AirBnB data analysis

(Italy, XIII Aurelia)

Charlie's Angels

Handling Missing Values

Categorical Variables: Mode Imputation

- Variables: `host_location` and `host_is_superhost`
- Method: Mode imputation for distribution preservation.

```
#host location
table(neighbourhood_data1$host_location)

# Impute missing values with the mode
mode_host_location <- names(sort(table(neighbourhood_data1$host_location), decreasing = TRUE))[1]
neighbourhood_data1$host_location[is.na(neighbourhood_data1$host_location)] <- mode_host_location
table(neighbourhood_data1$host_location)
```

Numerical Variables: Mean Imputation

- Variables: `host_response_rate` and `host_acceptance_rate`
- Method: Convert placeholder strings to numeric, mean imputation.

Decision Points & Strategic Actions

- Removal decisions for `bathrooms_text` and `bedrooms`.
- Imputation for `reviews_per_month` and `review_scores`.

Outliers & Data Optimization

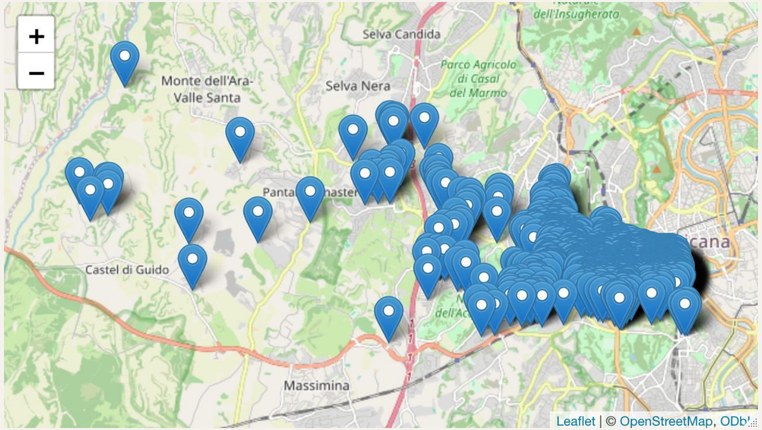
```r
# Calculate quartiles and IQR
Q1 <- quantile(neighbourhood_data$minimum_nights, 0.25)
Q3 <- quantile(neighbourhood_data$minimum_nights, 0.75)
IQR <- Q3 - Q1

# Define upper and lower bounds for outliers
lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR

# Identify potential outliers
outliers <- neighbourhood_data$minimum_nights < lower_bound | neighbourhood_data$minimum_nights > upper_bound

# Print values of potential outliers
print(neighbourhood_data$minimum_nights[outliers])
```

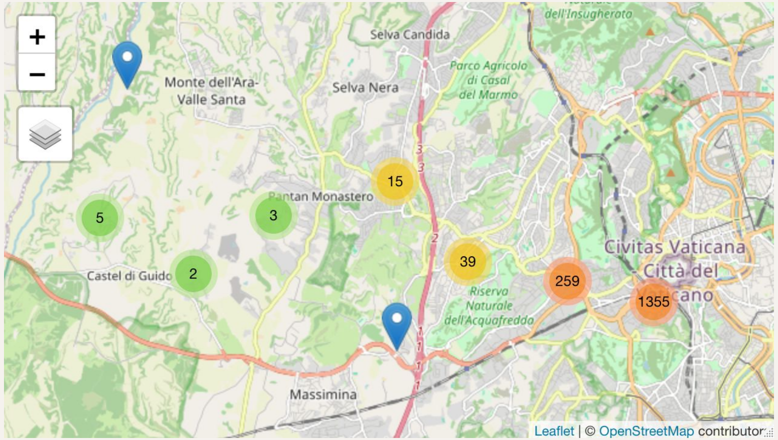
```
> table(neighbourhood_data1$minimum_nights)

  1   2   3   4   5   6   7  10  14  15  18  20  25  28  30  31  60  61  90 364 365
548 673 350  43  27   5   9   7   1   4   2   2   1   1  16   2   2   1   2   1   1
```

```r
class(neighbourhood_data1$instant_bookable)
summary(neighbourhood_data1$instant_bookable)
neighbourhood_data1$instant_bookable <- ifelse(neighbourhood_data1$instant_bookable == "t", 1, 0)
#we converted t f to 1 and 0
```

Mapping

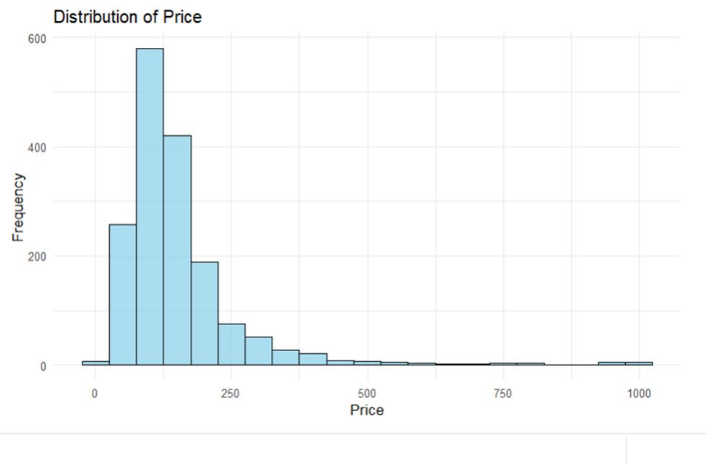
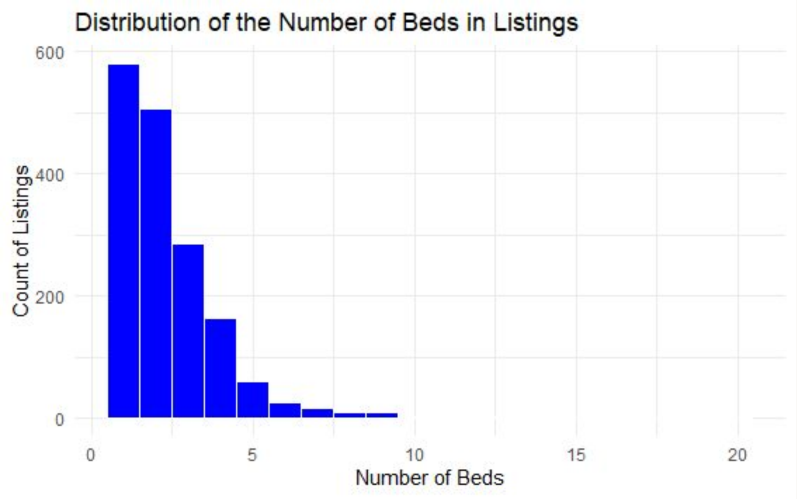
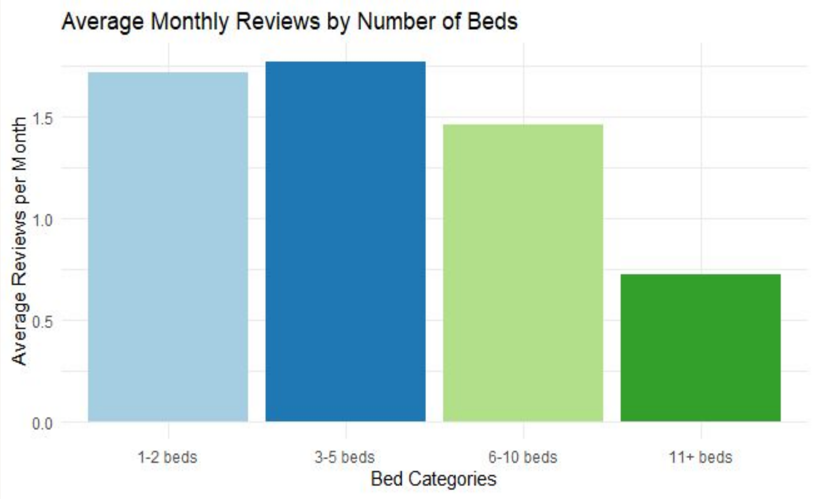
- Leaflet package for an interactive property map
- Tmap and sf packages, enabling the creation of a more detailed map
- Highlighted concentrated property clusters, particularly around Citta del Vaticano



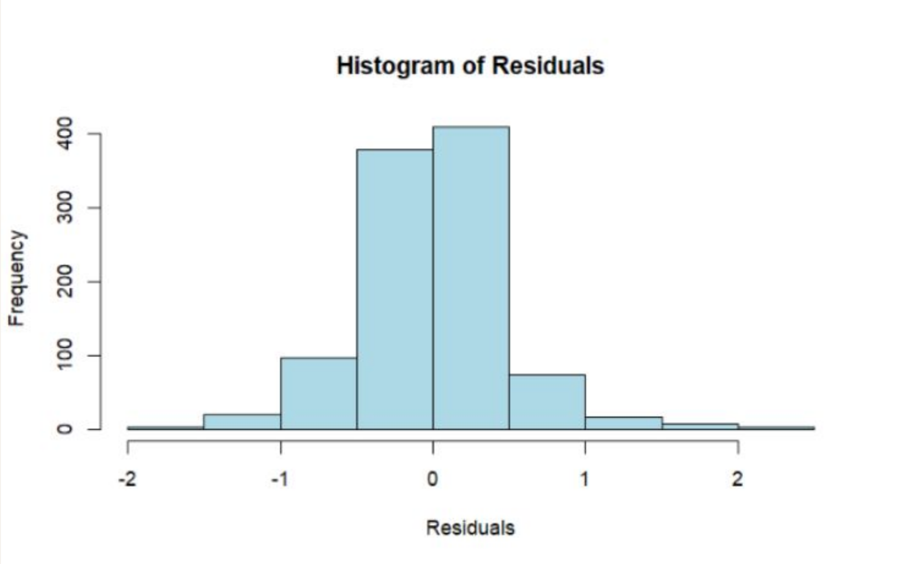


Summary Statistics for Price Variable

- Price is a critical variable influencing consumer choices in Airbnb listings.
- Five-number summary: Min, 1st Qu., Median, Mean, 3rd Qu., Max.

```
```{r}

summary(df$price)

```
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   20.0    90.0   125.0   153.4   174.0   986.0 
```

Histogram of Price Variable



- Histogram shows right-skewed distribution with higher concentration at lower prices.

"Entire home/apt" has highest mean and median prices.

```
room_type_summary
        room_type price.Mean price.Median
Entire home/apt   163.1760     133.5000
     Hotel room   136.1000     118.0000
   Private room   126.8511     100.0000
```

Verified hosts tend to have higher average and median prices.

```
> identify_verified
  host_identity_verified price.Mean price.Median
1                      f   138.1770     100.0000
2                      t   154.4764     126.0000
> 
```

Correlation with “beds”

```
[1] "Correlation Matrix:"
          price      beds
price 1.0000000 0.3401174
beds  0.3401174 1.0000000
```

Also, a quick glance at ‘beds’





Prediction - Regression Modeling

- Filtering out non-significant variables for multiple linear regression, based on domain knowledge.
- Splitting data into training (60%) and test sets (40%).
- Addressing multicollinearity issues and simplifying the model.

Highly Correlated Variables: Example

	review_scores_rating	review_scores_accuracy	review_scores_cleanliness	review_scores_checkin	review_scores_communication	review_scores_location	review_scores_value
review_scores_rating	1	0.606575061	0.585691076	0.494459361	0.557390227	0.396742444	0.598366026
review_scores_accuracy	0.606575061	1	0.760600935	0.708424243	0.769518116	0.530953598	0.789113579
review_scores_cleanliness	0.585691076	0.760600935	1	0.596016783	0.703337795	0.456591476	0.725936375
review_scores_checkin	0.494459361	0.708424243	0.596016783	1	0.776057338	0.509316278	0.662220969
review_scores_communication	0.557390227	0.769518116	0.703337795	0.776057338	1	0.553107094	0.723073371
review_scores_location	0.396742444	0.530953598	0.456591476	0.509316278	0.553107094	1	0.623236332
review_scores_value	0.598366026	0.789113579	0.725936375	0.662220969	0.723073371	0.623236332	1

Regression Model Results (Backward Elimination)

```
Residuals:
     Min       1Q   Median       3Q      Max 
-1.79498 -0.27075  0.00876  0.24009  2.10327 

Coefficients:
                                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                5.2665933  0.3578884  14.716  < 2e-16 ***
host_identity_verifiedt                    0.1507163  0.0650601   2.317 0.020734 *  
room_typeHotel room                       -0.0707140  0.1181726  -0.598 0.549715    
room_typePrivate room                     -0.3168582  0.0735292  -4.309 1.80e-05 ***
bathrooms_text0 baths                      0.0788431  0.4459748   0.177 0.859712    
bathrooms_text0 shared baths              -0.9489209  0.4459758  -2.128 0.033609 *  
bathrooms_text1 bath                      -0.8993893  0.2991314  -3.007 0.002709 ** 
bathrooms_text1 private bath              -0.7269655  0.2911509  -2.497 0.012694 *  
bathrooms_text1 shared bath               -1.0131384  0.3015677  -3.360 0.000811 ***
bathrooms_text1.5 baths                   -0.6579431  0.3094491  -2.126 0.033740 *  
bathrooms_text1.5 shared baths            -1.0808374  0.3566645  -3.030 0.002507 ** 
bathrooms_text2 baths                     -0.5842525  0.2999519  -1.948 0.051724 .  
bathrooms_text2 shared baths               0.2187651  0.3619769   0.604 0.545744    
bathrooms_text2.5 baths                   -0.6529565  0.3394504  -1.924 0.054700 .  
bathrooms_text3 baths                     -0.1478475  0.3102109  -0.477 0.633752    
bathrooms_text3 shared baths              -0.3980365  0.5578942  -0.713 0.475731    
bathrooms_text3.5 baths                    0.4621453  0.5584062   0.828 0.408092    
bathrooms_text4 baths                     -0.1093698  0.3572856  -0.306 0.759584    
bathrooms_text5 baths                     -0.4030341  0.3562364  -1.131 0.258180    
bathrooms_text5 shared baths              -1.9885235  0.5723717  -3.474 0.000535 ***
bathrooms_text5.5 baths                   -0.0954118  0.5626610  -0.170 0.865382    
bathrooms_text6 baths                     -2.2488136  0.5735305  -3.921 9.44e-05 ***
bathrooms_text6 shared baths              -2.0834693  0.5750055  -3.623 0.000306 ***
bathrooms_text7 baths                      0.1123378  0.4666547   0.241 0.809815    
bathrooms_textHalf-bath                   -0.8135016  0.5654793  -1.439 0.150584    
bathrooms_textShared half-bath            -0.8480498  0.4452953  -1.904 0.057144 .  
beds                                       0.0372340  0.0125411   2.969 0.003061 ** 
minimum_nights                            -0.0086070  0.0030610  -2.812 0.005026 ** 
has_availability                          -0.4283036  0.1102587  -3.885 0.000109 ***
availability_365                           0.0009797  0.0001311   7.472 1.76e-13 ***
number_of_reviews_ltm                     -0.0033300  0.0008064  -4.130 3.95e-05 ***
review_scores_rating                       0.0865730  0.0354458   2.442 0.014767 *  
instant_bookable                           0.1376459  0.0315439   4.364 1.42e-05 ***
calculated_host_listings_count_private_rooms  0.0205479  0.0088505   2.322 0.020457 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4719 on 973 degrees of freedom
Multiple R-squared:  0.341,	Adjusted R-squared:  0.3187 
F-statistic: 15.26 on 33 and 973 DF,  p-value: < 2.2e-16
```

R-squared= 0.341
Adjusted R-squared= 0.3187
F-statistic= 15.26
P-value < 2.2e-16

Significant predictors:
host_identity_verified, room_type (specific categories), bathrooms_text (specific categories), beds, minimum_nights, has_availability, availability_365, number_of_reviews_ltm, review_scores_rating, instant_bookable, and calculated_host_listings_count_private_rooms.

Residuals

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.79498 -0.27075  0.00876  0.24009  2.10327
```



Residuals follow a normal distribution

Accuracy Measures

```
pred = predict (model_1, train_df)
accuracy(pred, train_df$price)

```
```

               ME   RMSE   MAE   MPE   MAPE
Test set 146.8047 187.2087 146.8047 95.62446 95.62446
```

```
pred = predict (model_1, valid_df)
accuracy(pred, valid_df$price)

```
```

               ME   RMSE   MAE   MPE   MAPE
Test set 151.5474 194.9045 151.5474 95.72861 95.72861
```

- RMSE and MAE and all other error measures slightly higher for the validation set.
- No issue of overfitting.

KNN

- Engineered target variable indicating **wifi availability** (chosen amenity)

- Tuned k value via cross-validation for optimal model performance

- Achieved test accuracy of **97%** in predicting wifi availability

k	Accuracy	Kappa
5	0.9722079	-0.001351351
7	0.9732079	0.000000000
9	0.9732079	0.000000000
11	0.9732079	0.000000000
13	0.9732079	0.000000000
15	0.9732079	0.000000000
17	0.9732079	0.000000000
19	0.9732079	0.000000000
21	0.9732079	0.000000000
23	0.9732079	0.000000000

Fictional Host and their 7 nearest neighbors

Description: df [7 × 2]

	amenities <fctr>	host_name <chr>
13	TRUE	Dani&Pietro
189	TRUE	Somrit
218	TRUE	Maria Luisa
345	TRUE	Cristiano
642	TRUE	Luca
849	TRUE	Valerio
1007	TRUE	Catia

7 rows

- Accommodate 6 people
- A total listing count of 6
- Price for each listing is 300$
- 3 beds per accommodation

Naive Bayes Classification

- **Binned** continuous rating variable for use in classification (into 4 equal frequency bins)

- Carefully **selected predictors** representing property and host factors

- Demonstrated model application by **predicting rating** for fictional apartment

1	2	3	4
438	527	302	412

````
```{r}
#New apartment
new_apt <- data.frame(
  host_identity_verified = 0,
  property_category = "Private Room",
  binned_property_type = "Private Room",
  beds_grouped = 2
)

print(new_apt)
```
````

review_scores_rating_bin	**Min_Rating**	**Max_Rating**
<fctr>	<dbl>	<dbl>
1	0.00	4.67
2	4.68	4.82
3	4.83	4.92
4	4.93	5.00

Classification Tree

```
# Ensure consistency in factor levels between train and test data
for (var in selected_vars) {
  train_data[[var]] <- as.factor(train_data[[var]])
  test_data[[var]] <- factor(test_data[[var]], levels = levels(train_data[[var]]))
}
```

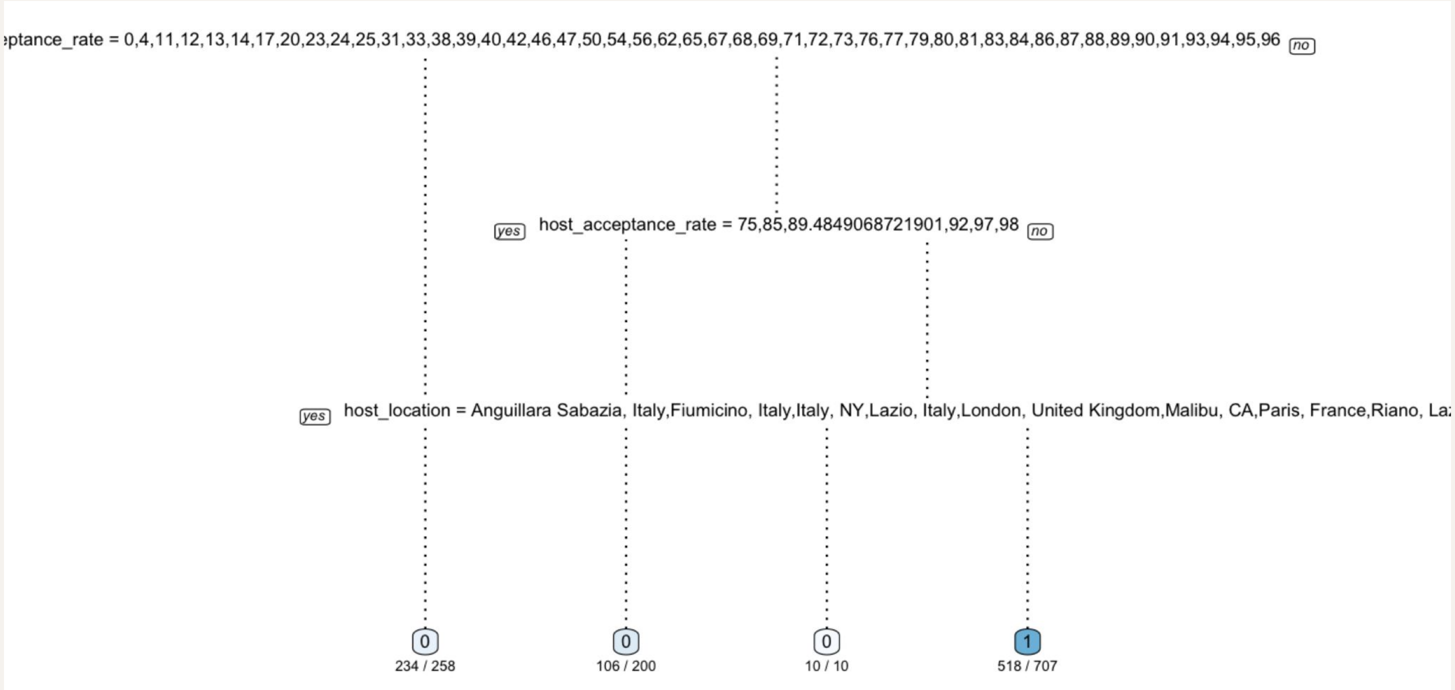
Objective: Predict instant bookability using a classification tree.

1. Data Preparation and Exploration
   - Converted categorical variables to factors.
   - Selected important variables for analysis.

2. Data Splitting
   - Split dataset into training and testing sets.
   - Ensured consistency in factor levels between them.

3. Classification Tree Construction
   - Applied pruning to avoid overfitting.
   - Visualized the pruned tree with enhanced aesthetics using `rpart.plot`.

```
> print(var_importance)
                                  Overall
host_acceptance_rate           121.457399
host_location                    4.610755
host_name                      282.773149
host_response_time              66.435290
host_since                     489.518391
neighborhood_overview          326.352425
host_id                          0.000000
host_response_rate               0.000000
host_is_superhost                0.000000
host_listings_count              0.000000
host_total_listings_count        0.000000
host_has_profile_pic             0.000000
host_identity_verified           0.000000
latitude                         0.000000
longitude                        0.000000
```




```r
# Evaluate model performance
confusion_matrix <- table(predictions, test_data$instant_bookable)
accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
print(paste("Accuracy:", round(accuracy, 2)))
```

```
[1] "Accuracy: 0.66"
```

Tree Cross Validation

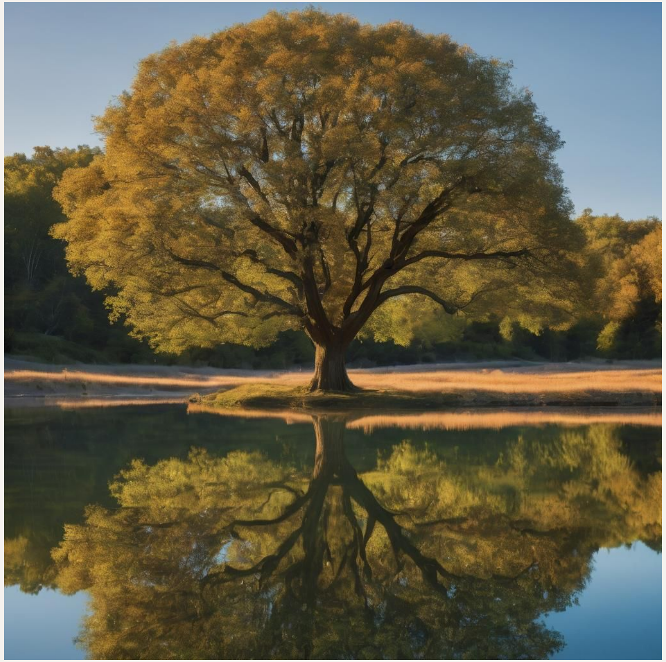
Objective: Determine the ideal size of the classification tree using cross-validation.

Created a grid of `cp` values ranging from 0.001 to 0.1.

Implemented 5-fold cross-validation with the created `cp` grid.

Extracted the best `cp` value from the cross-validation results.

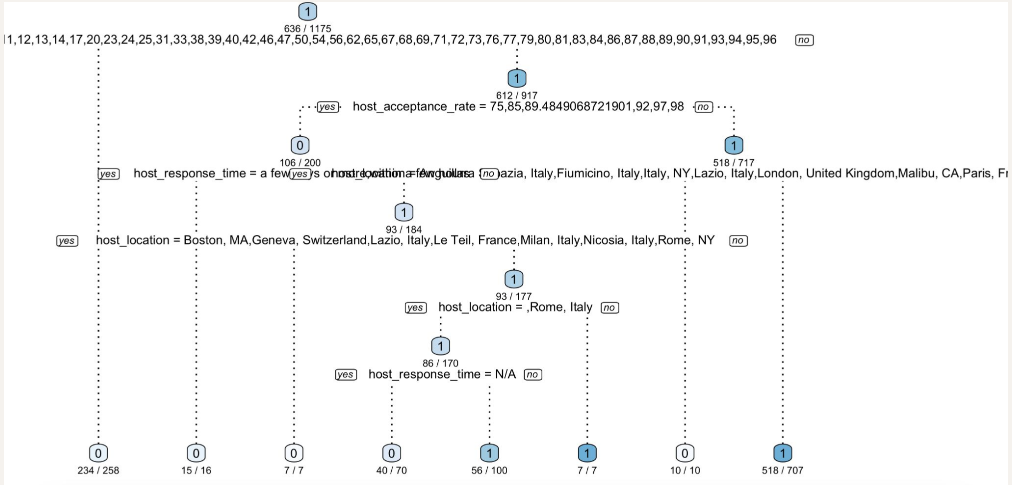
Constructed the tree model using the optimal `cp` value.



Mirror, mirror on the wall, who's the optimal tree of them all?

Visualization Process
Enhanced aesthetics and interpretability.

Tree Visualization
Displayed the resulting tree plot with the chosen graphical parameters.



Summarized the cross-validated tuning process and visualization.

Emphasized the importance of choosing an optimal tree size for better model performance.

```
# Create a grid of complexity parameter (cp) values
cp_grid <- expand.grid(cp = seq(0.001, 0.1, by = 0.001))

# Perform cross-validated tuning of the rpart model
fit_control <- trainControl(method = "cv", number = 5)
cv_results <- train(instant_bookable ~ ., data = train_data, method = "rpart",
                    trControl = fit_control, tuneGrid = cp_grid)

# Extract the best cp value
best_cp <- cv_results$bestTune$cp

# Build the tree with the optimal size
tree_model_optimal <- rpart(instant_bookable ~ ., data = train_data, method = "class", cp = best_cp)
```

```
[1] "Overall Model Accuracy: 0.67"
```

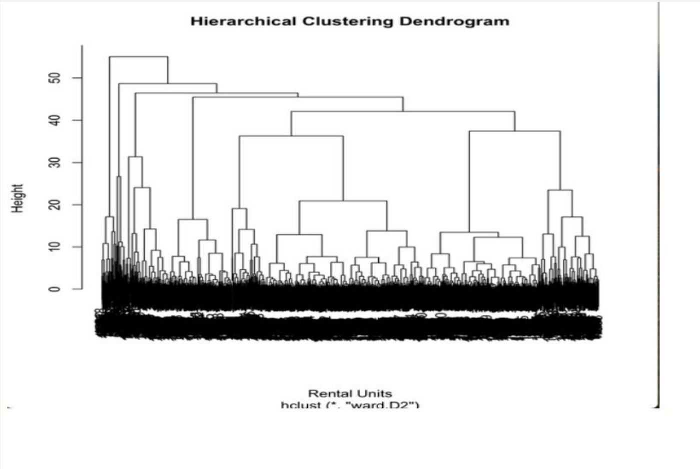
```
[1] "Complexity Parameter: 0.00834879406307978"
```

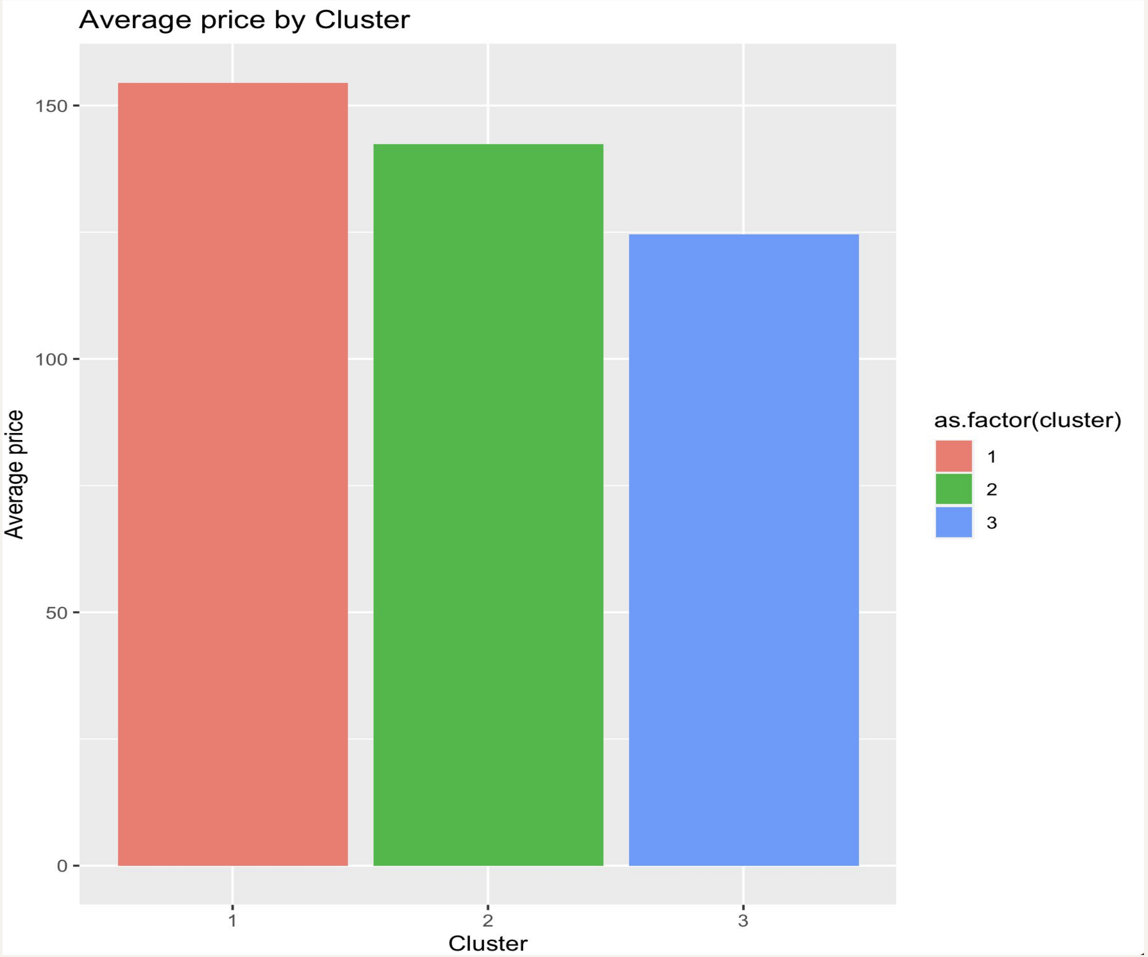
Cluster Analysis

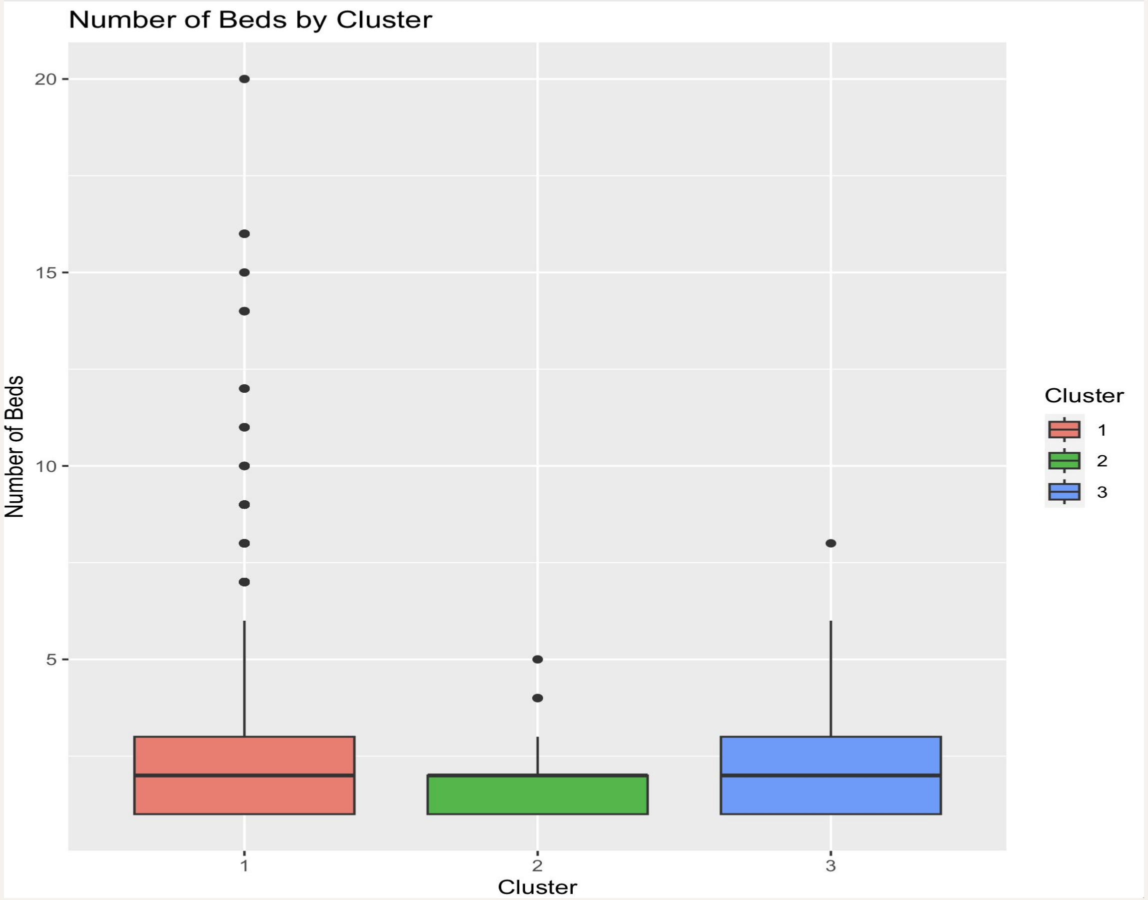
- **Key Variables:**
  - Accommodation, Pricing, Host Responsiveness, Guest Reviews
- **Methodology:**
  - Hierarchical clustering (Ward's linkage)
- **Cluster Decision:**
  - Three clusters chosen for balance
- **Insights for Stakeholders:**
  - Valuable for owners, hosts, and guests
- **Resulting Benefits:**
  - Nuanced understanding of the neighborhood rental market

Cluster Analysis Insights

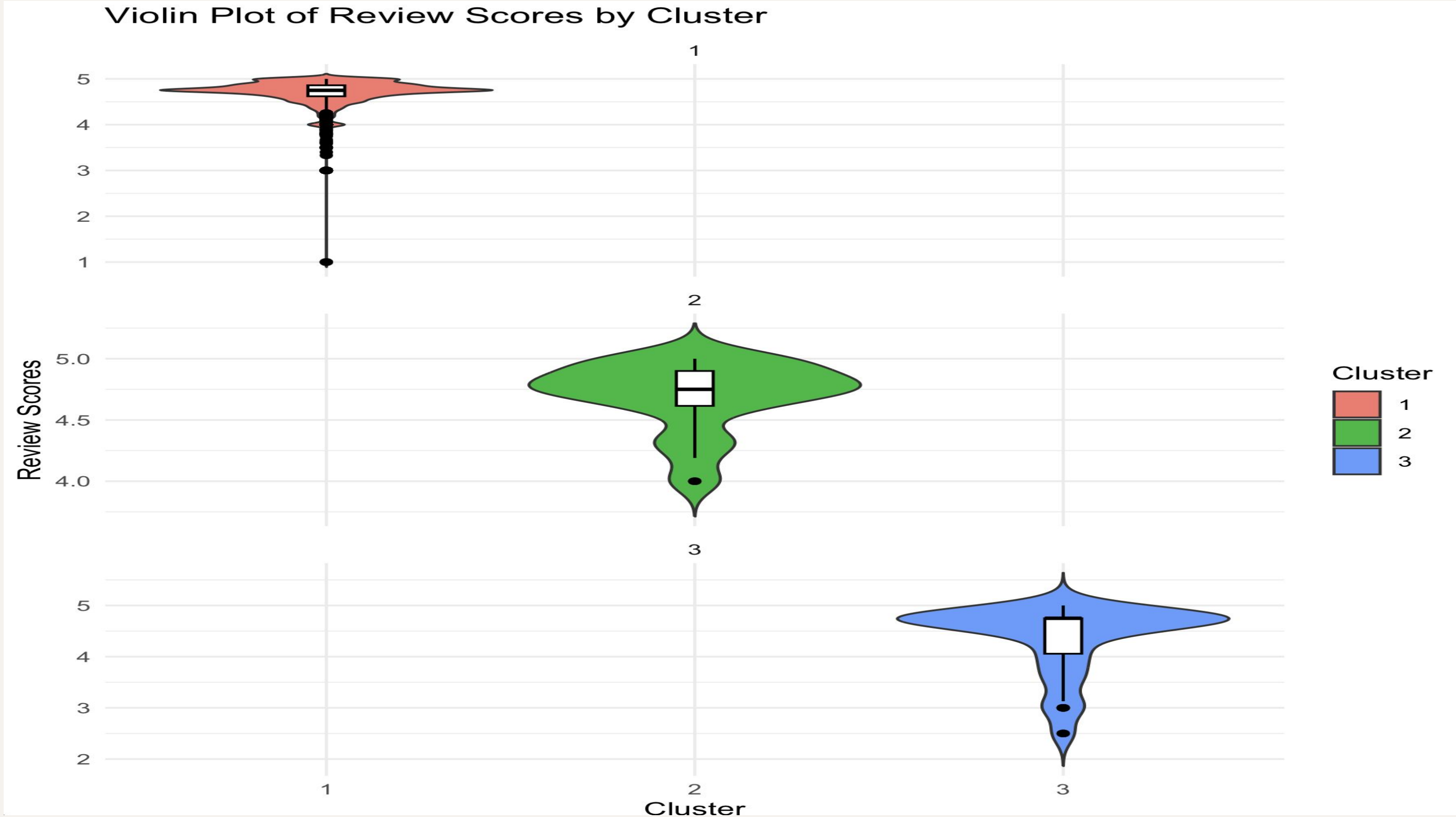
- Identified 3 distinct rental property clusters using hierarchical clustering
- Cluster 1 represents expensive, luxurious retreats
- Cluster 2 offers comfortable urban accommodations
- Cluster 3 caters to budget-conscious travelers




Average price by Cluster
150
100
50
0
Average price
1
2
3
Cluster
as.factor(cluster)
1
2
3



Number of Beds by Cluster
Number of Beds
20
15
10
5
1
2
3
Cluster
Cluster
1
2
3


Violin Plot of Review Scores by Cluster
1
2
3
Review Scores
Cluster
Cluster
1
2
3
1
2
3
4
5
4.0
4.5
5.0
2
3
4
5

Conclusion

- **Project Challenges and Impact :**

- Project exposed team to messy, real-world scraped data
- Required extensive data cleaning and interpretation effort
- Findings empower hosts, property managers to optimize listings

- **Skills Developed:**
  - Data cleaning, interpretation, visualization and analysis

- **Real-World Benefits:**
  - Our work helps hosts set good prices and keeps guests happy. It also provides useful info for people looking for Airbnb stays. Plus, it can help Airbnb make its platform better and even guide policymakers in making fair rules.

Thank You!