ONLINE MASTERS IN **DATA SCIENCE**

DSC 207 - PYTHON FOR DATA SCIENCE

COVID-19 SPREAD ANALYSIS

İLKAY ALTINTAŞ, PH.D.

UC San Diego

COMPUTER SCIENCE & ENGINEERING

HALICIOĞLU DATA SCIENCE INSTITUTE

By the end of this video, you will be able to:

- Generate valuable statistics about a Covid-19 spread dataset
- Summarize basics steps in data science applied to a sample dataset
- Visualize predictive algorithms along with understanding process of generating insight

Covid-19 Case Study - Acquiring Data

Data on COVID-19 (coronavirus) by *Our World in Data*

📂 **Download our complete COVID-19 dataset : CSV | XLSX | JSON**

Our complete COVID-19 dataset is a collection of the COVID-19 data maintained by *Our World in Data*. We will update it daily throughout the duration of the COVID-19 pandemic (more information on our updating process and schedule here). It includes the following data:

Metrics	Source	Updated	Countries
Vaccinations	Official data collated by the Our World in Data team	Every weekday	218
Tests & positivity	Official data collated by the Our World in Data team	Weekly	178
Hospital & ICU	Official data collated by the Our World in Data team	Daily	47
Confirmed cases	JHU CSSE COVID-19 Data	Daily	216
Confirmed deaths	JHU CSSE COVID-19 Data	Daily	216
Reproduction rate	Arroyo-Marioli F, Bullano F, Kucinskas S, Rondón-Moreno C	Daily	191
Policy responses	Oxford COVID-19 Government Response Tracker	Daily	187
Other variables of interest	International organizations (UN, World Bank, OECD, IHME...)	Fixed	241

Source: https://github.com/owid/covid-19-data/tree/master/public/data

- Initial step in the Data Science Process - fetch/collect the appropriate data

ACQUIRE

- Import raw dataset into your analytics platform

PREPARE

- Explore & Visualize
- Perform Data Cleaning

ANALYZE

- Feature Selection
- Model Selection
- Analyze the results

REPORT

- Present your findings

ACT

- Use them

Covid-19 Case Study - *Prepare*

```
[4] df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 177844 entries, 0 to 177843
Data columns (total 16 columns):
 #   Column                   Non-Null Count   Dtype
---  ------                   --------------   -----
 0   iso_code                 177844 non-null  object
 1   continent                167420 non-null  object
 2   location                 177844 non-null  object
 3   date                     177844 non-null  object
 4   total_cases              171501 non-null  float64
 5   new_cases                171279 non-null  float64
 6   new_cases_smoothed       170110 non-null  float64
 7   total_cases_per_million  170707 non-null  float64
 8   population               176733 non-null  float64
 9   population_density       158415 non-null  float64
 10  median_age               146933 non-null  float64
 11  gdp_per_capita           146418 non-null  float64
 12  female_smokers           111363 non-null  float64
 13  male_smokers             109836 non-null  float64
 14  handwashing_facilities   72165 non-null   float64
 15  human_development_index  143233 non-null  float64
dtypes: float64(12), object(4)
memory usage: 21.7+ MB
```

- Visualizing various features and their types with pandas functions

Covid-19 Case Study – *Prepare: Clean*

```
df['new_cases'] = df['new_cases'].fillna(0)
df['total_cases'] = df['total_cases'].fillna(0)
df['new_cases'] = df['new_cases'].clip(lower=0)
df['new_cases_smoothed'] = df['new_cases_smoothed'].clip(lower=0)
df['total_cases'] = df['total_cases'].clip(lower=0)
```

- Why do we need to clean data?
  i. Handle missing entries
  ii. Filter unwanted outliers
  iii. NULLs

- How do we clean Covid-19 data?
  i. Negative value for “new_cases” doesn’t make sense
  ii. Clip negative values to zero

