**Bigdata Pyspark Assignment Solution**

## **Problem Statement:**

Given 3 datasets related to online products sales in the year 2019 for months Jan, Feb and March for a given store.

1. Cleanse the data removing blank rows:

Data cleaning is a crucial step in the data analysis process to ensure that the data is accurate, consistent, and reliable for further analysis.

Data Inspection: On carefully inspecting the sales datasets, I found missing values, anomalies, and inconsistencies in data.

**Handling Missing Values:**

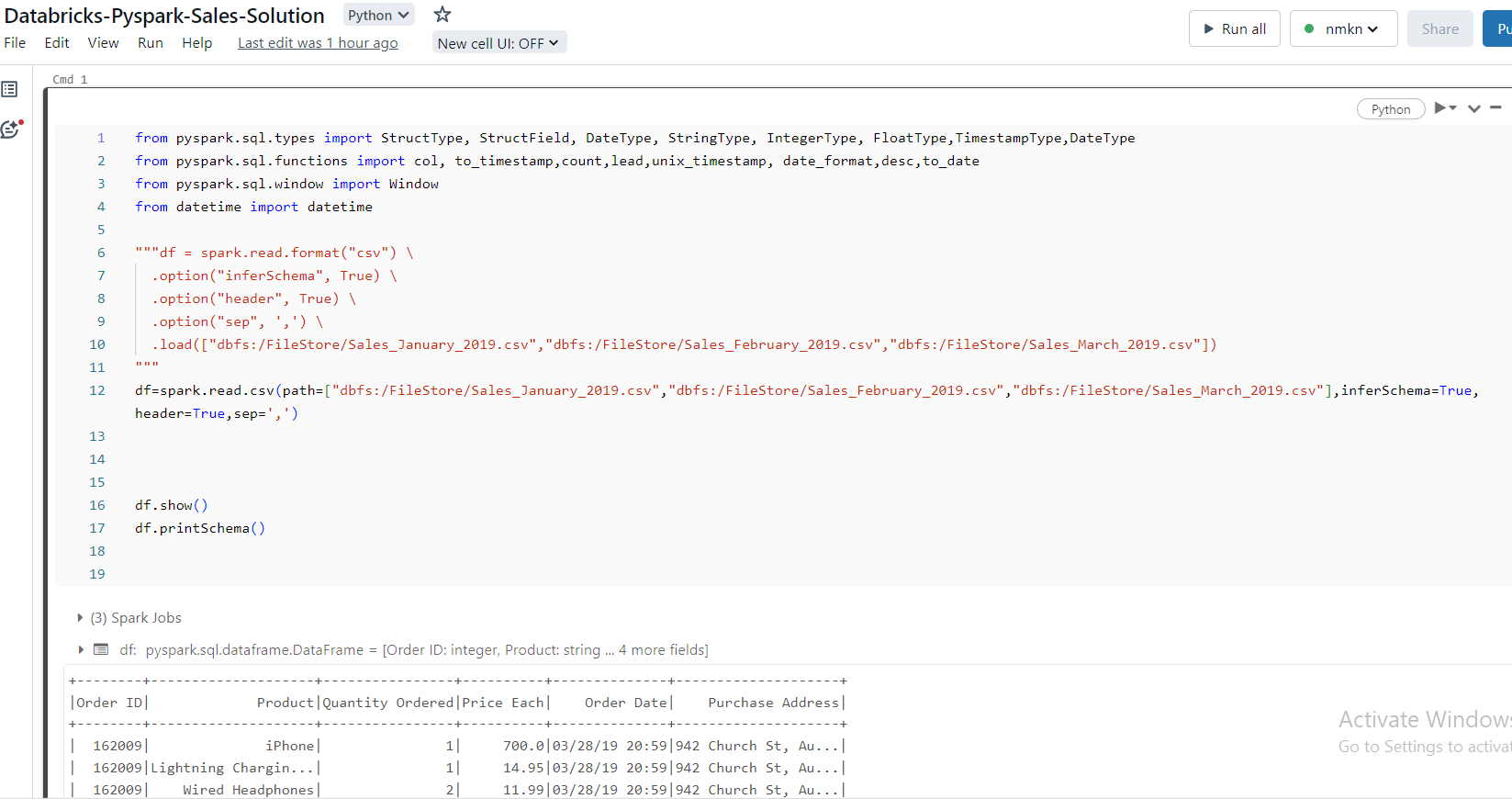
* Identified columns with missing data.
* Decided on an appropriate strategy to deal with missing values by removing rows containing missing values since they won't significantly impact the analysis.

