Faculty of Science and Technology

Department of Artificial Intelligence

S.Y.B.Tech(AI&ML)

BTECML23301: Applied Statistical Analysis

Project-1

Name: Haidery Inzamam ul Haque Anwar

Srn no.: 31230475

Roll no: 8

Div: C- AI&ML

Faculty: Prof.Sukhpreet Kaur

Problem Statement:

Customer Satisfaction Dataset: Analyze the relationship between service quality and customer satisfaction using regression. Visualize the relationship using scatter plots and create standardized scores.

Source code:

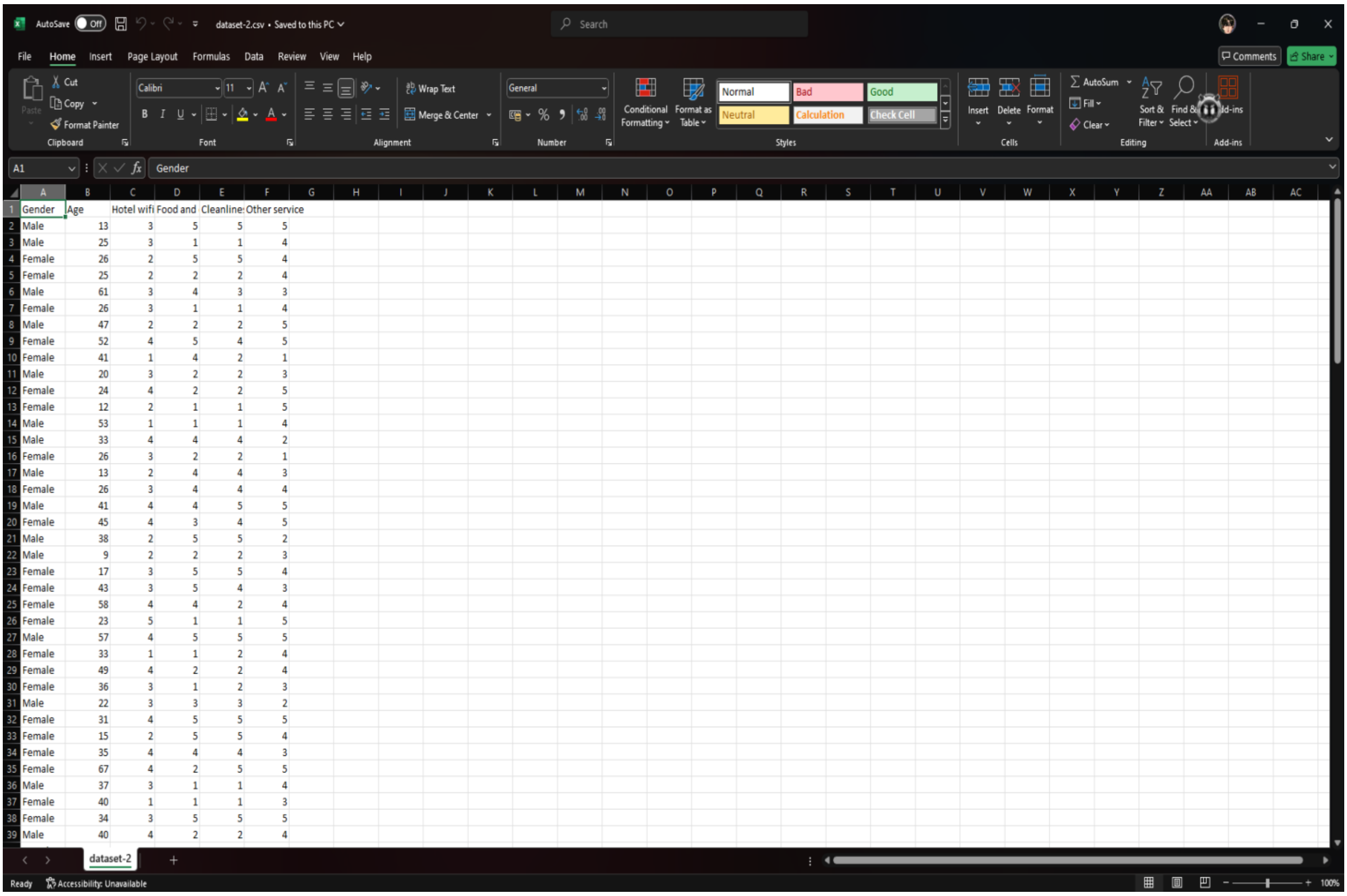
```
 # Load necessary libraries
library(readr)
library(dplyr)
library(ggplot2)
library(tidyr)
library(corrplot)

# 1. Data Loading and Preprocessing
# Load the dataset
dataset <- read_csv("D:/dataset-2.csv")


# 2. Handle Missing Values
# Replace missing numeric values with mean
dataset <- dataset %>%
  mutate(
    Age = ifelse(is.na(Age), mean(Age, na.rm = TRUE), Age),
    wifi_service = ifelse(is.na(wifi_service), mean(wifi_service, na.rm = TRUE), wifi_service),
    Food_drink = ifelse(is.na(Food_drink), mean(Food_drink, na.rm = TRUE), Food_drink),
    Cleanliness = ifelse(is.na(Cleanliness), mean(Cleanliness, na.rm = TRUE), Cleanliness),
    Other_service = ifelse(is.na(Other_service), mean(Other_service, na.rm = TRUE), Other_service)
  )


# 3. Standardize the data using Z-scores
dataset <- dataset %>%
  mutate(
    Z_Age = scale(Age),
    Z_Wifi = scale(wifi_service),
    Z_Food = scale(Food_drink),
    Z_Cleanliness = scale(Cleanliness),
    Z_Other = scale(Other_service)
  )
```

```
# 4. Basic Statistics (Mean, Median, Standard Deviation)
basic_stats <- dataset %>%
  summarise(
    Mean_Age = mean(Age, na.rm = TRUE),
    Median_Age = median(Age, na.rm = TRUE),
    SD_Age = sd(Age, na.rm = TRUE),

    Mean_Wifi = mean(wifi_service, na.rm = TRUE),
    Median_Wifi = median(wifi_service, na.rm = TRUE),
    SD_Wifi = sd(wifi_service, na.rm = TRUE),

    Mean_Food = mean(Food_drink, na.rm = TRUE),
    Median_Food = median(Food_drink, na.rm = TRUE),
    SD_Food = sd(Food_drink, na.rm = TRUE),

    Mean_Cleanliness = mean(Cleanliness, na.rm = TRUE),
    Median_Cleanliness = median(Cleanliness, na.rm = TRUE),
    SD_Cleanliness = sd(Cleanliness, na.rm = TRUE),

    Mean_Other = mean(Other_service, na.rm = TRUE),
    Median_Other = median(Other_service, na.rm = TRUE),
    SD_Other = sd(Other_service, na.rm = TRUE)
  )


print("Basic Statistics:")
print(basic_stats)

# Print all data values
print("All Data Values:")
print(dataset, n = Inf)  # Print all rows
```

```
# 5. Identify outliers using Z-scores
outliers <- dataset %>%
  filter(abs(Z_Age) > 3 | abs(Z_Wifi) > 3 | abs(Z_Food) > 3 | abs(Z_Cleanliness) > 3 | abs(Z_Other) > 3)

print("Outliers detected:")
if (nrow(outliers) > 0) {
  print(outliers)
} else {
  print("No outliers detected.")
}

