Results.Rmd

Foqia Shahid

12/17/2020

**Introduction**

Preface:

Gender inequality has been a long-standing issue in Pakistan and the lack of access to education for girls has been a key component of this broader issue. Recently, Pakistan ranked 151 out of 153 countries on the Global Gender Gap Index Report 2020. The report is a strong indicator of the limited opportunities for women in Pakistan. For my project, I wanted to to particularly examine the gender disparity prevalent in schools across Pakistan.

About the Data:

The data-set was compiled by Mesum Raza Hemani (Source: <https://www.kaggle.com/mesumraza/pakistan-education-performance-dataset>) and originally extracted from the surveys of Alif Ailaan & Aser Pakistan on Primary Schooling Performance in Pakistan. Alif Ailaan & Aser is an organization that collects data and information on children in Pakistan to obtain accurate estimates on their schooling and basic education (<http://aserpakistan.org/index.php>).

The data gives information about primary & secondary schooling in Pakistan. It includes demographic information about 145 cities including population, area in km^2, province etc. It also includes relevant information about the primary and secondary schools in the cities such as the number of schools, % Boys and girls enrolled etc.

For the purposes of this project, we will consider Percentage Disparity\* (percent boys enrolled - percent girls enrolled) as the dependent variable and examine the following as independent variables (predictor variables):

1)Population

2)Region\*

3)Total schools/ km^2\* (as a measure of proximity of schools)

4)School Infrastructure Score (Arithmetic average of electricity, drinking water, toilet availability, boundary wall and satisfactory condition of school building indicators. Each of the 5 indicators carry equal weightage)

5)Learning score (derives by averaging the literacy rate and basic competence in reading and writing)

6)Enrollment score (derived from the underlying net access data from government agencies)

\*Variables that were created from the data for this project

Context:

We will explore the relationship of Percentage Disparity and the predictor variables as mentioned above. In doing so, we may expect to find negative correlations between Percentage Disparity and Population, School Infrastructure Score, Learning Score, Enrollment Score and Schools/km^2. We may also find varying disparity across different regions. Some factors may carry more weightage in our model than others. Particularly, Total schools/ km^2, Learning score, Enrollment score and School Infrastructure Score may be more strongly related to Percentage Disparity.

**Step 0: Install packages that we will need**

library(readr)  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──

## ✓ ggplot2 3.3.2 ✓ dplyr 1.0.2  
## ✓ tibble 3.0.3 ✓ stringr 1.4.0  
## ✓ tidyr 1.1.2 ✓ forcats 0.5.0  
## ✓ purrr 0.3.4

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

**Step 1: The Data**

PakEdu <- read\_csv("Consolidated (Educational Dataset).csv")

## Parsed with column specification:  
## cols(  
## .default = col\_double(),  
## `% Boys Enrolled` = col\_character(),  
## `% Girls Enrolled` = col\_character(),  
## `All Four Facilities` = col\_character(),  
## `Any One Facility` = col\_character(),  
## `Any Three Facilities` = col\_character(),  
## `Any Two Facilities` = col\_character(),  
## `Boundary wall, Building condition satisfactory, Drinking water and 2 more (clusters)` = col\_character(),  
## `Show Sheet` = col\_character(),  
## `Table of Contents` = col\_character(),  
## `Other Factors Measure Value` = col\_number(),  
## `Analysis Level Selector` = col\_character(),  
## `Color By Measure Name` = col\_character(),  
## Country = col\_character(),  
## City = col\_character(),  
## `Educational Budget Spend of GDP` = col\_character(),  
## `No Facility` = col\_character(),  
## `Pakistan Economic Growth` = col\_character(),  
## Province = col\_character()  
## )

## See spec(...) for full column specifications.

head(PakEdu)

