Exploratory Data Analysis (EDA) in Databricks:

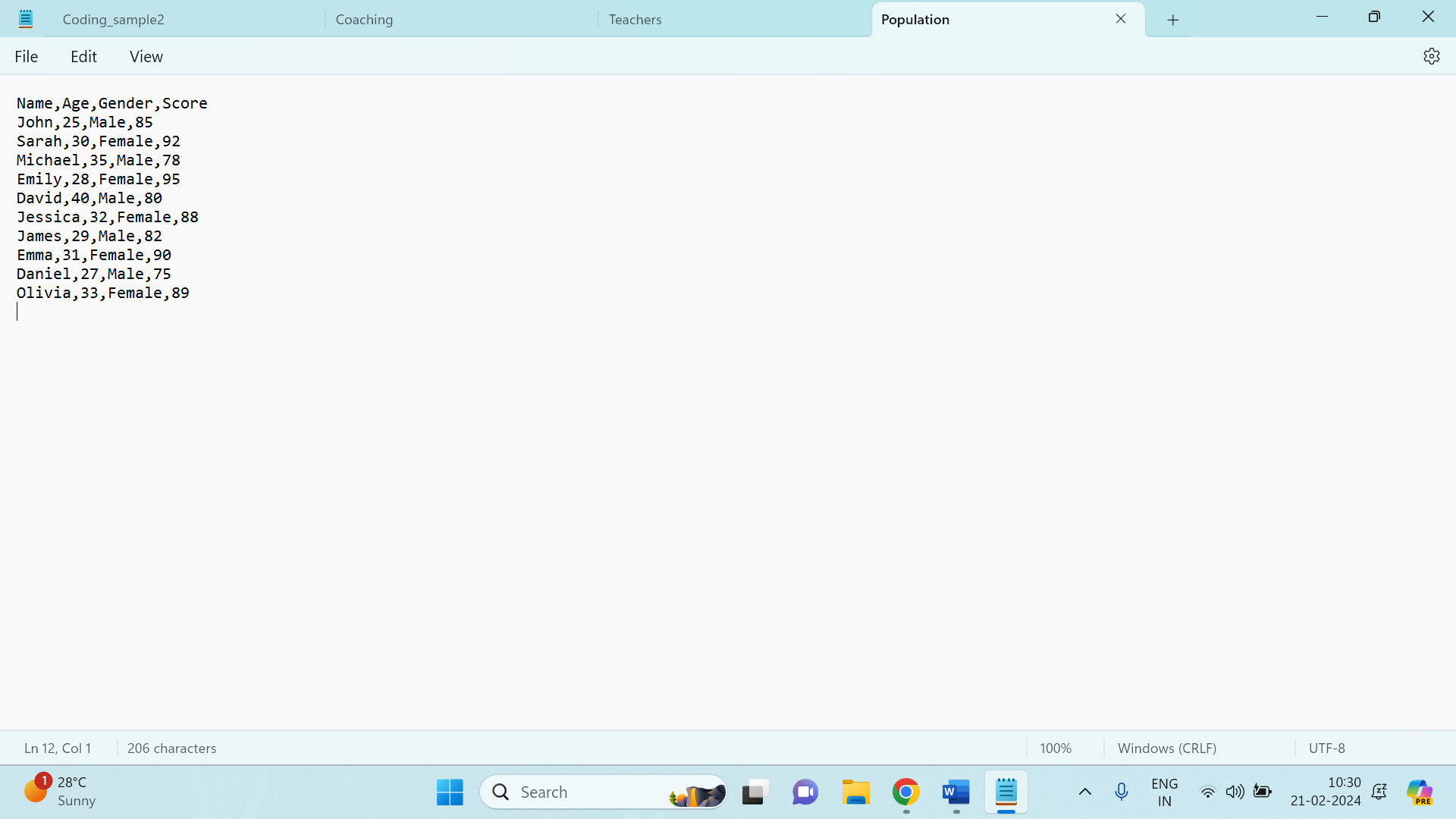
Databricks provides a powerful platform for conducting exploratory data analysis (EDA) on large datasets using Apache Spark. Exploratory Data Analysis (EDA) in Databricks refers to the process of visually and statistically exploring datasets to understand their key characteristics, identify patterns, detect anomalies, and formulate hypotheses for further analysis. 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.