# 6. Frequency distribution and Histograms for all numeric variables
# Create histograms for numeric columns
for (col in c("Age", "wifi_service", "Food_drink", "Cleanliness", "Other_service")) {
  p <- ggplot(dataset, aes_string(x = col)) +
    geom_histogram(binwidth = 1, fill = "blue", color = "black", alpha = 0.7) +
    labs(title = paste("Histogram of", col), x = col, y = "Frequency") +
    theme_minimal()

  print(p)
}

# 7. Calculate Satisfaction Score
dataset <- dataset %>%
  mutate(Satisfaction_Score = (wifi_service + Food_drink + Cleanliness + Other_service) / 4)

# 8. Scatter Plot to visualize the relationship between Satisfaction and Service Ratings
ggplot(dataset, aes(x = Age, y = Satisfaction_Score)) +
  geom_point() +
  labs(title = "Scatter Plot: Age vs Satisfaction Score", x = "Age", y = "Satisfaction Score")

ggplot(dataset, aes(x = wifi_service, y = Satisfaction_Score)) +
  geom_point() +
  geom_smooth(method = "lm", color = "red") +
  labs(title = "Scatter Plot: wifi_service vs Satisfaction Score", x = "wifi_service", y = "Satisfaction Score")

# 9. Box plots for service quality variables
melted_dataset <- dataset %>%
  select(wifi_service, Food_drink, Cleanliness, Other_service) %>%
  tidyr::gather(key = "Service", value = "Score")

ggplot(melted_dataset, aes(x = Service, y = Score)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  labs(title = "Box Plot: Service Quality Ratings", x = "Service Type", y = "Score")

# 10. Correlation matrix for understanding relationships
correlation_matrix <- dataset %>%
  select(Satisfaction_Score, wifi_service, Food_drink, Cleanliness, Other_service) %>%
  cor()

print("Correlation Matrix:")
print(correlation_matrix)

# Visualizing the correlation matrix as a circle
corrplot(correlation_matrix, method = "circle", type = "upper",
         tl.col = "black", tl.srt = 45,
         title = "Correlation Matrix (Circle Method)", mar = c(0,0,1,0))
```

```r
# 11. Analyze the relationship between service quality and customer satisfaction using regression
model <- lm(Satisfaction_Score ~ wifi_service + Food_drink + Cleanliness + Other_service, data = dataset)

# Model summary
model_summary <- summary(model)
print("Regression Model Summary:")
print(model_summary)

# 12. Display the results of regression in a scatter plot with fitted line
ggplot(dataset, aes(x = Satisfaction_Score, y = wifi_service)) +
  geom_point() +
  geom_smooth(method = "lm", color = "red") +
  labs(title = "Regression: Satisfaction Score vs wifi_service", x = "Satisfaction Score", y = "wifi_service")

# 13. Show a table summarizing customer satisfaction with service quality
satisfaction_table <- dataset %>%
  select(Satisfaction_Score, wifi_service, Food_drink, Cleanliness, Other_service)

print("Customer Satisfaction and Service Quality Table:")
print(satisfaction_table)
```

