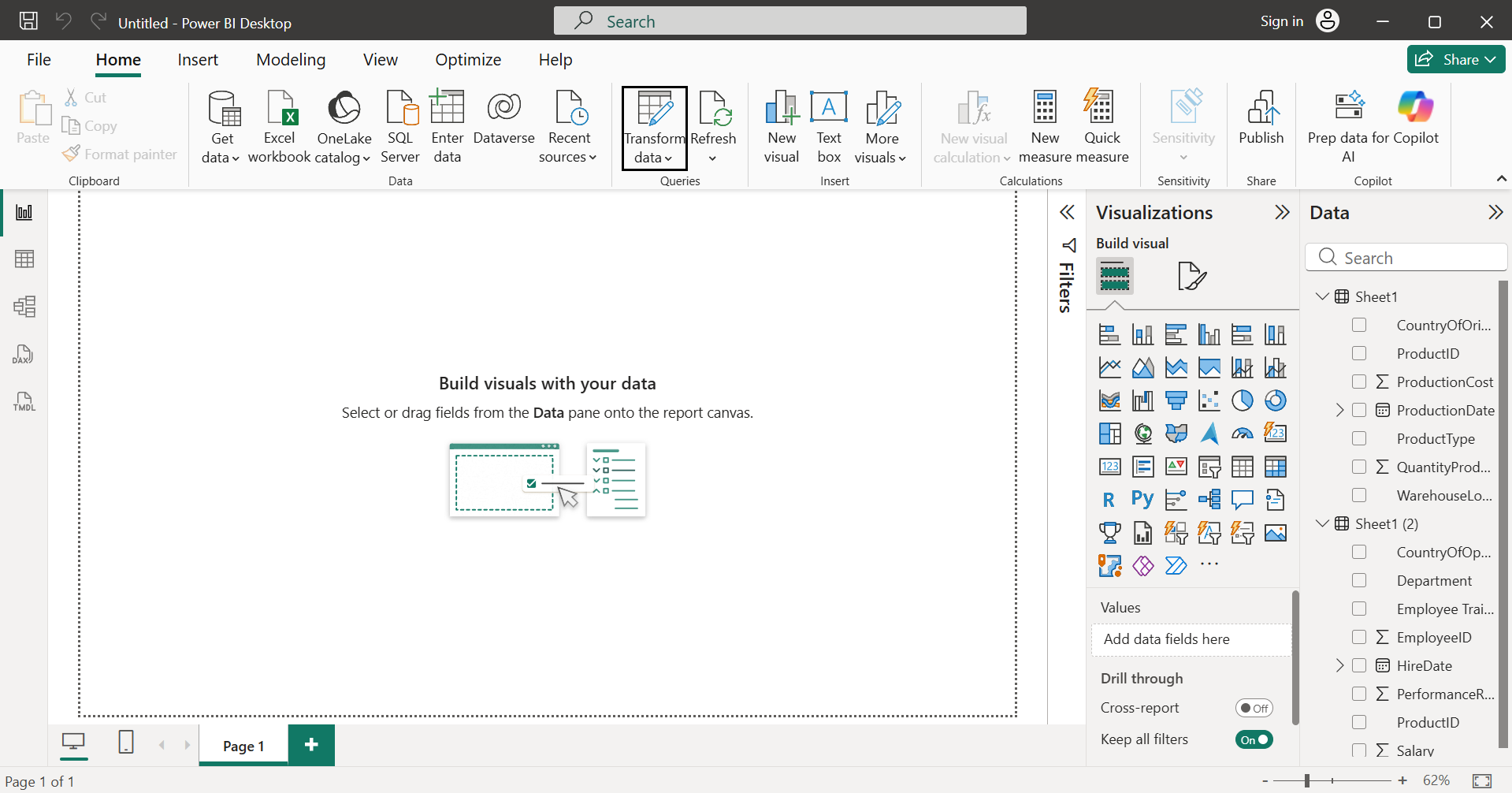
## **Part 1: Data Cleaning, Modeling, and DAX in Power BI**

### **1. Data Importing and Preliminary Examination**

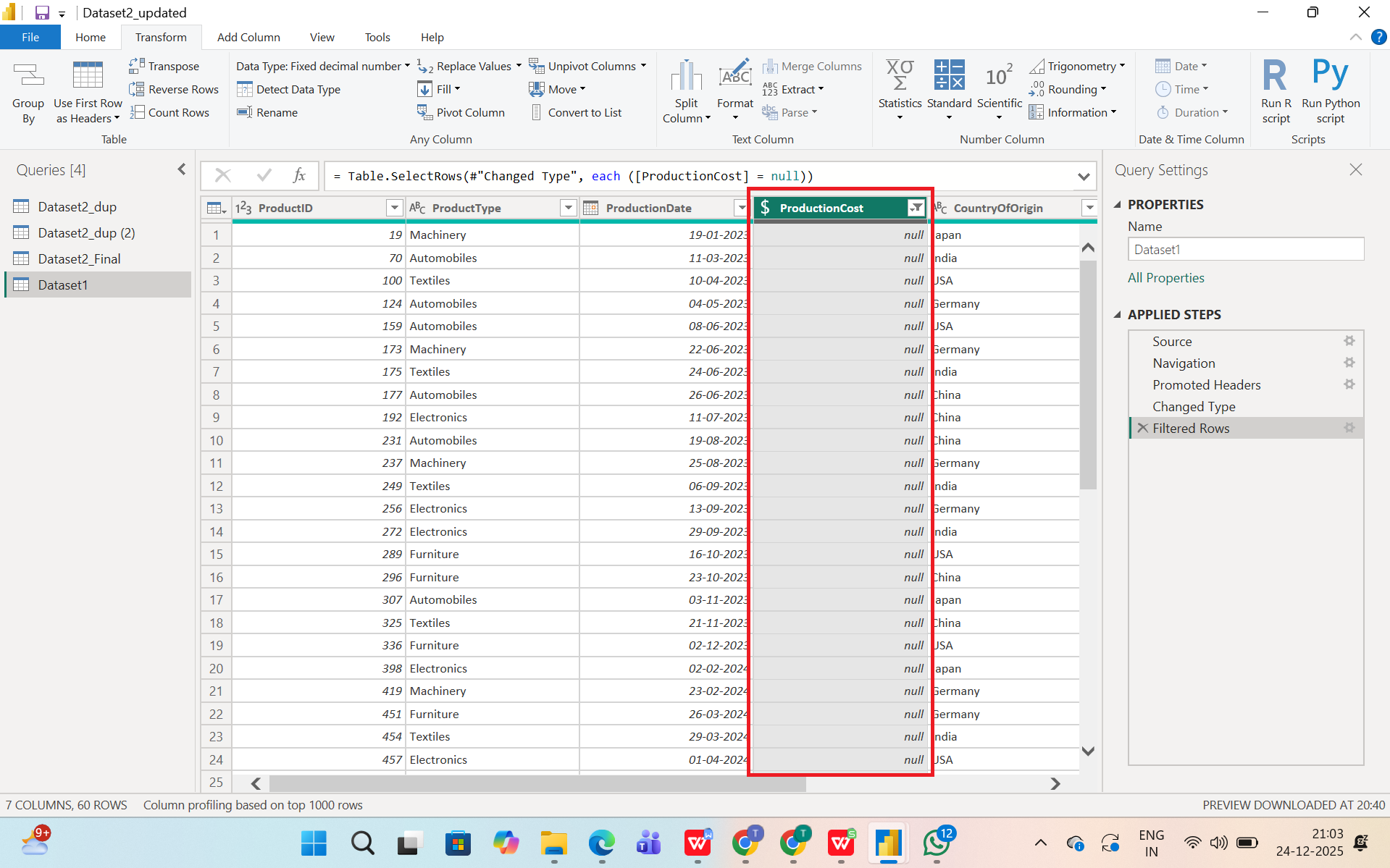
**Import both datasets into Power BI. Perform a preliminary examination of the data. Are there any anomalies or inconsistencies?**

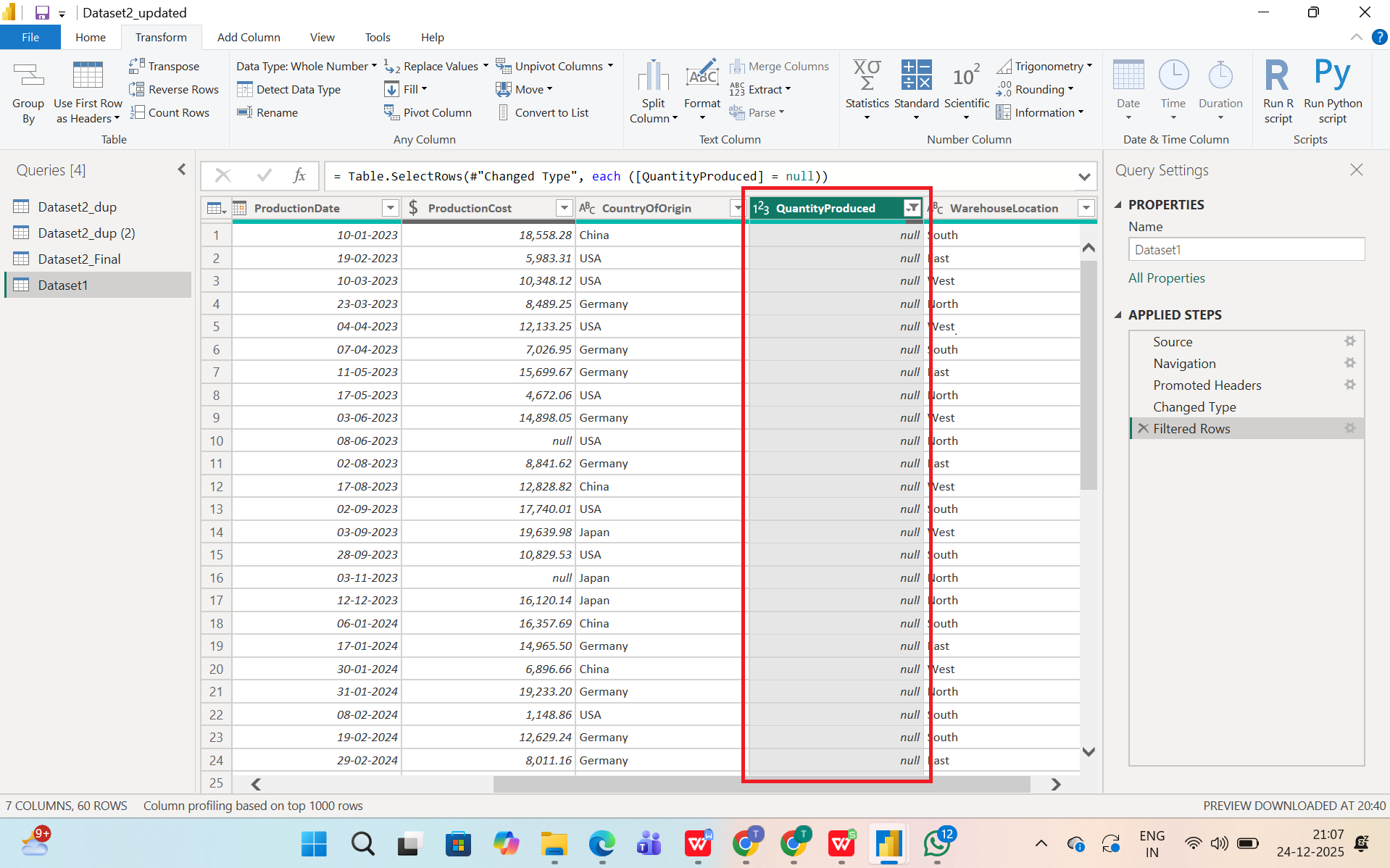
Both Excel files (ManufacturingDataset1 and ManufacturingDataset2) were imported into Power BI using the Excel connector and loaded into Power Query for examination. The production dataset contains anomalies such as **“000000”** values in the ProductionCost column, missing or invalid quantities (null and -999), and incorrect warehouse location codes **(e.g., -999, \*^, blanks)**. The employee dataset shows null values in Salary and PerformanceRating, invalid department codes **(-999, \*^, blanks)**, and a denormalized Employee Training Record column that stores multiple trainings in a single cell. These inconsistencies were identified as data‑quality issues to be addressed in the cleaning phase.​

