rm(list=ls())

library(readxl)

library(dplyr)

library(ggplot2)  
library(GGally)

# Read the dataset  
df <- read\_excel("learning\_loss\_covid\_data\_dictionary and database.xlsx", sheet = "Learning loss database")  
  
# Explore the dataset  
glimpse(df)

## Rows: 41  
## Columns: 24  
## $ country <chr> "Argentina", "Australia", "Bangladesh", "Belgium", "Botswa…  
## $ loss <dbl> 0.44, 0.01, 0.28, 0.23, 0.29, 0.32, 0.45, 0.17, 0.25, 0.22…  
## $ weeks <dbl> 39, 8, 63, 9, 11, 38, 40, 26, 14, 7, 23, 9, 8, 10, 21, 8, …  
## $ deaths <dbl> 288.00, 60.37, 17.84, 284.67, 118.60, 323.07, 18.28, 119.0…  
## $ gdp <dbl> 8585.690, 51680.300, 2270.350, 45189.400, 6348.830, 6814.8…  
## $ private <dbl> 26.08010, 29.53870, 23.90350, 54.17840, 6.93809, 18.65700,…  
## $ internet <dbl> 86, 90, 25, 92, 64, 81, 33, 97, 88, 70, 70, 91, 97, 55, 24…  
## $ hlo <dbl> 457.418, 544.360, 340.088, 340.088, 417.656, 465.452, 491.…  
## $ vac <dbl> 65.0681, 64.0000, 17.0000, 87.0000, 12.0000, 64.0000, 82.0…  
## $ stringency <dbl> 67.0396, 61.0000, 70.0000, 50.0000, 50.0000, 60.0000, 71.0…  
## $ union <dbl> 27.7, 13.7, 11.9, 49.1, 30.0, 13.0, 9.6, 29.4, 17.1, 44.3,…  
## $ polity <dbl> 9, 10, -6, 8, 8, 8, -4, 10, 10, -7, 7, 9, 10, 8, 1, 10, 10…  
## $ loggdp <dbl> 9.057853, 10.852830, 7.727688, 10.718620, 8.756026, 8.8268…  
## $ logweeks <dbl> 3.663562, NA, 4.143135, 2.197225, 2.397895, 3.637586, 3.68…  
## $ weeks2 <dbl> 1521, 64, 3969, 81, 121, 1444, 1600, 676, 196, 49, 529, 81…  
## $ pop <dbl> 4.5e+07, 2.6e+07, 1.7e+08, 1.1e+07, 2.3e+06, 2.1e+08, 1.6e…  
## $ schooling <dbl> 9.51, 11.54, 5.91, 10.69, 9.55, 7.89, 4.72, 12.32, 9.78, 7…  
## $ wbcode <chr> "ARG", "AUS", "BGD", "BEL", "BWA", "BRA", "KHM", "CAN", "C…  
## $ region\_code <chr> "Latin America and the Caribbean", "Advanced Economies", "…  
## $ hlo25 <dbl> 18.29670, 21.77440, 13.60350, 13.60350, 16.70620, 18.61810…  
## $ oecd <dbl> 0, 1, 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 0, 1…  
## $ covid <dbl> 413.418, 543.360, 312.088, 317.088, 388.656, 433.452, 446.…  
## $ high <dbl> 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 1, 1, 1, 0, 1…  
## $ covid25 <dbl> 16.53670, 21.73440, 12.48350, 12.68350, 15.54620, 17.33810…

summary(df)