**Dealing with Outliers:**

* Identified the outliers, the dataset contains header names or column names as rows, which are extreme values that deviate significantly from the rest of the data.
* Deleted the rows with outliers.

**Handling Inconsistent Data:**

* The data column ‘Order date’ contains inconsistent data, with some values has ‘MM/DD/YYYY hr:sec’ and some ‘DD-MM-YYYY hr:sec.’
* Converted the column to a consistent Timestamp Value ’YYYY-MM-DD hr:sec:ms’

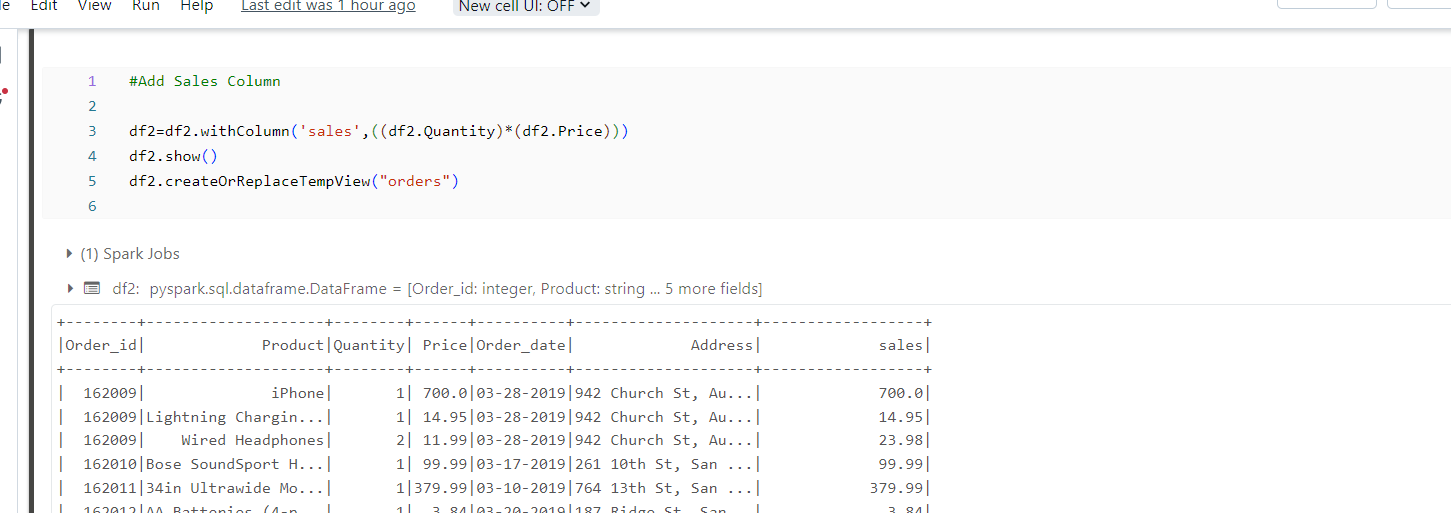


* Created New Data frame with new Column Names and Combining the 3 csv datasets:

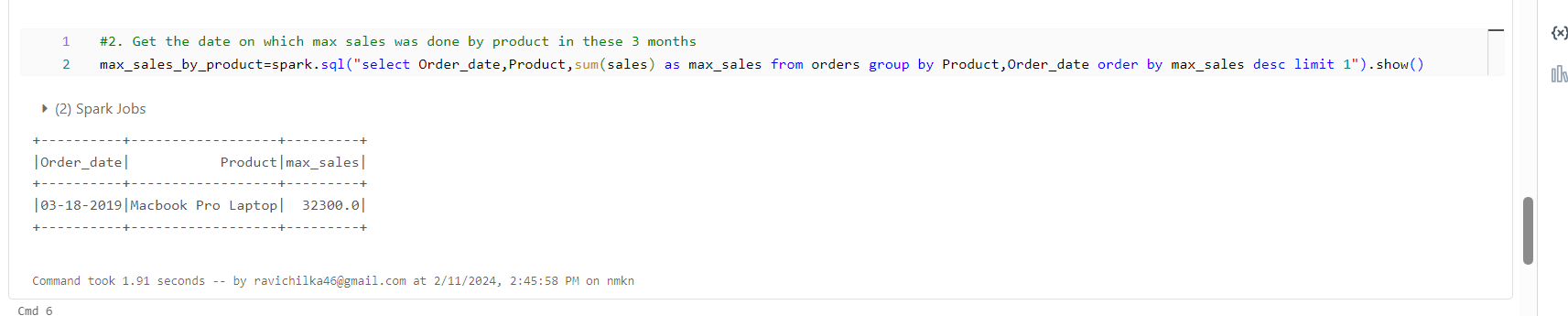




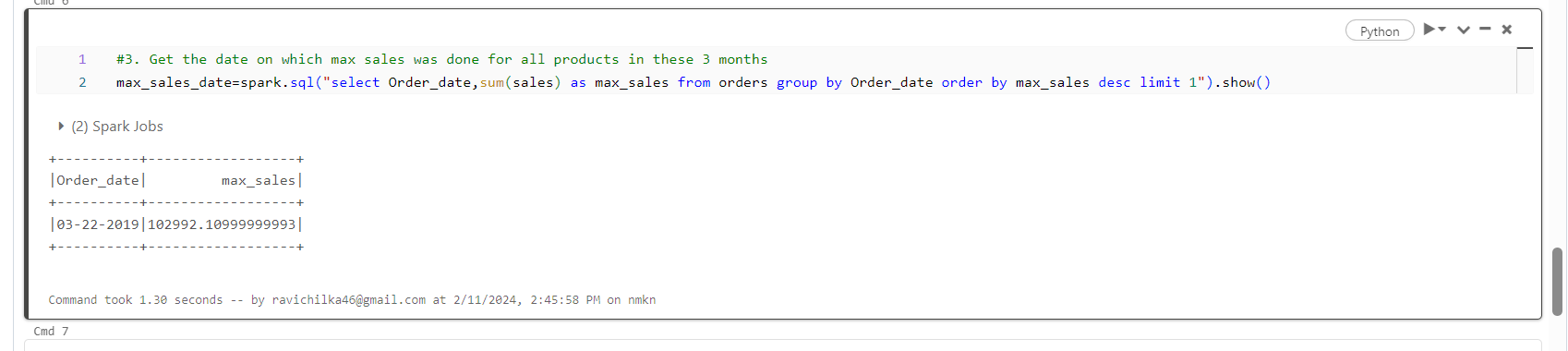
* Added Sales Column



1. Get the date on which max sales was done by a product in these 3 months:



1. Get the date on which max sales was done for all products in these 3 months:

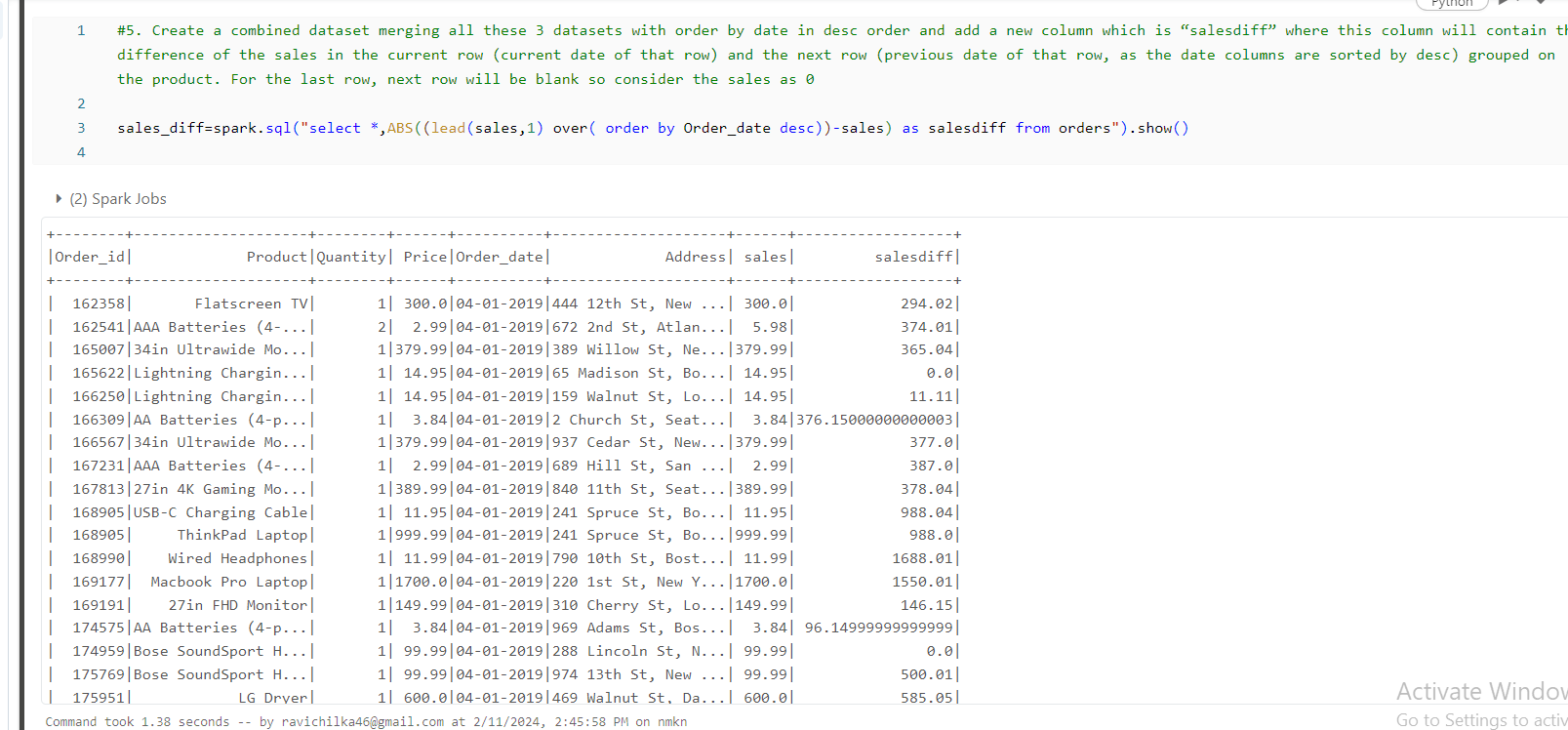


1. Get the average sales value for each product in these 3 months:

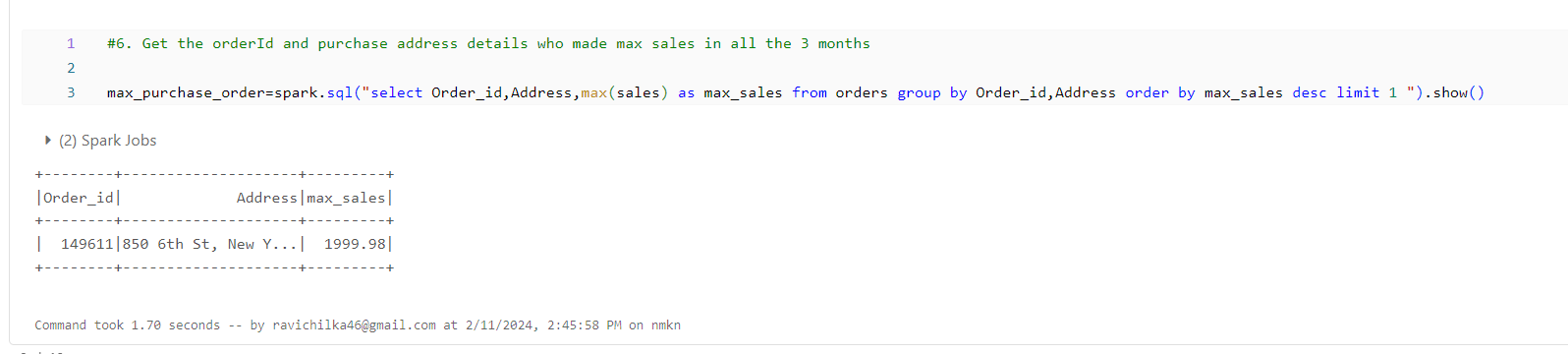


1. Create a combined dataset merging all these 3 datasets with order by date in desc order and add a new column which is “salesdiff” where this column will contain the difference of the sales in the current row (current date of that row) and the next row (previous date of that row, as the date columns are sorted by desc) grouped on the product.