Here are some common steps involved in EDA using Databricks:

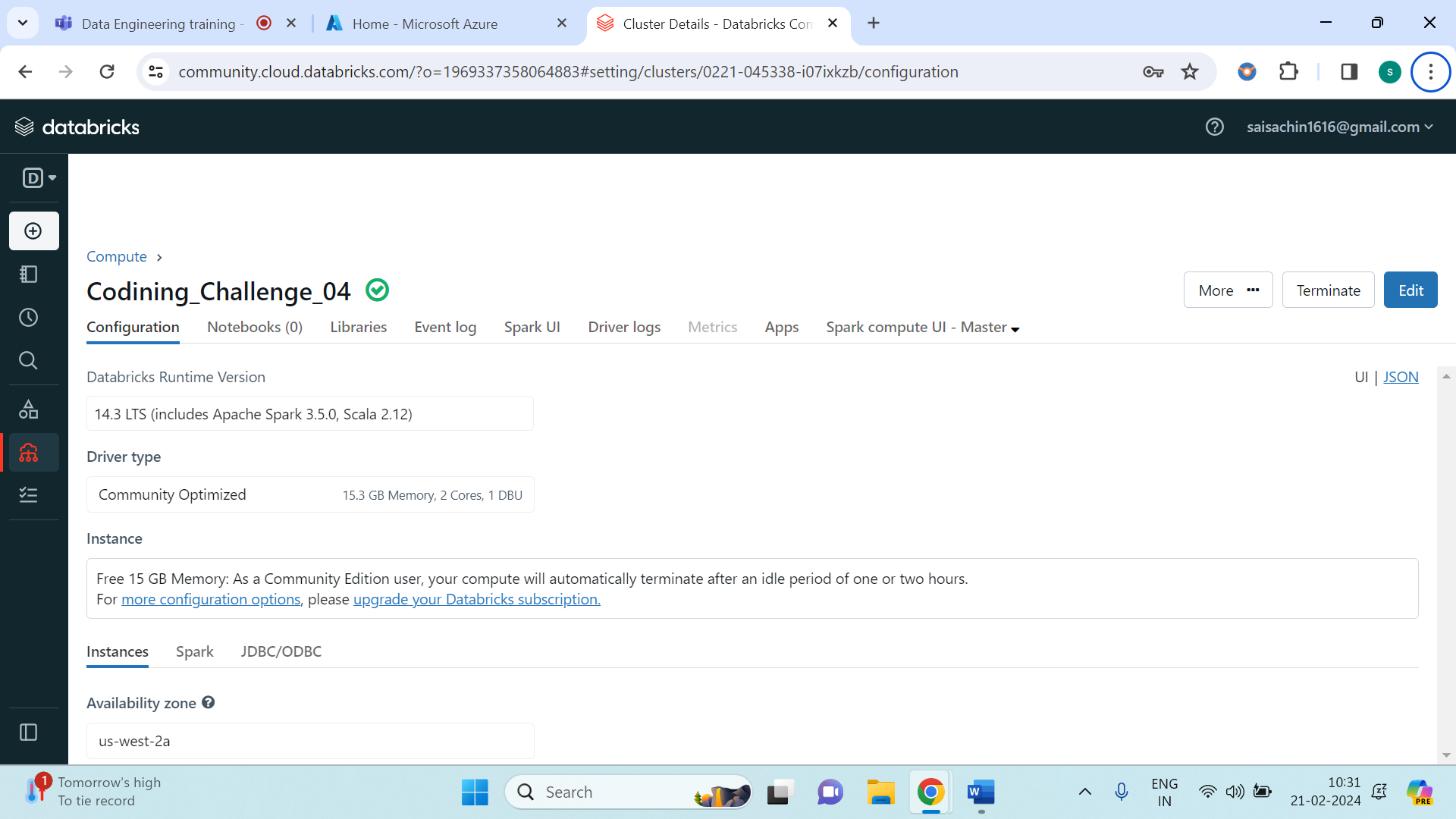
* Data Loading: Databricks allows users to load data from various sources such as databases, data lakes, or cloud storage into Spark DataFrames.
* Data Profiling: Users can generate summary statistics and descriptive analytics to understand the structure and distribution of data. Databricks offers tools for generating summary statistics, histograms, box plots, and other visualizations.
* Data Visualization: Databricks provides integration with popular visualization libraries like Matplotlib, Seaborn, and Plotly, allowing users to create insightful plots and charts to visualize patterns and relationships within the data.
* Data Cleaning: EDA often involves identifying and handling missing values, outliers, and inconsistencies in the data. Databricks offers powerful data manipulation and transformation capabilities using Spark SQL and DataFrame APIs.