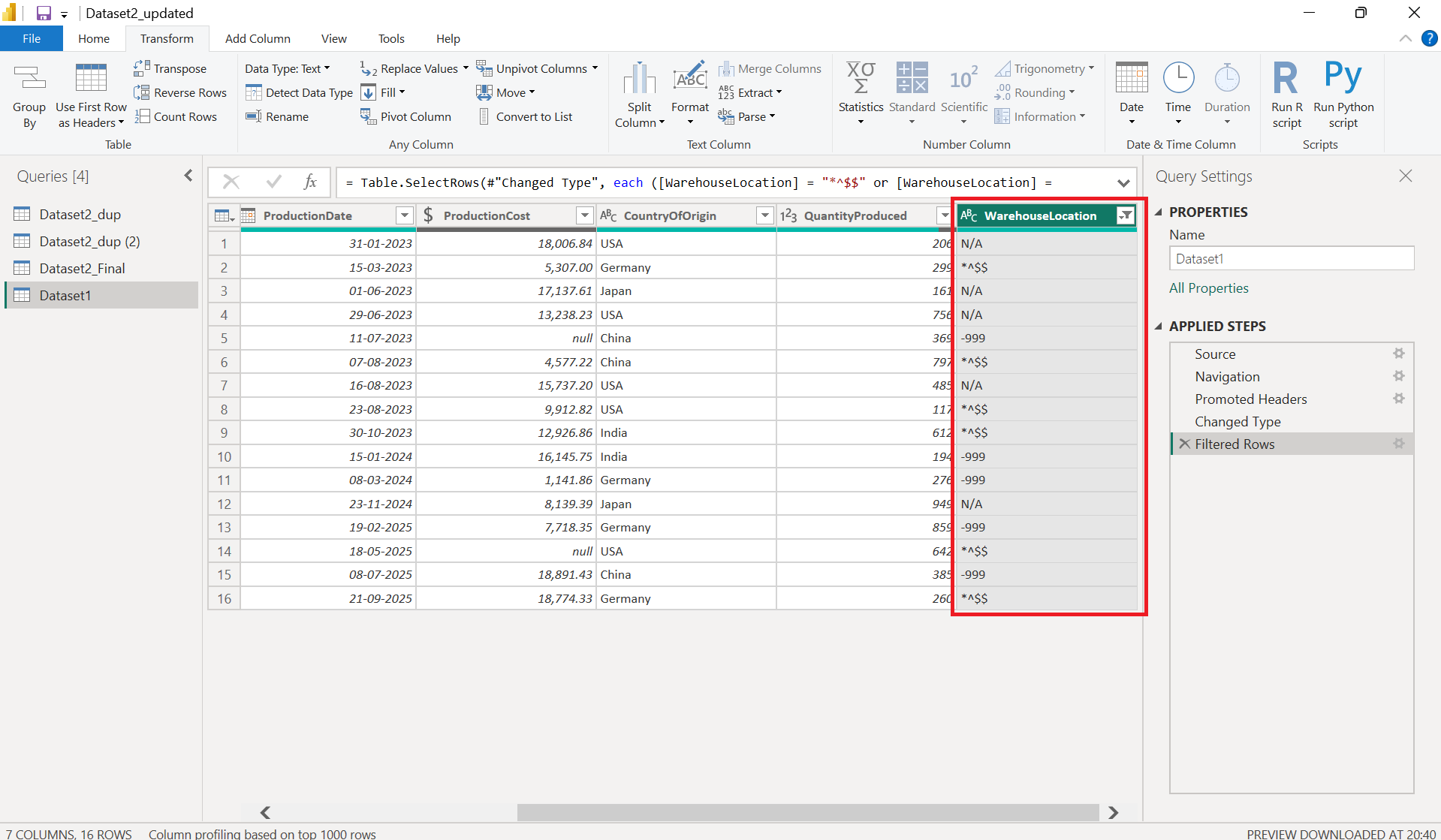
**Data import view:**



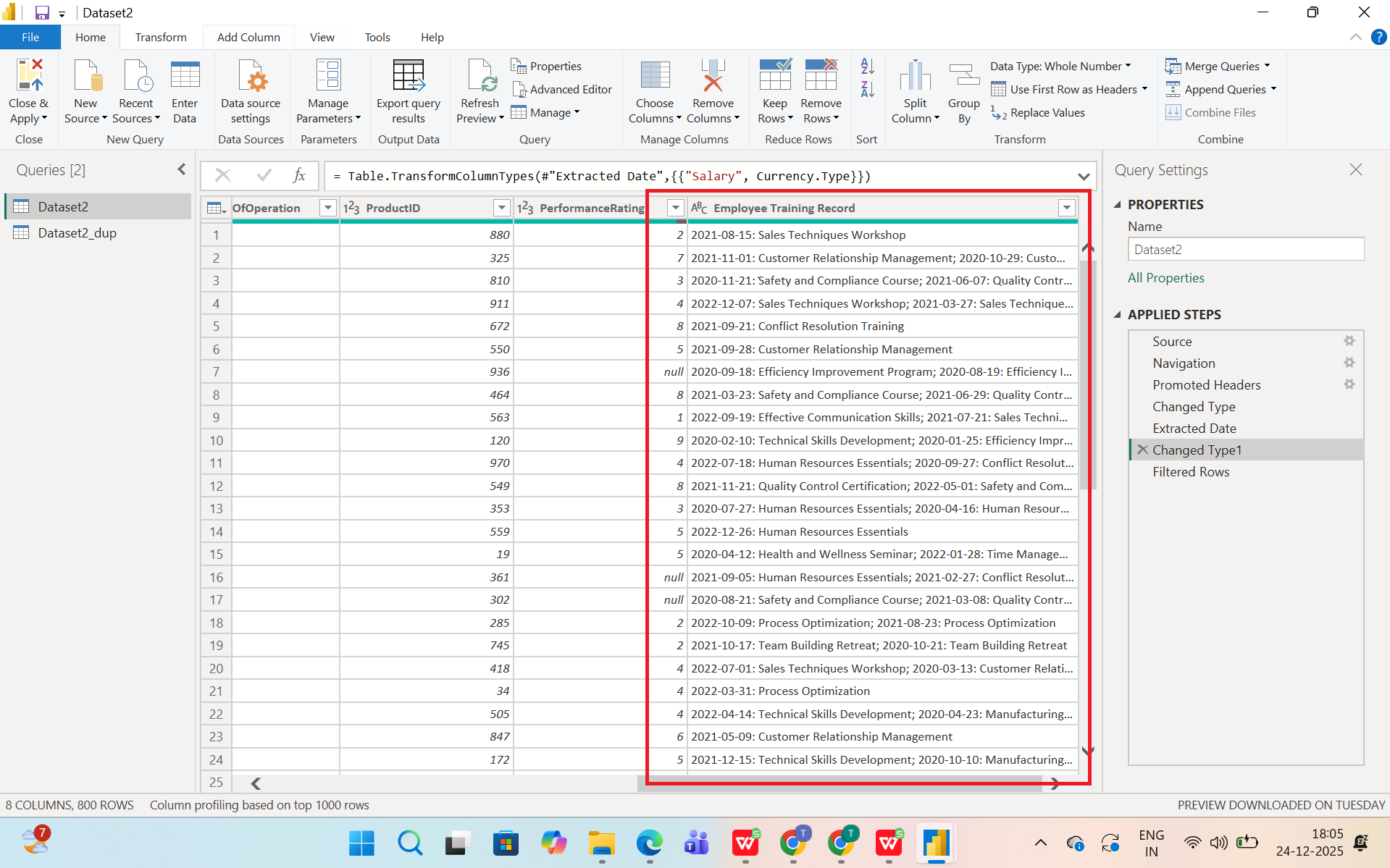
**Production dataset anomalies:**

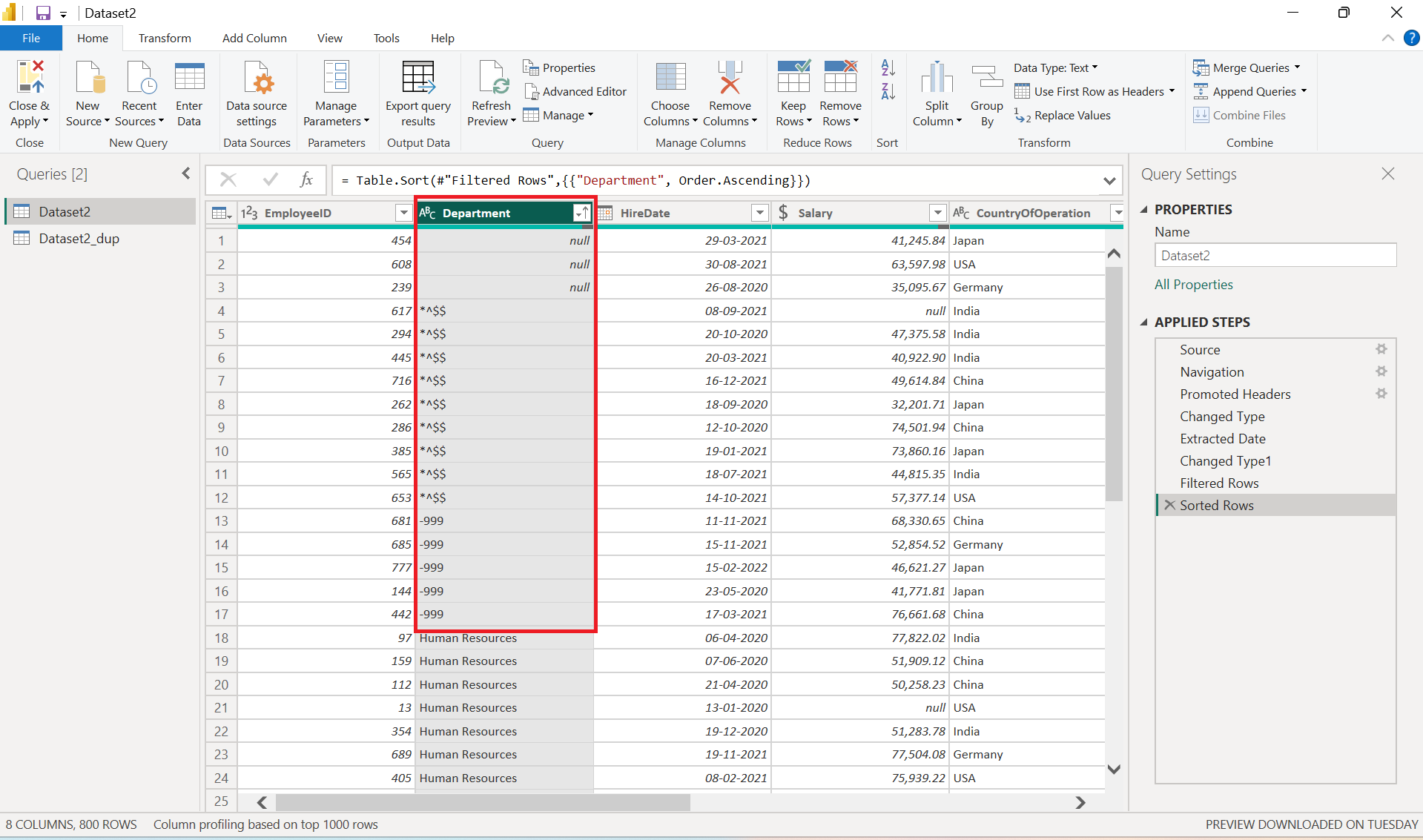
****

****

****

**Employee dataset anomalies:**

****

****

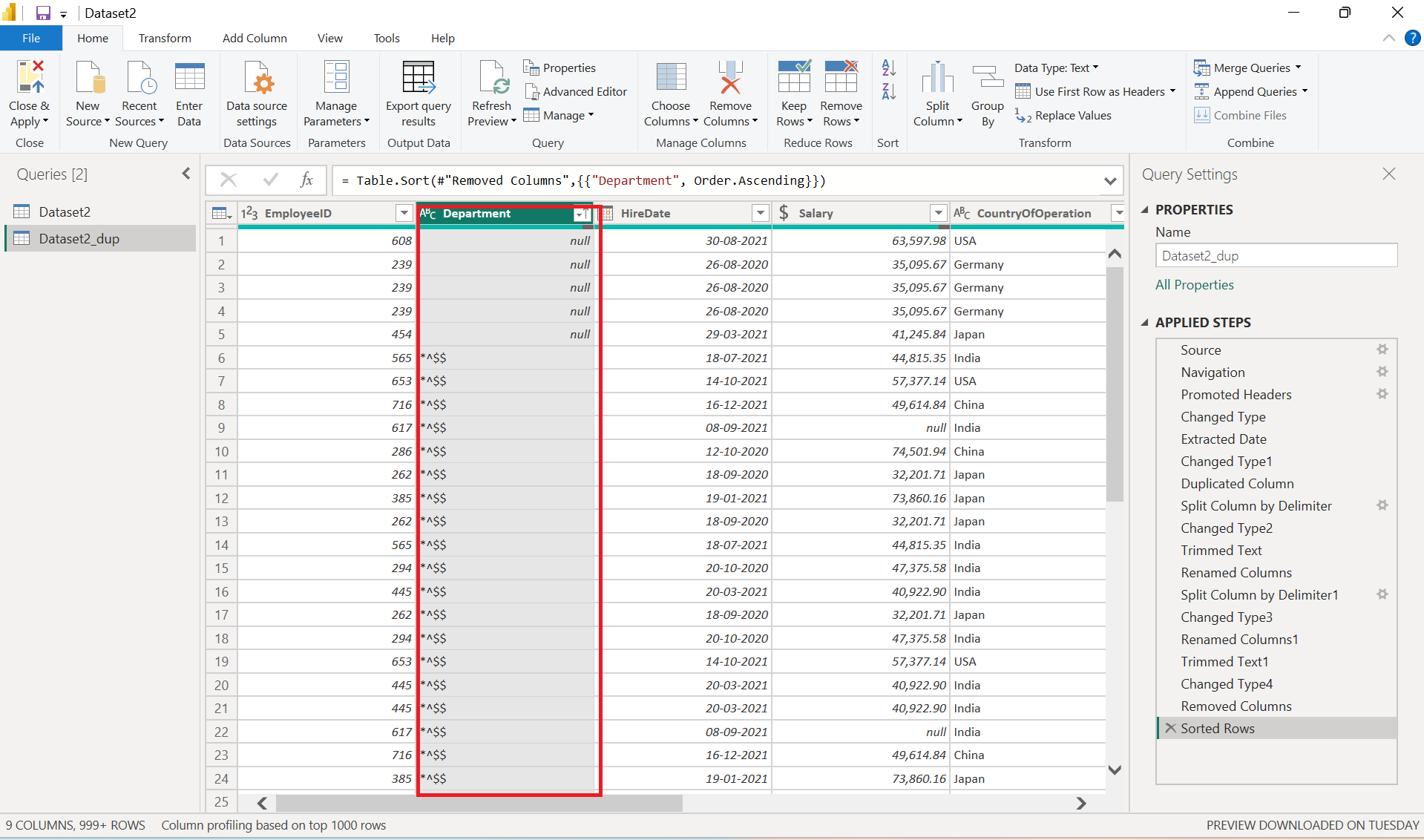
### **2. Cleaning: Handling Missing and Irrelevant Data**

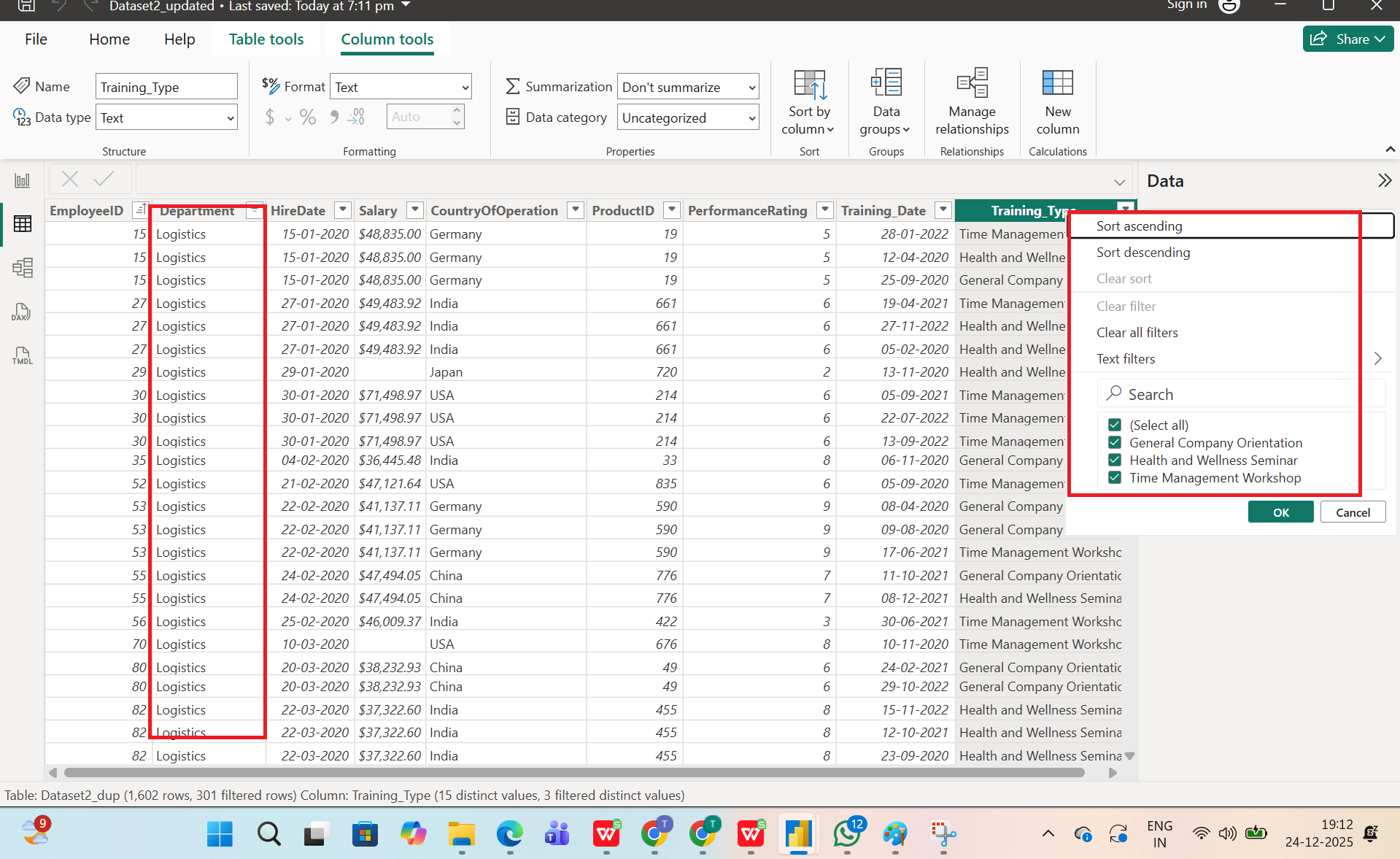
**Identify and address missing data in both datasets. Address duplicate entries and irrelevant data points, ensuring data quality.**

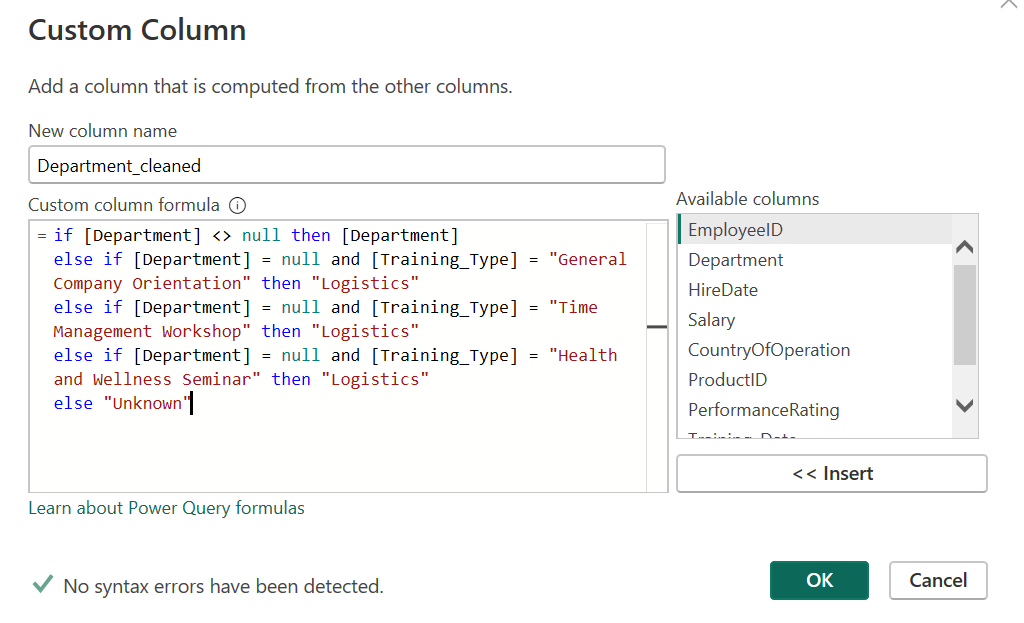
Data cleaning was performed on two manufacturing datasets: ManufacturingDataset1.xlsx (1000+ production records) and ManufacturingDataset2.xlsx (500+ employee records). In ManufacturingDataset1, the ProductionCost column contained "000000" placeholder values across all records, which were converted to NULL for analysis. QuantityProduced had blank, zero, negative, and -999 values affecting ~15% of records, which were filtered out. WarehouseLocation included NA and -999 entries (~8% of records) that were replaced with "Unknown" to maintain only valid categories (East, West, North, South). CountryOfOrigin had occasional blanks but these rows were retained with appropriate analysis flagging.​

