# PREDICT COVID LEARNING LOSS

BY

PRINCEWILL OKECHUKWU

Table of Contents

[1](#_Toc133925860)

[1. Introduction 3](#_Toc133925861)

[2. Literature review 3](#_Toc133925862)

[4. Analysis and Results 6](#_Toc133925863)

[4.1 Descriptive Analysis 6](#_Toc133925864)

[4.2 Correlation Analysis 11](#_Toc133925865)

[4.3 PREDICTIVE ANALYSIS 12](#_Toc133925866)

[5. Conclusion and Limitations 16](#_Toc133925867)

[REFERENCE 17](#_Toc133925868)

[APPENDIX 20](#_Toc133925869)

# 1. Introduction

The Covid –19 pandemic led to several sectors of human endeavours. One such sector is the educational sector. Approximately 90% of learners worldwide experienced disrupted education at the pandemic's peak, particularly those from marginalised populations (Reuge, 2021). A thorough analysis of recorded learning loss evidence documented since the beginning of the school closures between March 2020 and March 2022 finds even evidence of learning loss amounting to 0.17 of a standard deviation, equivalent to roughly a one-half years' worth of learning.(Data Catalog,2023 ; Patrinos, Vegas & Carter-Rau, 2022 ).

This dataset contains 41 countries as rows and 24 variables relating to the pandemic experience.

This report analyses the world bank dataset to extract insights about learning losses in certain regions, examines correlation among variables, and builds a prediction model to predict learning loss given input variables. The Cross-Industry Standard Process for Data Mining (CRISP-DM )(Provost & Faucett, 2013) and Business Analytics for Managers (Ghosh,2023) will guide our data mining methodology.

# 2. Literature review

'Learning Loss' refers to a specific or general Loss of knowledge and skills, most commonly due to extended gaps in a student's education (Engzell, Frey & Verhagen(2021).

Educational disruption has been reported in several countries of the world. Unesco(2020), cited in Donnelly & Patrinos (2021), reported that close to 1.6 Billion Learners in more than 190 Countries were affected by school closures.With some countries applying more stringent methods to curb infections and widespread illness among students (Deb et al., 2022).

A study by Engzell, Frey & Verhagen(2021) reports a drop in coursework during the lockdown, even for students with internet connection in developed economies like the Netherlands. Patrinos, Vegas & Carter-Rau (2022) argue that online education is an imperfect substitute for in-person learning, particularly for children from low-income families. Studies show that students in less educated households in Ghana and Netherlands suffered a 60% larger loss than their counterparts. Ullah and Alli (2022).

Conversely, there were positive improvements in other regions. Two Studies from countries in East Asia report learning improvements during the lockdown, Clark et al. (2021), whose study of Chinese students' performance during the pandemic, highlights that the very best students performed better with online learning, likewise in Japan, math achievement scores had increased by six months after COVID-related school closures (Asakawa & Ohtake, 2021) cited in (Patrinos, Vegas & Carter-Rau( 2022). Other reports from other regions suggest online technological support from top schools was satisfactory, like in Italy (Giovannella, Marcello and Donatella (2020); home learning support from schooled adults and access to learning resources were advantageous for Ghana students. (Sabates, Carter & Stern 2021)

**Variable Selection**

As mentioned in the literature review and sources in the reference list, The following variables are significant factors in learning loss during the COVID pandemic. Some of these variables will form a base for building intuition for the descriptive and predictive analysis.

Weeks schools closed & Stringency (Deb et al., 2022;Patrinos, Vegas & Carter-Rau, 2022 )

School Quality & Years of schooling (Ullah & Alli 2022;)

OECD (Dorn et al.,2020)

Private (Alam & Tiwari ,2021; Booth et al.,2020)

Internet (Clark et al., 2021; Asakawa & Ohtake, 2021; Giovannella, Marcello and Donatella, 2020 )

Although studies are limited, there is an intuitive link between a country's GDP and income with learning loss.(Engzell, Frey & Verhagen 2021 ; Josephson, Kilic & Michler,2021)

The region code indicates the geographical location of the samples, as countries are too diverse to be meaningful. (Patrinos, Vegas & Carter-Rau, 2022)

The transformations & interactions of these variables in multiple linear regression and where necessary.

**3. Methodology and Data**

The original dataset contains 41 countries and 24 demographic variables. Sixteen variables are used in the further analysis; these variables have been reduced to 16 based on research studies.

|  |  |
| --- | --- |
| **loss** | **Decline in test scores due to COVID school closures in standard deviations 2021** |
| **weeks** | Weeks schools closed on average 2021 |
| **gdp** | GDP per capita 2020 |
| **private** | Proportion of private schools in country, primary, 2019 |
| **internet** | Internet penetration rate - individuals using the Internet (% of population) |
| **hlo** | School quality |
| **stringency** | Stingency of lockdowns index, average 2021 |
| **loggdp** | Log of GDP per capita |
| **logweeks** | Log weeks schools closed |
| **pop** | Population |
| **schooling** | Years of schooling |
| **region\_code** | Region code |
| **hlo25** | hlo divided by 25 |
| **oecd** | OECD country |
| **covid** | hlo scores minus learning loss |
| **high** | high income country |
| **covid25** | covid divided by 25 |

Table 1: Table of Variables selected.

