**Discussion 3: Pandas**

1. What does the Pandas package do?

Pandas is an open source Python library that allows the handling of tabular data (i.e. explore, clean and process). The term originated from the econometrics term panel data and thus **PAN(**el)-**DA(**ta)-**S**.

At a high-level, Pandas works very much like a spreadsheet (i.e. think Microsoft Excel or Google Sheets) as you work with rows and columns. Pandas serves as one of the pillar libraries of any data science workflow as it allows you to perform processing, wrangling and munging of data. This is particularly important as many consider the data pre-processing stage to occupy as much as 80% of a data scientist’s time.

<https://towardsdatascience.com/how-to-master-pandas-for-data-science-b8ab0a9b1042>

1. How do you import the Pandas package?

Import pandas

1. What is a DataFrame?

A DataFrame is a 2-dimensional labeled data structure with columns of potentially different types. You can think of it like a spreadsheet or SQL table, or a dict of Series objects. It is generally the most commonly used pandas object

1. What is an index?

The (default) index is the corresponding row’s number starting at zero. A DataFrame’s index can be found at the far left of the DataFrame.

1. What does the .info() command do in Pandas?

Shows a succinct list of all of the columns, their datatypes, and whether they contain any null datapoints or not

1. How do you read the first 5 rows of a dataframe? The last 5?

.head() and .tail()

1. Why use Pandas versus Excel?

* Merge multiple csv/excel files
* Automating repeated data analysis work
* Applying advanced statistical functions
* Can handle more data

1. How do Pandas and SQL differ?

SQL is more efficient in querying data but it has less functions whereas in pandas, there might be lag for large volumes of data but it has more functions which enable us to manipulate data in an effective way. Python also has various data visualization libraries like matplotlib, seaborn, plotly, Altair etc which helps us to visualize hence helping us to analyze data more effectively.

Pandas is better if you intend to manipulate the data or plot it as it does it all in one place whereas in SQL we have to use Tableau for data visualization

<https://medium.com/analytics-vidhya/pandas-vs-sql-for-data-analysis-5a5cd8dc81d5>