Covid-19 Data Analysis of India

Pankaj & Sakshi

12/01/2021

This data is taken from Bing COVID-19 Tracker (www.bing.com/covid). Bing COVID-19 data includes confirmed, fatal, and recovered cases from all regions. The data is collected from multiple trusted, reliable sources, including the World Health Organization (WHO), Centers for Disease Control and Prevention (CDC), national and state public health departments, BNO News, 24/7 Wall St., and Wikipedia.

Importing Data and packages

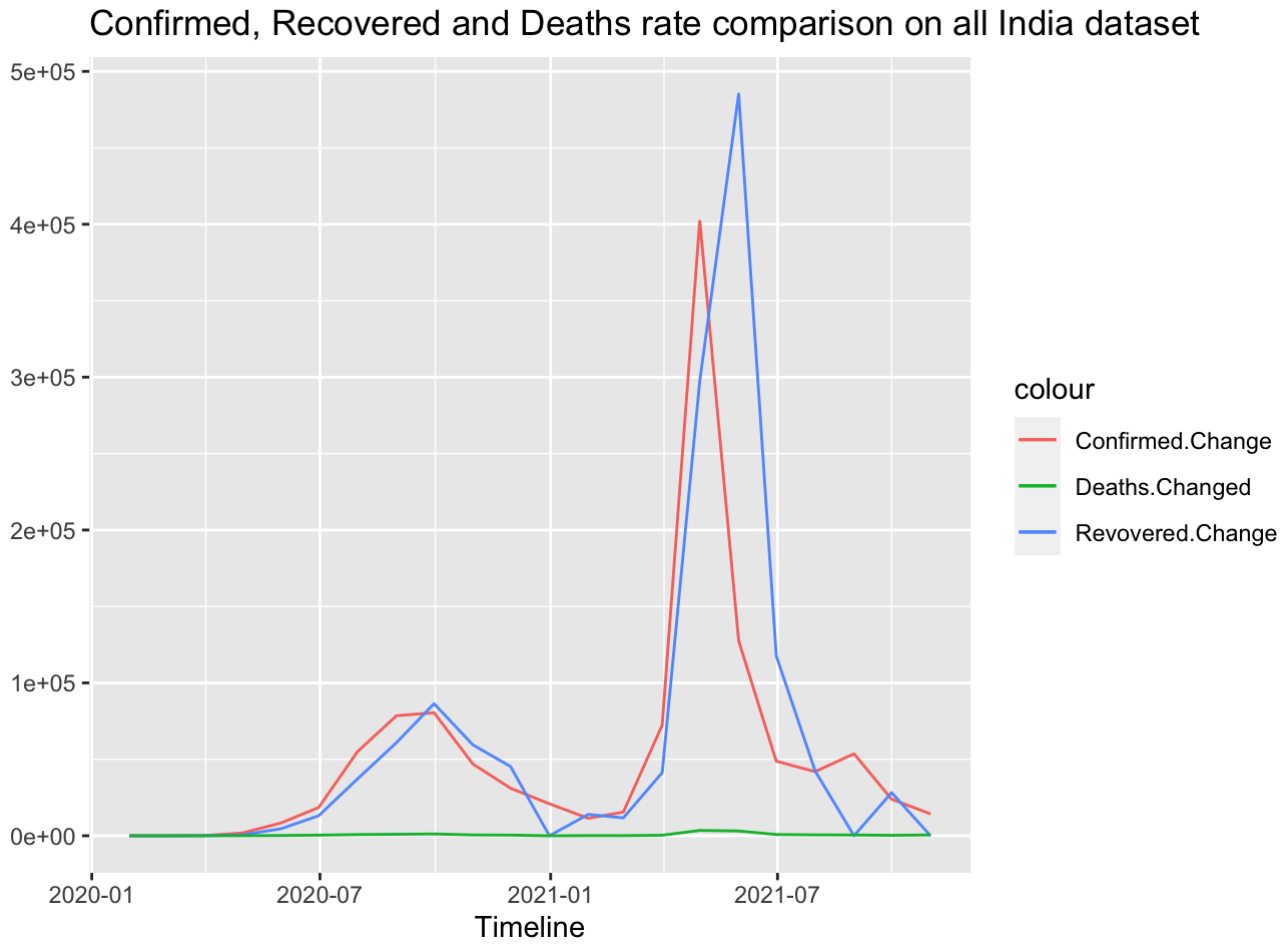
```
# add packages you need for this assignment
library("xlsx")
library(tidyverse)  # includes tibbles, ggplot2, dyplr, and more.
```

```
Demo2<- read.csv("~/Desktop/Demo2.csv")
ind<- read.csv("~/Desktop/ind.csv")
totaldata <- read.xlsx("~/Downloads/Indiaset.xlsx", 1, header=TRUE)
```