The research employed descriptive analysis, correlation analysis, and predictive analysis. Data preparation addressed data quality and tidiness issues, such as imputing missing data in log weeks and coding region\_code for analysis. Hypothesis tests (t-tests) is used to test the statistical significance of certain assumptions between the two income level groups.

**This Analytical Techniques to be used for the research:**

Descriptive analysis: Uses a step-wise exploration approach from Univariate, bivariate and multivariate distribution to explore dependent and independent variables to derive insights.

Correlation Analysis: Correlation analysis is conducted to identify associations between Learning Loss and the selected independent variables.

Predictive Analysis: Multiple linear regression analysis is used to model the relationship between loss and the independent variables, in an iterative process to optimise model performance and select statistically significant independent variables. The model is also evaluated using the cross-validation methods and comparing the root-mean squares of selected models in predicting.

# 4. Analysis and Results

Following Hanushek and Woessmann (2020), cited in Patrinos(2022), “we convert results presented in months of loss to standard deviations, with one school year of learning equal to 0.33 standard deviation”.

## 4.1 Descriptive Analysis

The following research involves descriptive, correlation and predictive.

* **Univariate exploration:**

The average loss reported for the dataset was 0.2268, which is over eight months of school learning. From the visualisation of the Loss variable in Figure 1a, it is seen that the distribution is right skewed, given there are two outliers, Nepal 0.8 and India 0.52, as seen in the 1b..

Chart, histogram

Description automatically generatedChart, box and whisker chart

Description automatically generated

Figure 1a: Histogram of Learning Loss. Figure 1b: Boxplot of Learning Loss

Nepal had 2.42 school years of learning loss, while India experienced 1.58 school years of learning Loss. Malawi,Cambodia and Argentina also made it to the Top 5 Learning Losses.

On the other end of countries that seem to manage the leaning loss, Sweden tops the list experiencing no weeks of loss experiencing no weeks of learning loss.Japan as discussed in the literature review, also made it to the Top 5 best-controlled learning loss countries shown to manage learning loss well.

Chart, bar chart

Description automatically generatedChart, waterfall chart

Description automatically generated

Figure 2: Shows the most impacted countries by the covid-19, the plot on the right shows the countries least impacted.

* **Bivariate Exploration**

Box plot of Learning Loss in High-income countries versus low-income countries

From the Box plot, It can be seen that the learning loss for low-income countries is more spread out than for High-Income countries; in fact, the maximum loss experienced by a high-income country is less than the median Loss for low-Income countries.   
a t-test is used to confirm the difference in means for the learning loss experienced in low income and high-income countries.

Chart, box and whisker chart

Description automatically generated

Figure 3: Boxplot of Learning Loss spread in Low- and High-income countries.

Hypothesis Test to confirm the difference in means for the High & Low income demographic learning loss:

Given the Null Hypothesis:

Ho : There was no difference in learning loss experienced in a low-income and high-income country

Ha: The learning loss experienced In High-income countries is less than that experienced in Low-income countries.

After computing the t-test, the p-value for the test is 0.000325; this means that the learning loss in high-income countries is less than the loss in low-income countries and did not happen by chance.

Also, The Plot of Learning Loss versus Weeks of school closure shows a moderate positive linear relationship with Learning Loss, With a correlation coefficient (r) of 0.644.

Chart, scatter chart

Description automatically generated

Figure 4: Learning Loss vs Weeks of school closure

* **Multivariate Exploration:**

The analysis of Schooling and the Internet in High-Income countries shows that all high-income counties are in the region of longer years of schooling and more internet penetration compared to lower-income countries. From the literature review, this was a key factor influencing learning loss, as longer years of schooling was associated in providing help to younger students.

Chart, scatter chart

Description automatically generated

Figure 5: Schooling Versus internet in High & Low-income Countries

The analysis of loss and weeks for each region shows that advanced economies had the last weeks of school closures and the most negligible learning loss compared to other regions. with south asia being the most affected, the region were countries with the loss outliers located.

Chart, scatter chart

Description automatically generated

Figure 6: Facet Wrap of Weeks vs Loss for each Region

## 4.2 Correlation Analysis

Correction matric depicts the associations between learning loss and the other independent variables.



Figure 7a: Correlation Matrix for Learning Loss and Independent Variables

Graphical user interface, application, table, Excel

Description automatically generated

Table 2: A tabular view of Correlation Matrix for Learning Loss cl

## 4.3 PREDICTIVE ANALYSIS

The correlation analysis is used as a guide to starting regression; several Iterations for the model building were implemented; here are the most relevant model iterations with the highest R-Squared for non-interaction and interaction between variables.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Iterations** | **Adjusted R-Squared** | **No. of Variables** | **Significant Variables** | **Interaction between 2 variables** |
| **1** | 1 | 16 | 2 | NO |
| **2** | 1 | 15 | 2 | NO |
| **3** | 0.467 | 6 | 2 | NO |
| **4** | 0.4748 | 5 | 2 | NO |
| **5** | 0.5367 | 4 | 3 | YES |
| **6** | 0.4943 | 3 | 2 | NO |

Table 3: A tabular view of Model Iterations

From the table above, in the first two models, the independent variables were highly multicollinear with the model even after taking off the highest p-values. It may seem that the model was overfitting to the data. A change of approach using research-tailored variables that had moderate to good correlation with the loss variable (+/- 0.4) while keeping correlations among other independent variables less than (+/-0.8). Comparing the summary for Model 4 Model 5 and Model 6.

