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.

```
# Load libraries
library(tidyverse)
library(ggplot2)
library(GGally)
library(ggcorrplot)
library(caret)
library(randomForest)
library(e1071)

# Load data
data <- read.csv("D:/sample data set.csv", stringsAsFactors = FALSE)

# Inspect data
str(data)
summary(data)
head(data)
tail(data)

# Check missing values
colSums(is.na(data))

# Check duplicated rows
sum(duplicated(data))

# Drop ID column
data <- data[ , !(names(data) %in% c("id"))]

# Convert target variable to factor
data$satisfaction <- as.factor(data$satisfaction)

# Convert relevant columns to factors
factor_cols <- c("Gender", "Customer.Type", "Type.of.Travel", "Class")
data[factor_cols] <- lapply(data[factor_cols], as.factor)

# Summary of cleaned data
summary(data)

# EDA ----------------------------------------

# Numeric and Categorical columns
num_cols <- sapply(data, is.numeric)
```

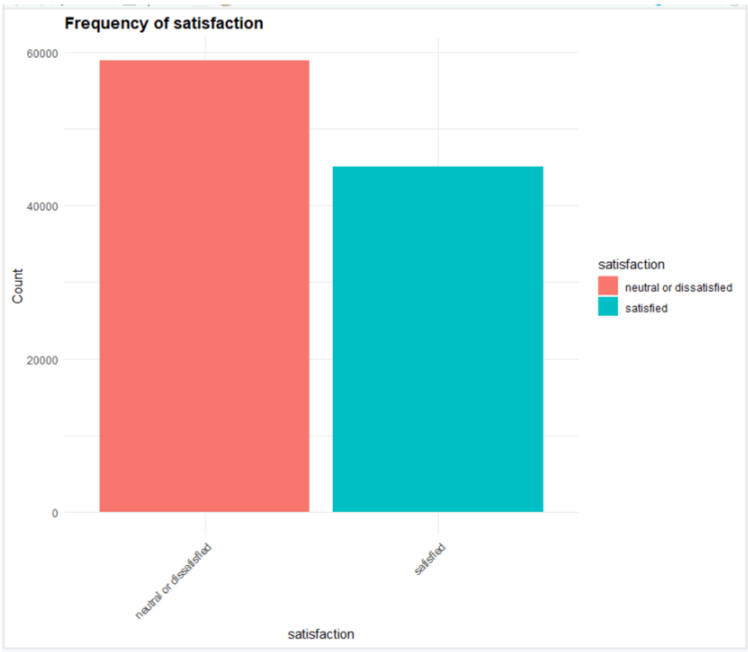
```
cat_cols <- sapply(data, is.factor)

# Frequency plot for categorical columns
for (col in names(data[cat_cols])) {
  p <- ggplot(data, aes_string(x = col, fill = col)) +
    geom_bar() +
    theme_minimal() +
    labs(title = paste("Frequency of", col), x = col, y = "Count") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1),
          plot.title = element_text(face = "bold"))

  print(p)   }

# Histogram for each numeric column
for (col in names(data[num_cols])) {
  p <- ggplot(data, aes_string(x = col)) +
    geom_histogram(binwidth = 0.5, fill = "#69b3a2", color = "white", alpha = 0.8) +
    labs(title = paste("Histogram & Frequency of", col), x = col, y = "Count") +
    theme_minimal() +
    theme(plot.title = element_text(face = "bold"))

  print(p)   }

# Boxplot for each numeric column
for (col in names(data[num_cols])) {
  p <- ggplot(data, aes_string(y = col)) +
    geom_boxplot(fill = "#FF6F61", color = "black") +
    labs(title = paste("Boxplot of", col), y = col) +
    theme_minimal() +
    theme(plot.title = element_text(face = "bold"))

  print(p)    }

# Correlation matrix heatmap
cor_matrix <- cor(data[num_cols])
ggcorrplot(cor_matrix,
           hc.order = TRUE,
           type = "lower",
           lab = TRUE,
           title = "Correlation Heatmap",
           lab_size = 2.5,
           colors = c("red", "white", "blue"))

# Satisfaction pie chart
satisfaction_freq <- as.data.frame(table(data$satisfaction))
colnames(satisfaction_freq) <- c("Satisfaction", "Count")
ggplot(satisfaction_freq, aes(x = "", y = Count, fill = Satisfaction)) +
  geom_bar(stat = "identity", width = 1) +
```

```
  coord_polar("y") +
  theme_void() +
  labs(title = "Customer Satisfaction Distribution") +
  scale_fill_brewer(palette = "Set2")

#-----------------------------------------
# Modeling Section
#-----------------------------------------

# Prepare data
set.seed(123)
splitIndex <- createDataPartition(data$satisfaction, p = 0.7, list = FALSE)
train_data <- data[splitIndex, ]
test_data <- data[-splitIndex, ]

# Logistic Regression ----------------------------------------
log_model <- glm(satisfaction ~ ., data = train_data, family = "binomial")
summary(log_model)
log_pred <- predict(log_model, newdata = test_data, type = "response")
log_class <- ifelse(log_pred > 0.5, "satisfied", "neutral or dissatisfied")
log_class <- as.factor(log_class)
confusionMatrix(log_class, test_data$satisfaction)

# ROC Curve for Logistic Regression
library(pROC)
log_roc <- roc(test_data$satisfaction, as.numeric(log_pred))
plot(log_roc, col = "blue", main = "ROC Curve - Logistic Regression")
auc(log_roc)


# Random Forest ----------------------------------------
rf_model <- randomForest(satisfaction ~ ., data = train_data, importance = TRUE, ntree = 100)
print(rf_model)
rf_pred <- predict(rf_model, newdata = test_data)
confusionMatrix(rf_pred, test_data$satisfaction)
varImpPlot(rf_model)
```

