**NAME :- ANIKET SANJAYKUMAR BIYANI**

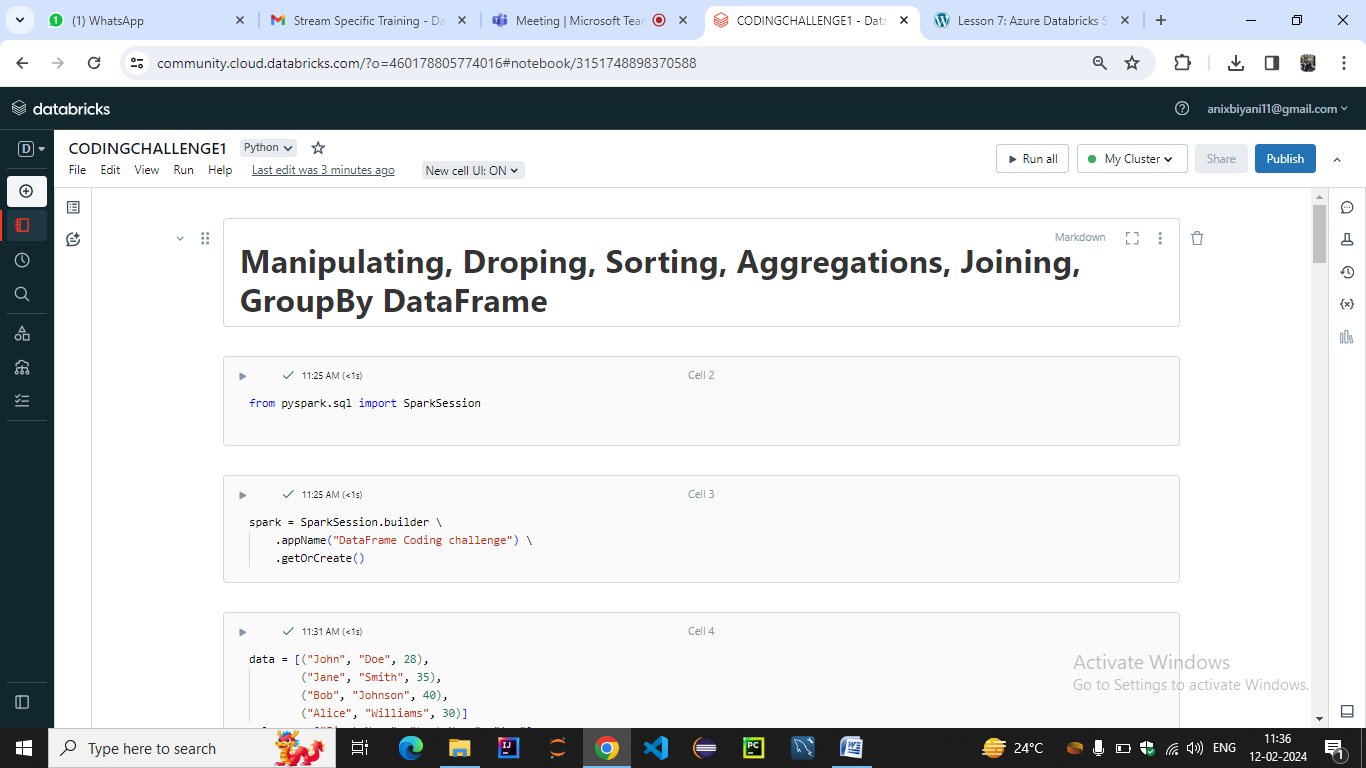
**DATA ENGINEERING BATCH -1**

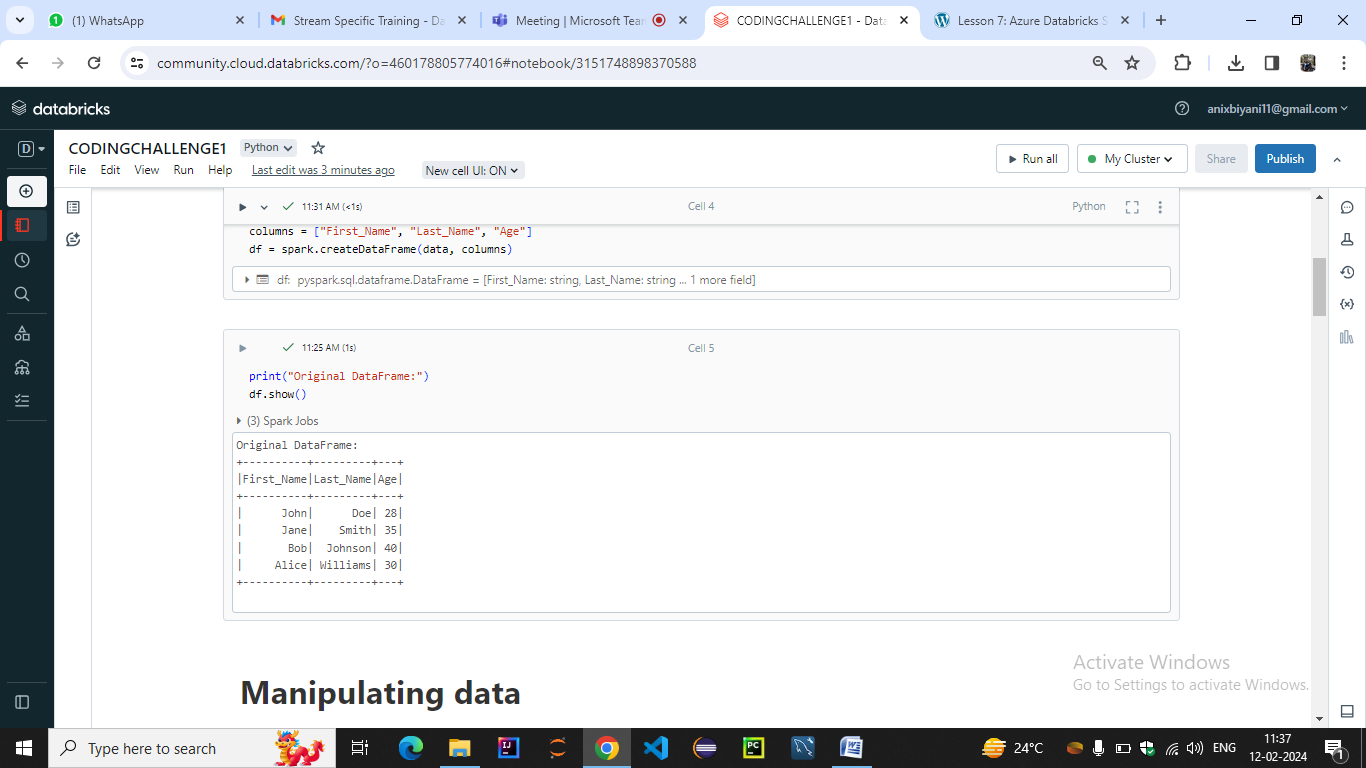
**PYSPARK CODING ASSESMENT**

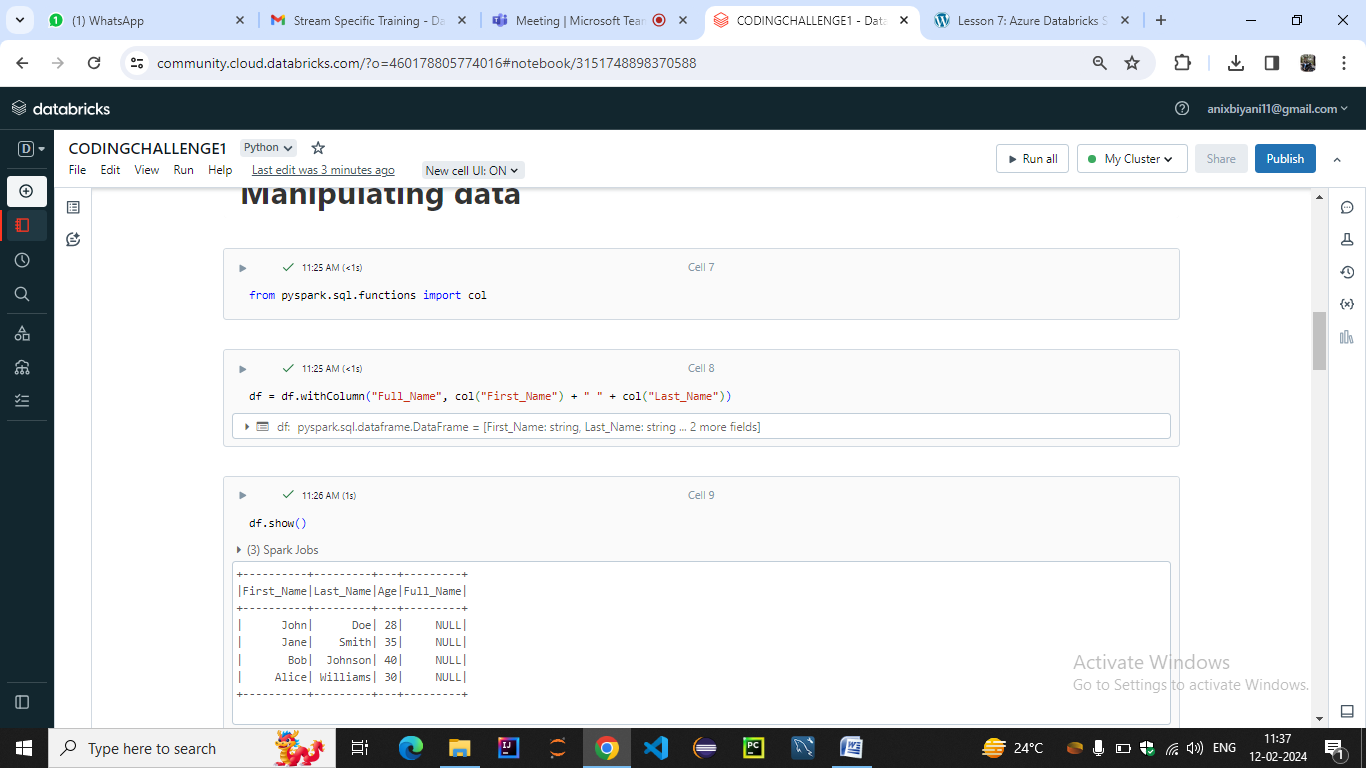
**Q1.** **Execute Manipulating, Droping, Sorting, Aggregations, Joining, GroupBy DataFrame**

* In PySpark, DataFrame manipulation involves several operations like filtering, dropping columns, sorting, aggregating, joining, grouping by.
* **Manipulating DataFrames**:

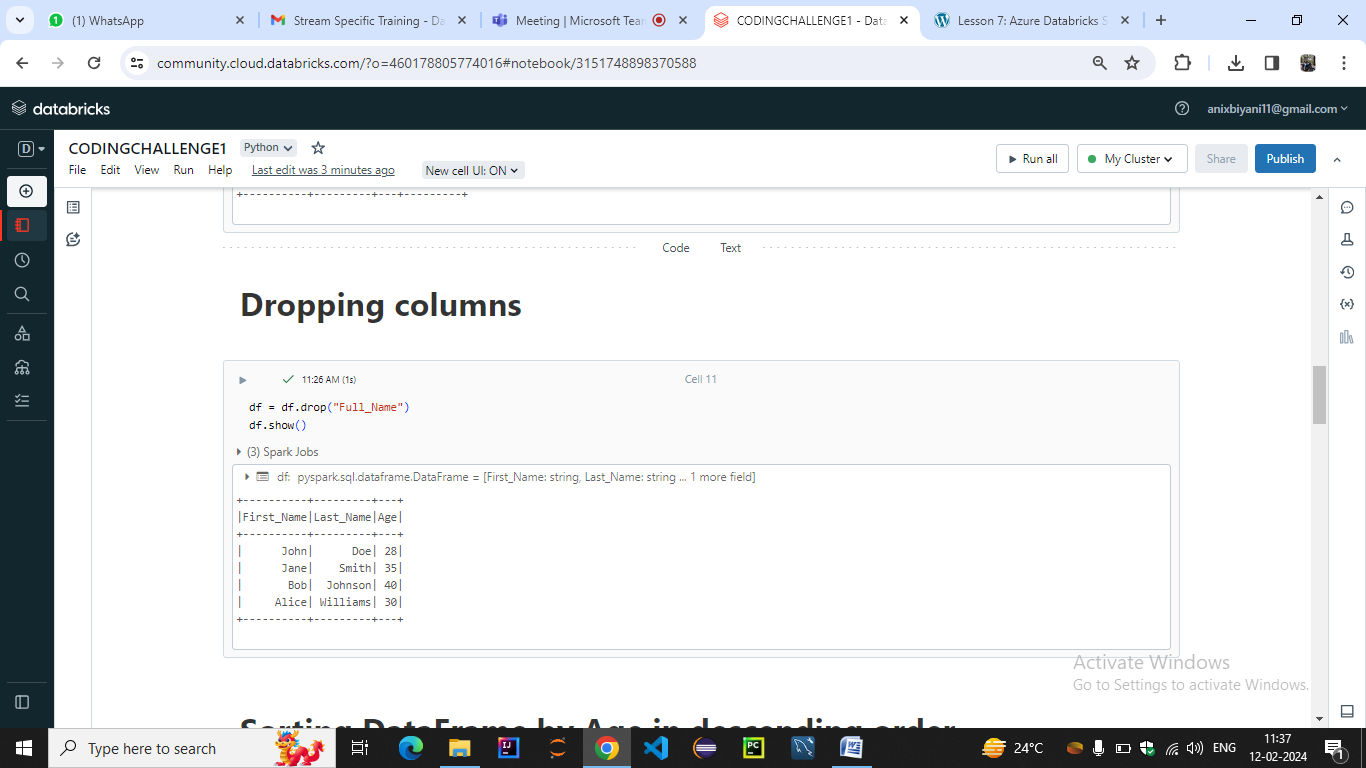
This involves operations like adding new columns, updating existing columns, or transforming data within the DataFrame. PySpark provides various functions like **withColumn()**, **select()**, **filter()**, etc., to manipulate DataFrame data.

