DAV Project Phase 4: Correlation Report

Course: Data Analysis and Visualization (CS-352)
Phase Lead: Usaira Shahbaz (NUM-BSCS-2022-27)
Group Members:
Ahmad Mustafa (NUM-BSCS-2022-37)
Zaheer Abbas (NUM-BSCS-2022-50)
Department: Computer Science, Namal University, Mianwali

April 2025

Phase 4: Correlation

1 Sentiment Analysis Introduction (Before Phase 4)

1.1 Why We Did Sentiment Analysis

The dataset contains a column called `review_comment_message`, where customers provide feedback. However, text data cannot be used directly in numerical analysis. Thus, we converted textual reviews into numerical sentiment scores:

- Positive reviews → Higher sentiment score (close to +1)
- Negative reviews → Lower sentiment score (close to -1)

Benefits of Sentiment Analysis:

- Quantifies customer emotions, making them measurable.
- Helps predict trends (e.g., positive reviews linked to higher ratings).
- Supports business decisions through customer satisfaction insights.
- Allows comparison between textual feedback and numerical features like price, payment value, and review scores.

Thus, sentiment analysis bridged the gap between qualitative (text) and quantitative (numbers) data.

1.2 Which Library Was Used: NLTK and VADER

1.2.1 NLTK (Natural Language Toolkit)

- A popular Python library for text processing and linguistic analysis.
- Provides tools for tokenization, stemming, tagging, and sentiment analysis.

1.2.2 VADER (Valence Aware Dictionary and sEntiment Reasoner)

- A specialized tool inside NLTK designed for short texts like reviews and comments.

Why We Chose VADER:

- Handles slang, emojis, capitalization, and punctuation well.
- Fast and does not require model training.
- Ready-to-use for English language text.

How VADER Works:

- **Lexicon Lookup:** Uses a dictionary (`vader_lexicon`) containing sentiment scores for words.
- **Text Parsing:** Identifies words and looks up their scores.
- **Intensity Adjustment:** Modifiers like "very", "extremely", and punctuation affect intensity.
- **Score Aggregation:** Outputs:
  - Positive Score
  - Neutral Score
  - Negative Score
  - Compound Score (summary between -1 and +1)

We mainly used the **compound score** for correlation analysis as it summarizes overall sentiment.

2 Phase 4: Correlation Analysis

2.1 Correlation Matrix

We calculated the Pearson correlation coefficients between important numerical features:

Feature 1	Feature 2	Correlation Coefficient
price	price	1.000
price	payment_value	0.645
price	review_score	0.0208
price	sentiment_polarity	0.0260
payment_value	payment_value	1.000
payment_value	review_score	-0.1098
payment_value	sentiment_polarity	-0.0628
review_score	review_score	1.000
review_score	sentiment_polarity	0.6006
sentiment_polarity	sentiment_polarity	1.000

2.2 Heatmap Visualization

(Insert heatmap visualization here.)

The heatmap visually shows stronger correlations with darker colors and weaker correlations with lighter colors.

2.3 Significant Correlations and Predictions

Feature 1	Feature 2	Correlation	Interpretation / Prediction
price	payment_value	0.645 (Strong Positive)	As price increases, payment value also increases.
review_score	sentiment_polarity	0.6006 (Strong Positive)	Positive comments are associated with higher review scores.
payment_value	review_score	-0.1098 (Weak Negative)	Slight negative relation: higher payments may cause slight dissatisfaction.
payment_value	sentiment_polarity	-0.0628 (Very Weak Negative)	Almost no relationship between payment value and comment sentiment.
price	review_score	0.0208 (Very Weak Positive)	Price does not significantly impact review scores.
price	sentiment_polarity	0.0260 (Very Weak Positive)	Price does not significantly impact comment sentiment.

3 Summary of Interpretation

- **Strong Positive:**
  - Price and Payment Value (0.645): More expensive products → Higher payments.
  - Review Score and Sentiment Polarity (0.6006): Happier comments → Higher ratings.
- **Very Weak Correlation:**
  - Price and Review Score / Sentiment Polarity: Price has little effect on review quality.
- **Weak Negative Correlation:**
  - Payment Value and Review Score: Slight dissatisfaction for higher payments.