## # A tibble: 6 x 51  
## `% Boys Enrolle… `% Complete Pri… `% Girls Enroll… `% Primary Scho…  
## <chr> <dbl> <chr> <dbl>  
## 1 54.77% 0.937 45.23% 0.0346  
## 2 62.50% 0.594 37.50% 0.360   
## 3 86.63% 0.939 13.37% 0.0215  
## 4 60.77% 0.827 39.23% 0.0419  
## 5 60.75% 0.592 39.25% 0.347   
## 6 58.48% 0.582 41.52% 0.174   
## # … with 47 more variables: `% Primary Schools with single teacher` <dbl>, `All  
## # Four Facilities` <chr>, `Any One Facility` <chr>, `Any Three  
## # Facilities` <chr>, `Any Two Facilities` <chr>, `Area (km²)` <dbl>, `Bomb  
## # Blasts Occurred` <dbl>, `Boundary wall, Building condition satisfactory,  
## # Drinking water and 2 more (clusters)` <chr>, `Boundary wall` <dbl>,  
## # `Building condition satisfactory` <dbl>, `MeasureGroup 1 Measures` <dbl>,  
## # `Color By Measure Value` <dbl>, `Show Sheet` <chr>, `Table of  
## # Contents` <chr>, `Other Factors Measure Value` <dbl>, `Analysis Level  
## # Selector` <chr>, `Color By Measure Name` <chr>, `Complete Primary  
## # Schools` <dbl>, Country <chr>, City <chr>, `Drinking water` <dbl>, `Drone  
## # attacks in Pakistan` <dbl>, `Education score` <dbl>, `Educational Budget  
## # Spend of GDP` <chr>, Electricity <dbl>, `Enrolment score` <dbl>, `Gender  
## # parity score` <dbl>, `Global Terrorism Index - Pakistan` <dbl>, `Learning  
## # score` <dbl>, `MeasureGroup 2 Measures` <dbl>, `No Facility` <chr>, `Number  
## # of Records` <dbl>, `Number of primary schools as % of total schools` <dbl>,  
## # `Number of primary schools` <dbl>, `Number of secondary schools as % of  
## # total schools` <dbl>, `Number of secondary schools` <dbl>, `Pakistan  
## # Economic Growth` <chr>, Population <dbl>, `Primary Schools with single  
## # classroom` <dbl>, `Primary Schools with single teacher` <dbl>,  
## # Province <chr>, `Retention score` <dbl>, `School infrastructure  
## # score` <dbl>, `Terrorist Attacks Affectees` <dbl>, Toilet <dbl>, `Total  
## # number of schools` <dbl>, Year <dbl>

**Step 2: Data Wrangling**

Eliminating variables that I will not be using in my model and creating a new dataset with renamed and reordered variables.

#Create a dataframe with province data  
PakEdu.province <- select(PakEdu, "City", province = "Province")  
PakEdu.province <- unique(PakEdu.province)  
PakEdu.province <- mutate(PakEdu.province, province = as.factor(province))  
  
PakEdu.data <- PakEdu %>%  
 select("City", "% Boys Enrolled", "% Girls Enrolled","Province", "Population", "Total number of schools", "Year", "School infrastructure score", "Area (km²)", "Enrolment score", "Learning score") %>%  
 mutate(Percent\_Boys\_Enrolled = as.numeric(gsub("[\\%,]", "", PakEdu$`% Boys Enrolled`))) %>%  
 mutate(Percent\_Girls\_Enrolled = as.numeric(gsub("[\\%,]", "", PakEdu$`% Girls Enrolled`))) %>%  
 mutate(Total\_Schools = as.numeric(as.character(PakEdu$`Total number of schools`))) %>%  
 mutate(Area = as.numeric(as.character(PakEdu$`Area (km²)`))) %>%  
 mutate(Learning\_score = as.numeric(as.character(PakEdu$`Learning score`))) %>%  
 mutate(Enrolment\_score = as.numeric(as.character(PakEdu$`Enrolment score`))) %>%  
 mutate(School\_infrastructure\_score = as.numeric(as.character(PakEdu$`School infrastructure score`))) %>%  
 mutate(Percentage\_Disparity = Percent\_Boys\_Enrolled - Percent\_Girls\_Enrolled) %>%  
 mutate (Province = as.factor(Province))  
  