****

****

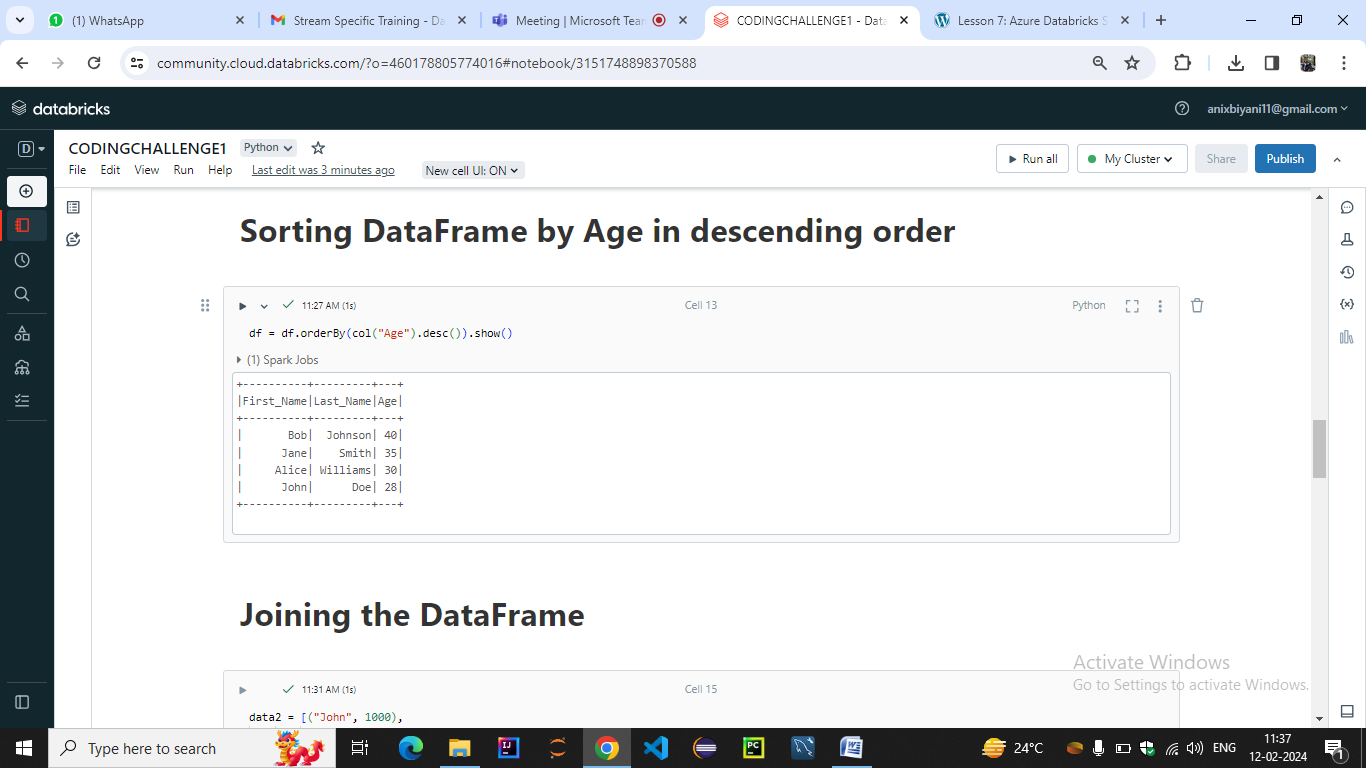
****

**Dropping Columns**: To remove unnecessary columns from the DataFrame. PySpark offers the **drop()** function to drop one or more columns from the DataFrame.

