DATABRICKS Assessment 1

Q) Exploratory data analysis (EDA) in Databricks &Visualizing data in Databricks.

Exploratory Data Analysis:

Exploratory Data Analysis (EDA) is a process of describing the data by means of statistical and visualization techniques in order to bring important aspects of that data into focus for further analysis. This involves inspecting the dataset from many angles, describing & summarizing it without making any assumptions about its contents.

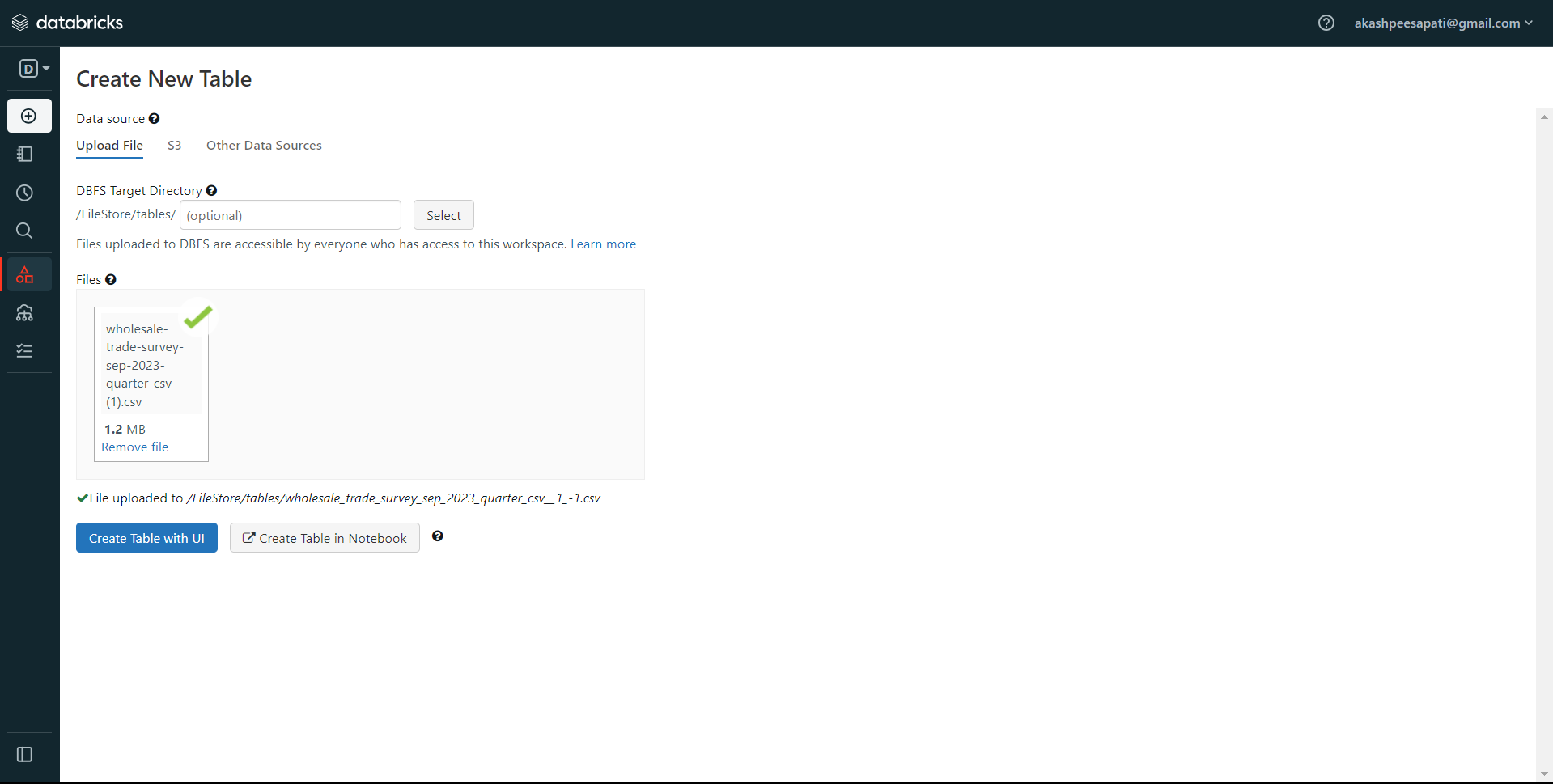
Databricks provides a unified analytics platform built on top of Apache Spark, which allows data scientists, analysts, and engineers to perform EDA efficiently at scale.