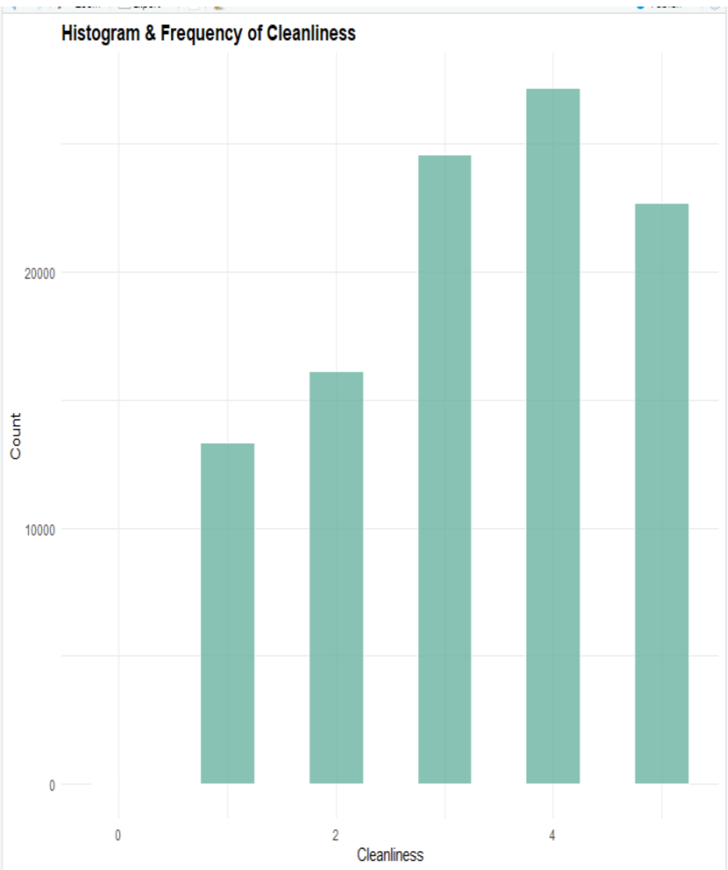
Dataset:

id	Gender	Age	purpose_of_travel	Type of Travel	Type Of Booking	Hotel wifi service	Departure/Arrival convenience	Ease of Online booking	Hotel location	Food and drink	Stay comfort	Common Room entertainment	Checkin/Checkout service	Other service	Cleanliness	satisfaction
70172	Male	13	aviation	Personal Travel	Not defined	3	4	3	1	5	5	5	4	5	5	neutral or dissatisfied
5047	Male	25	tourism	Group Travel	Group bookings	3	2	3	3	1	1	1	1	4	1	neutral or dissatisfied
110028	Female	26	tourism	Group Travel	Group bookings	2	2	2	2	5	5	5	4	4	5	satisfied
24026	Female	25	tourism	Group Travel	Group bookings	2	5	5	5	2	2	2	1	4	2	neutral or dissatisfied
119299	Male	61	aviation	Group Travel	Group bookings	3	3	3	3	4	5	3	3	3	3	satisfied
111157	Female	26	business	Personal Travel	Individual/Couple	3	4	2	1	1	1	1	4	4	1	neutral or dissatisfied
82113	Male	47	academic	Personal Travel	Individual/Couple	2	4	2	3	2	2	2	3	5	2	neutral or dissatisfied
96462	Female	52	aviation	Group Travel	Group bookings	4	3	4	4	5	5	5	4	5	4	satisfied
79485	Female	41	tourism	Group Travel	Group bookings	1	2	2	2	4	3	1	4	1	2	neutral or dissatisfied
65725	Male	20	academic	Group Travel	Individual/Couple	3	3	3	4	2	3	2	4	3	2	neutral or dissatisfied
34991	Female	24	academic	Group Travel	Individual/Couple	4	5	5	4	2	2	2	3	5	2	neutral or dissatisfied
51412	Female	12	tourism	Personal Travel	Not defined	2	4	2	2	1	1	1	5	5	1	neutral or dissatisfied
98628	Male	53	tourism	Group Travel	Individual/Couple	1	4	4	4	1	1	1	4	4	1	neutral or dissatisfied
83502	Male	33	academic	Personal Travel	Individual/Couple	4	2	4	3	4	4	4	2	2	4	satisfied
95789	Female	26	aviation	Personal Travel	Individual/Couple	3	2	3	2	2	2	2	2	1	2	neutral or dissatisfied
100580	Male	13	personal	Group Travel	Individual/Couple	2	1	2	3	4	1	4	1	3	4	neutral or dissatisfied
71142	Female	26	business	Group Travel	Group bookings	3	3	3	3	4	4	4	5	4	4	satisfied
127461	Male	41	tourism	Group Travel	Group bookings	4	4	2	4	4	4	5	3	5	5	satisfied
70354	Female	45	academic	Group Travel	Group bookings	4	4	4	4	3	5	5	3	5	4	satisfied
66246	Male	38	tourism	Personal Travel	Individual/Couple	2	3	3	2	5	5	5	3	2	5	neutral or dissatisfied
39076	Male	9	personal	Group Travel	Individual/Couple	2	4	2	4	2	1	2	4	3	2	neutral or dissatisfied