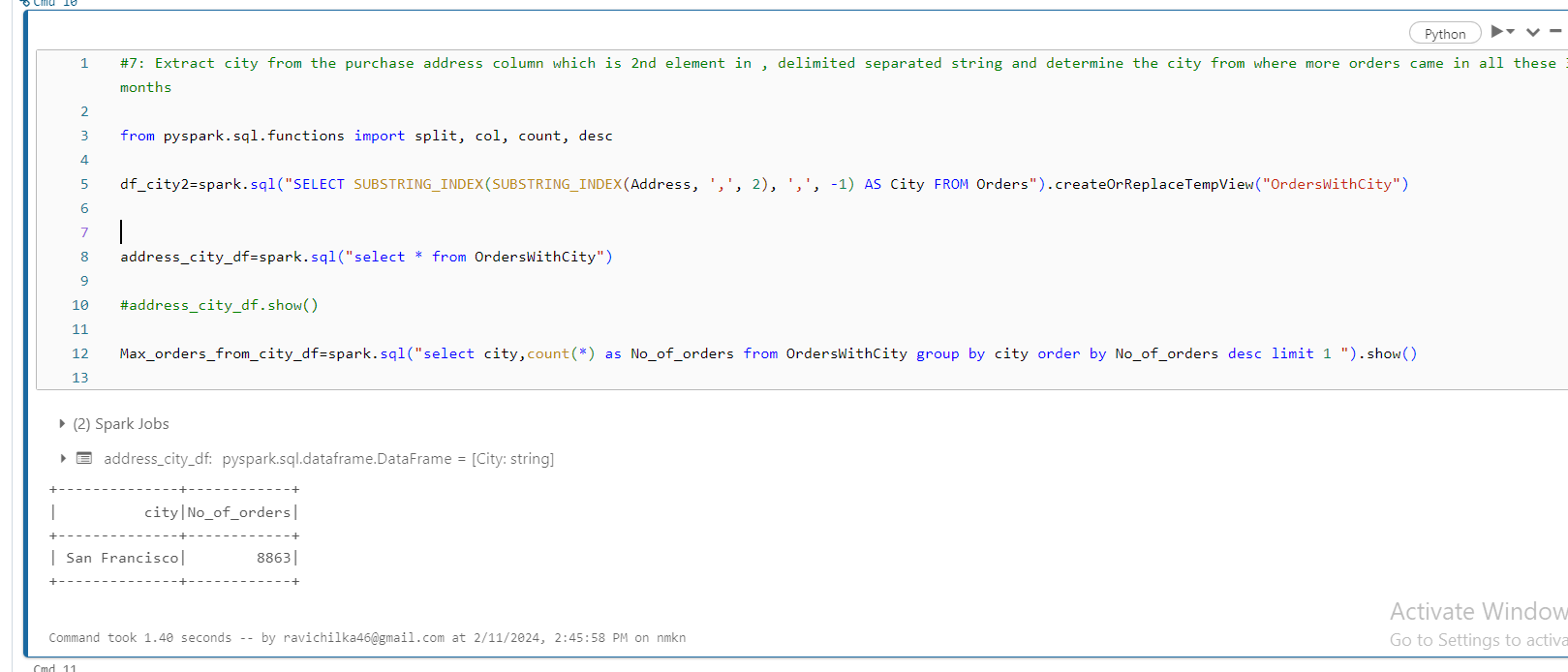
For the last row, next row will be blank so consider the sales as 0:



1. Get the ‘Order\_id’ and purchase address details who made max sales in all the 3 months:



1. Extract city from the purchase address column which is 2nd element in, delimited separated string and determine the city from where more orders came in all these 3 months:



1. Get the total order count details for each city in all the 3 months:



Email: [ravikishorchilka@gmail.com](mailto:ravikishorchilka@gmail.com)

Contact: 9145645606

Thank you for your time and consideration.

Sincerely,

Ravikishor Chilka