رواد مصر الرقمية

وزارة الاتصالات وتكنولوجيا المعلومات

Digital Egypt Pioneers Initiative (DEPI) Final Project

REPORT NAME: Milestone (2)

Sales Forecasting and Optimization

Group Members:

Basmala Ehab Mohamed Yousry

Mostafa Mahmoud Mohamed Elshahat

Mohab Mohamed Ibrahim Mohamed

Ziad Ahmed Gharieb

SHR2_AIS4_S2

Submission Date:

9 May 2025

Youth co.

رواد مصر الرقمية

وزارة الاتصالات وتكنولوجيا المعلومات

Table of Contents

Youth co.

Challenges

The challenge for us is how to analyze the data that plays a major role in the right decision that the decision maker will make to make an accurate decision that helps in achieving his goals by calculating the total sales and predicting future sales and knowing whether the business is on a downward or upward curve.

Data Cleaning

As we learned in milestone1, there are problems in the data. It is time to learn how to process the data in the right way that serves our results and our work.

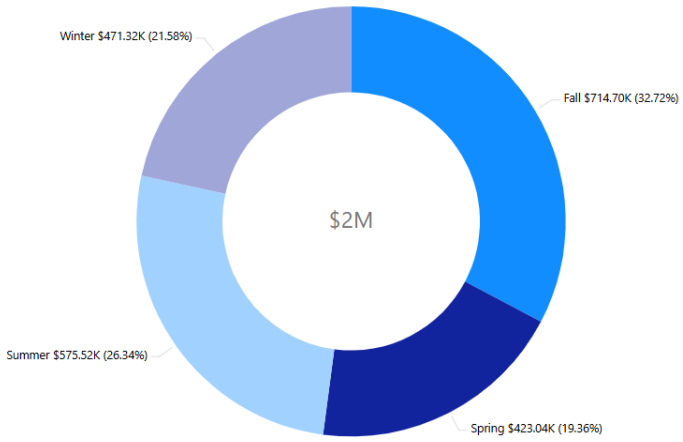
Steps we did follow:

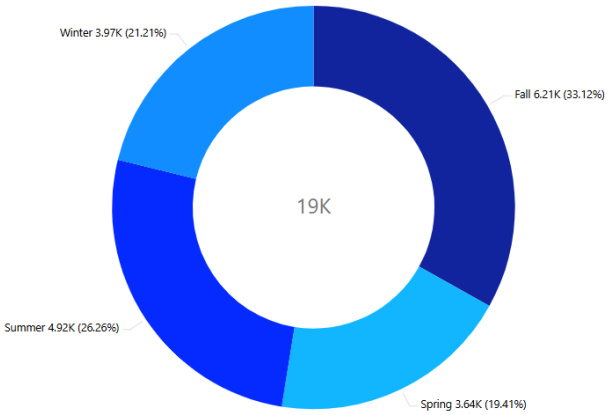
- **Handle Missing Values:**
  - Removed column called Postel code from data because more than 80% of data is null.
- **Duplicated Rows:**
  - We found that all rows do not have duplicated.
- **Removing unimportant features:**
  - We removed these features: Row ID – Postel Code – Customer Name – Product Name because we realized that not important in our Analysis
- **Extracting New Features:**
  - Extracting new columns called [Order Day – Order Month -Order Year – Order Month Name - Order Day Name] from existing column called Order Date
  - Creating column called Order Season from Order Month column by made function to did that
  - Creating column called Is_Black_Friday from Order Date to know if orders ordered on Black Friday or not
  - Creating a column called Ship_Days from column Order_Date and Ship_date to know how many days it will take to order reach to customer
- **Handling Outliers:**
  - Remove outlier data and became from 51290 row to 30652
    More than 20638 rows removed
- **Data consistency:**
  - We checked that no data is reflected have spelling errors

Data Analysis

- **statistical analysis:**
    - We detected that **positive** relationship between Sales & Profit equal **0.48** mean Increased sales lead to increased profit.
    - Sales & Shipping Cost**: Strong positive** relationship **0.77**, meaning higher sales increase shipping costs.
    - Quantity & Sales**: Moderate positive** relationship **0.31**, large quantity increases sales slightly
    - Discount & Profit**: Moderately negative** relationship **-0.32**, large discount reduces profit.
    - Profit & Shipping Cost: **Moderate positive** relationship **0.35**, higher shipping costs are associated with higher profit.