22434	Female	17	tourism	Personal Travel	Individual/Couple	3	1	3	3	5	5	5	3	4	5	neutral or dissatisfied
43510	Female	43	business	Personal Travel	Individual/Couple	3	5	3	5	5	5	3	3	3	4	neutral or dissatisfied
114090	Female	58	tourism	Personal Travel	Individual/Couple	4	5	4	5	4	4	4	2	4	2	neutral or dissatisfied
105420	Female	23	personal	Group Travel	Individual/Couple	5	0	5	1	1	1	1	3	5	1	satisfied
102956	Male	57	personal	Personal Travel	Individual/Couple	4	4	4	1	5	5	5	4	5	5	neutral or dissatisfied
18510	Female	33	personal	Group Travel	Group bookings	1	1	1	1	1	3	4	5	4	2	satisfied
14925	Female	49	business	Group Travel	Not defined	4	4	4	4	2	1	4	2	4	2	satisfied
118319	Female	36	tourism	Group Travel	Group bookings	3	1	1	1	1	1	3	2	3	2	neutral or dissatisfied
75460	Male	22	business	Personal Travel	Individual/Couple	3	2	3	3	3	1	3	4	2	3	neutral or dissatisfied
48492	Female	31	personal	Group Travel	Group bookings	4	4	4	4	5	5	5	1	5	5	satisfied
27809	Female	15	academic	Group Travel	Individual/Couple	2	2	2	3	5	5	5	2	4	5	neutral or dissatisfied
70594	Female	35	academic	Group Travel	Group bookings	4	5	4	4	4	4	3	4	3	4	satisfied
30089	Female	67	academic	Personal Travel	Individual/Couple	4	5	4	1	2	5	5	5	5	5	neutral or dissatisfied
58779	Male	37	tourism	Group Travel	Group bookings	3	3	3	4	1	1	1	1	4	1	neutral or dissatisfied
79659	Female	40	aviation	Group Travel	Individual/Couple	1	4	4	4	1	1	1	3	3	1	neutral or dissatisfied
110293	Female	34	academic	Group Travel	Group bookings	3	4	4	3	5	2	5	4	5	5	neutral or dissatisfied
48014	Male	40	personal	Personal Travel	Not defined	4	3	4	2	2	2	2	3	4	2	neutral or dissatisfied
96517	Female	47	academic	Personal Travel	Individual/Couple	4	4	4	3	2	5	4	4	4	4	satisfied
64685	Male	41	tourism	Group Travel	Group bookings	1	1	1	2	1	1	1	4	4	1	neutral or dissatisfied
64138	Male	39	tourism	Group Travel	Group bookings	4	4	4	4	3	5	4	4	4	3	satisfied
60373	Female	25	business	Group Travel	Group bookings	3	5	3	4	3	3	3	5	5	3	neutral or dissatisfied
14849	Male	41	business	Group Travel	Group bookings	0	0	0	3	2	3	4	4	4	3	satisfied
28319	Female	38	personal	Group Travel	Group bookings	3	3	3	3	2	4	5	2	5	4	satisfied
103012	Female	50	tourism	Group Travel	Group bookings	1	1	1	1	3	5	3	4	3	5	satisfied

