**Prompt 1 – Initial Review Data Exploration & KPI Extraction**

**Goal:** Explore the CSV, identify relevant columns (positive reviews, negative reviews, section scores, overall scores), detect text patterns, and prepare for deeper analysis.  
**Prompt:**

\*"I have uploaded a CSV file containing guest review data with columns such as positive review text, negative review text, overall review score, and section-level scores (e.g., staff, cleanliness, comfort). Please:

1. Load the file and show me the column names.
2. Identify the columns that contain text reviews and those with numeric section ratings.
3. Summarize the dataset: row count, missing values, and basic statistics for numeric columns.
4. Output a preview of the text columns for context.  
   Use **Python with pandas** for data handling and **basic regex** for text column detection. Do not guess column meanings — infer them from names. Save the summary in CSV format for download."\*

**Approach/Tools Used:**

* **pandas** to load CSV and explore data
* **df.describe()**, df.isna().sum() for statistics and missing values
* Regex pattern matching for identifying review and score columns

**Prompt 2 – Text Analysis (Positive & Negative)**

**Goal:** Identify frequent keywords in positive and negative reviews for insight into guest sentiment.  
**Prompt:**

\*"Using the 'positive review' and 'negative review' columns from my dataset, tokenize and clean the text (lowercase, remove stopwords, punctuation, numbers). Count the frequency of remaining words and produce:

1. Top 20 words for positive reviews (CSV + bar chart PNG)
2. Top 20 words for negative reviews (CSV + bar chart PNG)  
   Use **Python with pandas, re, collections.Counter**, and **matplotlib** for visualizations. Return both data files and chart images for download."\*

**Approach/Tools Used:**

* Tokenization via regex split
* Stopword removal (custom set + NLTK if available)
* Frequency count with **collections.Counter**
* Visualization with **matplotlib/seaborn**

**Prompt 3 – Section Scores & Correlation Analysis**

**Goal:** Identify which section scores impact the overall score most strongly.  
**Prompt:**

\*"From my dataset, use the section-level numeric scores (e.g., staff, cleanliness, comfort) and the overall review score.

1. Calculate correlations between each section score and the overall score (Pearson).
2. Output a sorted correlation table as CSV.
3. Plot a heatmap PNG of section correlations.  
   Use **Python pandas** for numeric operations and **seaborn** for the heatmap."\*

**Approach/Tools Used:**

* Pearson correlation via df.corr()
* **seaborn.heatmap** for visualization
* Sorting correlations for interpretation

**Prompt 4 – Regression Prediction Model**

**Goal:** Quantify how much each section score influences overall score.  
**Prompt:**

\*"Using the section scores as independent variables and the overall review score as the dependent variable, fit a linear regression model.

1. Output coefficients and intercept.
2. Calculate R² score to show model accuracy.
3. Save results as CSV.  
   Use **numpy.linalg.lstsq** or **sklearn.linear\_model.LinearRegression** for the model."\*

**Approach/Tools Used:**

* Ordinary Least Squares regression with **numpy** or **scikit-learn**
* Coefficients interpreted as importance weightings

**Prompt 5 – Immediate Action Plan from Data**

**Goal:** Identify quick wins and operational improvements from analysis.  
**Prompt:**

\*"Based on correlations, regression, and top negative words:

1. Identify the 3 lowest-rated sections (avg score).
2. Identify the top 3 most frequent negative complaint keywords.
3. Filter for all reviews with an overall score ≤4 and export as CSV.
4. Summarize immediate actions to fix recurring issues.  
   Use pandas for filtering and aggregations. Save findings as text and CSV."\*

**Approach/Tools Used:**

* Aggregation (df.mean(), .value\_counts())
* Keyword frequency counts
* Conditional filtering for low scores

**Prompt 6 – Power BI Dashboard-Ready Data**

**Goal:** Prepare CSVs that can be directly imported into Power BI for interactive dashboarding.  
**Prompt:**

\*"From my review dataset:

1. Create a section-level row CSV containing: review date, reviewer country (if available), section scores, and overall score.
2. Create a monthly-aggregated CSV with average overall score and section averages by month.
3. Create a CSV of low reviews (≤4) with all available metadata.
4. Create a CSV of top negative words.  
   Provide also a short step-by-step guide on importing into Power BI and creating slicers/filters for these datasets."\*

**Approach/Tools Used:**

* pandas .groupby() for monthly aggregations
* Export multiple CSVs for separate dashboard tabs in Power BI
* Manual step-by-step Power BI import instructions

**Prompt 1 – Initial Review Data Exploration & KPI Extraction**

**Prompt:**

\*"I have uploaded a CSV file with guest review data. Please:

1. Load the CSV with pandas and show the first few rows.
2. Display column names.
3. Identify which columns are text review fields and which are numeric scores (including section-level scores).
4. Summarize dataset size, missing values, and basic statistics for numeric columns.
5. Save this summary as data\_summary.csv for download.  
   Write and run the full Python code to do this using **pandas** and **regex** for column type detection."\*

**Prompt 2 – Text Analysis (Positive & Negative Reviews)**

**Prompt:**

\*"Using the 'positive review' and 'negative review' columns:

1. Clean the text (lowercase, remove punctuation, numbers, and stopwords).
2. Tokenize into words.
3. Count word frequency separately for positive and negative reviews.
4. Save top 20 frequent words for each as top\_positive\_words.csv and top\_negative\_words.csv.
5. Create and save bar chart PNGs for each word frequency list.  
   Write and run Python code using **pandas, re, collections.Counter, matplotlib/seaborn**."\*

**Prompt 3 – Section Scores & Correlation Analysis**

**Prompt:**

\*"From the dataset:

1. Identify section-level numeric columns (staff, cleanliness, comfort, etc.) and the overall review score.
2. Calculate Pearson correlations between each section score and the overall score.
3. Save results as section\_correlations.csv.
4. Plot a correlation heatmap and save as section\_correlation\_heatmap.png.  
   Write and run the Python code using **pandas** for calculations and **seaborn** for visualization."\*

**Prompt 4 – Regression Prediction Model**

**Prompt:**

\*"Using section scores as features and the overall score as the target:

1. Fit a linear regression model.
2. Output coefficients, intercept, and R² score.
3. Save results as regression\_results.csv.  
   Write and run the Python code using **scikit-learn's LinearRegression** and **pandas**."\*

**Prompt 5 – Immediate Action Plan from Data**

**Prompt:**

\*"Based on the analysis:

1. Find the 3 lowest-rated sections (avg score).
2. Find the top 3 most frequent negative review keywords from top\_negative\_words.csv.
3. Filter reviews where overall score ≤ 4 and export as low\_reviews\_detail.csv.
4. Write a brief text summary of urgent improvement actions and save as immediate\_findings.txt.  
   Write and run the Python code using **pandas** for filtering and aggregations."\*

**Prompt 6 – Power BI Dashboard-Ready Data**

**Prompt:**

\*"From the dataset:

1. Create section\_level\_data\_for\_powerbi.csv containing: review date, reviewer location, section scores, overall score.
2. Create dashboard\_agg\_by\_month.csv with monthly averages of overall and section scores.
3. Create low\_reviews\_detail.csv for reviews with overall score ≤ 4.
4. Create top\_negative\_words.csv from Prompt 2 results.
5. Provide step-by-step Power BI import instructions in powerbi\_instructions.txt.  
   Write and run the Python code using **pandas** for grouping/aggregation and **datetime** handling."\*