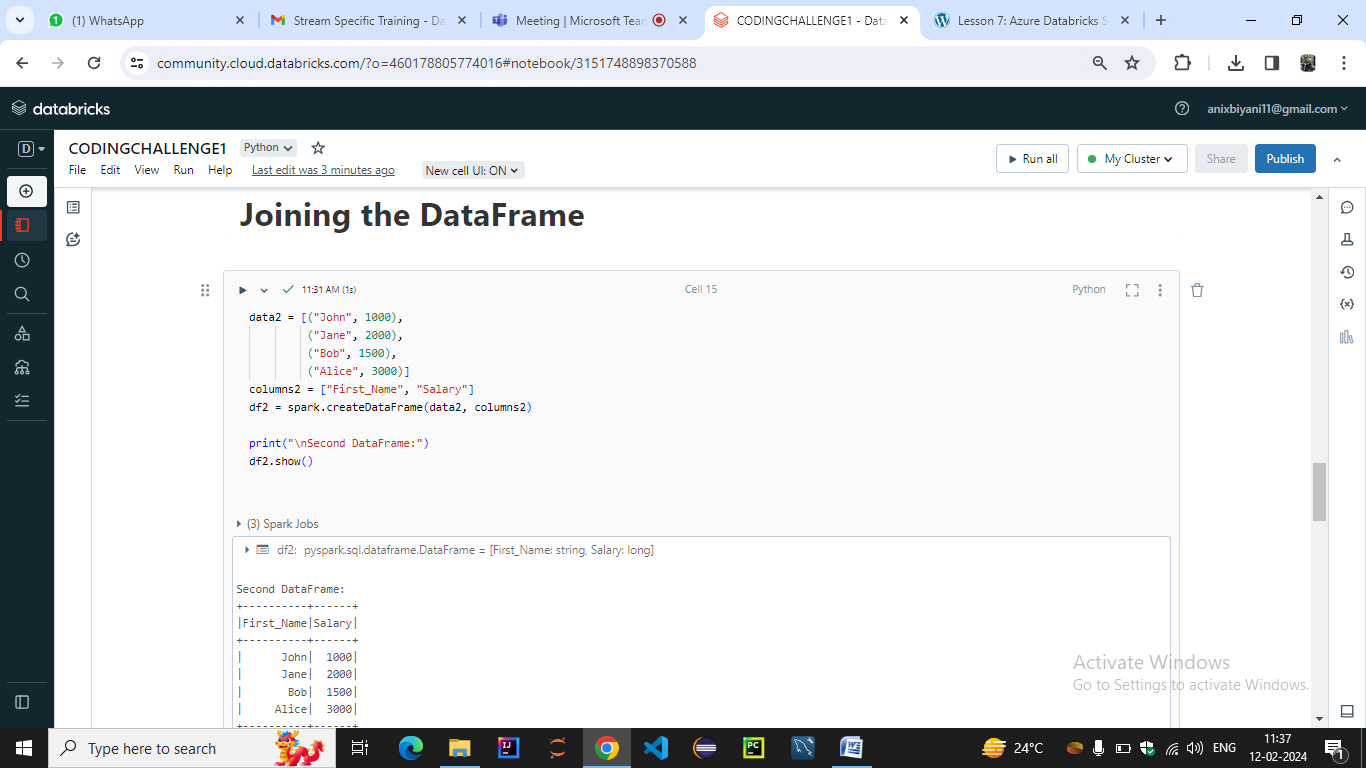
****

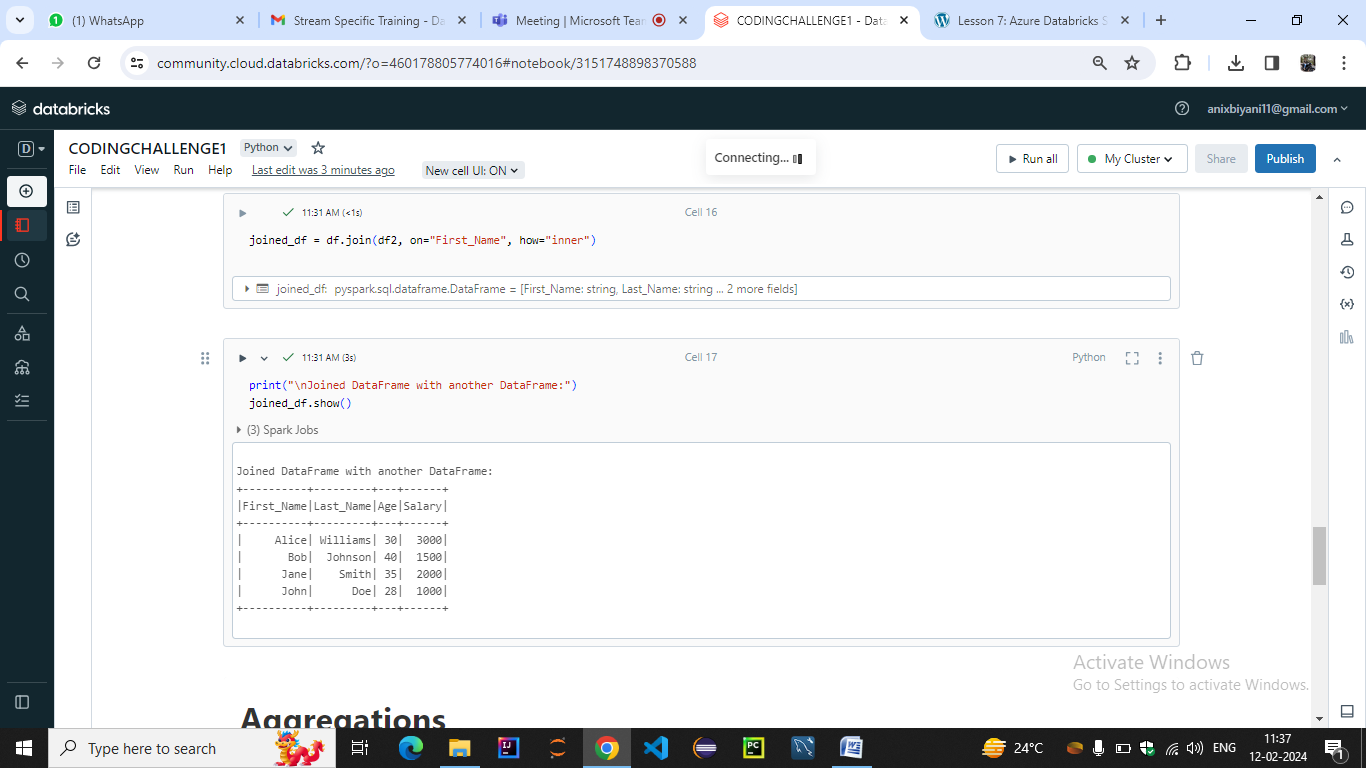
Sorting DataFrames:

Sorting rearranges the rows of a DataFrame based on one or more columns. PySpark's orderBy() or sort() functions can be used to sort the DataFrame in ascending or descending order based on specific column(s).

