**📘 Data Dictionary – Broker Performance Dataset**

**1. Broker\_ID**

* **Type**: String/ID
* **Description**: Unique identifier for each broker or agent.
* **Example**: BKR1023

**2. Broker\_Name**

* **Type**: Text
* **Description**: Full name of the broker/agent.
* **Example**: Sarah Johnson

**3. Region**

* **Type**: Categorical
* **Description**: Geographic location where the broker operates.
* **Example**: Ontario, Quebec, British Columbia

**4. Policies\_Sold**

* **Type**: Integer
* **Description**: Total number of insurance policies sold by the broker within the reporting period.
* **Example**: 125

**5. Total\_Premium**

* **Type**: Float (currency)
* **Description**: Total premium amount generated from policies sold.
* **Example**: $45,200.50

**6. Claims\_Filed**

* **Type**: Integer
* **Description**: Number of claims filed by clients under policies sold by the broker.
* **Example**: 12

**7. Claims\_Paid**

* **Type**: Integer
* **Description**: Number of claims that were successfully paid out.
* **Example**: 9

**8. Client\_Satisfaction\_Score**

* **Type**: Integer (1–10 scale)
* **Description**: Average client satisfaction rating for the broker (based on feedback surveys).
* **Example**: 8

**9. Commission\_Earned**

* **Type**: Float (currency)
* **Description**: Total commission earned by the broker, typically a percentage of Total\_Premium.
* **Example**: $7,850.00

**10. Month**

* **Type**: Date (or Text if messy)
* **Description**: Reporting month for the broker’s performance. (Contains some inconsistent/missing entries for cleaning practice.)
* **Example**: 2023-05, May-2023, 05/2023

**🏢 Broker Performance Data Cleaning & Preparation (Power Query)**

* **📌 Project Overview**
* In this project, I worked with a **mock dataset simulating broker performance in an insurance brokerage**. The dataset contained over **150 rows of data** with intentionally introduced errors such as missing values, inconsistent date formats, duplicates, and outliers. The goal of this project was to **clean, transform, and prepare the data** for reliable reporting and dashboard creation.

**🔧 Key Challenges & Cleaning Steps**

1. **Missing & Inconsistent Data**
   * I filled in the missing values in Policies\_Sold and replaced "N/A" values in Total\_Premium with the average as a replacement value
   * I standardized inconsistent date formats in the Month column (e.g., "2023-01", "Jan-2023", "01/2023", "Unknown") into a uniform YYYY-MM format.
2. **Data Validation & Consistency**
   * I checked and corrected cases where Claims\_Paid exceeded Claims\_Filed by adding a custom column to flag cases where business logic is broken.
   * I ensured Commission\_Earned was within a reasonable ratio compared to Total\_Premium to ensure commission earned was not greater than 50% of the Total Premium.
3. **Duplicate & Outlier Handling**
   * I removed duplicate broker records by Broker\_ID, Region, and Month.
   * I flagged brokers with unusually high Policies\_Sold compared to their regional peers to identify outliers and highlight top performers versus typical ranges.
4. **Transformations for Analysis**
   * I created new metrics such as Claims\_Ratio = Claims\_Paid / Claims\_Filed.
   * I aggregated performance data by **Region** and **Month**.
   * I ranked brokers by Total\_Premium and calculated average Client Satisfaction Score per region.

**🔹 Bonus Transformation Tasks**

* I created a new column Claims\_Ratio = Claims\_Paid / Claims\_Filed.
* I aggregated total Policies\_Sold and Total\_Premium **per Region**.
* I Calculated the **average Client Satisfaction Score** per region per month.

**📊 Outcome**

After cleaning, the dataset was transformed into a **structured, analysis-ready format**. The cleaned data provided:

* Reliable insights into broker productivity.
* Regional comparisons of sales and claims performance.
* A foundation for creating dashboards to track top performers and training needs.

👉 This project demonstrates my ability to **identify and resolve real-world data quality issues** using **Power Query**, ensuring that stakeholders can trust the insights drawn from the data.