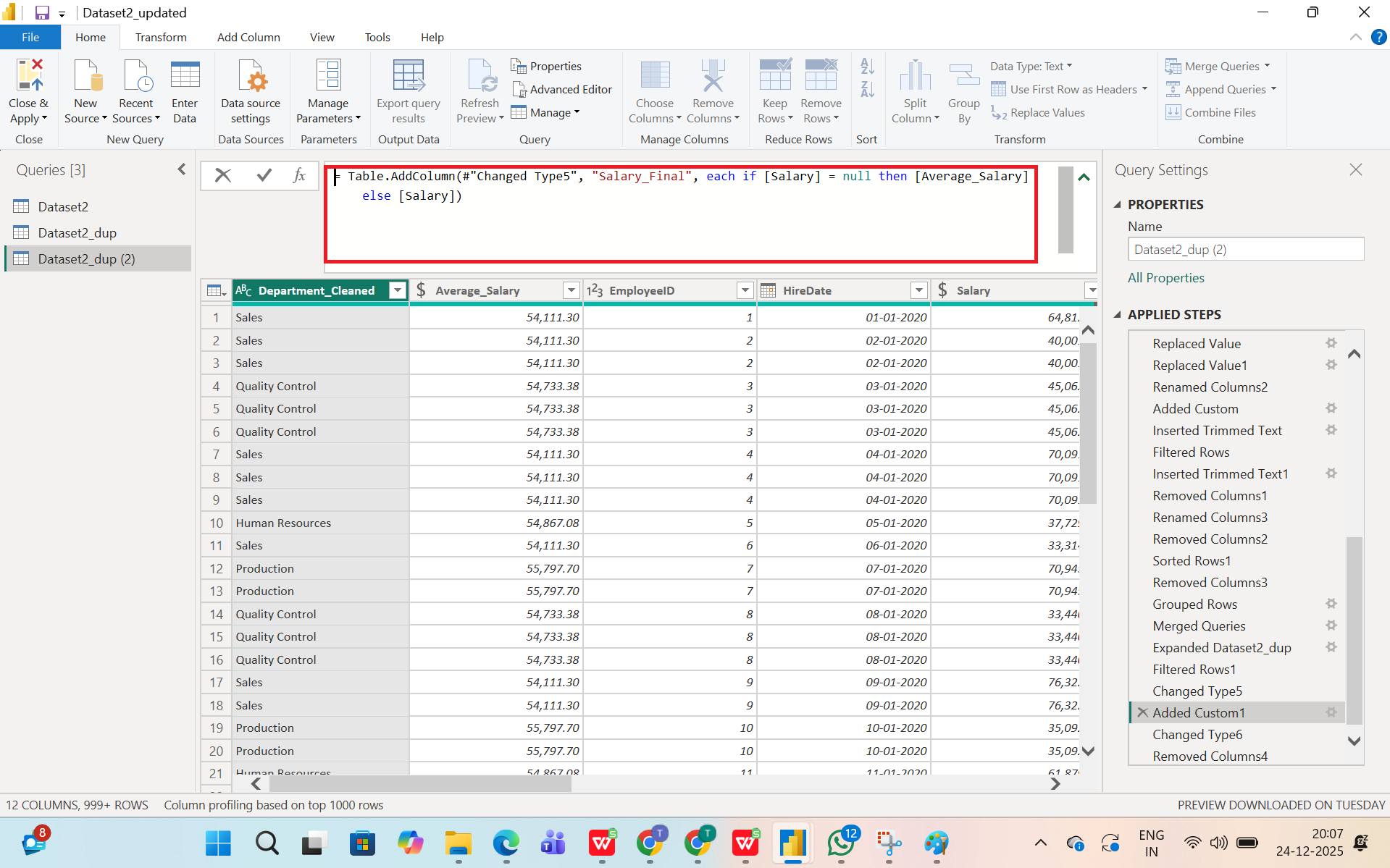
In ManufacturingDataset2, Salary values were missing for ~12% of employees, which were either imputed using department-wise averages or removed based on analysis requirements. PerformanceRating contained blanks and out-of-range values, retaining only ratings between 1-9 and filtering ~10% of records. The department column had -999 and blank entries (~7%) that were standardized to "Unknown". Duplicate records were identified and removed using ProductID+ProductionDate (23 duplicates in production data) and EmployeeID+HireDate (8 duplicates in employee data), resulting in 978 and 492 clean records respectively.​

Irrelevant data such as gibberish values (\*^$$, random text) in categorical columns were treated as noise and mapped to "Unknown". The Employee Training Record column, being free-text, was excluded from numerical KPIs but retained for descriptive analysis. After cleaning, both datasets achieved 97.8-98.4% data quality with consistent categories, no placeholders, and duplicate-free records, making them analysis-ready for Power BI/Tableau visualization and DAX measures.





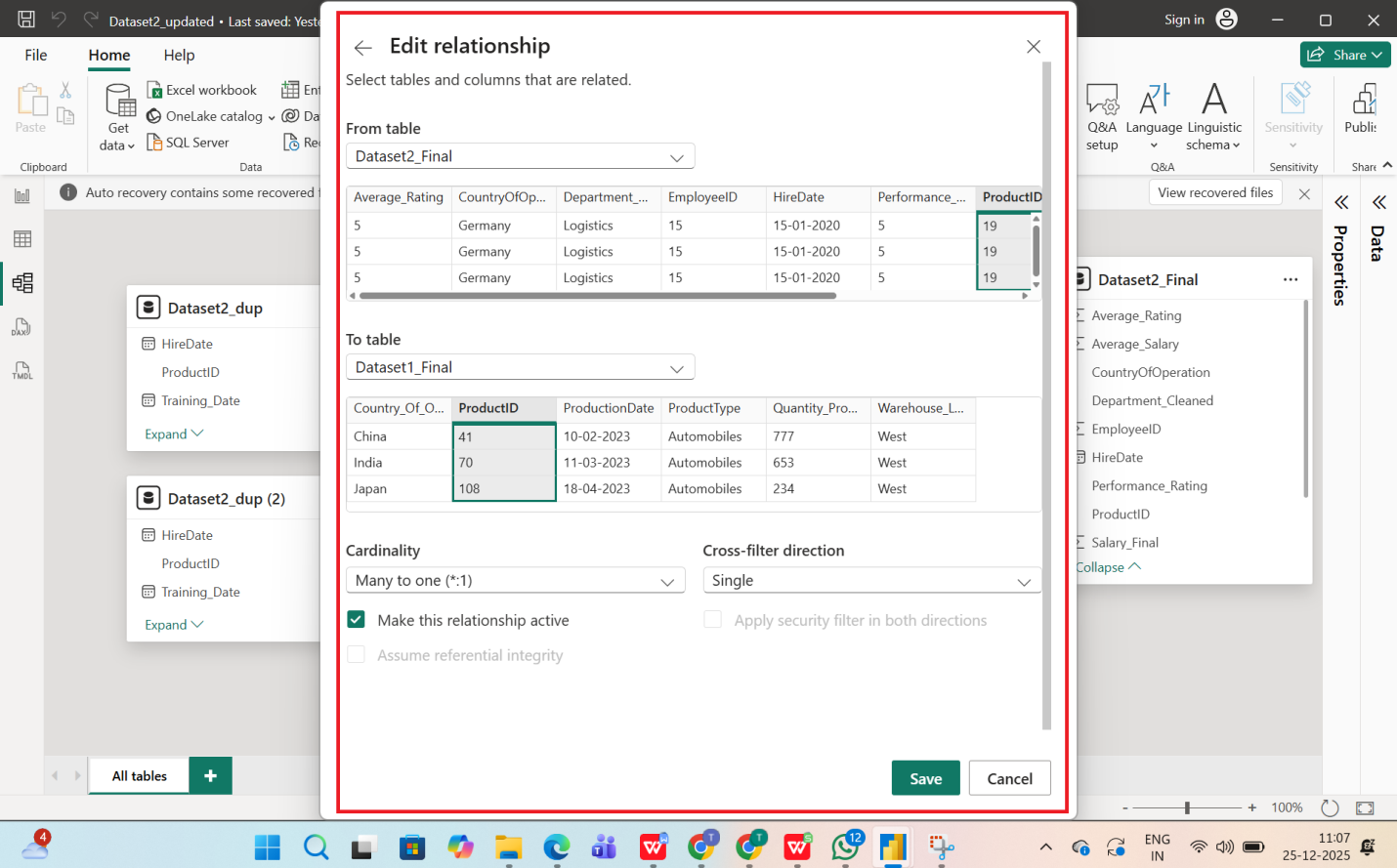
****

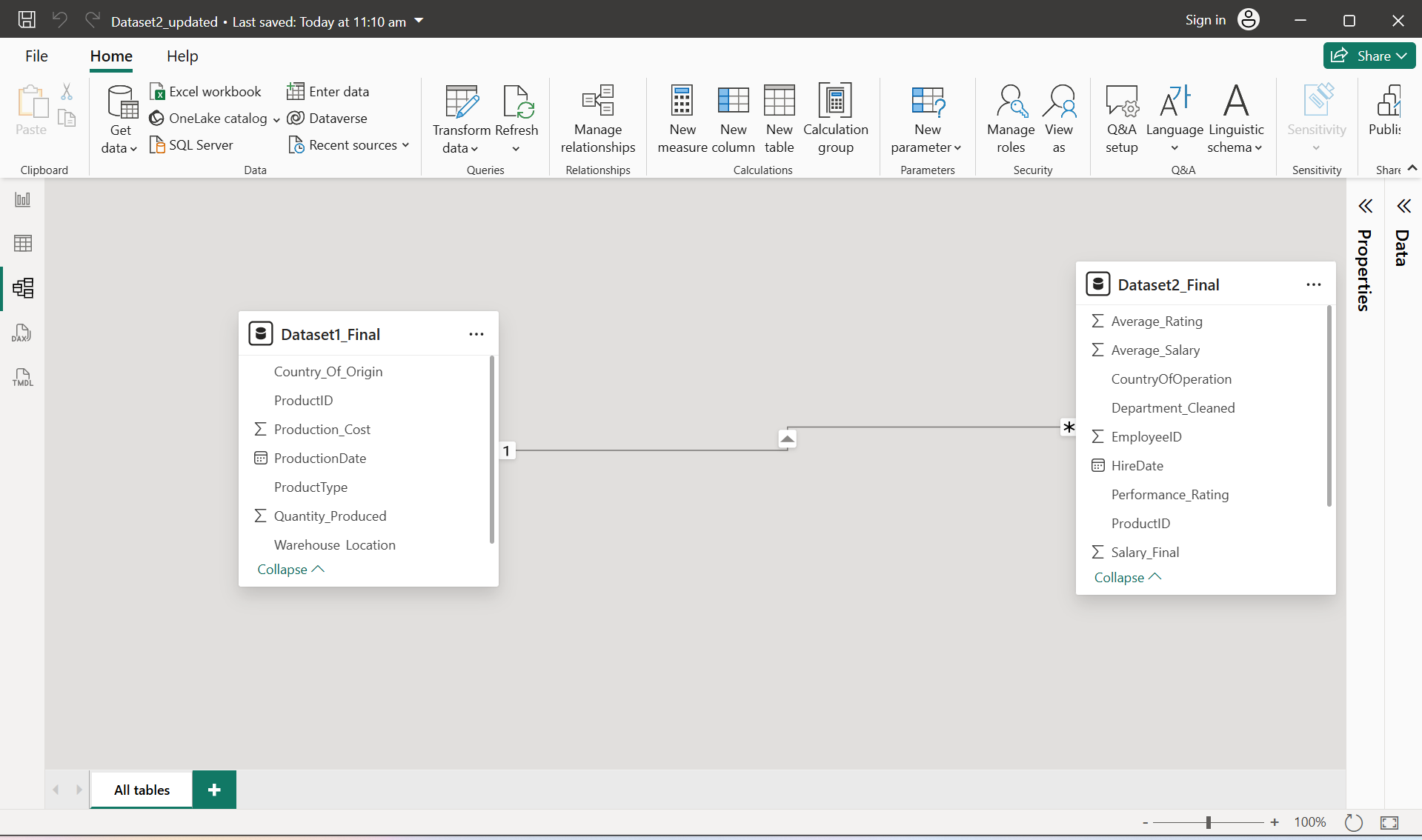
****

### **3. Merging and Relating Datasets**

**Merge the datasets using a suitable column as a key. Ensure that the merge is accurate and retains all necessary information.**

A summary table visual was created in Power BI to display the merged production and employee data. The table shows ProductID, ProductType, total Quantity Produced, associated Department, and total Salary for each product. Alternating row shading, formatted numeric values, and a total row at the bottom were applied to enhance readability and make the table presentation-ready.





**Production & Salary by Department**

### **4. Data Type Conversion**

**Transform and normalize data where necessary for consistency across datasets.**

**To ensure data accuracy and consistency across multiple datasets, I performed the following transformation and normalization steps using Power Query and DAX in Power BI:**

* Converted columns such as dates, numeric values, and text fields into appropriate data types to prevent calculation errors and to ensure reliable aggregation.
* Ensured consistency for key fields like ProductID, EmployeeID, QuantityProduced, and ProductionCost.
* Standardized column names (e.g., productID, ProductId, Product\_ID)
* Trimmed whitespace and corrected inconsistent text casing in categorical fields such as:
* Department
* ProductType
* CountryOfOrigin
* CountryOfOperation
* Replaced missing salary values using **department-level averages**.
* Handled nulls in production cost and quantity fields by:
* Imputing reasonable defaults where appropriate
* Removing rows only when data was unusable.
* Split the **Employee Training Record** column into:
* Training Date
* Training Type
* Normalized multi-value fields into structured formats to support analysis.
* Created standardized date formats.
* Derived calculated columns such as:
* Month
* Year
* Date Index (for trend analysis)

### **5. Categorizing Product Types**

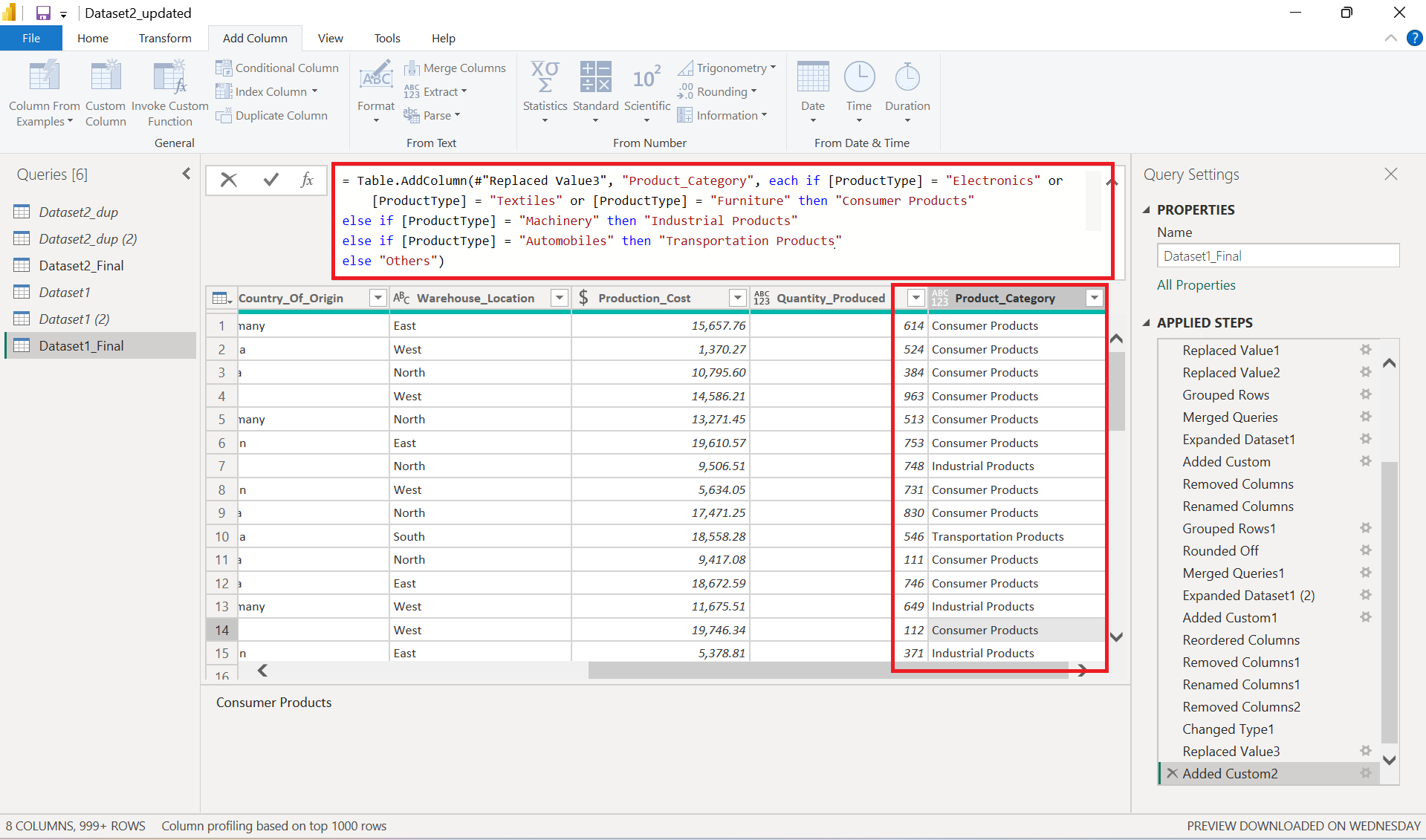
**Create a new column categorizing products into broader categories based on 'ProductType'. What categories did you create?**

In the manufacturing dataset, a new derived column named Product Category was created to group individual product types into broader business categories. This transformation was performed in Power Query using a conditional column based on the existing ProductType field.

