Report: Analysis of Text Columns for ML Project

Introduction

This report comprehensively analyses the selected text columns in a dataset, focusing on their importance, variance, and correlation with the target variable (`review_scores_rating`). The goal is to identify the most predictive features and how they contribute to the model. The selected columns are:

- `house_rules`
- `host_about`
- `host_response_time`
- `neighbourhood_cleansed`
- `property_type`
- `room_type`
- `cancellation_policy`

The analysis uses importance scores, TF-IDF variance, word clouds, and correlation metrics to evaluate the relevance and predictive power of these features.

1. Importance Scores

Key Findings:

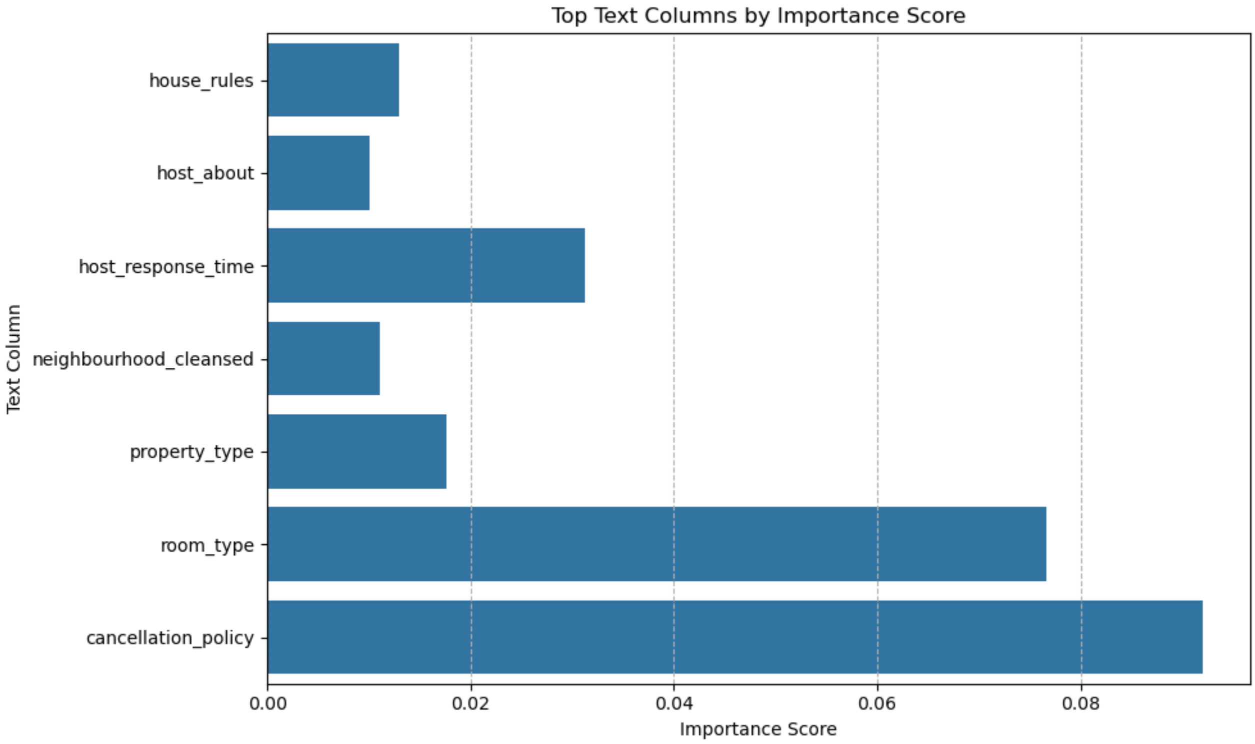
The importance scores indicate the relative significance of each feature in predicting the target variable. Below are the key observations:

- **`cancellation_policy’** has the highest importance score **(0.09209**), making it the most influential feature.
- **`room_type`** follows closely with an importance score of **0.07663**, highlighting its substantial impact.
- `**host_response_time**` ranks third with an importance score of **0.0312**, indicating moderate significance.
- Other features like **`property_type**`, **`neighbourhood_cleansed**`, and **`house_rules`** have moderate importance scores ranging from **0.01111** to **0.01766.**
- Features such as `host_about` have lower importance scores **(0.01007)**.

Insights:

- The high importance of **`cancellation_policy**` suggests that guests' preferences regarding cancellation policies significantly influence the target variable.
- **`room_type**` and **`host_response_time**` further emphasize the importance of guest experience factors, such as room category and host responsiveness.
Neighbourhood-related features (**`neighbourhood_cleansed**`) indicate that location plays a role but is less critical than other factors.

