Python Case Study

Behind the Scenes Clothing Retail

Behind the Scenes Clothing Retail Scenario

Behind the Scenes Clothing Retail (BTSCR) is a clothing retail company that sells and ships a variety of products to customers around the globe.

You have joined Behind the Scenes Clothing as a Data Analyst. You have been tasked by your manager to help senior management better understand the key drivers of the business. To achieve this, you will:

- Import and explore the data extracts available to you
- Combine, clean, and transform this information so it is ready for analysis and visualization
- Analyze key metrics of the business for Stores, Products, and Orders
- Visualize insights to help drive better communication and decision making

Load & Clean Data

Chapter Introduction

Preparing Data for Analysis

Load, explore, clean, and transform data to be ready for analysis and visualization

Import CSVs

Utilize Pandas and Numpy in a Jupyter notebook to import .CSV files

Goal

Build a single, clean DataFrame that allows us to dig into store, product, and order information

Try it!

Use the chapter instructions, template Jupyter Notebook, and source data to create a DataFrame

Questions

Answer questions to verify your code is working

Review

Follow our walkthrough through each task to review your work and learn best practices

Load & Explore Data – Task Overview

We have been provided with four data extracts to load into our Python Notebook and explore. We can see that, as in most cases working with data, the .csv files are not in perfect condition for analysis - so we are going to explore and clean the data.

1. Import the four .csv files into four Pandas data frames. When importing the .csv files, keep in mind the location of the files as well as the file path.
2. Check for null values in each of the data frames. Consider how we can most efficiently accomplish this task, reducing the redundancy in our code.
3. After identifying the null values, we are going to fill the values with new data. Assume that any unique identifiers should be sequential starting from 1 and unknown Shipping information can be considered ‘Expedited’.

Load & Explore Data – Topics to Review

The questions in this section require you to load data into multiple data frames, explore the imported data, and clean any missing values. You may wish to review the following material from Python Fundamentals.

Importing External .csv

For Loops

Cleaning Missing Data

Clean Data – Task Overview

Now that we have identified and dealt with missing values, we need to continue to clean and modify our data. We will need to add some additional information, as well as combine the data frames to be in a usable format.

1. Create a new data frame: “stores_df2”. The data frame will contain StoreIDs 11-15 and the Cities of Seoul, Tokyo, Denver, Miami, and San Diego.
2. Combine the two store data frames to create “stores_df”. The combined data frame should be formatted properly.
3. Identify and drop any duplicate rows from transactions_df.
4. Identify the column datatypes in shipping_df and transactions_df. Then, modify OrderID in shipping_df to an integer and modify the Year, Month, and Day columns in transactions_df to a string.
5. Split the Name column in products_df into two columns: Type and Colour. Keep in mind how we can add the result of this function back to the data frame.
6. Replace the values of 'Gray’ with 'Grey’ in the Colour column of products_df.

Clean Data – Topics to Review

The questions in this section require you to clean data using a few different techniques. You may wish to review the following material from Python Fundamentals.

Dictionaries

Pandas Series

Pandas DataFrames

Changing Data Types

Cleaning Duplicate Data

Cleaning Incorrect Data & Exporting

Concatenating DataFrames

Transform Data – Task Overview

We need to combine our data frames and create some additional columns. This will provide the basis of our analysis and visualization.

1. Create a new data frame: “orders_df”. The Transactions data is the focus of this data frame, but we will add additional information from our other data frames. When deciding how to join the data together, consider which type of joins are the most appropriate.
    - Create “order_df” by joining transactions_df with products_df. Order_df should contain all of the transactions, with the corresponding product information.
    - Create “shipping_stores_df” by joining shipping_df with stores_df. Shipping_stores_df should only contain full rows of information from shipping and stores.
    - Create “orders_df” by joining order_df and shipping_stores_df. Orders_df should only contain orders from the stores in shipping_stores_df.
2. Create a new column, Date, from the Year, Month, and Day columns in orders_df. Keep in mind the data type of the new column.

Transform Data – Task Overview

We need to combine our data frames and create some additional columns. This will provide the basis of our analysis and visualization.

