---

title: "Python Workshop: Data Subsetting"

---

**# Introduction**

For this week's assignment, you will be preparing data subsets for \*\*someone else to use\*\*. It's a hands-on approach to using Pandas' filtering and selection methods. You can work directly in this document for the assignment.

**# Setup**

First, make sure this document is in your `graph\_courses\_python` folder, where your virtual environment is activated and you have pandas installed.

Then, run the following code to load pandas, and set the display options for pandas dataframes:

```{python}

import pandas as pd

pd.set\_option('display.max\_columns', None)

```

**# Get the Data**

Download the lending club dataset from GitHub here[https://github.com/the-graph-courses/pbb\_2025\_q1\_materials/blob/main/week\_05\_workshop/data/loans\_full\_schema.csv](https://github.com/the-graph-courses/pbb\_2025\_q1\_materials/blob/main/week\_05\_workshop/data/loans\_full\_schema.csv) then place it in an appropriate place in your `graph\_courses\_python` folder.

Below, load the data into a pandas dataframe with the `read\_csv()` function from pandas.

```{python}

# Your code here

```

View the dataset in VS Code with the data viewer, so you can understand what it contains.

The dataset comes from Lending Club, a platform for peer-to-peer lending. Each row pertains to a specific loan made through the platform.

You can find the data dictionary here: [https://vincentarelbundock.github.io/Rdatasets/doc/openintro/loans\_full\_schema.html](https://vincentarelbundock.github.io/Rdatasets/doc/openintro/loans\_full\_schema.html).

Now, drop the `rownames` column from the dataframe, as it's not a useful variable, but an artifact of the data source.

```{python}

# Your code here

```

**# Create and Export Data Subsets**

For each subset below:

**1)** Use Pandas selection and filtering methods to create the required extract

**2)** Export each subset to a separate CSV file in an "outputs" folder that you create.

Here is a code snippet showing how to export a dataframe to a CSV file in an `outputs` folder:

```{python}

# | eval: false

# Export to CSV

df.to\_csv('outputs/my\_dataframe.csv', index=False)

```

**## Data Subset 1**: Extract demographic and loan information

Create a subset containing the following demographic information (employment title, employment length, state, homeownership status) and loan details (loan amount, loan term, and interest rate). You may need to consult the data dictionary to identify which variables you need. Call your subset `demog\_loan\_info`.

After creating the subset, rename the `term` column to `loan\_term\_months` to make it clear what the variable measures.

Then, export the subset to a CSV file.

```{python}

# Your code here

```

**## Data Subset 2**: Extract credit evaluation variables

Create a subset with all credit-related variables. Use filtering with the `filter()` method and the `regex` argument to select variables that contain "credit" in their names. Call your subset `credit\_vars`.

Then, export the subset to a CSV file.

```{python}

# Your code here

```

**## Data Subset 3**: Extract high-income borrowers

Create a subset of only high-income borrowers (annual income > $100,000). Call your subset `high\_income`.

Then, export the subset to a CSV file.

```{python}

# Your code here

```

Note that for this you do not have to subset any columns, as you did for the previous subsets.

**## Data Subset 4**: Extract high-risk loans

Create a subset of high-risk loans, defined as those meeting ANY of these criteria:

- Interest rate is greater than 15%

- Debt-to-income ratio is greater than 30 (use the individual debt-to-income, not joint)

- Have had any delinquencies in the last 2 years

Call your subset `high\_risk`. Then, export the subset to a CSV file.

```{python}

# Your code here

```

**## Data Subset 5**: Extract "prime" borrowers

Create a subset of prime borrowers, defined as those meeting ALL of these criteria:

- Grade A or B

- Debt-to-income ratio is less than 20

- Have NOT had any delinquencies in the last 2 years

Call your subset `prime`. Then, export the subset to a CSV file.

```{python}

# Your code here

```

## Optional Challenge: Export to Excel

For an extra challenge, export all subsets to a single Excel workbook with multiple sheets using the `pd.ExcelWriter()` function.

Here is an example code snippet showing how to do this using the `tips` dataset:

```{python}

# | eval: false

import openpyxl # important to install this package

import plotly.express as px

tips = px.data.tips()

# Create some sample subsets

sunday\_tips = tips[tips['day'] == 'Sun']

saturday\_tips = tips[tips['day'] == 'Sat']

friday\_tips = tips[tips['day'] == 'Fri']

# Use the ExcelWriter to export each subset to a separate sheet

with pd.ExcelWriter('outputs/tips\_subsets.xlsx') as writer:

sunday\_tips.to\_excel(writer, sheet\_name='Sunday', index=False)

saturday\_tips.to\_excel(writer, sheet\_name='Saturday', index=False)

friday\_tips.to\_excel(writer, sheet\_name='Friday', index=False)

```

```{python}

# Your code here

```

## Submission

Submit only your Quarto file or Python script with the completed code. The exports will be evaluated by reviewing your code, so you do not need to submit the data subsets.