Improving Inventory Management and Sales through Predictive Product Popularity Analysis

Christian Prado

Table of Contents

[Part A: Project Proposal for Business Executives 2](#_Toc1515453419)

[Letter of Transmittal 3](#_Toc1044087116)

[Project Recommendation 3](#_Toc1721750458)

[Problem Summary 3](#_Toc205640471)

[Application Benefits 3](#_Toc739883765)

[Application Description 3](#_Toc1967641075)

[Data Description 3](#_Toc1580263866)

[Objectives and Hypothesis 4](#_Toc1084915425)

[Methodology 4](#_Toc1455093807)

[Funding Requirements 4](#_Toc1716359054)

[Data Precautions 4](#_Toc821607173)

[Developer’s Expertise 4](#_Toc742297278)

[Part B: Project Proposal 4](#_Toc1516064891)

[Problem Statement 5](#_Toc602847551)

[Customer Summary 5](#_Toc411678477)

[Existing System Analysis 5](#_Toc1668713332)

[Data 5](#_Toc1911399030)

[Project Methodology 5](#_Toc1307313575)

[Project Outcomes 5](#_Toc1967172832)

[Implementation Plan 5](#_Toc306674583)

[Evaluation Plan 6](#_Toc849271423)

[Resources and Costs 6](#_Toc1695791095)

[Timeline and Milestones 6](#_Toc787992355)

[Part C: Application 6](#_Toc1767968421)

[Part D: Post-implementation Report 7](#_Toc22579049)

[A Business (or Organization) Vision 8](#_Toc1496854501)

[Datasets 8](#_Toc645781746)

[Data Product Code 8](#_Toc157285634)

[Objective (or Hypothesis) Verification 8](#_Toc1315546441)

[Effective Visualization and Reporting 8](#_Toc656892821)

[Accuracy Analysis 9](#_Toc1956352798)

[Application Testing 9](#_Toc871280264)

[Application Files 9](#_Toc1639588139)

[User Guide 9](#_Toc1125023771)

[Summation of Learning Experience 9](#_Toc1702287338)

# Part A: Project Proposal for Business Executives

## Letter of Transmittal

February 22nd, 2020

Wile E. Coyote - Chief technology officer

IntelliStock

101 N. Stock Blvd

Los Angeles, CA 90210

Dear Wile E. Coyote,

I'm writing to address a critical issue our company is experiencing: inventory management and sales. Our company has struggled to predict which products are popular with customers, resulting in stockouts and overstocking. Our sales and inventory management processes have been severely hampered as a result of this lack of information.

Our team has proposed a data product based on historical sales data that uses machine learning to predict product popularity. The proposed solution will identify patterns and relationships between products and customer behavior, allowing us to manage inventory better, avoid stockouts and overstocking, and promote the right products to customers.

Our proposed solution will benefit the organization significantly by allowing us to improve inventory management, avoid stockouts and overstocking, and increase sales. Furthermore, our solution will provide precise and actionable insights into product popularity, enabling us to optimize inventory management and product promotion.

The total cost of the project is estimated to be $248,000. This includes the cost of two data scientists and engineers for 20 weeks and any hardware, software, and third-party services that may be required.

Our team has extensive experience and expertise in data analytics and machine learning and a track record of completing projects in various industries. Our experience and expertise qualify us to create a solution for the organization's needs.

Sincerely,

Christian Prado

Data scientist

## Project Recommendation

### Problem Summary

Our proposed project aims to improve our organization's stock management and sales by predicting which products are most likely popular. The ability to predict product popularity is crucial to our success, allowing us to manage our inventory better and stock the right products in our stores. However, we need to effectively identify which products are most popular among customers, which can lead to out-of-stock and overstocking. The use of simple machine learning to predict product popularity is a concept that has been successfully practiced in various industries such as retail, e-commerce, and manufacturing.

We have been in business for over a decade. With such a long history, we have a large and difficult-to-manage inventory. We need to determine which of our products will most likely be popular as efficiently as possible to reduce stock issues and help us increase sales. We plan to develop a dashboard with a machine learning modal to analyze our customer behavior and purchase history.

### Application Benefits

Using the suggested solution will assist us in addressing the critical challenge of identifying the most popular products, which will be highly beneficial to our company. Our bottom line has been impacted by issues like out-of-stock and overstocking caused by this dilemma. We can evaluate our data using simple machine learning and precisely forecast which products will be popular. Enabling better inventory management, ensuring we have products on hand and reducing the possibility of future inventory problems.

Making data-driven decisions and remaining competitive in the retail industry will be aided by the proposed project. We can learn more about our customers' behavior, such as their purchase history or other relevant factors, which will be used to predict popular products. As a result, we can adjust our inventory and focus more on our marketing efforts with more knowledge about our products. The solution will also provide us with real-time insights and reports, allowing us to make more informed decisions faster and improve our inventory oversights and sales by implementing the suggested solution, increasing profitability and success in our industry.

### Application Description

The proposed data product will leverage collaborative filtering algorithms to predict which products will be popular among customers. *Collaborative filtering* is a machine learning algorithm that analyzes customer behavior, purchase history, and other factors to identify patterns and make predictions. Specifically, the algorithm will analyze data on past customer behavior, such as purchase history and browsing data, and make predictions on what products will be popular based on patterns in this data. The algorithm will be trained on a dataset of customer behavior and validated through a testing phase to ensure that it is accurate and effective.