3. Create a new column, Discount Pct, indicating a 30% discount for an order of socks that are green or grey.
4. Create five new metric columns in orders_df:
    - Sale Price = Calculate the Sale Price of an order after the Discount Pct has been applied to the Price
    - Net Profit = Sale Price - Cost
    - Total Revenue = Sale Price * Quantity
    - Total Cost = Cost * Quantity
    - Total Net Profit = Total Revenue - Total Cost

Transform Data – Topics to Review

The questions in this section require you to merge data frames and add additional columns to our data. You may wish to review the following material from Python Fundamentals.

Selecting Data with a Conditional Statement

Adding & Removing Data from a DataFrame

Concatenating DataFrames

Analyze & Visualize Data

Chapter Introduction

Analysis

Apply Pandas and Numpy packages for analysis

Visualization

Apply Seaborn and Matplotlib packages for data visualization

Goal

Dig deeper into the store, product, and order information of our company

Try it!

Use the chapter instructions, template Jupyter Notebook to analyze and visualize the data

Questions

Answer questions to verify your code is providing the correct responses

Review

Follow our walkthrough through each task to review your work and learn the pitfalls and best practices

Analyze Data – Task Overview

We can begin to generate insights from orders_df to help our senior management team. Based on the requirements, these can be singular values, specific rows of data, or aggregations of data. We will be extracting information from the following metric areas - Store, Product and Order.

Store Metrics

1. The transactions for a Store 1 in 2019.
2. The average Total Net Profit across all stores.
3. The number of orders for each Store and City.
4. The Total Cost by Store in descending order.
5. The Total Revenue in 2019 by store.
6. The StoreID and City with the highest and lowest Total Revenue.

Analyze Data – Task Overview

We can begin to generate insights from orders_df to help our senior management team. Based on the requirements, these can be singular values, specific rows of data, or aggregations of data. We will be extracting information from the following metric areas - Store, Product and Order.

Product Metrics

1. The number of black t-shirts sold.
2. The product that has the highest Total Revenue to Total Cost Ratio - the product that generates the most revenue per cost.
3. The top five products by Total Revenue and the top five products by Total Net Profit. Compare the products to see if they are the same. Consider how to sort these values and how we can use logical operators to validate if the top products for Total Revenue is the same for Total Profit.

Analyze Data – Task Overview

We can begin to generate insights from orders_df to help our senior management team. Based on the requirements, these can be singular values, specific rows of data, or aggregations of data. We will be extracting information from the following metric areas - Store, Product and Order.

Order Metrics

1. The first 200 transactions from orders_df.
2. The Total Net Profit for a specific row in orders_df.
3. The number of orders from Store 7.
4. The number of orders containing multiple products.
5. The mean Quantity for each combination of Colour and City in ascending order.

Analyze Data – Topics to Review

The questions in this section require you to analyze data to for the key business metrics of Store, Product, and Orders. You may wish to review the following material from Python Fundamentals.

Manipulating Lists

Built-In Functions

Selecting Data with loc() & iloc()

Selecting Data with a Conditional Statement

Grouping Data

Visualize Data – Task Overview

With the use of visuals, we can more easily analyze key performance metrics and identify trends and outliers. When creating visuals, consider the structure of the source data, the type of visuals that best suit the data, and how we can format the plots with labels and other options.

1. Create a bar plot that shows the Quantity sold of each ProductID.
2. Create a scatterplot that plots Total Revenue and Total Cost. The scatterplot should include:
    - hue that indicates the Total Net Profit of the order.
    - a horizontal line that indicates the mean Total Revenue for comparison.
3. Create a horizontal bar plot that shows the average Discount Pct for each City.
4. Create a line plot that shows the Total Net Profit by Date.

Visualize Data – Task Overview

With the use of visuals, we can more easily analyze key performance metrics and identify trends and outliers. When creating visuals, consider the structure of the source data, the type of visuals that best suit the data, and how we can format the plots with labels and other options.

5. Create a correlation matrix of relevant columns from orders_df. Then create a heatmap to visualize the correlation matrix.
6. Create a box plot of Total Revenue by Type of product to identify any outliers.
7. Further explore the outliers identified in the box plot. Use the IQR rule to return the outlier rows from orders_df. Consider the calculation of the IQR rule and how we can determine the value of the outlier boundaries.

Visualize Data – Topics to Review

The questions in this section require you to visualize data with a variety of plots. You will also identify outliers for further analysis. You may wish to review the following material from Python Fundamentals.

Identifying Outliers

Creating Box Plots

Creating a Correlation Matrix Heatmap

Creating Line Charts

Creating Bar Plots

Creating Scatter Plots