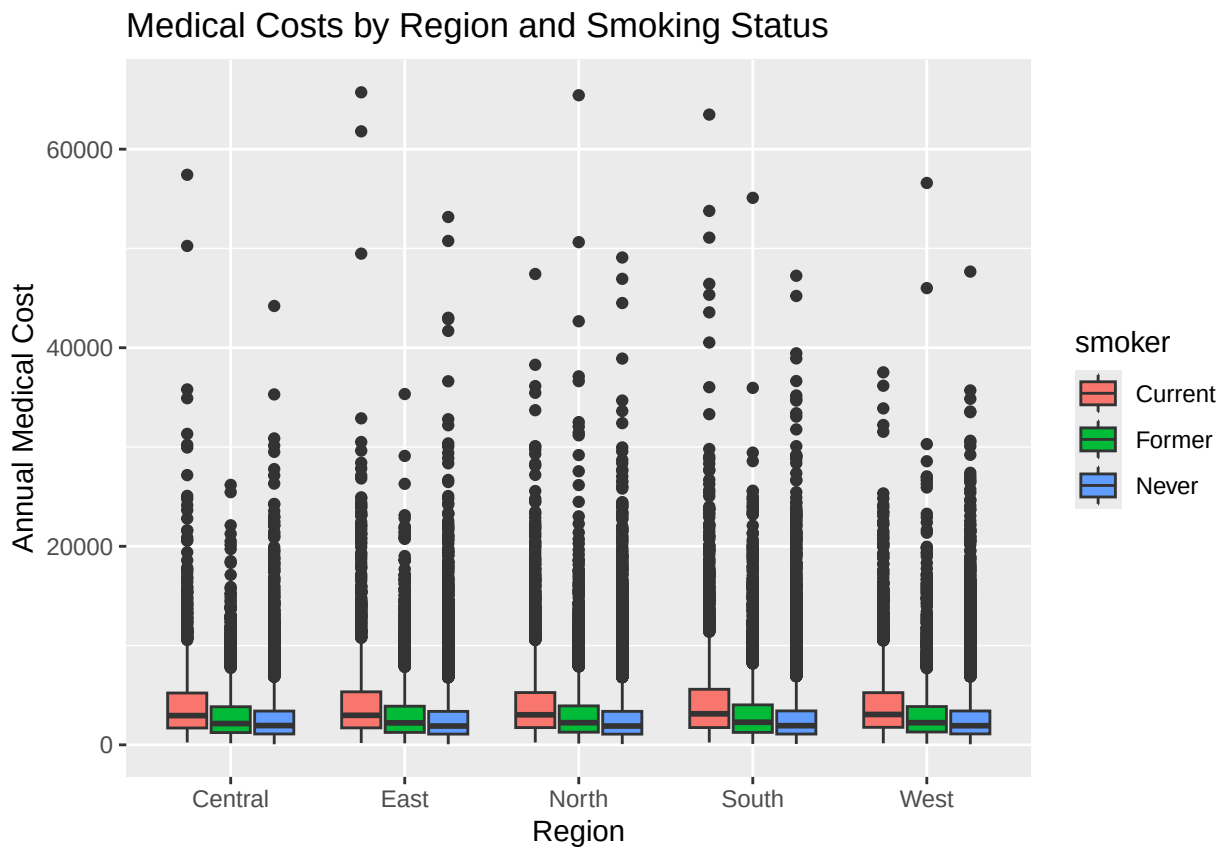
To start comparing smokers and non-smokers, I decided to group data by both region and smoker. Then I used summarise() function to calculate the mean of the medical cost for each group.

```
regional_smoker_wide <- regional_smoker_summary %>%
  pivot_wider(
    names_from = smoker,
    values_from = mean_charges
  )
regional_smoker_wide
```