Model 5 has the highest Adjusted R-Squared score amongst the three models, however, when we test for homoscedasticity (constant variance) and normality there appears to be a pattern in the residual plot compared to Model 4 and Model 6 which both appear to be randomly dispersed around 0 to +/-2. There is also presence of some outlier points shown in the 3 of them. One reason for this is the Outlier Learning Losses recorded for Nepal and India.

Table

Description automatically generatedTable

Description automatically generated

Figure 8a. Summary for Model 4, 8b, Summary for Model 5(right)

Table

Description automatically generated

Figure 8c. Summary for model 6

Chart, scatter chart

Description automatically generatedChart, histogram

Description automatically generated

Figure 8d. Residual and normality model 4

Chart, scatter chart

Description automatically generatedChart, histogram

Description automatically generated

Figure 8e. Residual and normality plot for model 5

Chart, scatter chart

Description automatically generatedChart, histogram

Description automatically generated

Figure 8f. Residual and normality plot for model 6

**Model Choice**

From the 3 models, at first glance, Model 5 appears to be the best model, because it has the highest Adjusted R-squared value of 0.5367 than Model 4 at 0.478 and Model 6 at 0.4943, However, this model has a tendency to overfit with the training data and not do well when it encounters a new data.  
**Model 6,** is the best choice model in this scenario, it explains more variability in Loss than Model 4, it also has a constant variance and it follows a normal distribution.

From the summary, Model 6 explains 49.43 % of the variability in the Loss variable.

Examine the coefficients in Model 4 :

Logweeks: Logweeks it is statistically significant and has a direct impact on the model , which is intuitive as Logweeks has a moderate correlation with learning Loss.

Covid: covid explains school quality minus learning loss , it also perfectly correlates with hlo, it has an inverse relationship with the model predictability which is also intuitive.

South Asia : Direct Relationship, South Asia correlates with loss

**Cross Validation with Model 5 & Model 6**

The code for the cross-validation is seen in the R-Code attached. Using a 70:30 train-test-split ratio, to train a new model with the same variables, with new training data, the Model 5 gives a higher adjusted R-Squared of 0.5542, while the Model 6 gives an adjusted r-square of 0.5419.  
When use to predict the validation data the Model 5 gives a Root Mean Square Error of 0.2575, while 0.18695 for the Model\_6. This proves Model 6 is better and does not overfit like Model 5.

# 5. Conclusion and Limitations

This report uses descriptive and predictive analytics to analyse the impact of covid-19 on learning loss for 41 countries. The insights describes what variables were most significant with the Loss variables. It reported that there is indeed a learning loss among high and Low-income demographics as thehigh-income countries have access to resources to mitigate the learning loss. However, the data has some limitations. The first limitation is the size of the data.

Since the adjusted R-Square is a function of a number of independent variables and sample size, a larger sample will result in a better-trained model. Additionally, it is a cross-sectional data and has no time series component in the data. Therefore, it is impossible to do a time series analysis of the data. Lastly, which the low R-squared value, the analysis may not account for all other relevant factors that are influencing learning loss.

# REFERENCE

Alam, A., & Tiwari, P. (2021). Implications of COVID-19 for low-cost private schools. \*UNICEF, Issue Brief\*, (8).

Asakawa, S., & Ohtake, F. (2021). Impact of Temporary School Closure Due to COVID-19 on the Academic Achievement of Elementary School Students. \*Discussion Papers in Economics and Business, 21-14\*, Osaka University, Graduate School of Economics. https://ideas.repec.org/p/osk/wpaper/2114.html

Booth, C., Villadsen, A., Goodman, A., & Fitzsimons, E. (2021). Parental perceptions of learning loss during COVID-19 school closures in 2020. \*British Journal of Educational Studies\*, 69(6), 657-673.

Clark, A. E., Nong, H., Zhu, H., & Zhu, R. (2021). Compensating for academic loss: Online learning and student performance during the COVID-19 pandemic. \*China Economic Review\*, 68, 101629.

Data Catalog (2023) Learning Loss COVID-19 2020-2022. Available at: https://datacatalog.worldbank.org/search/dataset/0063679/Learning-Loss-COVID-19-2020-2022 (Accessed: May 1, 2023).

Deb, P., Furceri, D., Ostry, J. D., & Tawk, N. (2022). The economic effects of COVID-19 containment measures. \*Open Economies Review\*, 33(1), 1-32.

Donnelly, R., & Patrinos, H. A. (2021). Learning loss during COVID-19: An early systematic review. \*Prospects\*, 1-9.

Dorn, E., Hancock, B., Sarakatsannis, J., & Viruleg, E. (2020). COVID-19 and learning loss—disparities grow, and students need help. \*McKinsey & Company\*, December 8, 6-7.

