AI Boot Camp

Transforming Data with Pandas

Module 4 Day 3

Class Objectives

By the end of class, you will be able to:

1. Create new columns in a DataFrame.
2. Use `Apply` to transform a column in Pandas.
3. Clean data with Pandas.
4. Use Pandas to answer abstract questions.

Instructor **Demonstration**

Creating New Columns

Reasons for Creating New Columns

- Mathematical operations between two columns, such as adding values together and inputting the sum of them in a new column.
- String manipulation to concatenate two string columns, perhaps combining a Name and Surname column into a Full Name column.
- Calculation of the difference between dates between two columns in order to determine time elapsed between the two.
- Data cleaning, whereby any trailing blank spaces are removed from a string column.
- Table visualization is not the only benefit of using Pandas DataFrames. Many of the functions and methods that come packaged with Pandas allow for quick and easy analysis of large datasets.

Activity:

Column Creation

In this activity, you will learn how to view numeric statistics on a DataFrame, find data about specific columns involving arithmetic operations, and create a new column using transformed data from an existing column.

Suggested Time:

10 Minutes

Activity:

Column Creation

- To convert `Membership (Days)` into `Membership (Weeks)`, the code simply takes the values stored within the initial column, divides them by seven, and then adds this edited Series into a newly created column:

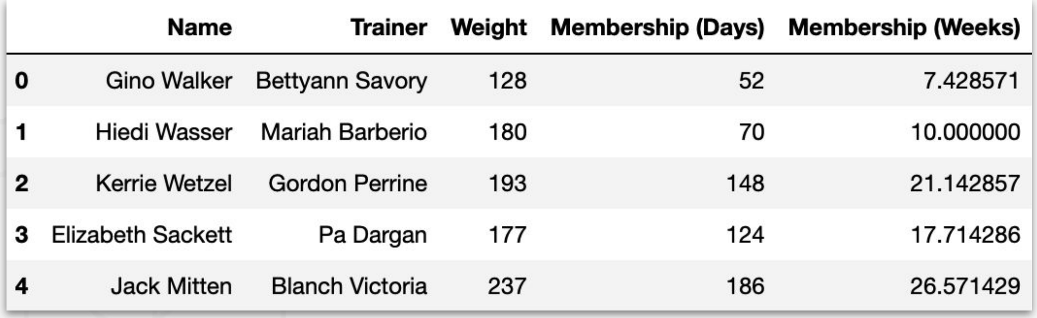
```
# Convert the membership days into weeks and then adding a column to the DataFrame

    weeks = training_df["Membership (Days)"]/7

    training_df["Membership (Weeks)"] = weeks

    training_df.head()
```

- The output is a DataFrame containing the newly created `Membership (Weeks)`:



	Name	Trainer	Weight	Membership (Days)	Membership (Weeks)
0	Gino Walker	Bettyann Savory	128	52	7.428571
1	Hiedi Wasser	Mariah Barberio	180	70	10.000000
2	Kerrie Wetzel	Gordon Perrine	193	148	21.142857
3	Elizabeth Sackett	Pa Dargan	177	124	17.714286
4	Jack Mitten	Blanch Victoria	237	186	26.571429

Suggested Time:
5 Minutes

Time’s up!

Let’s review

Instructor **Demonstration**

Apply Function

The `apply()` function

01

The `DataFrame(apply()` function is used to apply a function along an axis of the DataFrame or Series. It takes a function as an input, and applies the function to the whole DataFrame. For tabular data, the function requires specification of which axis the function should act on, where 0 represents the columns, and 1 represents the rows.

02

The `apply()` function is an essential function for data manipulation, which enables efficient versatile operations on DataFrames and Series.

The apply() function

```
#Using a Function on a DataFrame

import pandas as pd
df=pd.DataFrame({'Alpha' : [2, 4], 'Beta' :
[3, 5]})

def cube(x):
     Return x * x * x

df1=df.apply(cube)

print(df)
print(df1)
```

The output would look as follows:

```
     Alpha   Beta
0     2         3
1     4         5

     Alpha   Beta
0     8         27
1     64       125
```

The DataFrame, df, remains unchanged, while df1 is the result of the apply() function.

The `lambda` function

A lambda function is a concise way of calling a function without needing a formal function. Lambda functions are ad hoc functions that are generally used once, and allow quick operations on data elements.

```
import pandas as pd

df = pd.DataFrame({'A': [1, 2, 3]})

df['A_squared'] = df['A'].apply(lambda x: x**2)

print(df)

The output will be as follows:

      A     A_squared
0    1           1
1    2           4
2    3           9
```

Activity:

Apply Taxes

In this activity, you will practice using the `apply` function and `lambda` function to calculate the tax rate for utilities.

Suggested Time:

15 Minutes

Time’s up!

Let’s review

When dealing with massive datasets, it’s almost inevitable that we’ll encounter **duplicate rows**, **inconsistent spelling**, and **missing values**.

Activity:

Cleaning Data

In this activity, you will be required to perform data quality checks to ensure that the data is ready for analytical use. The objective of this activity is to learn how to clean data using Pandas native functions: **`count`**, **`value_counts`**, **`isnull`**, **`dropna`**, **`fillna`**, **`drop_duplicates`**, **`astype`**, and **`replace`**.

Suggested Time:

20 Minutes

Break

15 mins

Time’s up!

Let’s review

Questions?

Instructor **Demonstration**

Answering Abstract Questions

The Data Analysis Process

Data analytics follow a process involving the following steps:

01

Defining the questions
Questions may be vague initially. Refine them until they can be described in technical steps.

02

Determination of the analysis method
The type of question determines the method of analysis. Eg. Predictive models, time series analysis, or machine learning techniques.

03

Data collection
Involves the collection of 1st, 2nd, and 3rd party data, all of which combine both structured and unstructured data.

04

Data cleaning
Quality of data fed into machine learning algorithms need to be structured and of high quality.

05

Data analysis
Method of data analysis depends on the question being answered. Methods include: Descriptive, diagnostic, predictive and prescriptive analysis.

06

Sharing the results
Presentation method depends on audience. Needs to be concise and unambiguous as business decisions will be made based on data shared.

Data Analysis

01

Descriptive - What happened (Mean, Median, etc.)

02

Diagnostic - Why did this happen (Pattern Mining, etc.)

03

Predictive - What might happen, given past data (Machine learning)

04

Prescriptive - what should you do, given past records (recommendations, etc.)

Cisco Data Analysis Process

01

Stakeholder has Data / Defining the questions / Determine Analysis, Use Case

02

Gather Possibly Relevant Data

03

and Basic Analysis / Cleaning

- Can problem be solved with this data?

- Typically descriptive, sometimes small predictive model

04

Initial Report to primary stakeholder

05

Build Model / Application if applicable

06

Share Results to Further Stakeholders

Activity:

Answering Abstract Questions

Abstract question: Which utility's usage changed the most from 2013 to 2018?

Suggested Time:
30 Minutes

Time’s up!

Let’s review

Challenge

In this challenge, you’ll use order data from a wholesale business supply company to answer abstract questions about the product catalog, the clients, and the business. This will require using the full data analysis process!

Questions?

The End