

MZUZU UNIVERSITY

FACULTY OF SCIENCE, TECHNOLOGY AND INNOVATION

DEPARTMENT OF INFORMATION AND COMMUNICATION TECHNOLOGY

TO : MR STANLEY NDEBVU

FROM : NATHAN BVUMBWE

REG NUMBER : BSDS0322

LEVEL : 3

SEMESTER : 6

COURSE CODE : BICT3608

PROGRAME : DATA SCIENCE

COURSE TITLE : DATA VIRTUALIZATION

ASSIGNMENT : ONE

DUE DATE :27 JUNE 2025

DATA WAREHOUSE ETL AND FINDINGS REPORT

1. Data sources:

The data was sourced from three files namely: profiles.JSONL, sales.CSV and products files which was sourced using and API, this one I named it products. CSV for easy handling with python.

1. Data extraction techniques:

The data was extracted using the pandas and requests libraries. The first was used to load data into Jupiter notebook and manipulate it and the later used the API as input to fetch the data for the third dataset which I named products.CSV.

1. Data transformation & cleaning:

From the profiles.csv and sales.csv files I dropped duplicate values using the dropna function from pandas library.

I went as far as changing the column names to lowercase to bring consistency in the naming convections among the column names. This was done for all the datasets.

The columns which represented money, the data type was changed to float and set to 2 decimal places for consistency in financial reporting.

The columns which had dates as values where set to datetime format using the pandas library built-in function “to.datetime()”.

Specifically, for products.csv dataset, to drop duplicates, I had to identify hashable columns and remove duplicates because it contained some columns that had its values being dictionaries.

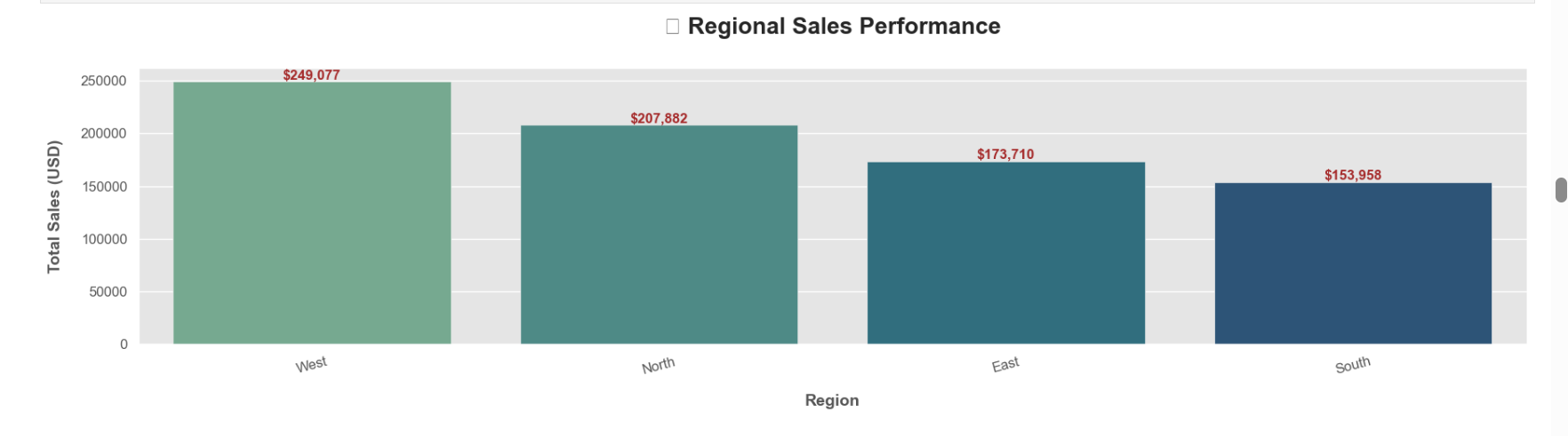
I dropped rows with null values on price and title for products dataset because these are critical for product identification and sales calculation.

Converted dates to datetime format just like in the other columns.