Overall, EDA in Databricks empowers data scientists and analysts to gain insights from their data efficiently, enabling them to make informed decisions and drive business value.

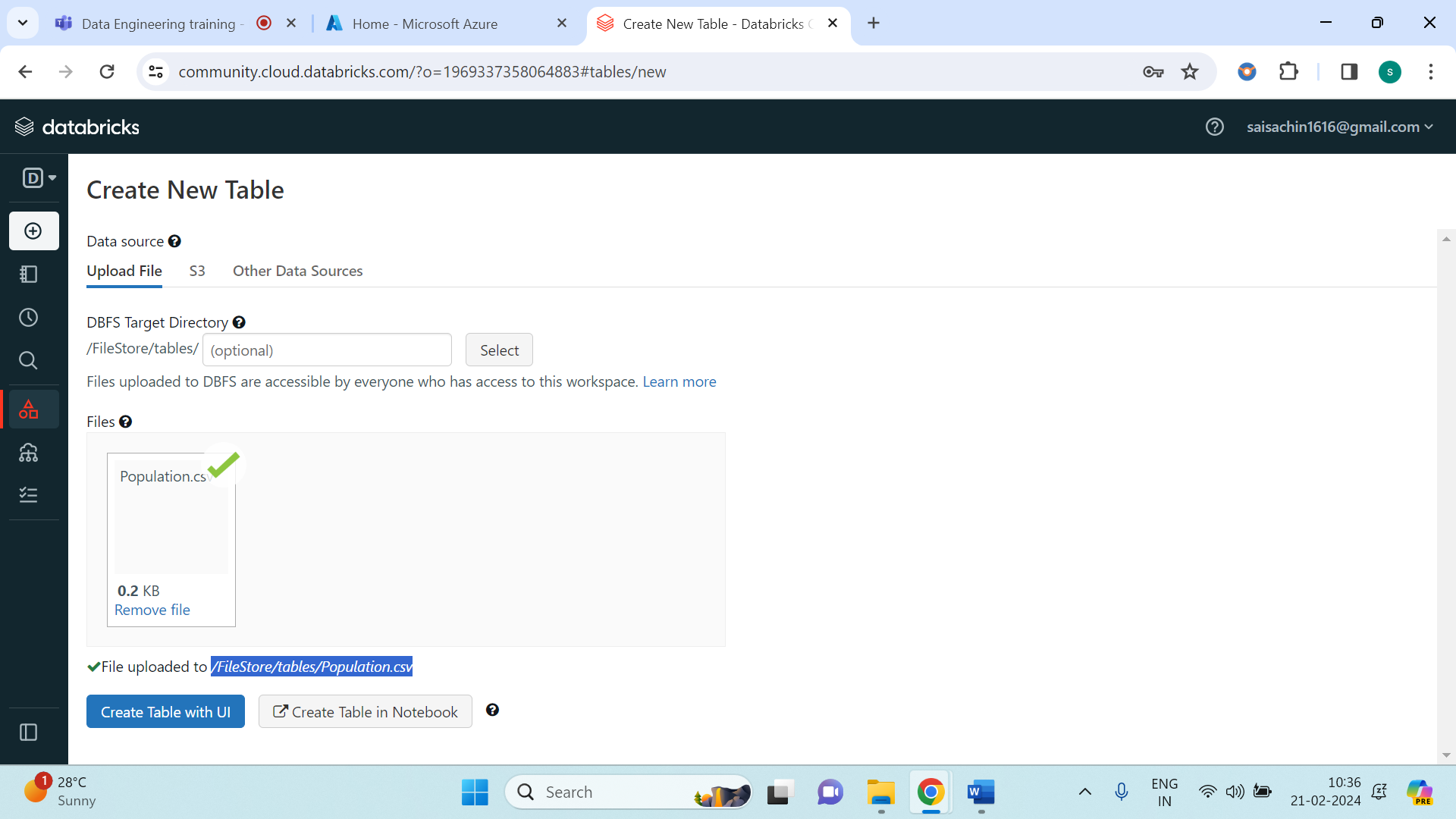
For Data Visualizing I will consider to create a csv file which consist of random data of a certain population in one street and name that file as **Population.csv** file and store it in local system. Now we can use Azure Databricks orelse Databricks community edition for Data Visualizing. I will use here Databricks community edition for Data Visualizing. The below image show the data kept in **Population.csv** file.