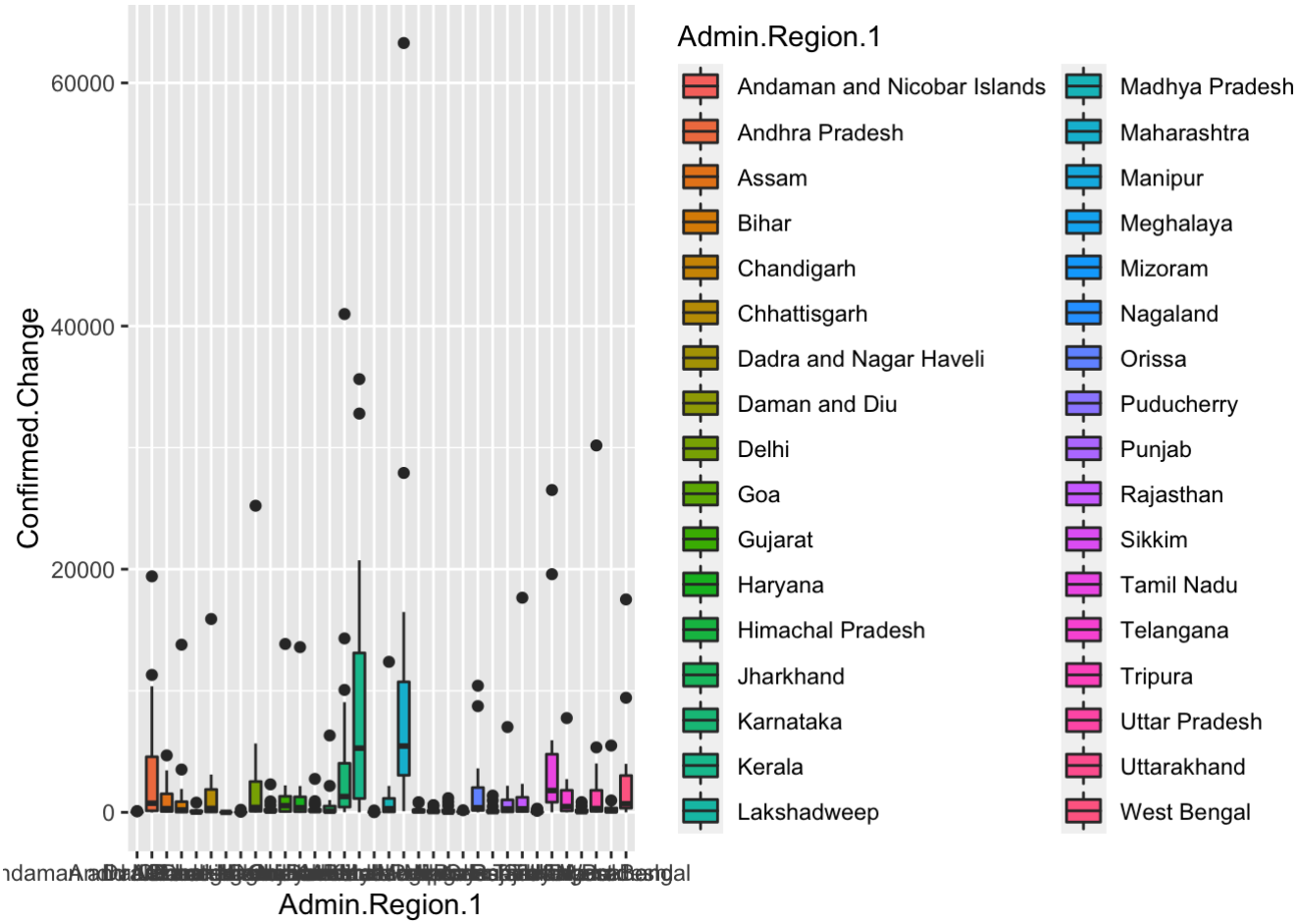
```
str(Demo2) #Structure of dataset
```

```
## 'data.frame':    642 obs. of  17 variables:
##  $ ID                : int  131295993 138635126 69421786 138635138 138635121 13129599
8 131295999 131296003 49177030 39015861 ...
##  $ Date              : chr  "4/30/21" "5/31/21" "9/30/20" "5/31/21" ...
##  $ Confirmed         : int  4665754 5761015 1400922 2123029 2618735 1174552 1282504 7
44602 431719 180298 ...
##  $ Confirmed.Change  : int  63282 14123 16476 26513 14304 25219 30180 15902 9601 5537
...
##  $ Deaths            : int  68813 95344 36662 24232 29090 16147 12570 8581 14994 7855
...
##  $ Deaths.Changed    : int  828 500 481 478 411 375 332 269 265 245 ...
##  $ Recovered         : int  3868976 5395370 1088322 1770503 2261590 1033825 928971 60
1161 256158 90911 ...
##  $ Revovered.Change  : int  69710 33000 19163 31223 44473 25288 32494 13677 7543 1951
...
##  $ Population.Density: int  950 950 950 1440 830 29260 2140 490 950 950 ...
##  $ Admin.Region.1    : chr  "Maharashtra" "Maharashtra" "Maharashtra" "Tamil Nadu"
...
##  $ Month             : chr  "Apr-21" "May-21" "Sep-20" "May-21" ...
##  $ Lattitude         : num  19.5 19.5 19.5 11 14.7 ...
##  $ Longitude         : num  76.1 76.1 76.1 78.4 76.2 ...
##  $ ISO2              : chr  "IN" "IN" "IN" "IN" ...
##  $ ISO3              : chr  "IND" "IND" "IND" "IND" ...
##  $ Country.Region    : chr  "India" "India" "India" "India" ...
##  $ Admin.Region.2    : logi  NA NA NA NA NA NA ...
```