The steps involved in EDA using Databricks are shown below:

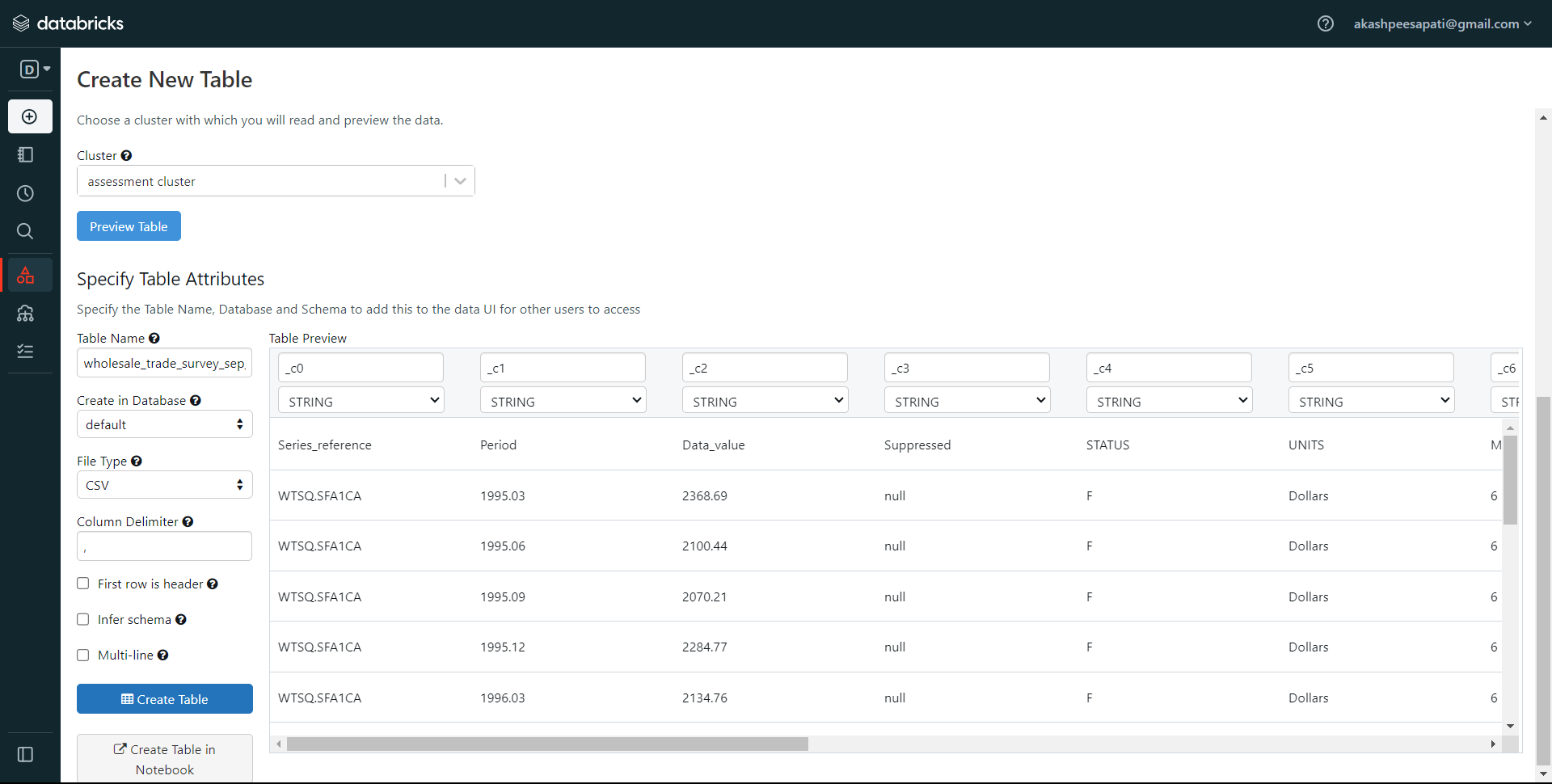
* **Data Loading:** In Databricks, you can bring data from different places like databases, data lakes, or cloud storage right into Spark DataFrames.
* **Statistical Summary:** With Databricks, users can quickly get an overview of their data by generating summary statistics and descriptive analytics. You can easily create histograms, box plots, and other visualizations to understand how the data is structured and distributed.
* **Data Visualization:** Databricks lets you use popular visualization libraries like Matplotlib, Seaborn, and Plotly. You can make histograms, box plots, scatter plots, and correlation matrices to see how your data is distributed and find relationships between different variables.
* **Data Cleaning:** During EDA, it's common to spot and manage missing values, outliers, and data inconsistencies. Databricks provides robust tools for data manipulation and transformation using Spark SQL and DataFrame APIs.
* **Feature Engineering:** Feature engineering involves creating new features from existing ones to improve model performance. Databricks offers functions for feature creation and transformation, enabling you to derive meaningful features from your data.
* **Data Transformation:** Perform data transformations such as scaling, normalization, or encoding categorical variables using Databricks DataFrame functions or MLlib transformers.
* **Interactive Exploration:** Make the most of Databricks' interactive notebooks to explore your data step by step, create visualizations, and improve your analysis as you uncover new insights along the way.