#Remove bad data points as neceassary  
PakEdu.data <- filter(PakEdu.data, Area > 0)  
PakEdu.data <- filter(PakEdu.data, Population > 0)  
PakEdu.data <- filter(PakEdu.data, Population < 13215631)  
PakEdu.data <- mutate(PakEdu.data, Schools\_per\_Area = Total\_Schools / Area)  
PakEdu.data <- PakEdu.data %>%  
 group\_by(City) %>% #so that data on cities repeated over years is averaged  
 summarise(  
 percentage\_disparity = mean(Percentage\_Disparity, na.rm=TRUE),  
 population = mean(Population),  
 school\_infrastructure\_score = mean(School\_infrastructure\_score, na.rm=TRUE),   
 schools\_per\_area = mean(Schools\_per\_Area),  
 learning\_score = mean(Learning\_score),   
 enrolment\_score = mean(Enrolment\_score)  
 )

## `summarise()` ungrouping output (override with `.groups` argument)

#Adding variable "Province" to data  
PakEdu.data <- PakEdu.data %>%   
 full\_join(PakEdu.province, by="City")  
  
#Creating a category for province: We will group provinces together, dividing Pakistan into Northern, Central and Southern regions.   
PakEdu.data <- PakEdu.data %>% mutate(region = case\_when(  
 province == "AJK" ~ "northern\_pakistan",  
 province == "Balochistan" ~ "southern\_pakistan",  
 province == "FATA" ~ "northern\_pakistan",  
 province == "GB" ~ "northern\_pakistan",  
 province == "KP" ~ "northern\_pakistan",   
 province == "Punjab" ~ "central\_pakistan",   
 province == "Sindh" ~ "southern\_pakistan",  
 province == "ICT" ~ "central\_pakistan" ))  
  
#Make region a factor  
PakEdu.data$region <- factor(PakEdu.data$region, c("northern\_pakistan",   
 "central\_pakistan",   
 "southern\_pakistan"))  
  
#Final DataSet  
PakEdu.data

## # A tibble: 145 x 9  
## City percentage\_disp… population school\_infrastr… schools\_per\_area  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 Abbo… 27.0 880666 59.2 0.928   
## 2 Atto… 15.3 1274935 87.5 0.188   
## 3 Awar… 70.8 118173 24.0 0.0213  
## 4 Badin 53.2 1136044 42.8 0.454   
## 5 Bagh 14.3 351415 27.9 0.783   
## 6 Baha… 22.5 2061447 80.2 0.271   
## 7 Baha… 27.0 2433091 86.2 0.0795  
## 8 Bark… 86.5 103545 19.7 0.173   
## 9 Bata… 65.7 307278 39.1 0.628   
## 10 Bhak… 40.8 1051456 82.2 0.165   
## # … with 135 more rows, and 4 more variables: learning\_score <dbl>,  
## # enrolment\_score <dbl>, province <fct>, region <fct>

**Step 3: Exploratory Univariate Data Analysis**

1. Explore the distribution of each variable using the appropariate numerical and graphical techniques. Is the distribution normal or skewed? What does it tell you about the behavior? Is this distribution what you expected to see? Why, or why not?

#getting numerical summary information for each variable.  
summary(PakEdu.data)

## City percentage\_disparity population   
## Length:145 Min. : 8.043 Min. : 33340   
## Class :character 1st Qu.:25.345 1st Qu.: 301633   
## Mode :character Median :40.547 Median : 834094   
## Mean :42.551 Mean :1046962   
## 3rd Qu.:56.355 3rd Qu.:1425572   
## Max. :98.045 Max. :6318745   
## NA's :24 NA's :24   
## school\_infrastructure\_score schools\_per\_area learning\_score enrolment\_score  
## Min. :14.54 Min. :0.005766 Min. :19.27 Min. :16.98   
## 1st Qu.:31.59 1st Qu.:0.093839 1st Qu.:40.02 1st Qu.:55.95   
## Median :57.71 Median :0.264823 Median :50.30 Median :67.92   
## Mean :56.59 Mean :0.317861 Mean :49.78 Mean :65.97   
## 3rd Qu.:83.09 3rd Qu.:0.440260 3rd Qu.:59.59 3rd Qu.:77.00   
## Max. :93.08 Max. :1.228916 Max. :72.04 Max. :95.07   
## NA's :24 NA's :24 NA's :24 NA's :24   
## province region   
## Punjab :36 northern\_pakistan:55   
## Balochistan:30 central\_pakistan :37   
## KP :25 southern\_pakistan:53   
## Sindh :23   
## FATA :13   
## AJK :10   
## (Other) : 8