Covid-19 Case Study - *Visualizing Data*

Fig. 1

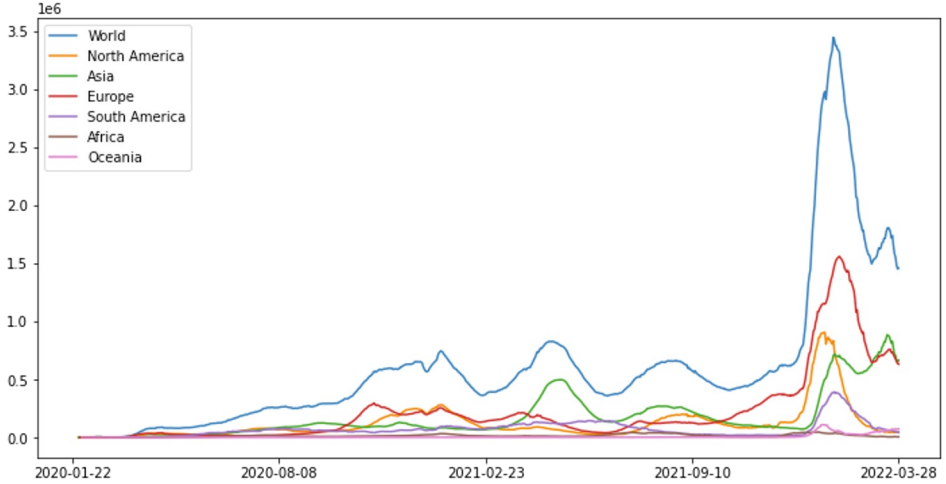
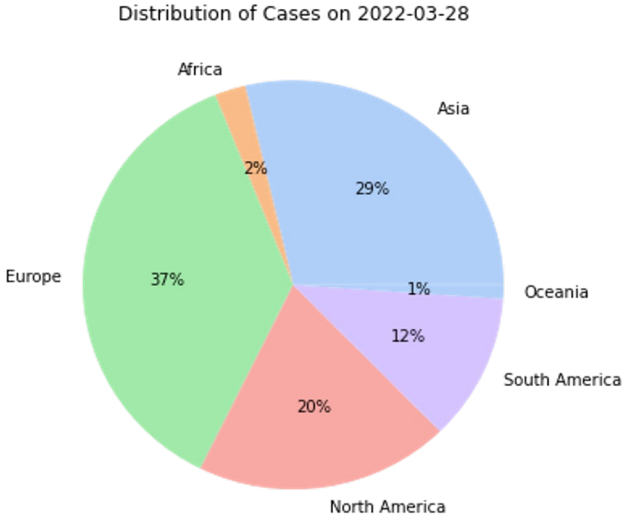
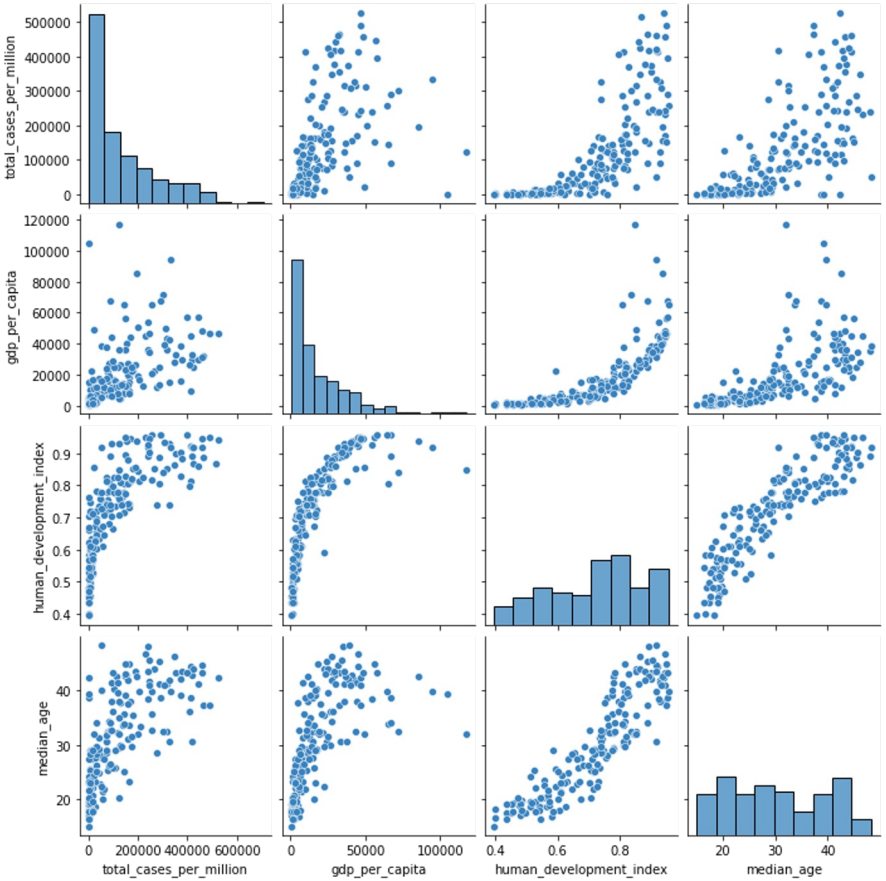


