Pivot Tables

This lesson will focus on how to create pivot tables in Python using pandas.

WE'LL COVER THE FOLLOWING

Pivot Tables

Pivot tables are a summary of the whole data that give us useful information. These tables reorganize the desired data in a different format. These tables can transform data from columns to rows or rows to columns, or group data by any attribute. Because of the reorganization and transformation of the data, these tables were given the name **pivot**. These tables can include statistics such as sum, mean, maximum, minimum, and many more.

Use cases

Pivot Tables are used to:

- Group data for business problems such as calculating sales by region or products
- Compare different classes of a data field such as comparing data for males and females
- Easily find out unique values in a field such as finding out different types of products for which we have data.
- Summarize complex tables

Example: Students alcohol consumption data

Let's look at pivot tables with an example. We will be using the Student

Alcohol Consumption Dataset. This dataset was made to understand how

alcohol consumption and other factors influence the grades of school students. We have grades for math class in the file *student-mat.csv*.

student-mat.csv

Now, a natural comparison that we may want to do is between schools. We can do that using the `pivot_table()` function in Pandas.

```
import pandas as pd
df = pd.read_csv('student-mat.csv')

# pivot table
table = df.pivot_table(index = 'school',aggfunc = 'mean')
print('\n\n',table)
```

We have used the `pivot_table()` function in **line 5** to calculate averages of all fields for the schools. We provide the name of the field for which we want to group our data as `index` to the function. Then we provide the aggregation function as `aggfunc=mean`. This averages all the fields.

But what if we are only interested in some of the fields?

```
import pandas as pd
df = pd.read_csv('student-mat.csv')

# choose which values to keep
table = df.pivot_table(index = 'school',values = ['G3','Walc','Dalc','studytime'],aggfunc = '
print('\n\n',table)

# see how stats vary with study time
table = df.pivot_table(index = 'school',values = ['G3'],columns = ['studytime'],aggfunc = 'me
print('\n\n',table)
```

In the case where we do not need all data fields, we provide the list of columns that we need to the function as `values` in **line 5**. We see that the school named `GP` performs better on average final grades(`G3`), their students

consume less alcohol on weekdays(`Dalc`) and weekends(`Walc`), and they study(`studytime`) more.

In **line 9**, we have provided `studytime` as `columns` to the function to see the final grades (`G3`) based on groups formed by `studytime` and `school`. By providing the column name `studytime` in a list, we tell the function to form columns according to the categories of `studytime`. This will give us the mean grade of all students, for each school, and for each category of `studytime`.

Now, what if we want to dive deeper into the data and see how different genders consume alcohol and perform in both schools?

```
import pandas as pd
df = pd.read_csv('student-mat.csv')

# pass a list as index
table = df.pivot_table(index = ['school','gender'],values = ['G3','Walc','Dalc','studytime'],
print(table)
```

We have provided two attributes that we need as `index` to the `pivot_table` function in **line 5**. We get a table that categorizes genders with schools and gives the average statistics for each category.

From the pivot table formed, we can see that, on average the males from school `MS` consume more alcohol than males from school `GP`. We can also conclude that females form school `GP` study the most, on average.

Now that we know how to create and interpret pivot tables in Python, we will focus on plotting the data in the next lesson.