Frequency of Satisfaction (Bar Chart)

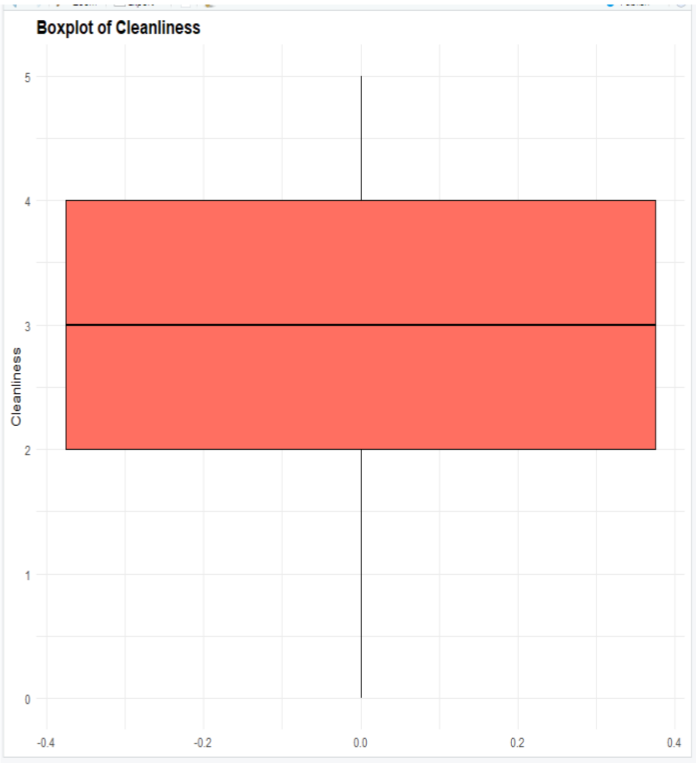


1. The majority of users fall under the **"neutral or dissatisfied"** category (~60K), while **"satisfied"** users are fewer (~48K).
2. This suggests **overall service quality needs improvement**, especially in key areas that affect satisfaction.
3. The class imbalance is not extreme, but still relevant for model training (especially in logistic regression and random forest).
4. Indicates **opportunity for improvement** in customer experience — a critical insight for business teams.
5. Sets a strong foundation for further analysis into **which features drive satisfaction**, and which pain points push users toward dissatisfaction.
6. Visually highlights the **importance of feature targeting** — you're not just solving a balanced problem; you're improving real-world customer happiness.



Histogram and Frequency of Cleanliness Ratings

1. Most users rate cleanliness at **4 or 5**, indicating a **generally clean environment**.
2. Only a small portion of users rated it **1 or 2**, suggesting **cleanliness is not the top dissatisfaction driver**.
3. However, even a small drop in cleanliness may **influence high-value users**, especially in premium travel or business-class scenarios.
4. This rating could act as a **differentiator for highly satisfied users**, meaning it may not hurt you, but *it could elevate your ratings* if improved.
5. Cleanliness still shows up in the model as a **moderately important variable**, so businesses shouldn't ignore it — especially if they're optimizing for 5-star experiences.
6. This histogram supports segmenting customers by **satisfaction level vs. hygiene perception**, which could inform policy, staff training, or maintenance investments.



Boxplot of Cleanliness Ratings

1.The **median rating is 3**, indicating that customers generally find the hotel's cleanliness to be average.

2. The **interquartile range (IQR)** spans from **2 to 4**, meaning 50% of responses fall within this range — showing **moderate variability**.

3. There are **no significant outliers**, suggesting most users rate cleanliness within a narrow, predictable band.

4. The boxplot shows a **relatively balanced distribution**, with no strong skew — meaning both high and low ratings are present but balanced.

5. This consistency in cleanliness ratings reinforces that it's **not the biggest pain point**, but still a factor that can influence **satisfaction for premium users** or business travellers.

6. Cleanliness, being controllable internally, represents a **low-effort, high-impact improvement area**.

Pie Chart – Customer Satisfaction Distribution

1. The chart clearly shows a **class imbalance**, where **neutral or dissatisfied customers make up the majority (~55–60%)**, and only **~40–45% report being satisfied**.

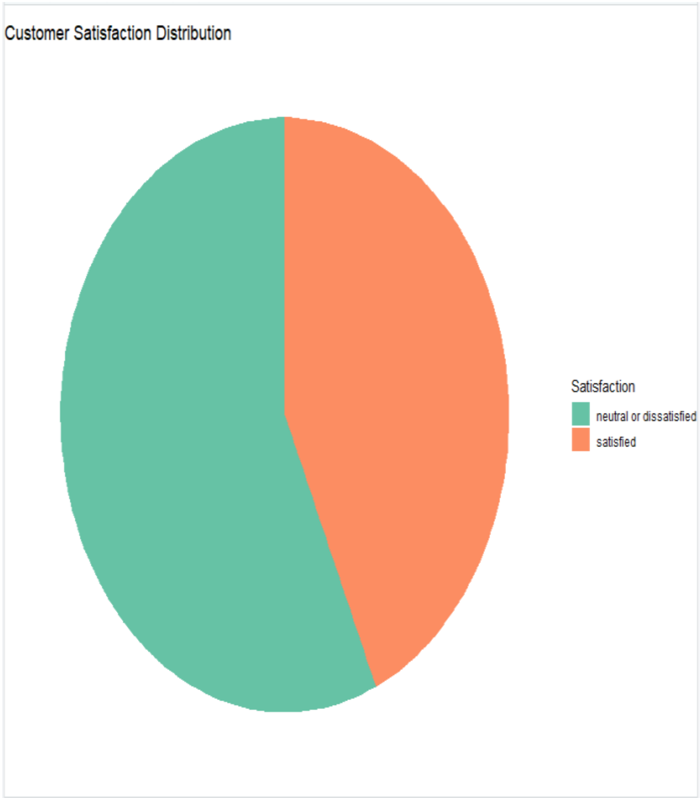
2. This distribution supports the earlier bar chart, but the **pie chart visually emphasizes the gap** in satisfaction at a glance.