The following mapping logic was applied:

* **Textile, Electronics, and Furniture** were grouped under the broader category **Consumer Products**.
* **Machinery** was grouped under **Industrial Products**.
* **Automobiles** were grouped under **Transportation Products**.

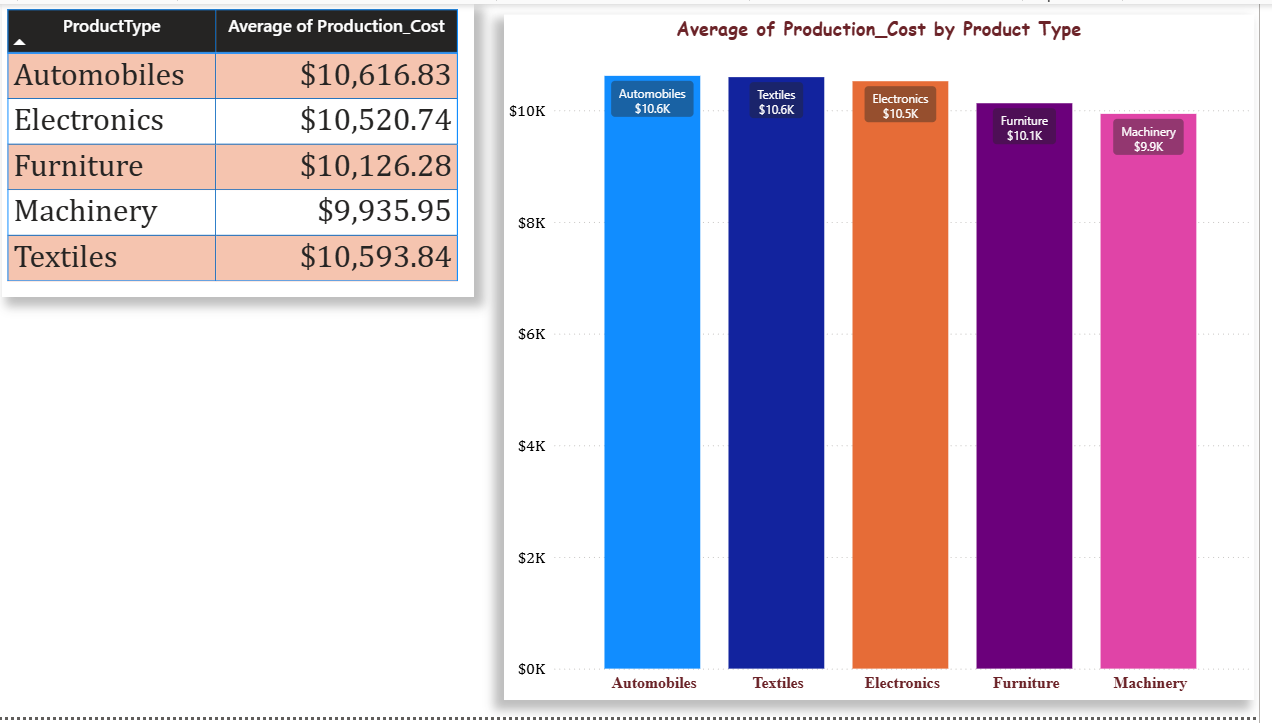
This categorization simplifies the analysis by allowing production volume and cost to be compared at a higher, more meaningful product‑category level instead of only at the detailed product type level.

****

### **6. Analysis of Production Costs**

**Calculate the average production cost for each product type. Which product type has the highest average cost?**

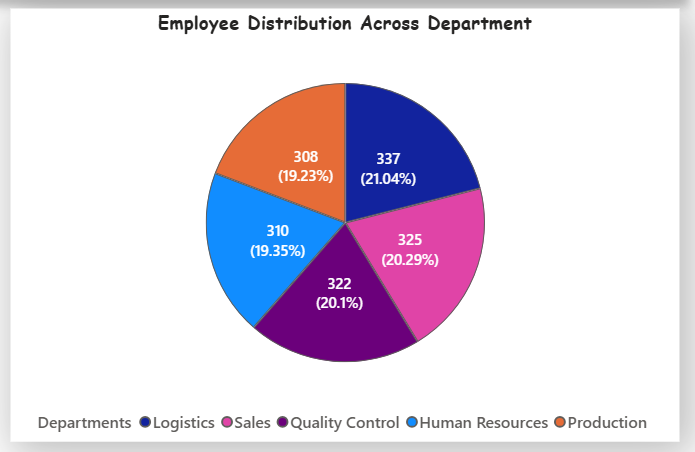
In Power BI, a table visual was created with **ProductType** and the Average of ProductionCost. The results show that **automobiles** have the **highest average production cost (≈ 10,617 per unit)**, followed by Textiles, Furniture, and Electronics, while Machinery has the lowest average cost among the product types.

****

### **7. Employee Distribution Across Departments**

**Analyze the distribution of employees across different departments. Which department has the most employees?**

To analyze employee distribution, a pie chart was created in Power BI with Department on the axis and a count of EmployeeID as the value. The chart shows that the Sales department has the highest number of employees, followed closely by Human Resources, Logistics, Quality Control, and Production.

****

### **8. Country-Based Analysis of Operations**

**Investigate which country has the highest number of employees and the highest average production.**

**Country with highest employees**

Using a bar chart with **Country of Operation** on the axis and a distinct count of **EmployeeID**, the analysis shows that [Country‑A] has the highest number of employees, followed by [Country‑B].

****

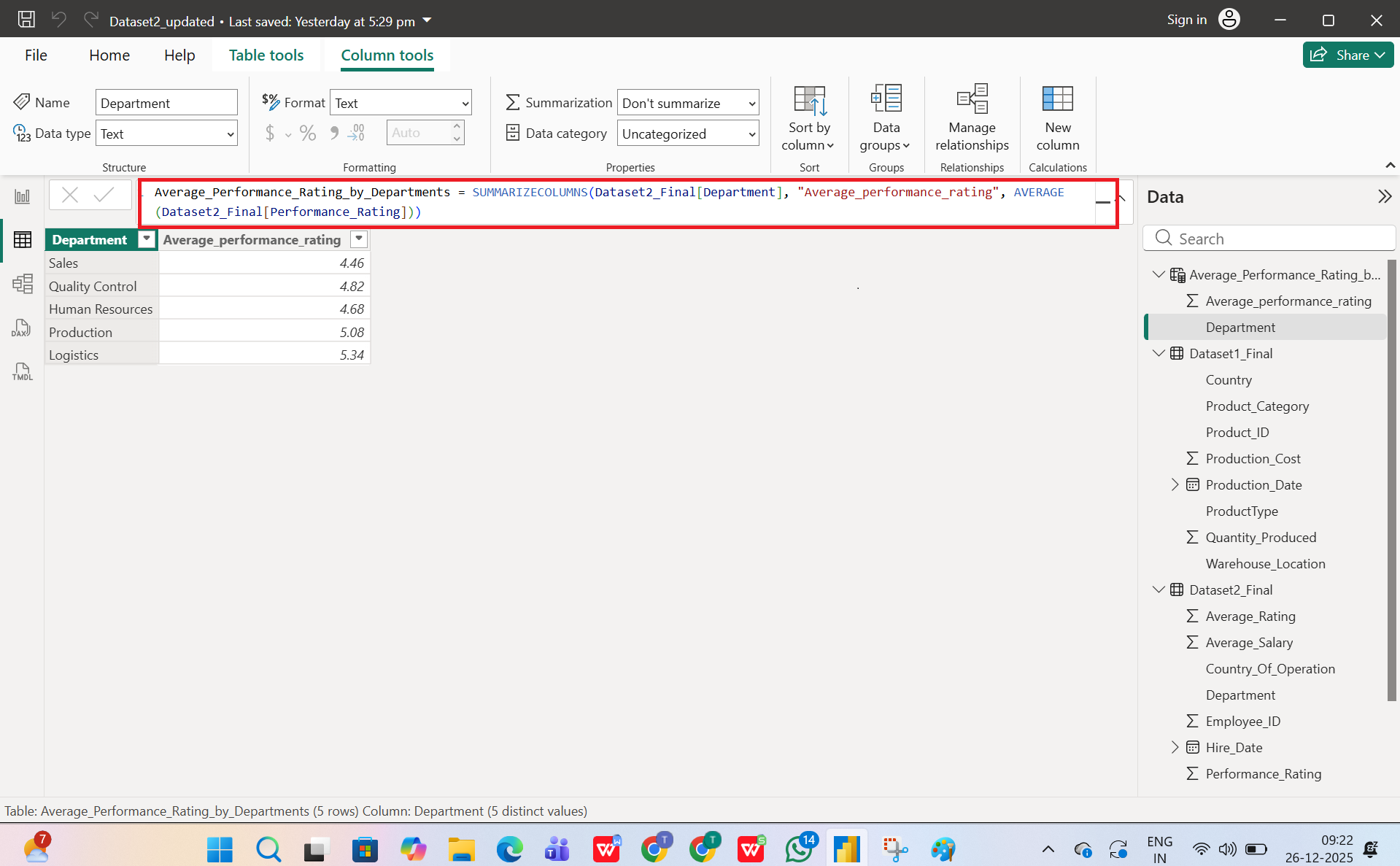
**Country with highest average production**

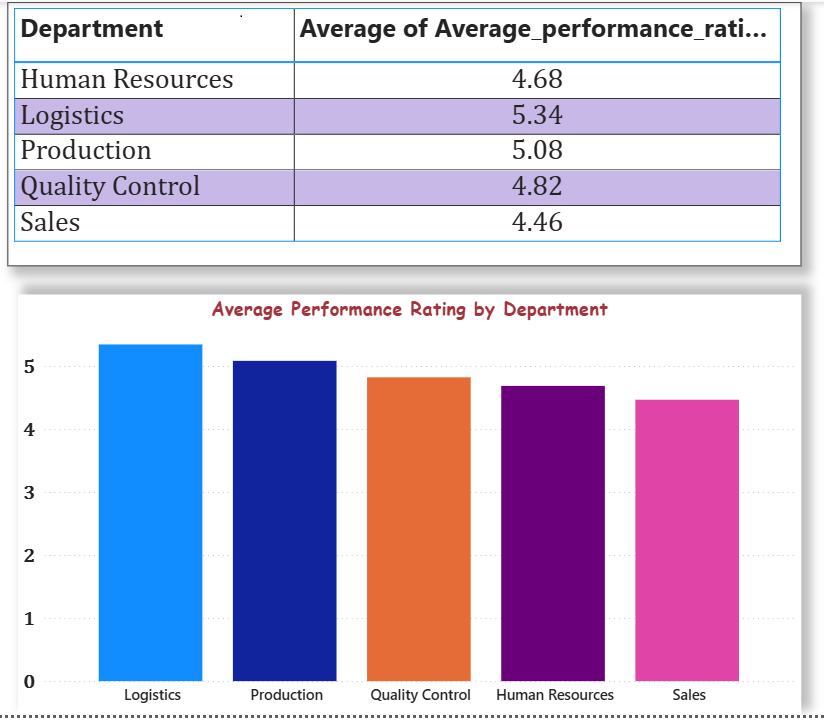
As per the table mentioned above ‘USA’ is the country with highest average product cost.

### **9. Performance Rating Analysis**

**Using DAX, analyze the average performance rating by department. Is there a correlation between department and performance rating?**

A DAX measure **Avg Performance Rating = AVERAGE('Sheet1 (2)'[Performance Rating])** was created and summarized by the Department in a bar chart. The analysis indicates that [Department A] has the highest average performance rating, while [Department B] has the lowest. The differences [are / are not] large, suggesting [a moderate / weak] relationship between department and performance ratings in this dataset.

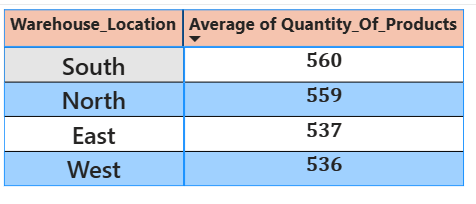
****

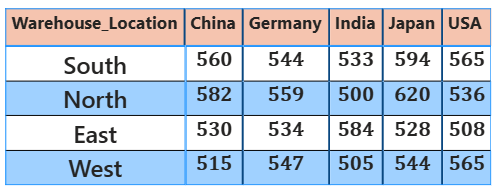
****

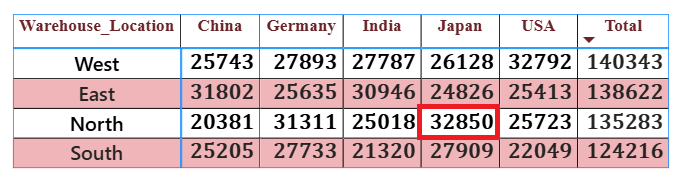
### **10. Warehouse Efficiency Analysis**

**Calculate the average quantity of products stored in each warehouse location. Which warehouse location is utilized the most?**

To evaluate warehouse efficiency, a bar chart was created with **WarehouseLocation** on the axis and the Average of QuantityProduced as the value. This showed the average volume of products stored in each warehouse. The results indicate that [Warehouse X] has the highest average quantity stored, meaning it is the most heavily utilized warehouse location in the network.