Lets execute this process using a sample csv file obtaining from any website.

**Step1:** Lets obtain a csv file named “wholesale\_trade\_survey\_sep\_2023\_quarter” and lets create the table using this csv file



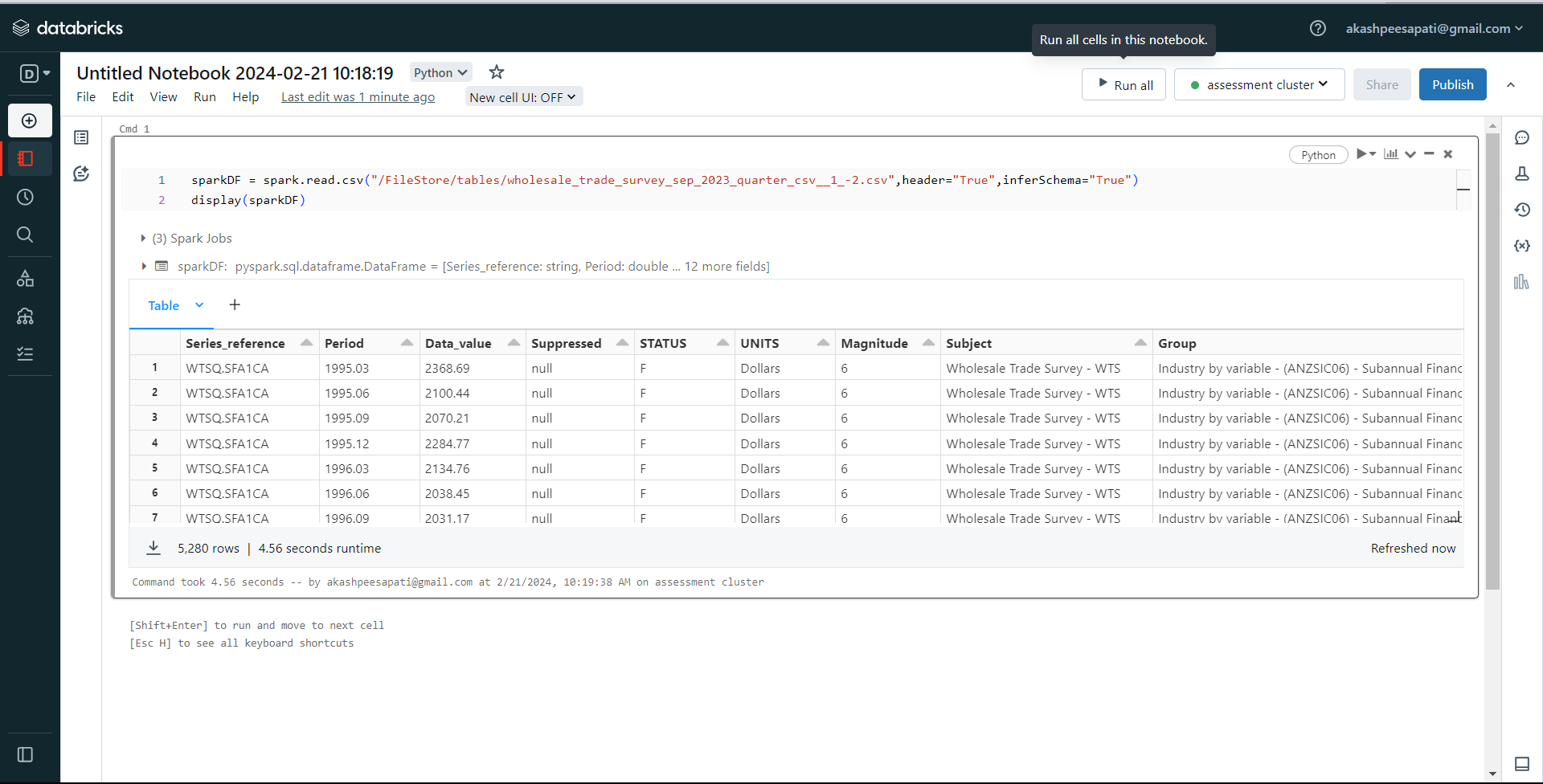
Now lets preview the table by selecting “Create table with UI”



