# TW3: Exploratory Data Analysis on COVID-19 Datasets

**Part 2:**

Submit a summary of your learning to Canvas. Your document should include:

* Full names of your team members who work on the assignment.
  + Xiaomei Xie
  + Lili Hao
* URL links to the notebook of each student on GitHub repo.
  + <https://github.com/xiaomeiX/TW3-dataAnalysis_on_COVID19>
  + <https://github.com/lhaoSeattleu/TW3-dataAnalysis_on_COVID19>
* A summary of what you learned from the teamwork assignment.

1. Select attributes (columns) you are interested (either US only or all data).

The dataset chosen in this teamwork is from:

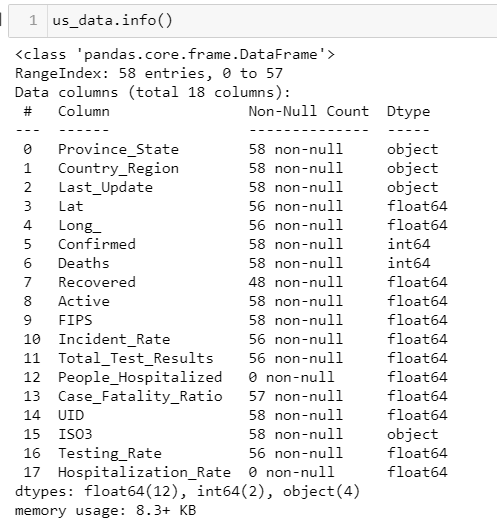
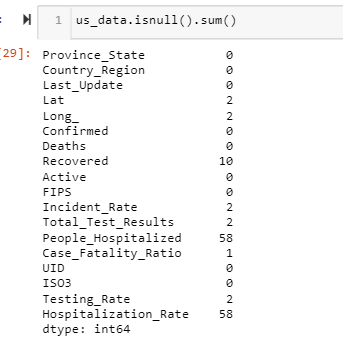
[COVID-19/12-31-2020.csv at master · CSSEGISandData/COVID-19 · GitHub](https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_daily_reports_us/12-31-2020.csv)

This dataset includes attributes:

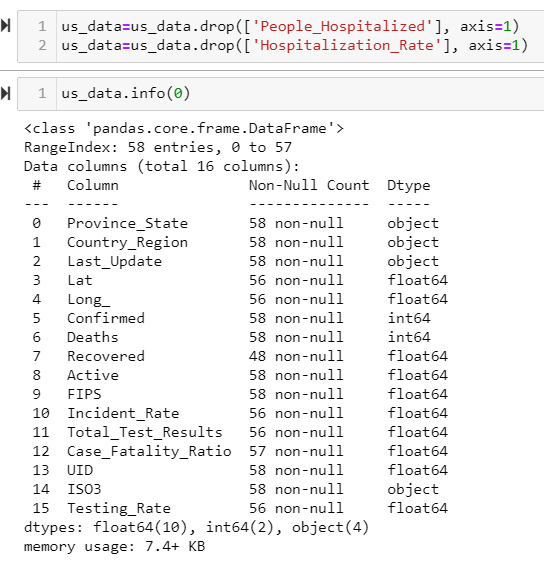
* **Province\_State** - The name of the State within the USA.
* **Country\_Region** - The name of the Country (US).
* **Last\_Update** - The most recent date the file was pushed.
* **Lat** - Latitude.
* **Long\_** - Longitude.
* **Confirmed** - Aggregated case count for the state.
* **Deaths** - Aggregated death toll for the state.
* **Recovered** - Aggregated Recovered case count for the state.
* **Active** - Aggregated confirmed cases that have not been resolved (Active cases = total cases - total recovered - total deaths).
* **FIPS** - Federal Information Processing Standards code that uniquely identifies counties within the USA.
* **Incident\_Rate** - cases per 100,000 persons.
* **Total\_Test\_Results** - Total number of people who have been tested.
* **People\_Hospitalized** - Total number of people hospitalized. (Nullified on Aug 31, see [Issue #3083](https://github.com/CSSEGISandData/COVID-19/issues/3083))
* **Case\_Fatality\_Ratio** - Number recorded deaths \* 100/ Number confirmed cases.
* **UID** - Unique Identifier for each row entry.
* **ISO3** - Officialy assigned country code identifiers.
* **Testing\_Rate** - Total test results per 100,000 persons. The "total test results" are equal to "Total test results (Positive + Negative)" from [COVID Tracking Project](https://covidtracking.com/).
* **Hospitalization\_Rate** - US Hospitalization Rate (%): = Total number hospitalized / Number cases. The "Total number hospitalized" is the "Hospitalized – Cumulative" count from [COVID Tracking Project](https://covidtracking.com/). The "hospitalization rate" and "Total number hospitalized" is only presented for those states which provide cumulative hospital data. (Nullified on Aug 31, see [Issue #3083](https://github.com/CSSEGISandData/COVID-19/issues/3083))