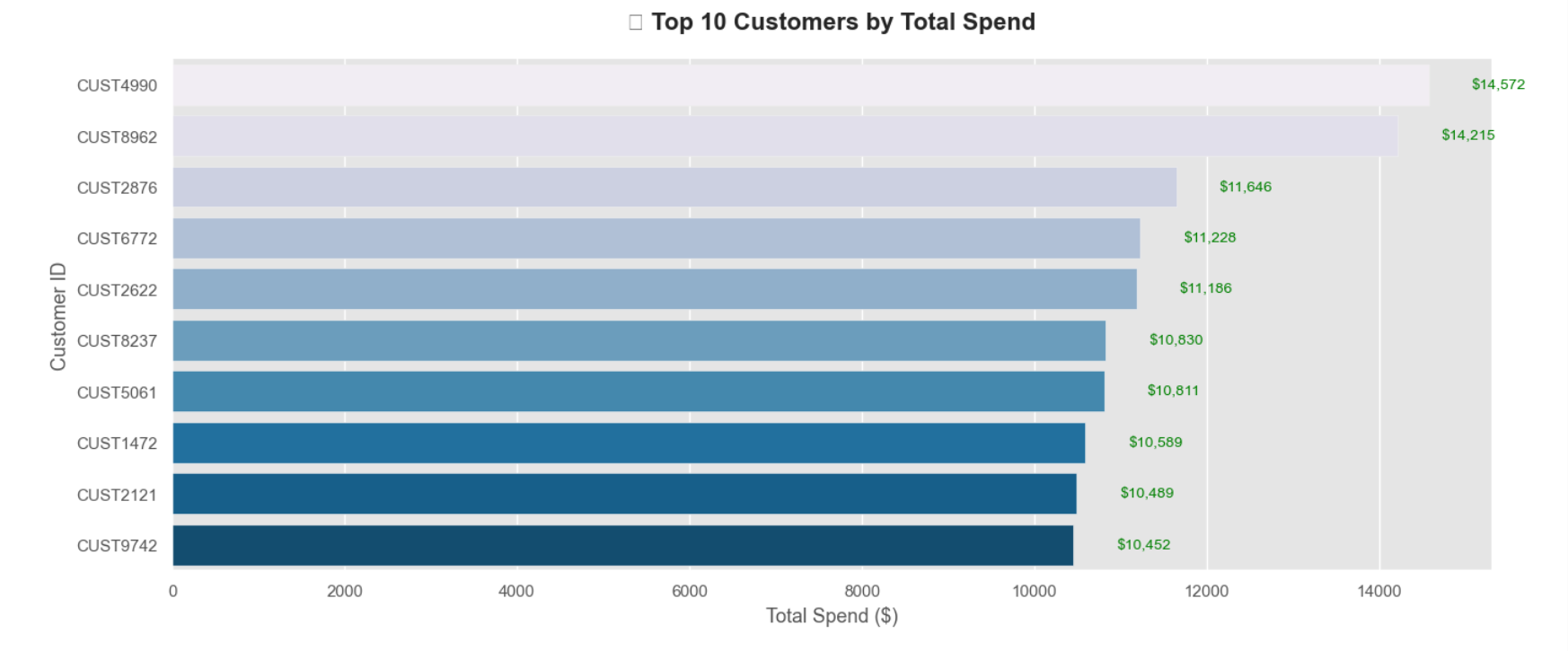
Last but not least, another dataframe was created that was a result of merging profiles and sales datasets on the customer\_id field thereby creating a consolidated DataFrame. This dtaframe combined customer information with their corresponding sales data. Only matching records from both Dataframes were included.

The four dataframes were then saved as csv files. The data was then loaded into Postgres where a data warehouse had been created.

1. Insights:
2. Total Sales by Region: it shows that the west region has the highest total sales of $249,077, indicating strong performance. The northern region follows with $207,882 while the south has the lowest sales among the four regions at $153,958.



1. Average purchase per customer: the top spender is CUST4940, with a total spending of $14,572 indicating they are a significant customer. The second one is CUST6862 with a total spend of $14,214.90. These customers are both male.



1. API data correlation with sales (i.e. region vs quantity sold):

I chose to check the correlation between region and quantity sold, this was so because theoretically different regions experience different weather conditions and therefore the idea was to check if weather has an effect on sales especially on quantity.

The results show that weather does not have a significant impact on the quantity of products sold in each region. Other factors are playing more crucial roles in sales performance. Therefore, there is need to explore other factors.