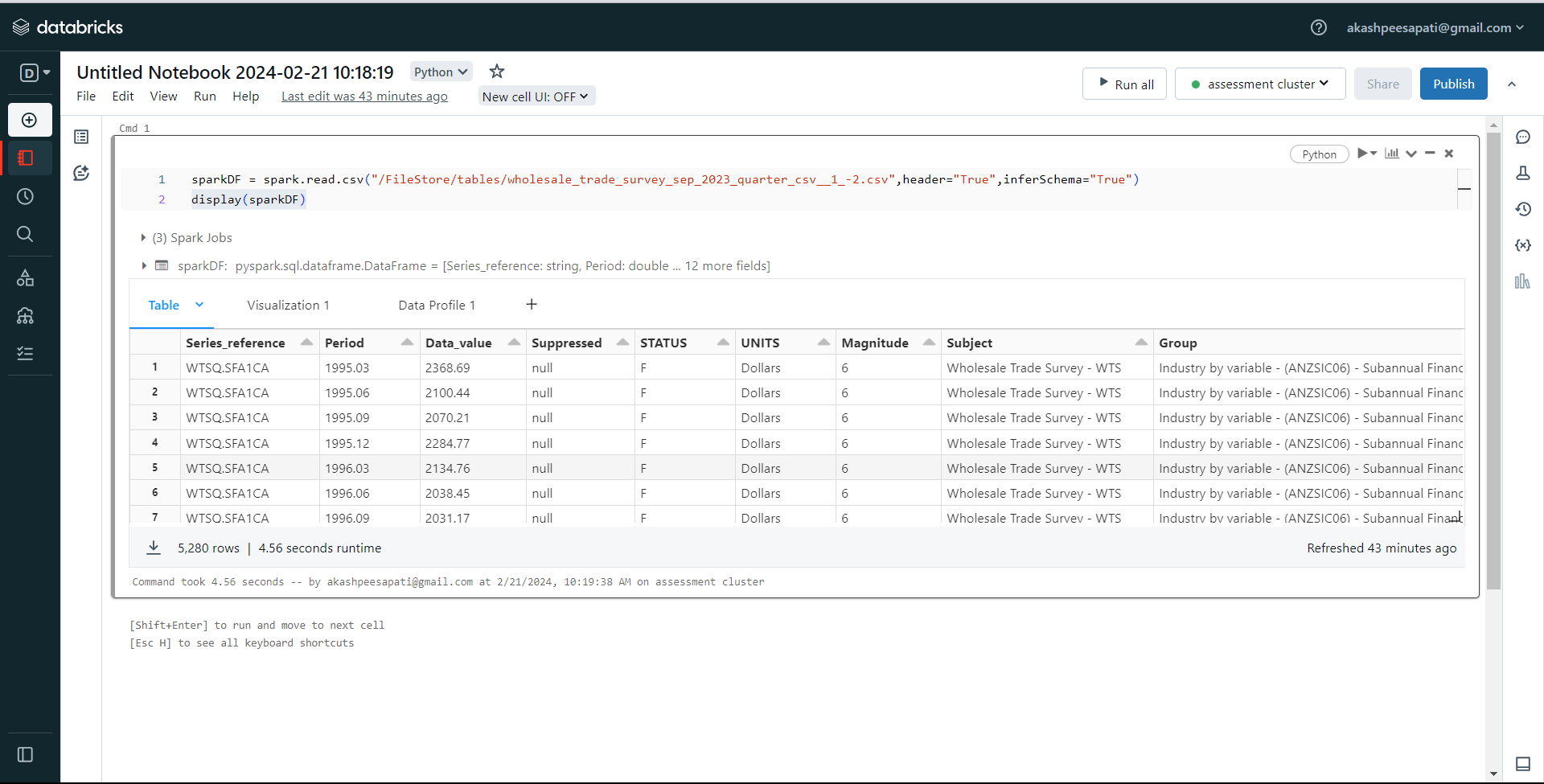
We have uploaded a csv file and we preview the table successfully.