1. (optional) If you find any missing data or/and invalid data in the selected attributes you would like to work on, apply tools to handle these data.

This dataset has missing data:

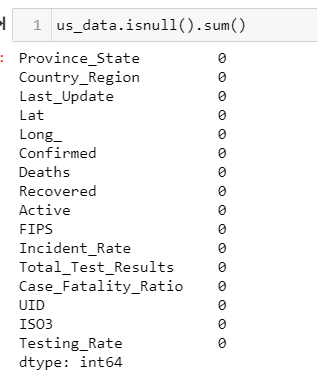
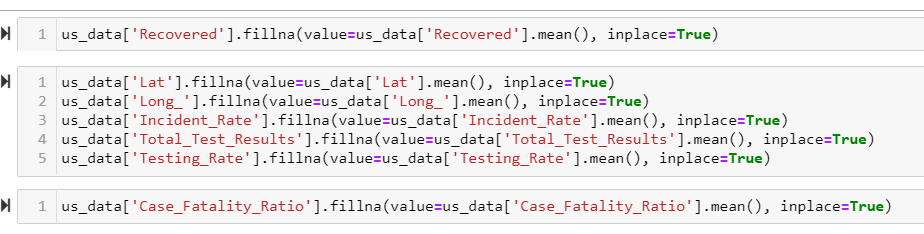


We decide to drop columns “People\_Hospitalized” and “Hospitalization\_Rate” because they have only NaN values.



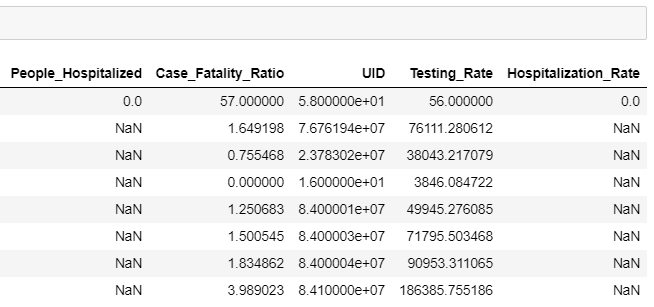
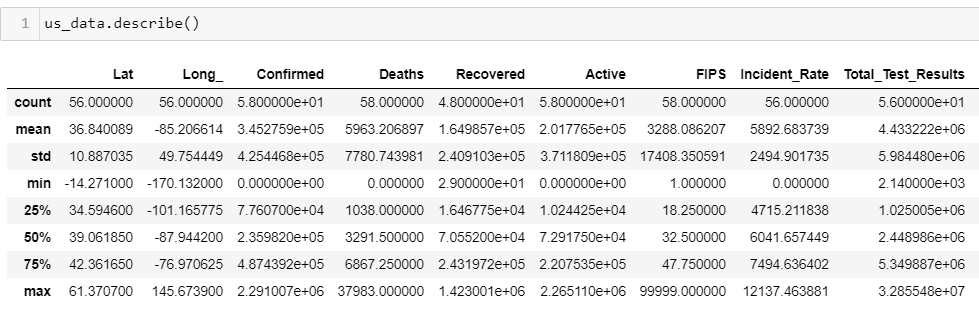
The rest of the columns have a few missing values, and we can replace them with mean values.

* The column “Recovered” has 10 missing values.
* The columns Lat, Long\_, Incident\_Rate, Total\_Test\_Results, and Testing\_Rate has only 2 missing values.
* The columns Case\_Fatality\_Ratio has only 1 missing values.