## country loss weeks deaths   
## Length:41 Min. :0.0000 Min. : 0.00 Min. : 1.09   
## Class :character 1st Qu.:0.1400 1st Qu.: 9.00 1st Qu.: 38.32   
## Mode :character Median :0.2200 Median :13.00 Median :119.97   
## Mean :0.2268 Mean :20.78 Mean :152.36   
## 3rd Qu.:0.2900 3rd Qu.:28.00 3rd Qu.:260.63   
## Max. :0.8000 Max. :93.00 Max. :492.49   
##   
## gdp private internet hlo   
## Min. : 636.3 Min. : 0.3852 Min. :10.00 Min. :229.2   
## 1st Qu.: 3597.2 1st Qu.: 5.1648 1st Qu.:55.00 1st Qu.:389.6   
## Median :12956.6 Median :10.2287 Median :78.00 Median :483.4   
## Mean :21709.2 Mean :14.8429 Mean :69.44 Mean :462.1   
## 3rd Qu.:39918.2 3rd Qu.:19.9098 3rd Qu.:90.00 3rd Qu.:543.1   
## Max. :87100.4 Max. :62.6325 Max. :97.00 Max. :569.0   
##   
## vac stringency union polity   
## Min. : 1.905 Min. :42.25 Min. : 4.70 Min. :-7.000   
## 1st Qu.: 36.000 1st Qu.:49.00 1st Qu.:11.90 1st Qu.: 7.000   
## Median : 72.000 Median :55.00 Median :16.30 Median : 9.000   
## Mean : 59.769 Mean :56.29 Mean :22.38 Mean : 7.024   
## 3rd Qu.: 82.000 3rd Qu.:62.32 3rd Qu.:29.10 3rd Qu.:10.000   
## Max. :115.000 Max. :72.00 Max. :67.00 Max. :10.000   
##   
## loggdp logweeks weeks2 pop   
## Min. : 6.456 Min. :1.946 Min. : 0.0 Min. :2.300e+06   
## 1st Qu.: 8.188 1st Qu.:2.250 1st Qu.: 81.0 1st Qu.:1.600e+07   
## Median : 9.469 Median :2.639 Median : 169.0 Median :4.500e+07   
## Mean : 9.216 Mean :2.812 Mean : 755.2 Mean :1.293e+08   
## 3rd Qu.:10.595 3rd Qu.:3.367 3rd Qu.: 784.0 3rd Qu.:8.300e+07   
## Max. :11.375 Max. :4.533 Max. :8649.0 Max. :1.400e+09   
## NA's :2   
## schooling wbcode region\_code hlo25   
## Min. : 2.300 Length:41 Length:41 Min. : 9.167   
## 1st Qu.: 7.610 Class :character Class :character 1st Qu.:15.583   
## Median : 9.840 Mode :character Mode :character Median :19.336   
## Mean : 9.438 Mean :18.483   
## 3rd Qu.:11.590 3rd Qu.:21.724   
## Max. :13.180 Max. :22.761   
##   
## oecd covid high covid25   
## Min. :0.0000 Min. :204.2 Min. :0.0000 Min. : 8.167   
## 1st Qu.:0.0000 1st Qu.:372.6 1st Qu.:0.0000 1st Qu.:14.903   
## Median :1.0000 Median :447.6 Median :0.0000 Median :17.903   
## Mean :0.5854 Mean :439.4 Mean :0.4146 Mean :17.576   
## 3rd Qu.:1.0000 3rd Qu.:526.1 3rd Qu.:1.0000 3rd Qu.:21.044   
## Max. :1.0000 Max. :567.0 Max. :1.0000 Max. :22.680   
##

head(df)

## # A tibble: 6 × 24  
## country loss weeks deaths gdp private inter…¹ hlo vac strin…² union  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Argentina 0.44 39 288 8586. 26.1 86 457. 65.1 67.0 27.7  
## 2 Australia 0.01 8 60.4 51680. 29.5 90 544. 64 61 13.7  
## 3 Bangladesh 0.28 63 17.8 2270. 23.9 25 340. 17 70 11.9  
## 4 Belgium 0.23 9 285. 45189. 54.2 92 340. 87 50 49.1  
## 5 Botswana 0.29 11 119. 6349. 6.94 64 418. 12 50 30   
## 6 Brazil 0.32 38 323. 6815. 18.7 81 465. 64 60 13   
## # … with 13 more variables: polity <dbl>, loggdp <dbl>, logweeks <dbl>,  
## # weeks2 <dbl>, pop <dbl>, schooling <dbl>, wbcode <chr>, region\_code <chr>,  
## # hlo25 <dbl>, oecd <dbl>, covid <dbl>, high <dbl>, covid25 <dbl>, and  
## # abbreviated variable names ¹​internet, ²​stringency

tail(df)