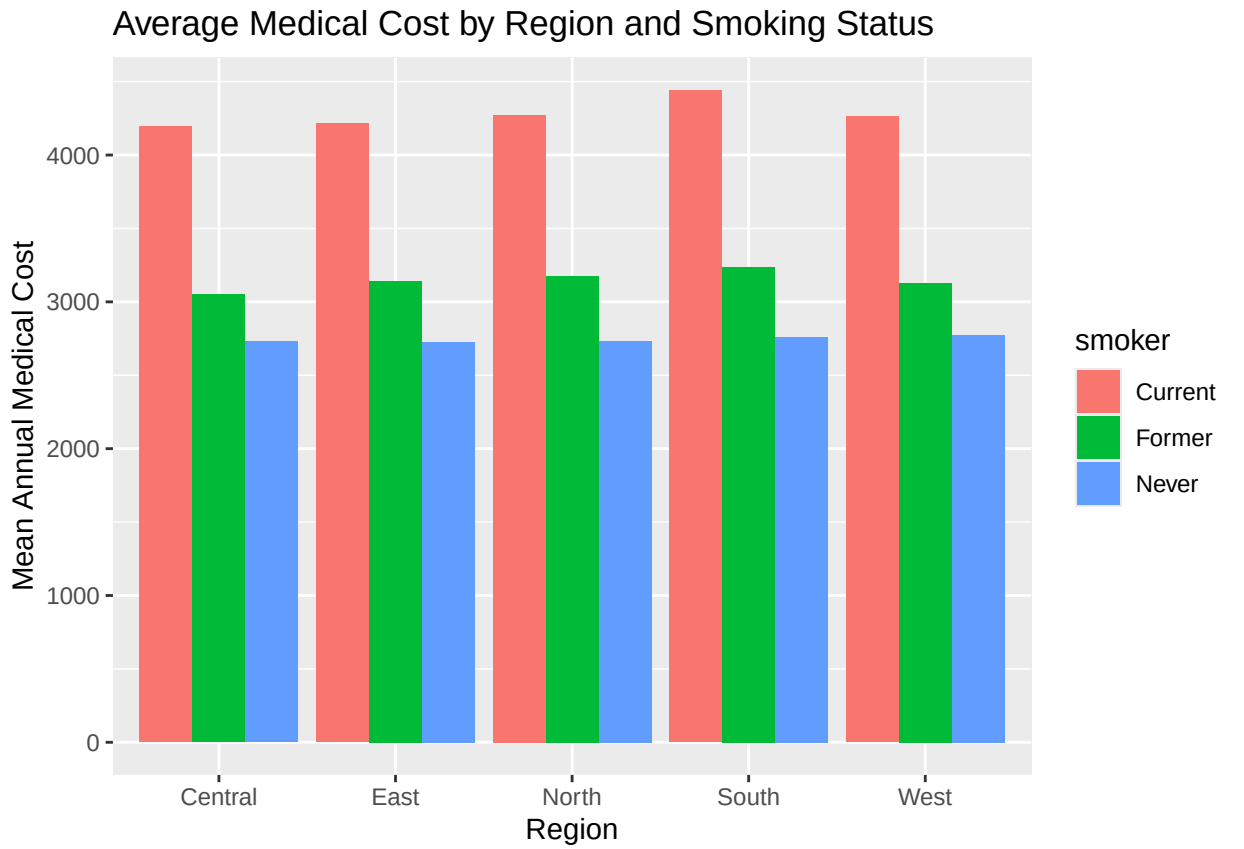
```
## # A tibble: 5 x 4
## # Groups:   region [5]
##   region  Current Former Never
##   <fct>     <dbl>  <dbl> <dbl>
## 1 Central   4194.  3050. 2729.
## 2 East      4215.  3141. 2727.
## 3 North     4273.  3174. 2732.
## 4 South     4441.  3238. 2761.
## 5 West      4263.  3129. 2772.
```

After creating the summary table. I used pivot_wider() to display smoker categories into separate columns. So, by doing this, it is now easier to analyze the difference between smokers and non-smokers for every region.