Engzell, Frey, & Verhagen (2021). Learning inequality during the Covid-19 pandemic.

Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., & Tu, X. M. (2014). Log transformation and its implications for data analysis. \*Shanghai Archives of Psychiatry\*, 26(2), 105–109. https://doi.org/10.3969/j.issn.1002-0829.2014.02.009

Giovannella, C., Marcello, P., & Donatella, P. (2020). The effects of the Covid-19 pandemic on Italian learning ecosystems: The school teachers' Perspective at the steady state. \*ID&A Interaction Design & Architecture (s)\*, 45, 264-286.

Hasan, N., & Bao, Y. (2020). Impact of "e-Learning crack-up" perception on psychological distress among college students during COVID-19 pandemic: A mediating role of "fear of academic year loss". \*Children and youth services review\*, 118, 105355.

Hevia, F. J., Vergara-Lope, S., Velásquez-Durán, A., & Calderón, D. (2022). Estimation of the fundamental learning loss and learning poverty related to the COVID-19 pandemic in Mexico. \*International Journal of Educational Development\*, 88, 102515.

Josephson, A., Kilic, T., & Michler, J. D. (2021). Socioeconomic impacts of COVID-19 in low-income countries. \*Nature Human Behaviour\*, 5(5), 557-565.

Patrinos, H. A., Vegas, E., & Carter-Rau, R. (2022). An analysis of COVID-19 student learning loss.

Provost, F., & Fawcett, T. (2013). Data Science for Business: What you need to know about data mining and data-analytic thinking. "O'Reilly Media, Inc.".

Reuge, N., Jenkins, R., Brossard, M., Soobrayan, B., Mizunoya, S., Ackers, J., ... & Taulo, W. G. (2021). Education response to COVID-19 pandemic, a special issue proposed by UNICEF: Editorial review. \*International Journal of Educational Development\*, 87, 102485.

Sabates, R., Carter, E., & Stern, J. M. (2021). Using educational transitions to estimate learning loss due to COVID-19 school closures: The case of Complementary Basic Education in Ghana. \*International Journal of Educational Development\*, 82, 102377.

Ullah, H., & Ali, J. (2022). Impact of COVID-19 pandemic on the schooling of public and private school students in Pakistan. \*Education 3-13\*, 50(8), 1085-1094.

# APPENDIX

getwd()

setwd('Desktop/Buiness Analytics /BE 277/')

#packages:

install.packages("dplyr")

install.packages("ggplot2")

library(dplyr)

#Understand the Business Problem

#Data importation

learnLoss <- read\_excel('learning\_loss\_r\_upload.xlsx')

#View(learnLoss)

#Data inspection - From the data we want to check for missing data and any incorrect data types

#summary(learnLoss)

length(which(is.na(learnLoss)))

#We have two missing values in LogWeeks , on for australia and one for sweden, to aviod the Log Zeoro issure we can a small constant of 1 to the week values and

#take the log, this doesn't affect the data much

constant <- 1

learnLoss$logweeks2 <- log(learnLoss$weeks + constant)

View(learnLoss)

#Remove the incomplete logweeks

learnLoss <- subset(learnLoss, select = -logweeks)

View(learnLoss)

#Check for complete data

length(which(is.na(learnLoss))) #Data Complete

#check Summmary Statistics for Learn Loss table

summary\_stats <- summary(learnLoss)

summary\_stats

#Data Exploration

#Univariate Exploration (Single variable)

#Histogram of Loss

#summary(learnLoss$loss)

interval\_loss <- seq(0,0.8,0.1)

?hist

hist(learnLoss$loss,breaks = interval\_loss,right = TRUE,ylim = c(0,14),main = 'Histogram of Learning Loss')

#the plot of loss is skewed to the right, with the futhermost loss at 0.8,

#Investigating further with a box plot , if there are any outliers.

boxplot(learnLoss$loss, main = "Box Plot of Learning Loss", ylab = "Learning Loss")

#From the boxplot we can see that 2 outliers are present to see the countries

outliers\_loss <- learnLoss[(learnLoss$loss >= 0.50),c(1:2)]

View(outliers\_loss)

#We can see that the Countries that have Outlier Losses are : India, Nepal.with 0.52 & 0.8 respectively

Nepal\_loss = 0.8/0.33

Indian\_Loss = 0.52/0.33

#Years of loss : according to Hanushek and Woessmann(2020) cited in Patrinos(2022), Nepal experienced 2.42 school years of learning loss while india experienced 1.58 school years of learning Loss

#Not removing the outliers from the data because they are not extremely high, plus to also preserve the data quantity

#Histogram

# 2.What are the 5 Countries with the Highest Learning Loss and Ones with the Lowest Learning Loss

#Using Functions to avoid repetitive codes

my\_plots <- function(type,plot\_data,high,marker, ylimit, ax\_size, xlabel, ylabel, title) {

if(type == 'Barchart'){

bars <- barplot(high ,

names.arg = marker,

ylim = ylimit,

cex.names = ax\_size,

xlab = xlabel,

ylab = ylabel,

main = title)

text(x = bars, y = high, labels = high, pos = 3, cex = 0.8, col = "black")

return(bars)