## # A tibble: 6 × 24  
## country loss weeks deaths gdp private inter…¹ hlo vac strin…² union  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Spain 0.05 10 245. 27056. 31.5 93 528. 82 55 12.5  
## 2 Sweden 0 0 201. 52300. 11.0 95 555. 79 46 65.2  
## 3 Switzerla… 0.2 8 162. 87100. 5.50 94 555. 75 49 14.4  
## 4 Turkey 0.37 40 120. 3597. 4.98 57 483. 72 57 9.9  
## 5 Uganda 0.14 39 8 822. 20 20 360. 1.90 62.3 10.3  
## 6 United St… 0.22 8 323. 15419. 17.1 73 470. 89 57 10.3  
## # … with 13 more variables: polity <dbl>, loggdp <dbl>, logweeks <dbl>,  
## # weeks2 <dbl>, pop <dbl>, schooling <dbl>, wbcode <chr>, region\_code <chr>,  
## # hlo25 <dbl>, oecd <dbl>, covid <dbl>, high <dbl>, covid25 <dbl>, and  
## # abbreviated variable names ¹​internet, ²​stringency

# Correlation analysis  
cor(df[,c("loss", "weeks", "deaths", "gdp", "private", "internet", "hlo", "vac", "stringency", "polity", "loggdp", "logweeks", "weeks2", "pop", "schooling")])

## loss weeks deaths gdp private  
## loss 1.00000000 0.6439601 -0.087412823 -0.4534252 0.10111895  
## weeks 0.64396009 1.0000000 -0.203612072 -0.5104835 0.23548652  
## deaths -0.08741282 -0.2036121 1.000000000 0.1356103 0.20343067  
## gdp -0.45342521 -0.5104835 0.135610252 1.0000000 -0.12071436  
## private 0.10111895 0.2354865 0.203430675 -0.1207144 1.00000000  
## internet -0.43766538 -0.5442657 0.558927679 0.7332139 0.05848097  
## hlo -0.31095444 -0.2090989 0.438837120 0.6180698 -0.16748112  
## vac -0.32211449 -0.4157630 0.322254211 0.5248547 0.06408129  
## stringency 0.17720486 0.3763417 -0.009921679 -0.2360164 0.24829000  
## polity -0.22238324 -0.3581960 0.385626446 0.4559629 0.10282960  
## loggdp -0.51677001 -0.5774919 0.443629056 0.8798855 -0.04289585  
## logweeks NA NA NA NA NA  
## weeks2 0.52052535 0.9350358 -0.211247849 -0.3563483 0.30645092  
## pop 0.20355246 0.3977607 -0.225963681 -0.2228288 0.17636546  
## schooling -0.56504059 -0.6073859 0.460519732 0.6302510 -0.10045382  
## internet hlo vac stringency polity  
## loss -0.43766538 -0.31095444 -0.32211449 0.177204861 -0.2223832  
## weeks -0.54426566 -0.20909886 -0.41576302 0.376341737 -0.3581960  
## deaths 0.55892768 0.43883712 0.32225421 -0.009921679 0.3856264  
## gdp 0.73321386 0.61806977 0.52485467 -0.236016433 0.4559629  
## private 0.05848097 -0.16748112 0.06408129 0.248290001 0.1028296  
## internet 1.00000000 0.70597291 0.66514166 -0.099423899 0.5638881  
## hlo 0.70597291 1.00000000 0.65181825 0.121669072 0.3377571  
## vac 0.66514166 0.65181825 1.00000000 0.231166527 0.1777499  
## stringency -0.09942390 0.12166907 0.23116653 1.000000000 -0.2949984  
## polity 0.56388806 0.33775711 0.17774989 -0.294998405 1.0000000  
## loggdp 0.93668854 0.75088024 0.67409015 -0.105366364 0.5137149  
## logweeks NA NA NA NA NA  
## weeks2 -0.41450359 -0.09255770 -0.30497727 0.331800453 -0.2432483  
## pop -0.16861667 0.01689962 0.04593586 0.360121761 -0.3305298  
## schooling 0.80035296 0.59002387 0.62958443 -0.125179606 0.5556036  
## loggdp logweeks weeks2 pop schooling  
## loss -0.51677001 NA 0.5205253 0.20355246 -0.5650406  
## weeks -0.57749188 NA 0.9350358 0.39776067 -0.6073859  
## deaths 0.44362906 NA -0.2112478 -0.22596368 0.4605197  
## gdp 0.87988553 NA -0.3563483 -0.22282877 0.6302510  
## private -0.04289585 NA 0.3064509 0.17636546 -0.1004538  
## internet 0.93668854 NA -0.4145036 -0.16861667 0.8003530  
## hlo 0.75088024 NA -0.0925577 0.01689962 0.5900239  
## vac 0.67409015 NA -0.3049773 0.04593586 0.6295844  
## stringency -0.10536636 NA 0.3318005 0.36012176 -0.1251796  
## polity 0.51371495 NA -0.2432483 -0.33052985 0.5556036  
## loggdp 1.00000000 NA -0.4233097 -0.17566301 0.8007495  
## logweeks NA 1 NA NA NA  
## weeks2 -0.42330972 NA 1.0000000 0.55982222 -0.4562046  
## pop -0.17566301 NA 0.5598222 1.00000000 -0.2199156  
## schooling 0.80074949 NA -0.4562046 -0.21991559 1.0000000

