Task-by-Task Guide

If you'd like a little more support while completing this project, explore this step-by-step resource to get additional hints and resources to help you along each task of this project.

Task 0 - Start with a Plan in Mind

Start With A Plan in Mind

Before you begin, consider taking a step back to plan your steps. Properly planning your project, or scoping, will greatly benefit you; scoping creates structure while requiring you to think through your entire project before you begin. You should start by stating the goals for your project, then gathering the data, and considering the analytical steps required. A proper project scope can be a great road map for your project, but keep in mind that some analyses you start may become dead ends which will require you to adjust your plan.

Task 1 - Load the Data

For this project, we will be working with online retail data.

In this project, we'll use a data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered online retail store.

For the specific example project, you have been given a single .xlsx file:

- **Online Retail.xlsx** - contains data about an online retail store in the UK

Hint

Open **Online Retail.xlsx** with pandas. The dataset provided has the following columns of

data:

Column name	Description
InvoiceNo	Invoice number of the transaction
StockCode	Unique code of the product
Description	Description of the product
Quantity	Quantity of the product in the transaction
InvoiceDate	Date and time of the transaction
UnitPrice	Unit price of the product
CustomerID	Unique identifier of the customer
Country	Country where the transaction occurred

Read over the pandas read_excel() documentation for a refresher on how to load and look at the dataset.

Task 2 - Explore the Data

Once you have your data, it's a good idea to get acquainted with it. You should show some summary statistics and visually examine your data. Don't forget to write out some insights that you have gained along with your analysis.

Hint

You can start to build graphs from the data by first importing Matplotlib or seaborn and then making some plots!

In this task, you might ask yourself questions such as, "Are there specific months or days of the week that have higher sales?" before analyzing data and creating visuals to showcase your findings.

More Resources:

- The National Institute of Standards and Technology's (NIST) EDA Introduction.

Task 3 - Clean and Validate the Data

After loading and exploring the data we have gained a better understanding of what is included in our dataset. A good next step may be to clean or validate the data as needed if it may help with our visualizations or analysis down the line.

Hint

A few common data-cleaning methods include:

- Dealing with missing, incorrect, or duplicate data
- Fixing structural errors
- Formatting data
- Removing or dealing with outliers

Consider exploring some common pandas techniques such as isnull(), fillna(), and drop().

Task 4 - Analyze the Data

Once the data has been cleaned and validated and appears to be in good shape, we can continue to analyze the data further.

Hint

Be sure to consider the main questions you were looking to answer when scoping out the project. A few examples of what you may want to consider analyzing and visualizing price per neighborhood or price per room type (shared room, an entire place, etc.).

Consider exploring some common Matplotlib or seaborn plots to help with your analysis and visualizations.

Task 5 - Findings and Conclusions

Finally, we can wrap up the project. You can write a conclusion about your process and any key findings.

Hint

The main components that you will want to include:

- What did you learn throughout the process?
- Are the results what you expected?
- What are the key findings and takeaways?