}

}

# 2a)

#Top 5 Countries with the Highest Loss

my\_plots('Barchart',top5\_loss,top5\_loss$loss,top5\_loss$country,c(0,1),0.65,'countries','Learning Loss','Top 5 Learn Loss')

#2b)

#Countries with Least Loss

least5\_loss <- learnLoss[order(learnLoss$loss,decreasing = FALSE)[1:5],c(1,2)]

least5\_loss

my\_plots('Barchart',least5\_loss,least5\_loss$loss, least5\_loss$country,c(0,0.05),0.65,'countries','Learning Loss','Least 5 Learn Loss')

#3) The distribution of weeks school closed

summary(learnLoss$weeks)

#On average Schools closed for 20 weeks

interval\_weeks <- seq(0,96,12)

hist(learnLoss$weeks,breaks = interval\_weeks,right = TRUE,xlim = c(0,100),main = 'Histogram for Weeks of School Closure')

box

boxplot(learnLoss$weeks,main="Plot of weeks school closed",ylab='Weeks closed')

outliers\_weeks <- learnLoss[order(learnLoss$weeks,decreasing = TRUE)[1:2],c(1,3)]

View(outliers\_weeks)

#Explanation: The histogram for weeks is Right-skewed with two outliers, These countries are India and Bangladesh with 93 and 63 weeks closed respectively

#Bivariate Exploration

#4)

#What is the relationship between Learning Loss and Weeks closed

plot(learnLoss$weeks,learnLoss$loss, pch=18,

cex=0.9,

col="#69b3a2",xlab='Weeks closed',ylab='Learning Loss',main = 'Learning Loss vs Weeks Closed ')

abline(lm(learnLoss$loss~learnLoss$weeks), col="red")

cor(learnLoss$loss,learnLoss$weeks)

#Explanation: From the Plot of Learning Loss versus Weeks of school closure, It can be seen that there is a good positive linear relationship with Learning Loss

#Explanation: From the Plot of Learning Loss versus Weeks of school closure, It can be seen that there is a good positive relationship with Learning Loss

#4b) sINCE INTERNET IS AN INPORTANT FACTOR FOR HOME STUDY, LET US SEE IF LOSS NEGATIVELY CORRELATES WITH HIGH Internet

#What is the relationship between Learning Loss and Weeks closed

plot(learnLoss$internet,learnLoss$loss, pch=18,

cex=0.9,

col="#69b3a2",xlab='Internet',ylab='Learning Loss',main = 'Learning Loss vs Internet Access ')

abline(lm(learnLoss$loss~learnLoss$internet), col="red")

#The Graph of Learning Loss vs internet access shows some inverse relationship. The more interent access in a country, the lesser the learning loss

#Hypothesis Tests for Learning Loss for High Income and Low Income demography

High\_income <- learnLoss[(learnLoss$high == 1),2]

low\_income <- learnLoss[(learnLoss$high == 0),2]

low\_income

#Hypothesis Tests

# Perform the t-test

loss\_income\_test<- t.test(High\_income, low\_income,alternative='less')

loss\_income\_test

#The P value for the test is 0.000325 , this means that the loss in low income countries is less than high income countries and did not happen by chance

#Mean there is difference in average Learning

#5) What is the distribution of loss for High income and Low Income countries

?boxplot

boxplot(learnLoss$loss~learnLoss$high,xlab="0: Low-income 1:High-income",ylab = 'Learning Loss',main='Loss by Low & High Income Country')

#Explanation: From the Box plot, It can be seen that the Learning Loss for Low income countries are more spread out than for High Income countries,

#in fact maximum loss experience by a high income country is less the median Loss for Low Income Country

#5b) #Schooling and Internet in High Income countries

#Adding Income Level to the dataframe

learnLoss$income\_level <- ifelse(learnLoss$high == 1,'High','Low')

library(ggplot2)

ggplot(learnLoss, aes(internet, schooling, colour = income\_level)) +

geom\_point()

#Examine the Region Variable

regionCat <- table(learnLoss$region\_code)

regionCat

barplot(regionCat)

regionCatProp <- prop.table(regionCat)

regionCatProp

regionBar <- barplot(regionCat,main = 'Number of countries in regions',las=2,ylim = c(0 , 20),ylab = 'count of countries')

abline(h=0,col='black') #this adds a line to height for height 0

text(0.75,regionCat[1]+1,regionCat[1])

text(1.8,regionCat[2]+1,regionCat[2])

text(3.0,regionCat[3]+1,regionCat[3])

text(4.2,regionCat[4]+1,regionCat[4])

text(5.4,regionCat[5]+1,regionCat[5])

text(6.6,regionCat[6]+1,regionCat[6])

#REGION VS iNCOME LEVEL

loss\_table <- table(learnLoss$income\_level,learnLoss$region\_code)

loss\_table

#oecd is the row , high income country is column

?barplot

barplot(loss\_table,col = c('green','grey'),ylab = 'frequency',cex.names = 0.35,las=2,main='Income Level By Region')

legend(x="topright",fill=c("green","grey"),legend=c('High\_income','Low\_Income'),cex=0.7,inset=0)