# Scatter plot matrix  
ggpairs(df[,c("loss", "weeks", "deaths", "gdp", "private", "internet", "hlo", "vac", "stringency", "polity", "loggdp", "logweeks", "weeks2", "pop", "schooling")])

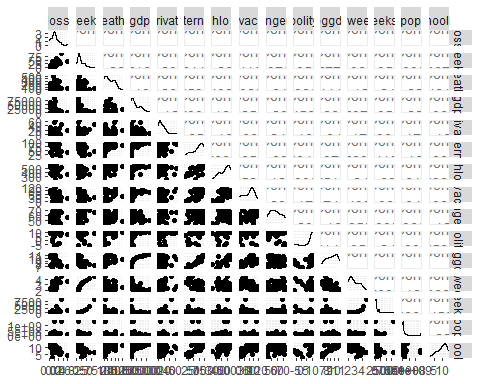
## Warning in ggally\_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removed 2 rows containing missing values  
  
## Warning in ggally\_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removed 2 rows containing missing values  
  
## Warning in ggally\_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removed 2 rows containing missing values  
  
## Warning in ggally\_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removed 2 rows containing missing values  
  
## Warning in ggally\_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removed 2 rows containing missing values  
  
## Warning in ggally\_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removed 2 rows containing missing values  
  
## Warning in ggally\_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removed 2 rows containing missing values  
  
## Warning in ggally\_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removed 2 rows containing missing values  
  
## Warning in ggally\_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removed 2 rows containing missing values  
  
## Warning in ggally\_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removed 2 rows containing missing values  
  
## Warning in ggally\_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removed 2 rows containing missing values

## Warning: Removed 2 rows containing missing values (geom\_point).  
## Removed 2 rows containing missing values (geom\_point).  
## Removed 2 rows containing missing values (geom\_point).  
## Removed 2 rows containing missing values (geom\_point).  
## Removed 2 rows containing missing values (geom\_point).  
## Removed 2 rows containing missing values (geom\_point).  
## Removed 2 rows containing missing values (geom\_point).  
## Removed 2 rows containing missing values (geom\_point).  
## Removed 2 rows containing missing values (geom\_point).  
## Removed 2 rows containing missing values (geom\_point).  
## Removed 2 rows containing missing values (geom\_point).

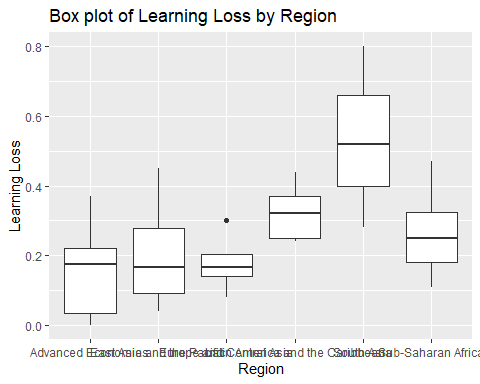
## Warning: Removed 2 rows containing non-finite values (stat\_density).

## Warning in ggally\_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removed 2 rows containing missing values  
  
## Warning in ggally\_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removed 2 rows containing missing values  
  
## Warning in ggally\_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removed 2 rows containing missing values

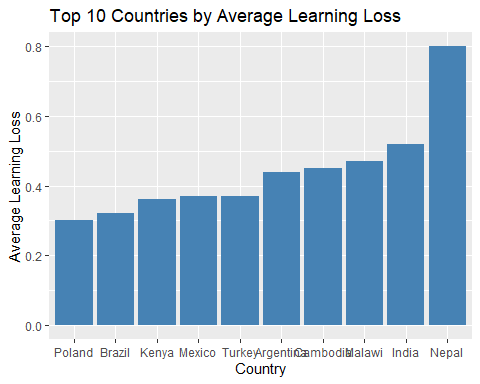
## Warning: Removed 2 rows containing missing values (geom\_point).  
## Removed 2 rows containing missing values (geom\_point).  
## Removed 2 rows containing missing values (geom\_point).



