Basic Data Analysis and Visualisation

Duration: 30 minutes

Learning Objectives:

- Be able to analyse, clean and visualise data to draw conclusions!

Numpy, Pandas, Matplotlib

In our very last lesson, I'm going to introduce you to some basic data analysis concepts. Please be aware that the course is not a data analysis course at its core, so this will be more of a quick intro and a call for further exploration!

We already saw how we can read in data, and using some basic Python logic we were able to make some assumptions or figure out specifics about our Amazon data set. What we are going to do now, is leverage multiple packages to help us with reading in data and analyse it!

Let's install both `pandas` and `matplotlib`!

In our terminal, enter `pip install matplotlib` and `pip install pandas` (or `pip3`)

This should install every package and dependency we need!

According to the website, pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language. This will enable us to read in files and clean our data very easily!

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.

Let's try to use them both! Again, in the following section, we will distill you the basic knowledge on how to use these, but the possibilites are endless - you can always read the documentation for both of these packages and the Python language in general to improve or customise your solutions!

Reading in from CSV using Pandas

Even though it's not always great to rely on packages, the pandas package makes this considerably easier! Let's try it with our .csv file! Create a new file called `data_visualisation.py`

As per convention, we can import our packages under certain aliases, in the case of pandas, the usual way is to use `pd`.

> Note that you might have to run Python files directly from the terminal - try `python data_visualisation.py` / `python3 data_visualisation.py`

```
import pandas as pd

bestsellers = pd.read_csv('bestsellers with categories.csv')
print(bestsellers.head())
```

This will give us an easy way to import CSV files, and the `head()` method gives us the first 5 columns. The main use for this is to confirm the structure of the data.

> Note that the first column represents each row with an index number - it helps us find specific records.

Cleaning data

Before we can get started with visualizing data, we have to learn about something important - data cleaning!

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

There are a couple of things to clean in our data set too. For example, there are some mispellings for authors, and also there is quite a lot of duplications, because certain books were bestselling ones in multiple years.

Solving some of the spelling mistakes would be a lot of work, but if you feel like it, give it a go - we will focus on the bigger issue of multiple duplicate rows!

We can see quite a lot of duplicate rows by calling the `.tail()` method - this gives us the last 5 items.

How can we get rid of the duplicates? The easiest way to do this is to use the built-in method `drop_duplicates()`. This takes in 2 arguments - we can set the column to check for duplicates (best fit would be `name`) and we can also tell it which one to keep. Last makes sense, as these are usually the latest values, which means prices might be the most accurate.

```
import pandas as pd

bestsellers = pd.read_csv('bestsellers with categories.csv')
bestsellers = bestsellers.drop_duplicates(subset='Name', keep='last')
print(bestsellers.tail())
```

Fantastic, duplicates are gone!

Now we can start figuring out some visualisations. Let's take a look at 2 different tasks!

1. Create a bar chart showing the author with the most amount of bestselling titles in the given years.
2. Create a pie chart showing the distribution between fiction and non-fiction books!

First off, we need to create our dataset that we want to inspect. In the first task, we are looking for a list of the top selling authors - we should set a limit, so let's say we need to look for the top 10! Second, we need to count how many titles are there with the same author!

We could write this with only Python, but it would take an extensive amount of time - instead, let's utilise pandas built-in tools!

```
import pandas as pd

bestsellers = pd.read_csv('bestsellers with categories.csv')
bestsellers = bestsellers.drop_duplicates(subset='Name', keep='last')

number_of_books_written = bestsellers.groupby('Author')[['Name']].count(
print(number_of_books_written)
```

Fabtastic - now we just need to create a barchart using matplotlib!

```
import pandas as pd
import matplotlib.pyplot as plt

bestsellers = pd.read_csv('bestsellers with categories.csv')
bestsellers = bestsellers.drop_duplicates(subset='Name', keep='last')

number_of_books_written = bestsellers.groupby('Author')[['Name']].count(
print(number_of_books_written)
plt.bar(number_of_books_written.Author,
        number_of_books_written.Name,
        color ='maroon',
        width = 0.4)
plt.xlabel("Authors")
plt.ylabel("Number of bestselling books")
plt.title("Number of bestselling books by author")
plt.show()
```

Last, but not least, let's create a pie chart showing the distribution between fiction and non-fiction for bestsellers! For this, we will need all genres, and count how many times they occur in our dataset!

Comment out the previous code, or create a new file for the pie chart.

```
import pandas as pd
import matplotlib.pyplot as plt

bestsellers = pd.read_csv('bestsellers with categories.csv')
bestsellers = bestsellers.drop_duplicates(subset='Name', keep='last')

number_of_books_by_genre = bestsellers.groupby('Genre')[['Name']].count(

plt.pie(number_of_books_by_genre.Name, labels=number_of_books_by_genre.Ge
plt.show()
```

Conclusion

In this lesson, you have seen how to use popular data analytics and visualisation packages to gather insight about data, and learned the fundamentals of data cleaning.

This was just a short introduction into the world of data, there's plenty more to explore - maybe another time! Until then, feel free to discover new things on your own!