The product will be built with Python and Sckit-learn programming languages and frameworks and hosted on a cloud platform like Google Colab. The solution will be accessible via an easy-to-use web interface, allowing users to input data and generate real-time reports and insights. The solution will also be designed to be scalable, allowing it to handle large amounts of data while also supporting our organization's long-term growth. Overall, the solution's technical details center on leveraging cutting-edge machine-learning techniques and cloud-based technology to provide an accurate and efficient solution for predicting product popularity.

### Data Description

This project's raw data will be sourced from our organization's internal databases, which store information on customer behavior, purchase history, and other relevant factors. The data will be preprocessed to remove any missing or irrelevant information and transformed so that the collaborative filtering algorithm can analyze it.

Purchase history and customer addresses will be among the quantitative and nominal variables included in the data. Furthermore, the data structure, such as a relational database or a data warehouse, will be organized for efficient analysis.

The prediction of product popularity is the dependent variable for this project, while the independent variables will include customer location, purchase history, and product attributes. The collaborative filtering algorithm will use these variables to identify customer behavior patterns and predict which products will be popular.

Data anomalies and outliers may exist due to incomplete or incorrect data or changes in customer purchasing behavior over time. These constraints will be addressed during the project's preprocessing and validation phases, ensuring that the data used to train and validate the algorithm is accurate and effective. Furthermore, the data will be updated and refined regularly to ensure the algorithm remains accurate and current.

### Objectives and Hypothesis

This project aims to develop a machine-learning model that accurately predicts which products will be popular among our organization's customers. We want to improve our prediction accuracy to manage inventory, reduce stockouts and overstocking, and improve our overall sales performance.

One hypothesis for this project is that by analyzing customer purchase history and behavior, we can identify patterns that will accurately predict which products will be popular among our customers. We can create a predictive model with the desired level of accuracy by analyzing this data with a collaborative filtering algorithm. As a result, the collaborative filtering algorithm is expected to predict which products are most popular among customers accurately. This model will improve inventory management and sales performance in our organization. Throughout the project, the hypothesis will be tested to ensure that it is correct and that the model produces the desired results.

### Methodology

We will employ an agile methodology to develop and implement the machine learning model for this project. Agile methodology is appropriate for this project because it allows flexibility and adaptation to change requirements and priorities. It also emphasizes collaboration and communication, which are critical when working with non-technical stakeholders.

Planning, data preparation, model development, testing, and implementation will be the phases of the project methodology. We will define the project objectives and scope, identify the data sources and tools required, and establish the project timeline and budget during the planning phase. During the data preparation phase, we will collect and clean the data, identify any missing or incorrect data, and perform any required feature engineering.

The collaborative filtering algorithm will create the machine learning model during the model development phase. This will entail choosing the right features, fine-tuning hyperparameters, and validating the model's accuracy. We will then evaluate the model's performance and make any necessary adjustments during the testing phase. Finally, during the implementation phase, we will incorporate the model into the organization's existing systems and processes and create necessary documentation and training materials for stakeholders. We will use agile methodologies to manage the development process throughout the project, with regular reviews and feedback loops to ensure the project stays on track and meets the desired outcomes.

### Funding Requirements

The proposed project requires **$248,000** in funding for the 20-week duration. This will cover the costs of data collection, cleaning, and analysis of hardware and software, as well as the salaries of two data scientists and two engineers. The purchase of high-performance computers and storage devices, as well as data communication and networking equipment, is included in the hardware cost. In addition, licenses for various machine learning tools and platforms and other software tools required for developing and deploying the proposed solution are included in the software costs.

Furthermore, the project requires hiring two data scientists and two engineers. The data scientists will be in charge of creating the machine-learning algorithms. Simultaneously, the engineers will integrate them into the organization's systems. The project's development cost is estimated to be $192,000, including salaries, benefits, and overhead costs.

Finally, licenses for third-party tools and platforms are required for the project to develop and deploy machine-learning models. These are estimated to cost **$45,000** each.

### Data Precautions

Kaggle is a data-sharing platform where businesses and organizations can share their data with data scientists and other professionals looking to build machine learning models, gain insights, and improve their skills. The data shared on Kaggle is open to the public, so anyone who wants to use it can. As a result, Kaggle has no sensitive or protected data restrictions. Furthermore, any data shared on Kaggle is voluntarily done by the organizations or individuals who own it. They usually share it to promote data science research and innovation. However, even if the data on Kaggle is not sensitive or protected, appropriate measures must be taken to ensure privacy and security.

### Developer’s Expertise

As a data scientist, I am skilled in software development and interacting with big datasets. I hold a master's degree in computer science with a concentration in data science. In my current position, I have created and implemented multiple machine-learning programs for many companies and different industries, all of which have greatly improved those businesses.

My experience in this field qualifies me to provide a solution that meets your organization's needs. With my machine learning and data science expertise, I can help your company predict which products will be popular among our customers and improve inventory management, increasing sales and revenue. Furthermore, my experience working with sensitive data and adherence to ethical and legal considerations will ensure that the project is carried out responsibly and securely.

**Part B: Project Proposal**

## Problem Statement

Inventory management and sales have become significant challenges for businesses in today's world of online shopping. Businesses require assistance in tracking which products are most popular among customers, which can lead to overstocking or stockouts. To address this issue, we propose a machine learning-based solution that predicts product popularity among customers. Businesses will benefit from the proposed solution by improving inventory management, ensuring product availability, and lowering costs.