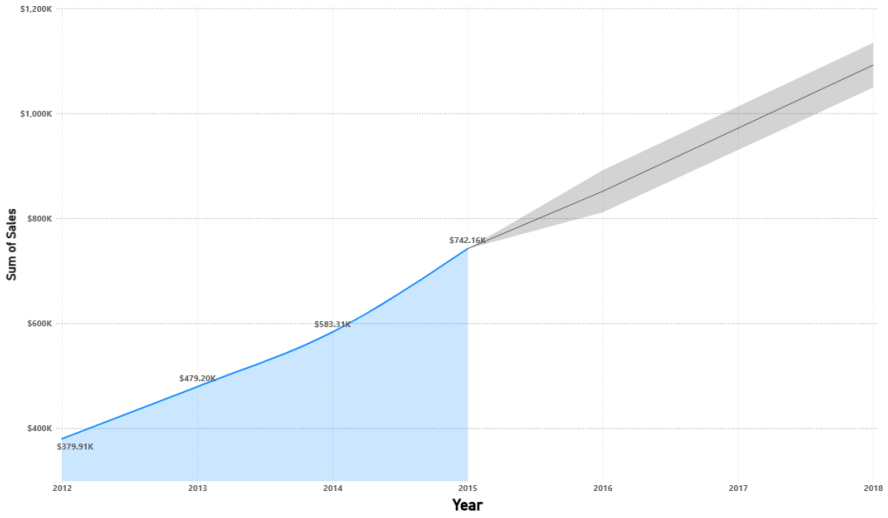
- **Seasonality Analysis:**
    - We get Total sales & Orders in **Fall season** is a highest percentage equal **$714.70k (32.72%)** after that second is **Summer $575.52k (26.34%)** after that coming **Winter** and **Spring**



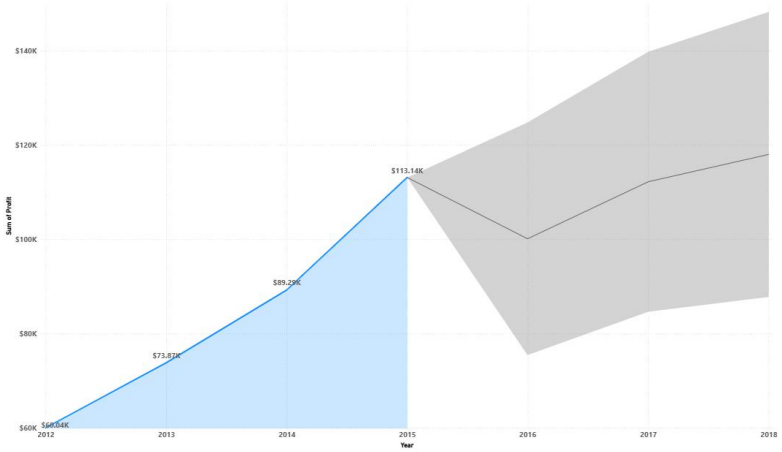


Data Visualization

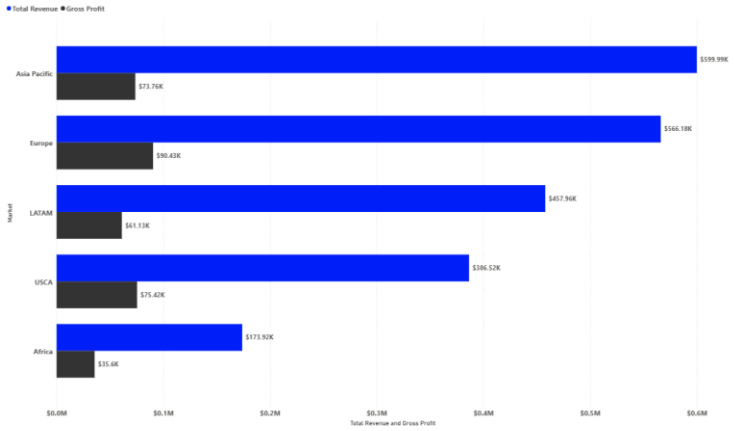
- **Analysis of Sales & Profit Overtime:**
    - Total sales increase year by year we see that in 2012 total sales is $379.91k , 2013 is $479.20k , in 2014 equal $583.31k and in 2015 equal $742.16k and so on. We predict that in 2018 total sales will reach $1,134,689M



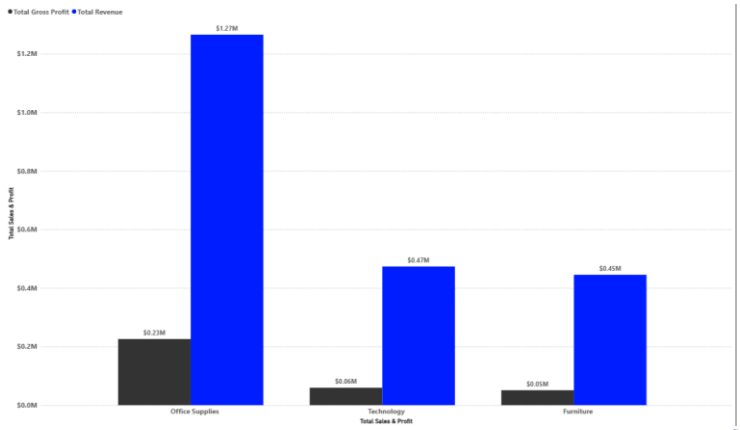
- Total Profit year by year profit increase a little bit profit from 2012 to 2015 Increase at a $73k almost and we predict that in 2018 profit will increase to be $148,263k