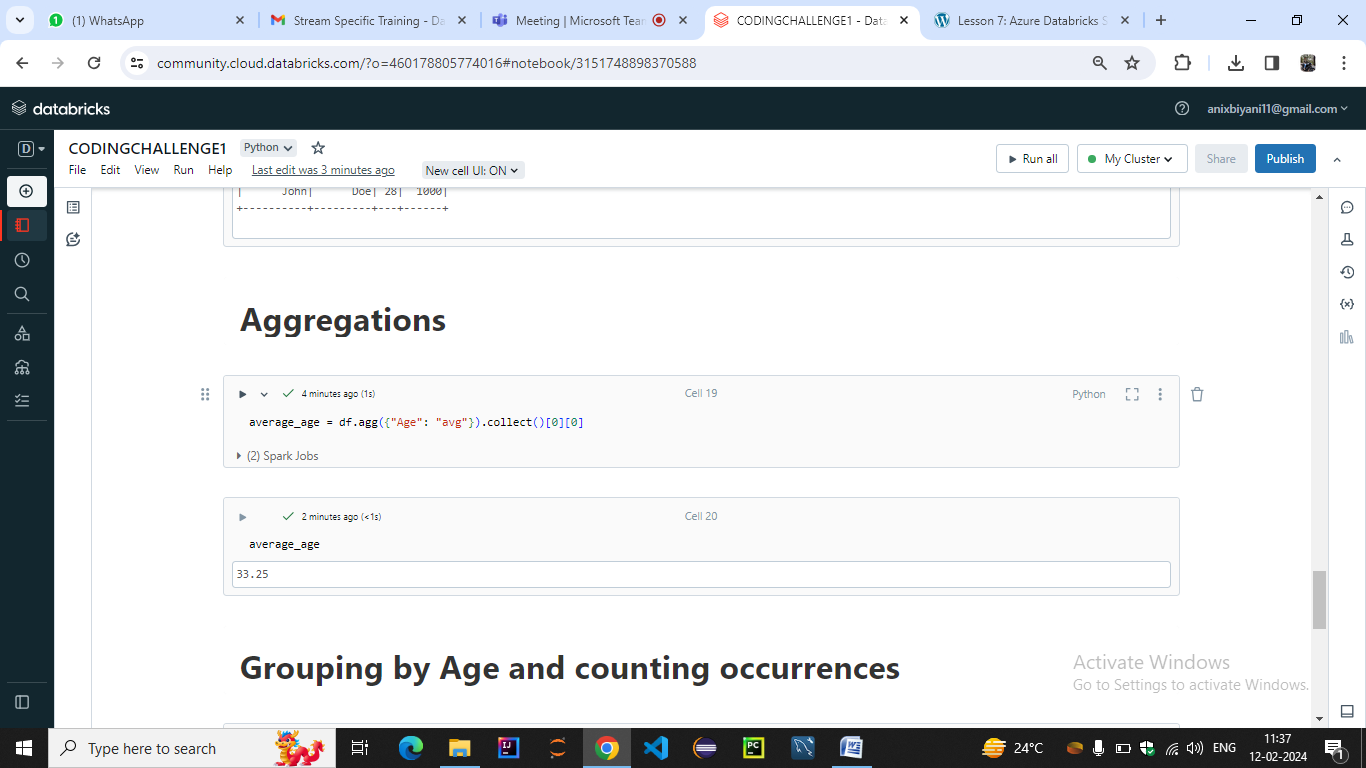
****

**Joining DataFrames**: Joining combines data from two DataFrames based on a common key or column. PySpark supports different types of joins such as inner join, outer join, left join, right join, etc., using functions like **join()**, **inner()**, **outer()**, **left()**, **right()**, etc

