

Assignment Chapter 6 - Data Wrangling with Python

Instructions

1. This assignment is split into 2 parts. Part 1 will focus on data inspection. Part 2 will involve wrangling and exporting a dataset.
2. Please answer Part 1 questions in the boxes provided.
3. For Part 2, export your clean and wrangled data as an “assignment.csv” file.
4. Please submit the assignment through the TalentLabs Learning System. You will need to submit a zip file which contains this word document (with answers) and the wrangled data assignment.csv file.

Part 1 – Data Inspection with Python

For parts 1 and 2 of this assignment you will need to use the GlobalTemperatures.csv dataset.

Question 1.1:

How many rows and columns make up this dataset?

Rows: 942712
Columns: 7

Question 1.2:

How many duplicated rows are there?

0

Question 1.3:

How many columns have missing values?

6

Question 1.4:

Is there a pattern to the missing data?

No

</talentlabs>

Question 1.5:

How many unique countries and cities are there in the dataset?

Countries: 81
Cities: 375

Question 1.6:

What is the date range of the data? (Use dd.mm.yyyy date format for your answer)

01.11.1743 – 01.09.2013

Part 2 – Data Wrangling with Python

In this final part of the assignment, your task is to prepare the GlobalTemperatures.csv dataset for analysis. Carry out the actions below to wrangle this dataset. Once finished, create a zip folder submission.zip which contains your wrangled dataset and this word document with the answer to questions in Parts 1, 2 and 3 of this assignment. Make sure the wrangled dataset is named **assignment.csv**. Good luck!

Data wrangling tasks:

- Rename the dt column to Date.
- Convert the format of the Date column to datetime.
- Drop any rows which contain missing values.
- Combine the Latitude and Longitude columns into one “Location” column. Make sure the values are separated by a comma and a space, e.g.: 57.05N, 10.33E
- Drop the Latitude and Longitude columns.
- Filter the data to only contain rows with Australia and Brazil countries.
- Sort the data by date with the most recent dates first.
- Export the data as a csv file called “assignment.csv”. Do not include the index column:
`df.to_csv(“assignment.csv”, index=False).`