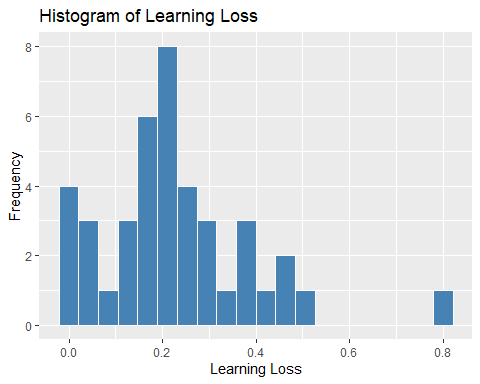
# Box plot of loss by region\_code  
ggplot(df, aes(x = region\_code, y = loss)) +  
 geom\_boxplot() +  
 ggtitle("Box plot of Learning Loss by Region") +  
 xlab("Region") +  
 ylab("Learning Loss")



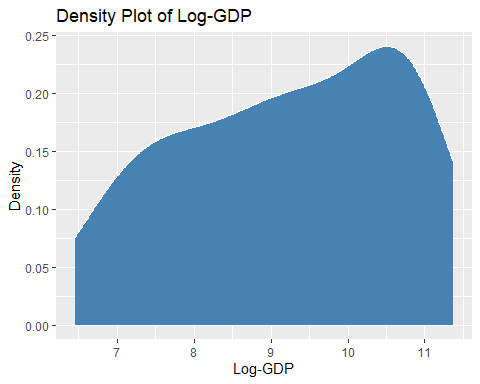
# Bar plot of average loss by country  
df\_avg\_loss <- df %>% group\_by(country) %>% summarize(avg\_loss = mean(loss, na.rm = TRUE)) %>% arrange(desc(avg\_loss))  
ggplot(df\_avg\_loss[1:10,], aes(x = reorder(country, avg\_loss), y = avg\_loss)) +  
 geom\_bar(stat = "identity", fill = "steelblue") +  
 ggtitle("Top 10 Countries by Average Learning Loss") +  
 xlab("Country") +  
 ylab("Average Learning Loss")



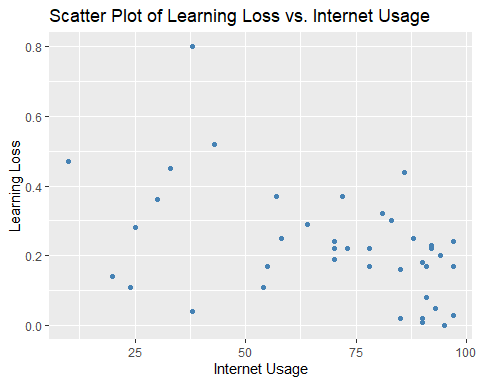
# Histogram of learning loss  
ggplot(df, aes(x = loss)) +  
 geom\_histogram(bins = 20, fill = "steelblue", color = "white") +  
 ggtitle("Histogram of Learning Loss") +  
 xlab("Learning Loss") +  
 ylab("Frequency")



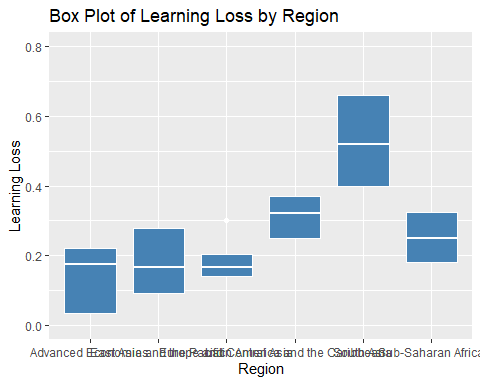
# Density plot of log-gdp  
ggplot(df, aes(x = loggdp)) +  
 geom\_density(fill = "steelblue", color = "white") +  
 ggtitle("Density Plot of Log-GDP") +  
 xlab("Log-GDP") +  
 ylab("Density")