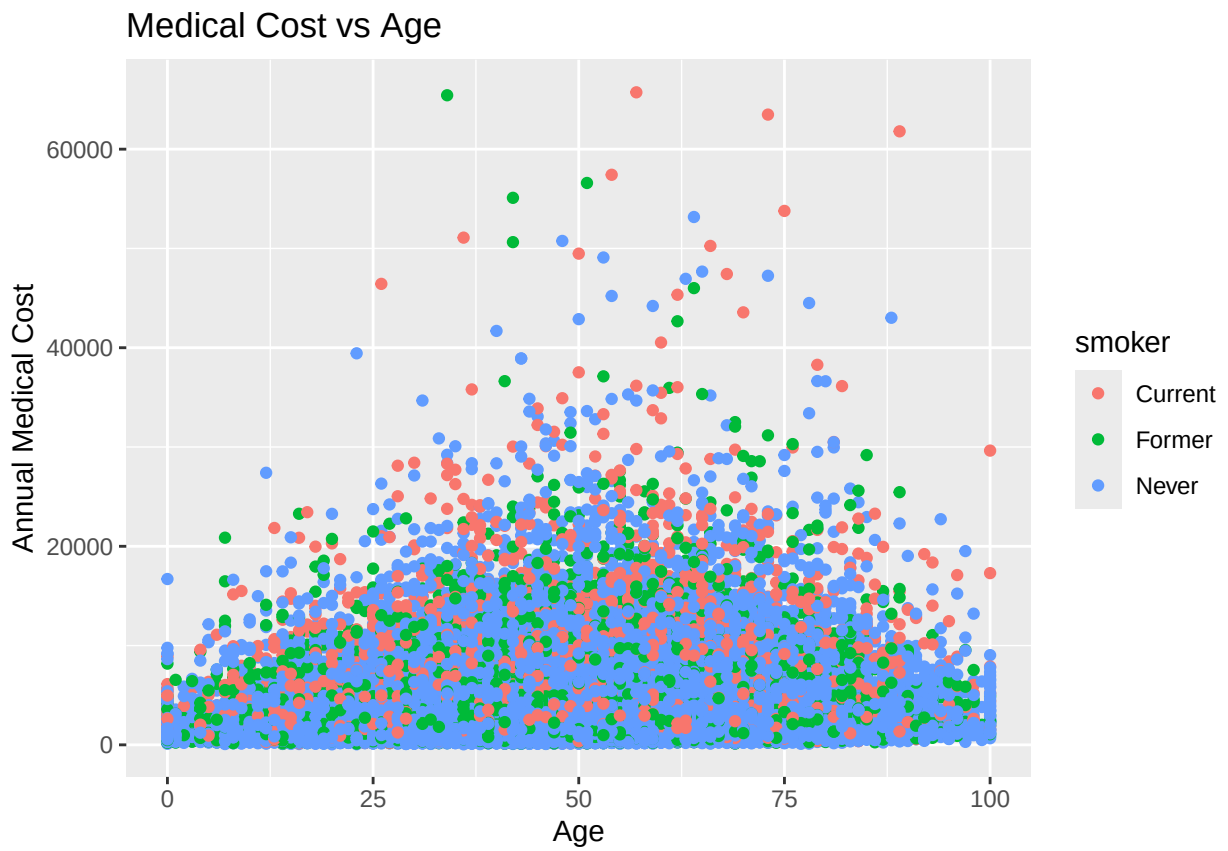
```
# Visualization 1
ggplot(insurance, aes(x = region, y = annual_medical_cost, fill = smoker)) +
  geom_boxplot() +
  labs(
    title = "Medical Costs by Region and Smoking Status",
    x = "Region",
    y = "Annual Medical Cost"
  )
```



