**Exercise: Profiling a dataset**

[Other Company Products](https://d3c33hcgiwev3.cloudfront.net/qN3VArnYRYGA7gwKxGcUMA_d943011ab38641d2967072272c49d1e1_Other-Company-Products.xlsx?Expires=1709856000&Signature=PcSBiQAkJsxGRTfTeC4McngfheVAu46qA53BdDnjw0XL~W4I-zBz1o-kfEJ-DC0UldJHxhuh3OhsB9sH5dYqchn01nsrP4DJxnMmP76udqVQYX6ytpoEopZs6b0txrL6z31ir7c1XU02vNftzUQTNkK98CAXCUDj2Hcu2yWwzvk_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/qN3VArnYRYGA7gwKxGcUMA_d943011ab38641d2967072272c49d1e1_Other-Company-Products.xlsx?Expires=1709856000&Signature=PcSBiQAkJsxGRTfTeC4McngfheVAu46qA53BdDnjw0XL~W4I-zBz1o-kfEJ-DC0UldJHxhuh3OhsB9sH5dYqchn01nsrP4DJxnMmP76udqVQYX6ytpoEopZs6b0txrL6z31ir7c1XU02vNftzUQTNkK98CAXCUDj2Hcu2yWwzvk_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

**Introduction**

By now, you should have a good understanding of the concept of profiling data and its practical applications to identify data anomalies. You covered the three primary data profiling operations, column quality, column distribution and column profile. In this exercise, you can apply your knowledge by using Microsoft Power Query to examine the valid, error, empty, min, max, unique, and distinct values in Excel spreadsheet rows, allowing you to identify the anomalies in the data.

**Case study**

Adventure Works has recently acquired another bicycle business. Adventure Works’ CEO, Jamie Lee, has assigned a task to the sales department to ensure that the product data from the newly acquired company is validated, revised in quality factors, and ready for importing to the current company products. Your manager, Adio Quinn, has requested that you examine the data using Power Query by the factors of profile, quality, and distribution.

Adventure Works has provided you with an Excel file containing the newly acquired company’s product data called *Other Company Products.xlsx*. The dataset has some empty data in its **ProductKey** column. You also need to assess the distribution of the products by the product categories and detect potential anomalies in the **Price** column. To complete your task successfully, you must import the Excel file, transform the data in Power Query and assess the **Column quality**, **Column distribution**, and **Column profile** options in the **Data Preview** group.

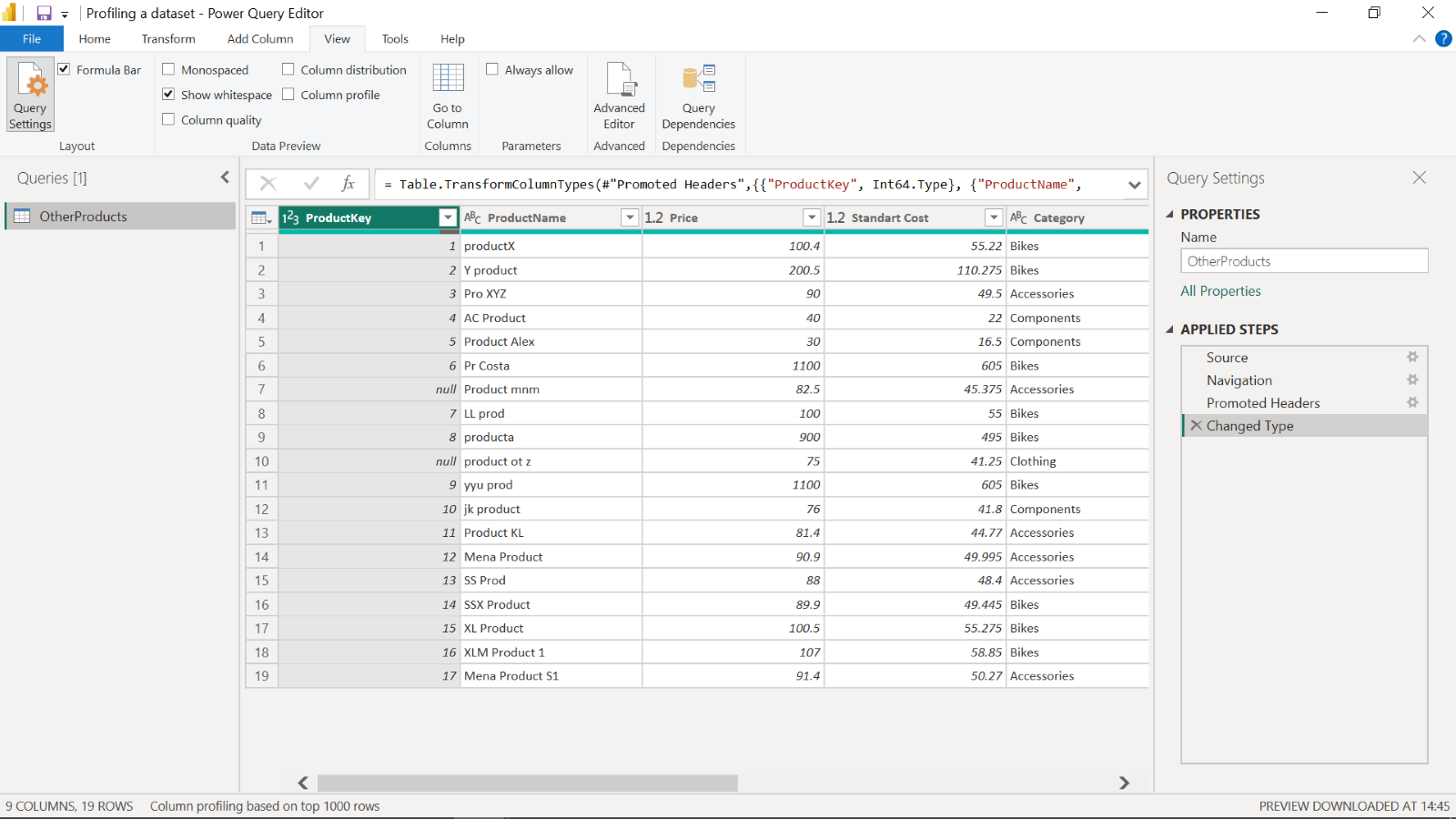
* This exercise aims to help you to understand how to identify data anomalies by profiling data.
* By the end of this exercise, you’ll understand how to profile data in Power Query, and how to identify data anomalies.