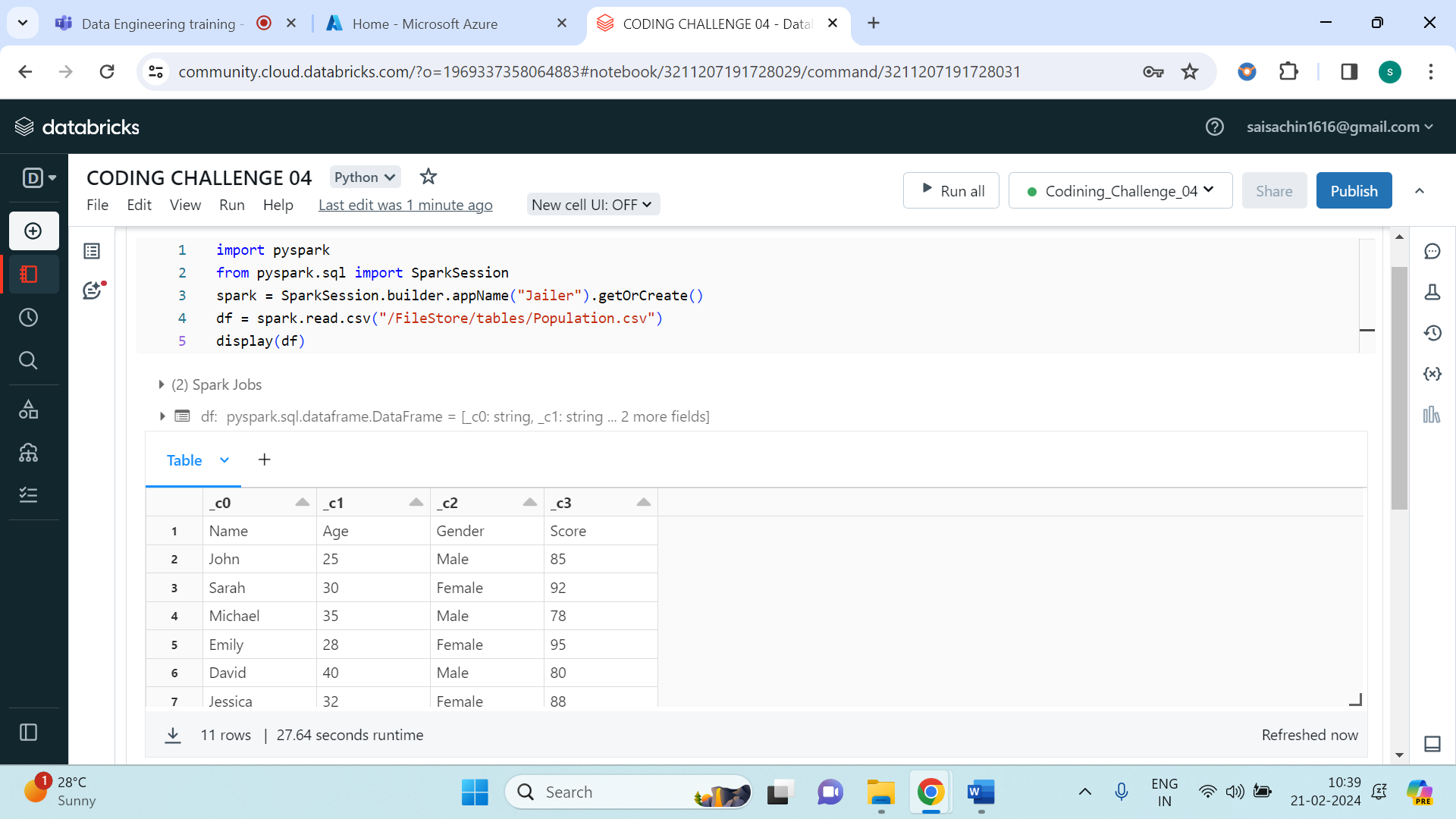
Now we will open Databricks community edition and start to run a cluster. The below images shows us after successfully running the cluster.



