Project Documentation: Covid-19 Data Analysis

Ignatius S. Condro Atmawan

Introduction

In the Covid-19 Data Analysis, the data procedure is like this:

1. Select variables of interest, from the data
2. Clean the data
3. Calculate descriptive statistics
4. Visualize the data with diagrams or other images

The data source comes from
https://www.kaggle.com/datasets/hendratno/covid19-indonesia/data
However, the data has noises after it is downloaded.

In order to process the data, the class CovidData is developed. This class is put in the script "covid_data_process.py" and the main file "covid19_main.py" is using the methods from the class to process Covid-19 data.

CovidData Class

Attributes

The attributes for this class are

- file_name:str, as the name of the file to be analyzed
- current_directory: str,
- output_directory: str
- df: pandas data frame
- columns: list, containing the list of columns of the data frame
- df_oi_dict: dict, containing the dictionary of the data frame of interest
- df_oi_statistics: dict, containing the dictionary of the descriptive statistics

Methods

Methods for selecting variables

- select_column(self, list_of_columns: list)
- get_dataframe_of_interest_based_on_string()

Methods for cleaning data

- omit_empty_data(self)
- clean_numeric_data(self, list_of_columns: list)

Methods for calculating statistics

- calculate_descriptive_statistics(self, list_of_columns: list)
- save_descriptive_statistics(self)

Methods for data visualization

- plot_from_saved_dict(self, x_col: str, y_col: str, dict_key: str)

Other methods

- read_csv_data(self, file_name: str)
- save_csv_data(self, file_name: str)
- assign_directory(self)

Data Processing

The data processing procedures is

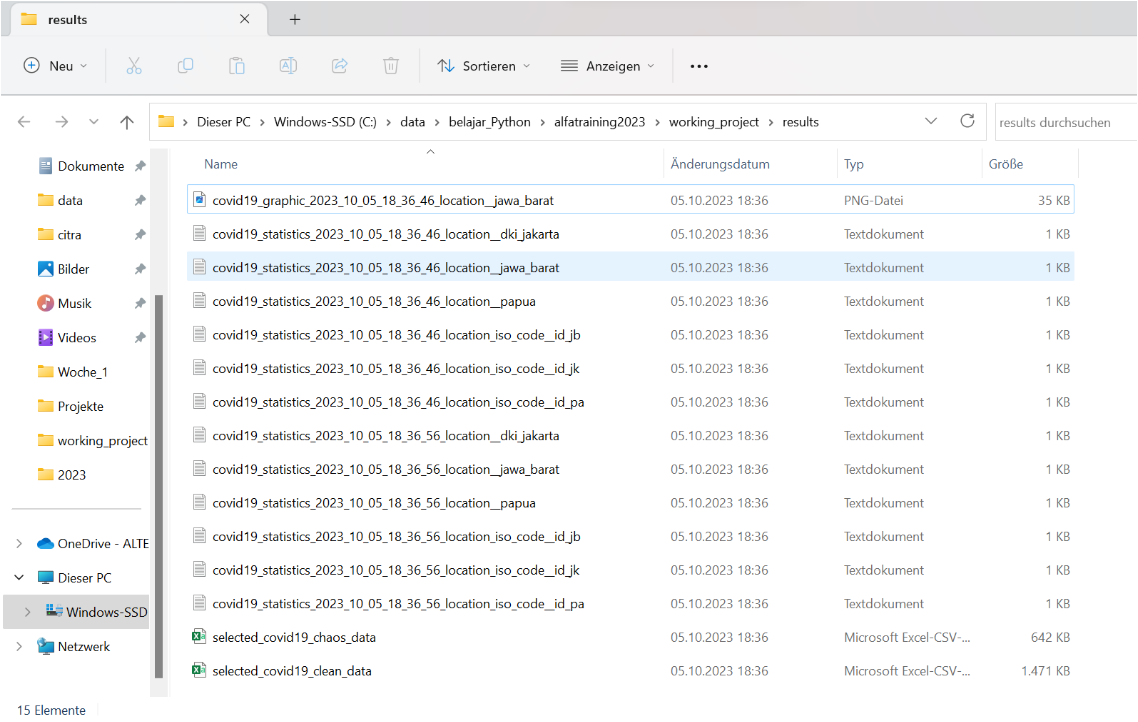
1. Open two CSV files, with clean data as control and chaos data to test the data filtering/cleaning functions
   - "covid_19_indonesia_time_series_all.csv"
   - "Ignatius_covid_19_indonesia_time_series_all_chaos.csv"
2. Select the variables of interest:
   - for clean data: "New Cases", "New Deaths", "New Recovered", "New Active Cases"
   - for chaos data: "New Cases.1", "New Deaths.1", "New Recovered.1", "New Active Cases.1"
   - for both data: "Date", "Location ISO Code", "Location",
3. Cleaning the data using "clean_numeric_data" and "omit_empty_data"
4. Select the location of interest, e.g. "Jawa Barat", "DKI Jakarta", "Papua"
5. Calculate statistics based on our selection
6. Visualize the results

