QM563 - Final Project

YOUR NAME HERE

12/14/2020

Please submit this .Rmd file and compiled (Knit) output (as .html or .pdf)

Scenario: Welcome to Care4All PCPs, the largest Primary Care Network in the state of California! We are glad to have someone with your analytical prowess on board. We are interested in ensuring that our patients get the highest quality care at the fairest price. To that end, we hired a consultant to perform an analysis of hospitals in California to help us understand 1) Which hospitals are the highest quality? 2) Which hospitals charge the most/least?

Based on our request, the consultant provided data and code for each of those questions. While this was helpful, we want to rewrite the code in a different language and explain it in detail (no comments or explanations were provided). Then, we would like to extend this work to learn about the relationship between health quality and cost for our patients. Therefore, we have laid out 3 tasks.

Your Tasks:

Task 1: Describe hospital quality ratings in California Using code written in R, 1a) Explain the code then 1b) Translate that code into Python, improving it as necessary to best answer the question

Task 2: Describe procedure prices at hospitals in California Using code written in Python, 2a) Explain the code, then 2b) Translate that code into R, improving it as necessary to best answer the question

Task 3: Combine Data and Create Visualization Use the data from the first two tasks to determine the relationship between price and quality.

Hints and Advice

- The most important thing is that you understand the function of the code and can write code in another language that gives the equivalent output - there is no single correct solution to these tasks
- If you are unsure about what a particular block of code (out of a larger chunk) does, run that bit in isolation and note the changes to the output.
- Don't forget to check the Code Companions and live class slides for explanations of functions or equivalencies between the two languages.

Task 1: Hospital quality ratings in the state of California

For this task, you are given a .csv from *data.medicare.gov/data/hospital-compare* to help answer the question. This dataset contains identifying information for the hospital (Provider ID, Hospital Name, Phone Number, etc.) as well as Medicare-determined quality metrics (Overall rating, national comparison, safety of care, etc.).

1a) Explain the code

Explain in as much detail as possible what the code is doing and how it arrives at the final output table. How does this address the task of describing quality ratings of hospitals in California? Filtering the data for california allows us to look at the top ratings for hospitals using the head function once the data is filtered.