The proposed project will concentrate on implementing a supervised machine-learning algorithm called Collaborative Filtering. Customer behavior data is used in collaborative filtering to predict other customers' preferences. The algorithm will analyze customer purchasing habits and recommend products likely to be popular with other customers. The solution will also reveal the most popular products, allowing our company to adjust its inventory and marketing strategies accordingly. As a result of this solution, our company can optimize its inventory levels and increase sales.

## Customer Summary

The proposed project will concentrate on implementing a supervised machine-learning algorithm called Collaborative Filtering. Customer behavior data is used in collaborative filtering to predict other customers' preferences. The algorithm will analyze customer purchasing habits and recommend products likely to be popular with other customers. The solution will also reveal the most popular products, allowing our company to adjust its inventory and marketing strategies accordingly. As a result of this solution, our company can optimize its inventory levels and increase sales.

IntelliStock can reap significant benefits by implementing a data product that predicts product popularity using machine learning and collaborative filtering algorithms. The product will enable them to make more informed decisions about which products to stock, resulting in reduced inventory, increased sales, and higher profits. Furthermore, the data product will integrate seamlessly into their existing systems and processes, making it simple to use and saving time.

## Existing System Analysis

We currently use an inventory management system to monitor product inventory levels, sales, and customer purchase history. On the other hand, the system must provide insight into which products are most popular among customers, resulting in low stock and overstocking issues. This information should always be included to make inventory management easier. Furthermore, the current system lacks machine learning algorithms for forecasting product demand, which could improve inventory management and sales forecasting.

The proposed solution will predict product demand using machine learning algorithms, allowing for better inventory management and sales. This new system will fill the gaps left by the current inventory management system by providing real-time data on product popularity, demand, and other inventory management factors. The system will also recommend promotions and assist marketing efforts based on product popularity and other factors. This new system will make data-driven decisions using machine learning, resulting in more effective inventory management, higher sales, and higher profits.

## Data

This project's data set includes customer purchase history, product details, and ratings. First, the data will be collected and preprocessed from the organization's existing transactional database before being fed into the collaborative filtering algorithm. Then, we will establish the data collection and processing pipelines during the design phase to ensure efficient and accurate data handling. Finally, the collected data will be managed and stored in a data warehouse for easy access and maintenance.

We will then use cleaning techniques such as data correction, outlier detection, and data transformation to deal with incomplete and missing data anomalies. These techniques then improve the accuracy and relevance of our predictive model.

We can then ensure that the data is constantly monitored and evaluated throughout the development and maintenance phases for completeness and accuracy. This will help maintain quality and the organization's continued delivery of the requested results.

## Project Methodology

We will use the Agile methodology for this project to develop and deploy our application. Agile is an iterative and collaborative methodology that promotes continuous development and delivery of working software. In addition, it enables us to be flexible and responsive to changing requirements throughout the project.

The first phase is the Planning phase. During this phase, we will define the project's scope, identify stakeholders, and create a backlog of requirements. Next, we will establish project milestones, set a timeline, and determine the resources required to complete the project.

In the second phase, we will focus on Analysis and Design. During this phase, we will refine the requirements and create user stories. We will analyze the data to identify trends and patterns and design the data model and user interface.

The third phase is Development. We will start building the application using the appropriate programming languages and frameworks during this phase. We will implement features in sprints and conduct regular testing to ensure product quality.

The fourth phase is Testing. During this phase, we will conduct various types of testing, including unit, integration, and system. We will ensure that the product meets the requirements and is high quality.

The fifth phase is Deployment. During this phase, we will deploy the application to the production environment and ensure it is stable and scalable. In addition, we will provide support and maintenance to ensure that the application is running smoothly and meeting the users' needs.

## Project Outcomes

Jupyter Notebook was selected since it can compile all the required files in one location. Furthermore, Jupyter Notebook allows easy data exploration and analysis while also providing an easy-to-use interface for sharing our project. In addition, we can guarantee that a fully functional application is provided and can accurately predict popularity based on our sales data.

* **Data preprocessing and cleaning code:** This step will clean the raw data set by removing duplicates, dealing with any missing data, and helping in removing outliers and is fit for analysis.
* **Code for exploratory data analysis:** This step will provide insights into the data, including visualizations that reveal trends, patterns, and correlations using charts and graphs.
* **Machine learning code:** This step will contain the algorithms and models used to forecast product popularity based on the sales data. It will also include the code that evaluates and adjusts the model's performance as needed to ensure maximum accuracy.
* **User guide:** This will provide step-by-step instructions on how to use the application and troubleshoot common problems. It will be simple to use and accessible to all users.

A comprehensive user guide will be created to assist users in adopting and understanding the application. The user guide will contain instructions on the application's features and functionality and step-by-step instructions for using it. The user guide will be written in English, including some helpful hints and troubleshooting suggestions for technical and non-technical users.

## Implementation Plan

The project will be implemented using an agile methodology with iterative development and stakeholder feedback. The following phases will comprise the overall strategy:

* **Data Gathering and Preparation:** During this phase, raw data is collected from the identified data sources and prepared for use in the project. Cleaning the data, dealing with missing values and outliers, and transforming the data to the appropriate format are all part of the preparation process.
* **Model Development:** The collaborative filtering modal will be built in Jupyter Notebook during this phase. The algorithm will be trained and tested on the pre-processed data to ensure it produces the desired result.
* **Integration and testing:** During this phase, the collaborative filtering algorithm is coded into the user interface, and the system is tested for errors. User acceptance testing will also be performed to ensure that the system dashboard is simple to use.
* **Deployment and Maintenance:** The final phase entails deploying the application and providing end-user maintenance support.