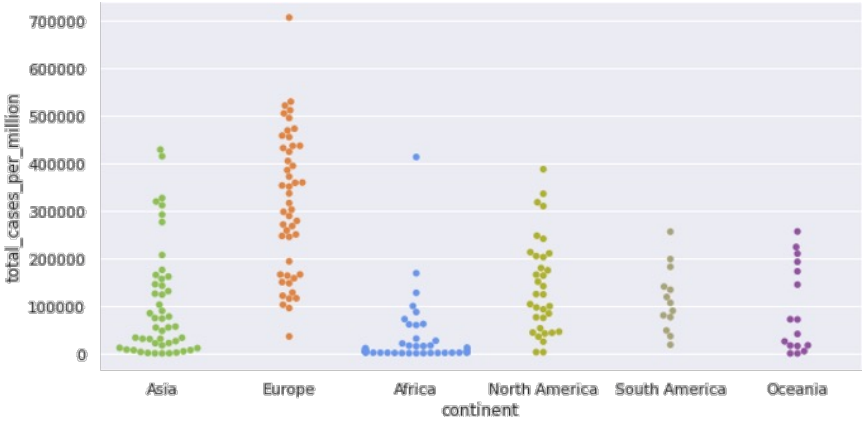
Fig. 2

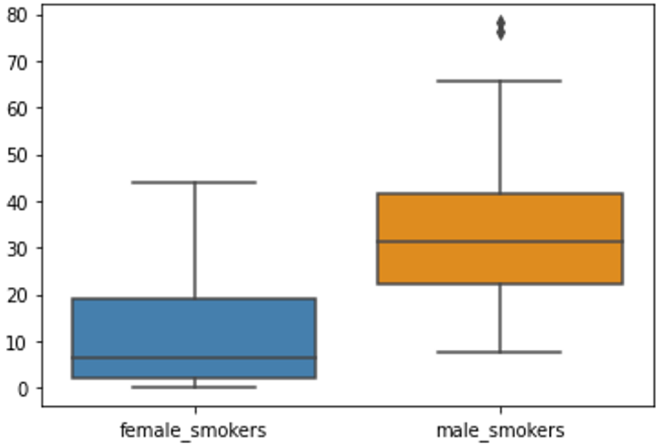


- We have global time series data on Covid-19 cases throughout the world. We create visualizations on this, to better understand data distribution.
- Figure 1 is a line plot, demonstrating comparative rise/fall of Covid-19 cases across the globe during the given time.
- Figure 2 is a distribution of Covid-19 cases by date.
- Contd.

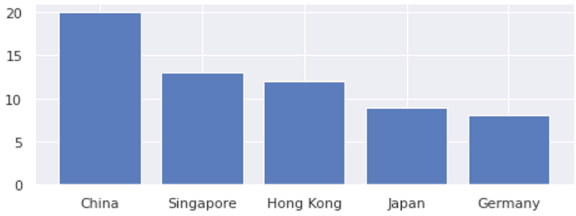
Covid-19 Case Study – *More Visualization*



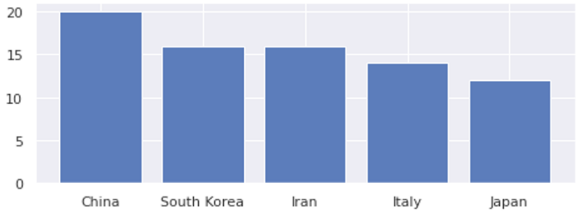




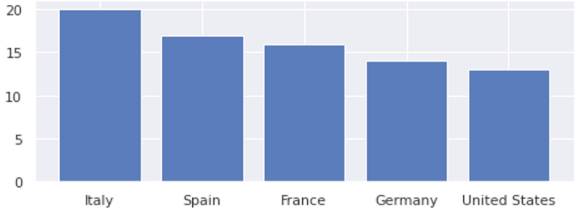
Covid-19 Case Study – *Spread Analysis*



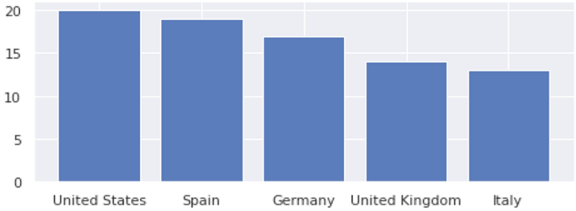
Week 3-5



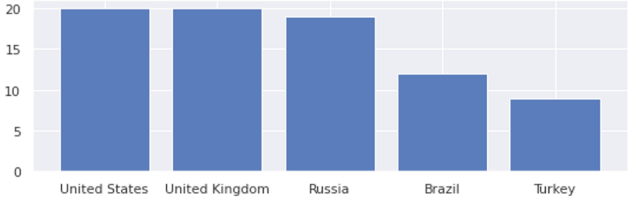
Week 5-8



Week 8-11



Week 11-13



Week 13-17

- **How did Covid-19 spread in initial few weeks?**
  i. Frequency of occurrence of a country in new_cases ranking is considered as a sustained measure of spread in the region.

Is it possible to improve this measure?

Covid-19 Case Study – *Predictive Models: Linear Regression*

- For a given set of attributes (e.g., median_age, population_density, etc.), how do we predict number of cases?
- Given demographic information about a region, can we predict rise of Covid-19 cases?

Covid-19 Case Study - *More Models: Decision Tree Classifier*

- For a given set of attributes (e.g., median_age, population_density, etc.), predict the continent from where the data point is coming.
- This model helps in identifying unlabeled/mislabeled data for future data points.