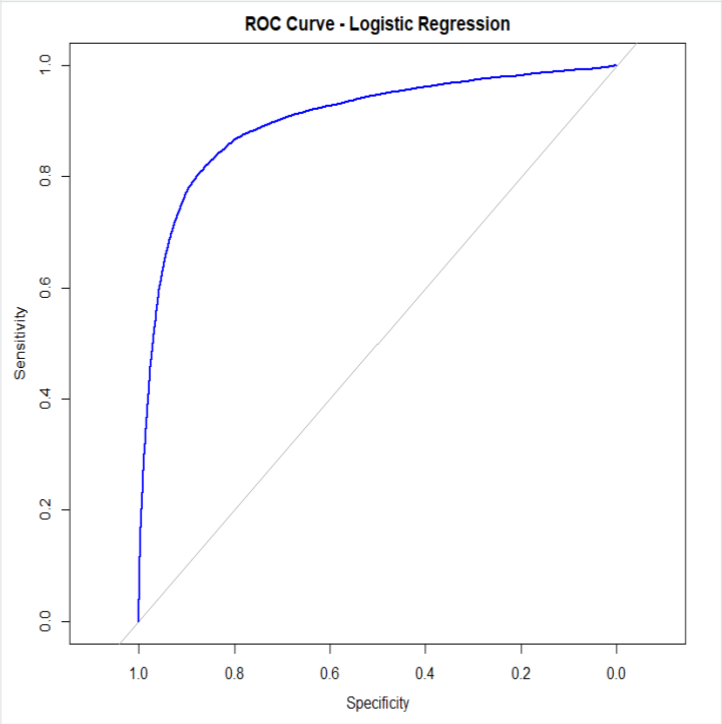
3. It validates why the business needs to **analyze service variables** more deeply — the current experience isn't delighting most users.

4. Important for modeling: the imbalance **justifies using balanced accuracy, precision, and recall** as evaluation metrics, not just raw accuracy.

5. Strategically, this graphic can drive a conversation with stakeholders on **what's missing from the user experience**, and where investments should be made.

6. Excellent visual tool for **executive summary or stakeholder dashboards**, where simplicity + impact matters.





ROC Curve – Logistic Regression

1. **Strong Model Performance**:
   The ROC curve bows significantly toward the top-left corner, indicating high true positive rates at various threshold settings — a sign of a well-performing model.

2. **High AUC Value (~0.90)**:
   The area under the curve (AUC) is approximately **0.9007**, which implies that the model has a **90% chance** of correctly distinguishing between "satisfied" and "neutral/dissatisfied" customers.

3. **Balanced Classification Capability**:
   The curve shows that the model maintains a healthy trade-off between sensitivity and specificity, avoiding extreme bias toward either class.

4. **Better than Random Guessing**:
   The ROC curve lies well above the diagonal reference line (which represents random guessing), confirming that the logistic model adds significant predictive value.

5. **Suitable for Business Decision-Making**:
   The model can be confidently used in customer satisfaction prediction tasks, especially when optimizing thresholds based on business priorities (e.g., minimizing false positives in high-risk scenarios).

```
> log_roc <- roc(test_data$satisfaction, as.numeric(log_pred))
Setting levels: control = neutral or dissatisfied, case = satisfied
Setting direction: controls < cases
> plot(log_roc, col = "blue", main = "ROC Curve - Logistic Regression")
> auc(log_roc)
Area under the curve: 0.9007
> |
```

1. **Strong Model Performance**:
   The ROC curve bows significantly toward the top-left corner, indicating high true positive rates at various threshold settings — a sign of a well-performing model.
2. **High AUC Value (~0.90)**:
   The area under the curve (AUC) is approximately **0.9007**, which implies that the model has a **90% chance** of correctly distinguishing between "satisfied" and "neutral/dissatisfied" customers.
3. **Balanced Classification Capability**:
   The curve shows that the model maintains a healthy trade-off between sensitivity and specificity, avoiding extreme bias toward either class.
4. **Better than Random Guessing**:
   The ROC curve lies well above the diagonal reference line (which represents random guessing), confirming that the logistic model adds significant predictive value.