Results

The results will be saved with date and time:

- The descriptive statistics are saved as text files
- The data visualization is saved in figures of PNG format,

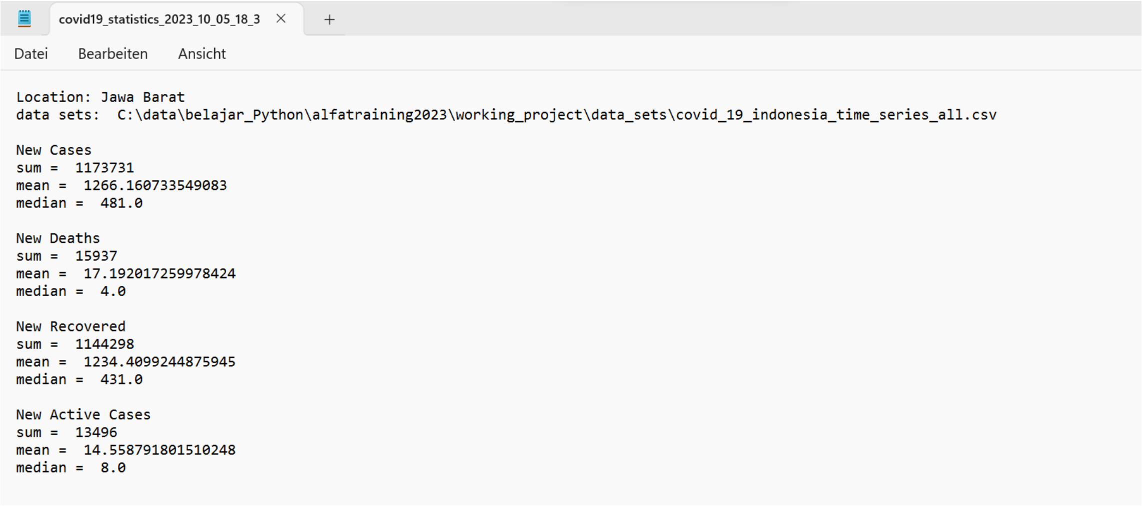
They are like this



Name	Änderungsdatum	Typ	Größe
covid19_graphic_2023_10_05_18_36_46_location__jawa_barat	05.10.2023 18:36	PNG-Datei	35 KB
covid19_statistics_2023_10_05_18_36_46_location__dki jakarta	05.10.2023 18:36	Textdokument	1 KB
covid19_statistics_2023_10_05_18_36_46_location__jawa_barat	05.10.2023 18:36	Textdokument	1 KB
covid19_statistics_2023_10_05_18_36_46_location__papua	05.10.2023 18:36	Textdokument	1 KB
covid19_statistics_2023_10_05_18_36_46_location_iso_code__id_jb	05.10.2023 18:36	Textdokument	1 KB
covid19_statistics_2023_10_05_18_36_46_location_iso_code__id_jk	05.10.2023 18:36	Textdokument	1 KB
covid19_statistics_2023_10_05_18_36_46_location_iso_code__id_pa	05.10.2023 18:36	Textdokument	1 KB
covid19_statistics_2023_10_05_18_36_56_location__dki jakarta	05.10.2023 18:36	Textdokument	1 KB
covid19_statistics_2023_10_05_18_36_56_location__jawa_barat	05.10.2023 18:36	Textdokument	1 KB
covid19_statistics_2023_10_05_18_36_56_location__papua	05.10.2023 18:36	Textdokument	1 KB
covid19_statistics_2023_10_05_18_36_56_location_iso_code__id_jb	05.10.2023 18:36	Textdokument	1 KB
covid19_statistics_2023_10_05_18_36_56_location_iso_code__id_jk	05.10.2023 18:36	Textdokument	1 KB
covid19_statistics_2023_10_05_18_36_56_location_iso_code__id_pa	05.10.2023 18:36	Textdokument	1 KB
selected_covid19_chaos_data	05.10.2023 18:36	Microsoft Excel-CSV-...	642 KB
selected_covid19_clean_data	05.10.2023 18:36	Microsoft Excel-CSV-...	1.471 KB

