Python Project Report

Course: Certificate in Data Analytics for Business

By Anna Koldova

## GitHub URL

<https://github.com/AnnaKoldova/AnnaKoldova.git>

## Abstract

As part of the course for Data Analytics for Business, Python Programming Language has been introduced as a means to manipulate and analyse large datasets as effectively as possible, as well as produce visualisations in form of graphs and charts.

During the eight weeks of the course, basic Python competences were introduced, such as importing data from various sources, data manipulation with Pandas and Numpy. Course has though how to extract data using Python coding language, and what commands are available for execution.

Data visualisation with Matplotlib and Seaborn has been introduced and, together with various forms of visualisation clarified in order to achieve the desired outcome.

All beforementioned functionalities have been used in the below detailed Python Project and implemented with Python code.

## Introduction

The purpose of this Project is to demonstrate how to use Python to analyse business data and produce valuable insights based on a real dataset. This Python Project is aimed at analysing a dataset of Store sales using Python Programming Language, and PyCharm has been selected as preferred application to execute Python Commands.

As part of this project, data has been uploaded, filtered, and visualised with Python Pandas, Matplotlib and Seaborn packages. In addition, data has been cleaned by identifying duplicates and grouped by.

With the help of basic Python commands and presented dataset it was possible to derived some key datapoints and insights described in more detail below.

## Dataset

Chosen dataset was sourced from “Kaggle”1. Dataset is kept in a CSV format and contains various details of products sold in online store, categorized by product type, order priority, shipment method, customer segment, profit and sales figures associated with each order and customer, as well as country region from where orders originated.

This dataset is simple enough for the first Python project yet has sufficient number of rows and columns which allows to exercise the most important Python commands and identify interesting patterns and correlations in dataset.

This Dataset could potentially utilise the capabilities of Machine Learning. Machine Learning can study customer behaviour at the time of purchasing. For example, if a customer orders the same combination of products all the time with the same cyclicity, Machine Learning can prompt specific customer with their favourite products every time they log in to the Store.

AI algorithms can also propose similar products when searched product is stocked out, or prompt a customer with additional product that can be purchased in combination with the product already placed into the basket.

## Implementation Process

PyCharm has been selected as the most effective environment for Python code implementation. PyCharm provides tips as well as highlights errors and suggests how these errors can be corrected.

Additional Python Interpreter packages were required for future code execution and implementation. To perform the data analysis, Pandas, Numpy, Matplotlib and Seaborn packages were installed via Settings.

Dataset that was chosen to be analysed with Pandas was downloaded as a CSV file which also needed to be uploaded to Python. CSV library was uploaded using csv.reader object command successfully.

Visualisation was performed with a mix of seaborn and matplotlib commands, where Figures have been plotted by Seaborn and printed (plt.show()) via MatplotLib. More detailed overview of visualised output is presented in Results section.

Dataset was further refined by grouping most common order combination of 2 products, with the help of “Intertools” and “Collections”.

There were duplicates identified by order number, where the same order number could contain several different Product Categories purchased, generating different Sales or Profit figures, as well as generated from different Region and shipped by different Ship Modes.