The rollout will be carried out in stages. The algorithm and user interface will be tested in a controlled environment with a small group of stakeholders during the first phase. The system will be scaled up in the second phase to include more data and a larger group of stakeholders. Access to data sources and the availability of personnel and tools are all project dependencies.

Testing will take place at various stages of development. Unit testing, system testing, and acceptance testing will all be performed. The project will be distributed via the client's internal network.

## Evaluation Plan

Verification methods are critical for ensuring that the project stays on track and that the product's quality is maintained. The verification methods will vary depending on the stage of development.

* **Gathering of requirements:** Several verification methods will be used to correctly identify the client's needs during the requirements gathering phase. To learn about their needs, we will conduct stakeholder interviews, review existing documentation and data, and administer surveys.
* **Design:** We will use manual and automated techniques to ensure the design meets the requirements during the design phase. We will perform code and design reviews and use automated tools to ensure compliance with coding standards and requirements.
* **Implementation:** Various techniques will be used to test the code's quality and functionality. Code reviews, unit testing, and integration testing are examples of these. In addition, we will use tools to automate the testing process and ensure that the code meets the requirements.
* **Deployment:** We will use various techniques to ensure the application works as expected in the production environment. We will perform load, stress, and performance testing to ensure the application is ready for production.

The verification methods used at each stage will help ensure that the project proceeds as planned and that the application meets the client's needs.

Following project completion, the validation method will include a series of tests to ensure that the developed application meets the desired objectives and performs the intended functions accurately:

1. The data used to train the machine learning model will be validated to ensure completeness and accuracy.
2. The model's performance is evaluated using statistical techniques such as accuracy, rmse, and mae scores.
3. Users will test the application to ensure it is intuitive, user-friendly, and meets our requirements.

## Resources and Costs

**Hardware:**

* Two servers with sufficient storage (1tb minimum) and computational power (2x 3090 GPUs) to handle large amounts of data.
* 4 workstations and 2 laptops for data preprocessing, model development, and evaluation.

**Software:**

* A machine learning framework: TensorFlow and Scikit-learn.
* A programming language: Python.
* A data visualization tool: Matplotlib and Seaborn.
* A data preparation tool: Pandas and NumPy.

**Work hours:**

* Data collection and preprocessing: 260 hours
* Model development and training: 180 hours
* Model evaluation and optimization: 60 hours
* Deployment and integration: 40 hours
* Monitoring and Maintenance: 60 hours

**Third-party services:**

* Cloud storage from Amazon S3 Storage for storing and processing data.
* Cloud computing resources from Amazon EC2 for running the machine learning models.
* Third-party open-source data cleaning and preprocessing services (no cost), if necessary.

|  |  |  |
| --- | --- | --- |
| **Resource** | **Description** | **Cost** |
| Hardware | Servers and workstations | $11,000 |
| Software | proprietary software | $5,000 |
| Work hours | Two data scientists and two engineers for 20 weeks | $192,000 |
| Third-party services | Cloud storage and compute resources (20 weeks) | $40,000 |
|  | **Total** | $248,000 |

## Timeline and Milestones

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Milestone** | **Start Date** | **End Date** | **Duration** | **Dependencies** | **Resources Assigned** |
| Requirements Gathering | 01/20/2023 | 02/02/2023 | 12 days | None | Data Analyst, Project Manager |
| Data Collection | 02/03/2023 | 02/16/2023 | 10 days | Requirements Gathering | Data Analyst, IT Support |
| Data Cleaning | 02/17/2023 | 03/02/2023 | 12 days | Data Collection | Data Analyst |
| Exploratory Data Analysis | 03/03/2023 | 03/23/2023 | 15 days | Data Cleaning | Data Analyst |
| Model Development | 03/24/2023 | 04/20/2023 | 20 days | Exploratory Data Analysis | Data Scientist |
| Model Testing | 04/21/2023 | 05/05/2023 | 10 days | Model Development | Data Scientist, IT Support |
| Model Deployment | 05/06/2023 | 05/26/2023 | 15 days | Model Testing | Data Engineer, IT Support |
| User Guide | 05/27/2023 | 06/08/2023 | 10 days | Model Deployment | Technical Writer |
| Final Report | 06/09/2023 | 06/15/2023 | 5 days | User Guide | Project Manager, Data Analyst |

# 

# Part C: Application

See *Application Files* in **Part D: Post-implementation Report** for files and how to access this project.

# Part D: Post-implementation Report

## A Business (or Organization) Vision

A company wishes to increase sales by better understanding its customers' purchasing habits and identifying its best-performing products. Therefore, significant sales data must be analyzed and visualized to gain insights.

