Metyis Data Engineer Technical Assignment

Project name: Creating Sales ETL

Author: Konrad Wroński

1. Project Description

This project invloves creating an ETL pipeline using PySpark. The process should read all of the files, merge them and write them to *cleansed* folder. The files are representing sales of a retail company. On top of data loading and processing, exploratory data analysis was performed.

2. Approach

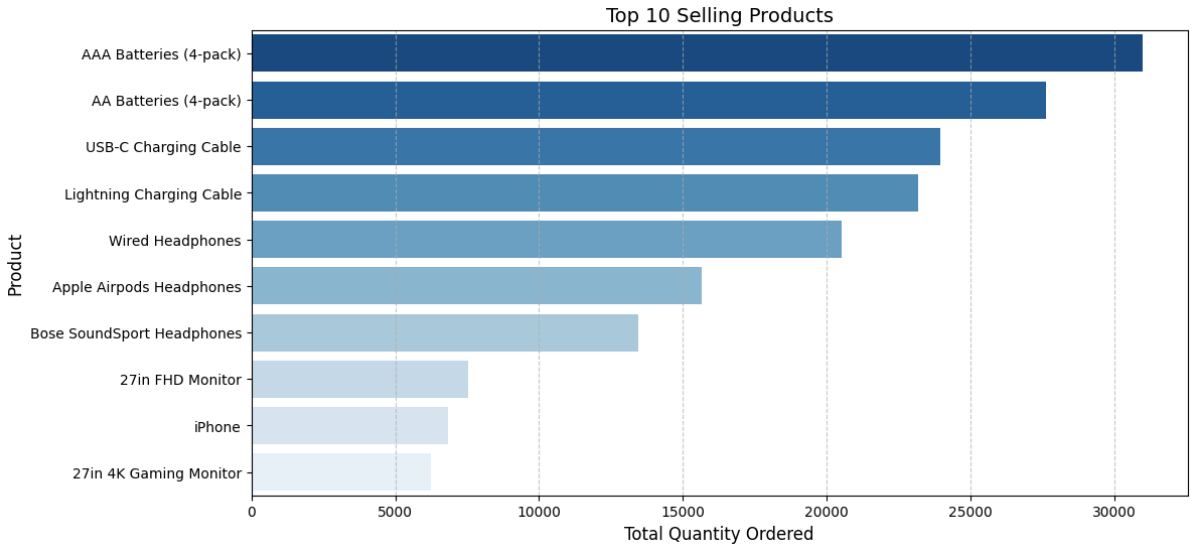
1. Creating PySpark session
2. Fixing column naming – lowercase and removing whitespaces
3. Creating schema and empty data frame to store merged files
4. Custom *load* function to process the csv files
5. *For loop* to interatively load and merge files
6. Handling missing values – removal of the rows, as missing values occur in multiple columns simultaneusly
7. Deduplication of the data using *Window function*
8. Data transformation – spliting the *Order Date* column for easier analysis
9. Loading the data into *cleansed* folder as parquet files with month and year partition
10. Data quality check – asserting that no data loss occurred with *assert*.

Parquet file has been chosen as a destination format as it is: columnar and supports compression

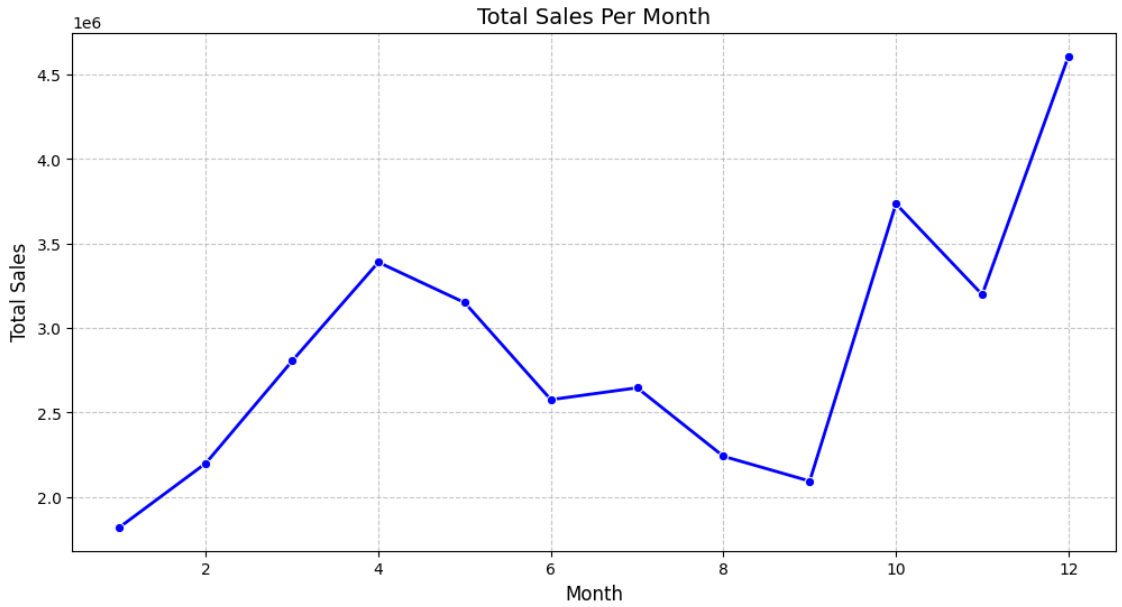
Partitioning strategy for this file is per month and year. The reason for this choice is that sales data are often used to compare certain time frames. The function used is partitionBy().