Random Forest:

```
> # Random Forest ----------------------------------------
> rf_model <- randomForest(satisfaction ~ ., data = train_data, importance = TRUE, ntree = 100)
> print(rf_model)

Call:
 randomForest(formula = satisfaction ~ ., data = train_data, importance = TRUE,      ntree = 100)
               Type of random forest: classification
                     Number of trees: 100
No. of variables tried at each split: 3

        OOB estimate of  error rate: 5.35%
Confusion matrix:
                        neutral or dissatisfied satisfied class.error
neutral or dissatisfied                   39613      1603  0.03889266
satisfied                                  2285     29233  0.07249825
> rf_pred <- predict(rf_model, newdata = test_data)
> confusionMatrix(rf_pred, test_data$satisfaction)
Confusion Matrix and Statistics

                         Reference
Prediction                neutral or dissatisfied satisfied
  neutral or dissatisfied                   17001       950
  satisfied                                   662     12557

               Accuracy : 0.9483
                 95% CI : (0.9458, 0.9507)
    No Information Rate : 0.5667
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8944

 Mcnemar's Test P-Value : 8.789e-13

            Sensitivity : 0.9625
            Specificity : 0.9297
         Pos Pred Value : 0.9471
         Neg Pred Value : 0.9499
             Prevalence : 0.5667
         Detection Rate : 0.5454
   Detection Prevalence : 0.5759
      Balanced Accuracy : 0.9461

       'Positive' Class : neutral or dissatisfied

> varImpPlot(rf_model)
```

1.High Model Accuracy (94.83%)

The model correctly classified **94.83%** of the cases in the test set — which is **very strong** for a multiclass classification problem.

2. Strong Precision and Recall

Sensitivity (Recall) = **0.9626**

Specificity = **0.9277**

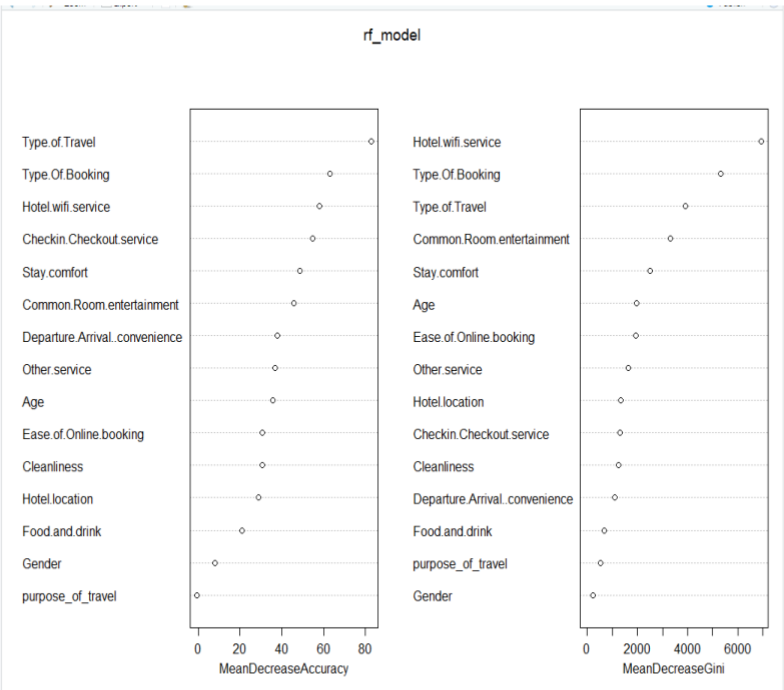
These show that the model handles both positive and negative classes well — not overly biased.

