

DATA SCIENCE FOR ECONOMISTS

ECON 220 LAB

Jafet Baca-Obando

Week 3, Introduction to Python (II) – 09/12/2025

Outline

- Setup
- Loading data
- Creating variables
- Grouped statistics
- Basic visualization

Setup check

- Local installation of VS Code, Python, Jupyter extension?
- If not:
 - Use [Google Colab](#).

The Economist's Toolkit & Setup

- Import the libraries we'll be using today:

```
import pandas as pd  
import numpy as np  
import matplotlib.pyplot as plt
```

✓ 3.2s

Python

Loading a dataset

- Download the file **flights.csv** from Canvas.
- Find its path on your laptop.
- Recall: **Use Pandas to load a CSV file using its path.**

```
path = "C:/Users/jbaca/OneDrive/Documents/2. Ph.D. in Economics/Courses/  
Semester 7 - Fall 2025/ECON 220 - Data Science for Economists - Lab/Lectures/  
Week 3/flights.csv"  
flights = pd.read_csv(path)
```

✓ 0.0s

Python

Creating new variables

- Sometimes the most interesting data isn't in the original dataset—you must **create** it.
- Let's measure the delay time of departures and arrivals.

```
flights["arr_delay"] = flights["sched_arr_time"] - flights["arr_time"]
flights["dep_delay"] = flights["sched_dep_time"] - flights["dep_time"]
flights.head(10)
```

 Open 'flights' in Data Wrangler

Python

Basic visualization: histogram

- How does the distribution of departure delay times look like?
- A **histogram** is a perfect tool for this, as it shows how frequently different values occur.

```
plt.figure()  
flights['dep_delay'].hist(bins=50, color='blue')  
plt.title('Histogram of Delay in Departure Time')  
plt.xlabel('Proportion')  
plt.ylabel('Frequency')  
plt.show()
```

Python

Comparing group means with .groupby()

- Which destination airport has the **highest average arrival delay**?
- The **.groupby()** method: Think of it like sorting your data into different buckets (e.g., one bucket for each destination airport) and then performing a calculation on each bucket.

```
destination_arr_delay = flights.groupby('dest')['arr_delay'].mean()  
sorted_destinations = destination_arr_delay.sort_values(ascending=False)  
sorted_destinations.head(10)
```

✓ 0.0s ━ Open 'sorted_destinations' in Data Wrangler

Python

Comparing group variances with .groupby()

- Which destination airport has the **lowest variance arrival delay**?

```
destinations_var = flights.groupby('dest')['arr_delay'].mean()  
sorted_vars = destinations_var.sort_values(ascending=True)  
sorted_vars.head(10)
```

✓ 0.0s ━ Open 'sorted_vars' in Data Wrangler

Python

Means of multiple variables

- You aren't limited to analyzing one variable at a time.
- By passing a **list of column names**, you can get descriptive statistics for multiple variables side-by-side.

```
for airport in ['ATL', 'ORD', 'LAX']:
    flights["arr_delay_" + airport] = np.where(flights["dest"] == airport, flights["arr_delay"], np.nan)
] ✓ 0.0s

flights[["arr_delay_ATL", "arr_delay_ORD", "arr_delay_LAX"]].describe()
] ✓ 0.0s
```

Converting Jupyter notebook to HTML

- You have to submit a Jupyter notebook together with an HTML file for the data exercises.

```
!jupyter nbconvert --to html "Week 3. Introduction to Python (II).ipynb"
```

Python

Recap

- **Created new variables** like arr_delay and dep_delay to make our dataset more useful.
- **Visualized histograms** to understand the frequency of our new variables and learned to filter data to improve our charts.
- We used .groupby() to **ask powerful comparative questions**, like finding the average delay by destination.

To-do list

- DataCamp
 - Complete “Intermediate Python” course
 - Upload certificate on Canvas: 09/13/2025
 - **Complete Data Exercise 1**
 - Upload Jupyter notebook (.ipynb file) and HTML file on **September 20**.



Attendance Check 1