Project 3: Data Manipulation with Pandas

John Wesley Mathis

Dr. Anthony Choi

June ##, 2024

ECE/SSE 591, Summer 2024

Table of Contents:

[Deliverable Table 4](#__RefHeading___Toc2569_331549091)

[1. Introduction 5](#__RefHeading___Toc2573_331549091)

[2. Top Movie Data Analysis 6](#__RefHeading___Toc2577_331549091)

[3. Conclusion 7](#__RefHeading___Toc2609_331549091)

[References 8](#__RefHeading___Toc3692_331549091)

Table of Figures:

# **Deliverable Table**

The purpose of this table is to provide a complete view of the concepts covered in chapter 3 of *"Python Data Science Handbook"* (VanderPlas, 2016) and provide a general page location and/or project name for where the topic was demonstrated. This is not an exhaustive list. Some areas may be covered by multiple projects. This is meant only as a means to show specific areas where the topic is demonstrated and to ensure that every area was covered.

|  |  |
| --- | --- |
| Deliverables | Location |
| Introducing Pandas Objects |  |
| Data Indexing and Selection |  |
| Operating on Data in Pandas |  |
| Handling Missing Data |  |
| Hierarchical Indexing |  |
| Combining Datasets: Concat and Append |  |
| Combining Datasets: Merge and Join |  |
| Aggregation and Grouping |  |
| Pivot Tables |  |
| Vectorized String Operations |  |
| Working with Time Series |  |
| High-Performance Pandas: eval() and query() |  |

# 1. Introduction

Python has a rich repository of libraries that aid scientists and researchers in data analysis and manipulation. One of the most common libraries in use is Pandas, which is built on top of NumPy and provides a higher-level, and more flexible interface for data handling. While NumPy excels at efficient numerical computations with arrays, Pandas introduces data structures like Series and DataFrame that offer a more intuitive means to work with structured data.

Because of Pandas’ Series and DataFrame objects, data scientists have an indispensable tool to handle, clean and manipulate data in tabular form. These objects support a wide range of operations, from simple data aggregation and filtering, to complex time-series analysis. The library’s ability to handle missing data, merge datasets, and perform group-by operations adds significant value to Python’s data manipulation kit.

This report aims to demonstrate my proficiency in Python data manipulation techniques as covered in Chapter 3 of the “Python Data Science Handbook” by Jake VanderPlas (2016). This report attempts to illustrate the core concepts and functionalities of the Pandas library. The code presented in this report was developed using Visual Studio Code with Jupyter Notebook extensions. I will provide detailed explanations, highlighting key features and operations that make Pandas an essential tool for data analysis.

# 2. Top Movie Data Analysis

# **3. Conclusion**

This report documents my journey in mastering NumPy which involved grasping fundamental concepts such as arrays, computations using universal functions and broadcasting, incorporating comparisons and boolean logic into arrays, indexing and sorting arrays, and learning how to create structured arrays.

Through the process of coding, many errors were encountered. Due to the simplicity of the projects and exercises, all of the errors were syntax errors. My experience is mainly in C and C++ so I am used to the nuance of adding semicolons to the end of each line. Many times in Python, I found myself accidentally adding semicolons to the ends of blocks and statements, confused why my code block continually showed errors. Other examples of syntax errors I encountered include forget a quotation, forgetting a colon when defining a function or loop, or simply misspelling a variable that I named.

Engaging in mini-projects and working through the exercises has proved to be a valuable source in aiding me to take the theory of Python fundamentals and the NumPy package and put the concepts into practice and think about practical data science applications. This has allowed me to grasp the concepts that make the NumPy package a valuable tool for efficiently managing large sets of data.

# References

1. Lowe, S., Mathis, J., & Wall, N. (2019). *Humans vs. Zombies Lab.* Unpublished paper, Mercer University.
2. VanderPlas, J. (*2016*).  *Python Data Science Handbook*. O’Reilly Media. Retrieved from https://jakevdp.github.io/PythonDataScienceHandbook/index.html