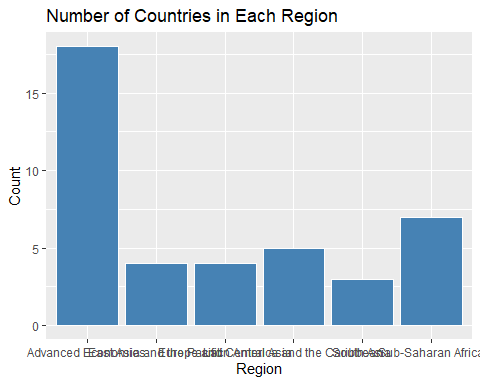
# Scatter plot of learning loss vs. internet usage  
ggplot(df, aes(x = internet, y = loss)) +  
 geom\_point(color = "steelblue") +  
 ggtitle("Scatter Plot of Learning Loss vs. Internet Usage") +  
 xlab("Internet Usage") +  
 ylab("Learning Loss")



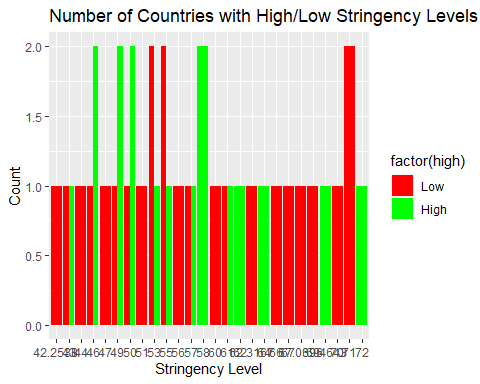
# Box plot of learning loss by region  
ggplot(df, aes(x = factor(region\_code), y = loss)) +  
 geom\_boxplot(fill = "steelblue", color = "white") +  
 ggtitle("Box Plot of Learning Loss by Region") +  
 xlab("Region") +  
 ylab("Learning Loss")



# Bar chart of number of countries in each region  
ggplot(df, aes(x = factor(region\_code))) +  
 geom\_bar(fill = "steelblue", color = "white") +  
 ggtitle("Number of Countries in Each Region") +  
 xlab("Region") +  
 ylab("Count")



# Bar chart of number of countries with high and low stringency levels  
ggplot(df, aes(x = factor(stringency), fill = factor(high))) +  
 geom\_bar(position = "dodge") +  
 ggtitle("Number of Countries with High/Low Stringency Levels") +  
 xlab("Stringency Level") +  
 ylab("Count") +  
 scale\_fill\_manual(values = c("red", "green"), labels = c("Low", "High"))



#Predictive Analytics Linear regression  
fit8 <- lm(loss ~ gdp + internet + hlo + vac + stringency + union + polity + loggdp + pop + schooling, data = df)  
summary(fit8)

##   
## Call:  
## lm(formula = loss ~ gdp + internet + hlo + vac + stringency +   
## union + polity + loggdp + pop + schooling, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.27945 -0.05299 -0.00044 0.07140 0.34239   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.715e-01 5.021e-01 1.935 0.0625 .  
## gdp 1.268e-06 2.759e-06 0.459 0.6492   
## internet 3.724e-03 3.291e-03 1.132 0.2668   
## hlo 2.579e-04 4.194e-04 0.615 0.5432   
## vac 1.114e-04 1.184e-03 0.094 0.9257   
## stringency 2.623e-03 3.232e-03 0.812 0.4234   
## union 1.231e-03 1.775e-03 0.693 0.4935   
## polity 5.699e-03 6.523e-03 0.874 0.3892   
## loggdp -1.211e-01 8.711e-02 -1.390 0.1747   
## pop 2.170e-11 8.561e-11 0.254 0.8016   
## schooling -2.742e-02 1.590e-02 -1.724 0.0949 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.14 on 30 degrees of freedom  
## Multiple R-squared: 0.4216, Adjusted R-squared: 0.2288   
## F-statistic: 2.187 on 10 and 30 DF, p-value: 0.04781

ggplot(df, aes(x = loggdp, y = loss)) +  
 geom\_point(color = "steelblue") +  
 geom\_smooth(method = "lm", se = FALSE) +  
 ggtitle("Scatter Plot of Learning Loss vs. Log-GDP with Linear Regression Line") +  
 xlab("Log-GDP") +  
 ylab("Learning Loss")

## `geom\_smooth()` using formula 'y ~ x'

