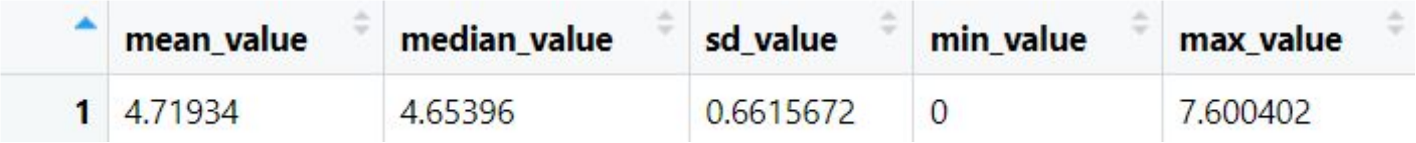
AD699 Team Project

Statistics

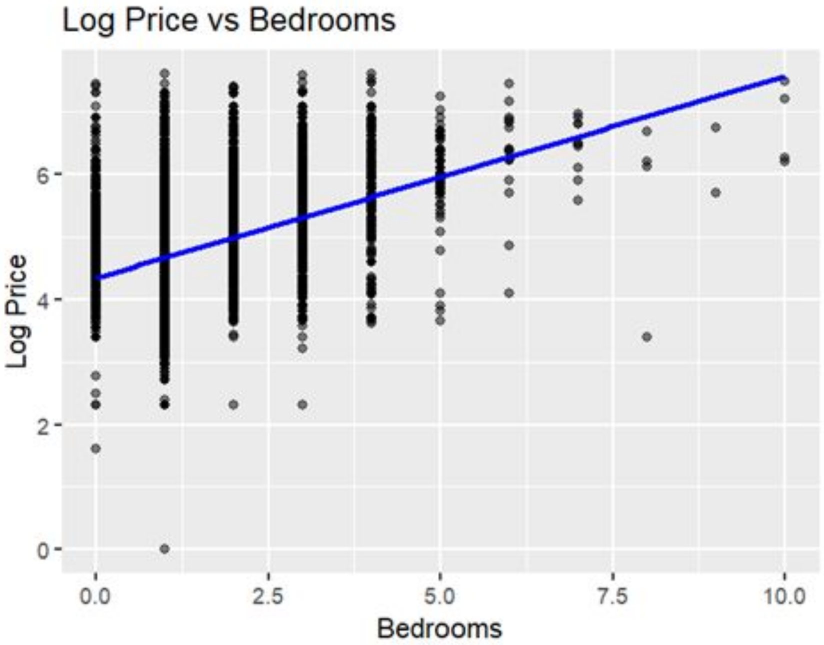


	mean_value	median_value	sd_value	min_value	max_value
1	4.71934	4.65396	0.6615672	0	7.600402

The summary statistics for log_price in the NYC dataset reveal key characteristics of the price distribution. The mean log price is 4.719, while the median is 4.653, indicating that the distribution is slightly right-skewed, as the mean is slightly higher than the median. The standard deviation (0.66) suggests moderate variability in property prices.

The minimum log price is 0, which is unusual and may indicate missing or incorrect data entries that need further investigation. On the other hand, the maximum log price is 7.60, showing the upper bound of property prices in the dataset.

Visualization



The scatter plot shows a positive correlation between log price and number of bedrooms, more bedrooms generally mean higher prices. A linear trend line supports this, but variability and outliers suggest that other factors also impact pricing.

Prediction:

Goal: Predicting Airbnb listing prices using a linear regression model.

Process:

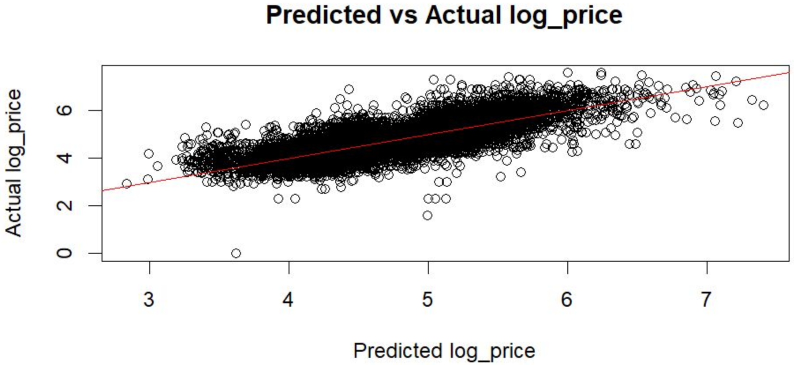
- Selected relevant variables: bedrooms, bathrooms, accommodates, reviews, ratings, location, room type, cancellation policy.
- Removed irrelevant info (like listing ID, name, and description).
- Converted categorical variables (room type, cancellation policy).
- Removed insignificant variables using p-value > 0.05 and checked multicollinearity with VIF.