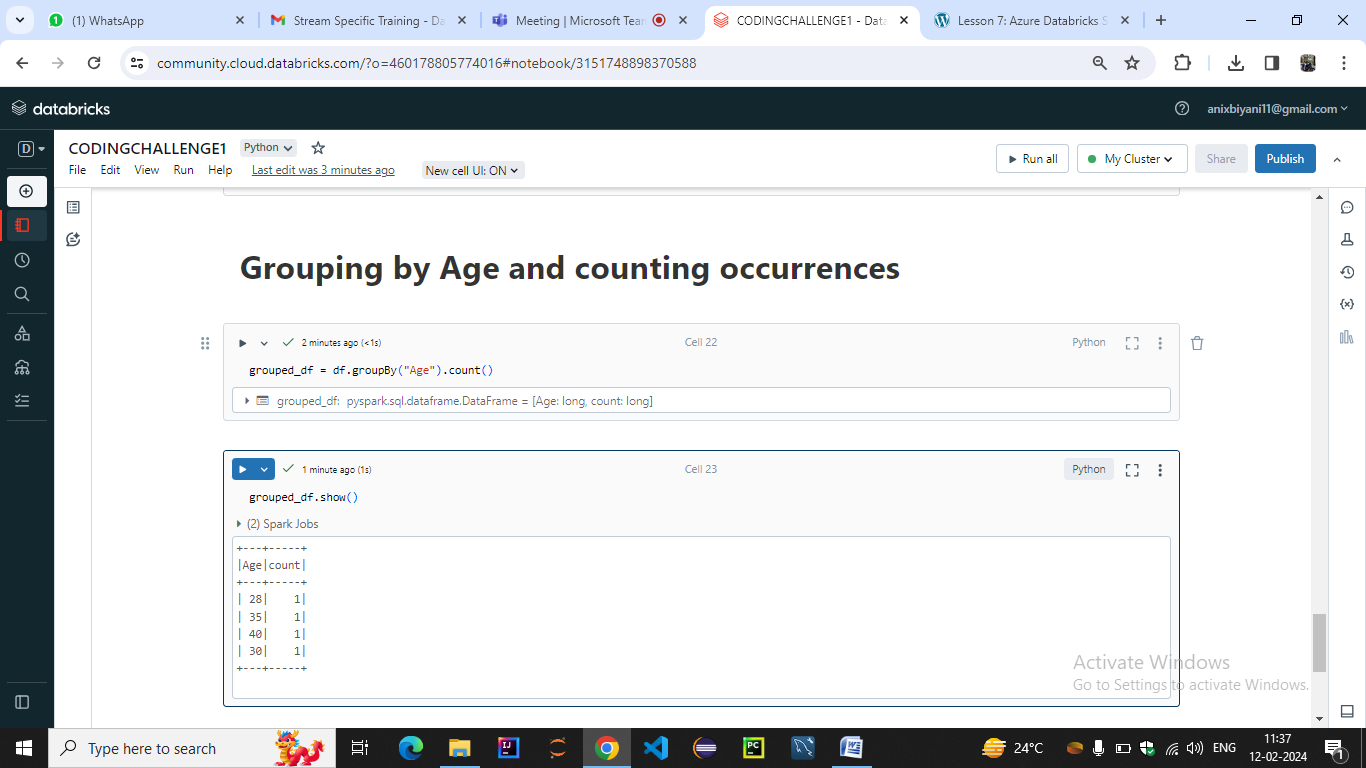
****

****

**Aggregations**: Aggregations involve computing summary statistics like sum, count, average, etc., on the DataFrame data. PySpark provides functions like **groupBy()** followed by **agg()** or **groupBy()** combined with aggregate functions like **sum()**, **count()**, **avg()**, etc., to perform aggregations.



**GroupBy:** Grouping data involves splitting the DataFrame into groups based on some criteria (like values of one or more columns) and then applying aggregate functions to each group. PySpark's groupBy() function is used for this purpose, followed by aggregate functions to compute statistics for each group.

****