ISC 4241 – Activity #1

Team Number: Team 7

Team Captain: Sriharsha Aitharaju

Team Members: Daniel Rodriguez

Fernando Sosa

Kzzy Centeno

Robert Law

Activity on

PART I (20 Points) Programming

Problem 1.1 (8 Points) Read the EXCLE file "COVID_08312020.csv"

Problem 1.2 (8 Points) Produce a scatter plot using "TotalCases" and "TotalDeaths" and impose a loess line on the top of the data.

Problem 1.3 (8 Points) Produce a scatter plot using "ToTCases_1M" and "TotDeath_MPOP" and impose a loess line on the top of the data.

Problem 1.4 (8 Points) Produce a table with the following summary statistic including minimum, mean, median, variance, standard deviation, maximum, and skewness for the following five variables "ToTCases_1M", "TotDeath_MPOP", "TotalCases", "TotalDeaths", and "TotalTested". **(Note: Display only three decimal place)**

Problem 1.5 (8 Points) Obtain both the Spearman correlation and the Pearson correlation between the following variables "ToTCases_1M", "TotDeath_MPOP", "TotalCases", "TotalDeaths", and "TotalTested".

PART II (10 Points) Fill in Blank

1. Suppose that $\{x_1, x_2, x_3, \cdots, x_n\}$ be a set of data and $x_{(15)} = 5$, $x_{(16)} = 7$, and $x_{(17)} = 8$, the median of this data set is 6 if n = 30 and the median of this data is 7 if n = 31.
2. Suppose that $\{x_1, x_2, x_3, \cdots, x_n\}$ be a set of data and $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 100$ and n = 26, the sample variance of this data set is 4.
3. The points at distances 1.5 times of IQR (Inner Quartile Range) from each hinge mark the inner fences of the data set.
4. Tom is interested in finding out the salary of students graduated from UCF in the past three years. He collected data from one thousand students graduated from UCF. The data he collected including their major, their graduation year, their gender, their salary, and their GPA. Tom's study is a Classification with 1000 observations and 5 predictors.
5. Jennifer has a data set to perform an analysis; however, you cannot find any response variable in this set of data. The analysis performed by Jennifer should be a (supervised learning / non-supervised learning).
6. Steve fit a model on a set of data. After perform data exploration analysis, he decided to assume that the data come from normal population and the relationship between the response variable and a set of predictors should be linear. The analysis perform by Steve should be (parametric analysis / nonparametric analysis / cluster analysis).
7. Lori likes to know the relationship between a given predictor and the response variable. Lori is interested in (prediction / inference) problem.

ISC 4241 - Activity 1, Part 1

Problem 1.1

In [5]:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import statistics
```

In [6]:

```
covid = pd.read_csv('COVID_08312020.csv')
```

In [7]:

```
covid.head(10)
```

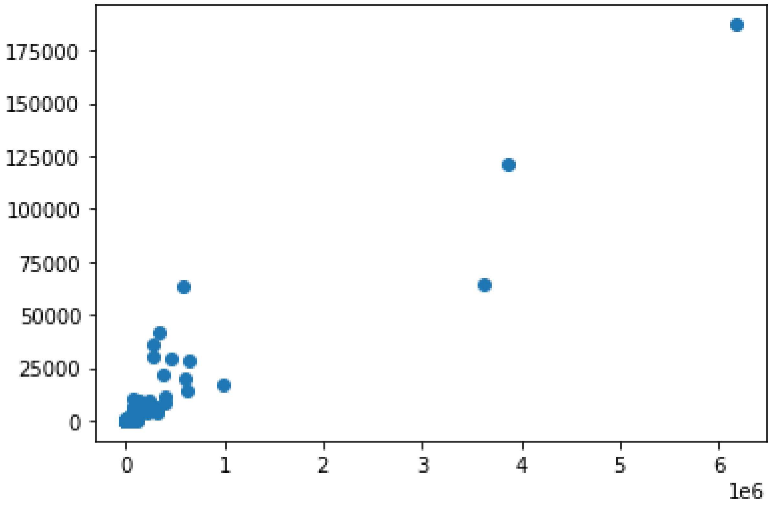
Out[7]:

	Country	Total Cases	Total Deaths	TOTCases_1M	TOTDeath_!M	TotalTested
0	Afghanistan	38162	1402	977	36	102598
1	Albania	9380	280	3260	97	57618
2	Angola	2624	107	79	3	64747
3	Argentina	408426	8457	9023	187	1242269
4	Armenia	43750	877	14760	296	205450
5	Australia	25670	611	1005	24	6167592
6	Austria	27166	733	3013	81	1172092
7	Azerbaijan	36309	531	3576	52	917027
8	Bahrain	51574	189	30150	110	1100729
9	Bangladesh	310822	4248	1884	26	1537749

Problem 1.2

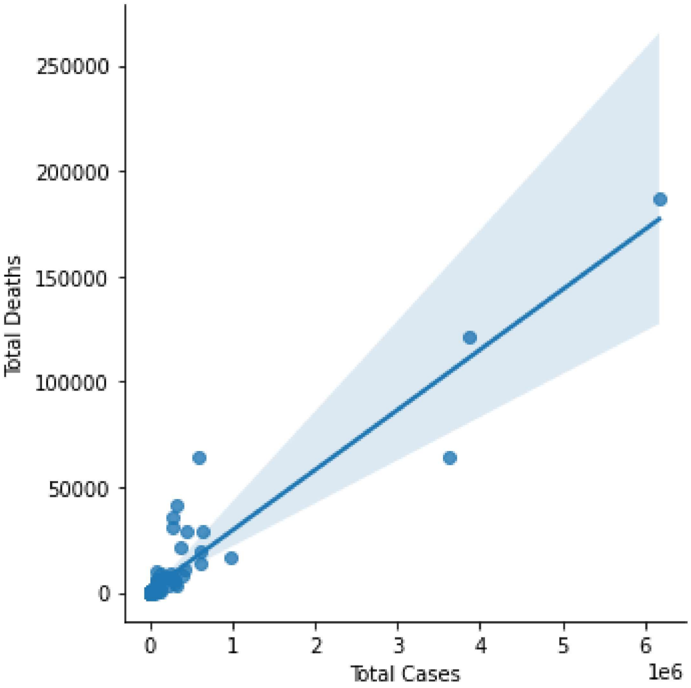
In [8]:

```
plt.scatter(covid['Total Cases'], covid['Total Deaths'])
plt.show()
sns.lmplot(x='Total Cases', y='Total Deaths', data=covid)
```



Out[8]:

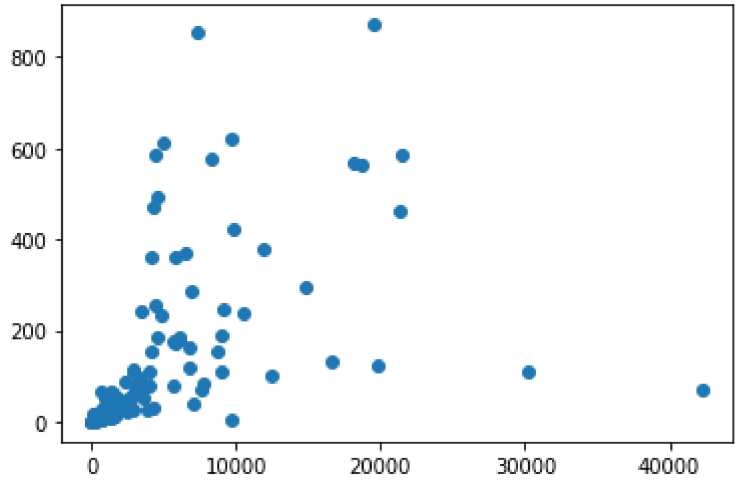
```
<seaborn.axisgrid.FacetGrid at 0x7f60eb8de410>
```



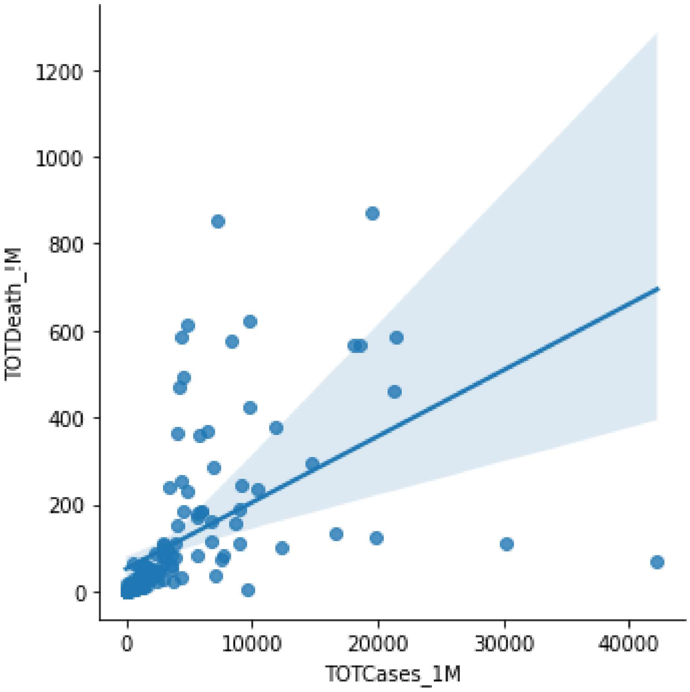
Problem 1.3

In [9]:

```
plt.scatter(covid['TOTCases_1M'], covid['TOTDeath_!M'])
plt.show()
sns.lmplot(x='TOTCases_1M', y='TOTDeath_!M', data=covid)
```



```
Out[9]: <seaborn.axisgrid.FacetGrid at 0x7f60eb86add0>
```



Problem 1.4

```
In [37]: from numpy import minimum
         mean = [covid['Total Cases'].mean(), covid['Total Deaths'].mean(), covid['TOTCases_1M'
         mean = [round(item,3) for item in mean]

         median = [covid['Total Cases'].median(), covid['Total Deaths'].median(), covid['TOTCas
         min1 = [covid['Total Cases'].min() , covid['Total Deaths'].min(), covid['TOTCases_1M']
         max1 = [max(covid['Total Cases']) , max(covid['Total Deaths']), max(covid['TOTCases_1M

         std = [statistics.stdev(covid['Total Cases']), statistics.stdev(covid['Total Deaths'])
         std = [round(item,3) for item in std]

         var = [statistics.variance(covid['Total Cases']), statistics.variance(covid['Total Dea
         var = [round(item,3) for item in var]
```

```
skew = [covid['Total Cases'].skew(skipna=True), covid['Total Deaths'].skew(skipna=True
skew = [round(item,3) for item in skew]
```

In [38]:

```
data = [mean, median, min1, max1, std, var, skew]
data
df = pd.DataFrame({
            'mean' : mean,
             'median': median,
              'minimum': min1,
              'maximum': max1,
              'variance': var,
              'standard deviation': std,
              'skewness': skew

}, index= ['Total Cases', 'Total Deaths', 'TOTCases_1M', 'TOTDeath_!M', 'TotalTested']

df
```

Out[38]:

	mean	median	minimum	maximum	variance	standard deviation	skewness
Total Cases	181486.137	24367.0	355	6173236	4.767454e+11	6.904675e+05	6.836
Total Deaths	6091.115	411.0	1	187224	4.393447e+08	2.096055e+04	6.343
TOTCases_1M	4177.388	1789.0	11	42230	3.814673e+07	6.176304e+03	3.066
TOTDeath_!M	115.187	34.0	0	871	3.215569e+04	1.793200e+02	2.229
TotalTested	3141261.633	404944.0	120	90410000	1.280726e+14	1.131691e+07	6.328

Note for Output: Variance and Standard Deviation are rounded to 3 decimal places but the whole number is too large to fit in table output.

Problem 1.5

In [14]:

```
print('\nPearson Correlation Coefficient on Columns')
print(covid.iloc[: , 1:].corr(method='pearson'))
print('\nSpearman Correlation Coefficient on Columns')
print(covid.iloc[: , 1:].corr(method='spearman'))
```

```
Pearson Correlation Coefficient on Columns
              Total Cases  Total Deaths  TOTCases_1M  TOTDeath_!M  TotalTested
Total Cases      1.000000      0.940320     0.306869     0.361500     0.659495
Total Deaths     0.940320      1.000000     0.310425     0.525759     0.620081
TOTCases_1M      0.306869      0.310425     1.000000     0.524348     0.129914
TOTDeath_!M      0.361500      0.525759     0.524348     1.000000     0.190367
TotalTested      0.659495      0.620081     0.129914     0.190367     1.000000

Spearman Correlation Coefficient on Columns
              Total Cases  Total Deaths  TOTCases_1M  TOTDeath_!M  TotalTested
Total Cases      1.000000      0.919164     0.735747     0.719670     0.736226
Total Deaths     0.919164      1.000000     0.643341     0.794517     0.668932
TOTCases_1M      0.735747      0.643341     1.000000     0.889098     0.456534
TOTDeath_!M      0.719670      0.794517     0.889098     1.000000     0.448563
TotalTested      0.736226      0.668932     0.456534     0.448563     1.000000
```