#Group By AVERAGE Learning LOSS FOR EACH REGION

df\_loss <- group\_by(learnLoss,region\_code)

region\_loss\_mean <- summarise(df\_loss,region\_mean = mean(loss))

region\_mean\_table <- region\_loss\_mean[order(region\_loss\_mean$region\_mean,decreasing = TRUE),]

View(region\_mean\_table)

#my\_plots('Barchart',region\_loss\_mean,region\_loss\_mean$my\_mean, region\_loss\_mean$region\_code,c(0,0.6),0.35,'reg','Average Loss','Average Loss by Region')

#Facet Wrap - Learning Loss vs Weeks In Each Region

ggplot(data = learnLoss) +

geom\_point(mapping = aes(x = weeks, y = loss)) +

facet\_wrap(~ region\_code, nrow = 2)

#It can be seen that a MAJORITY OF COUNTRIES FOR aDVANCED ECONOMMICS , East Asia and Pacific, Europe and Central Asia has their Leaning lOSS BELOW 25 weeks, South Asia and Latain America had

#had a large number of countries experiencing lEANING lOSS PAST 25 WEEKS.

#Predictive Analytics

#To begin with Predictive analysis, we check correlation

#Since R-SQAURED is a function of the number of independent variable and sample size

#a larger sample size will give a better model

#Using the Systematic Model Building Approach

#correlation exceeding +/- 0.7 may indicate Multicollinearity ,

#Using Trial and Error

#Variables to be used from literature review , Correlation with loss and Logic

learnLossPred <- learnLoss[,c(2,3,5,6,7,8,10,13,16,19,20,21,22,23,24)]

View(learnLossPred)

write.csv(learnLossPred,'shortlist\_loss.csv')

#correlation : Correlation is important to determine association as it measure the extent of linearity between two variables

?cor

corr\_matrix <- round(cor(learnLossPred),2)

View(corr\_matrix)

#Variables to be used from literature review , Correlation and Logic

write.csv(corr\_matrix,'Corr\_matrix.csv')

install.packages("corrplot")

library(corrplot)

corrplot(corr\_matrix, method = "circle")

#adding region\_code to selected variables, for its a geographical location element

learnLossPredR <- learnLoss[,c(2,3,5,6,7,8,10,13,16,18,19,20,21,22,23,24)]

View(learnLossPredR)

#Creating Dummy variables for region code , using Latin America & the Caribbean as reference

learnLossPredR$advanced\_eco <- ifelse(learnLossPredR$region\_code == 'Advanced Economies',1,0)

learnLossPredR$south\_asia <- ifelse(learnLossPredR$region\_code == 'South Asia',1,0)

learnLossPredR$east\_asia\_pac<- ifelse(learnLossPredR$region\_code == 'East Asia and the Pacific',1,0)

learnLossPredR$europe\_cent\_asia <- ifelse(learnLossPredR$region\_code == 'Europe and Central Asia',1,0)

learnLossPredR$sub\_sah\_africa <- ifelse(learnLossPredR$region\_code == 'Sub-Saharan Africa',1,0)

#Removing region code because we now have dummy variables

learnLossPredR <- subset(learnLossPredR, select = -region\_code)

View(learnLossPredR)

View(cor(learnLossPredR))

write.csv(learnLossPredR,'loss4Reg.csv')

corrplot(cor(learnLossPredR), method = "circle")

#Using the systematic model Building approach

#Model 1

#Build Regression With all selected Independent varibles, to serve as Base model before tuning

model1 <- lm(loss~weeks+gdp+private+internet+hlo+stringency+loggdp+schooling+hlo25+oecd+covid+high+covid25+logweeks2+advanced\_eco+south\_asia+east\_asia\_pac+europe\_cent\_asia+sub\_sah\_africa,data = learnLossPredR)

summary(model1)

#Model 1: Adjusted R-Square of 1 , this is rare, and the model may be overfitting or there exists high multicollinearity in the data

#Model99:

#Model 1 showed significant for hlo and internet, weeks is highest unsignificant

model99 <- lm(loss~gdp+private+internet+hlo+stringency+loggdp+schooling+hlo25+oecd+covid+high+covid25+logweeks2+advanced\_eco+south\_asia+east\_asia\_pac+europe\_cent\_asia+sub\_sah\_africa,data = learnLossPredR)

summary(model99)

#model99 r square is 1

model98 <- lm(loss~gdp+private+internet+hlo+stringency+loggdp+schooling+hlo25+oecd+covid+high+logweeks2+advanced\_eco+south\_asia+east\_asia\_pac+europe\_cent\_asia+sub\_sah\_africa,data = learnLossPredR)

summary(model98)

#model98 r square is 1

model97 <- lm(loss~gdp+private+internet+hlo+stringency+loggdp+schooling+hlo25+covid+high+covid25+logweeks2+advanced\_eco+south\_asia+east\_asia\_pac+europe\_cent\_asia+sub\_sah\_africa,data = learnLossPredR)

summary(model97)