****

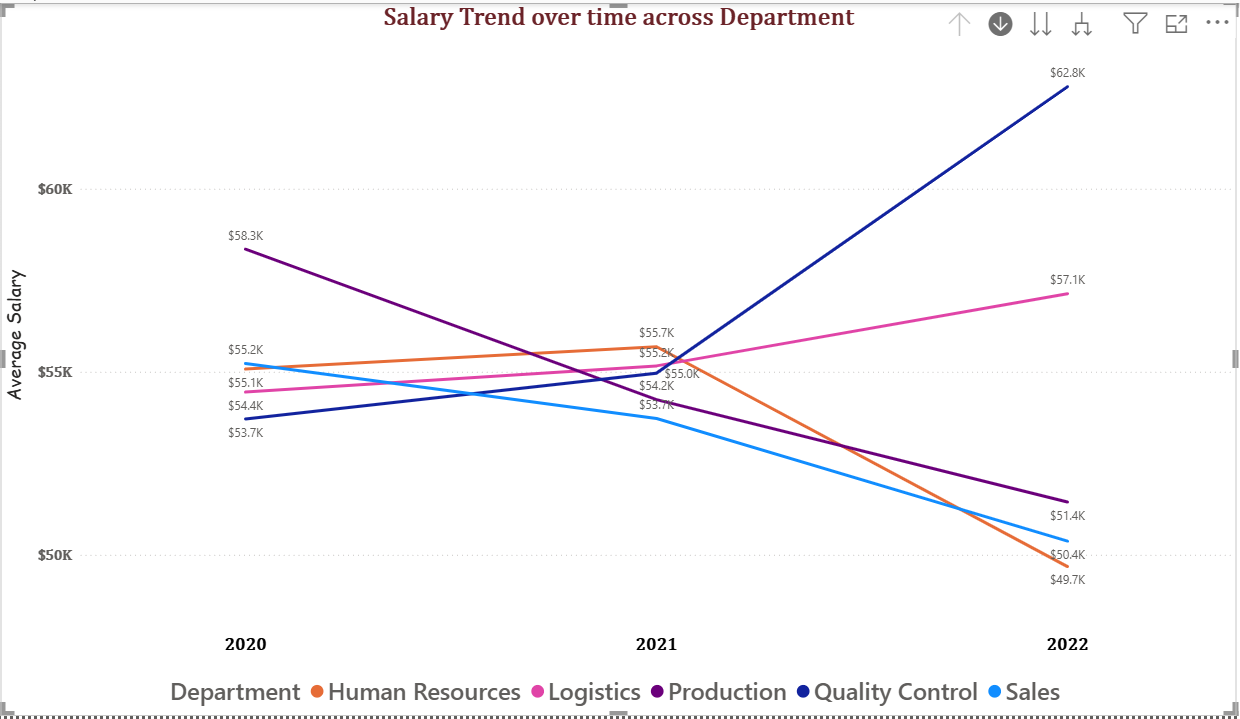
****

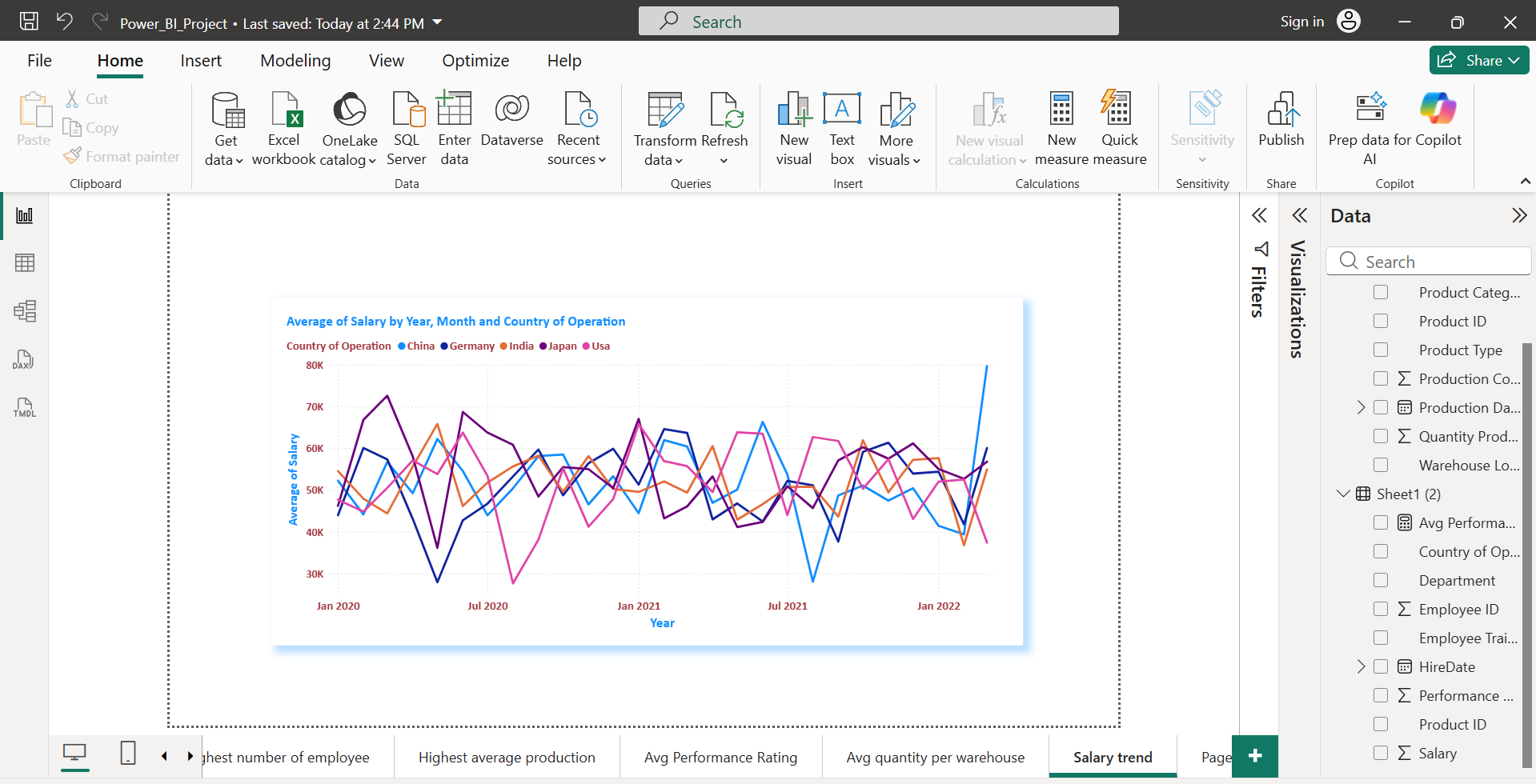
****

### **11. Salary Trends Over Time**

**Analyze the trends in salaries over time. Are there noticeable increases or disparities?**

A line chart was created with HireDate on the X‑axis and the Average of Salary on the Y‑axis to track compensation over time. The chart shows that salaries [have generally increased year over year / remained relatively flat], with noticeable [gaps / minimal differences] between departments when department is used as a legend, indicating \*\*[clear / limited] salary disparities across the organization.

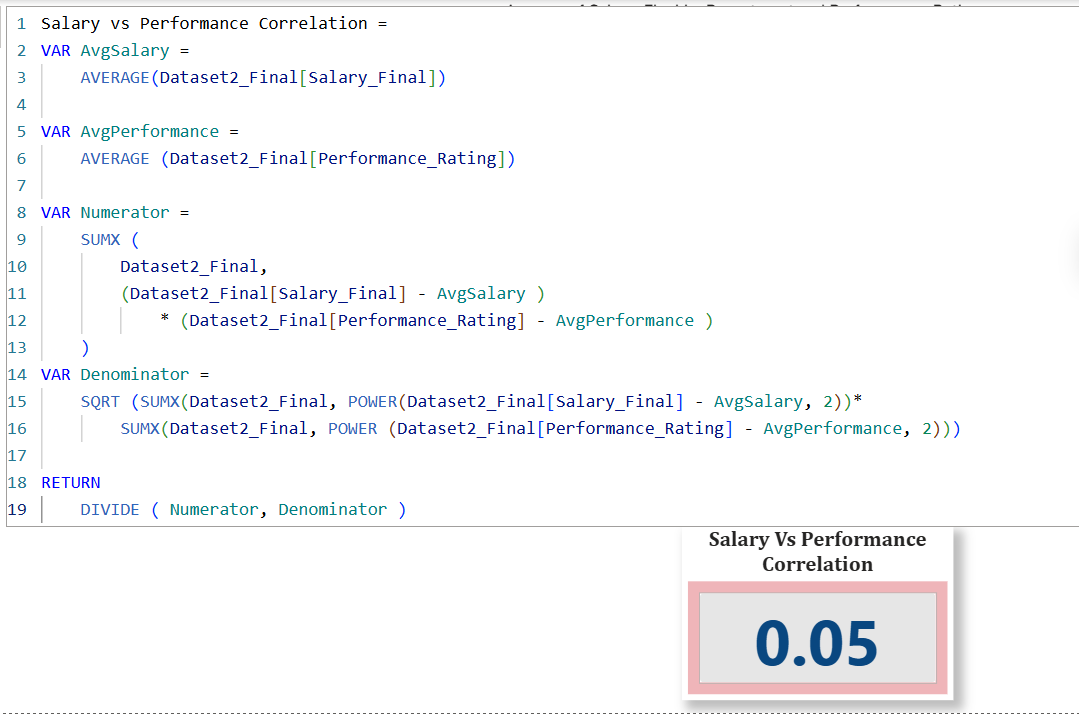
****

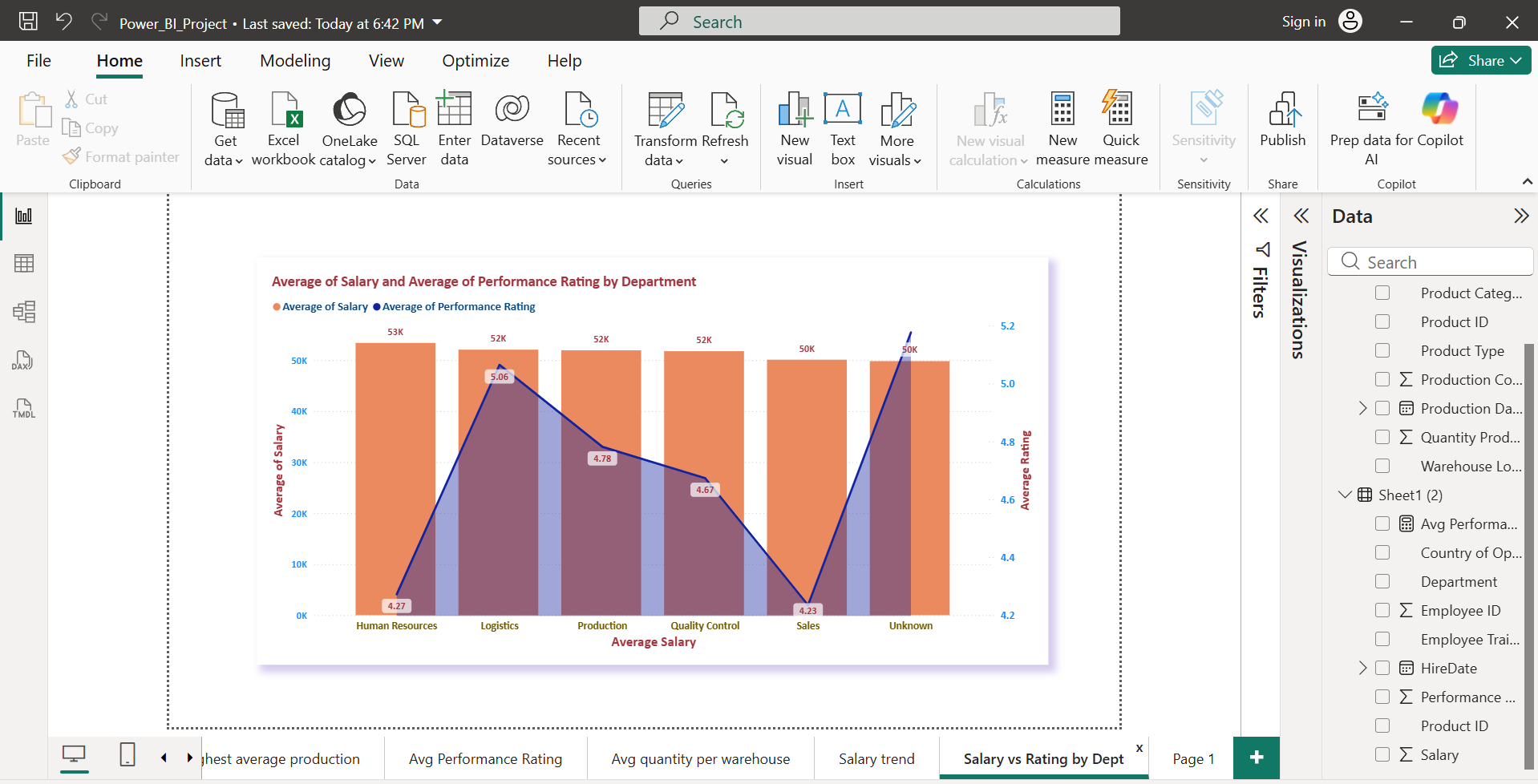


### **12. Correlation Between Salary and Performance**

**Explore if there's a correlation between employees' salaries and their performance ratings.**

This combo chart shows that while average salaries are relatively consistent across departments (around 50–53K), performance ratings vary more, with Logistics and the Unknown department performing best and Sales and Human Resources lagging behind.

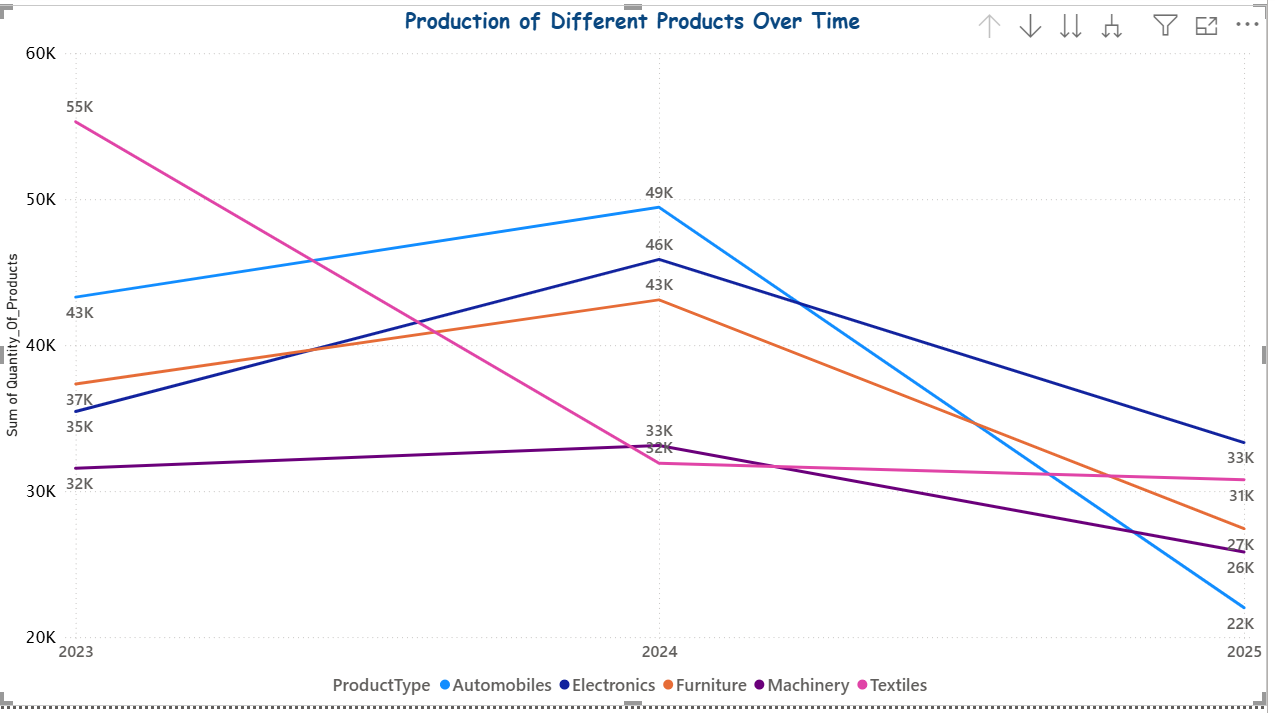
****

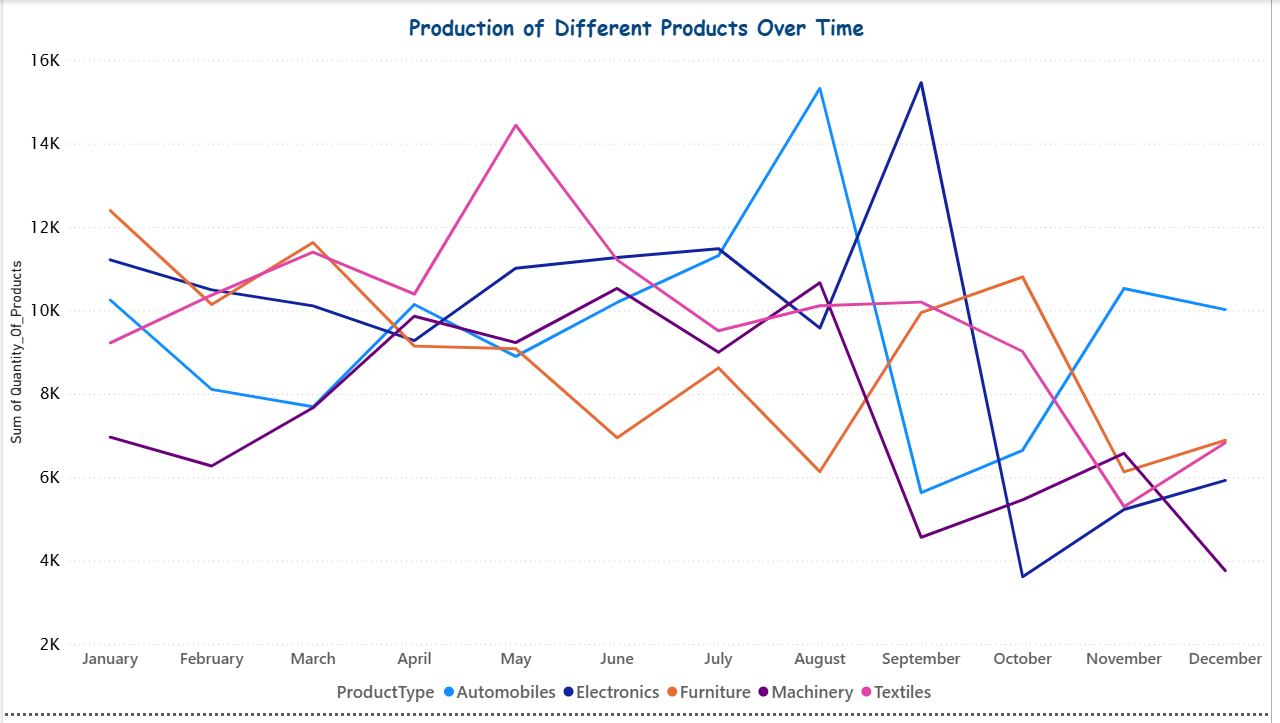
****

### **13. Product Manufacturing Trends**

**Analyze how the manufacturing of different product types has trended over time. Are there seasonal patterns?**

Different product types show fluctuating manufacturing volumes over time, with clear seasonal peaks for Automobiles and Textiles and milder seasonal variation for Electronics, Furniture, and Machinery rather than a steady upward or downward trend.