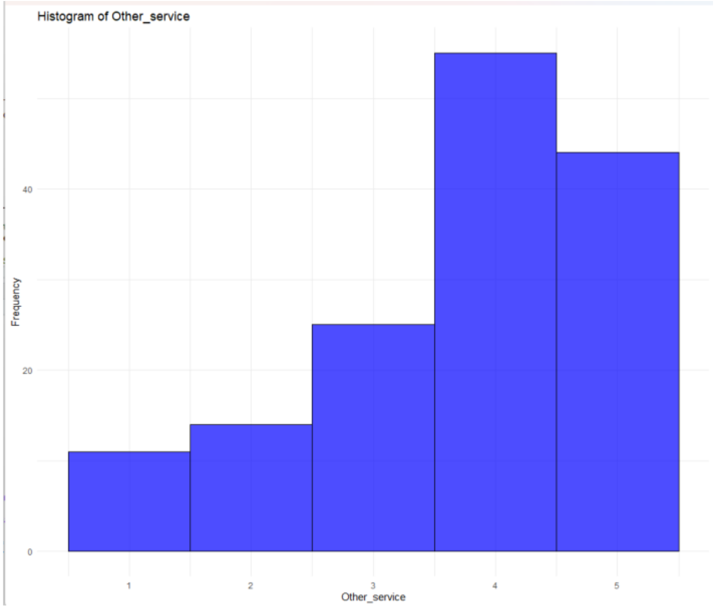
Dataset:



dataset-2.csv • Saved to this PC

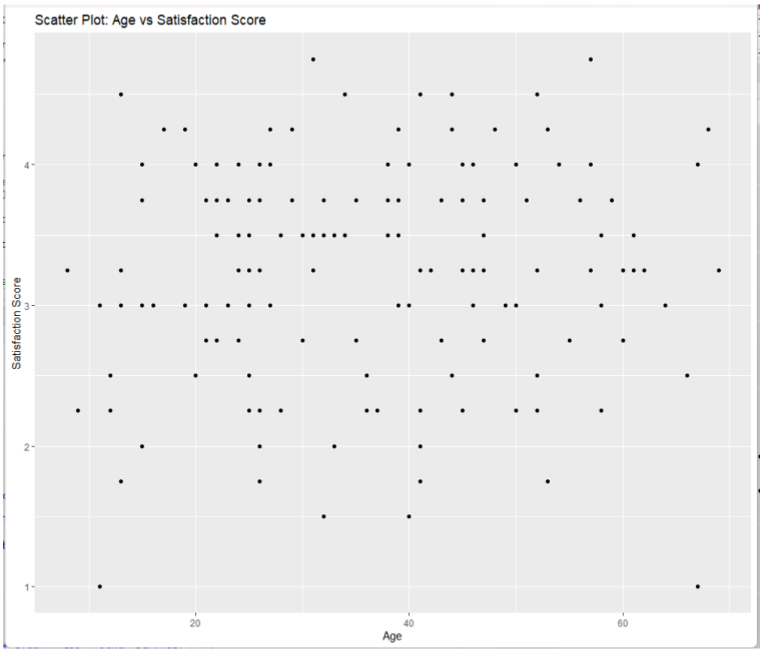
Gender	Age	Hotel wifi	Food and	Cleanline:	Other service
Male	13	3	5	5	5
Male	25	3	1	1	4
Female	26	2	5	5	4
Female	25	2	2	2	4
Male	61	3	4	3	3
Female	26	3	1	1	4
Male	47	2	2	2	5
Female	52	4	5	4	5
Female	41	1	4	2	1
Male	20	3	2	2	3
Female	24	4	2	2	5
Female	12	2	1	1	5
Male	53	1	1	1	4
Male	33	4	4	4	2
Female	26	3	2	2	1
Male	13	2	4	4	3
Female	26	3	4	4	4
Male	41	4	4	5	5
Female	45	4	3	4	5
Male	38	2	5	5	2
Male	9	2	2	2	3
Female	17	3	5	5	4
Female	43	3	5	4	3
Female	58	4	4	2	4
Female	23	5	1	1	5
Male	57	4	5	5	5
Female	33	1	1	2	4
Female	49	4	2	2	4
Female	36	3	1	2	3
Male	22	3	3	3	2
Female	31	4	5	5	5
Female	15	2	5	5	4
Female	35	4	4	4	3
Female	67	4	2	5	5
Male	37	3	1	1	4
Female	40	1	1	1	3
Female	34	3	5	5	5
Male	40	4	2	2	4

Output:



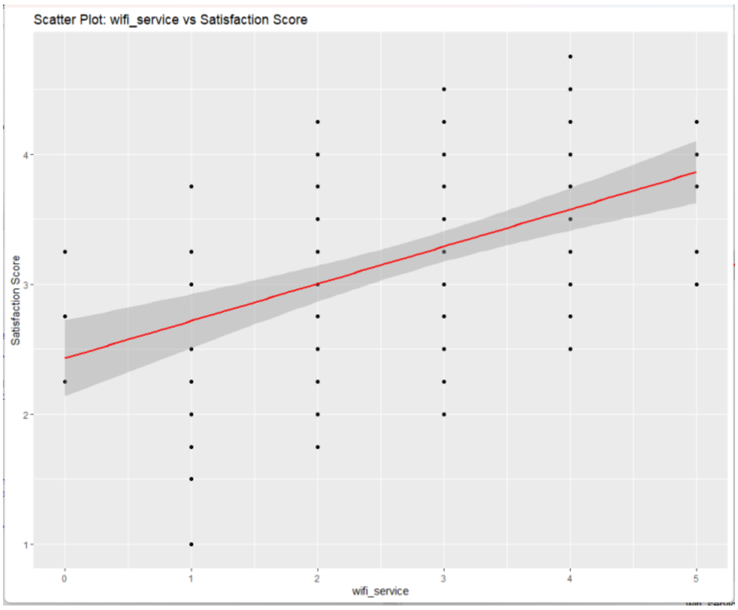
Histogram of Other_service:

- This histogram illustrates the frequency distribution of the "Other_service" variable, which represents the service quality score in a customer satisfaction context.
- The majority of ratings fall between 3 and 5, with higher counts in the 4-5 range, indicating that most customers rated this service favorably.
- The shape suggests a skew towards higher service quality ratings.



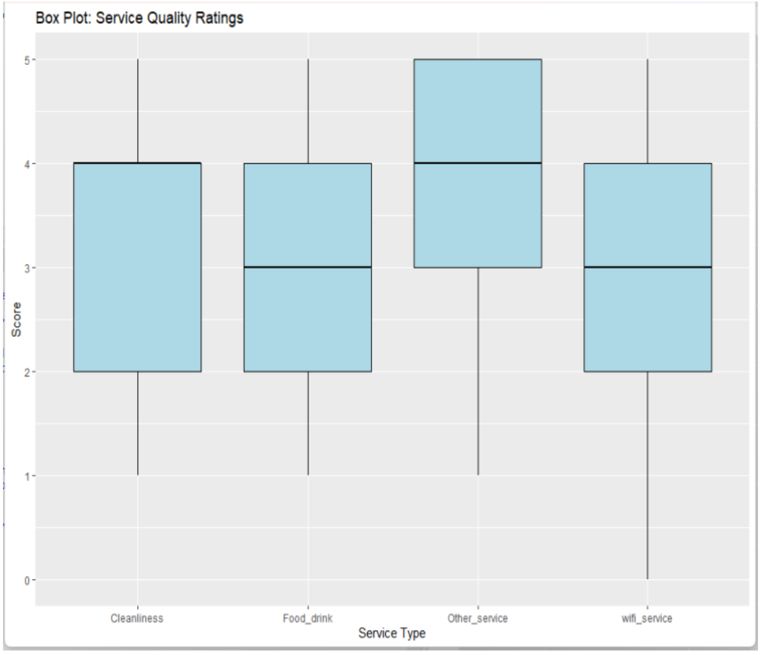
Scatter Plot: Age vs Satisfaction Score:

- This scatter plot depicts the relationship between customer age and satisfaction scores.
- There is no clear linear trend, as the points are widely scattered, suggesting a weak or no direct correlation between age and overall customer satisfaction.
- Different age groups show similar satisfaction levels, ranging from 1 to 5, without concentration in any specific area.



Scatter Plot: WiFi Service vs Satisfaction Score:

- This scatter plot shows the relationship between WiFi service ratings and satisfaction.
- The positive trend suggests that higher WiFi service ratings are associated with higher satisfaction scores.
- There is a relatively consistent pattern in the data, though with some variability.
- The red regression line indicates the predicted relationship between the two variables.
- Some scatter around the line reflects differing levels of satisfaction despite similar WiFi service ratings.

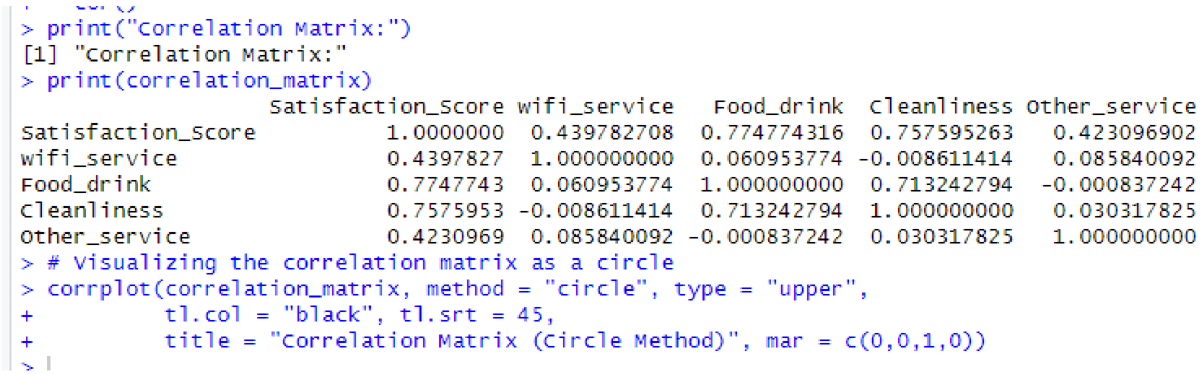


Box Plot: Service Quality Ratings:

- The box plot compares the distribution of ratings across four different service quality types: cleanliness, food and drink, other service, and Wi-Fi service.

- Cleanliness appears to have the highest median score, while food and drink show more variability in ratings.

- The Wi-Fi service shows a wider range, with lower ratings indicated by its lower median and more extreme outliers, suggesting variability in service quality perception across categories.



```
> print("Correlation Matrix:")
[1] "Correlation Matrix:"
> print(correlation_matrix)
                   Satisfaction_Score wifi_service   Food_drink  Cleanliness Other_service
Satisfaction_Score          1.0000000  0.439782708  0.774774316  0.757595263   0.423096902
wifi_service                0.4397827  1.000000000  0.060953774 -0.008611414   0.085840092
Food_drink                  0.7747743  0.060953774  1.000000000  0.713242794  -0.000837242
Cleanliness                 0.7575953 -0.008611414  0.713242794  1.000000000   0.030317825
Other_service               0.4230969  0.085840092 -0.000837242  0.030317825   1.000000000
> # Visualizing the correlation matrix as a circle
> corrplot(correlation_matrix, method = "circle", type = "upper",
+          tl.col = "black", tl.srt = 45,
+          title = "Correlation Matrix (Circle Method)", mar = c(0,0,1,0))
> 
```

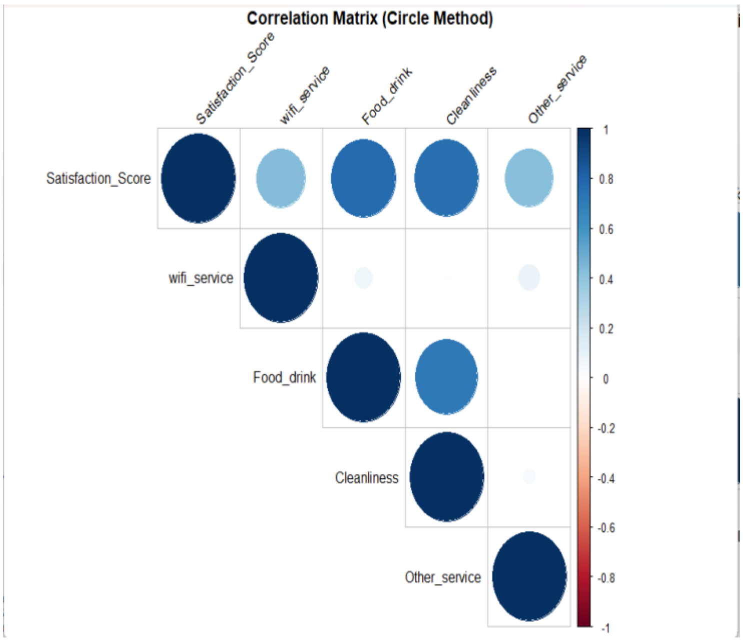
Correlation Matrix:

1. Statisfaction Score:
   - The satisfaction score has the highest correlation with food & drink (0.7747) and cleanliness (0.7759), indicating that improvements in these two areas will most likely increase overall satisfaction.
   - WiFi service has a moderate positive correlation with satisfaction (0.4397), implying that it also impacts customer satisfaction but less than food & drink and cleanliness.
2. WiFi Service: WiFi service shows a positive but relatively low correlation with food & drink (0.0609) and a weak, almost negligible negative correlation with cleanliness (-0.0086), suggesting that WiFi service ratings are mostly independent of these factors.

3. Food & Drink:
   - Food & drink has a strong positive correlation with cleanliness (0.7142), implying that customers who rate food & drink highly tend to also rate cleanliness highly.
   - It is less correlated with other services (0.0037), meaning other services don't affect the food & drink rating significantly.