**Instructions**

Create a new Power BI project called **Exercise – Profiling a dataset.** Follow the prompts below to complete the exercise.

**Step 1: Download the Excel files**

Download the *Other Company Products.xlsx* file, which you will use in this exercise. This file is available at the top of this exercise.

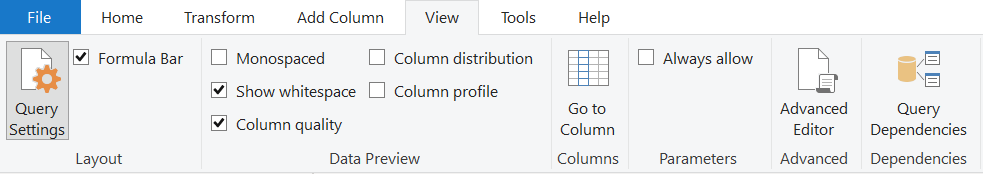


**Step 2: Open the Power Query Editor**

In Power BI, select **Get Data** and then select **Transform data** to open the Power Query editor and import your dataset, *Other Company Products*.

**Step 3: Detect empty values in ProductKey column:**

1. There are some empty values in the spreadsheet’s **ProductKey** column.
2. To detect empty and invalid values, you need to assess column quality, on the **View** ribbon tab, from inside the **Data Preview** group, check **Column Quality**. The column quality feature allows you to easily determine the percentage of valid, error, or empty values found in columns.



1. Note amount **Valid**, **Error** and **Empty** rows percentage values of the **ProductKey** column

**Step 4: Assess the distribution of product categories:**

1. There are many categories in **Product** list and you need to find out how the data is distributed by the category data.
2. To assess column distribution, on the **View** ribbon tab, from inside the **Data Preview** group, check **Column Distribution** and note the amount of **distinct** values and **unique** values. Check **Column Profile** while keeping **Column Distribution** checkbox as checked and note the number of Bikes, Accessories, Components and Clothing.

**Step 5: Detect potential anomalies in the Price column**

1. You have to assess **Price** column in the **Product** list and you need to find out **min**, **max**, **mean** values and the distribution of the values.
2. To detect potential anomalies and assess column distribution for the Price column, on the **View** ribbon tab, from inside the **Data Preview** group, check **Column Profile** while keeping **Column Distribution** checkbox as checked.
3. Note the min, max, mean values for the **Price** column and also note assess the column distribution.

**Conclusion**

Well done. You have now successfully assessed your datasets by the factors of profile, quality and distribution and detected potential anomalies using Power Query.