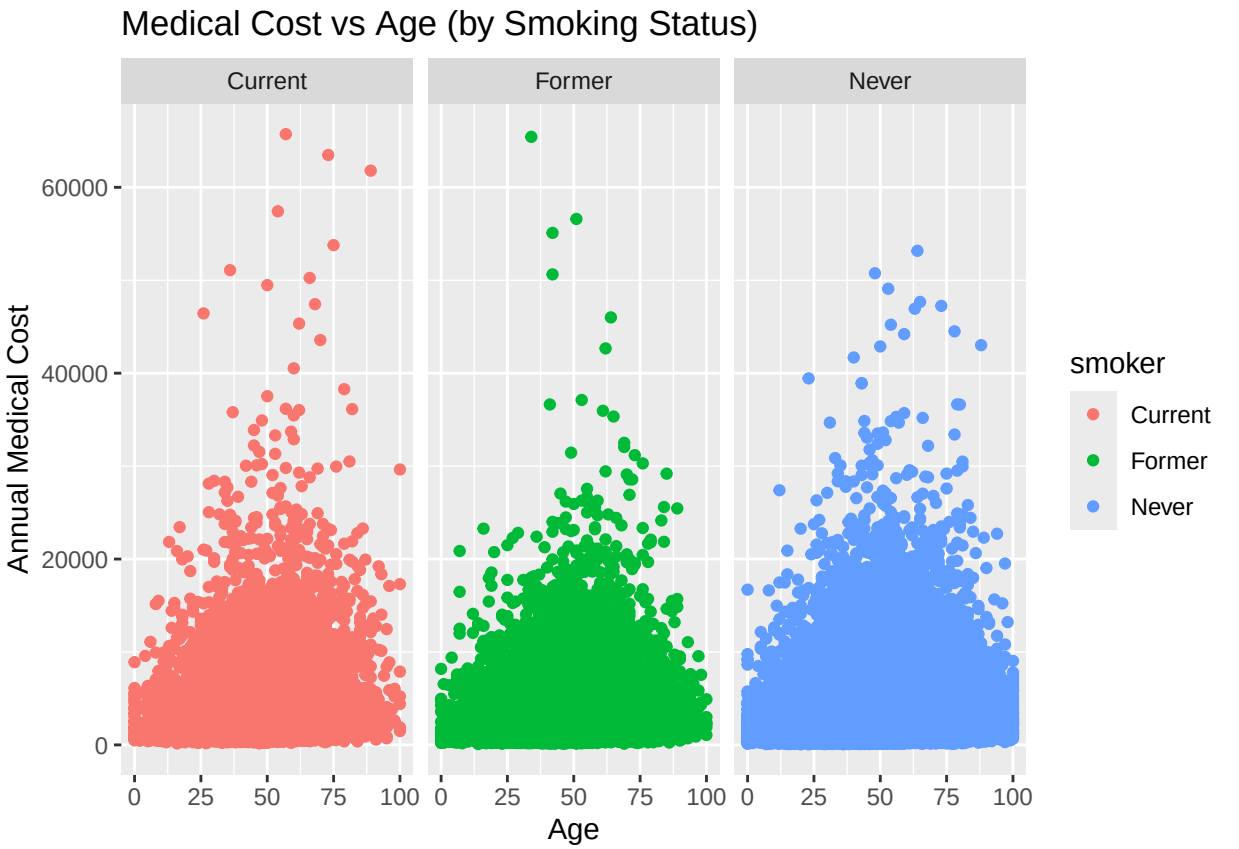

```
# Visualization 2
ggplot(regional_smoker_summary, aes(x = region, y = mean_charges, fill = smoker)) +
  geom_col(position = "dodge") +
  labs(
    title = "Average Medical Cost by Region and Smoking Status",
    x = "Region",
    y = "Mean Annual Medical Cost"
  )
```