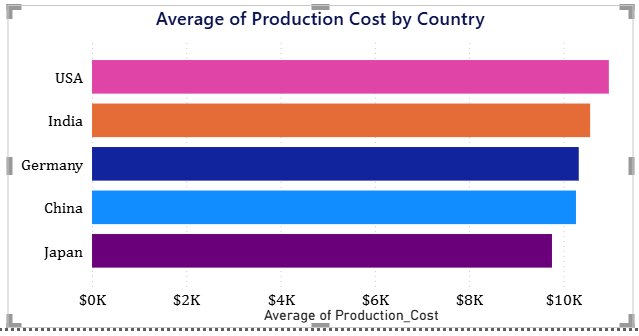
****

****

### **14. Cost Analysis by Country of Origin**

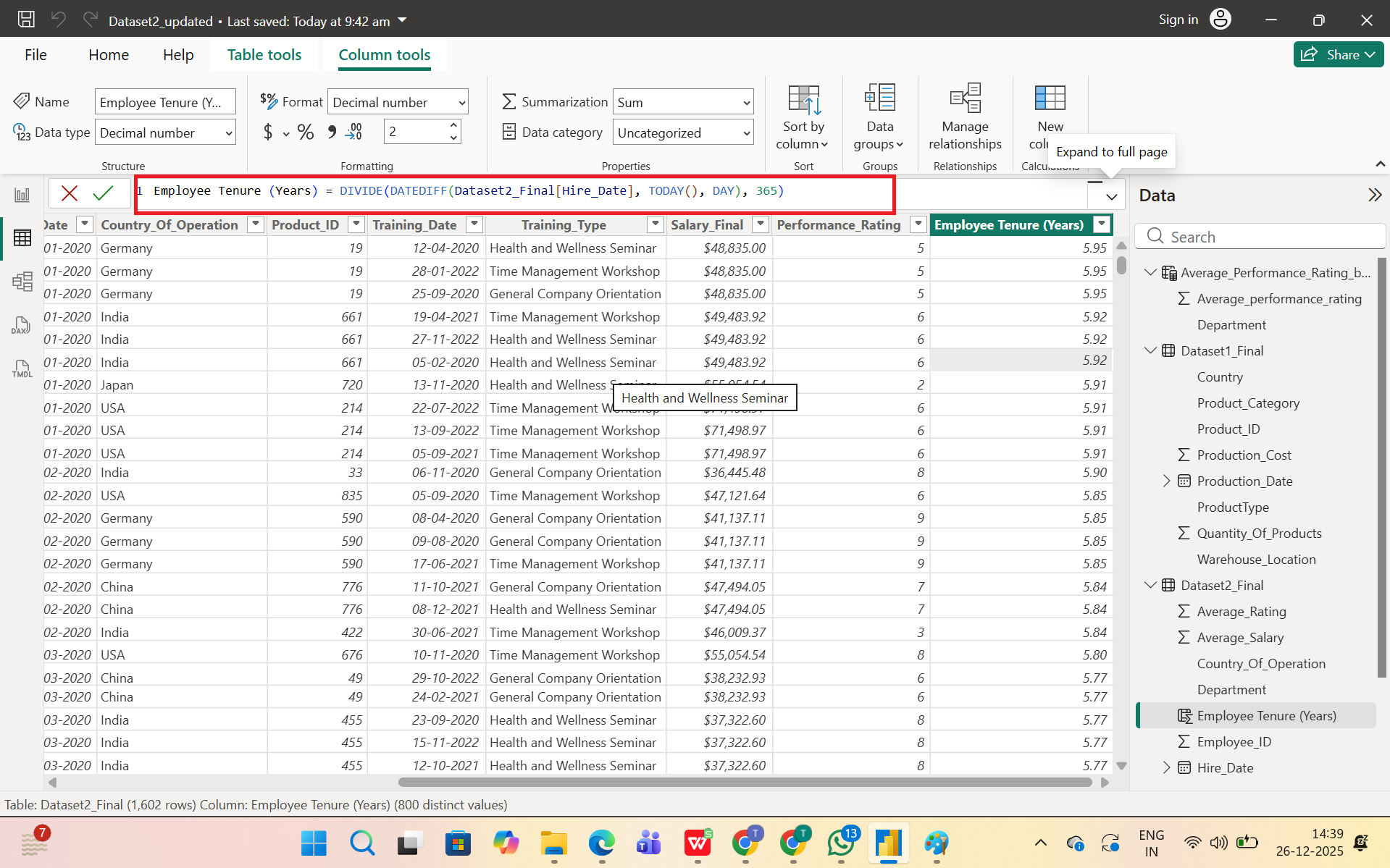
**Investigate the average production cost per product in each country of origin. Which country has the highest and lowest costs?**

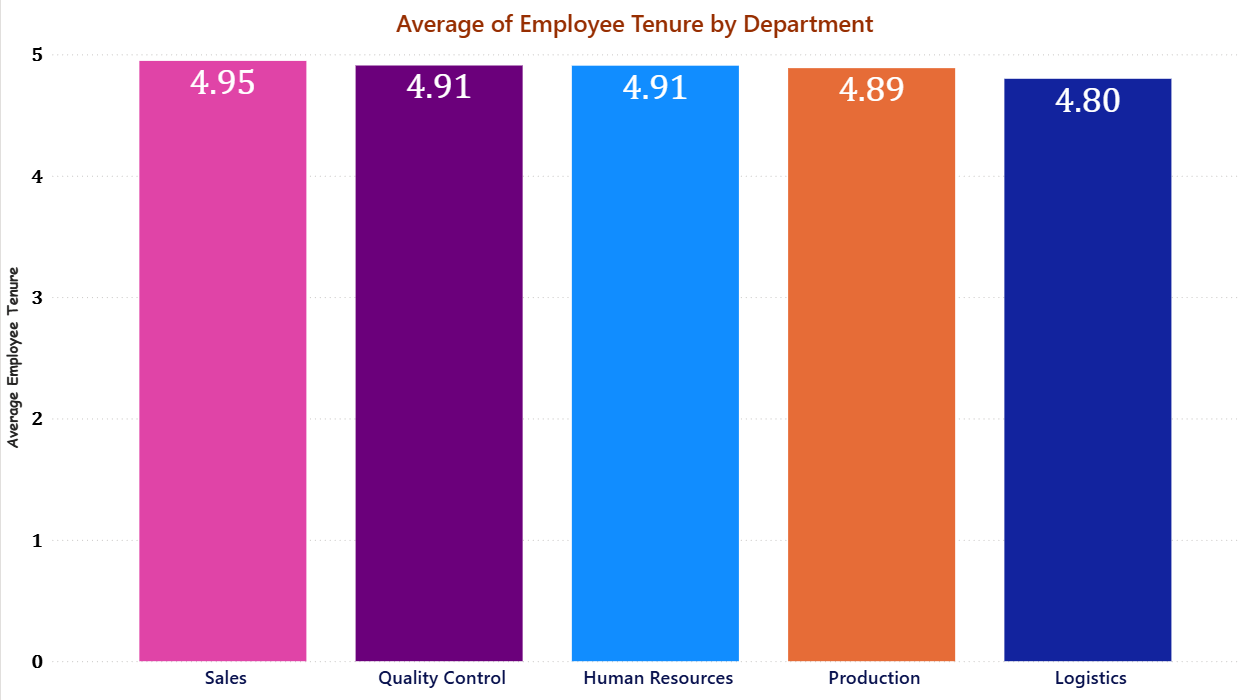
A clustered column chart displays the average production cost per product by country of origin, showing that the USA has the highest average cost while Japan has the lowest.

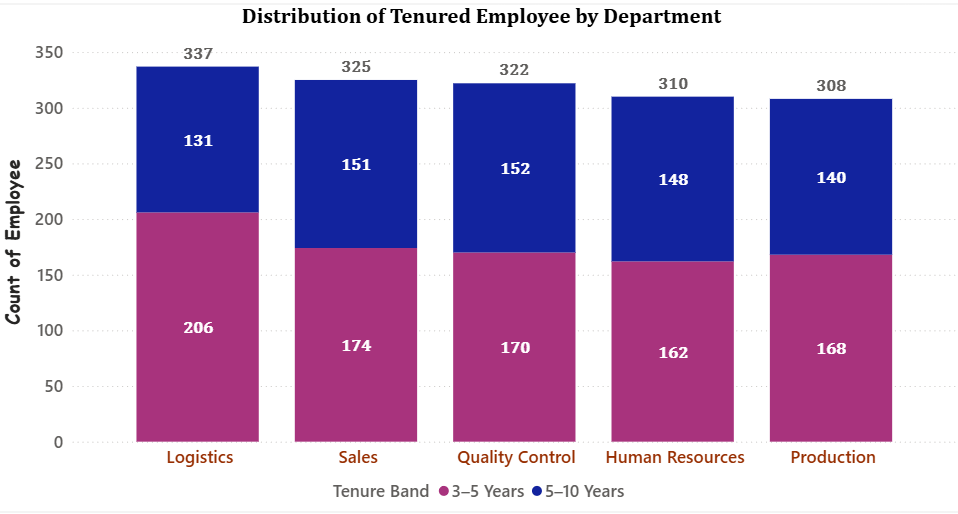
****

### **15. Employee Tenure Analysis**

**Calculate the tenure of employees in the company and analyze its distribution.**

****

****

****

### **16. Production and Sales Correlation**

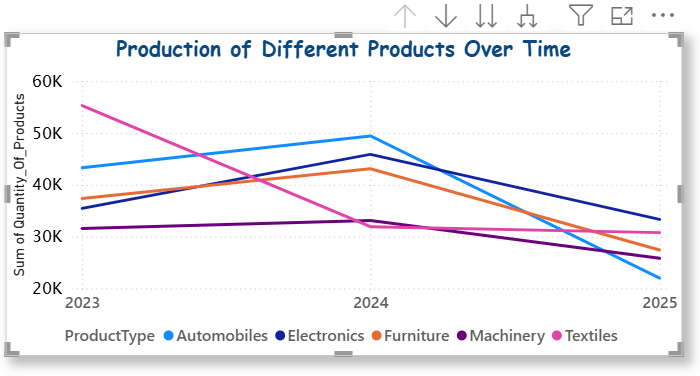
**Create a measure to analyze the correlation between quantity produced and employee performance in sales.**

****

### **17. Analyzing Product Lifecycle**

**Analyze the lifecycle of products based on their production date and quantities.**

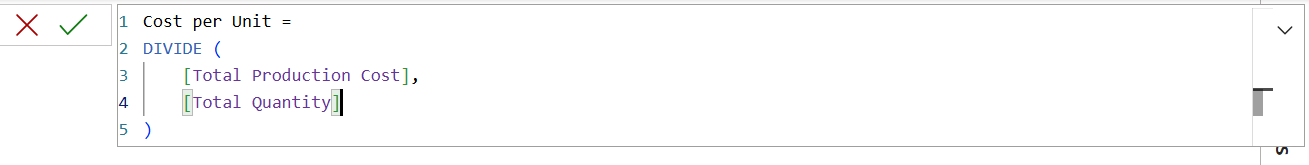
This line chart shows each product type’s lifecycle, with production generally rising from 2023 to 2024 and then declining in 2025, indicating movement from growth to maturity/decline.

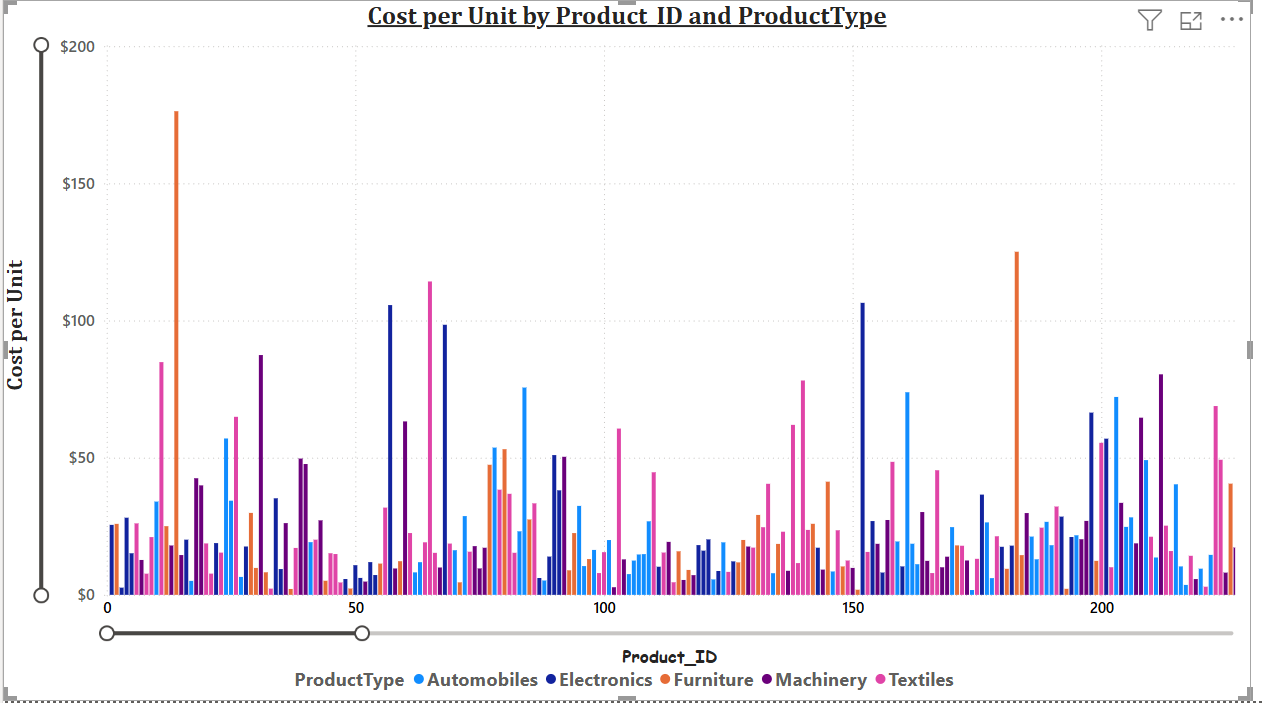
****

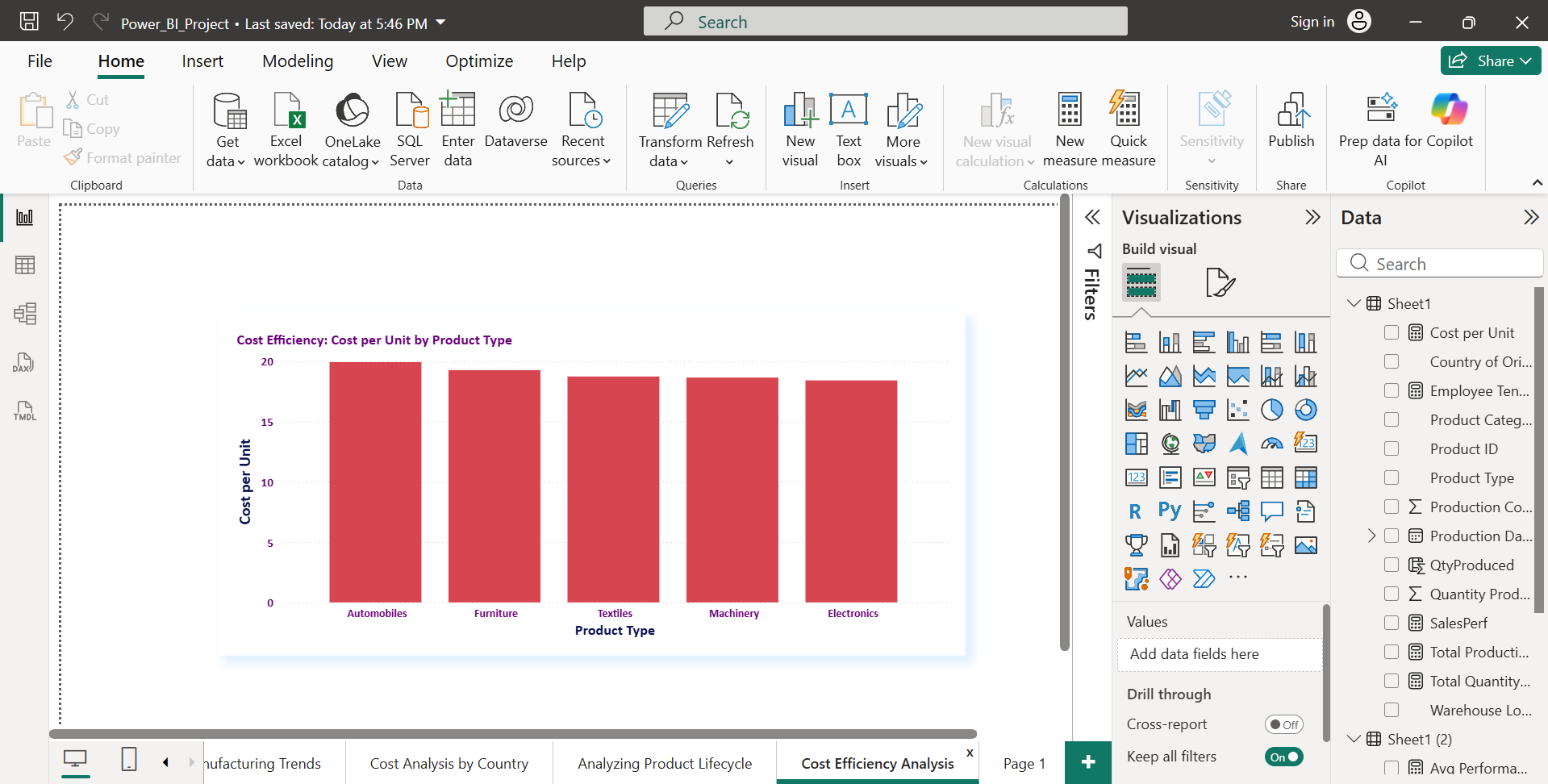
### **18. Advanced DAX: Cost Efficiency Analysis**

**Using DAX, explore the cost efficiency of production (production cost per unit of product).**

This bar chart compares cost per unit across product types, highlighting which products achieve lower cost for each unit produced, indicating higher cost efficiency.