**Exemplar: Profiling a dataset**

[Other Company Products](https://d3c33hcgiwev3.cloudfront.net/qN3VArnYRYGA7gwKxGcUMA_d943011ab38641d2967072272c49d1e1_Other-Company-Products.xlsx?Expires=1709856000&Signature=PcSBiQAkJsxGRTfTeC4McngfheVAu46qA53BdDnjw0XL~W4I-zBz1o-kfEJ-DC0UldJHxhuh3OhsB9sH5dYqchn01nsrP4DJxnMmP76udqVQYX6ytpoEopZs6b0txrL6z31ir7c1XU02vNftzUQTNkK98CAXCUDj2Hcu2yWwzvk_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/qN3VArnYRYGA7gwKxGcUMA_d943011ab38641d2967072272c49d1e1_Other-Company-Products.xlsx?Expires=1709856000&Signature=PcSBiQAkJsxGRTfTeC4McngfheVAu46qA53BdDnjw0XL~W4I-zBz1o-kfEJ-DC0UldJHxhuh3OhsB9sH5dYqchn01nsrP4DJxnMmP76udqVQYX6ytpoEopZs6b0txrL6z31ir7c1XU02vNftzUQTNkK98CAXCUDj2Hcu2yWwzvk_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

**Overview**

In the exercise *Profiling a dataset,* you put into practice your understanding of how to profile a data set and detect potential anomalies in Power Query.

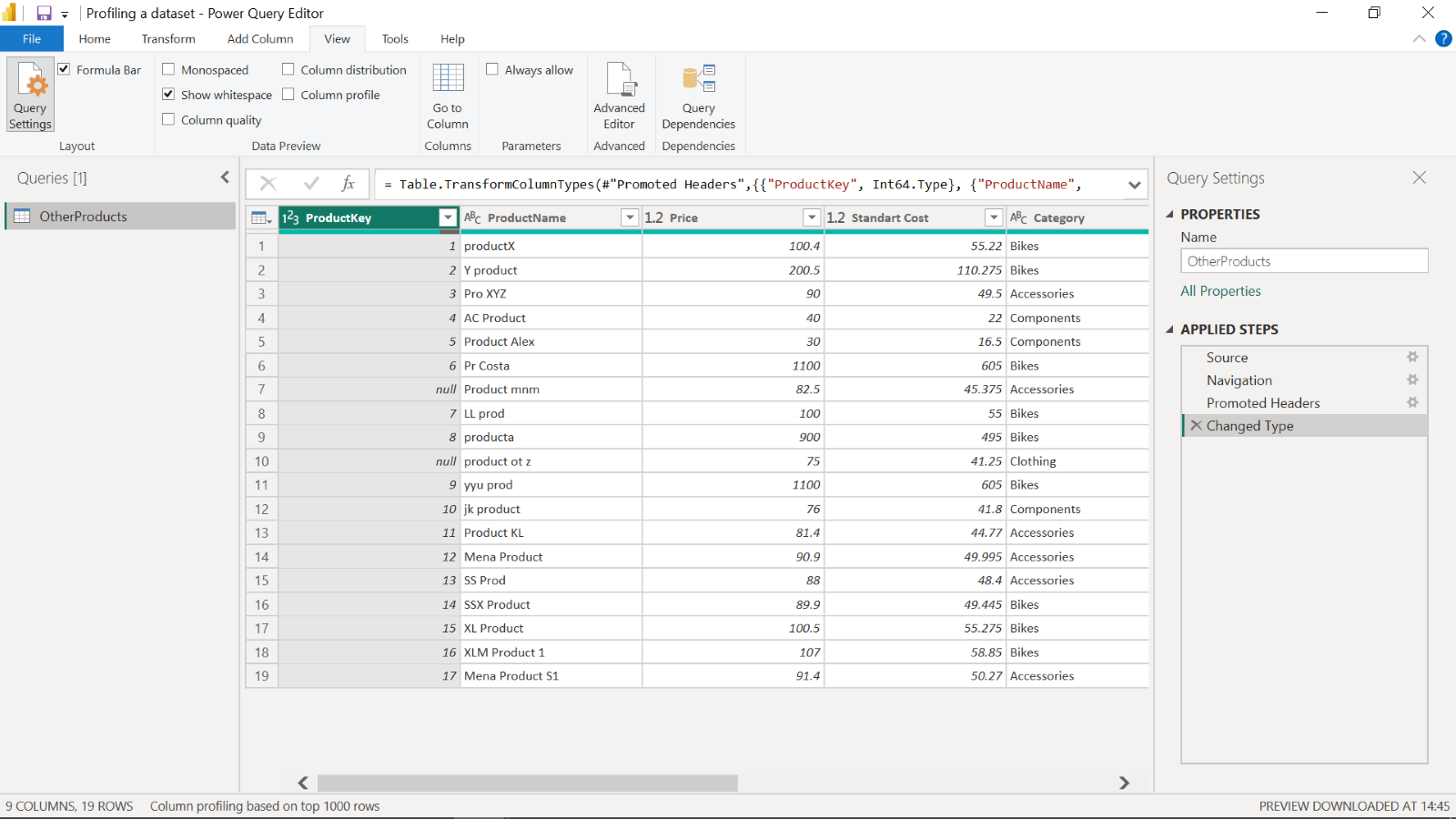
Your objective for this exercise was to prepare a worksheet for analysis by completing the following tasks:

* Download the Excel Files.
* Open the Power Query Editor.
* Detect empty values in the **ProductKey** column.
* Assess the distribution of product categories.
* Detect potential anomalies in the **Price** column.