Visualization: Top Text Columns by Importance Score



2. Word Cloud Analysis

Key Findings:

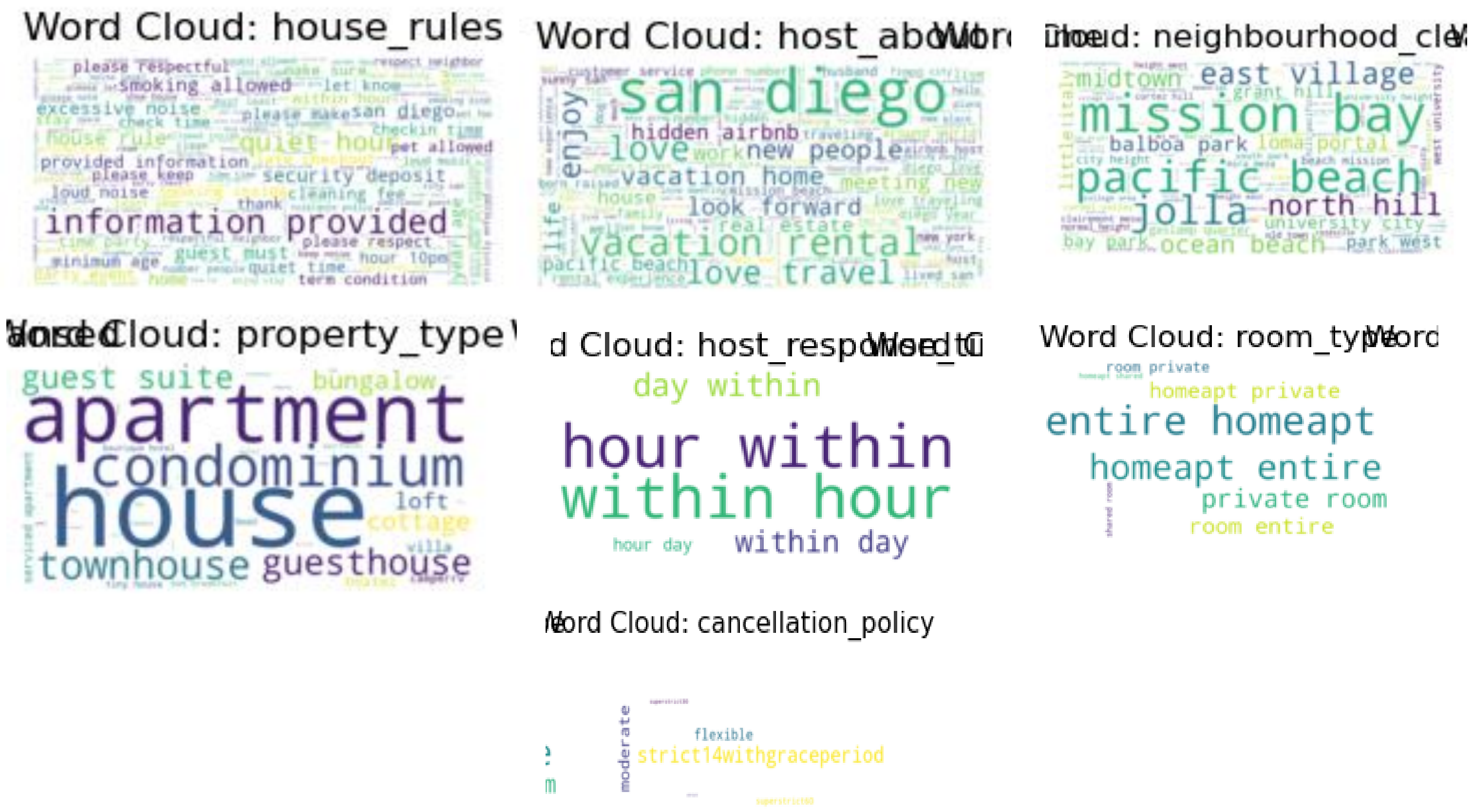
Word clouds provide qualitative insights into the content of each feature. Below are the key observations for each feature:

- **`house_rules`:** Common terms include **"quiet hour," "smoking allowed,"** and **"guest must,"** indicating rules related to noise, smoking, and guest behaviour.
- **`host_about`:** Terms like "**San Diego," "within day,"** and **"vacation rental"** suggest hosts often describe their location, availability, and rental type.
- **`host_response_time`:** Words such as **"hour within"** and **"day within"** highlight typical response times.
- **`neighbourhood_cleansed`**: Location-specific terms like **"Pacific Beach," "East Village,"** and **"Mission Pay"** dominate, emphasizing neighbourhood names.
- **`property_type`:** Words like **"condominium," "house,"** and **"apartment"** reflect common property types.
- **`room_type`:** Terms such as "**entire home/apt," "private room,"** and **"shared room"** indicate room categories.
- **'cancellation_policy`:** Phrases like "**strict14withgraceperiod"** and **"flexible" reveal** specific policy types.