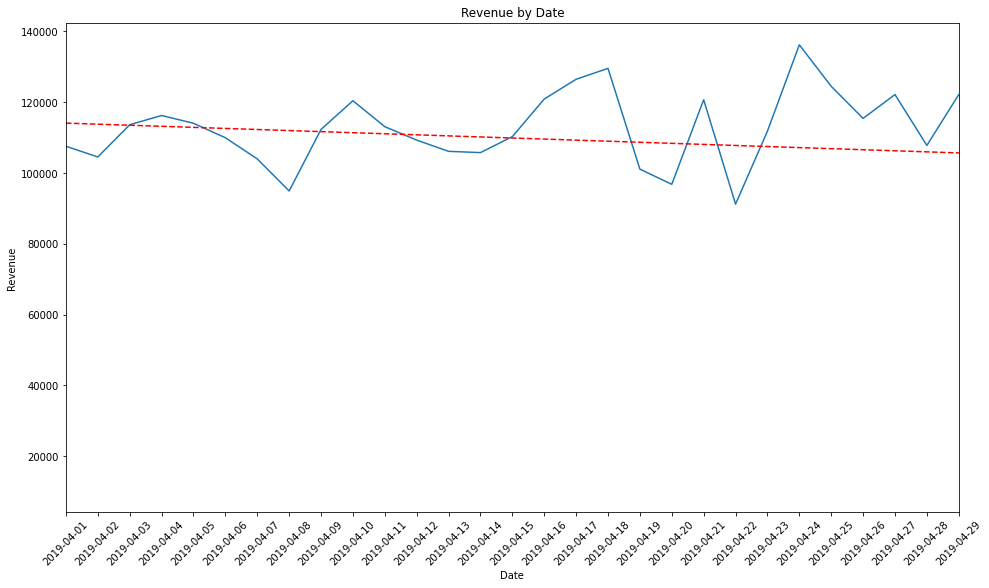
The sales data analysis and visualizations provide the company with tools for analyzing and visualizing sales data. The application, for example, can assist the company in identifying top-performing products, sales trends, and the relationship between product features such as price and popularity. The company can make data-driven decisions to improve its sales and customer satisfaction with this information.

To use the application, the user must first export their sales data to a CSV file and then import it into the application. The application can then present the user with various data visualizations and analyses. The product popularity dashboard, for example, can be used by the user to identify the top-performing products and their correlation with other features. Furthermore, the scatter plot can be used to compare predicted versus actual values for a product, and the heatmap can identify the relationship between different features.

The sales by weekday chart can assist the user in identifying sales trends by weekday, and the time series plot can show revenue trends over a specific period. The parallel coordinate plot allows the user to compare sales data across multiple dimensions. Finally, the histogram or density plot can be used to depict the distribution of a particular feature, such as price.

The user can gain insights into their sales data and make data-driven decisions to improve their sales and customer satisfaction using various visualizations and analyses.

The revenue by date time series plot is one example of a visualization in the application. The plot depicts revenue trends over a specific period, allowing the user to identify sales trends and seasonality. In addition, a trend line was added to the plot to demonstrate how sales decrease over time.



## Datasets

The data for our project was a CSV file containing over 186,000 rows of sales data found on Kaggle. The data included the fields such as the order date, the product name, the price, the quantity ordered, the customer's address, and product ratings.

We first loaded the CSV file into a pandas DataFrame using the read csv function. Then, the data was cleaned up by manually removing any missing values and duplicate rows inside excel.

The data processing involved adding new columns to the dataset, such as the User ID. We also converted the 'Order Date' column to a datetime format to be easily manipulated later. This enabled us to study data trends based on the day of the week, month, or a specific time.

The processed data was then used for data visualizations such as a time series plot and correlation heatmaps to help gain insights into product popularity, sales trends, and other topics. Our project files contain the data, which can be accessed for further analysis or built upon.

## Data Product Code

The code used in our project involved several steps in data analysis and application development. First, we began processing the raw data by adding User ID and Rating columns. We then analyzed the data using various descriptive methods and visualizations, such as bar charts, scatter plots, heatmaps, and time series plots, to gain insights into the most popular products and sales by weekday. These descriptive methods assisted us in identifying trends and patterns in the data, which aided in the development of our application.

In addition to descriptive methods, we used non-descriptive methods to make predictions and recommendations, including machine learning algorithms such as Singular Value Decomposition (SVD) and Matrix Factorization. These methods were selected because they suit collaborative filtering and provide accurate predictions by identifying patterns and relationships between users and products. We trained and tested these methods by dividing our data into training and testing sets and evaluating their accuracy using root mean square error (RMSE) and mean absolute error (MAE).

Overall, the analysis aided in selecting and improving our descriptive and non-descriptive methods by identifying trends and patterns in the data and laying the groundwork for making accurate predictions and recommendations. This enabled us to create an application that can provide personalized recommendations to users based on their preferences and purchase history, resulting in a better user experience.

