Most Asked Questions

Lecture: Exploratory Data with Pandas
Lecture date: 02-08-2025

Top 10 Technical Queries

Query: What is the difference between `df.isnull()` and `df.isnull().sum()`? **Answer:** `df.isnull()` returns a DataFrame of the same shape as the original, with `True` for each missing value and `False` otherwise. `df.isnull().sum()` is more practical, as it sums the `True` values (which are treated as 1) for each column, providing a concise count of missing values per column.

Query: What is the difference between `.loc` and `.iloc`? **Answer:** The `.loc` accessor is label-based, meaning you use row and column labels (e.g., `df.loc[0, 'Name']`). The `.iloc` accessor is integer-based, meaning you use integer positions (0-based) for both rows and columns (e.g., `df.iloc[0, 1]`).

Query: What does the `object` datatype mean in a DataFrame? **Answer:** In Pandas, `object` is a catch-all dtype for arbitrary Python objects. In practice you'll often see it used for string (text) columns, but it can also hold mixed types—lists, dicts, custom Python objects, even numbers—so it's not limited strictly to textual data.

Query: Sir used `&` operator to check two conditions when filtering. What is the operator for OR? **Answer:** The operator for `OR` is the vertical bar `|`. For example, to filter for passengers who are either male or survived, you would use `df[(df['Sex'] == 'male') | (df['Survived'] == 1)]`.

Query: What is the purpose of the `~` (tilde) operator? Why is `~df['sex'].isin(['male'])` used instead of simply filtering for 'female'? **Answer:** The `~` is the negation operator. It inverts a boolean condition. Using `~df['sex'].isin(['male'])` is a robust way to select all rows where the 'sex' is not 'male'. This is particularly useful in datasets with more than two categorical values, and it correctly handles any other non-male entries.

Query: What is a correlation matrix and what do the values mean? **Answer:** A correlation matrix is a table that shows the pairwise correlation coefficients between multiple variables. The values range from -1 to 1. A value close to 1 indicates a strong positive linear relationship, -1 indicates a strong negative linear relationship, and 0 indicates no linear relationship.

Query: What is the difference between `.agg()` and `.transform()`? **Answer:** Both can be used after a `groupby`, but they differ in output shape and intent:

- **`.agg()`** (aggregate) applies one or more summary functions to each group and by default reduces each group to a single value per function. If you supply multiple functions, you may get a DataFrame with a hierarchical column index. In most cases, the result has one row per group (i.e., it changes the shape of the data).

- **`.transform()`** applies a function to each group but **must** return an output of the same length as the group. The result aligns **row-for-row** with the original DataFrame, so shape is preserved and you can assign the transformed values back into your original index.

Query: What is the difference between `df[...][...]` and `df.loc[... , ...]`? **Answer:** `df.loc` is the recommended method for label-based indexing as it is more explicit and prevents unexpected behavior, such as returning a copy of a slice instead of a view. `df[...][...]` is often a chain of operations that can lead to a `SettingWithCopyWarning` and is generally discouraged for modification.

Query: Why are we getting different values for `df.describe()` than the instructor? **Answer:** This can happen if your dataset is different from the instructor's. This is common if the instructor has already performed data cleaning or manipulation (e.g., dropping rows with missing values) before running the `describe()` command.

Query: I am getting a `NameError: name 'pd' is not defined`. Why? **Answer:** This error means the program does not recognize `pd`. This almost always happens because you have not imported the Pandas library at the beginning of your script or notebook with the command `import pandas as pd`.

Top 10 Non-Technical Queries

Query: Where can I find all the PDFs and other materials, such as the Google Colab notebooks, used by teachers during classes? **Answer:** All class materials, including PDFs, code notebooks, and data files, will be uploaded to the Learning Management System (LMS) after each session. You can access them in the "Resources" or "Files" section of the respective class module.

Query: Can we get a link to the dataset so we can code along with the sir? **Answer:** Yes, the dataset link is often provided in the chat during the session so you can follow along. The link is also available in the Google Colab notebook, which is shared on the LMS after the class. For this session, the link to the Titanic dataset was: `https://raw.githubusercontent.com/mwaskom/seaborn-data/master/titanic.csv`.

Query: When will attendance be taken? I was completely present, could you please check? **Answer:** Attendance is automatically tracked. If there is an issue with your attendance record, please raise a ticket on the LMS with the session details for review.

Query: Where can I submit the link to the exercises given in previous classes? **Answer:** The submission process for exercises will be communicated by the program team. Please check the instructions on the LMS for the correct method to submit your work, which may involve a dedicated submission link or a specific section on the platform.

Query: I'm new here, when was the last class? **Answer:** All session recordings and materials are uploaded to the LMS. You can catch up on any missed sessions by viewing the recordings and reviewing the provided resources.

Query: What is the answer to the exercises Sir just gave? **Answer:** The solutions to the exercises and the completed Colab notebooks are typically shared on the LMS after the session. This gives you an opportunity to attempt the problems yourself first, which is a crucial part of the learning process.

Query: What if I have an exam tomorrow, can you please not schedule class on Sunday mornings? **Answer:** Program schedules are carefully planned to accommodate a large group of students and working professionals. All live sessions are recorded and made available on the LMS, allowing you to watch anytime and anywhere.

Query: What if I get disconnected during class due to a power outage or another issue? Will my attendance still be marked? **Answer:** Attendance is logged based on your presence in the session. If you get disconnected, the system will record the time you were present. Please raise a ticket on the LMS with your details, and a program coordinator will assist you with the attendance for that session.

Query: Can you please provide the answers to the exercises? **Answer:** The solutions to the exercises and the completed Colab notebooks are typically shared on the LMS after the assignment deadline is passed.