Add comments in the code and a several sentence summary to complete this task.

```
# Filter the hospital type by acute care hospitals in the state of california by pulling the list of ho
hosp_names = hosp_info %>%
  filter(`Hospital Type` == "Acute Care Hospitals") %>%
  filter(State == "CA") %>%
  pull(`Hospital Name`)


# Filter for several columns as they are renamed for pulling a smaller data set, and then filter for ho
hosp_info_CA =
hosp_info %>%
  rename(Hospital = `Hospital Name`,
         Provider_ID = `Provider ID`,
         Safety = `Safety of care national comparison`,
         Effectiveness = `Effectiveness of care national comparison`
         ) %>%
  filter(Hospital %in% hosp_names, State == "CA") %>%
  mutate(Overall_Rating = as.numeric(`Hospital overall rating`)) %>%
  drop_na(Overall_Rating)
```

```
## Warning: There was 1 warning in ‘mutate()‘.
## i In argument: ‘Overall_Rating = as.numeric(‘Hospital overall rating‘)‘.
## Caused by warning:
## ! NAs introduced by coercion
```

```
# Use the head function to display the top 7 rows for hospitals in California based on their overall ra
hosp_info_CA %>%
  arrange(desc(Overall_Rating), Hospital) %>%
  head(7)
```

```
## # A tibble: 7 x 30
##   Provider_ID Hospital              Address City  State ‘ZIP Code‘ ‘County Name‘
##         <dbl> <chr>                 <chr>   <chr> <chr>      <dbl> <chr>
## 1       50145 COMMUNITY HOSPITAL O~ 23625 ~ MONT~ CA         93940 MONTEREY
## 2       50357 GOLETA VALLEY COTTAG~ 351 S ~ SANT~ CA         93111 SANTA BARBARA
## 3       50238 METHODIST HOSPITAL O~ 300 W ~ ARCA~ CA         91006 LOS ANGELES
## 4       50396 SANTA BARBARA COTTAG~ 400 WE~ SANT~ CA         93102 SANTA BARBARA
## 5       50424 SCRIPPS GREEN HOSPIT~ 10666 ~ LA J~ CA         92037 SAN DIEGO
## 6       50324 SCRIPPS MEMORIAL HOS~ 9888 G~ LA J~ CA         92037 SAN DIEGO
## 7       50281 ALHAMBRA HOSPITAL ME~ 100 S ~ ALHA~ CA         91801 LOS ANGELES
## # i 23 more variables: ‘Phone Number‘ <dbl>, ‘Hospital Type‘ <chr>,
## #   ‘Hospital Ownership‘ <chr>, ‘Emergency Services‘ <lgl>,
## #   ‘Meets criteria for meaningful use of EHRs‘ <lgl>,
## #   ‘Hospital overall rating‘ <chr>, ‘Hospital overall rating footnote‘ <chr>,
## #   ‘Mortality national comparison‘ <chr>,
## #   ‘Mortality national comparison footnote‘ <chr>, Safety <chr>,
## #   ‘Safety of care national comparison footnote‘ <chr>, ...
```

```r
# Group the data by Overall_Rating and Safety, then and use the count function of the number of occurre
hosp_info_CA %>%
  group_by(Overall_Rating, Safety) %>%
  count()
```

```
## # A tibble: 17 x 3
## # Groups:   Overall_Rating, Safety [17]
##    Overall_Rating Safety                           n
##             <dbl> <chr>                        <int>
##  1              1 Above the national average       1
##  2              1 Below the national average       7
##  3              1 Same as the national average     1
##  4              2 Above the national average      10
##  5              2 Below the national average      26
##  6              2 Not Available                    6
##  7              2 Same as the national average    46
##  8              3 Above the national average      36
##  9              3 Below the national average      16
## 10              3 Not Available                    4
## 11              3 Same as the national average    69
## 12              4 Above the national average      24
## 13              4 Below the national average       4
## 14              4 Not Available                    3
## 15              4 Same as the national average    17
## 16              5 Above the national average       5
## 17              5 Same as the national average     1
```

```r
# Write the filtered and processed hospital information for California to a CSV file
write_csv(hosp_info_CA, 'hosp_info_CA.csv')
```

1b) (Translation to Python, see .ipynb)

Task 2: Hospital Costs in the state of California

Motivating Question : **Which hospitals charge the most/least?**

For this task, you are given a .csv from *https://data.cms.gov/Medicare-Inpatient/Inpatient-Prospective-Payment-System-IPPS-Provider/97k6-zzx3* to help investigate hospital costs in California. The dataset contains identifying information for the hospital (Provider ID, Hospital Name, Address, Zip Code), the diagnosis-related group (DRG), and associated costs (Average Total Payments, Average Medicare Payments)

Average Total Payments: The average of Medicare payments to the provider for the DRG including the DRG amount, teaching, disproportionate share, capital, and outlier payments for all cases. Also included are co-payment and deductible amounts that the patient is responsible for.

2a) (Code Explanation, see.ipynb

2b) Translate the Python Code to R

Translate the provided code from Python to R, improving it if neccessary to best address the question: **Which hospitals cost the most/least?**. San Francisco General hosptial costs the most and Arroyo Grande Community Hospital costs the least.