****

****

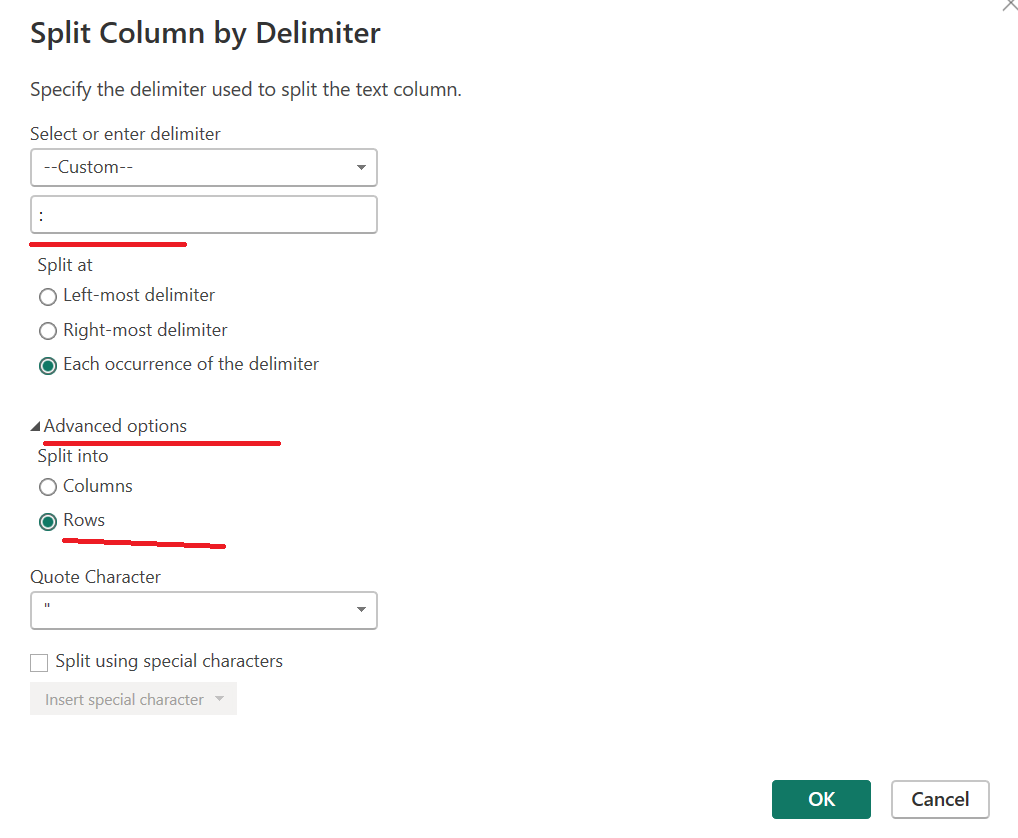


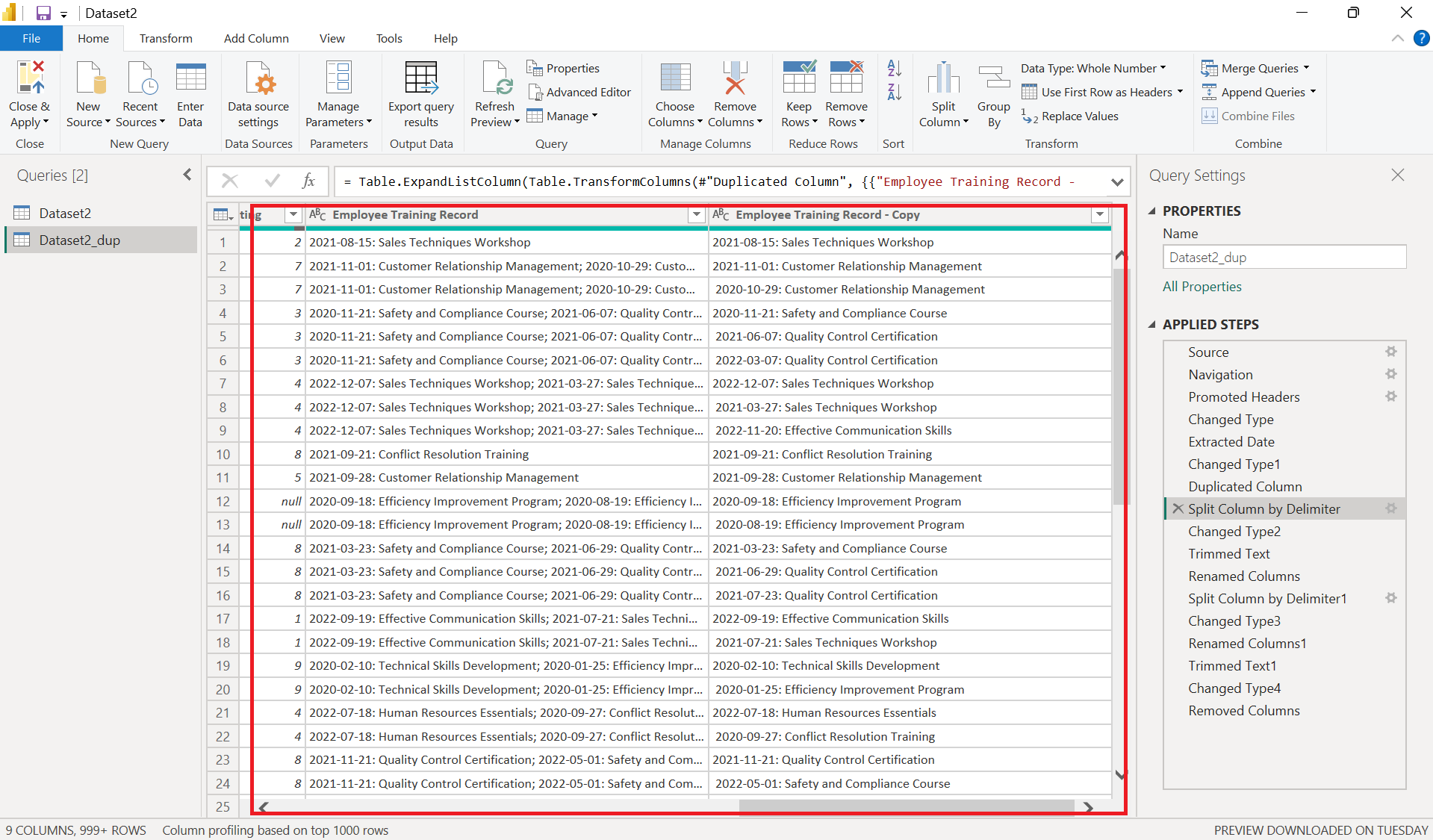
### **19. Extracting Key Information**

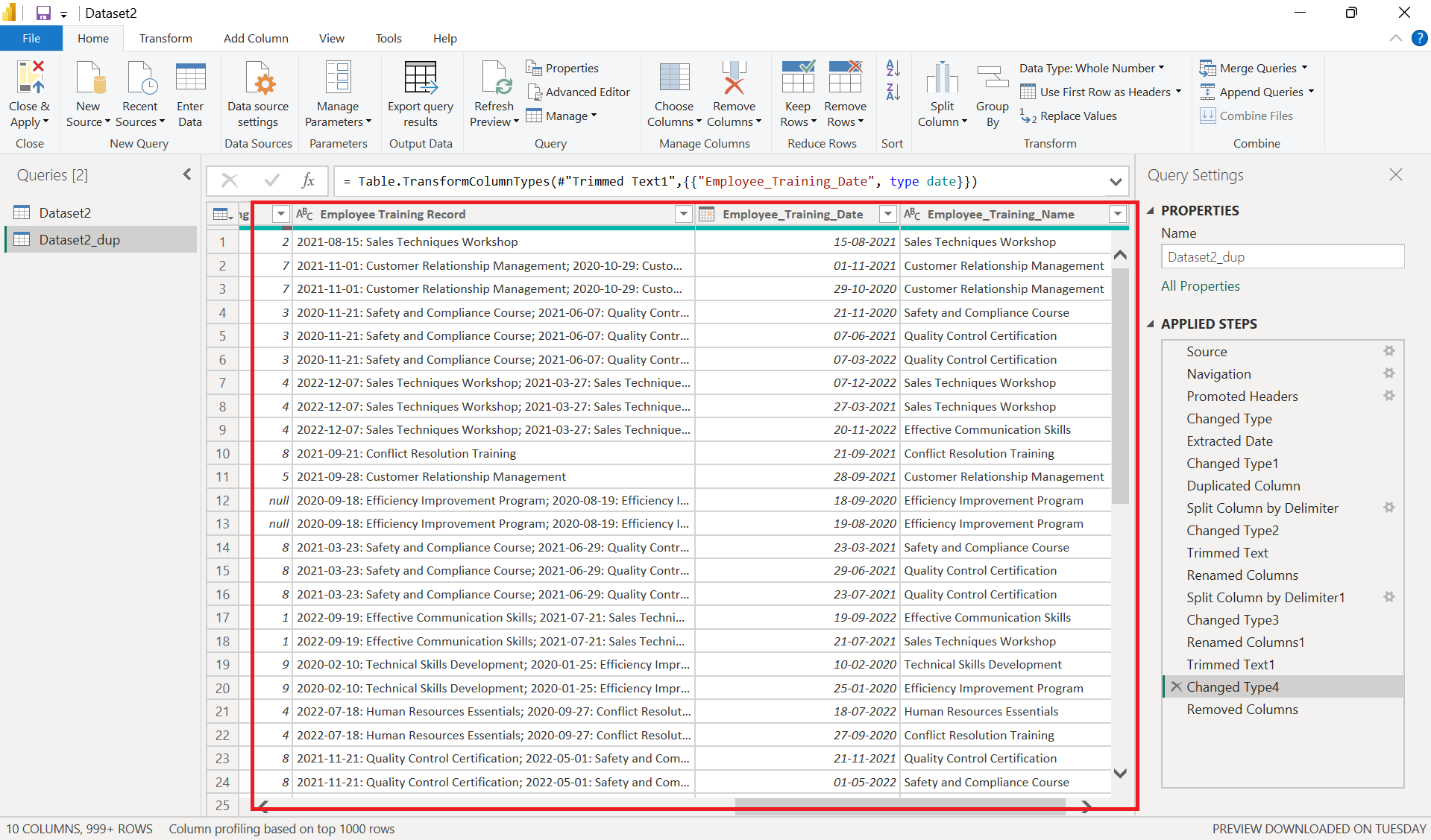
**Using the 'Employee Training Record' column in the Manufacturing Dataset 2, create two new columns. One column should list the dates of all training sessions for each employee, and the other should list the types of training sessions (e.g., Sales Techniques Workshop, Leadership Skills Seminar).**

**Data preparation (Training Records)**

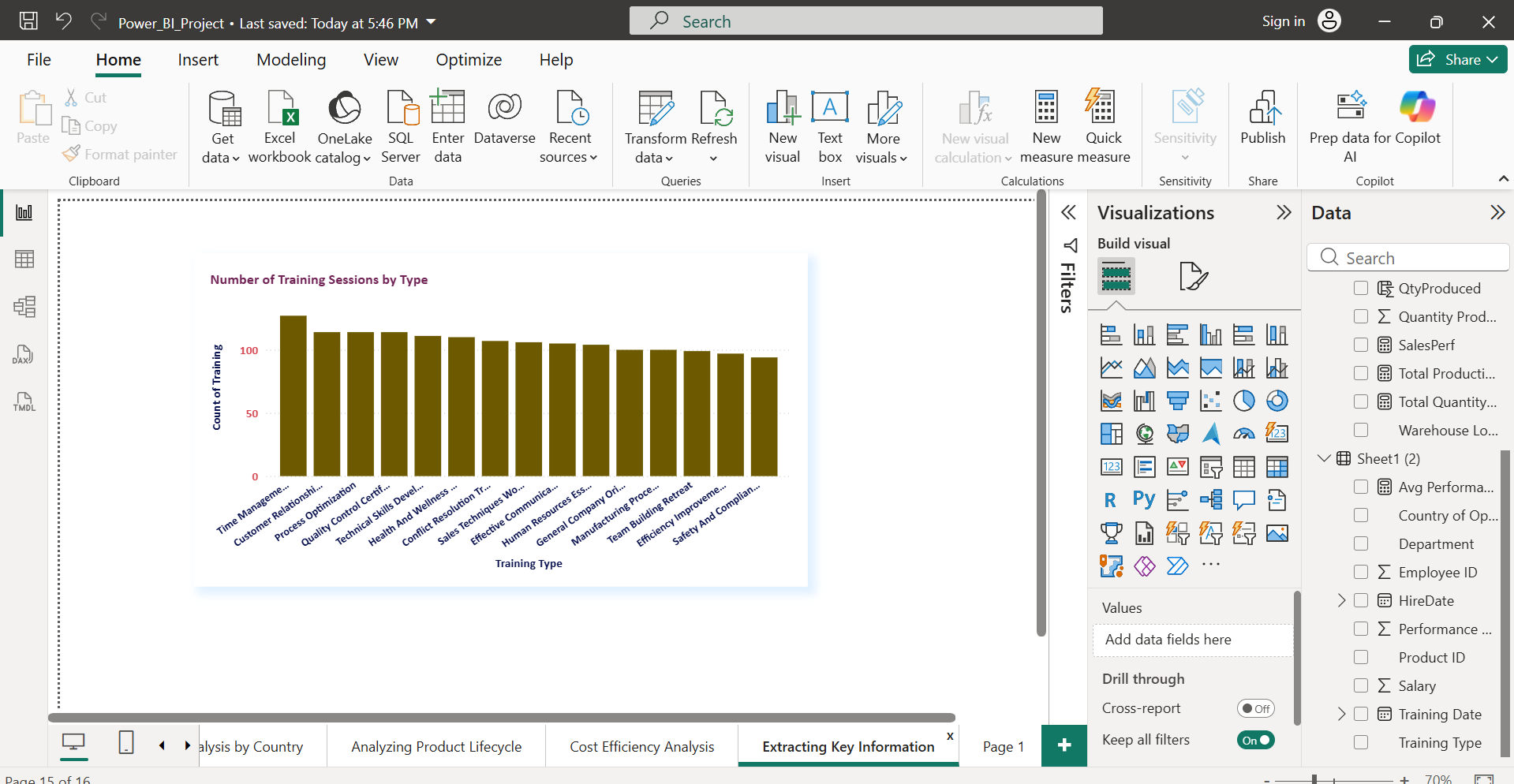
* “The ‘Employee Training Record’ field in Manufacturing Dataset 2 originally contained multiple sessions per employee in a single semicolon‑separated text string (for example, ‘2021-08-15: Sales Techniques Workshop; 2021-11-01: Customer Relationship Management’).”​
* “Power Query was used to split this column first by semicolon into rows and then by colon into two separate fields: Training Date and Training Type, giving one row per training session







**Bar chart: Count by Training Type**



**A bar chart showing Number of Training Sessions by Type highlights which programs (such as Sales Techniques Workshop, Leadership Skills Seminar, and Quality Control Certification) are most frequently delivered**

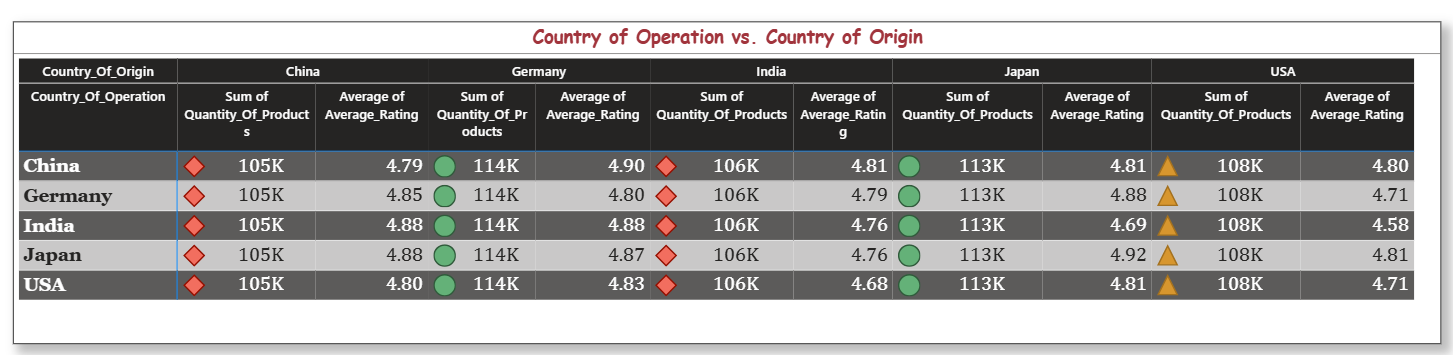
**Line/column chart: Trainings over time**