3. Data Exploration

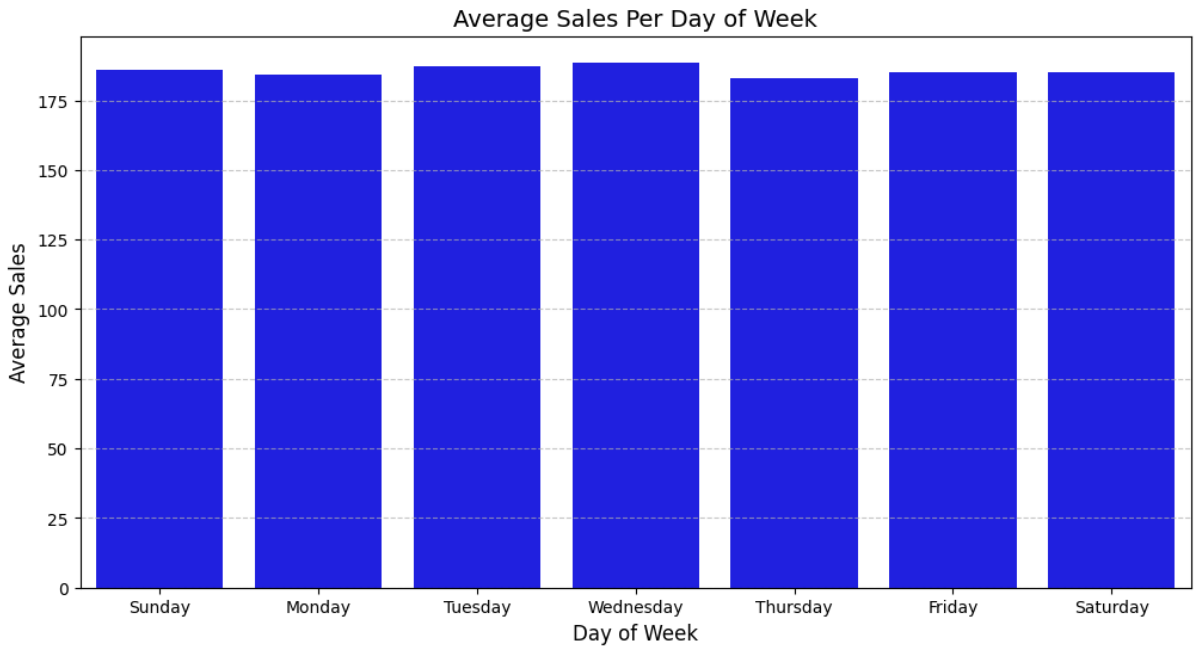
For a big retail company analysing its data is a crucial part of the decisioning process. Having substancial amount of data about sales can help to spot trends or help in resources allocation.



Above visualization shows the best selling products. A vital information that a company can use to modify its product offering. The best selling items in terms of quantity are batteries with over 50000 sold units. Charging cables are also sellling very well with both USB-C and Lightning selling over 22500 units.



One of the most important metrics for every retail company is an analysis of their sales for each month to spot seasonal trends. As seen on the plot the highest sales this company reach were in December and October. The lowest in January and September.



Analysing which day of the week the sales are the biggest can provide very important insights that can support marketing decisioning or resource allocation in company. In this case there is no correlation between sales and specific day of the week.

4. Summary and Suggestions

The pipeline for sales data has been made. Loading and mergning the data has been wrapped into one function. Data transfomation involved fixing the column names. Data cleaning process relied on dedcuplicating and removing rows with missing values. Finally data has been loaded to destination folder in parquet format while being partitioned by month and year.

For further analysis this data is suitable for machine learning to obtain even more sophisticated insights. With the time series structure of data company can forecast their sales using moving averages, exponential smoothing or ARIMA models.