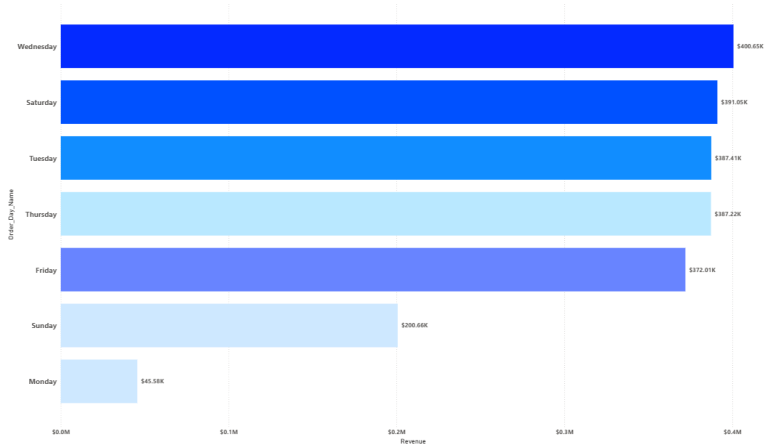
- Here we get that Market Asia is the highest market inside Sales = $599.99k but is not the highest in profit.
  And the lowest market is Africa in both sales and profit



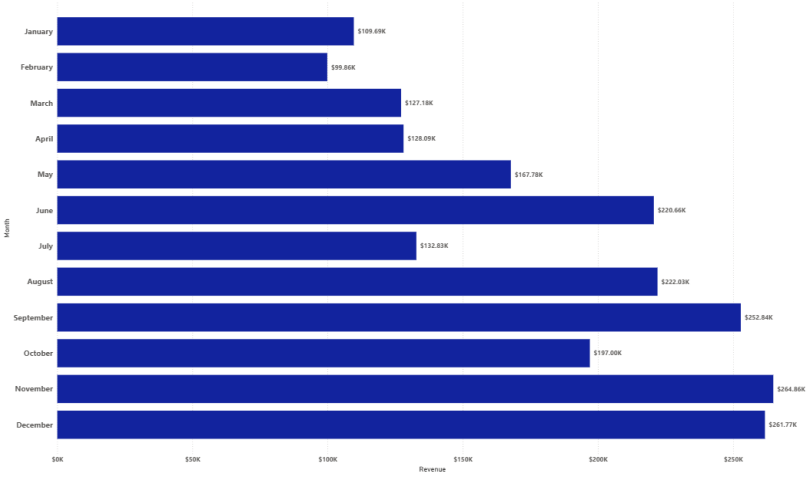
- Sales and profit by categories the category Office Supplies is achieve sales to equal $1.27M and Profit $23k is the highest category by the way
  And the lowest category inside sales & profit we will get Furniture



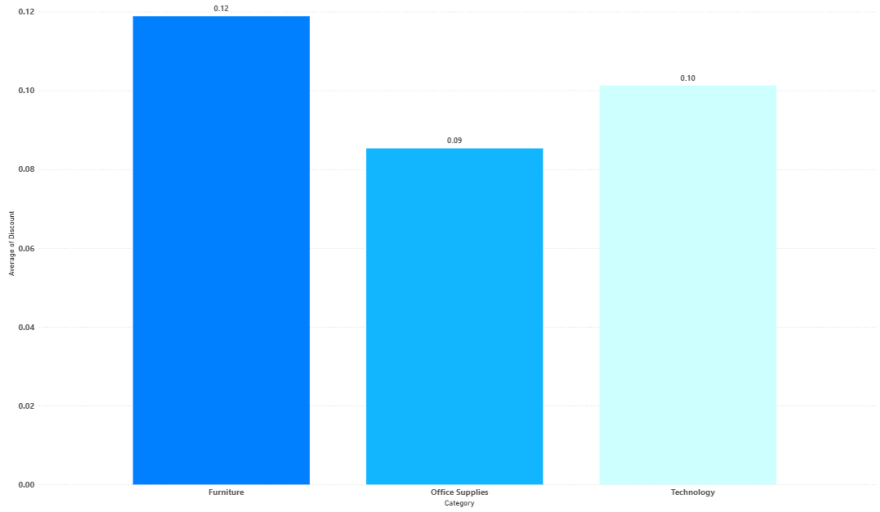
- Now we get the best day that sales is much we get Wednesday & Saturday is the high between $390k to $400k & Sunday, Monday is the lowest level between $45k to $120k