Key Coefficients Insight:

- 📈 More bedrooms/bathrooms/accommodates = higher price.
- 📉 Private/Shared rooms = lower price vs. entire home.
- 📍 Latitude ↑ / Longitude ↓ → Higher prices in certain areas.
- ✅ Some strict cancellation policies → Higher prices.

Model Performance:

- **R-squared:** 0.645 → Model explains **64.5%** of price variability.
- **RMSE:** 0.387 → Predictions are off by ~0.39 (log scale).



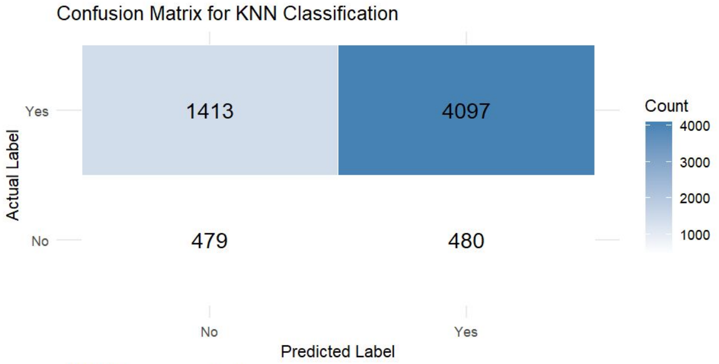
Classification: KNN

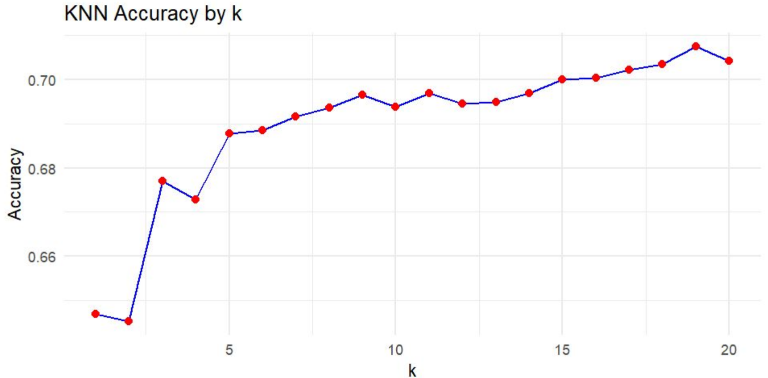
Goal: Predicting if a listing charges a **cleaning fee** (Yes/No) using **K-Nearest Neighbors**.

Process:

- Converted `cleaning_fee` from True/False to Yes/No.
- Selected numeric features + room type (one-hot encoded).
- Scaled features & split data (80/20 train-test).
- Tested `k = 1 to 20` → Best **k = 19**

Accuracy Rate: 70.7%

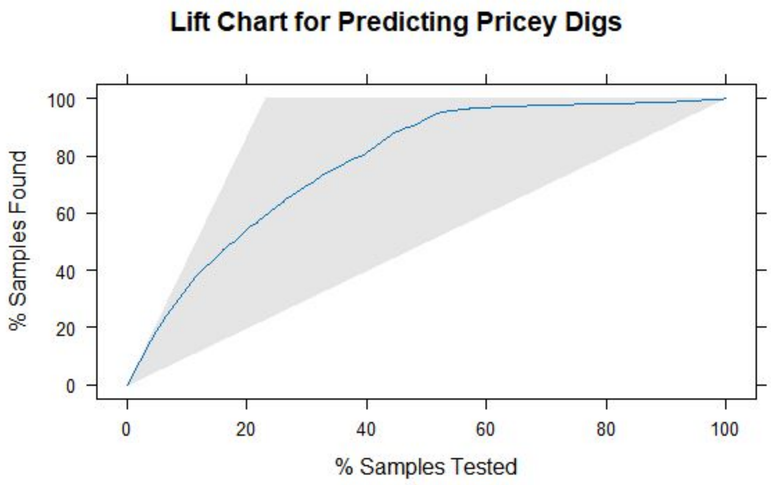




Classification: Naive Bayes

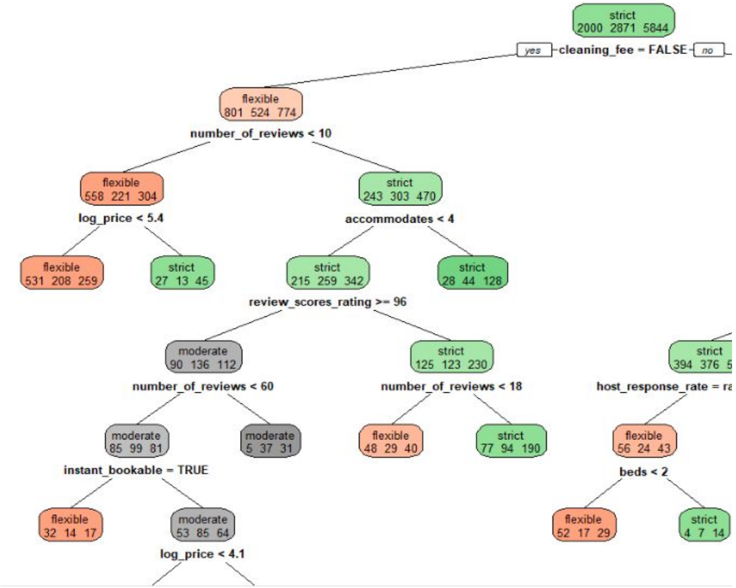
Applied Naive Bayes classification model to classify prices of airbnb listings:

- Levels: student budget, below average, above average, pricey digs
- Predictors: room type, accommodates, bathrooms, bedrooms, cancellation policy
- Model performs better than random guessing



Accuracy Rate: 52%

Classification Tree



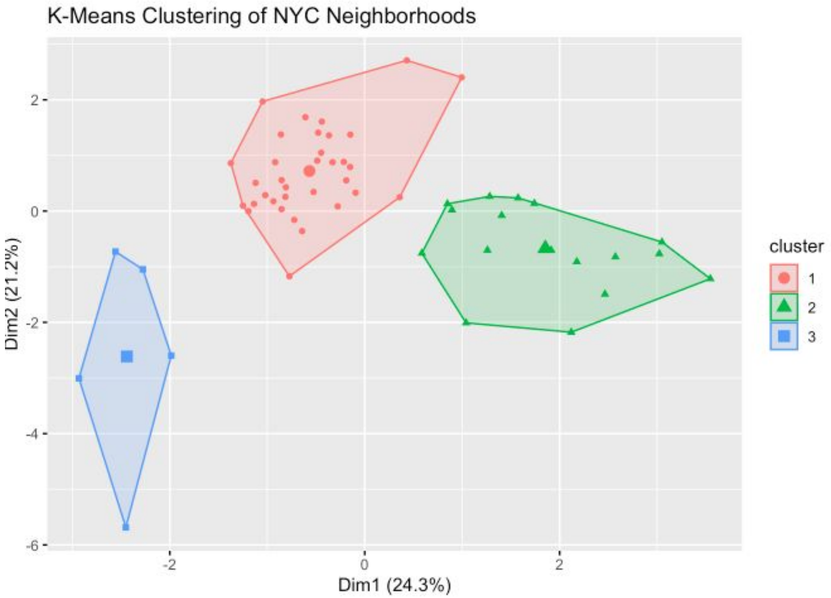
Applied a Classification Tree model to predict cancellation policy categories: strict, moderate, flexible .

Recommendations:

- Utilize model to identify factors that drive host decisions.
- Help users identify listings faster and easier that suit their needs.

Accuracy Rate: 57%

Clustering



Applied K-Means Clustering on engineered features at the neighborhood level:

- Pricing: Average listing price
- Demand: Average number of reviews
- Customer Experience: Review scores
- Operational Traits: Bedroom-to-guest ratio, cleaning fee presence.
- Market Density: Listings per neighborhood

Segment	Characteristics
Cluster 1 – *High-Demand Urban Core*	Mid-priced, highest reviews, largest volume
Cluster 2 – *Premium Markets*	High price, low volume, highest cleaning fee ratio
Cluster 3 – *Budget-Friendly Zones*	Lowest price, moderate reviews, higher space per guest

Recommendations

Clustering: Leverage cluster profiles to tailor pricing strategies, refine amenity offerings, and prioritize investment in high-growth neighborhoods.

Segment	Strategic Recommendation
Cluster 1 – *High-Demand Urban Core*	Optimize yield through dynamic pricing & service consistency
Cluster 2 – *Premium Markets*	Focus on luxury positioning & premium experience
Cluster 3 – *Budget-Friendly Zones*	Promote value offers & extended stays

Conclusion

In this project, we analyzed NYC Airbnb listings using regression, classification, and clustering to understand price drivers, predict pricing, and segment the market.

The multiple linear regression model was the most effective for predicting log_price, explaining 65% of the variation ($R^2 = 0.645$, RMSE = 0.387) using key variables like bedrooms, bathrooms, review scores, and room type.

Classification models like Naive Bayes and decision trees were less accurate (52% and 57%) due to assumptions and difficulties handling imbalanced categories.

The KNN model for cleaning fee classification had 70.74% accuracy but performed poorly for the minority class ("no"), making it unreliable for unbalanced data.

K-means clustering revealed three distinct neighborhood groups—luxury, high-demand, and budget-friendly—offering valuable market-level insights for strategic planning.

Overall, multiple regression is the best model for analyzing and predicting Airbnb listing prices in NYC due to its accuracy and interpretability.