1 Introduction

Factor Analysis (FA) is a statistical technique used to uncover latent structures—referred to as **factors**—that explain correlations among observed numerical variables. By reducing the dimensionality of a dataset, FA highlights hidden relationships, simplifies complex data, and supports more efficient downstream analytics.

In this project, Factor Analysis was applied to the cleaned dataset that merged and stored at merged data. The resulting analytical outputs were stored in the **Gold layer**, following the project's **Medallion Architecture** (Bronze → Silver → Gold), ensuring reliable, curated, and high-quality data for further analysis.

2 Methodology

2.1 Load Cleaned Data

The cleaned Silver-layer dataset was loaded for statistical analysis.

2.2 Select Numeric Variables

Only numerical features were retained, as Factor Analysis requires continuous numeric inputs.

2.3 Standardize the Data

Standardization transformed all numeric features to have zero mean and unit variance, ensuring each variable contributes equally to the factor extraction process.

2.4 Apply Factor Analysis

A Factor Analysis model was configured to extract three latent factors. The model decomposed the standardized dataset into:

- **Factor Scores**: Representation of each observation in terms of latent factors.
- **Factor Loadings**: Contribution strength of each variable to each factor.

2.5 Save Outputs

Factor scores and loadings were saved to the Gold layer for downstream analytics.

3 Results

3.1 Factor Scores

- **File:** `gold/factor_analysis_scores.parquet`
- **Description:** Numerical representation of each observation in terms of the three extracted factors.

3.2 Factor Loadings

- **File:** `gold/factor_analysis_loadings.parquet`
- **Description:** Contribution strength of each variable to each factor.

4 Visualizations

4.1 Factor Loadings Heatmap

Other variables (humidity, wind speed, rain, visibility) have low loadings, indicating minor influence.

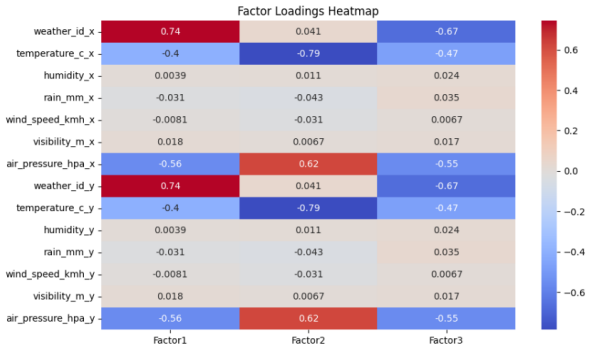

Figure 1: HeatMap

Figure 2: Factor Loadings Heatmap

4.2 Scatter Plots of Factor Scores

4.2.1 Factor 1 vs Factor 2

Points are dispersed, indicating independence between these factors. Useful for clustering and dimensionality reduction.

Figure 3: F1 Vs F2

Figure 4: Factor 1 vs Factor 2 Scatter Plot

4.2.2 Factor 2 vs Factor 3

Points show a strong linear trend, indicating high correlation. Suggests potential redundancy; these factors may represent overlapping constructs.

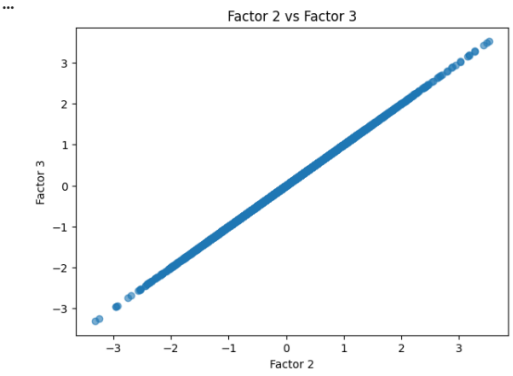

Figure 5: F2 Vs F3

Figure 6: Factor 2 vs Factor 3 Scatter Plot

4.2.3 Factor 1 vs Factor 3

Points are moderately dispersed, indicating partial independence. Factor 1 contributes unique information relative to Factor 3.

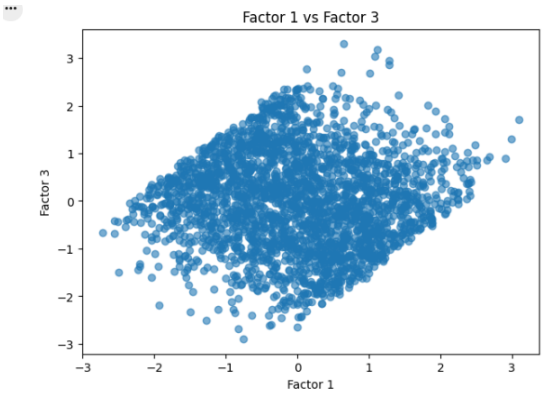

Figure 7: F1 Vs F3

Figure 8: Factor 1 vs Factor 3 Scatter Plot

5 Conclusion

The Factor Analysis revealed three latent factors capturing key patterns in the weather-related dataset:

1. General weather conditions
2. Pressure-temperature dynamics
3. Secondary weather patterns

Scatter plots indicate that Factor 2 and Factor 3 are highly correlated, which may require further consideration in downstream modeling. Factors 1 and 3 provide complementary information, supporting dimensionality reduction and exploratory insights.

All factor scores and loadings are saved in the Gold layer, ensuring availability for predictive modeling, clustering, or further statistical analysis.