## Results

#### Chart 1

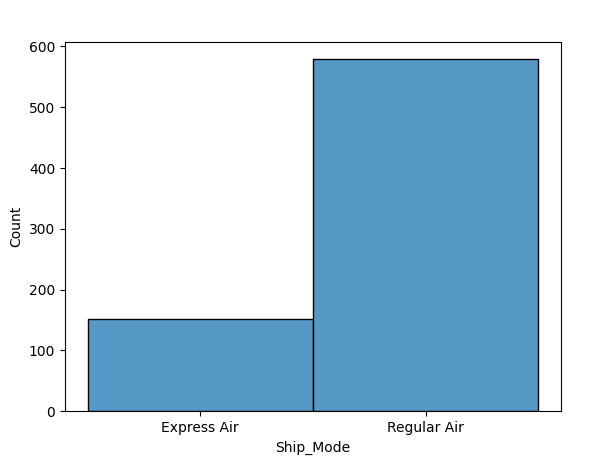
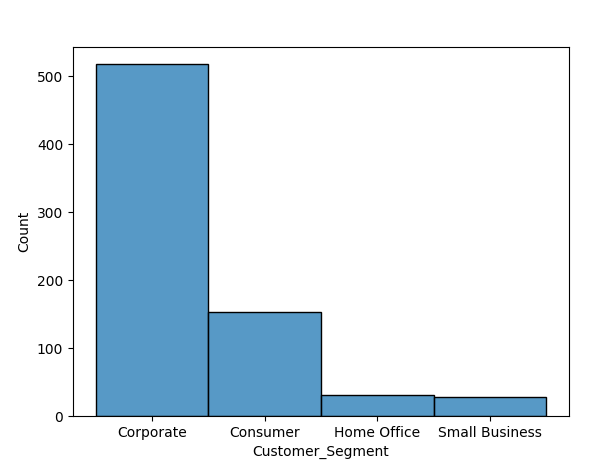


Chart 1 provides a simple visualisation of the number of orders dispatched by either Express or Regular Air. This chart suggests that majority of customers prefer Standard Air mode. Based on this information the Store can decide whether to continue offering Express Air shipping, could it be offered for cheaper in order to increase Express Air uptake, or perhaps offer an occasional discount to customer frequently ordering ‘Express Air’.

#### Chart 2



The above chart provides a Customer Segment breakdown. It is visible that the highest number of customer fall into ‘Corporate’ segment. Second highest is ‘Consumer’ segment; however, the number of customers is about three times smaller than in ‘Corporate’.

Knowing the most prevalent customer segment at the store allows to target specific offerings tailored for the ‘Corporate’ consumers’ needs.

#### Chart 3

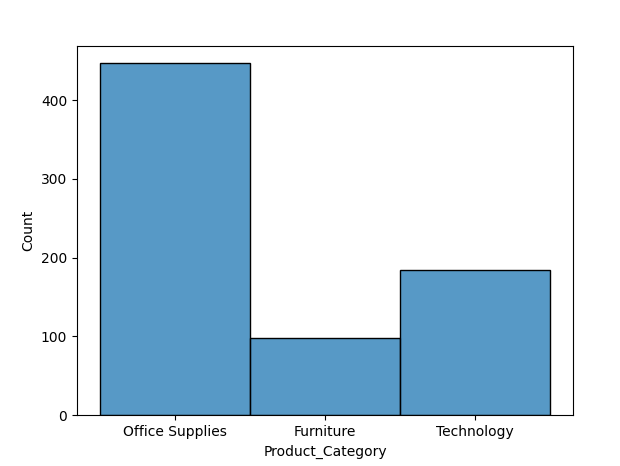
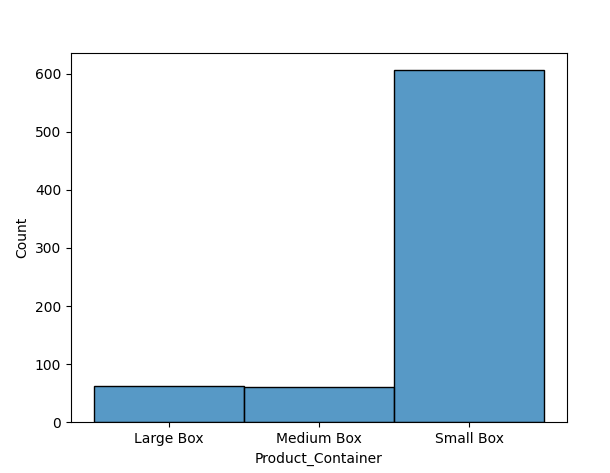


Chart 3 provides a breakdown of Product categories and count of orders for each. ‘Office Supplies’ is the most popular product category, which can also be logically explained by the fact that the largest customer segment is ‘Corporate’.

Based on the chart output, the Store can consider expanding ‘Office Supplies’ product range, as well as offer special discounts for bulk orders or joint orders with ‘Technology’ or ‘Furniture’ category, since the latter is the second largest product category, and can also be associated with ‘Corporate’ customers.