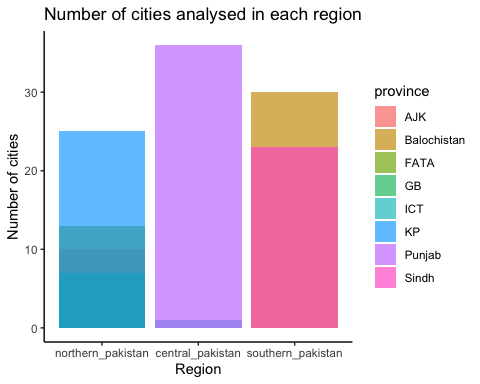
table(PakEdu.data$province)

##   
## AJK Balochistan FATA GB ICT KP   
## 10 30 13 7 1 25   
## Punjab Sindh   
## 36 23

summary(PakEdu.data$region)

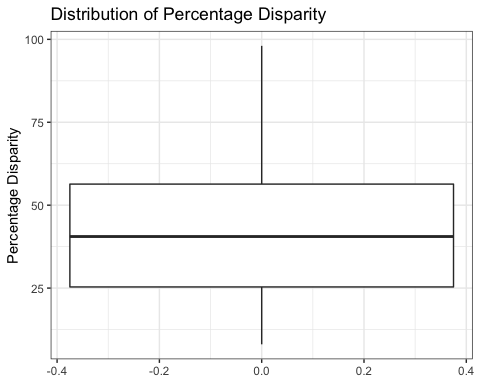
## northern\_pakistan central\_pakistan southern\_pakistan   
## 55 37 53

ggplot(data = PakEdu.data, mapping = aes(x = region, fill = province)) +   
 geom\_bar(alpha = 2/3, position = "identity") +  
 labs(title = "Number of cities analysed in each region", x = "Region", y = "Number of cities") + theme\_classic()



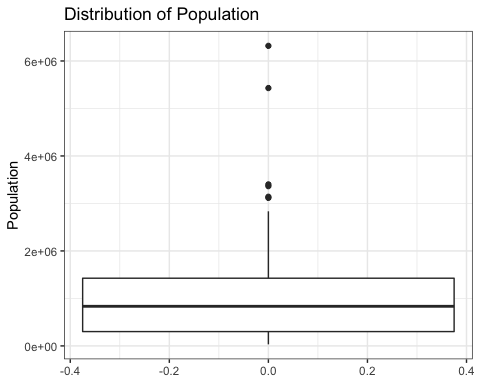
#Distribution of percentage\_disparity  
ggplot(PakEdu.data, aes(y = percentage\_disparity)) +  
 geom\_boxplot() +  
 labs(title = "Distribution of Percentage Disparity", y = "Percentage Disparity") + theme\_bw()

## Warning: Removed 24 rows containing non-finite values (stat\_boxplot).



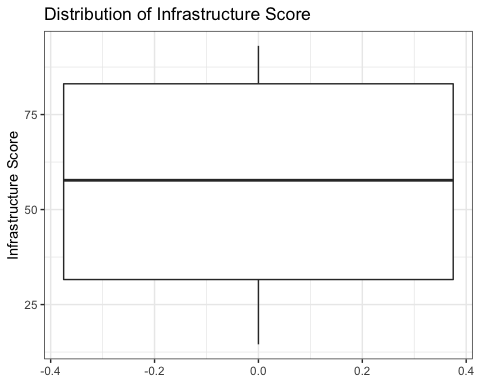
#Distribution of Population  
ggplot(PakEdu.data, aes(y = population)) +  
 geom\_boxplot() +  
 labs(title = "Distribution of Population", y = "Population") + theme\_bw()