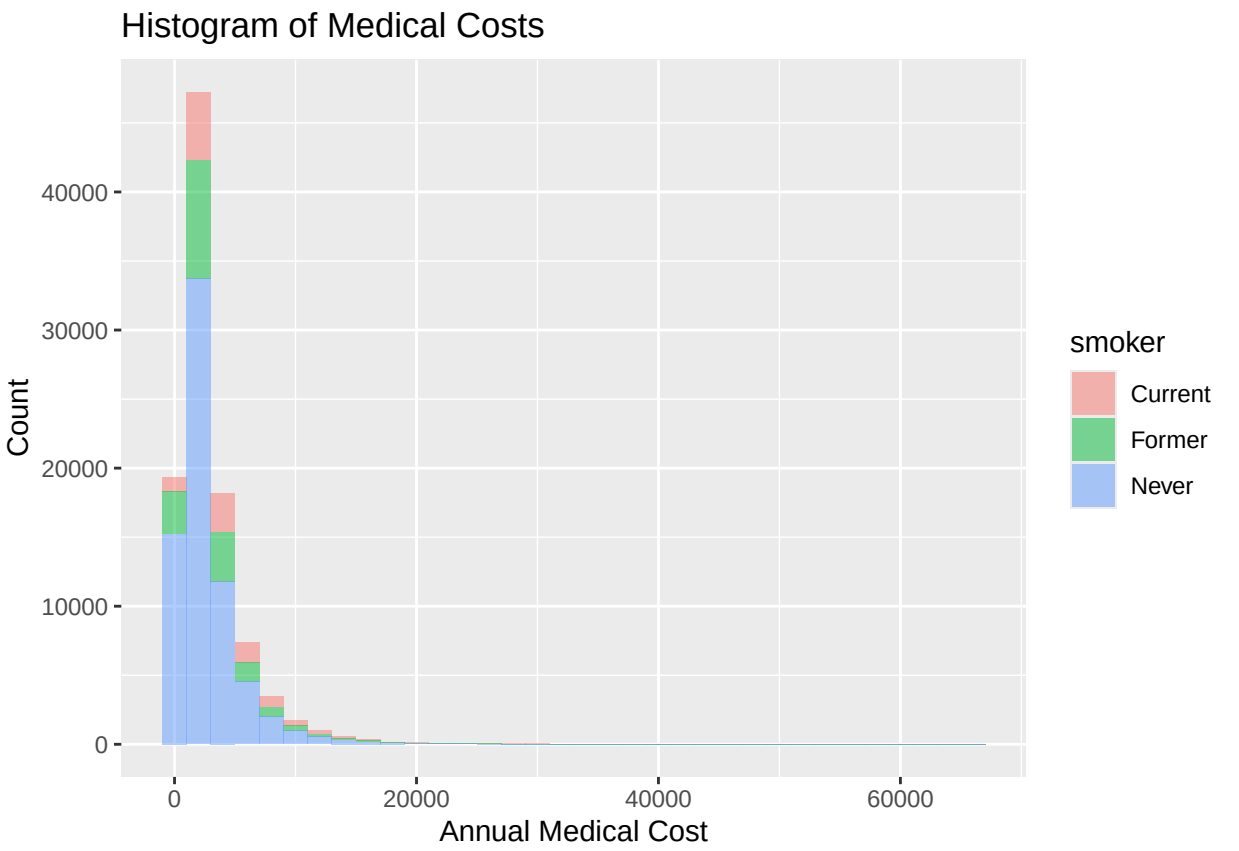

```
# Visualization 3
ggplot(insurance, aes(x = age, y = annual_medical_cost, color = smoker)) +
  geom_point() +
  labs(
    title = "Medical Cost vs Age",
    x = "Age",
    y = "Annual Medical Cost"
  )
```




```r
# Visualization 3.1
ggplot(insurance, aes(x = age, y = annual_medical_cost, color = smoker)) +
  geom_point() +
  facet_wrap(~ smoker) +
  labs(
    title = "Medical Cost vs Age (by Smoking Status)",
    x = "Age",
    y = "Annual Medical Cost"
  )
```




```r
# Visualization 4
ggplot(insurance, aes(x = annual_medical_cost, fill = smoker)) +
  geom_histogram(binwidth = 2000, alpha = 0.5) +
  labs(
    title = "Histogram of Medical Costs",
    x = "Annual Medical Cost",
    y = "Count"
  )
```



Results and Interpretation Across all four different visualizations I made, the same strong pattern appears - we can see how smokers consistently have higher medical insurance costs than former smokers and non-smokers. In the boxplot by region, this ranking repeats the same way in every location, which means that smoking habit has much more influence than geographical region. This supports He(2024)(link), who identified smoking as the most important predictor of medical costs in the United States.