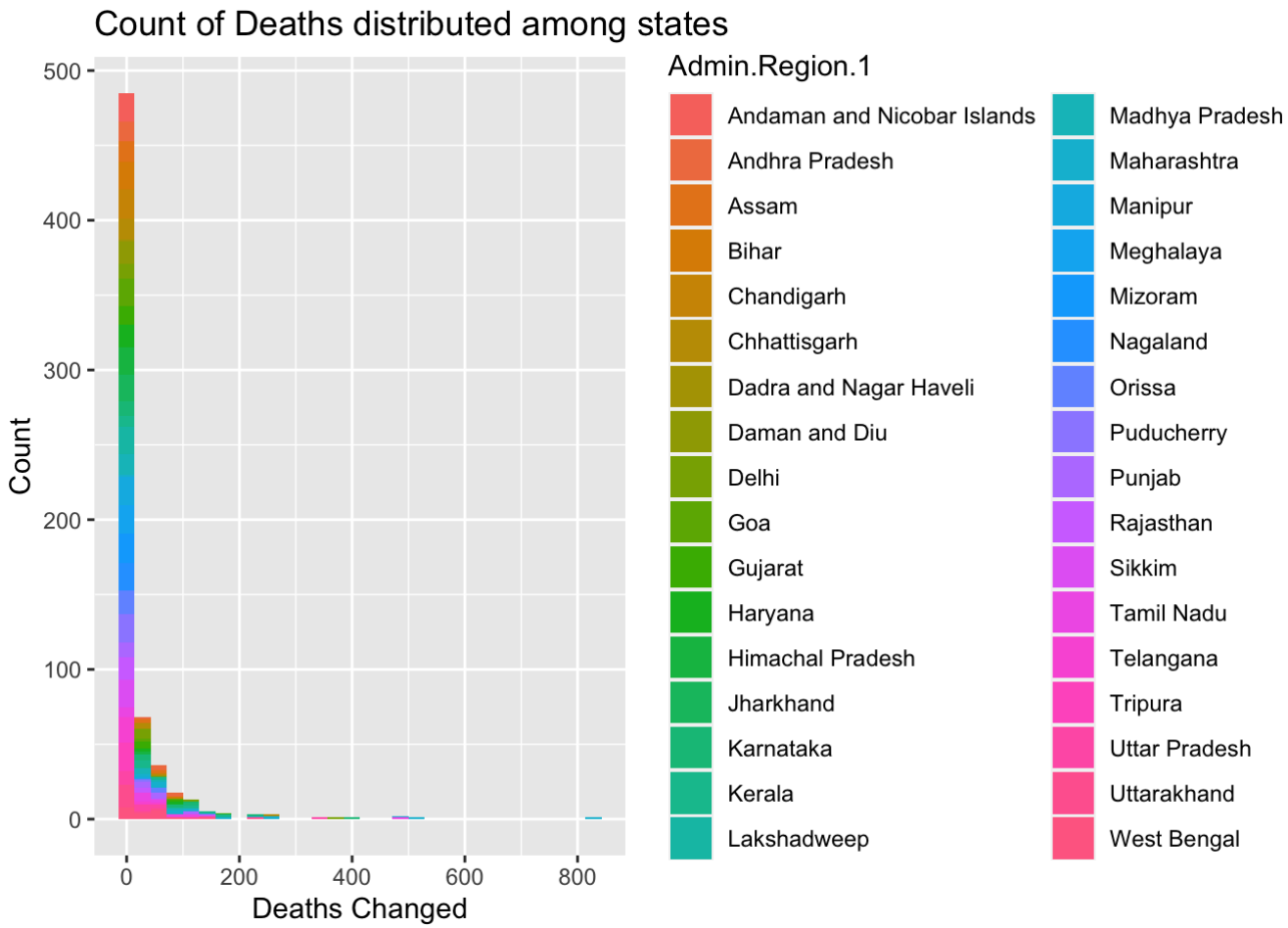
```
plot1 <-
  ggplot(totaldata, aes(x=Date)) +
  geom_line(aes(y = Confirmed.Change, color = "Confirmed.Change")) +
  geom_line(aes(y = Revovered.Change, color="Revovered.Change")) +
  geom_line(aes(y = Deaths.Changed, color="Deaths.Changed"))

print(plot1 +ggtitle("Confirmed, Recovered and Deaths rate comparison on all India datas
et")+ labs(y=" ", x = "Timeline"))
```



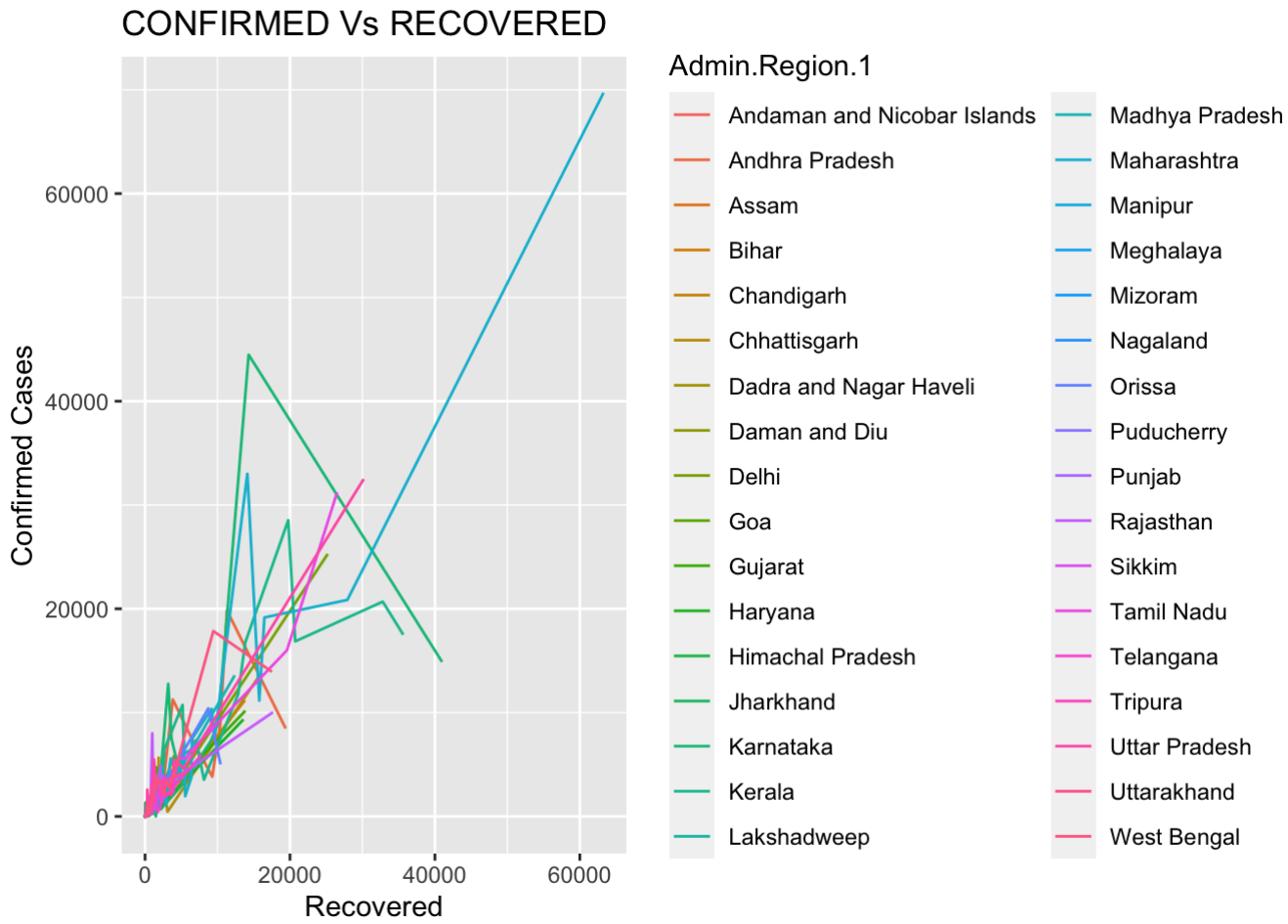

```
ggplot(data = Demo2, mapping = aes(x = Admin.Region.1, y = Confirmed.Change)) +
  geom_boxplot(aes(fill = Admin.Region.1))
```



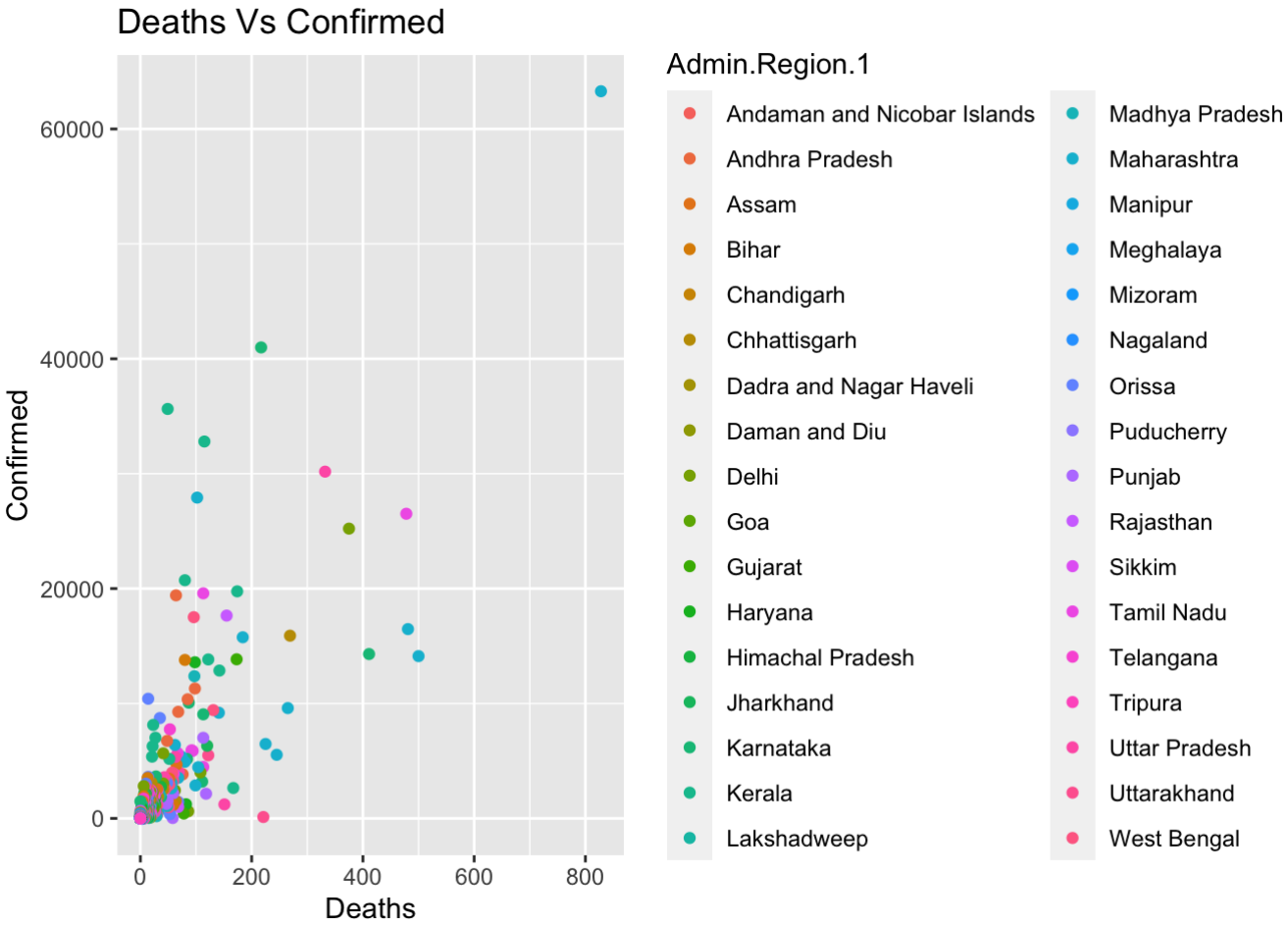

```
plot2 <- ggplot(data = Demo2, aes(x = Deaths.Changed)) +
  geom_histogram(aes(fill = Admin.Region.1),bins = 30)
print(plot2 +ggtitle("Count of Deaths distributed among states")+ labs(y="Count", x = "D
eaths Changed"))
```




```
plot3 <- ggplot(data = Demo2, mapping = aes(x = Confirmed.Change)) +
  geom_line(aes(y = Revovered.Change, color = Admin.Region.1))
print(plot3 +ggtitle("CONFIRMED Vs RECOVERED")+ labs(y="Confirmed Cases", x = "Recovere
d"))
```

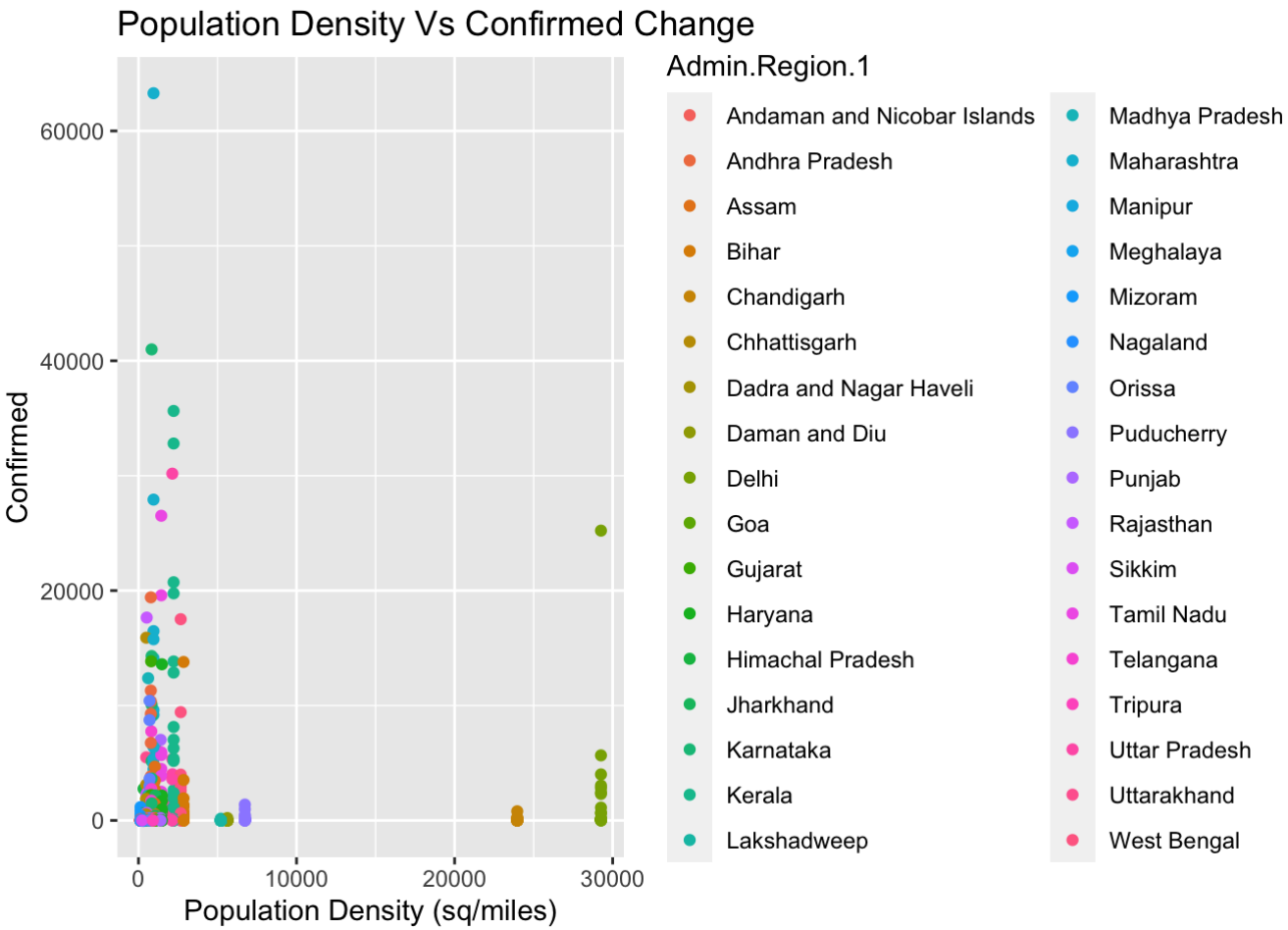



```
plot4 <-  ggplot(data = Demo2) +
  geom_point(mapping = aes(x = Deaths.Changed, y = Confirmed.Change,colour = Admin.Regio
n.1))
print(plot4 +ggtitle("Deaths Vs Confirmed")+ labs(y="Confirmed", x = "Deaths"))
```




```
plot5 <-ggplot(data = Demo2) +
  geom_point(mapping = aes(x = Population.Density, y = Confirmed.Change,color = Admin.Re
gion.1))
print(plot5 +ggtitle("Population Density Vs Confirmed Change")+ labs(y="Confirmed", x =
"Population Density (sq/miles)"))
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



Hypothesis Testing -

- This Hypothesis testing is done on statement that the average number of Confirmed changed cases in the duration of January 2020 - October 2020 is same as Confirmed Changed cases in January 2021 - October 2021.
- null hypothesis, denoted by H_0
- alternative hypothesis, denoted by H_1
- Assuming variance is equal.

In our binomial example, we may state

$$H_0 : \mu_w 1 = \mu_w 2,$$
$$H_1 : \mu_w 1 \neq \mu_w 2$$

```
wave1 <- ind[1:10,]  #Data from Jan 2020 to Oct 2020
wave2 <- ind[13:22,] #Data from Jan 2021 to Oct 2021
```

```
w1 <- wave1$Confirmed.Change
w2 <- wave2$Confirmed.Change
t.test(w1,w2, var.equal = TRUE, conf.level = .95)
```

```
##
##  Two Sample t-test
##
## data:  w1 and w2
## t = -1.3449, df = 18, p-value = 0.1954
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -133610.35   29314.15
## sample estimates:
## mean of x mean of y
##   28989.6   81137.7
```

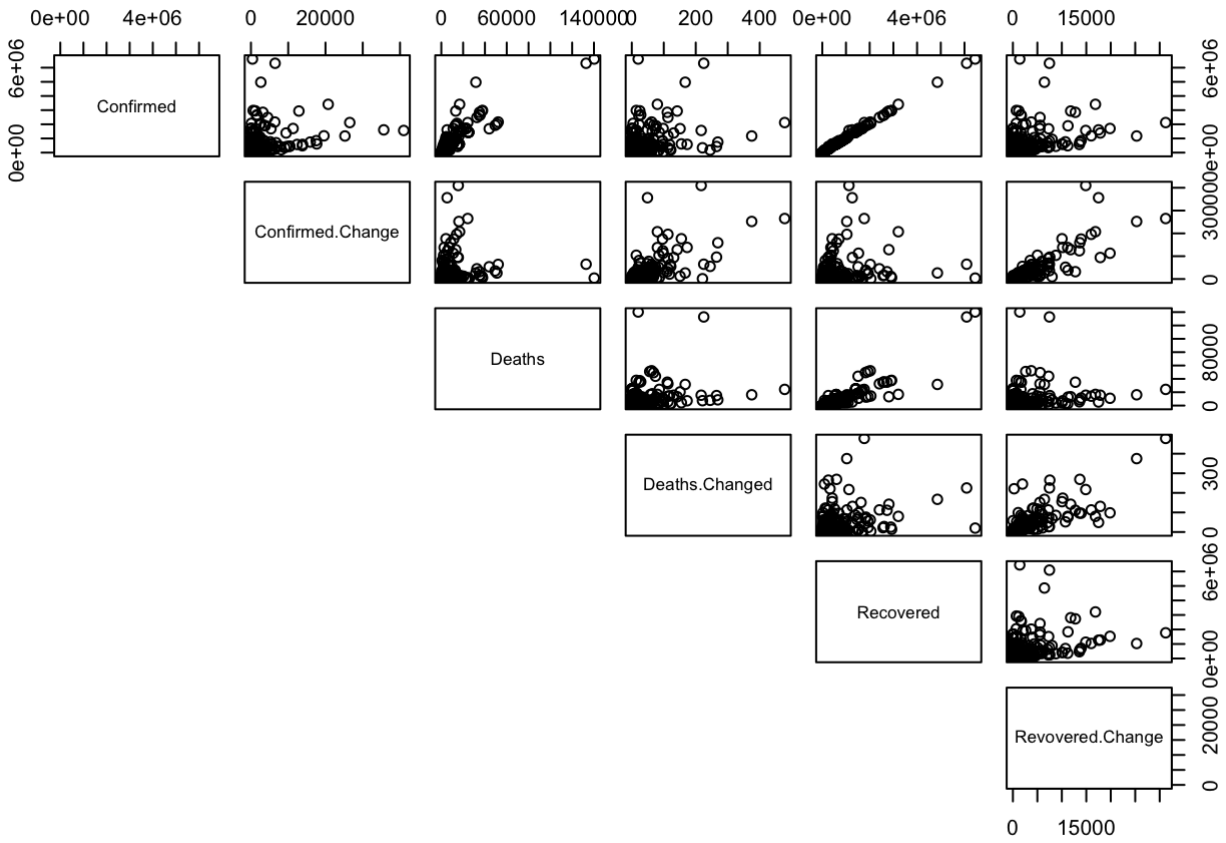
In the above test we used t-test to compare mean of two sample sets with 95% confidence Interval and the resulted P-value is 0.1954 which is comparitively very large to Alpha (0.05). Since the P-value is larger than alpha we will accept the H_0 which stated that both means are equal.

Regression

```
library(ISLR)
data(Demo2)
```

```
## Warning in data(Demo2): data set 'Demo2' not found
```

```
i <- sample(2, nrow(Demo2), replace=TRUE, prob=c(0.8, 0.2))
Demo2Training <- Demo2[i==1,]
Demo2Test <- Demo2[i==2,]
pairs(Demo2Training[,3:8],lower.panel = NULL)
```



From the above scatterplot matrix we can determine that “Recovered” has direct relationship with “Confirmed cases” and “Deaths”.

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
# Create a null model
intercept_only <- lm(Confirmed ~ 1, data=Demo2Training[,3:8])
# Create a full model
all <- lm(Confirmed~., data=Demo2Training[,3:8])
# perform forward step-wise regression
forward <- stepAIC (intercept_only, direction='forward',scope = formula(all))
```

```
## Start:  AIC=13561.45
## Confirmed ~ 1
##
##                     Df  Sum of Sq        RSS   AIC
## + Recovered          1 2.4286e+14 6.6470e+11 10588
## + Deaths             1 1.9562e+14 4.7909e+13 12744
## + Revovered.Change   1 4.5778e+13 1.9775e+14 13458
## + Deaths.Changed     1 4.1404e+13 2.0212e+14 13470
## + Confirmed.Change   1 2.6484e+13 2.1704e+14 13505
## <none>                            2.4353e+14 13562
##
## Step:  AIC=10588.02
## Confirmed ~ Recovered
##
##                     Df  Sum of Sq        RSS     AIC
## + Confirmed.Change   1 5.9264e+11 7.2068e+10  9470.3
## + Revovered.Change   1 4.6217e+11 2.0254e+11  9991.1
## + Deaths.Changed     1 3.4650e+11 3.1820e+11 10218.7
## <none>                            6.6470e+11 10588.0
## + Deaths             1 3.0134e+08 6.6440e+11 10589.8
##
## Step:  AIC=9470.27
## Confirmed ~ Recovered + Confirmed.Change
##
##                     Df  Sum of Sq        RSS    AIC
## + Deaths             1 1.6225e+10 5.5843e+10 9343.7
## + Deaths.Changed     1 6.6721e+09 6.5396e+10 9423.3
## + Revovered.Change   1 1.0451e+09 7.1023e+10 9464.9
## <none>                            7.2068e+10 9470.3
##
## Step:  AIC=9343.71
## Confirmed ~ Recovered + Confirmed.Change + Deaths
##
##                     Df  Sum of Sq        RSS    AIC
## + Deaths.Changed     1 3648188631 5.2195e+10 9311.7
## + Revovered.Change   1 2401099877 5.3442e+10 9323.6
## <none>                            5.5843e+10 9343.7
##
## Step:  AIC=9311.66
## Confirmed ~ Recovered + Confirmed.Change + Deaths + Deaths.Changed
##
##                     Df Sum of Sq        RSS    AIC
## + Revovered.Change   1 437807213 5.1757e+10 9309.4
## <none>                           5.2195e+10 9311.7
##
## Step:  AIC=9309.41
## Confirmed ~ Recovered + Confirmed.Change + Deaths + Deaths.Changed +
##     Revovered.Change
```

```
summary(forward)
```

```
##
## Call:
## lm(formula = Confirmed ~ Recovered + Confirmed.Change + Deaths +
##     Deaths.Changed + Revovered.Change, data = Demo2Training[,
##     3:8])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -136801   -1217    -554     188   66123
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.602e+02  5.222e+02   1.073   0.2839
## Recovered        1.001e+00  1.667e-03 600.878  < 2e-16 ***
## Confirmed.Change 8.432e+00  2.430e-01  34.697  < 2e-16 ***
## Deaths           1.109e+00  9.706e-02  11.430  < 2e-16 ***
## Deaths.Changed   7.128e+01  1.770e+01   4.026 6.55e-05 ***
## Revovered.Change 7.177e-01  3.497e-01   2.052   0.0406 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10190 on 498 degrees of freedom
## Multiple R-squared:  0.9998, Adjusted R-squared:  0.9998
## F-statistic: 4.685e+05 on 5 and 498 DF,  p-value: < 2.2e-16
```

```
data(Demo2)
```

```
## Warning in data(Demo2): data set 'Demo2' not found
```

```
fitlm <- lm(Confirmed.Change ~ Revovered.Change, data=Demo2)
summary(fitlm)
```

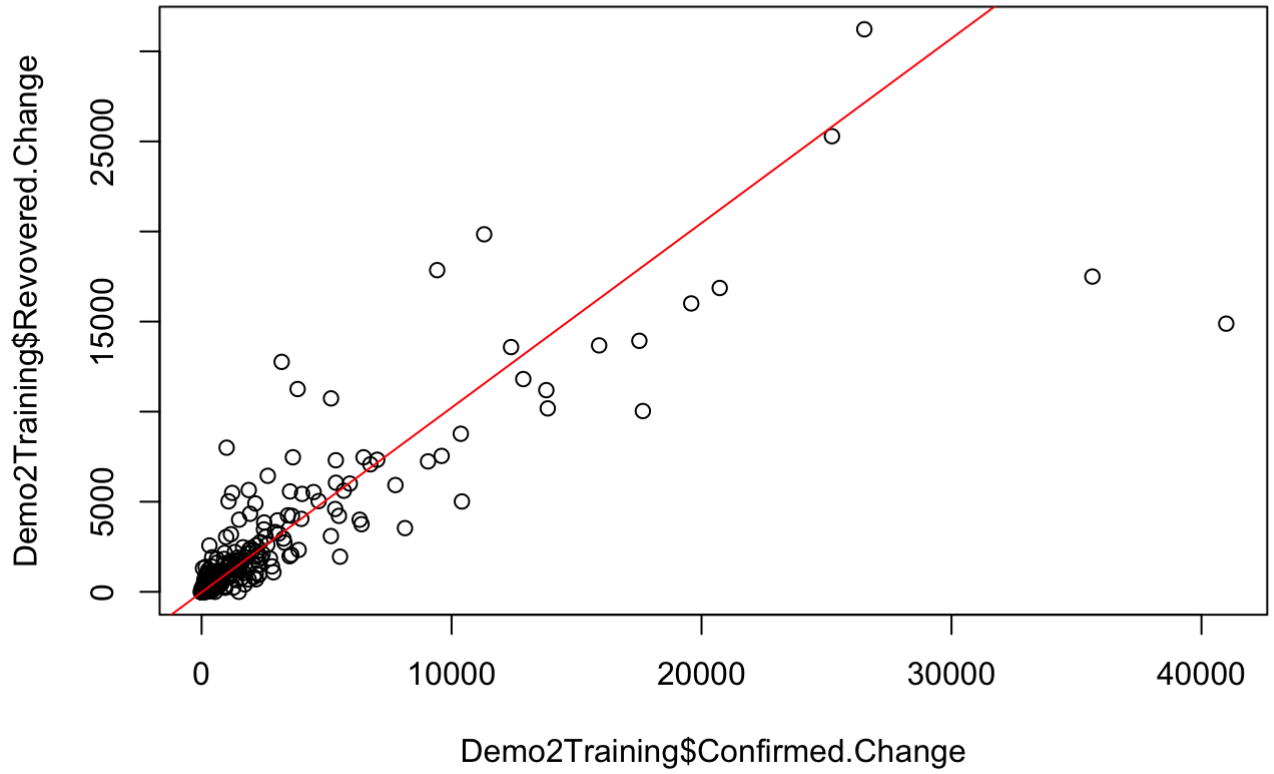
```
##
## Call:
## lm(formula = Confirmed.Change ~ Revovered.Change, data = Demo2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24722.9   -206.7   -171.4    -85.3  27814.1
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      172.22254   96.03506   1.793   0.0734 .
## Revovered.Change   0.87367    0.01787  48.902   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2299 on 640 degrees of freedom
## Multiple R-squared:  0.7889, Adjusted R-squared:  0.7885
## F-statistic:  2391 on 1 and 640 DF,  p-value: < 2.2e-16
```

The above result shows that

```
fitlm <- lm(Demo2Training$Confirmed.Change~ Demo2Training$Revovered.Change, data=Demo2Tr
aining[,3:8])
summary(fitlm)
```

```
##
## Call:
## lm(formula = Demo2Training$Confirmed.Change ~ Demo2Training$Revovered.Change,
##     data = Demo2Training[, 3:8])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9851.7   -76.4    22.2    54.3 25761.6
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -25.41651   91.15759  -0.279     0.78
## Demo2Training$Revovered.Change   1.02485    0.02578  39.755   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1895 on 502 degrees of freedom
## Multiple R-squared:  0.7589, Adjusted R-squared:  0.7585
## F-statistic:  1580 on 1 and 502 DF,  p-value: < 2.2e-16
```

```
plot(Demo2Training$Confirmed.Change, Demo2Training$Revovered.Change)
abline(fitlm, col="red")
```



Checking MAE and MSE

```
library(MLmetrics)
```

```
##
## Attaching package: 'MLmetrics'
```

```
## The following object is masked from 'package:base':
##
##     Recall
```

```
ypred <-predict(object = fitlm, newdata = Demo2Test[,3:8])
```

```
## Warning: 'newdata' had 138 rows but variables found have 504 rows
```

```
summary(ypred)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   -48.99   -12.09   140.10  1342.38  1000.20 31973.36
```

```
MAE(y_pred = ypred, y_true = Demo2Test$Confirmed.Change)
```

```
## Warning in y_true - y_pred: longer object length is not a multiple of shorter
## object length
```

```
## [1] 3509.131
```

```
MSE(y_pred = ypred, y_true = Demo2Test$Confirmed.Change)
```

```
## Warning in y_true - y_pred: longer object length is not a multiple of shorter
## object length
```

```
## [1] 66949960
```

> The above results shows MAE(Mean Absolute Error) of 2474.65 and MSE(Mean Squared Error) of 30067300.

Thankyou