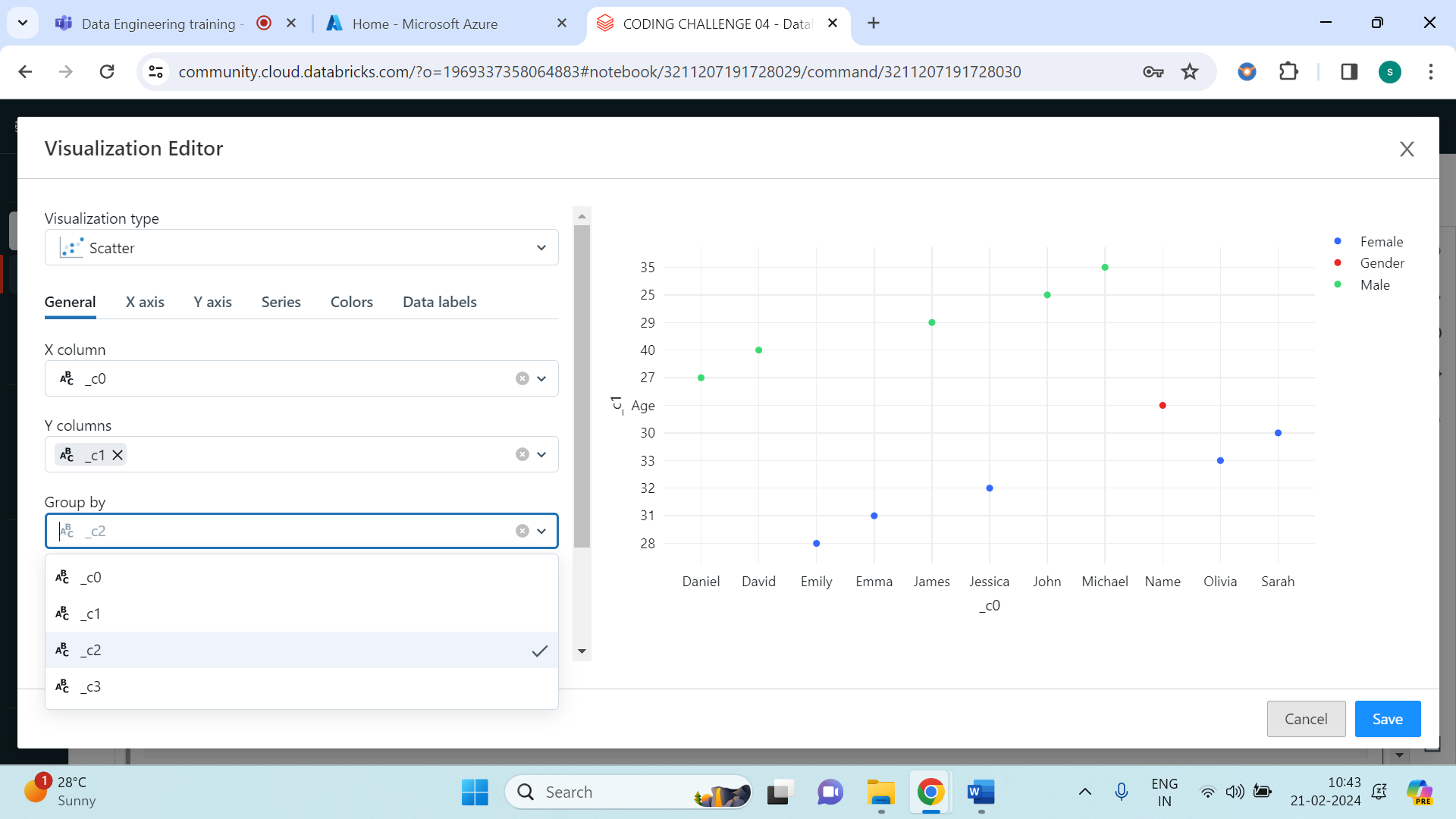
Then we will insert the Population.csv file in the table and copy the path where the file is uploaded in the Databricks (*/FileStore/tables/Population.csv*).



Later on, we will start to write a small piece of code for Data Visualizing as shown below.



From the above image, click on ‘+’ symbol and select Visualization and select Visualization type, select X, Y and Group by (optional) column. And below shows us the Visualization of Scatter graph.



Also we can perform Data Profile, by again clicking on ‘+’ symbol on the notepad result and select Data Profile.

