Home assignment 1.

1. Compute statistics (over countries) of total number of **confirmed** cases on the 10’th day since 50 confirmed cases:

a. Mean

b. Median

c. Max

d. Min

*If " March, 13" is the first day with >= 50 confirmed cases, then "March, 22" is the 10th day.*

2. Compute statistics (over countries) of total number of deaths on last available day:

a. Mean

b. Median

c. Max

d. Min

3. What was the average number of **new** cases for the last 10 days in Germany?

4. Compute case fatality rate (death to total cases ratio) for the last available day in countries with more than 10 000 reported cases (in total).

a. What is the biggest case fatality rate? Write **percentage** rounded to 2 decimal places.

b. What is the lowest? Write **percentage** rounded to 2 decimal places.

c. Plot a scatter plot: Total number of cases vs Case fatality rate, color points according to the country.

5. On which weekday most cases were reported in France on average? On which weekday least cases were reported in Italy on average?

Write all numbers rounded to 2 decimal places.

Suggestions

1. Use Aggregation.

2. Use Aggregation twice, second time with empty groupby. Use the fact that the total number of deaths in the country is maximal on the last available day.

3. Use Differentiate; use Filter Examples.

4. Use Join; Use Generate Attribute; use Filter Examples.

5. Use Generate Attribute, you can get weekday using the following function: *date\_str\_custom(Date, "E")*, where Date is the name of the column with the date; use Pivot.

You are provided with the starter process to download and preprocess latest available COVID data:

1.<https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv>

2.<https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_global.csv>

3.<https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_recovered_global.csv>

However for the purpose of this assignment you should use data from April 18, 2020 **(files attached on wiki page).**

You should submit your answers to the google form (link on wiki page).

Home assignment 2.

Analyze Titanic data.

I. Start with basic EDA (Exploratory data analysis):

1. Compute **average** `Age` of passengers and **number of passengers** who survived and not survived grouped by `Sex` and `Passenger Class` (24 numbers);
2. What can you say about survivors based on the resulting table (open question), e.g. what is the surviving ratio for females in First class compared to the Second and Third?

*This answer is limited to 150 words.*

1. What is the **average** number of males and females on all boats (rounded to the closest integer)?

*Do not forget to filter out all `?` in `Life Boat` attribute.*

II. Proceed with feature generation.

1. **Drop the column** `Life Boat`.
2. Generate new attribute `Family size`: sum up `No of Parents or Children on Board` and `No of Siblings or Spouses on Board` and add 1 (for passenger himself, thanks to @pianovanastya). What is the **average** family size? In **which class** did the biggest family travel?

*Do not drop original attributes.*

1. It seems that `Passenger Fare` is total among all passengers with the same `Ticket Number`: **create new attribute** `Single passenger fare`. For every passenger you need to compute the number of passengers with the same `Ticket Number` and then use this number as a divisor for `Passenger Fare`.

*Do not drop the original attribute.*

1. Impute missing values: for numerical attributes use **averaging** over three groups: `Passenger Class`, `Sex`, `Embarkation Port`; for every numerical attribute create separate column that contain 1 for imputed value and 0 for originally presented.

*This step is mainly for practicing your groupby/join skills. In real tasks this kind of imputation is relatively rare.*

1. Pre-process categorical attributes: For every categorical attribute create a separate column that contains 1 for a missing value and 0 for originally presented. One-hot encode categorical attributes with less than 20 unique values, drop other categorical attributes; drop original attributes.
2. Set the role of the `Survived` attribute to `label`.

III. Finish by building a classification model using preprocessed data

1. Compute classification accuracy on a train-test setup:
   1. Create a Cross Validation block, fix the random\_state parameter to 2020.
   2. Use a decision tree with `maximal depth` = 7; uncheck `apply pruning` box; leave all other parameters by default.
   3. Use accuracy as a performance metric
2. Analyze the resulting confusion matrix, which error is larger: Type I or Type II?
3. Provide a short analysis of the results, based on your answers III.2-III.3. E.g. What are the splitting features of the first 3 levels of the best tree (up to 7 attributes)? Do these results coincide with your intuition? You may include some misclassified examples along with explanations why they were misclassified.

*This answer is limited to 250 words.*

Suggestions.

I.1 Use the Aggregation block.

I.3 Use the Aggregation block twice.

II.2 Use Generate attribute block.

II.3 Use Aggregation block with `count` aggregation function; use Join block

II.4 Use example block from Seminar 3.

II.5 Use One Hot encoding block

II.6 Use Set Role block

III.1 Use example from Seminar 3.

**Upload your solution** (.py, .r, .ipynb, .rpm). Answers without an uploaded solution file will not be graded!

You should submit your answers to the google form (link on wiki page).