**Step2:** Now lets create a notebook where we need to the display the table data by writing the code in the notebook. Here we have copied the file path ie “/FileStore/tables/wholesale\_trade\_survey\_sep\_2023\_quarter\_csv\_\_1\_-2.csv” and we displayed it by using “displayDF”.

The display(sparkDF) command is typically used in Databricks notebooks to visualize the contents of a Spark DataFrame named sparkDF.

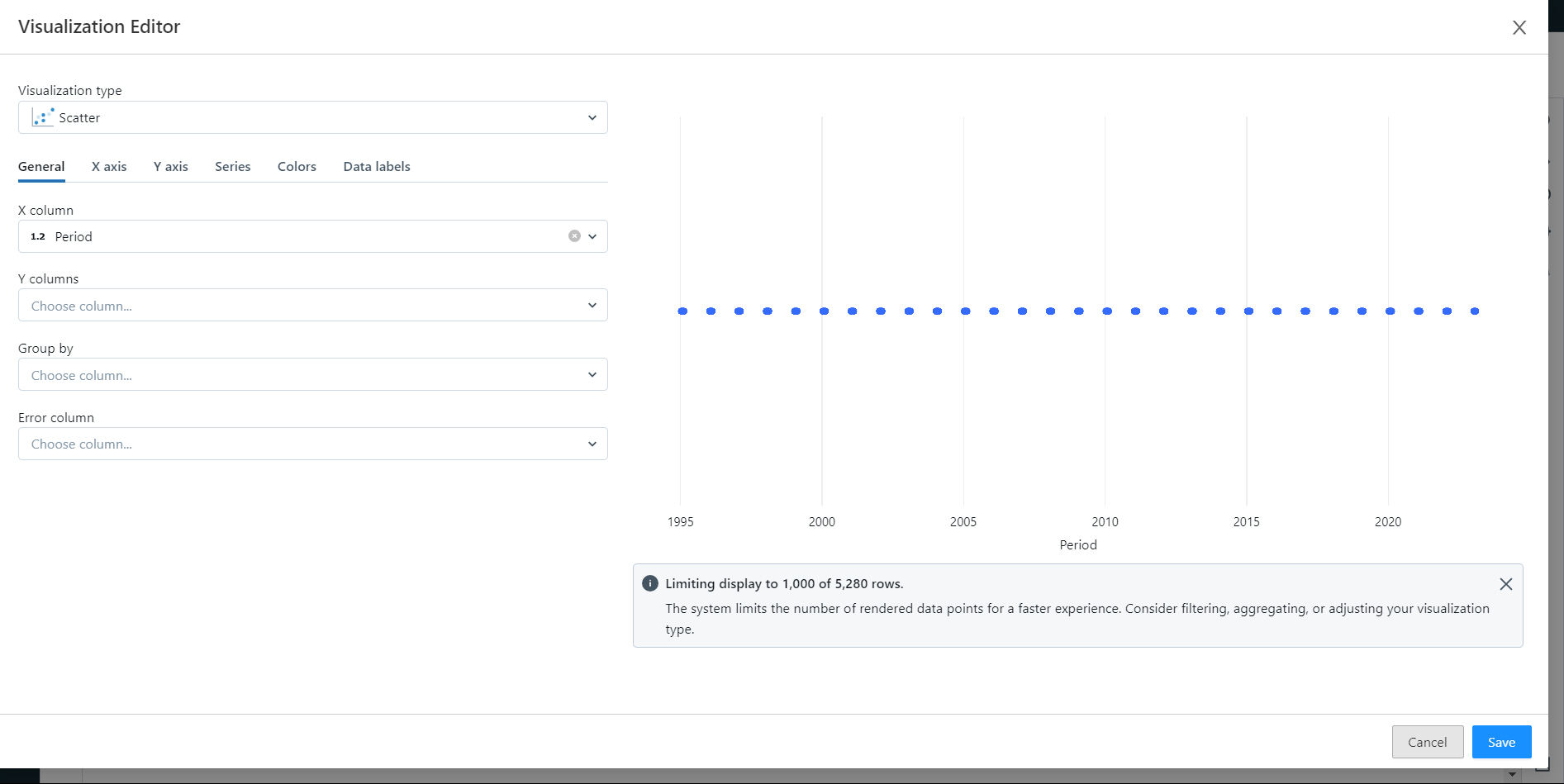


**STEP3:** Now we need to visualize the data. For that, click on the plus sign (+) and choose "Visualization.

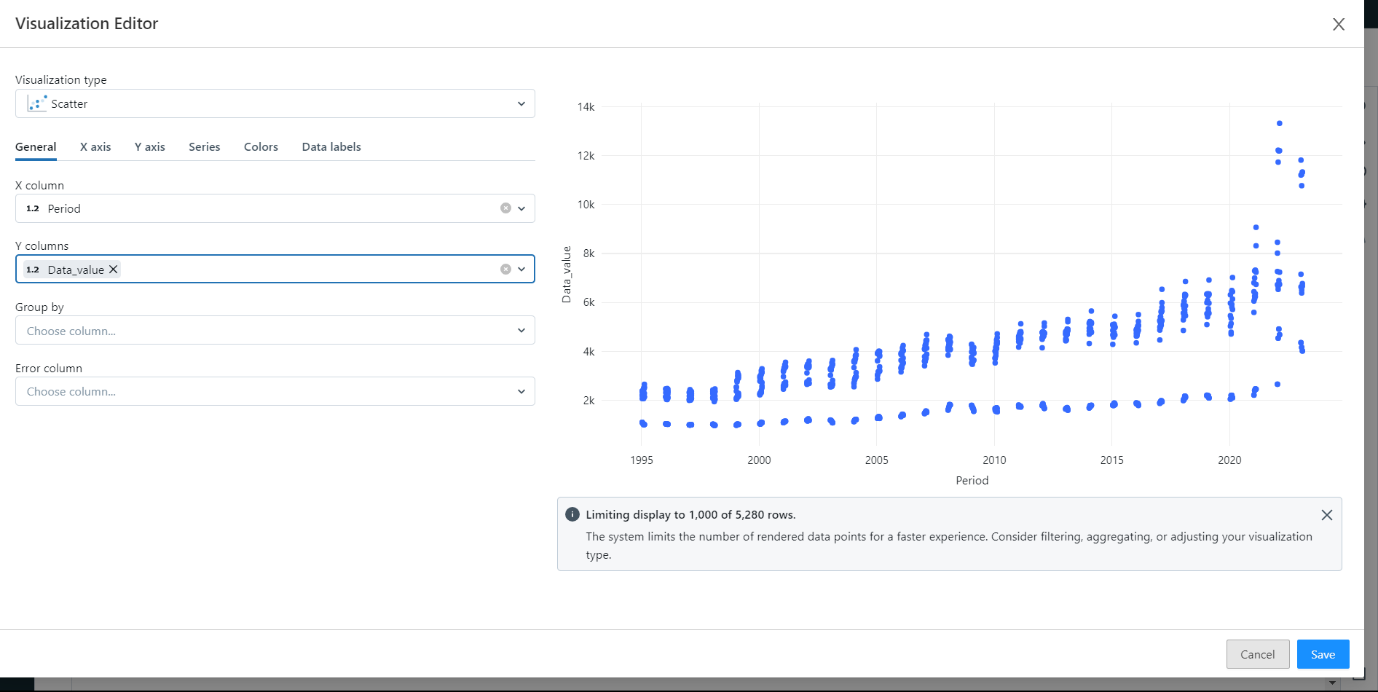


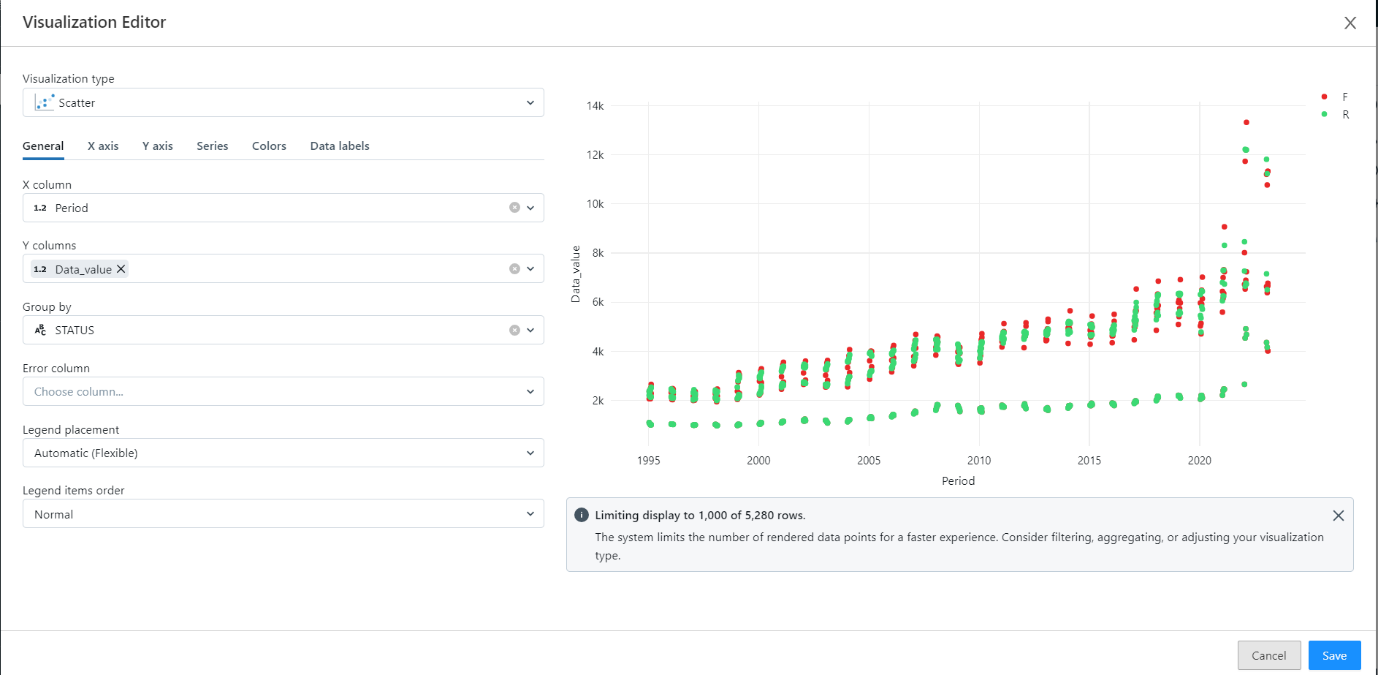
**STEP 4:** Next, select the type of visualization, we used “Scatter” and designate the 'X' and 'Y' columns. Additionally, you have the option to select a "Group by" parameter if needed.