This reading provides a step-by-step guide for completing these tasks, accompanied by screenshots for easy comparison with your own copy.

**Step 1: Download Excel files**

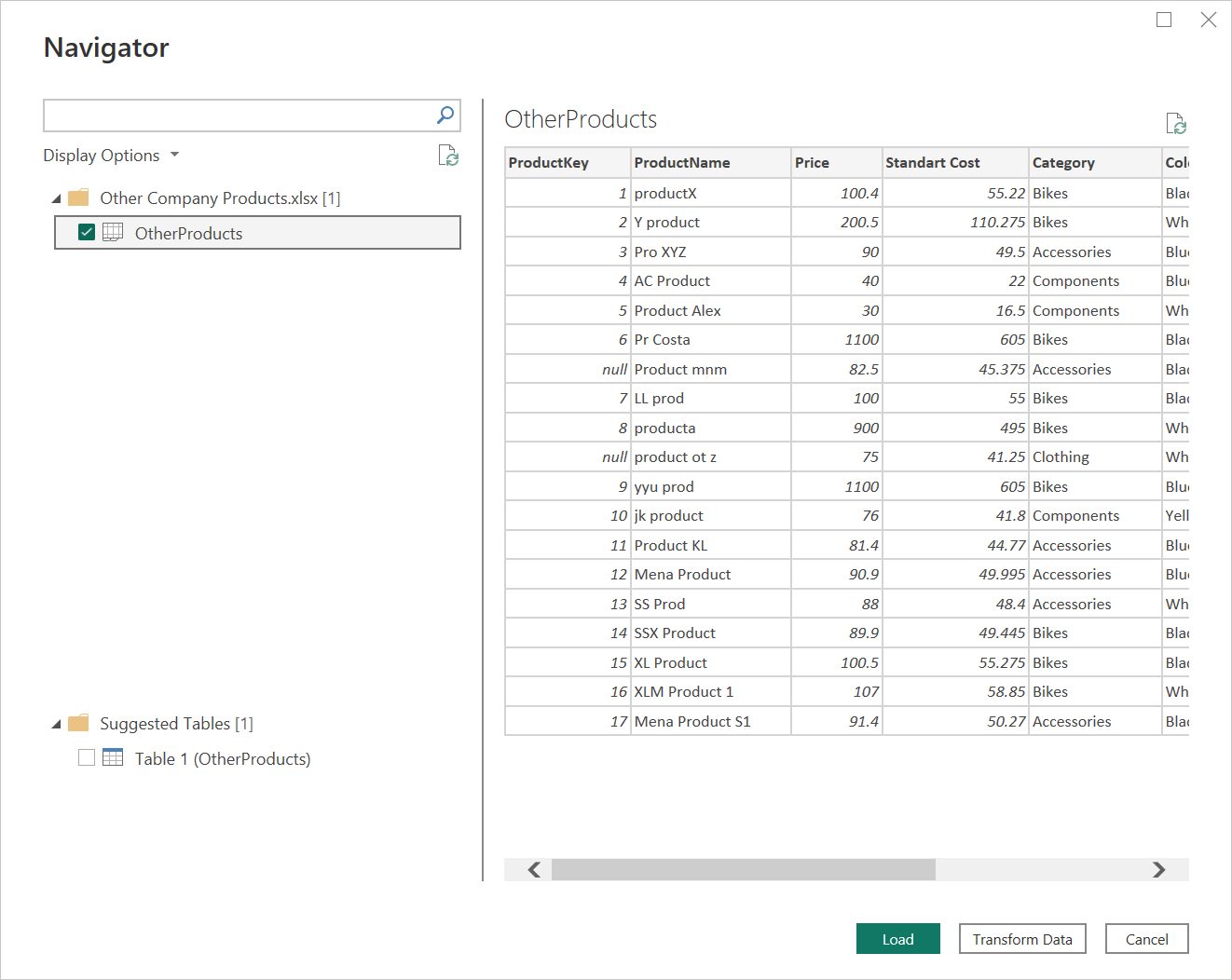
* Download the *Other Company Products.xlsx* file, which you will use in the exercise.The file is available at the top of this reading.



**Step 2: Open the Power Query Editor**

Open the Power Query editor and import your dataset, *Other Company Products*.

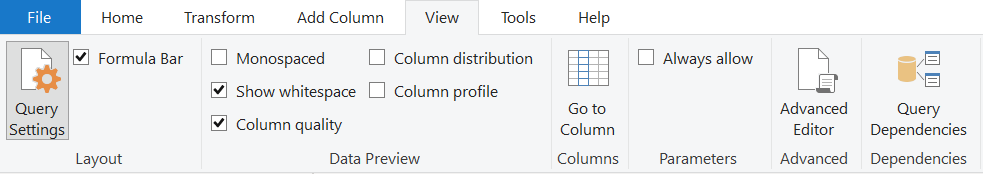
1. Navigate to the **Home** ribbon tab at the top of the **Power BI** window.
2. Select the **Excel Workbook** button in the **Data** group, in the middle of the toolbar.
3. Select *Other Company Products.xlsx* and select **Transform Data** in the opened window.



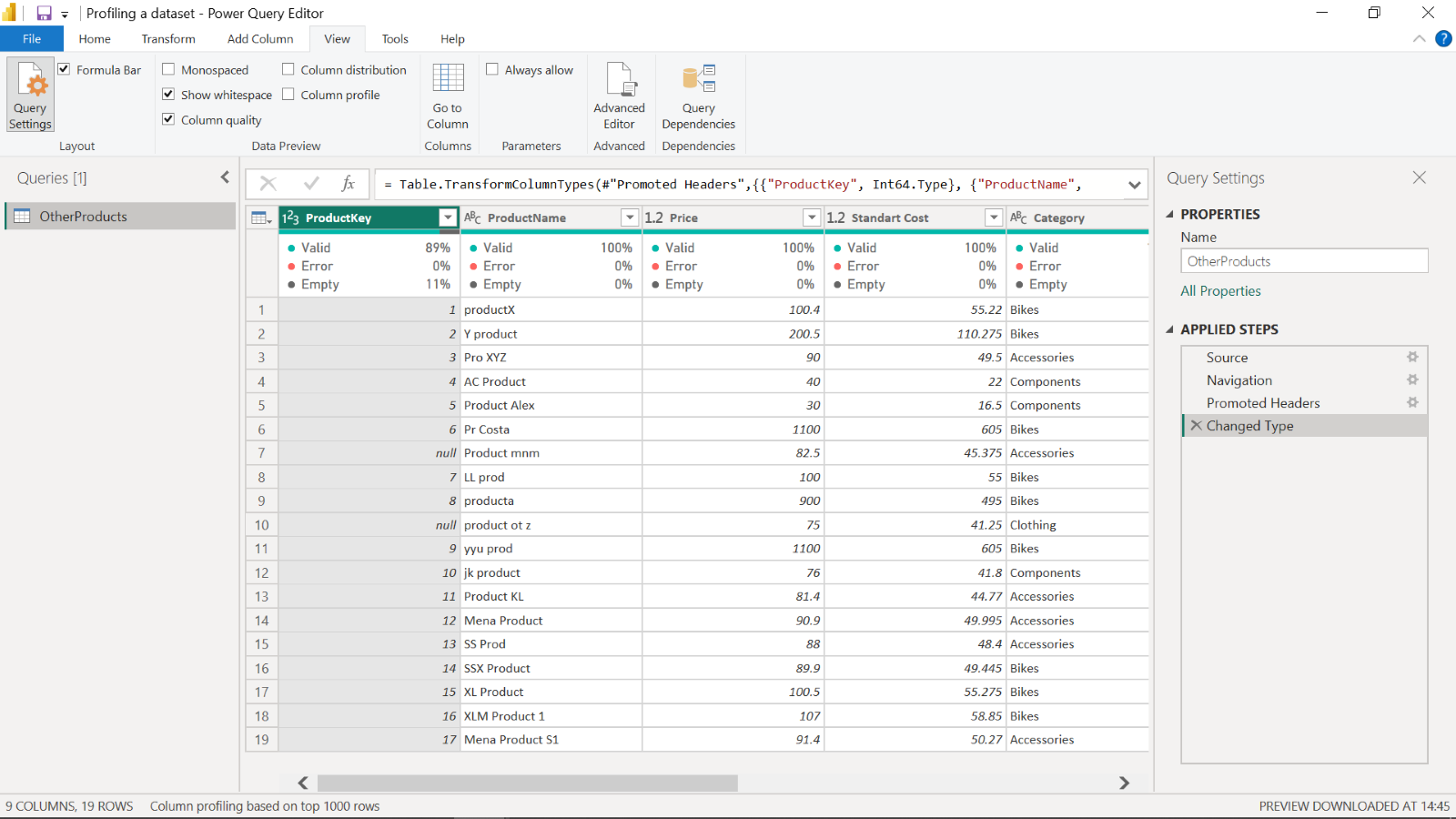
1. The **Power Query Editor** window opens. You can now begin profiling the data.

**Step 3: Detect empty values in ProductKey column**

1. There are empty values in **ProductKey** column.
2. To detect empty and invalid values, you need to assess column quality, on the **View** ribbon tab, from inside the **Data Preview** group, and select the **Column Quality** checkbox. **The column quality feature allows you to easily** determine the percentage of valid, error, or empty values found in columns.

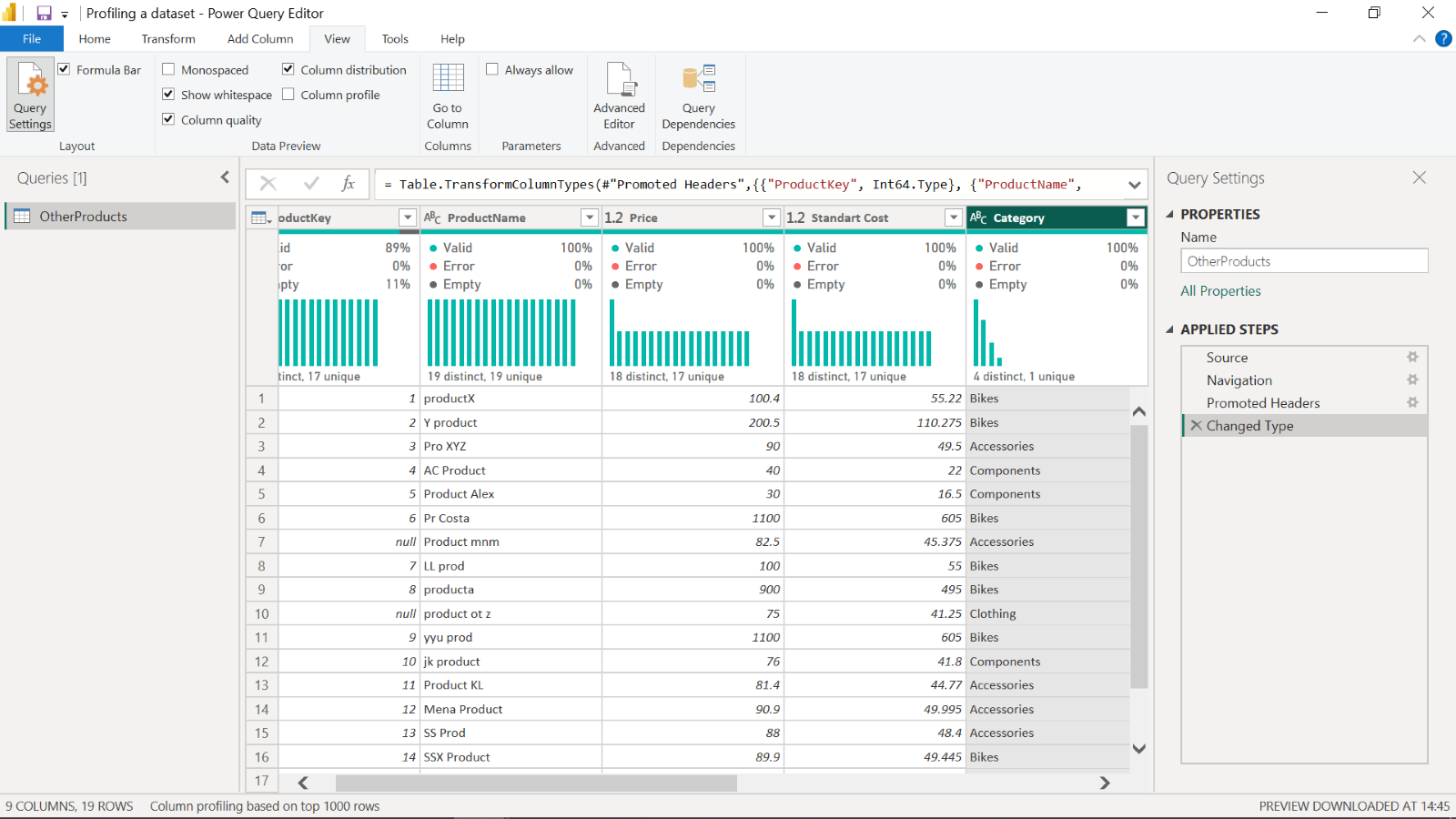


1. Note that 89% of the values are **Valid**, 0% of the values are **Error** and 11% of the values are **Empty** rows in the **ProductKey** column.

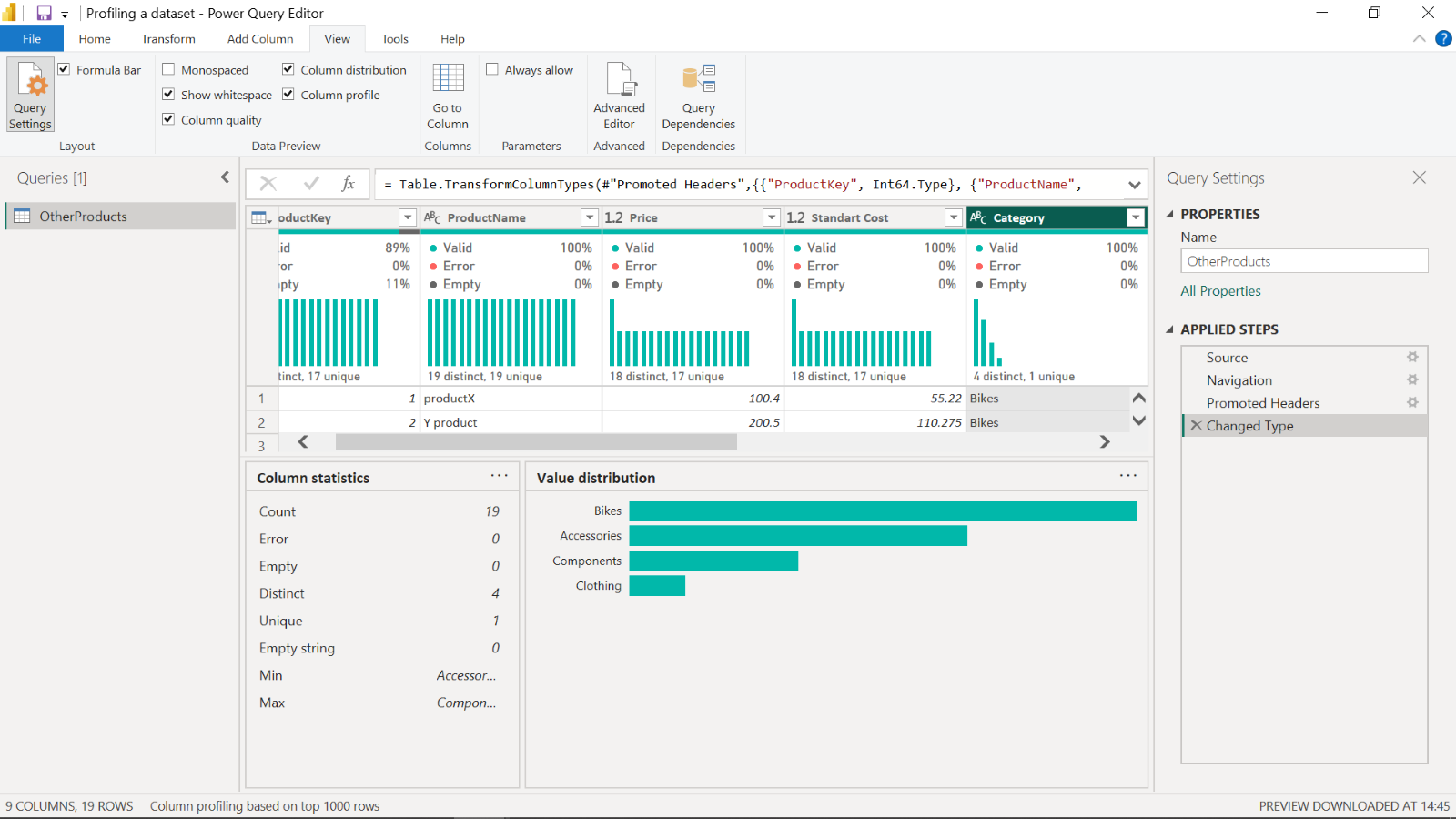


**Step 4: Assess the distribution of product categories**

1. There are many categories in the **Product** list, and you need to find out how the data is distributed by the category data.
2. To assess column distribution, on the **View** ribbon tab, from inside the **Data Preview** group, check the **Column Distribution** checkbox. Note that there are 4 **distinct** values and 1 **unique** value.