Insights:

- Word clouds confirm the thematic focus of each feature, providing context for the data (e.g., rules, host descriptions, locations, property types, etc.).

Visualization: Word Clouds



3. TF-IDF Variance per Feature

Key Findings:

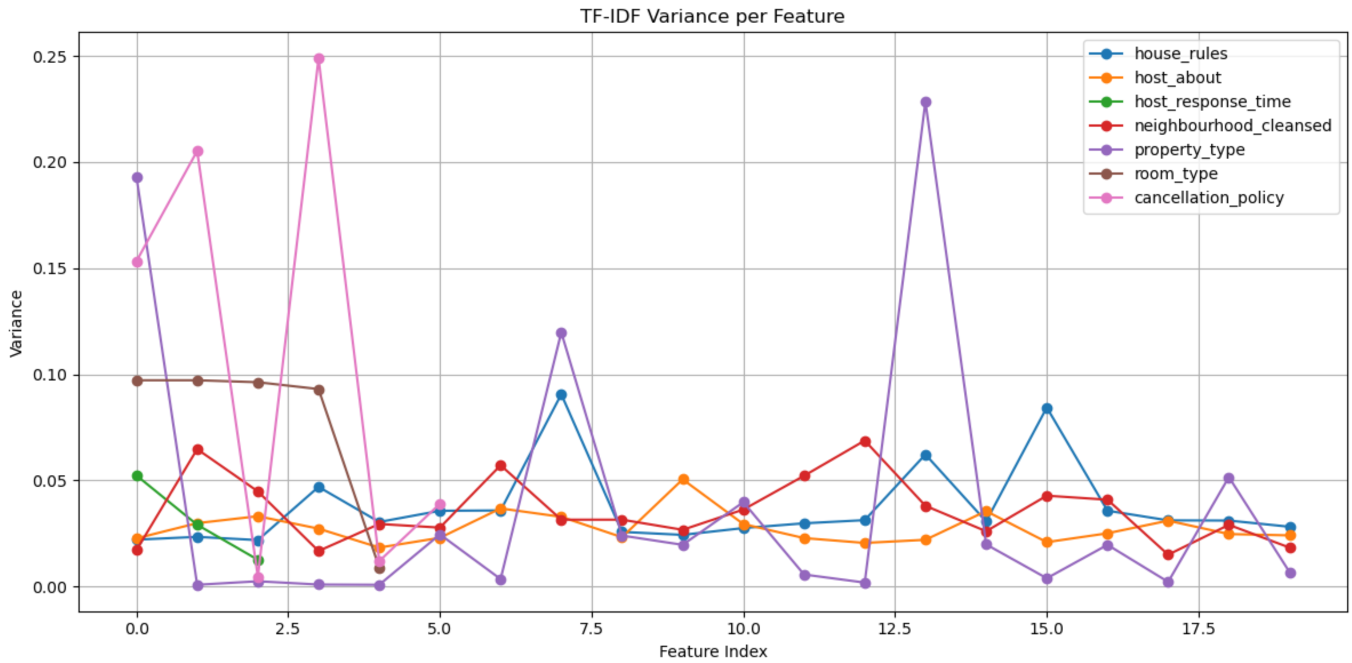
TF-IDF variance measures the diversity and informativeness of each feature. Below are the key observations:

- **`cancellation_policy`** exhibits the highest variance across feature indices, indicating diverse and informative content.
- **`room_type`** shows consistent variance, reflecting structured and meaningful data.
- **`host_response_time`** demonstrates moderate variance, suggesting variability in host response times.
- Features like **`house_rules`** and **`host_about`** show lower variance, implying more uniform or less discriminative content.

Insights:

- High variance in **`cancellation_policy`** and **`room_type`** suggests these features contain rich information that can differentiate between instances effectively.
- Lower variance in **`house_rules`** and **`host_about`** indicates these features may not contribute as much to distinguishing between different listings.