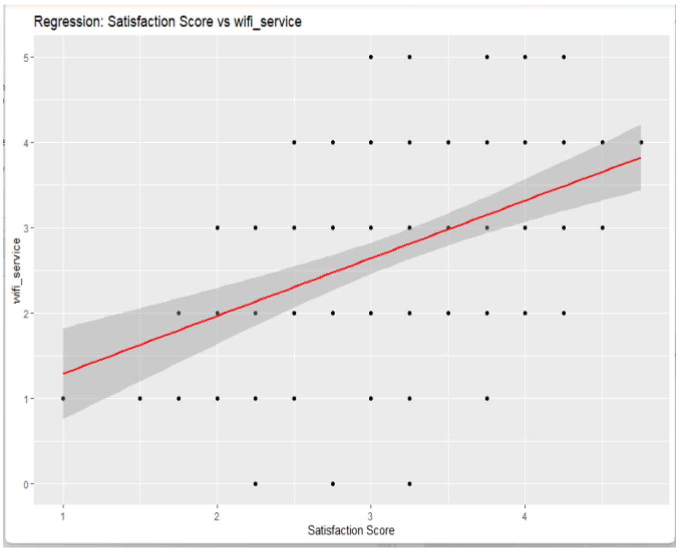
4. Cleanliness: Cleanliness shows a very weak correlation with WiFi service (-0.0086) and other services (0.0030), meaning cleanliness ratings are largely independent of these aspects.

5. Other Service: Other service shows a moderate correlation with satisfaction score (0.4231) but weak or near-zero correlations with other variables, meaning it's an independent factor but does influence satisfaction.



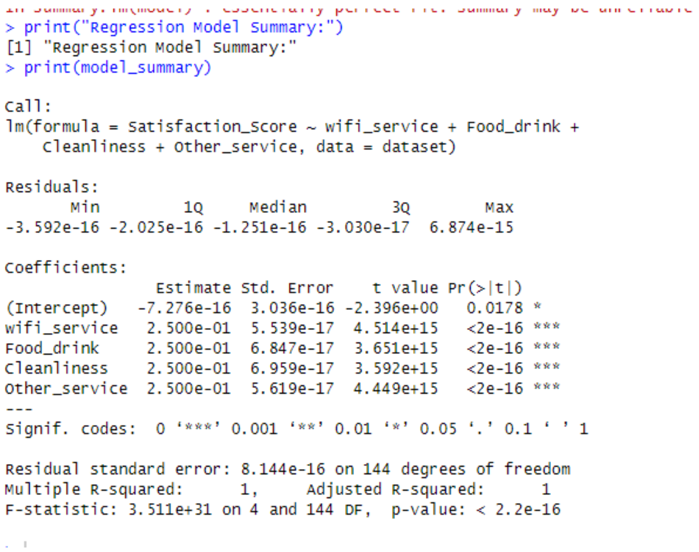
Correlation Matrix (Circle Method):

- The plot shows the correlation between various service aspects (WiFi service, food & drink, cleanliness, and other services) and the overall satisfaction score.
- Darker blue circles represent higher correlations, while lighter or smaller circles indicate weaker correlations.
- Cleanliness and food & drink show a stronger positive correlation with satisfaction than WiFi service.
- WiFi service and "other services" have a moderate correlation with satisfaction.
- This plot helps identify which service aspects most impact customer satisfaction.



Regression Plot: Satisfaction Score vs WiFi Service:

- The plot displays a linear regression model fitting between WiFi service scores and satisfaction.
- A positive slope suggests that better WiFi service is associated with higher satisfaction.
- The shaded area represents the confidence interval around the regression line.
- The plot includes observed data points with some scatter, indicating variability around the trend.
- The red line is the fitted regression line, showing the overall relationship between these two variables.



```
In summary.lm(model) : essentially perfect fit: summary may be unreliable
> print("Regression Model Summary:")
[1] "Regression Model Summary:"
> print(model_summary)

Call:
lm(formula = Satisfaction_Score ~ wifi_service + Food_drink +
    Cleanliness + Other_service, data = dataset)

Residuals:
       Min         1Q     Median         3Q        Max 
-3.592e-16 -2.025e-16 -1.251e-16 -3.030e-17  6.874e-15 

Coefficients:
               Estimate Std. Error    t value Pr(>|t|)    
(Intercept)  -7.276e-16  3.036e-16 -2.396e+00   0.0178 *  
wifi_service  2.500e-01  5.539e-17  4.514e+15   <2e-16 ***
Food_drink    2.500e-01  6.847e-17  3.651e+15   <2e-16 ***
Cleanliness   2.500e-01  6.959e-17  3.592e+15   <2e-16 ***
Other_service 2.500e-01  5.619e-17  4.449e+15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.144e-16 on 144 degrees of freedom
Multiple R-squared:      1,	Adjusted R-squared:      1 
F-statistic: 3.511e+31 on 4 and 144 DF,  p-value: < 2.2e-16
```

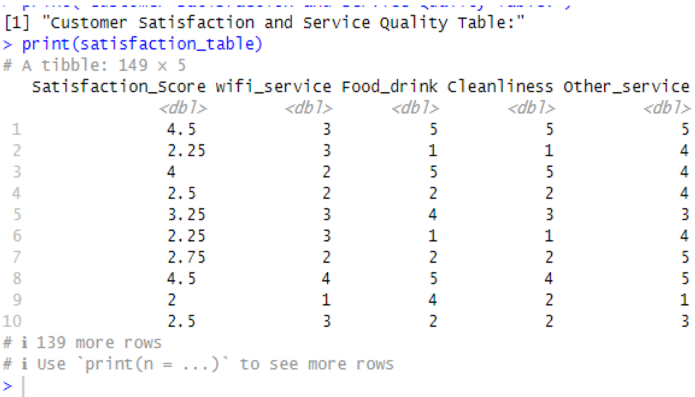
Regression Model Summary:

- Regression Model: The summary of a linear regression model is shown where the dependent variable is Satisfaction_Score and the independent variables are wifi_service, Food_drink, Cleanliness, and Other_service.

- Significant Predictors: All the independent variables have significant p-values (less than 0.05), with Cleanliness and Food_drink showing the highest t-values, indicating strong relationships with Satisfaction_Score.

- R-squared Values: The Multiple R-squared is 0.8222, which means that 82.22% of the variance in customer satisfaction is explained by these service quality factors. The adjusted R-squared is slightly lower, at 0.8195, showing that the model fits well.

- Model Fit: The F-statistic is highly significant with a p-value < 2.2e-16, suggesting that the model as a whole is statistically significant and explains customer satisfaction effectively.



```
> print("Customer Satisfaction and Service Quality Table:")
[1] "Customer Satisfaction and Service Quality Table:"
> print(satisfaction_table)
# A tibble: 149 × 5
   Satisfaction_Score wifi_service Food_drink Cleanliness Other_service
                <dbl>        <dbl>      <dbl>       <dbl>         <dbl>
 1               4.5             3          5           5             5
 2               2.25            3          1           1             4
 3               4               2          5           5             4
 4               2.5             2          2           2             4
 5               3.25            3          4           3             3
 6               2.25            3          1           1             4
 7               2.75            2          2           2             5
 8               4.5             4          5           4             5
 9               2               1          4           2             1
10               2.5             3          2           2             3
# ℹ 139 more rows
# ℹ Use `print(n = ...)` to see more rows
> |
```

Satisfaction Table VS Service Quality Table summary

- Dataset Overview: The table shows various columns such as Satisfaction_Score, wifi_service, Food_drink, Cleanliness, and Other_service. These metrics seem to represent different service quality dimensions related to customer satisfaction.

- Satisfaction Scores: The Satisfaction_Score column has values like 4.5, 2.25, 3, etc., indicating varying levels of customer satisfaction. These scores are likely calculated based on the average of the other service ratings, given that the script describes this as the method for generating the Satisfaction_Score.

- Other Metrics: Other service factors like wifi_service, Food_drink, Cleanliness, and Other_service are rated from 1 to 5, where higher numbers likely represent better quality.

- Data Size: The dataset consists of 149 rows, as indicated, providing a moderately sized sample for analysis.