CIS400 Assignment 1 report

Huahao Shang
hushang@syr.edu
SUID: hushang

1.

(a) Three files are successfully read using read_csv and read_xls and three files don't have NA data.

(b) There are 3140 rows after the merge

(c) There are 3093 rows and 23 columns

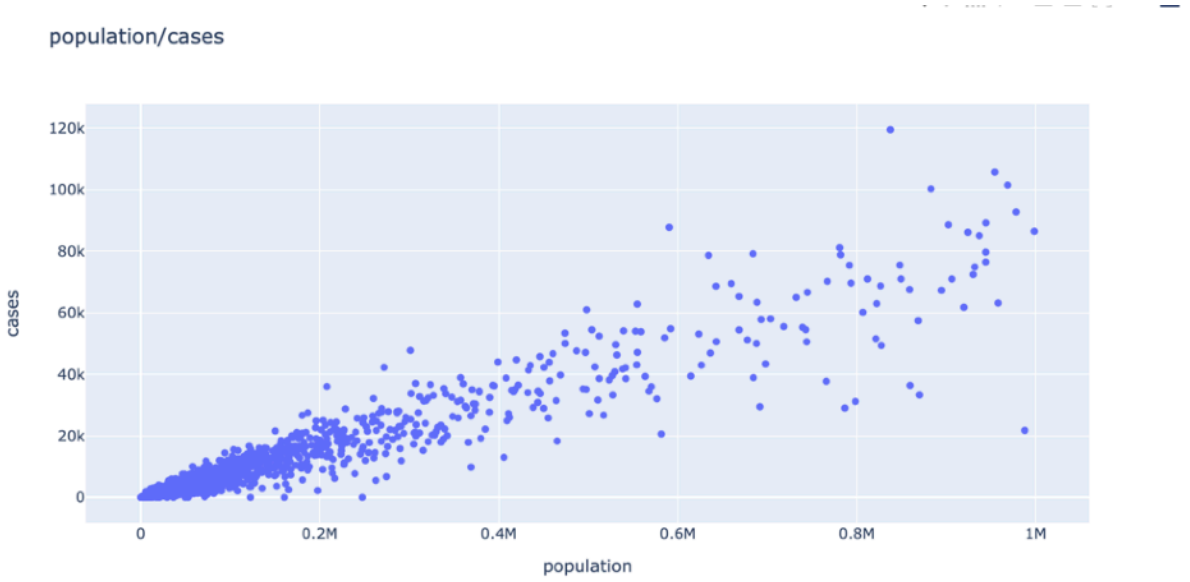
2.
(a)
fips:discrete
contry-x:categorical
state_x: categorical
state_code_x:categorical
male: discrete
famale: discrete
median_age: discrete
population: discrete
female_percentage:discrete
lat_x:continuous
long_x:continuous
Areaname: categorical
LND010200D: continuous
county_y:categorical
state_y: categorical
lat_y: continuous
Long_y: continuous
date: categorical
cases: discrete
stste_code_y: categotical
deaths: discrete
populatin-density: discrete
case-ratio: discrete

(b)
The histogram are not very surprised.
'male','female','median_age','population','female_percentage','cases','deaths','population_density','case_ratio' are obtained. The data and the graph are in a expected looking.

(c)
For heat map, I expected the cases, case_ratio and death are related to the population and population density. And Land area, longitude and latitude are not correlated to any of the cases and death numbers.

The real heat map turns out as expected.

(d).



The trend is that cases will grow as population getting larger.

3
(a).
Train data length: 2072
Test data length: 1021

(b)
R^2 is 0.90
Beta 1 is 0.07 This indicates it will be a smooth slope
Beta 0 is 248.21

(c)

MAE is 1418.24

The error seems fine, from the shown pred_cases and real cases, they are not have to much difference.

(d)

```
print(mean_absolute_error)
```

Out[13]:

	pred_cases	cases
448	6,880.76	7273
667	1,545.67	1287
3139	808.99	617
230	479.63	97
3093	3,672.06	3834

```
Mean absolute error is
1418.240890654271
```

```
# define function to import wig libraries
```