Provide your insights from the output of the code.

```r
# Load required libraries
library(dplyr)

# Read the data from the CSV file into a data frame
costs <- read.csv("Inpatient_Prospective_Payment_System__IPPS__Provider_Summary_for_the_Top_100_Diagnosi


# Rename columns for better readability
costs <- costs %>%
  rename(
    DRG  = `DRG.Definition`,
    Total_Cost = `Average.Total.Payments`,  # Note spaces around ' Average Total Payments '
    Count_Discharges = `Total.Discharges`   # Note spaces around ' Total Discharges '
  )


# Split the DRG column into two separate columns, DRG_Code and DRG_Description
costs <- costs %>%
  separate(DRG, into = c("DRG_Code", "DRG_Description"), sep = " - ", remove = TRUE)


# Calculate the average DRG cost for each DRG_Code
costs <- costs %>%
  group_by(DRG_Code) %>%
  mutate(Avg_DRG_Cost = mean(Total_Cost, na.rm = TRUE))

# Calculate the cost difference between the Total_Cost and the average DRG cost for each DRG_Code
costs <- costs %>%
  mutate(Cost_Diff = Total_Cost - Avg_DRG_Cost)

# Group the 'costs' data frame by 'DRG_Code', summing the 'Count_Discharges' for each group,
# and then arrange the result in descending order based on the sum of 'Count_Discharges'
top_drgs <- costs %>%
  group_by(DRG_Code) %>%
  summarize(Count_Discharges = sum(Count_Discharges)) %>%
  arrange(desc(Count_Discharges))

# Retrieve the first element (DRG_Code) from the sorted 'top_drgs' data frame
top_drgs$DRG_Code[1]
```

```
## [1] "470"
```

```r
# Filter the 'costs' data frame using a query to select rows where 'DRG_Code' is '470' and 'Provider St
# and then write the filtered data to a CSV file named "Hip_Replacement_Costs_by_Hosp.csv" without incl
filtered_costs <- costs %>%
  filter(DRG_Code == '470' & `Provider.State` == 'CA')

write.csv(filtered_costs, "Hip_Replacement_Costs_by_Hosp.csv", row.names = FALSE)

hipcosts = read_csv("Hip_Replacement_Costs_by_Hosp.csv")
```

```
## Rows: 235 Columns: 15
## -- Column specification --------------------------------------------------------
```

```
## Delimiter: ","
## chr (6): DRG_Description, Provider.Name, Provider.Street.Address, Provider.C...
## dbl (9): DRG_Code, Provider.Id, Provider.Zip.Code, Count_Discharges, Average...
##
## 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.
```

```r
# Read the data
hip_costs <- read.csv("Hip_Replacement_Costs_by_Hosp.csv")

# Sort hospitals by Total_Cost in ascending order
sorted_costs <- hip_costs %>%
  arrange(Total_Cost)

# Display hospitals with the highest total cost
cat("Hospitals with the highest total cost:\n")
```

```
## Hospitals with the highest total cost:
```

```r
print(sorted_costs %>% tail(1))
```

```
##     DRG_Code                                                  DRG_Description
## 235      470 MAJOR JOINT REPLACEMENT OR REATTACHMENT OF LOWER EXTREMITY W/O MCC
##     Provider.Id                  Provider.Name Provider.Street.Address
## 235       50228 SAN FRANCISCO GENERAL HOSPITAL     1001 POTRERO AVENUE
##     Provider.City Provider.State Provider.Zip.Code
## 235 SAN FRANCISCO             CA             94110
##     Hospital.Referral.Region.Description Count_Discharges
## 235                   CA - San Francisco               22
##     Average.Covered.Charges Total_Cost Average.Medicare.Payments Avg_DRG_Cost
## 235                124488.6   33777.27                  32366.63     14566.93
##     Cost_Diff
## 235  19210.34
```

```r
# Display hospitals with the lowest total cost
cat("\nHospitals with the lowest total cost:\n")
```

```
##
## Hospitals with the lowest total cost:
```

```r
print(sorted_costs %>% head(1))
```

```
##   DRG_Code                                                  DRG_Description
## 1      470 MAJOR JOINT REPLACEMENT OR REATTACHMENT OF LOWER EXTREMITY W/O MCC
##   Provider.Id                    Provider.Name Provider.Street.Address
## 1       50016 ARROYO GRANDE COMMUNITY HOSPITAL        345 S HALCYON RD
##   Provider.City Provider.State Provider.Zip.Code
## 1 ARROYO GRANDE             CA             93420
##   Hospital.Referral.Region.Description Count_Discharges Average.Covered.Charges
## 1                 CA - San Luis Obispo               168                77875.84
##   Total_Cost Average.Medicare.Payments Avg_DRG_Cost Cost_Diff
## 1   12802.02                  11703.07     14566.93 -1764.909
```

Task 3: What is the relationship between cost and quality?

Is it the case that "you get what you pay for"? Now that we have completed some preliminary analyses of the cost and quality of Hospitals in the state of California, we would like to take a look at their relationship jointly. That is, we would like to see how cost relates to quality by combining the output from the first two questions.

With the language of your choosing (either R or Python),

3a) Join/Merge together Cost and Quality tables

Join together the resulting tables from tasks 1 and 2. What type of join did you perform and why? How many hospitals were removed (if any) due to the type of join? I performed an inner join to have an exact match between the columns provided by the unique key.

```
library(dplyr)

# Rename columns using rename() before joining
#data_frame1 <- data_frame1 %>% rename(new_column_name = old_column_name)
hosp_info_CA <- hosp_info_CA %>% rename(provider.id = Provider_ID)
costs <- costs %>% rename(provider.id = Provider.Id)

# Inner Join
merged_data <- inner_join(hosp_info_CA, costs, by = "provider.id")

# Other types of joins (left_join, right_join, full_join) can also be used similarly


# Inner Join
merged_data <- inner_join(hosp_info_CA, costs, by = "provider.id")


# Print the merged data to check the result
print(merged_data)
```

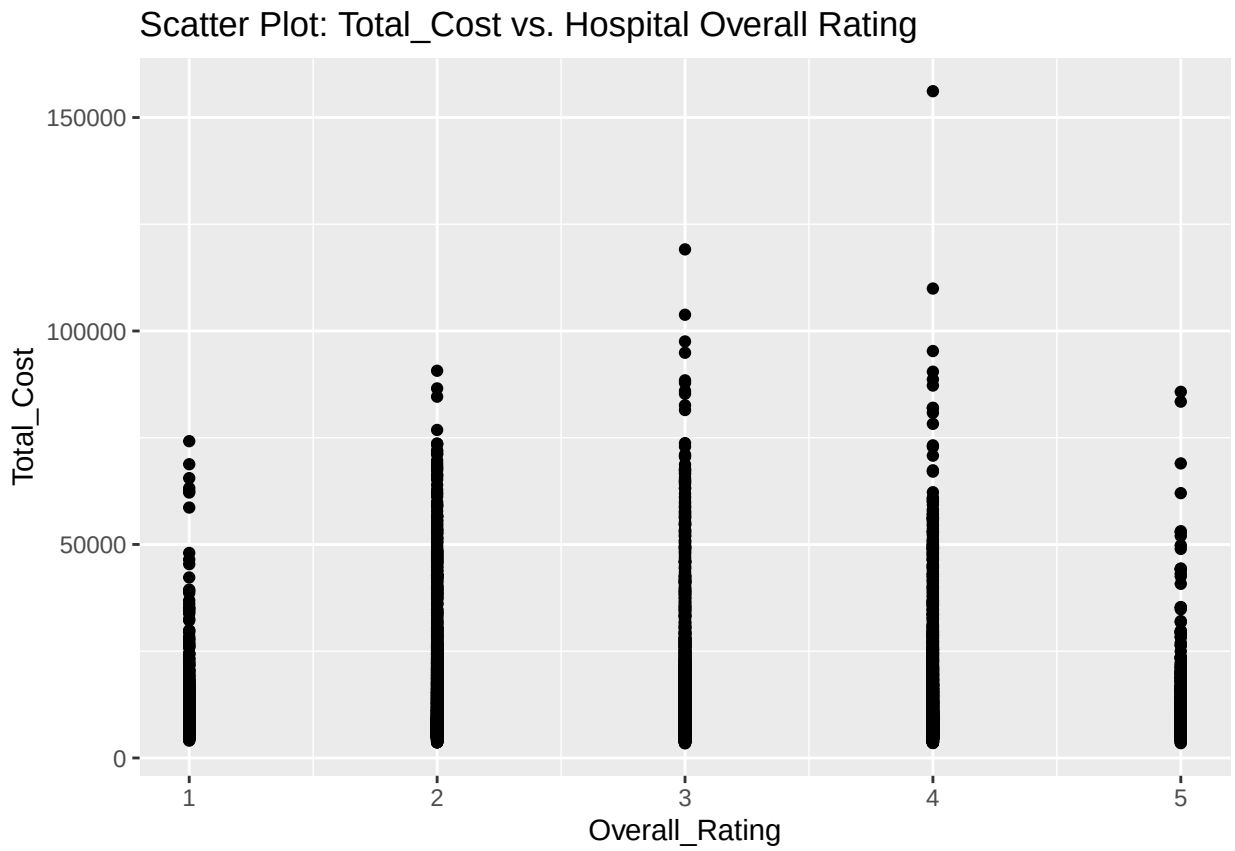
```
## # A tibble: 12,704 x 44
##    provider.id Hospital             Address City  State `ZIP Code` `County Name`
##          <dbl> <chr>                <chr>   <chr> <chr>      <dbl> <chr>
##  1       50054 SAN GORGONIO MEMORI~ 600 NO~ BANN~ CA         92220 RIVERSIDE
##  2       50054 SAN GORGONIO MEMORI~ 600 NO~ BANN~ CA         92220 RIVERSIDE
##  3       50054 SAN GORGONIO MEMORI~ 600 NO~ BANN~ CA         92220 RIVERSIDE
##  4       50054 SAN GORGONIO MEMORI~ 600 NO~ BANN~ CA         92220 RIVERSIDE
##  5       50054 SAN GORGONIO MEMORI~ 600 NO~ BANN~ CA         92220 RIVERSIDE
##  6       50054 SAN GORGONIO MEMORI~ 600 NO~ BANN~ CA         92220 RIVERSIDE
##  7       50054 SAN GORGONIO MEMORI~ 600 NO~ BANN~ CA         92220 RIVERSIDE
##  8       50054 SAN GORGONIO MEMORI~ 600 NO~ BANN~ CA         92220 RIVERSIDE
##  9       50054 SAN GORGONIO MEMORI~ 600 NO~ BANN~ CA         92220 RIVERSIDE
## 10       50054 SAN GORGONIO MEMORI~ 600 NO~ BANN~ CA         92220 RIVERSIDE
## # i 12,694 more rows
## # i 37 more variables: `Phone Number` <dbl>, `Hospital Type` <chr>,
## #   `Hospital Ownership` <chr>, `Emergency Services` <lgl>,
## #   `Meets criteria for meaningful use of EHRs` <lgl>,
## #   `Hospital overall rating` <chr>, `Hospital overall rating footnote` <chr>,
```

```
## #   `Mortality national comparison` <chr>,
## #   `Mortality national comparison footnote` <chr>, Safety <chr>, ...
```

3b) Create a Visualization

Using the insights you gained from the sections above, create a visualization to address the question. Provide a detailed explanation of the insights gained from your visualization.

```
# Example for Total Cost vs. Hospital Overall Rating
  ggplot(merged_data, aes(x = Overall_Rating, y = Total_Cost)) +
  geom_point() +
  labs(title = "Scatter Plot: Total_Cost vs. Hospital Overall Rating")
```



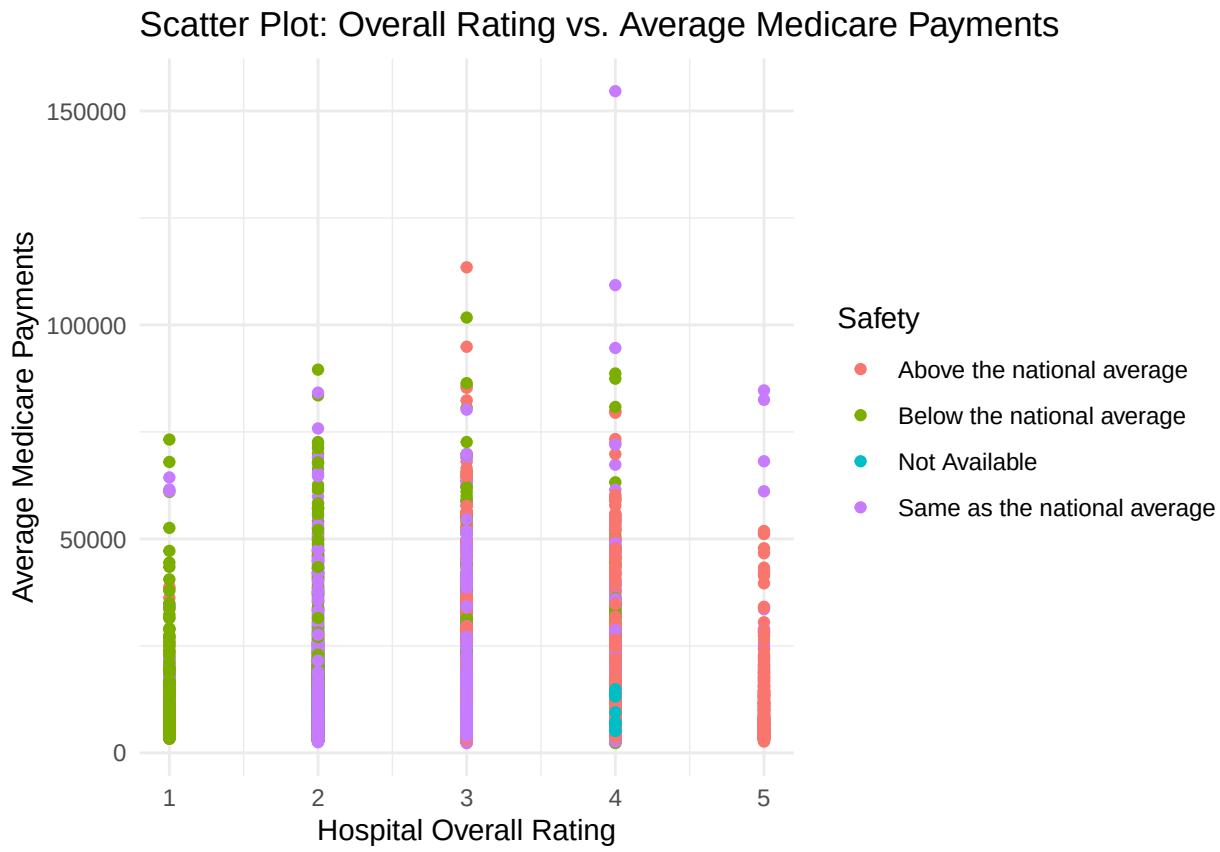
Insight: It seems that there is no correlation between hospital total costs and overall rating for hospitals. This means that expensive hospitals don't necessarily provide better care via the overall rating.

3c) Extend the insights from above

With the code and data you used in the previous tasks as a base, provide additional insights that augment those from the previous task (3b).

```
# Scatter plot: Average Medicare Payments vs. Overall Rating
ggplot(merged_data, aes(x= Overall_Rating, y= `Average.Medicare.Payments`, color=Safety)) +
  geom_point() +
  labs(title='Scatter Plot: Overall Rating vs. Average Medicare Payments',
```

```
       x='Hospital Overall Rating',
       y='Average Medicare Payments') +
  theme_minimal()
```