#model98 rsquare is 1

model96 <- lm(loss~gdp+private+internet+hlo+stringency+loggdp+schooling+hlo25+covid+high+logweeks2+advanced\_eco+south\_asia+east\_asia\_pac+europe\_cent\_asia+sub\_sah\_africa,data = learnLossPredR)

summary(model96)

model95 <- lm(loss~gdp+private+hlo+stringency+loggdp+schooling+hlo25+covid+high+logweeks2+advanced\_eco+south\_asia+east\_asia\_pac+europe\_cent\_asia+sub\_sah\_africa,data = learnLossPredR)

summary(model95)

model94 <- lm(loss~gdp+private+hlo+stringency+loggdp+schooling+hlo25+covid+high+logweeks2+advanced\_eco+east\_asia\_pac+europe\_cent\_asia+sub\_sah\_africa,data = learnLossPredR)

summary(model94)

model93 <- lm(loss~gdp+private+hlo+stringency+schooling+hlo25+covid+high+logweeks2+advanced\_eco+east\_asia\_pac+europe\_cent\_asia+sub\_sah\_africa,data = learnLossPredR)

summary(model93)

model92 <- lm(loss~gdp+private+hlo+stringency+hlo25+covid+high+logweeks2+advanced\_eco+east\_asia\_pac+europe\_cent\_asia+sub\_sah\_africa,data = learnLossPredR)

summary(model92)

model91 <- lm(loss~gdp+private+hlo+stringency+covid+high+logweeks2+advanced\_eco+east\_asia\_pac+europe\_cent\_asia+sub\_sah\_africa,data = learnLossPredR)

summary(model91)

model90 <- lm(loss~gdp+private+hlo+stringency+covid+logweeks2+advanced\_eco+east\_asia\_pac+europe\_cent\_asia+sub\_sah\_africa,data = learnLossPredR)

summary(model90)

model89 <- lm(loss~gdp+private+hlo+stringency+covid+advanced\_eco+east\_asia\_pac+europe\_cent\_asia+sub\_sah\_africa,data = learnLossPredR)

summary(model89)

model88 <- lm(loss~private+hlo+stringency+covid+advanced\_eco+east\_asia\_pac+europe\_cent\_asia+sub\_sah\_africa,data = learnLossPredR)

summary(model88)

model87 <- lm(loss~private+stringency+covid+advanced\_eco+europe\_cent\_asia+sub\_sah\_africa,data = learnLossPredR)

summary(model87)

model86 <- lm(loss~private+hlo+stringency+covid+advanced\_eco+europe\_cent\_asia,data = learnLossPredR)

summary(model86)

model85 <- lm(loss~hlo+stringency+covid+advanced\_eco+europe\_cent\_asia,data = learnLossPredR)

summary(model85)

#removing hlo

model84 <- lm(loss~ covid + stringency+advanced\_eco,data = learnLossPredR)

summary(model84)

model83 <- lm(loss~+covid+advanced\_eco,data = learnLossPredR)

summary(model83)

model82 <- lm(loss~covid25+schooling,data = learnLossPredR)

summary(model82)

#Model 3:

#selecting variables with moderate correlation of greater than +/- 0.4 correlation with loss, Model 4, we will adjust based on Multicollinearity

#These are : weeks ,GDP,Internet,logGDP, schooling, Oecd,Covid,high,Covid25,logweeks2,advanced economics, South Asia

model\_3 <- lm(loss~weeks+gdp+internet+loggdp+schooling+oecd+covid+high+covid25,logweeks2+advanced\_eco+south\_asia,data = learnLossPredR)

summary(model\_3)

#model\_3 r-square is 1 , there is perfect multicollinearity,with 5 Coefficients not defined because of singularities

#This means that one or more independent variables (predictors) are

#linearly dependent on other independent variables in the model, causing the model to be over-fitting

#Model 4:

#We want to optimize for a good correlation with the dependent variable and minimize correlation with other independent variables,

#Adjusting for Multicollinearity , the following variables have moderate and above correlation with Loss , and show Multicorrelation less than +/-0.8

#logweeks #loggdp, #covid #high #southasia, #advanced\_eco

#model\_4 <- lm(loss~weeks+gdp+internet+loggdp+schooling+oecd+covid+high+logweeks2+advanced\_eco+south\_asia,data = learnLossPredR)

install.packages("car")

library(car)

model\_4 <- lm(loss ~ logweeks2 + loggdp + covid + high + south\_asia + advanced\_eco, data = learnLossPredR)

summary(model\_4)

model\_4\_vif <- vif(model\_4)

model\_4\_vif

#removing the highest not-significant variable from Model 4 , Loggdp, and re run 5

model\_4 <- lm(loss ~ logweeks2 + covid + high + south\_asia + advanced\_eco, data = learnLossPredR)

summary(model\_4)

model\_5\_vif <- vif(model\_5)

model\_5\_vif

model\_6 <- lm(loss ~ logweeks2 + covid + south\_asia, data = learnLossPredR)

summary(model\_6)

model\_6\_vif <- vif(model\_6)

model\_6\_vif

#substituting covid for covid25

model\_7 <- lm(loss ~ logweeks2 + covid25 + loggdp + south\_asia + advanced\_eco, data = learnLossPredR)

summary(model\_7)

model\_7\_vif <- vif(model\_7)

model\_7\_vif

#Schooling is also a good predictor of learning loss, from literature review, as higher years of schooling for adults can be of help to students and motivate then to learn