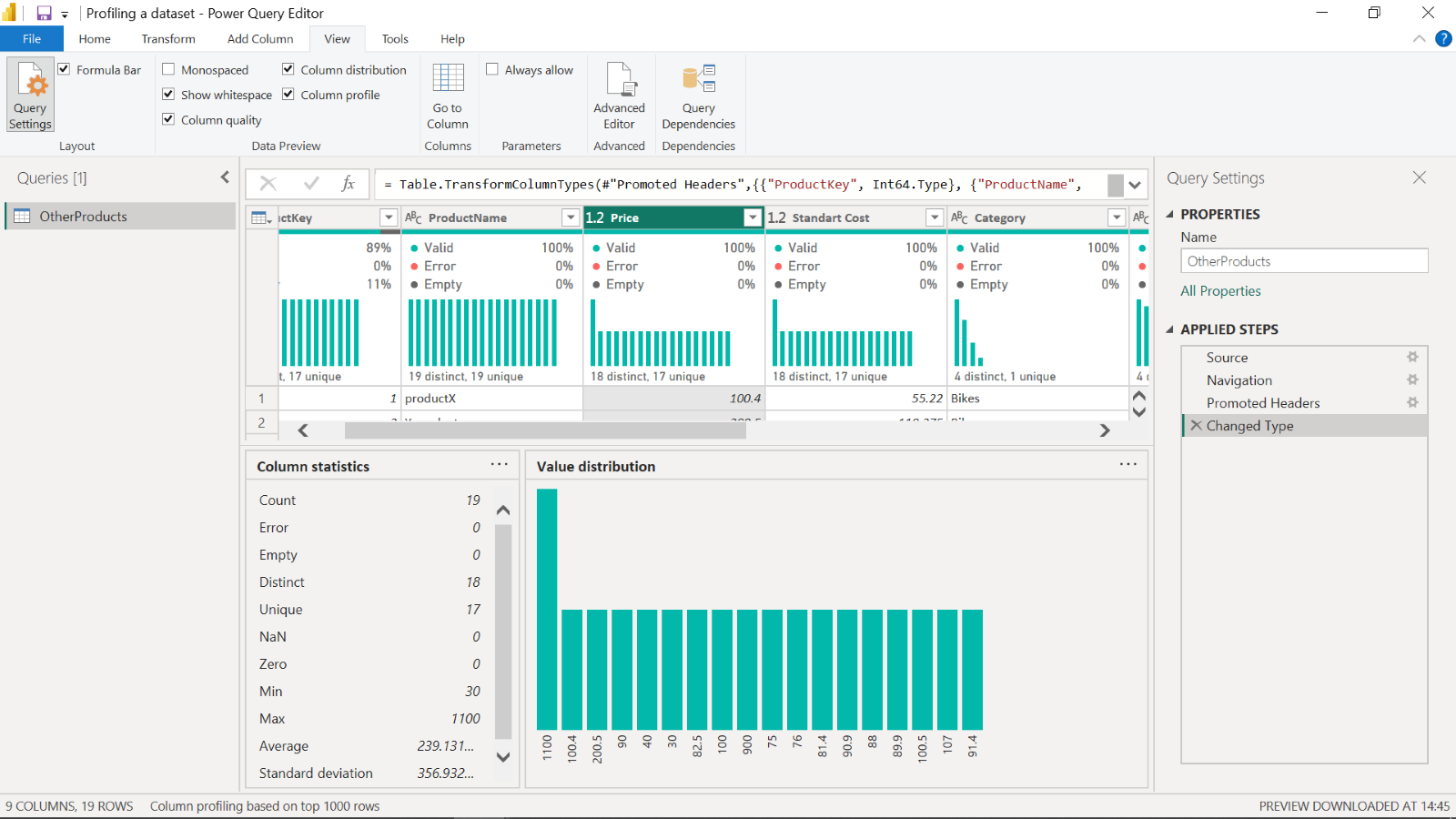


1. Check the **Column Profile** checkbox, while keeping **Column Distribution** checkbox as checked. Note that there are 9 Bikes, 6 Accessories, 3 Components and 1 Clothing categories.



**Step 5: Detect potential anomalies in the price column**

1. You have to assess the **Price** column in **Product** list and you need to find out min, max, mean values and the distribution of the values. To detect potential anomalies and assess column distribution for the **Price** column, on the **View** ribbon tab, from inside the **Data Preview** group, check **Column Profile** while keeping **Column Distribution** checkbox as checked.
2. Note that min value is 30, max value is 1100, and average is 239.13 for the **Price** column.
3. There are 18 distinct and 17 unique values, which means there are 2 products with the same price.
4. When you assess the value distribution, it can be considered as normal distribution and there are some outliers like 30, 40, 900 and 1100. The aim of this assessment is to find a potential anomaly in the **Price** column.



**Conclusion**

Your objective for this exercise was to apply techniques for profiling data to detect potential anomalies. In this context, you learned how to use Power Query Editor to check column quality, column distribution and column profile.