Data Analysis Exercise 2

February 19, 2023

1 Data Analysis and Manipulation Exercise For Machine Learning Module Series

2 Getting Started

In this module, we will perform data analysis on the Netflix series using Python. We will be using the Pandas library to load and analyze the data and the Seaborn library for data visualization.

2.1 Dataset

We will be using the "Netflix Movies and TV Shows" dataset from Kaggle. The dataset contains information about movies and TV shows available on Netflix as of 2019. The dataset can be downloaded from the following link: https://www.kaggle.com/shivamb/netflix-shows.

2.1.1 Loading the Data

To load the data into a Pandas DataFrame, we will use the `read_csv()` method.

Write code to load the data into a DataFrame:

[1]:
```
import pandas as pd

# Load the data into a DataFrame
...
```

[1]:
```
Ellipsis
```

2.1.2 Exploring the Data

To explore the data, we can use various Pandas methods.

Write code to display the first five rows of the DataFrame:

[2]:
```
# Display the first five rows of the DataFrame
...
```

The `head()` method is used to display the first five rows of the DataFrame.

2.1.3 Data Cleaning

Before we can analyze the data, we need to clean the data.

Write code to drop the columns that are not required: e.g., ['show_id', 'description', 'date_added']

[3]:
```
# Drop the columns that are not required
...
```

The `drop()` method is used to drop the `show_id`, `description`, and `date_added` columns.

Write code to display the number of missing values in each column:

[4]:
```
# Display the number of missing values in each column
...
```

The `isnull()` method is used to find the missing values, and the `sum()` method is used to find the number of missing values in each column.

2.1.4 Data Visualization

To visualize the data, we can use various Seaborn methods.

Write code to create a bar chart of the number of TV shows and movies:

[6]:
```
import seaborn as sns
import matplotlib.pyplot as plt

# Create a bar chart of the number of TV shows and movies
...
```

[6]:
```
Ellipsis
```

The `countplot()` method from the Seaborn library is used to create the bar chart, and the `x` parameter is set to `type` to indicate that we want to create a bar chart of the `type` column.

Write code to create a scatter plot of the release year and the duration of the movies:

[7]:
```
# Create a scatter plot of the release year and the duration of the movies
...
```

The `scatterplot()` method from the Seaborn library is used to create the scatter plot. The `x` parameter is set to `release_year`, the `y` parameter is set to `duration`, and the `data` parameter is set to a subset of the data that contains only movies.

2.2 Conclusion

In this module, we performed data analysis on the Netflix series using Python. We started by loading the data into a Pandas DataFrame, and then we explored the data using various data visualization techniques. We then cleaned the data by dropping the columns that are not required and removing the missing values. Finally, we visualized the data by creating a bar chart of the number of TV shows and movies and a scatter plot of the release year and the duration of the movies.

This module covers only the basics of data analysis using Python. There are many more techniques and libraries available in Python for data analysis, and you can further

[ ]:

```

```