Machine learning algorithms can help identify purchasing patterns of ‘Corporate’ customers and propose special deals automatically.

#### Chart 4



The following bar chart reflects on the number of orders sold broken down by Product Container. Majority of orders are shipped in ‘Small Box’ container, which also correlates well with Chart 3, which identified ‘Office Supplies’ being the most purchased category.

#### Chart 5

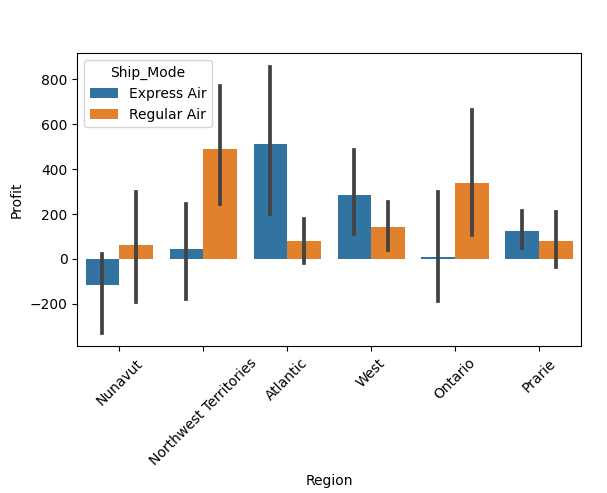


Chart 5 analyses the Profit generated in each Region of the country, as well as broken down by ‘Ship Mode’. The most profitable regions are Northwest Territories and Atlantic, where ‘Regular Air’ shipments have generated higher Profits. In Atlantic region ‘Express Air’ shipments are more profitable. Nunavut region is operating with loss, specifically in the ‘Express Air’ shipments.

Based on the above chart, the Store may decide to change their strategy with regards to Ship Mode offering depending on the Region where customer is based. For example, the seller can introduce a discounted rate for ‘Express Air’ shipping. This can encourage more sales in that category in Northwest Territories, Nunavut and Ontario. Perhaps these Regions are more price sensitive and are ready to wait for their purchase longer, however with a small discount they would be willing to proceed with ‘Express Air’.

#### Chart 6

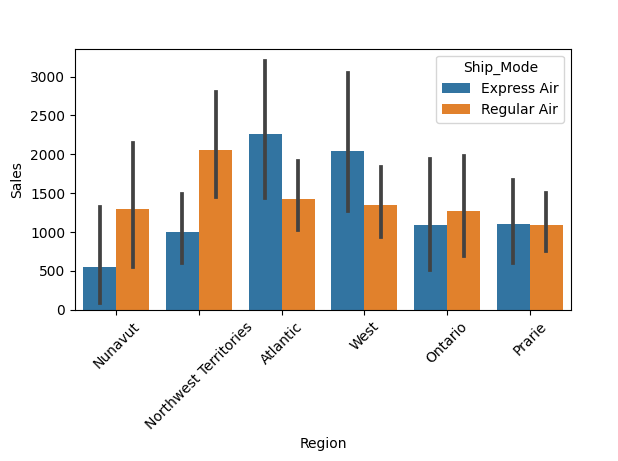


Chart 6 is similar to Chart 5; however, it analyses Sales by Region, broken down by Ship Mode. Similarly, to previous Chart, it is visible that Atlantic region makes by far the most Sales comparing to other Regions. Again, ‘Express Air’ in Atlantic is preferred Ship Mode and generates most Sales. It should be noted that Sales revenue generated does not directly correlate with Profits generated in the same area. For example, in Nunavut, Profit on ‘Express Air’ shipments were negative, whereas Sales stayed in positive, meaning that overhead costs of handling Nunavut sales is much higher and not as economically viable as in other Regions.