## Objective (or Hypothesis) Verification

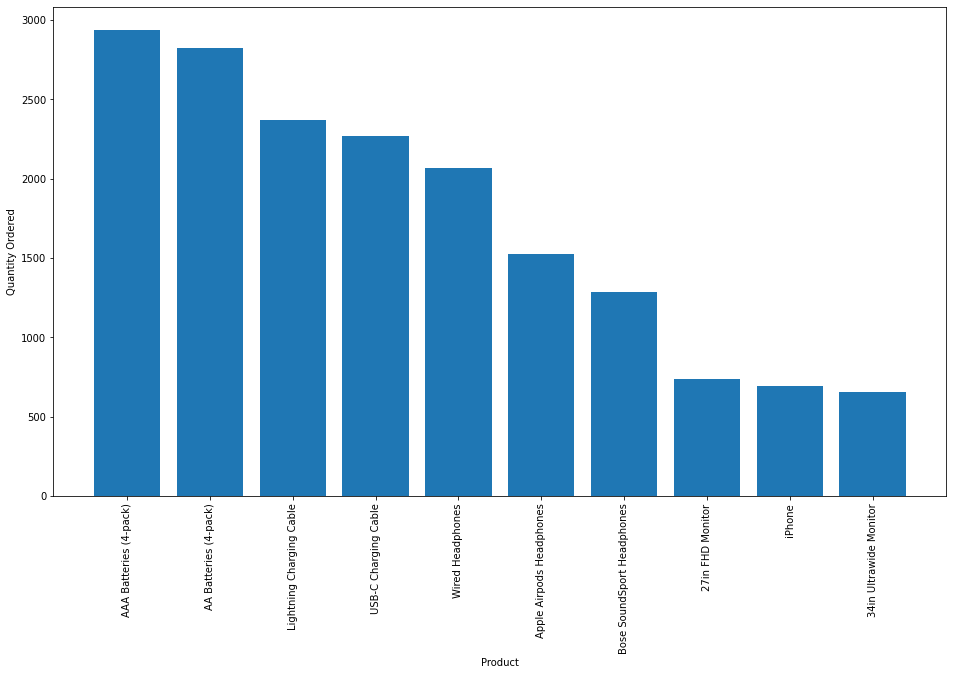
The project aimed to identify popular products and sales trends for a given dataset of e-commerce transactions using data analysis techniques. The hypothesis was that we could identify the most popular products and their sales trends by analyzing transaction data.

Various techniques, such as data visualization, trend analysis, and machine learning, were used to complete this project. For example, machine learning was used to identify popular products, determine sales trends for those products, and even forecast some future sales. As a result, we successfully identified key trends and patterns that can help make informed decisions about inventory planning, product pricing, and marketing strategies.

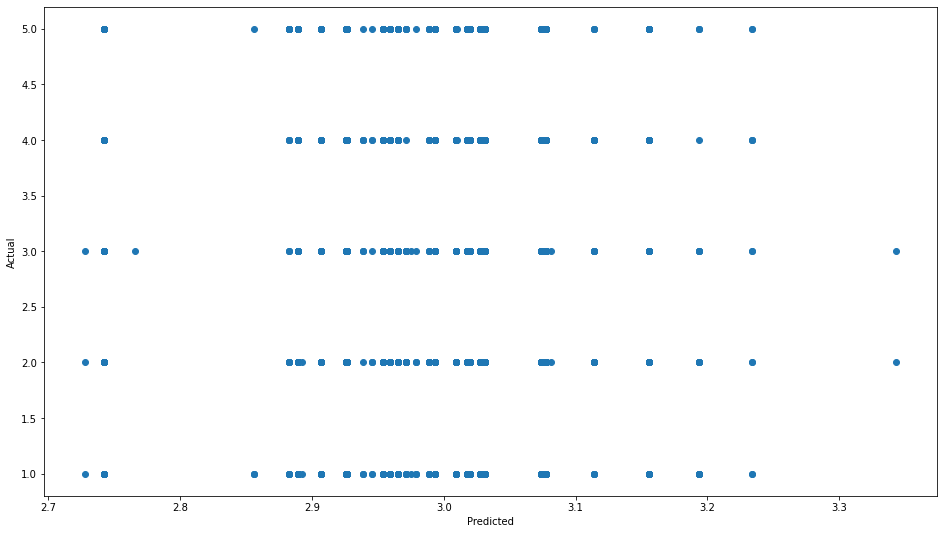
Finally, the project's goal was met, and the results showed that they could aid in improving business operations. Furthermore, the project's techniques effectively identified critical insights from data and provided valuable business recommendations.

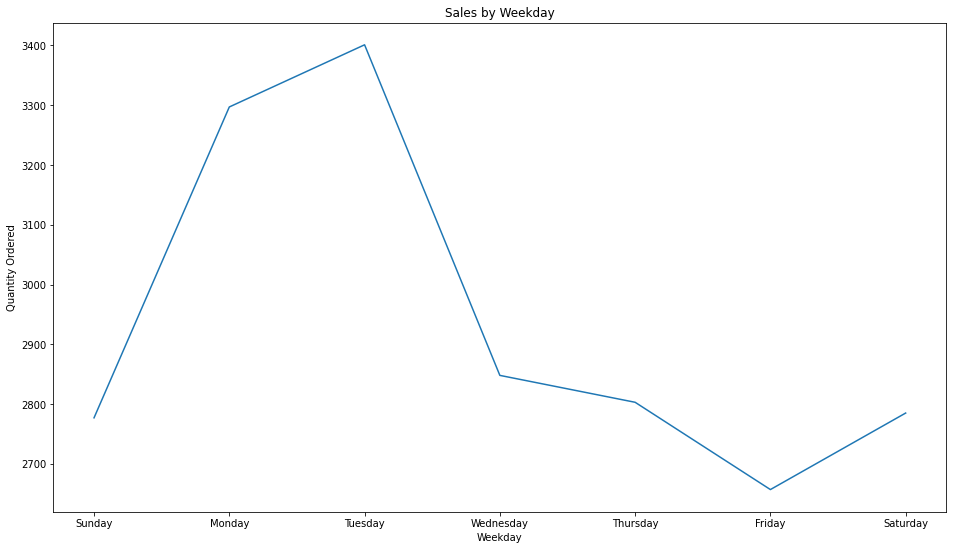
## Effective Visualization and Reporting

The descriptive and visualization methods were important in the data exploration, analysis, and summary during the development process. In addition, we used various visualizations to understand better the data and its distribution, which aided us in selecting the best models and features for training and testing our non-descriptive methods.

