Week 1: Exploring Data

Syllabus Acknowledgment

After you have read the Syllabus, please click the Syllabus acknowledgment button above. This will take you to a one-question quiz asking if you have completed the task. Once you have selected your response, click save and submit. Upon completion, you will have access to all Week 1 material.

Introduction

Contents of the Week

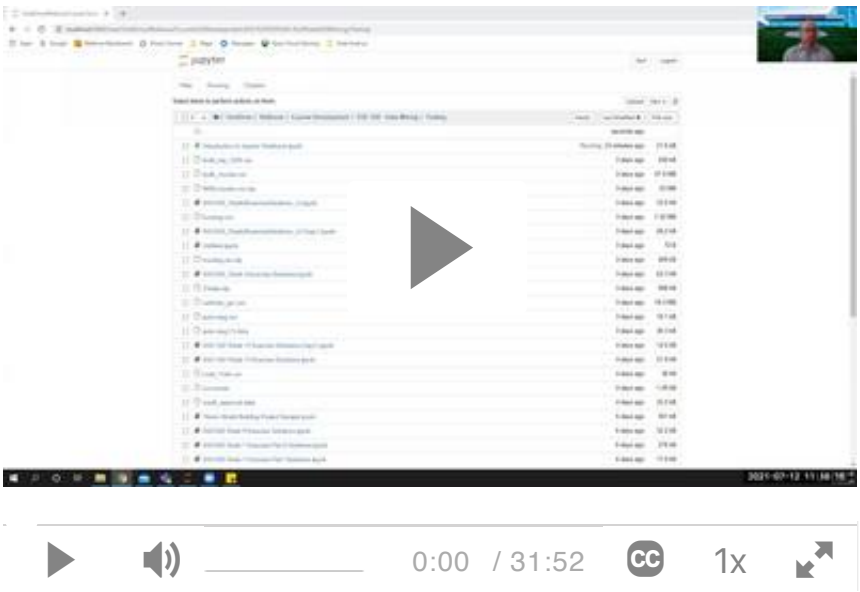
- Introduction
- Syllabus Acknowledgment
- Introduction to Jupyter Notebook Video
- Readings
- Supplemental Materials
- Additional Resources
- 1.1 Discussion/Participation
- 1.2 Exercise: Exploring a Pandas Data Frame
- Term Project Information

Topics

A. Pandas and NumPy
B. Vectors, arrays, and matrices
C. Importing data/creating data frames
D. Summarizing and exploring data
E. Filtering and aggregating data
F. Working with missing data values

Introduction to Jupyter Notebook Video

Watch this video for a primer on how to use Jupyter Notebook.



Readings

Read the following:

- Chapters 1-4 of *Machine Learning with Python Cookbook*

Supplemental Materials

All of the materials below are from external sources. Authorship and ownership are indicated within the sources themselves.

Readings | **Videos**

- What is NumPy?
- Pandas API Reference
- Pandas Working with Missing Data
- Pandas groupby
- Pandas Sort
- How to Import a CSV File into Python using Pandas
- Difference between Pandas VS NumPy

Additional Resources

- GitHub: GitHub is a great resource to share the data science/programming content you have created. You are not required to use GitHub in this course, but it is a great resource to post your assignments to be used for future coursework and to show off your work to potential employers. You are strongly encouraged to keep your term project content on GitHub.
- Anaconda: Anaconda is a tool that allows you to easily install several other data science tools including Jupyter Notebook/Lab and R Studio. It also can simplify the process of installing new libraries.
- Jupyter Notebook: Jupyter Notebook is a tool that allows you to create nice visual representations of Python and R code and outputs. This is a great tool to help you tell a story with your data, an extremely important skill. In addition to code, you can add Markdown and LaTeX to Jupyter Notebook files. You are strongly encouraged to use Jupyter Notebook for your assignments in this course.
- Kaggle: Kaggle is a resource where you participate in predictive modeling and analytics competitions and find interesting data sets.

1.1 Discussion/Participation

Here are optional topics for discussion via Teams this week. Remember, these topics aren't required, but if you are struggling to know what to post about, these can be used to initiate discussion!

1. What is NumPy? What is it used for?
2. What is Pandas and what are some of the most useful Pandas functions?
3. What is data aggregation and how is this usually done with Pandas?
4. What are some filtering methods in Pandas?
5. How does Pandas represent missing data?
6. What is the difference between NumPy and Pandas? How do they work together?
7. What is meant by data wrangling?

1.2 Exercise: Exploring a Pandas Data Frame

Download the Video Game Sales with Ratings dataset from this link: Video Game Sales with Ratings.

1. Load the dataset as a Pandas data frame.
2. Display the first ten rows of data.
3. Find the dimensions (number of rows and columns) in the data frame. What do these two numbers represent in the context of the data?
4. Find the top five games by critic score.
5. Find the number of video games in the data frame in each genre.
6. Find the first five games in the data frame on the **SNES** platform.
7. Find the five publishers with the highest total global sales. **Note**: You will need to calculate the total global sales for each publisher to do this.
8. Create a new column in the data frame that calculates the percentage of global sales from North America. Display the first five rows of the new data frame.
9. Find the number **NaN** entries (missing data values) in each column.
10. Try to calculate the median user score of all the video games. You will likely run into an error because some of the user score entries are a non-numerical string that cannot be converted to a float. Find and replace this string with NaN and then calculate the median. Then, replace all NaN entries in the user score column with the median value.

Submission Instructions

Click the title above to submit your assignment.

This exercise is due by Sunday 11:59 PM.

Submit your code, output, and answers at the link above. Comment all your code and answer any questions that are asked in the instructions. It is perfectly fine to answer a question by displaying output from your code, but make sure you are displaying the appropriate output to answer the question. I would recommend using and submitting a Jupyter Notebook, but this is not required.

View the rubric for this Assignment by clicking on the link below:

Exercise Rubric

Term Project Information

This course requires you to complete the term project. See the "Term Project" link in the menu on the left-hand side of the page.