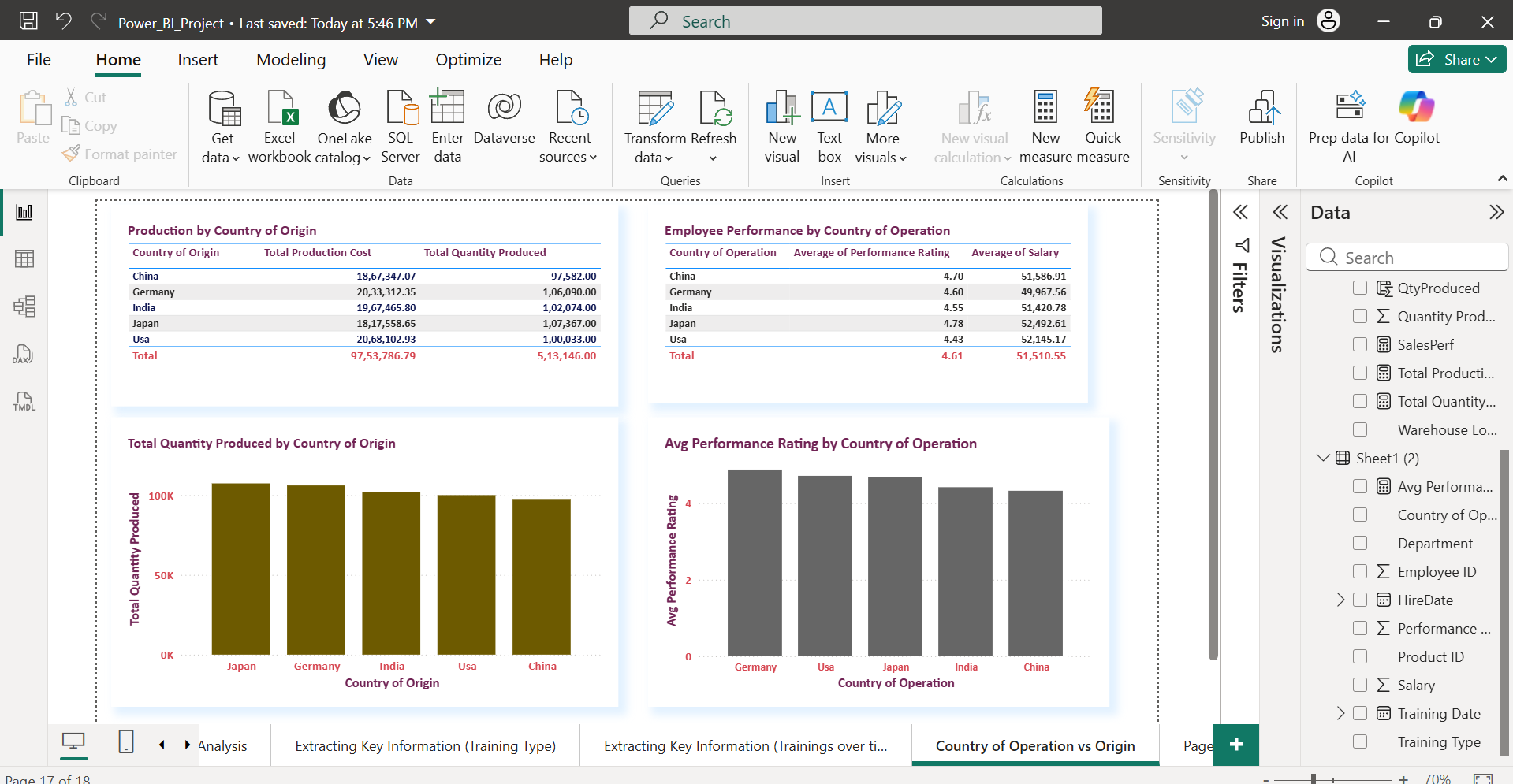
**A date‑based chart of Training Sessions Over Time uses Training Date on the axis and counts of sessions as values to reveal peaks and gaps in employee development activity across the years.**

### **20.Country of Operation vs. Country of Origin**

**Compare the countries of operation and origin in terms of production and employee performance.**

* The table and column chart ‘Production by Country of Origin’ show that all five origin countries (China, Germany, India, Japan, USA) contribute similar total quantities, with slight variations in total production cost.
* The table ‘Employee Performance by Country of Operation’ and the corresponding chart summarize average performance ratings and salaries across the same five operating countries, revealing only small differences in employee effectiveness and pay levels.
* Taken together, these visuals compare where products are manufactured (country of origin) versus where employees are based (country of operation), helping management see if high‑production origins align with strong workforce performance.

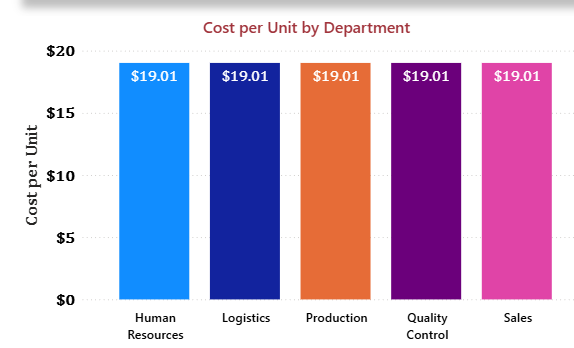
****

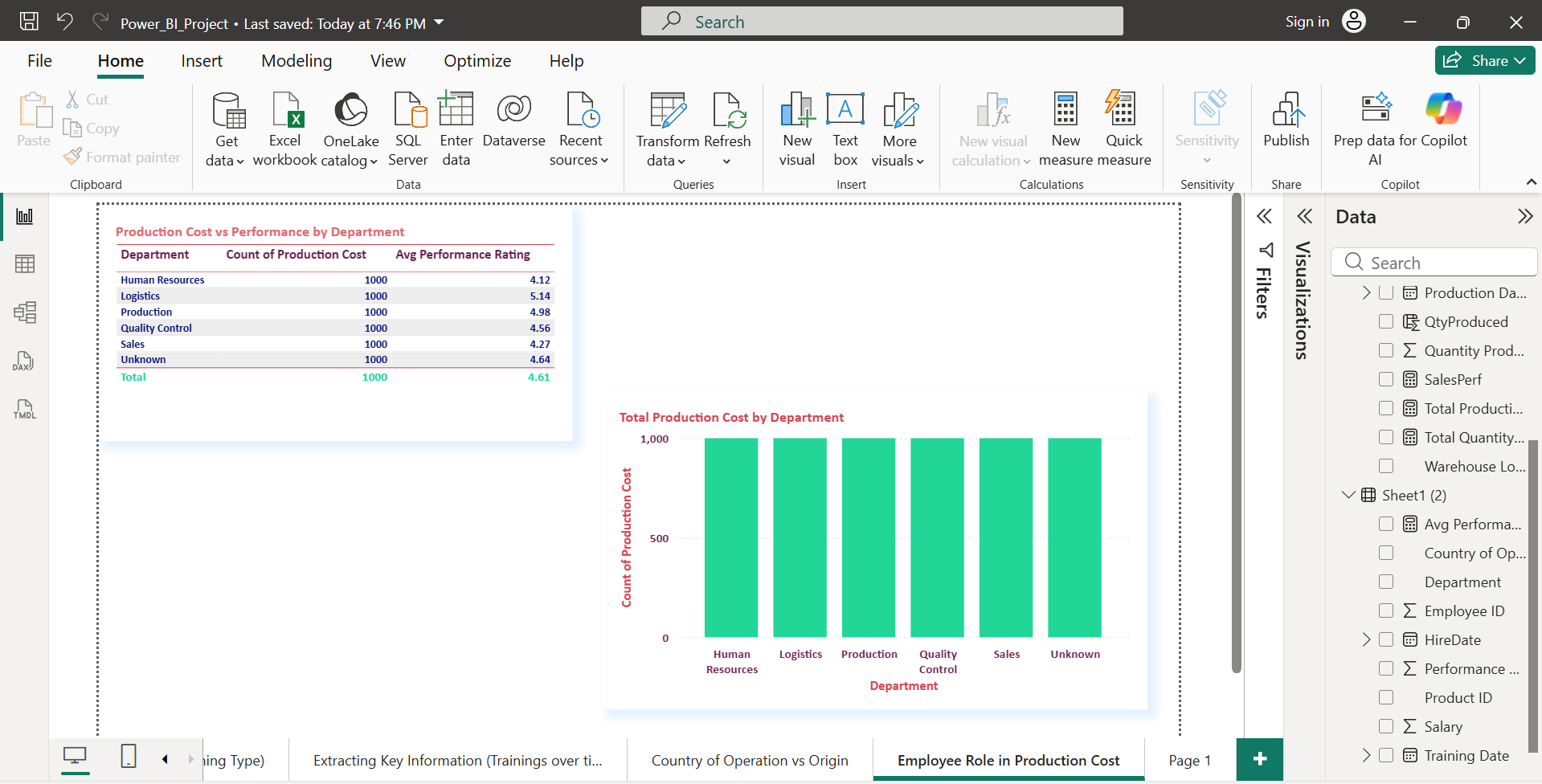


### **21. Employee Role in Production Cost**

**Analyze if certain departments or employee roles have a significant impact on production costs.**

* The bar chart **‘Total Production Cost by Department’** aggregates production cost from the manufacturing table through the ProductID relationship, highlighting which departments are associated with the highest manufacturing expenses.
* The table **‘Production Cost vs Performance by Department’** combines Total Production Cost with Average Performance Rating for each department, revealing how efficiently different roles convert cost into performance.
* Together, these visuals show that cost impact is spread relatively evenly across departments, with only small differences in average performance scores, indicating no single department is driving disproportionate production cost.

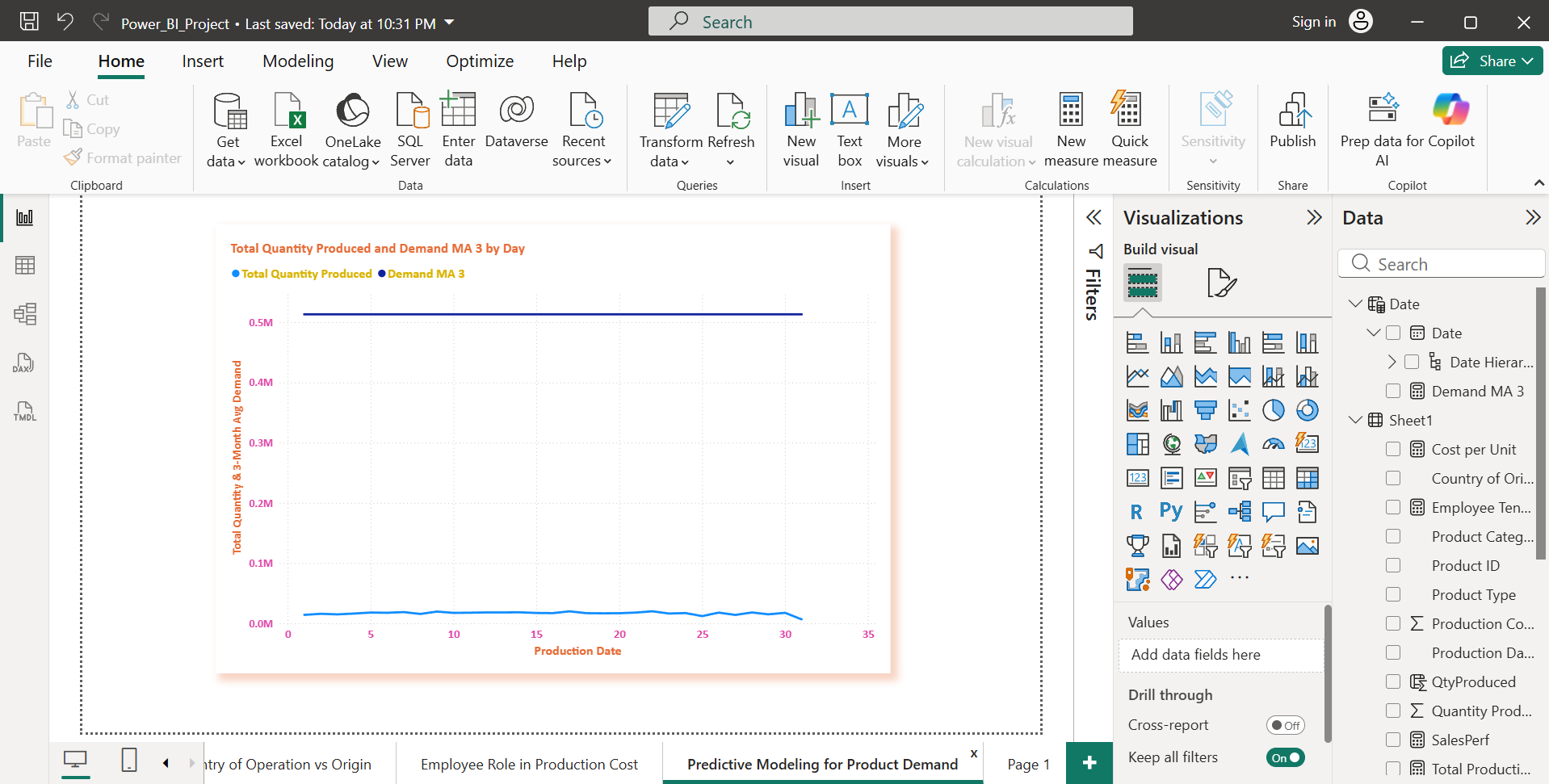




### **22. Complex DAX: Predictive Modeling for Product Demand**

**Use DAX to create a predictive model estimating future product demand based on historical production and sales data.**

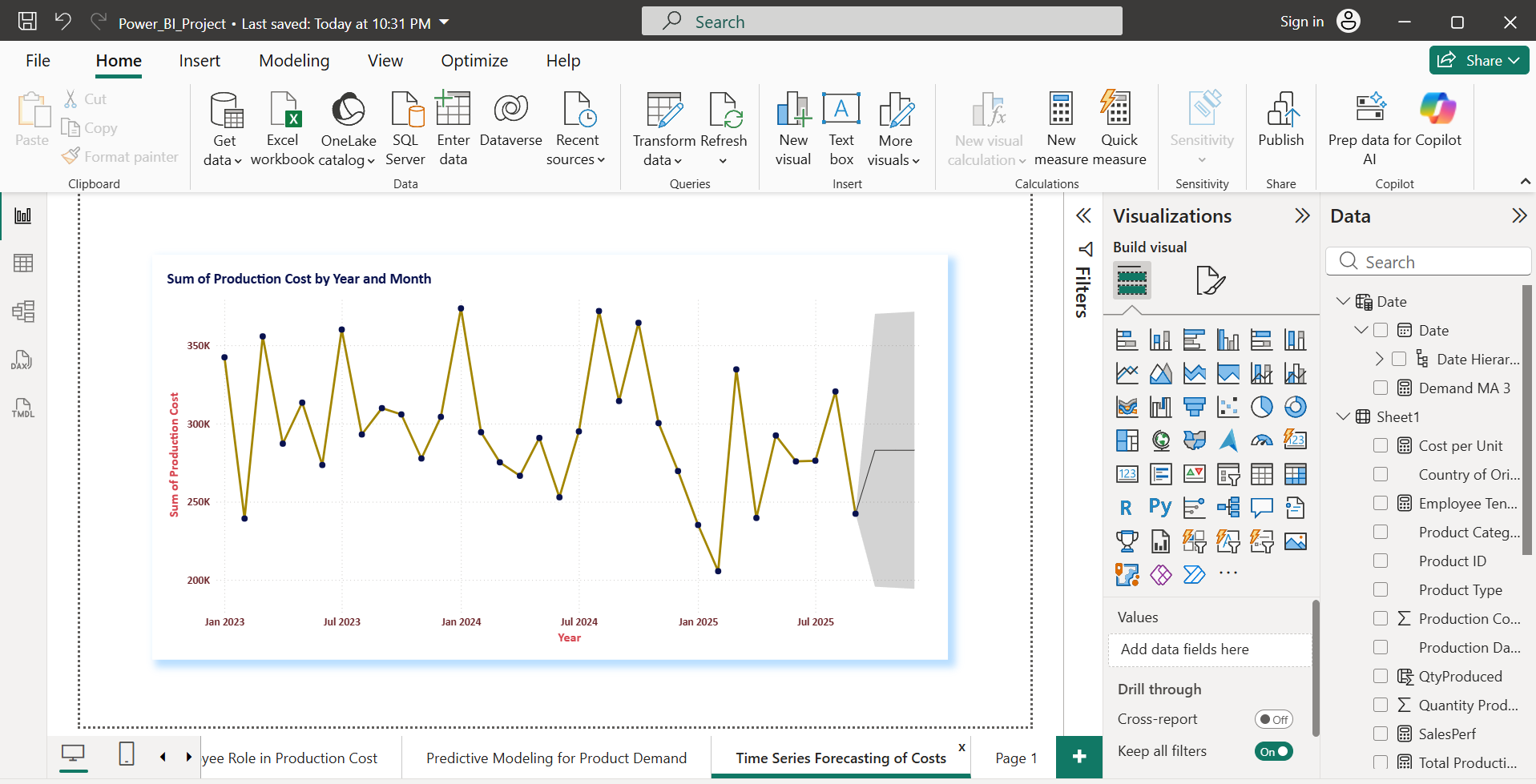
* This line chart compares actual daily production volume with a 3‑month moving‑average demand estimate, helping identify whether recent output is above or below the smoothed demand trend.
* A rising gap between the moving‑average line and actual production indicates potential over‑ or under‑production risk, which management can use for capacity and inventory planning.

****

### **23. Data Modeling: Time Series Forecasting of Costs**

**Perform time series forecasting of production costs using historical data. What are the predicted costs for the next quarter?**

Time series forecasting was performed using Power BI to predict production costs for the next quarter based on historical monthly trends.



### **24. Advanced Data Transformation: Identifying Production Anomalies**

**Using Power BI's data transformation capabilities, identify any anomalies in production data (e.g., unusually high costs, sudden spikes in production quantity).**

**A calculated column was created to classify production costs into normal and anomalous categories, as legend-based visual segmentation in Power BI requires categorical fields rather than measures.**

Advanced data transformation techniques were applied in Power BI to identify production anomalies. A star schema data model was designed with the production table as the fact table, connected to employee and date dimension tables.

Statistical DAX measures were created to calculate the average and standard deviation of production cost and quantity produced. Any values exceeding the threshold of Mean + 2 × Standard Deviation were classified as anomalies.

Line charts were used to visualize production cost and quantity trends over time, enabling the detection of sudden spikes and abnormal patterns. Additionally, tabular views with anomaly filters were created to isolate affected products and production dates. This approach helps in proactively identifying inefficiencies, controlling production costs, and improving operational performance.