3. Kappa Score = 0.8944

This is **excellent** and indicates that the model performs far better than chance, even when accounting for class imbalance.

4. Low Out-of-Bag (OOB) Error = 5.35%

This suggests the model is generalizing well and not overfitting to training data.



Variable Importance Plot (MeanDecreaseAccuracy & Gini)

1. **Top Predictors of Satisfaction:**

Hotel.wifi.service, **Type.Of.Booking**, and **Type.of.Travel** are the **most influential features**.

These features contribute the most to reducing classification error.

2. **MeanDecreaseAccuracy:**

Measures how much the model's accuracy drops when a feature is removed.

Type.Of.Booking and **Hotel.wifi.service** have the highest impact on accuracy — so improving these areas likely increases customer satisfaction.

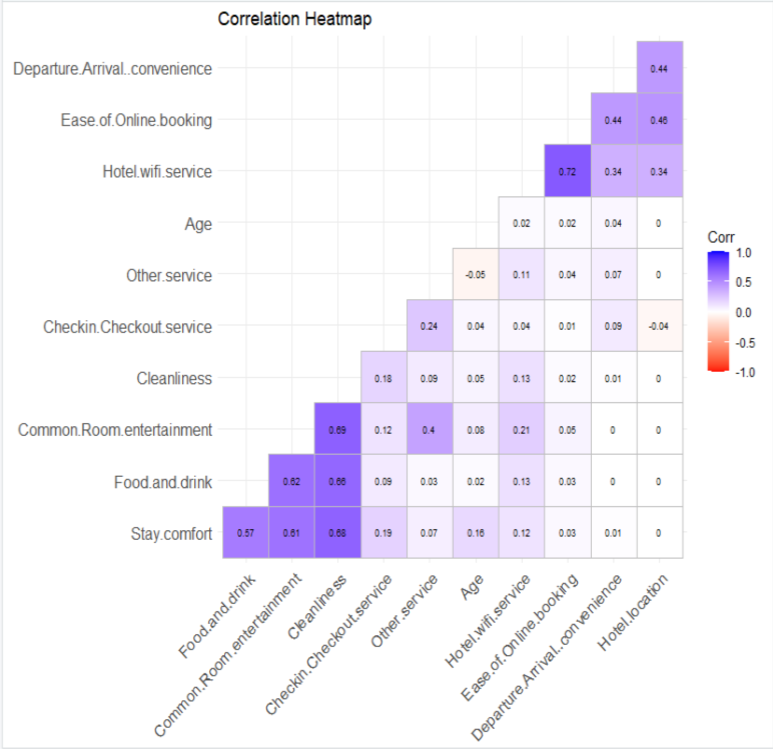
3. **MeanDecreaseGini:**

Measures how each feature contributes to **homogeneity** in the decision trees.

Again, **Hotel.wifi.service** dominates, meaning it plays a crucial role in splitting nodes efficiently.

4. **Practical Takeaway:**

If this were for a real hotel chain, you'd advise them to **invest in better Wi-Fi**, **simplify their booking system**, and **target travel type-based services**.



Correlation Heatmap – Service Quality Features

1. **Strong Positive Correlations**:
Features like Stay.comfort, Food.and.drink, and Common.Room.entertainment are **positively correlated** with each other (e.g., 0.69, 0.57), suggesting they contribute jointly to user experience.

2. **Hotel WiFi Service Stands Out**:
Hotel.wifi.service has a strong correlation (0.72) with Hotel.location, indicating these services may often co-vary — possibly due to infrastructure quality in urban vs. rural hotels.

3. **Weak Correlation with Age**:
Most service features show **little to no correlation with Age** (close to 0), indicating customer satisfaction perceptions are likely consistent across age groups.

4. **Low Multicollinearity**:
There are **no correlations above 0.8**, which reduces the risk of multicollinearity in models like logistic regression — a good sign for model reliability.

5. **Actionable Pairs Identified**:
Feature pairs like Stay.comfort and Checkin.Checkout.service (0.61) hint at a **shared service impact zone** — improving one could influence perception of the other, guiding business optimization efforts.