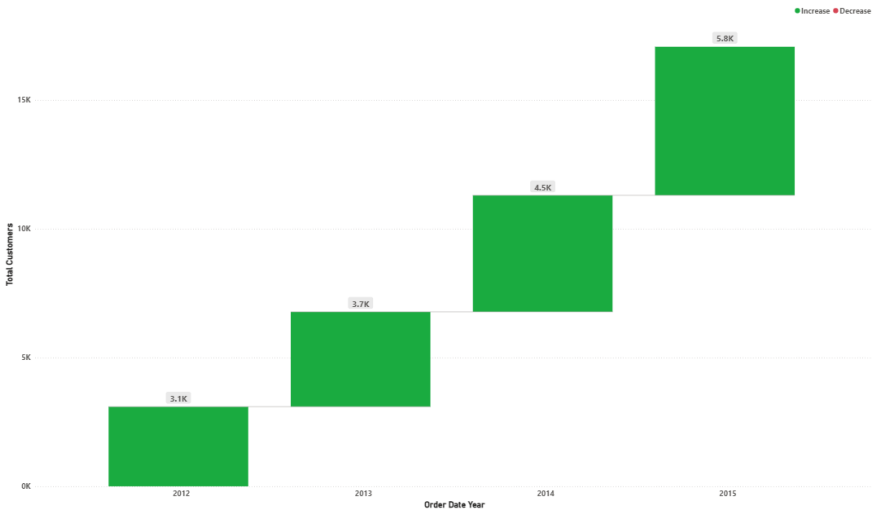
- Last not least we get the best month that sales is high in November & December sales are between $260k to $264k
  And the bad months are January & February are between $99k to $109k



- We know the main factor in making a profit & sales is not good is the high discount. The average discount between all categories is between 9% to 12%



- The last thing we want to discuss is over year the total number of customers is high we will get in 2012 number of customer is 3.1k and in 2015 the total is 5.8k



Key Findings and Insights

- **Sales:**
  There has been a significant increase in sales volume over the years, especially in 2015, and it is expected that in 2018 sales will exceed 1 million.
- **Profit:**
  The profit margin increases by a very small percentage due to:

- The extremely high volume of discounts
- And the very high shipping cost.

- **Market:**

  The sales rate in the Asian and European markets is good, even the Latin American market is somewhat acceptable, but the African market is very low in terms of sales because of the poor countries' expectation of the proposal to make a relative reduction to increase profits there.

- **Increase number of customers:**

  Customer growth rate is a very good indicator over the years.

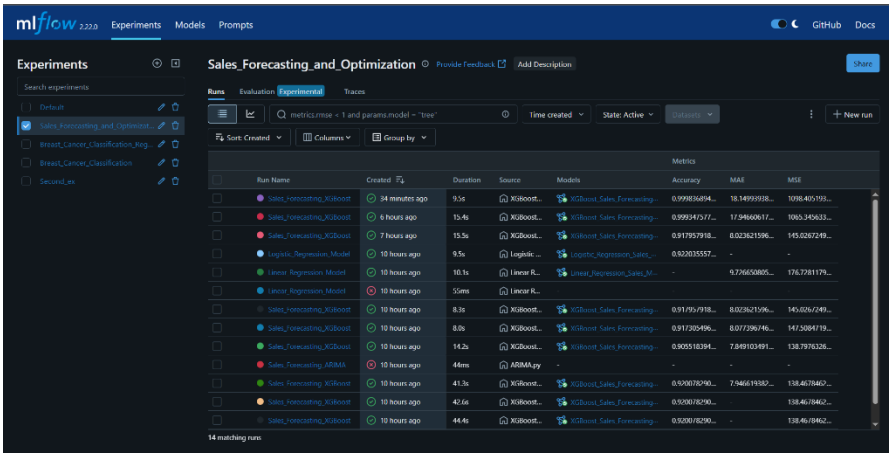
- **Sales in Months & Days:**

  It is suggested to make promotions on days when there are not many sales to attract customers to buy.

Forecasting Model Deployment

- Streamlit UI: This image shows the interface of a sales forecasting application using XGBoost, allowing you to enter variables such as shipping method and order date.

- **MLFlow**: This image contains a mlflow interface showing sales forecasting experiments using XGBoost, with performance metrics such as RMSE and MAE. and We see their versions and the ones that came to the models. And at the same time, we save or log the metrics of these models so that we can compare them.


mlflow 2.22.0
Experiments
Models
Prompts
GitHub
Docs
Experiments
Sales_Forecasting_and_Optimization
Provide Feedback
Add Description
Share
Runs
Evaluation
Experimental
Traces
New run
Sort: Created
Columns
Group by
Metrics
Run Name
Created
Duration
Source
Models
Accuracy
MAE
MSE
14 matching runs