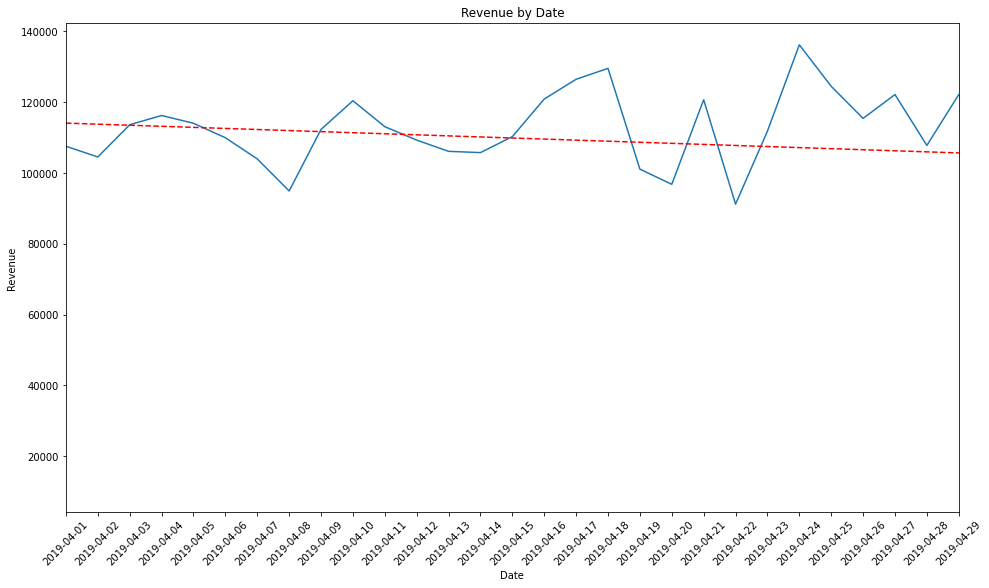


To start, we used a bar chart that listed the most popular products, which enabled us to see which products were in high demand. This chart helped the development of our non-descriptive method by clearly understanding the most profitable products, allowing us to focus on these products when developing our models.

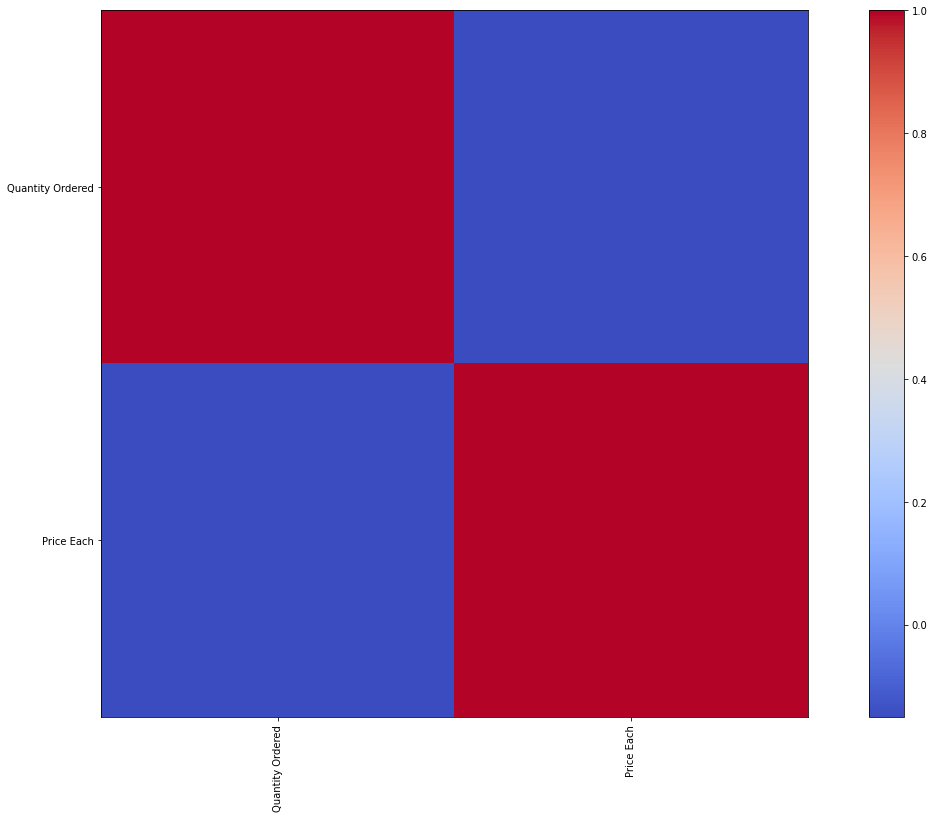


Second, we used a scatter plot to visualize the relationship between our predicted and actual values. The plot helped us determine whether our model accurately predicted the values and whether there was a correlation between the predicted and actual values.

The Sales by Weekday chart was instrumental in providing insight into customers' purchasing patterns during the week. The chart showed that sales were highest during the weekdays, with a significant drop on weekends. This information was valuable for the business. It allowed them to optimize their staff scheduling and inventory planning, ensuring they had enough personnel and products to meet the higher demand during the weekdays.



The Revenue by Date chart assisted the company in tracking its revenue over time, allowing them to analyze how its sales performed daily. The company identified trends in their sales and analyzed how their business was performing throughout April by visualizing the data in a time series plot. The chart also included a trendline, which assisted in determining whether sales were trending upward or downward, providing the company with valuable insights into the future of its business. By tracking revenue daily, the company can make data-driven decisions about pricing, promotions, and inventory planning, maximizing profits and improving business performance.



Finally, we also incorporated a heatmap to see the relationship between the popularity and price of the products, which also helped us understand the connection between them. In addition, this visualization enabled us to decide which features to include in our model.

The heatmap was essential to understanding the data, categorizing, and selecting appropriate features and models for our non-descriptive methods. And as a result, we were able to make decisions and develop effective models that could accurately predict product popularity and revenue.

## Accuracy Analysis

We used the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) in our project to evaluate the accuracy of our non-descriptive method. The RMSE measures the average difference between predicted and actual ratings. The MAE is a metric that measures the absolute difference between predicted and actual ratings. The RMSE in our model was 1.4329, and the MAE was 1.2317. These results show that our model can predict product ratings with a reasonable degree of accuracy.

We used the Surprise library to train a Singular Value Decomposition (SVD) model on our dataset to demonstrate the accuracy. The data were divided into training and testing sets, and the model was then trained on that training set. The model was tested on the testing set, and predictions were obtained. We calculated the model's RMSE and MAE on the testing set, which was 1.4121 and 1.2154, respectively. These metrics indicate that our model was competent in making accurate predictions.

Consider a user who bought a Macbook Pro Laptop and gave it four stars. Our model predicted that this product would receive a rating of 4.13 stars. This indicates that our model correctly predicted the rating for this product.

## Application Testing

The application was tested by running various test cases with sample data with errors and evaluating the outputs against the expected results. The test cases that we used included

* data input validation,
* verifying the correct functioning of the algorithms and visualizations, and
* assessing the overall performance and accuracy of the system.

The developers conducted the testing using a combination of automated testing scripts and manual testing.

By adding a comment sections, the testing results were used to identify and fix any bugs or issues in the code, optimize the algorithms for better performance, and improve the overall user experience of the application..

## Application Files

If you're attempting to run the application on a Windows 10 machine, you will need the following:

1. Python 3 installed on your machine (you can download and install it from the official Python website: https://www.python.org/downloads/).
2. Jupyter Notebook or Jupyter Lab (you can install it using pip: *pip install jupyter*).
3. Required Python libraries: pandas, numpy, matplotlib, scikit-learn, surprise, fbprophet, plotly (you can install them using pip: *pip install pandas numpy matplotlib scikit-learn surprise fbprophet plotly*).

Alternatively, you can follow the *user guide* section below to run the environment in the cloud.

The files are organized in the following way:

1. *sales\_data.csv*: This is the dataset used in the project.
2. *sales\_predictions\_cap\_capstone.ipynb*: This is the Jupyter Notebook file containing the project code.
3. The Jupyter Notebook was run on Google Colab Jupyter Notebooks, so you can also access the notebook directly from Google Colab (https://colab.research.google.com/) by uploading the *sales\_predictions\_cap\_capstone.ipynb* file along with the *sales\_data.csv*.

## User Guide

Here is a guide to execute and use the sales prediction application:

1. Open the Google Colab website in your preferred web browser.
2. Click on the "New notebook" button to create a new notebook.
3. Upload the "sales\_predictions\_cap\_capstone.ipynb" notebook to Google Colab by clicking on the "Files" icon on the left sidebar, then clicking on the "Upload to session storage" button and selecting the file from your local machine.
4. Click on the "Runtime" menu, then "Change runtime type" and select "GPU" as the hardware accelerator for faster processing.
5. Run the notebook by clicking on "Runtime" then "Run all" to generate the sales predictions and see the results.
6. Input the sales data in the "sales\_data.csv" file in the format shown in the file. Make sure the file is uploaded to the notebook's storage in Google Colab.
7. Once the model has been trained and the predictions have been generated, you can view the results in the output cell of the notebook.
8. Use the various visualizations and graphs generated in the notebook to analyze the sales data and make informed decisions about inventory management, product pricing, and sales forecasting.

Note: The necessary libraries are already located in the first column cell in Google Colab, so there is no need to download or install any additional software or libraries once ran.

## Summation of Learning Experience

My previous four years of experience as a Software Developer gave me a strong foundation in problem-solving skills, which was essential to completing the capstone. I approached the project carefully and thoroughly because I was familiar with the development life cycle. In addition, prior experience with data manipulation, analysis, and visualization and various tools and technologies, including Python and its libraries, aided me in navigating the data and developing an appropriate solution.

In addition to my previous experience, this project necessitated continuous learning and upskilling. To complete the project, I learned to manipulate and visualize data using several new libraries and technologies, including Plotly and Pandas. I also had to learn about machine learning ideas and methods and how to use the scikit-learn library to implement them. Constant learning also helped me improve my skills and knowledge in data analysis and machine learning, which will definitely help me in future projects.

The capstone helped me realize that learning is lifelong and that even though the project was on a smaller scale, it solved a real-world problem. I also learned how to approach a problem, develop a viable solution, and validate the results. I could use everything I learned while at WGU and from past knowledge to solve and document the project effectively.