Data Science I

Exercise 1: Getting Started With Python & Jupyter Notebooks

Submitted by: <your names here>

Intro

Hey there, we are glad you are interested in data science!

These exercises are designed to support the lectures by letting you apply some of the ideas we cover to actual data. However, we deliberately do not saturate your schedule with our assignments, because data science is all about exploration and we want you to have time for picking up challenges on your own. So please feel free to look up interesting data sets and analyze them when you have the time!

We think discussion is the best way to learn and therefore encourage you to hand in your solutions and present them in our exercise sessions (every second Thursday, starting in Week 2). Please work in groups of 2 to 3 and upload your submissions no later than Monday 7:00 UTC in the week of the next exercise session. We chose this deadline to make sure that we have the time to actually look at your submissions before we meet for discussion. We will mark your submissions and make the number of achieved bonus points available online in Stud.IP, but please note that we will not give you back the corrected submissions: If you have questions on how your grading, please visit the Wednesday session and ask questions! 😉

We hope you have fun and learn something!

Part 1: Intro to Python & Jupyter

In the first exercise sheet, you will get an introduction to Python and Jupyter Notebooks, how to load and manipulate data, and how to a simple plot to visualize your findings. But first things first.

1.) Install Jupyter Notebook on your computer if it is not already installed. You can download and install Anaconda, which includes Jupyter Notebook, from the following website: https://www.anaconda.com/products/individual

2.) You are probably looking at a PDF document right now, but we will be working with notebook files from here on. So please open the `.ipynb` file we provided in Jupyter Notebook and continue working from there.

3.) In today's exercise, we will be looking at a data set on dinosaurs, because dinosaurs are awesome. So please create a new Markdown cell and write a short statement on your favorite dinosaur.
Example:

```
# The Best Dinosaur in the World
We are Jannik and Wolle and we could not yet agree on what the
most awesome dinosaur was. But we agreed that, going by its name,
 [dreadnoughtus](https://en.wikipedia.org/wiki/Dreadnoughtus) must
have been pretty awesome!
```

4.) Now lets change the name of the file you are working in to make it unique for your group. To generate a file name, execute the two code cells below by marking them and either clicking the run button at the top or by pressing `ctrl-enter`.
You should get the following as output:

```
exercise_1_schroeder_wingerath
```

In [ ]:
```python
# provide last names of your group members
last_names = ["Schröder", "Wingerath"]
exercise_sheet = 1
```

In [ ]:
```python
# generate file name for your submission
last_names = [name.lower() for name in last_names]
last_names.sort()
last_name_str = '_'.join(last_names)

print(f"exercise_{exercise_sheet}_{last_name_str}")
```

5.) Obviously, this does not reflect your group members, so please update the code cell with the variable `last_names` according to the names of your group members and execute the code for name generation again. Then rename your notebook file

using the name you just generated. While you are at it, please also update the "Submitted by" entry at the top of the file with the names of your group members.

6.) To avoid losing all your data (like it may or may not have happened to us the first time we created this assignment sheet 😅), click on "File" in the menu bar and select "Save and Checkpoint". Then, close Jupyter Notebook by clicking on the "Logout" / "Quit" button in the upper-right corner and reopen it. If everything worked fine, go to the next task. If everything is gone, please start all over 😉

7.) We'd like you to submit both PDF and notebook files for your exercises. To enable PDF export, please install the `nbconvert` and `pyppeteer` libraries. You can do so via commandline, but you can also do this from within this notebook like so:

```
!pip install <library>
```

In [ ]:
```
# install nbconvert
<your code here>
```

In [ ]:
```
# install pyppeteer
<your code here>
```

8.) Assuming everything went right, you are now able to generate a PDF file from your notebook using the following command:

```
!jupyter nbconvert --to webpdf --allow-chromium-download
<file-name>.ipynb
```

Note: This command may take a while when running for the first time. Also, we recommend using a chromium-based browser.

In [ ]:
```
# generate PDF:
<your code here>
```

Part 2: Loading and Working With Data [/ 30]

In the next part, you will learn how to load and work with data by using a small data set on dinosaurs.

1.) Lets start by importing a few libraries that we are going to use:

In [ ]:
```
# Import libraries
import pandas as pd
```

2.) Now download our dino data set from
https://raw.githubusercontent.com/kjanjua26/jurassic-park/main/data/data.csv and put

it in the same directory as your notebook file. Next, load the data into a data frame using the read_csv function from the pandas library and inspect the structure of the data set by playing around with the data frame (see this cheatsheet for more info on working with pandas).

```python
In [ ]: # read the data from the CSV file you downloaded
        dinosaurs = <your code here>
```

```python
In [ ]: # take a quick look at the data (example functions: head, tail, sample)
        <your code here>
```

```python
In [ ]: # look at some statistical information on the different attributes using th
        <your code here>
```

3.) Now that you have an idea of how the data looks, try to answer a few questions:

```python
In [ ]: # Is your favorite dinosaur captured in the data set?
        # (Hint: You can use the query function to filter the data by name.)
        <your code here>
```

```python
In [ ]: # How many dinosaurs are recorded for each individual diet?
        # (Hint: You can use the groupby function)
        <your code here>
```

```python
In [ ]: # Try sorting the data set by length to get the longest dinosaur.
        <your code here>
```

4.) You may have noticed that the length attribute is not a numerical type, so that the `length` attribute has only limited use as a sorting key or in computations. Create a new attribute `length_in_meters` to fix this problem. When you have done that, answer a few more questions.

```python
In [ ]: # Add a new column length_in_meters with data type float by parsing the met
        <your code here>
```

```python
In [ ]: # Now try finding the dinosaur with the biggest record at length again, thi
        <your code here>
```

```python
In [ ]: # Now compute the maximum recorded length per diet, similar to the groupby
        <your code here>
```

5.) Finally, visualize how many dinosaurs we have in the data set and how big they are:

```python
In [ ]: # Create a histogram plot with length_in_meters on the x-axis and the numbe
        <your code here>
```

Congratulations! You have done an analysis on dinosaurs!🦖

Part 3: Find & Analyze a Data Set! | / 40 |

Now that you can access data sets and do basic analyses, look up a data set yourself and find out something that might be of interest to somebody. Depending on the kind of data that you are looking at, you probably won't do the exact same analysis like we did for the dinosaur data set. But perhaps it serves you as an inspiration on how you could approach this analysis task.

If you don't know where to look for interesting data sets, you can try out the following sources:

- Kaggle
- GovData
- Statistisches Bundesamt (Federal Statistical Office of Germany)

Note: Targeting a data set that helps you in your project report might save you some work later 😉

1.) Please describe what data set you chose and what kind of information it contains. (Please provide a brief description and the code for loading your data.)

2.) Ask and answer at least 3 questions about the data set.

3.) Please create a plot that illustrates something you learned from the data analysis.

4.) Save your notebook and submit it (as both notebook and PDF file) along with the data that you downloaded.
Please make sure to **include the execution output in your submission**!