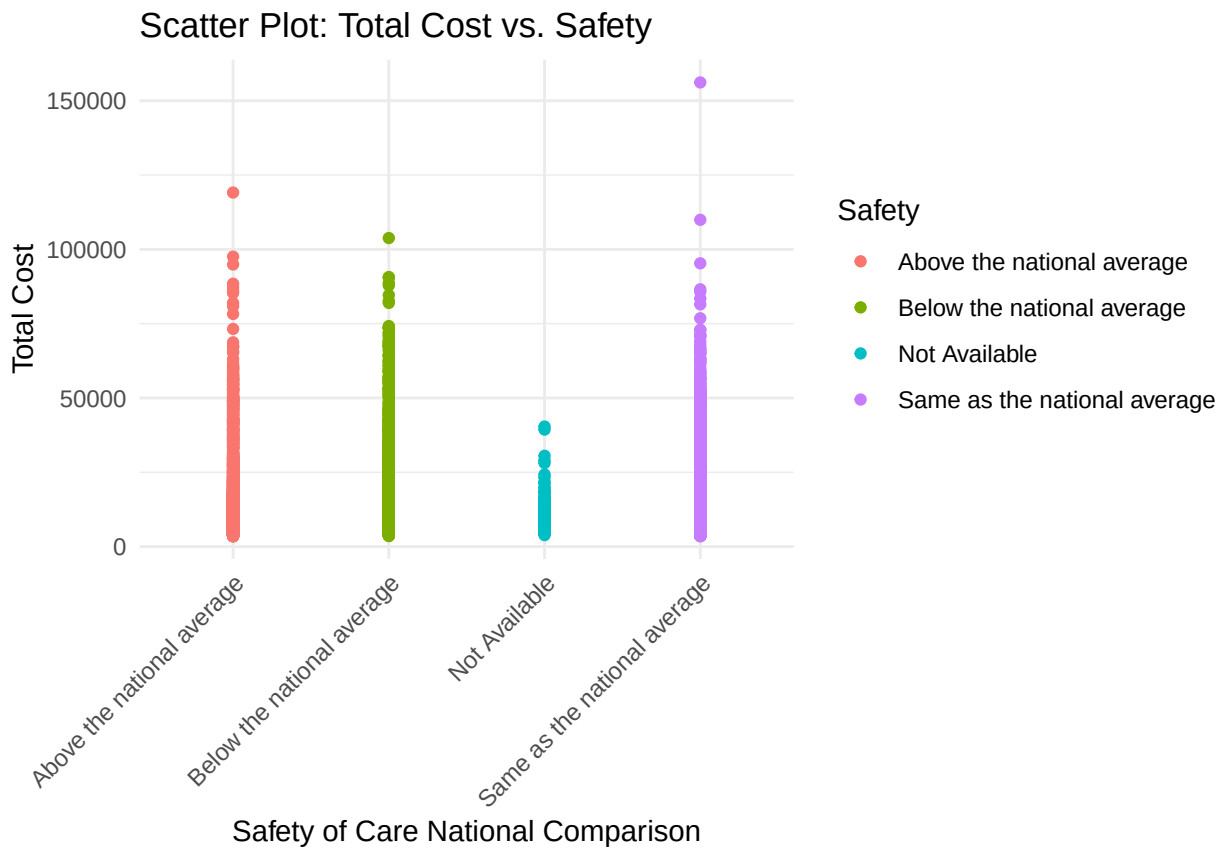
Question: Does the average Medicare payment have an impact on the hospital's overall rating?

Hypothesis: We hypothesize that hospitals receiving higher average Medicare payments might demonstrate a positive correlation with their overall ratings.

Insights: Correlation Patterns: The plot suggests a general trend where hospitals with higher overall ratings tend to receive higher average Medicare payments. This initial observation aligns with the hypothesis, indicating a potential positive correlation.

Potential Areas for Improvement/Implications for Policy: If a strong positive correlation is observed, it may have implications for healthcare policy. Hospitals receiving higher Medicare payments may be incentivized to maintain or improve overall quality, impacting reimbursement strategies.

```
# Scatter plot: Total Cost vs. Safety
ggplot(merged_data, aes(x= Safety, y= Total_Cost, color=Safety)) +
  geom_point() +
  labs(title='Scatter Plot: Total Cost vs. Safety',
       x='Safety of Care National Comparison',
       y='Total Cost') +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Spread out x-axis titling
```



Question: Does the safety of care rating impact the total cost incurred by hospitals?

Hypothesis: We hypothesize that hospitals with higher safety ratings might demonstrate lower total costs, suggesting a potential correlation between safety measures and cost-effectiveness.

Description of New Insights:

The visualization aids in assessing whether the observed pattern aligns with the hypothesis. Hospitals with better safety ratings consistently demonstrate lower total costs, so it supports the notion that investing in safety measures could contribute to cost-effectiveness.

Potential Areas for Improvement: Hospitals with high safety ratings and unexpectedly high total costs could signal areas for improvement in cost management strategies, even within the context of robust safety measures.

For example, you could consider: - Visualizing more variables from the datasets in tasks 1 and 2 - Leveraging the whole dataset to understand where the California providers stand nationally - Gather outside data (e.g. Census) and join with the data in this task (e.g. using Zip Code) - Create an interactive plot (plotly or ggplotly) to help explore an expanded dataset

Be sure to structure this response as:

```
1) Question or hypothesis
2) Code/Data Formatting and Plotting
3) Description of new insights
```