The bar chart makes this pattern even clearer. For every region (Central, East, North, West), current smokers have the highest mean cost, never smokers have the lowest, and former smokers always sit in the middle. This "moddle" position for former smokers matches the key findings of Fishman et al.(2003)(link), who showed that former smokers initially have higher medical costs after quitting, but they eventually drop to levels closer to non-smokers.

I decided to look at even more interactions between different factors and made a scatterplot of age versus annual cost, which suggests to us that medical cost goes up with age for everyone, but current smokers reach higher cost levels more frequently anyway. However, the most interesting observation is how former smokers are in the middle, showing an important transition between current and never smokers (which we can also observe when pivoting the data to a wide format table). This matches another important article by Izumi et al.(2001)(link), who found that smokers use more medical care overall, while former smokers tend to use less and less.

Finally, the histogram shows that current smokers are at the top, while never smokers cluster heavily at the bottom. Former smokers again fall in between. This pattern matches findings of Bareddgert et al.(1997)(link), suggesting that smokers can have up to 40% higher medical costs at a given age due to smoking-related illnesses.

Overall, all four graphs - and relevant articles - point to the same conclusion: smoking habits have a much larger impact on annual medical costs than region, and former smokers consistently show a "transactional" cost pattern between current and never smokers, exactly as reported in multiple studies.

Bonus: The faceted scatterplot, that Professor suggested, further clarifies the relationship by separating each smoking group into three different plots, making it easier to observe how smokers consistently experience higher costs across all ages.

Ethical Considerations I think when working with any type of medical data. It is important to think about privacy and fairness. Even though my data set is anonymous, it is still important to acknowledge the sensitivity of information that could potentially harm people if shared without any protection. Second, it is important to avoid stigma: this analysis is made not to punish or judge smokers, but to see some consequences of bad habits, stress, and again make sure to address any other possible factors that play an important role in medical insurance costs. I think it is, again, very important when raising questions about inequalities and unequal access to healthcare.

AI Use Statement I didn't use AI for this project at all. I tried to write answers for each question myself without paraphrasing it academically; however, I used the Google Grammarly extension to check my punctuation to make it easier for the reader. I used the library.fandm website and available databases to find articles. I used Google Search to look up overlap prevention in the second graph (position "dodge"). For the references part, I used this free APA generator: link

For presentation, I used powerpoint_presentation in the header, which helped me to generate pptx file easily.

References

- Mohan Krishna Thalla. (2025). *Medical Insurance Cost Prediction*. Kaggle.com. https://www.kaggle.com/datasets/mohankrishnathalla/medical-insurance-cost-prediction?resource=download
- Izumi, Y., Tsuji, I., Ohkubo, T., Kuwahara, A., Nishino, Y., & Hisamichi, S. (2001). Impact of smoking habit on medical care use and its costs: a prospective observation of National Health Insurance beneficiaries in Japan. *International Journal of Epidemiology*, *30*(3), 616–621. https://doi.org/10.1093/ije/30.3.616
- Fishman, P. A., Khan, Z. M., Thompson, E. E., & Curry, S. J. (2003). Health Care Costs among Smokers, Former Smokers, and Never Smokers in an HMO. *Health Services Research*, *38*(2), 733–749. https://doi.org/10.1111/1475-6773.00142
- He, Q. (2024). Research on Factors Influencing Medical Insurance Cost in the US. *Transactions on Economics, Business and Management Research*, *10*, 31–36. https://doi.org/10.62051/yj92kn72
- Barendregt, J. J., Bonneux, L., & van der Maas, P. J. (1997). The Health Care Costs of Smoking. *New England Journal of Medicine*, *337*(15), 1052–1057. https://doi.org/10.1056/nejm199710093371506