

Pandas Data Wrangling Project – Phase 3

Introduction

This is Phase 3 of the Pandas Data Wrangling Project. Because this phase builds on the work completed in Phases 1 and 2, it assumes that the dataset you are working with is in the state it would have been at the end of Phase 2. You have a couple of options for maintaining this continuity:

1. Continue in the same notebook: Build out Phase 3 in the same notebook you used for Phases 1 and 2, in which case you would simply continue to modify the core DataFrame.
 2. Start a new notebook: Export the final version of your DataFrame from Phase 2 to a .csv file, then import that file into a new notebook to begin Phase 3.
-

Task 1: Load the Services Datasets

Bruno now wants to analyze how specific *internet add-ons* (security, backup, device protection, and tech support) relate to the rest of our customer data.

The complication is that this services data was exported from two separate legacy systems, so we'll need to combine those exports into one DataFrame first.

Start by loading the first and second services datasets into separate DataFrames. These datasets - **Keiko Services Data 1.csv** and **Keiko Services Data 2.csv** – are available for you to download from the course resources.

Task 2: Combine the Services DataFrames

The datasets each have the same *structure* (columns and data types), so you will want to combine them “vertically”, so to speak.

Stack (combine) the two services DataFrames vertically into a single DataFrame.

Task 3: Merge Services Data into Customer Dataset

Merge the combined “services” DataFrame into our primary customer dataset, in such a way that every record in our customers_df is preserved, even if no matching services record exists.

Task 4: Remove Redundant Customer ID Column

The combined dataset that results from merging the two datasets now has *two* customer ID columns; remove one of them.

Task 5: Handle Missing Values in Services Columns

Bruno suspects that some of the customers in the original dataset may not have a matching record in the services dataset. Verify whether this is the case.

If it is, replace any Null values in the fields from the services dataset with the phrase “No internet service.”

Complete the following steps:

1. Check how many missing values were introduced in the newly-added services columns
 2. Replace Null values in the services columns with the phrase “No internet service”
-

Task 6: Load the Streaming Add-ons Dataset

We’re not done yet!

Bruno has now provided *another* distinct dataset containing customers’ streaming add-ons: **Keiko Streaming Data.csv**, which you’ll again find available for download in the course resources.

Load this dataset into a DataFrame.

Task 7: Merge Streaming Data into Customer Dataset

Just like before, you’ll want to merge this new dataset into the primary customer dataset in such a way that every record in the primary dataset is preserved, even if no matching streaming add-ons record exists.

Task 8: Remove Redundant Customer ID Column (Again)

Also like before, make sure to remove one of the redundant customer ID columns in the newly-merged dataset.

Task 9: Handle Missing Values in Streaming Columns

And finally, once again replace any Null values in the fields from the *add-ons* dataset with the phrase “No internet service.”

Task 10: Rename Columns to Snake Case

With all four datasets combined into a single DataFrame, all that’s left to do is rename those new columns to align with the “snake_case” we’ve been using thus far.

And preview the results of course!