#### Chart 7

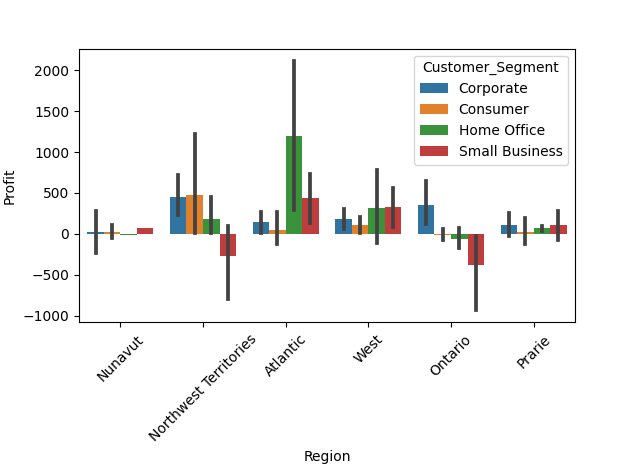


Chart 7 provides a visual description of Profits earned by Customer Segment broken down by Region. As it was already established with Chart 5, Atlantic is the most profitable region, and the biggest share of region’s profit is generated by ‘Home Office’ customer segment.

Based on Chart 7 and Chart 1, although ‘Corporate’ accounts for the largest customer segment in the Store, this segment does not generate the highest Profits for the Store. On the contrary, ‘Home Office’ being one of two smallest Customer Segments, bring higher proportion of profit.

Northwest Territories have two equally large customer groups – ‘Corporate’ and ‘Consumer’, which generate highest volumes in this Region.

‘Small Business’ segment operates at loss in two Regions – Ontario and Northwest Territories, these to regions also have higher proportion of ‘Regular Air’ shipments.

#### Chart 8

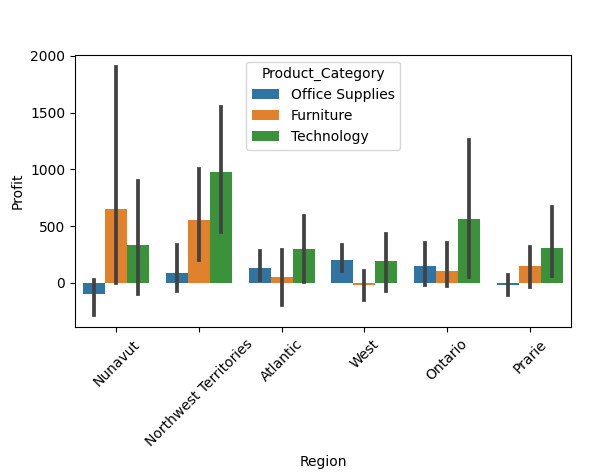


Chart 8 provides an analysis of the most profitable Product Category in each Region. In Atlantic, Ontario, Prairie and Northwest Territories are the most profitable from ‘Technology’ category. Second most profitable Product Category is ‘Furniture’ in Nunavut, Northwest Territories and Prairie.

## Insights

* ‘Office Supplies’ has the highest count of orders, but this Product Category generates the lowest Profits.
* ‘Corporate’ is the largest Customer Segment, yet it generates the lowest Profits for the Store.
* ‘Atlantis’ Region generates the highest Profit and Sales figures, whilst the biggest Customer Segment is ‘Home Office’ with preferred Ship Mode being ‘Express Air’.
* ‘Technology’ Product Category has lowest number of orders but earns the highest Profit in five Regions out of six.
* When analysing Customer Segments by region, Nunavut customers bring the lowest share of Profit, although their ‘Furniture’ sales are the highest out of all six Regions.

## References

1 – Kaggle Data source: <https://www.kaggle.com/takkimsncn/sales-store-sales-analysis/data>

2 – Stack Overflow community: <https://stackoverflow.com/questions/66066186/how-do-i-get-all-possible-combinations-of-a-list-to-put-into-a-dictionary-such>

3 – Seaborn plotting instructions: <https://seaborn.pydata.org/generated/seaborn.barplot.html>

4 – Matplotlib Cheat sheet: <https://s3.amazonaws.com/assets.datacamp.com/blog_assets/Python_Matplotlib_Cheat_Sheet.pdf>

5 – Seaborn Cheat sheet: <https://s3.amazonaws.com/assets.datacamp.com/blog_assets/Python_Seaborn_Cheat_Sheet.pdf>