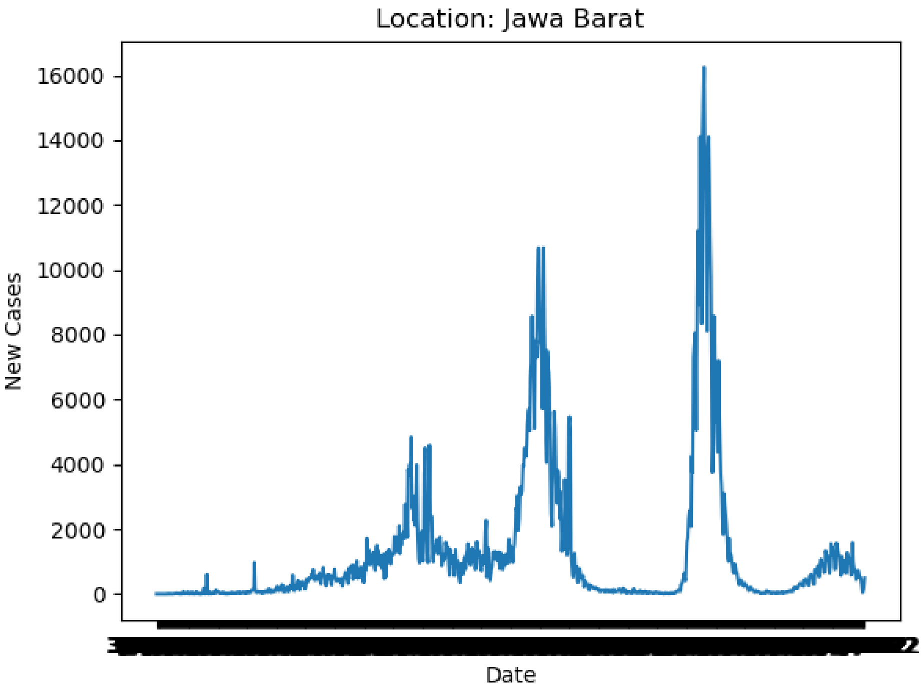
15 Elemente

The result of descriptive statistics is like this

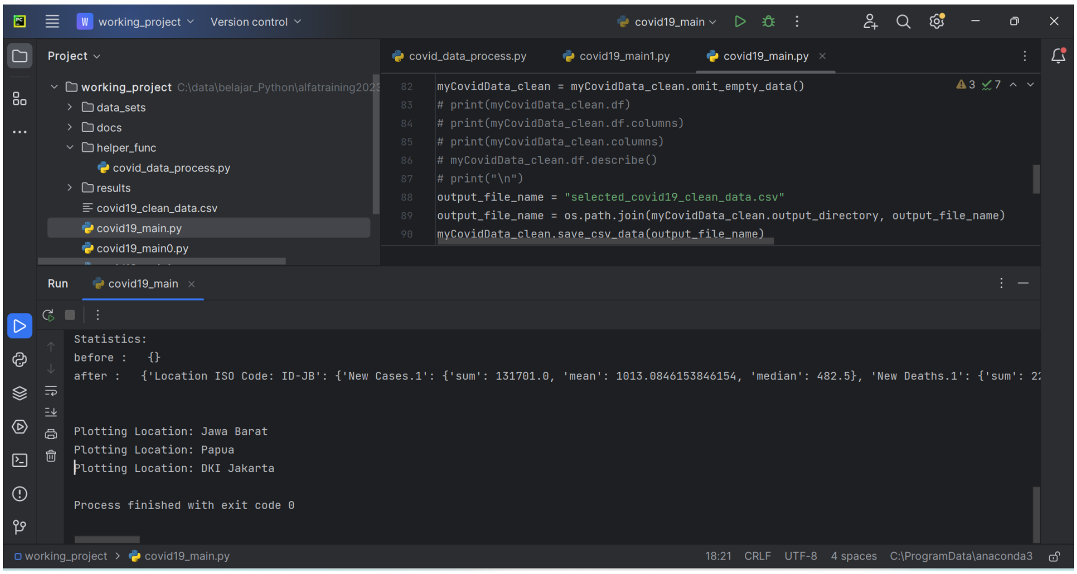


```
Location: Jawa Barat
data sets:  C:\data\belajar_Python\alfatraining2023\working_project\data_sets\covid_19_indonesia_time_series_all.csv

New Cases
sum =  1173731
mean =  1266.160733549083
median =  481.0

New Deaths
sum =  15937
mean =  17.192017259978424
median =  4.0

New Recovered
sum =  1144298
mean =  1234.4099244875945
median =  431.0

New Active Cases
sum =  13496
mean =  14.558791801510248
median =  8.0
```

The visualization result is like this



The Pycharm logs will look like this



```python
myCovidData_clean = myCovidData_clean.omit_empty_data()
# print(myCovidData_clean.df)
# print(myCovidData_clean.df.columns)
# print(myCovidData_clean.columns)
# myCovidData_clean.df.describe()
# print("\n")
output_file_name = "selected_covid19_clean_data.csv"
output_file_name = os.path.join(myCovidData_clean.output_directory, output_file_name)
myCovidData_clean.save_csv_data(output_file_name)
```

```
Statistics:
before :   {}
after :   {'Location ISO Code: ID-JB': {'New Cases.1': {'sum': 131701.0, 'mean': 1013.0846153846154, 'median': 482.5}, 'New Deaths.1': {'sum': 2


Plotting Location: Jawa Barat
Plotting Location: Papua
Plotting Location: DKI Jakarta

Process finished with exit code 0
```

Discussion

The data processing can clean the data, calculate statistics, and visualize the data. However the data cleaning process only deals with numerical data (integer) and it throws away a lot of data samples. If the string data can be cleaned, more data may be saved and not thrown away. The data visualization only shows minimum requirements, but the axis are not clearly visible.