1. Conduct exploratory data analysis on the selected data and visualize the data. This may include the following but not limited to:

(a) The mean, median and standard deviation



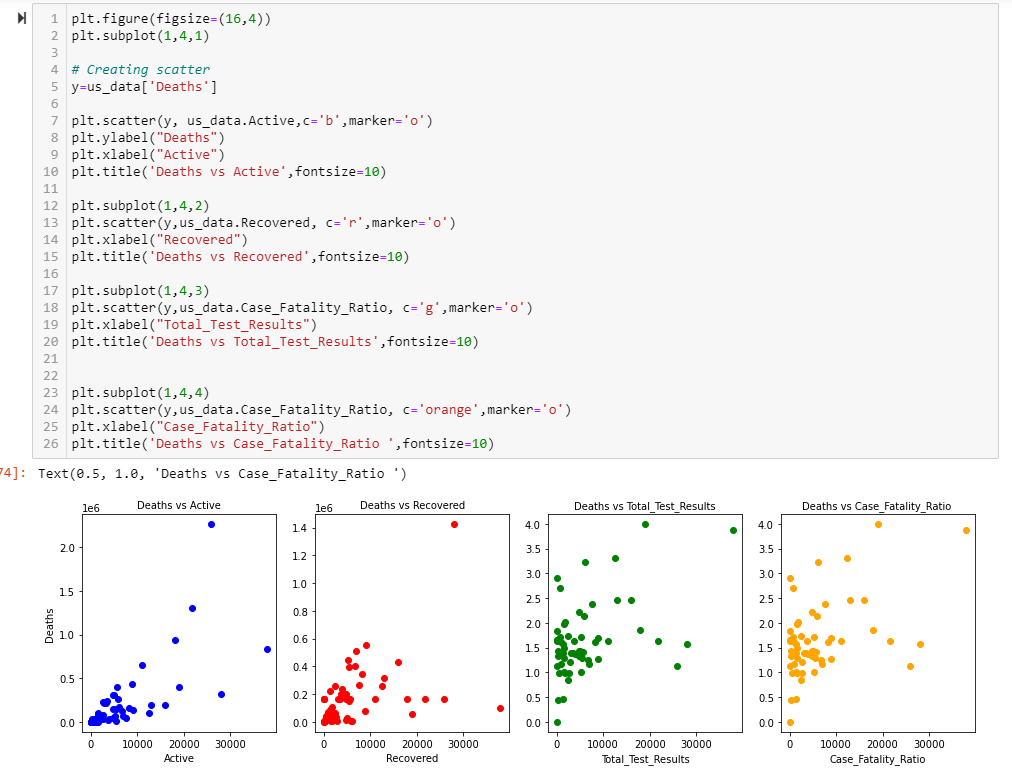
(b) Draw boxplots



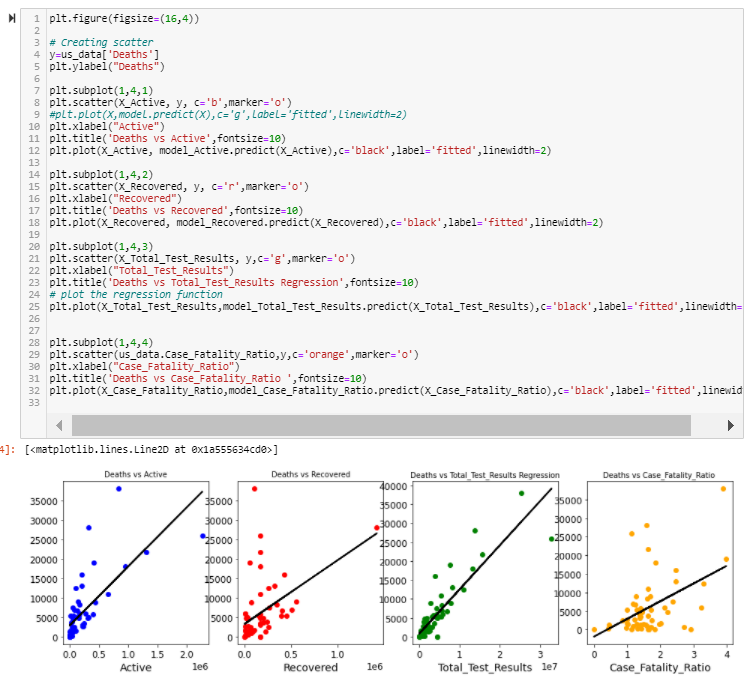
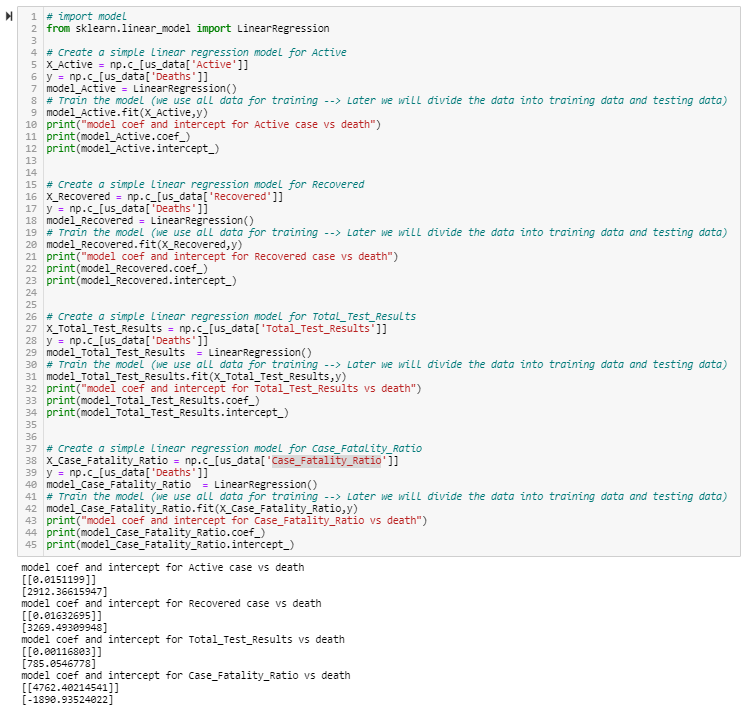
(c) Draw histograms



(d) Draw a scatter plot of the data



(e) Display simple linear regression function with a scatter plot --> You will see the data does not follow a simple linear function.

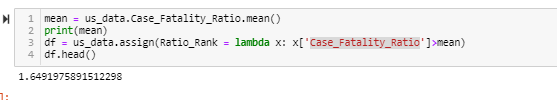


(f) Create a logistic regression function instead of a simple linear regression. See the web resources below:

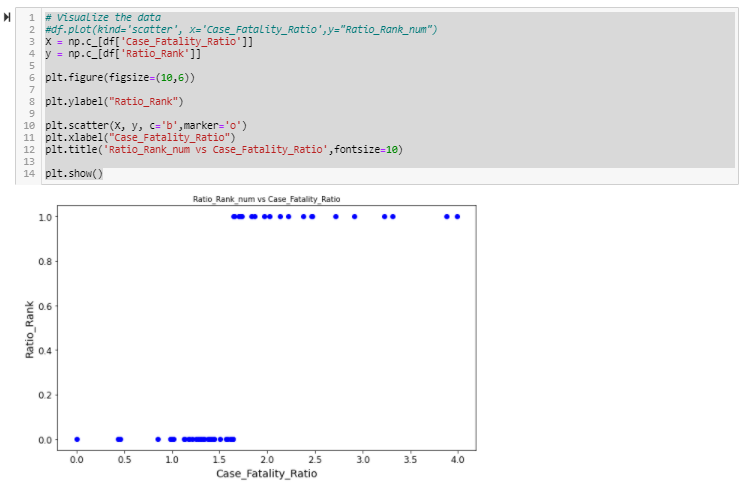
from sklearn.linear\_model import LogisticRegression

This time draw a scatter plot of the data together with the logistic regression function.

We set the Ratio\_Rank as 1 when Case\_Fatality\_Ratio is greater than its mean, and set Ratio\_Rank to 0 when Case\_Fatality\_Ratio is less than the mean.



The scatter plot shows the relationship between Ratio\_Rank and Case\_Fatality\_Ratio:



Lastly, we create a logistic regression model using “Recovered”, “Total\_Test\_Results” and “Deaths” and build the logistic regression model to predict Ratio\_Rank.