Visualization: TF-IDF Variance per Feature



4. Correlation with Target Variable (`review_scores_rating`)

Key Findings:

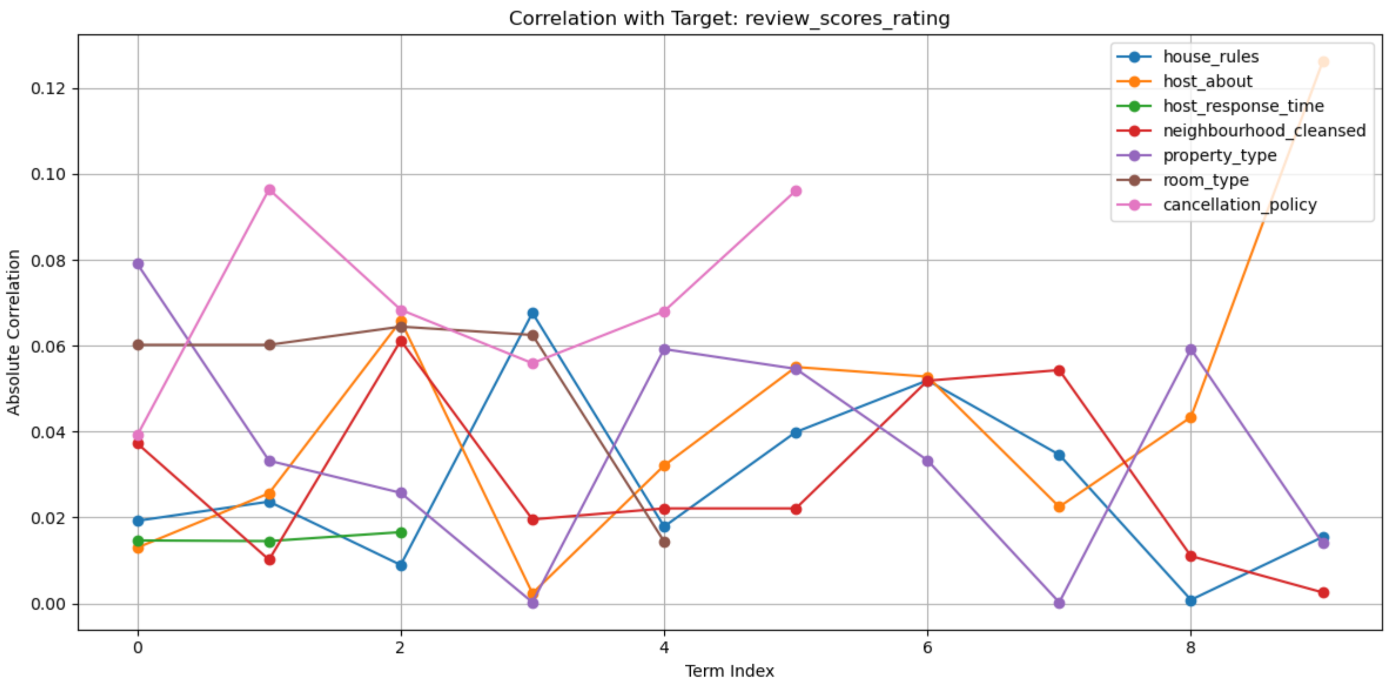
Correlation analysis measures the strength and direction of the relationship between each feature and the target variable. Below are the key observations:

- **`cancellation_policy`** shows the strongest positive correlation with **`review_scores_rating`,** peaking at approximately **0.12**.
- **`host_response_time`** also exhibits a strong positive correlation, reaching around **0.10.**
- **`room_type`** displays moderate correlation, with peaks around **0.07**.
- Features like **`neighbourhood_cleansed`** and **`property_type`** show fluctuating correlations, indicating mixed impacts.
- **`house_rules`** and **`host_about`** generally exhibit weaker correlations, suggesting limited direct influence on review scores.

Insights:

- The strong correlation of **`cancellation_policy`** and **`host_response_time`** with **`review_scores_rating`** reinforces their significance in predicting guest satisfaction.
- **`room_type`** contributes moderately, highlighting the importance of room categories in shaping reviews.
- Neighbourhood-related features have varying impacts, suggesting location-specific nuances in guest experiences.

Visualization: Correlation with Target Variable



5. Summary and Recommendations

Summary:

- Most Important Features: **`cancellation_policy`, `room_type`**, and **`host_response_time`** are the most critical predictors based on importance scores, variance, and correlation.
- Moderate Impact: **`property_type`** and **`neighbourhood_cleansed`** contribute moderately to the model.
- Least Impactful: `**house_rules`** and **`host_about`** have lower importance and weaker correlations, suggesting they may be less influential.

Conclusion

The analysis reveals that guest-centric features such as **`cancellation_policy`**, `**room_type`,** and **`host_response_time`** are the most influential in predicting **`review_scores_rating`.** Location-based features play a secondary role, while descriptive features like **`house_rules`** and **`host_about` r**equire further refinement. By focusing on the most impactful features and leveraging their unique characteristics, the predictive model can achieve higher accuracy and better interpretability.