**X-AXIS:**

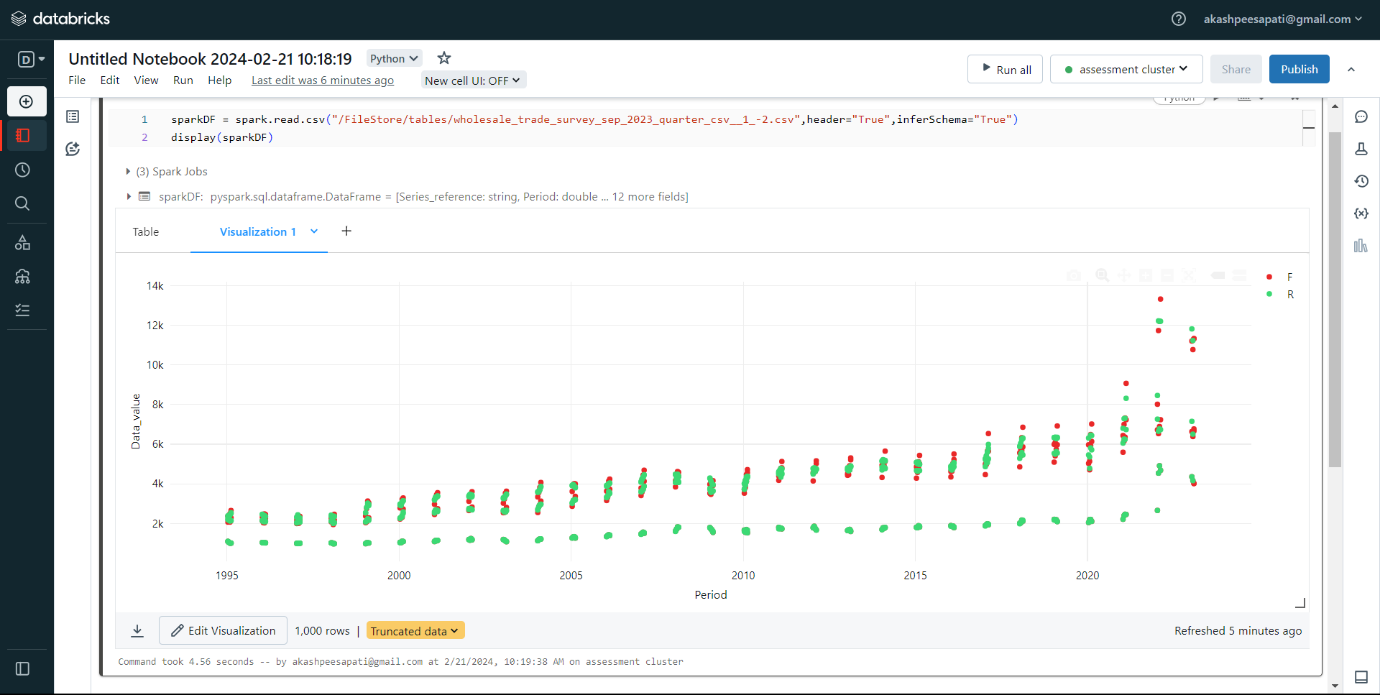


**Y-AXIS:**

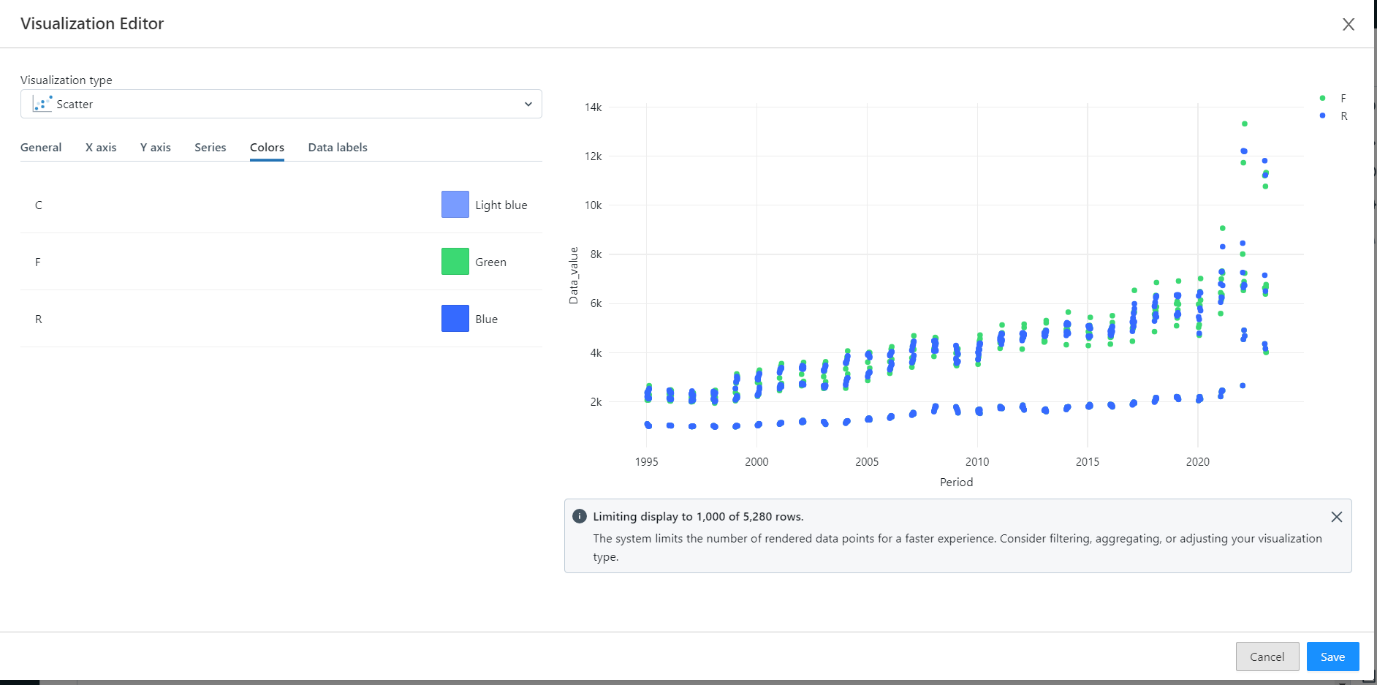


**GROUPBY:**  


**STEP 5:** After customizing your visualization, save it for future reference.

**STEP 6:** The image below illustrates the data visualization process.  
  


**STEP 7:** We can also change colours if we need by selecting desired colours



**STEP 8:** Also, we can observe Data Profile and the image is shown below:

