```
#IMPORTING AND UNDERSTANDING THE DATA

cov = read.csv(file.choose())
View(cov)

sapply(cov,class)
dim(cov)unique()
str(cov)
summary(cov)
row.names(cov)
colnames(cov)
unique(cov$continent)
unique(cov$location)

#SANITY CHECK
cov[is.na(cov),]

cov[duplicated(cov),]

cov[cov$new_cases<0,"new_cases"]  #There seems to be inappropriate data
entry but they are legitimate so no need to drop it.

cov[cov$new_deaths<0,"new_deaths"] #There seems to be inappropriate data
entry but they are legitimate so no need to drop it.

cov[cov$new_deaths<0,"total_deaths"]

cov[cov$new_deaths<0,"total_cases"]




#DATA MANIPULATION:

library(dplyr)

cov1 =cov %>%

select(continent,location,new_cases,total_deaths,total_cases,new_deaths,po
pulation,gdp_per_capita,hospital_beds_per_thousand,life_expectancy) %>%
  group_by(continent,location) %>%
  mutate(SpreadRate = sum(new_cases)/max(population)*100, DeathRate =
sum(new_deaths)/max(population)*100,EffectRate = (sum(new_deaths,na.rm =
T)/sum(new_cases,na.rm = T)*100),RecoveryRate = 100-
EffectRate,TotalCountRecorded = sum(new_cases),TotaldeathsRecorded
=sum(new_deaths),Population = population,RecoveryCount = sum(new_cases)-
sum(new_deaths)) %>%
  data.frame()

View(cov1)

cov[cov$location=='Anguilla',"new_cases"]
cov[cov$location=='Anguilla',"new_deaths"]
```

```
#SPLITTING DATA INTO TRAIN AND TEST SET

library(caTools)
set.seed(123)
split = sample.split(cov1$DeathRate,SplitRatio = 0.7)
train_data = subset(cov1, split== T)
test_data = subset(cov1,split == F)
View(train_data)
View(test_data)



#BUILDING THE LINEAR REGRESSION MODEL
classifier = lm(DeathRate ~., data =train_data)
summary(classifier)

library(car)
library(dplyr)
car::vif(classifier)

be = train_data %>%
  select(-location,- continent,-Population,-population,-new_deaths,-
life_expectancy,-RecoveryRate,-RecoveryCount,-new_cases,-total_cases,-
total_deaths )

classifier1 = lm(DeathRate ~., data = be)
summary(classifier1)
car::vif(classifier1)



#INTERPRETATION OF DATA BUILT

    #1) The R squared of this model is pretty good to help develop a
system to predict the number of the population that will be killed by
covid 19.
    #   The R squared value suggests that 72.01% of the variance in the
independent variables explain the target variable (Death Rate).

    #2) Some of the factors significantly impacting the target variable
(Death rate) are GDP per capita, hospital beds, spread rate, effect rate,
total counts recorded, and total deaths recorded.
    # In other to reduce the number of the population being killed by the
virus, policy interventions and stakeholders must sharpen their ax to
understand the mechanism between these significant independent variables
and the target variable.
    #The model also explains to us why some policies like social
distancing, wearing of masks and introduction of the vaccines are all key
to the curbing of the impact of covid on the population.

    #3) There exist a negative relationship between GDP per capita and
death rate. That is to say, the poorer a continent or country is, the
```

```
higher the possibility of losing a significant number of the population to
the virus and vice versa.
    # Recent studies and research allude to the fact that, many poor
countries do not have what it takes to build many well-equipped health
facilities, provide it people with good food and good drinking water which
have an indirect effect on their immune system.
    # Hence, exposing their population to high chance of dying after
contracting the virus. This also explains why countries like the United
State, Luxembourg, just to mention a few have recorded a high number of
covid cases, yet a low number of people being killed by the disease.

    #4) There exist a positive association between hospital beds and death
rate. Researchers at University of Minnesota and University of Washington
affirms to this insight. When the number of hospital beds at a health
facility increases, the impact of death on the population is also high.
    # This is because many health facilities that have a lot of hospital
beds are with few ICUs. Therefore, the hospitals can contain a lot of
people, but do not have ICUs to save them from dying.

    #5) There is a direct correlation between spread rate and death rate.
Spread Rate in this context is the percentage of the population that have
been infested by the disease. In continuity, countries or continents that
have many of its population contracting the disease are of high risk of
losing a chunk of the population to death.

    #6) There is an indirect relationship between effect rate and death
rate. This suggests that the more people contract the disease, the higher
the likelihood of losing a lot of them to the death trap of covid 19.
    # NB: Effect Rate means the percentage of the number of infested
people who have died out of covid

    #7) There exist a negative relationship between total number of cases
recorded and death rate. This means, although people are contracting the
disease, yet, we've not lost high significant number of the population to
death.

    #8) There is a positive association between total number of deaths
recorded and death rate. When high number of people die out of the virus,
it affects quite a high percentage of the continent or country's
population.
```

```
#PREDICTING THE MODEL USING THE TEST_DATA

valid = predict(classifier1,test_data)
summary(valid)
#convert valid into dataframe:
valid1 = data.frame(valid)
View(valid1)
```

```
#To include valid in the test_data for comparism:
final_data = cbind(test_data,valid1)
View(final_data)
head(final_data,5)



#To cal rmse
sqrt(mean((final_data$DeathRate - final_data$valid)^2,na.rm = T))



#The RMSE value is low (0.04477029) and for that matter,suggests that the
model built has the required predictive power to predict the outcome of
death rate .



write.csv(final_data,"CovidPortfolio.csv")
```