#Although it correlates with Log GDP we can test it effect in the equation.

model\_8 <- lm(loss ~ logweeks2 + covid25 + loggdp + south\_asia + advanced\_eco + schooling, data = learnLossPredR)

summary(model\_8)

model\_8\_vif <- vif(model\_8)

model\_8\_vif

#Model 8 didn't change much compared to model 7, let's add some interaction between schooling and advanced economy

learnLossPredR$adv\_sch\_intrt <- learnLossPredR$advanced\_eco \* learnLossPredR$schooling

View(learnLossPredR)

model\_9 <- lm(loss ~ logweeks2 + covid25 + loggdp + south\_asia + advanced\_eco + schooling+adv\_sch\_intrt, data = learnLossPredR)

summary(model\_9)

model\_9\_vif <- vif(model\_9)

model\_9\_vif

#model\_9 was lower in performance than model 8

#removing schooling variable and using only interaction

View(learnLossPredR)

model\_10 <- lm(loss ~ logweeks2 + covid25 + loggdp + south\_asia + advanced\_eco+adv\_sch\_intrt, data = learnLossPredR)

summary(model\_10)

model\_10\_vif <- vif(model\_10)

model\_10\_vif

#Model 10 is still subpar, trying another interaction with schooling and loggdp

learnLossPredR$sch\_lgdp\_intrt <- learnLossPredR$schooling \* learnLossPredR$loggdp

model\_11 <- lm(loss ~ logweeks2 + covid25 + loggdp + south\_asia + advanced\_eco+schooling+sch\_lgdp\_intrt, data = learnLossPredR)

summary(model\_11)

model\_11\_vif <- vif(model\_11)

model\_11\_vif

#model 11 is 0.45

#Model with interaction of schooling and loggdp minus schooling or gdp

model\_12 <- lm(loss ~ logweeks2 + covid25 + south\_asia + advanced\_eco+sch\_lgdp\_intrt, data = learnLossPredR)

summary(model\_12)

model\_12\_vif <- vif(model\_12)

model\_12\_vif

#model\_12 gives a shows an improvement with adjusted rsquare of 0.467 and all vif factors less than 5

#model 13 interaction with logweeks2 and south\_asia

learnLossPredR$lweeks\_asia\_intrt <- learnLossPredR$logweeks2 \* learnLossPredR$south\_asia

model\_13 <- lm(loss ~ logweeks2 + covid25 + loggdp + south\_asia + advanced\_eco+lweeks\_asia\_intrt+sch\_lgdp\_intrt, data = learnLossPredR)

summary(model\_13)

model\_13\_vif <- vif(model\_13)

model\_13\_vif

#the model gives an adjusted r-square of 0.5065,

#model 14 using hlo in place of covid because they are perfectly correlated

model\_14 <- lm(loss ~ logweeks2 + hlo + loggdp + south\_asia + advanced\_eco+lweeks\_asia\_intrt+sch\_lgdp\_intrt, data = learnLossPredR)

summary(model\_14)

model\_14\_vif <- vif(model\_14)

model\_14\_vif

#Trying the principle of parsimony, taking out loggdp ,hlo, advanced economy

model\_5 <- lm(loss ~ logweeks2 + south\_asia + lweeks\_asia\_intrt + sch\_lgdp\_intrt, data = learnLossPredR)

summary(model\_5)

model\_15\_vif <- vif(model\_15)

model\_15\_vif

#New resid

model\_5\_resid <- resid(model\_5)

model\_5\_resid

plot(fitted(model\_5),model\_5\_resid)

abline(0,0)

plot(density(model\_5\_resid))

#New Resid

model\_4\_resid <- resid(model\_4)

model\_4\_resid

plot(fitted(model\_4),model\_4\_resid)

abline(0,0)

plot(density(model\_4\_resid))

residreg3 <- resid(model\_4)

residreg3

plot(fitted(model\_4),residreg3)

abline(0,0)

plot(density(residreg3))

#MODEL 6

model\_6\_resid <- resid(model\_6)

model\_6\_resid

plot(fitted(model\_6),model\_6\_resid)

abline(0,0)

plot(density(model\_6\_resid))

#Cross validating for model 5 and Model 6

TData <- learnLossPredR[1:29,]

VData <- learnLossPredR[30:41,]

View(TData)

#creating a new model with same vairables

model\_five\_train <- lm(loss ~ logweeks2 + south\_asia + lweeks\_asia\_intrt + sch\_lgdp\_intrt, data = TData)

summary(model\_five\_train)

model\_six\_train <- lm(loss ~ logweeks2 + covid + south\_asia, data = TData)

summary(model\_six\_train)

#predict

Pred1 <- predict(model\_five\_train, VData)

Pred1

sqrt(mean((VData$loss-Pred1)^2))

Pred2 <- predict(model\_six\_train, VData)

Pred2

sqrt(mean((VData$loss-Pred2)^2))