## Warning: Removed 24 rows containing non-finite values (stat\_boxplot).



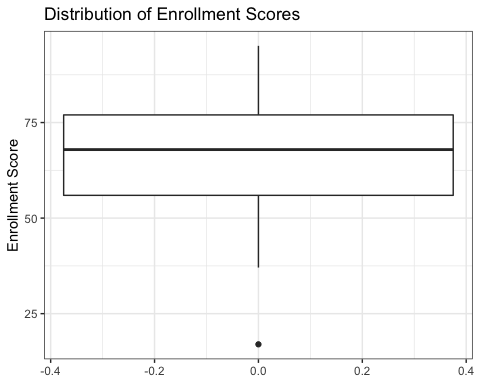
#Distribution of school\_infrastructure\_score  
ggplot(PakEdu.data, aes(y = school\_infrastructure\_score)) +  
 geom\_boxplot()+  
 labs(title = "Distribution of Infrastructure Score", y = "Infrastructure Score") + theme\_bw()

## Warning: Removed 24 rows containing non-finite values (stat\_boxplot).



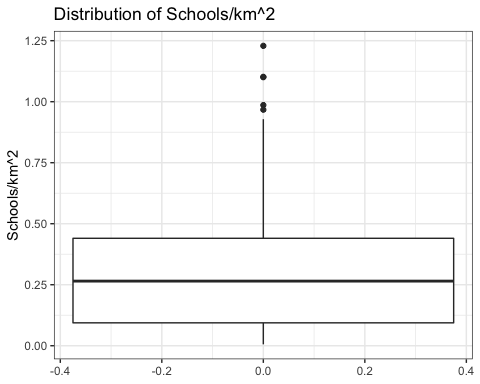
#Distribution of enrolment\_score  
ggplot(PakEdu.data, aes(y = enrolment\_score)) +  
 geom\_boxplot()+  
 labs(title = "Distribution of Enrollment Scores", y = "Enrollment Score") + theme\_bw()

## Warning: Removed 24 rows containing non-finite values (stat\_boxplot).



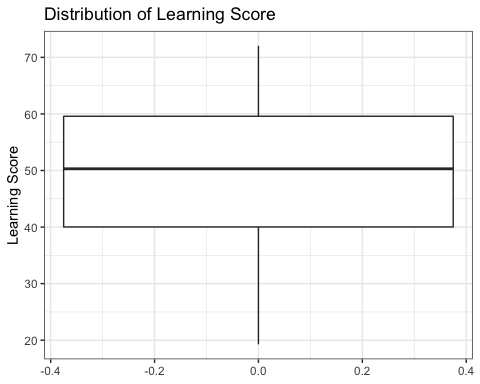
#Distribution of schools\_per\_area   
ggplot(PakEdu.data, aes(y = schools\_per\_area)) +  
 geom\_boxplot()+  
 labs(title = "Distribution of Schools/km^2", y = "Schools/km^2") + theme\_bw()

## Warning: Removed 24 rows containing non-finite values (stat\_boxplot).



#Distribution of learning\_score  
ggplot(PakEdu.data, aes(y = learning\_score)) +  
 geom\_boxplot()+  
 labs(title = "Distribution of Learning Score", y = "Learning Score") + theme\_bw()

## Warning: Removed 24 rows containing non-finite values (stat\_boxplot).

 Findings:

-Central Pakistan has the highest number of cities

-Large outliers are effecting the distribution of Population

-Scores are relatively normally distributed

**Step 4: Exploratory Bivariate Data Analysis**

1. Exploring the relationship between each predictor variable and outcome variable using numerical and graphical techniques. Do you see evidence of non-linearity? Do you see evidence of high leverage outliers?
2. In addition, we will want to examine the relationship between the predictor variables to confirm that we do not have issues with multicollinearity.

#Correlation table  
cor(PakEdu.data[,2:7], use = "pairwise.complete.obs")

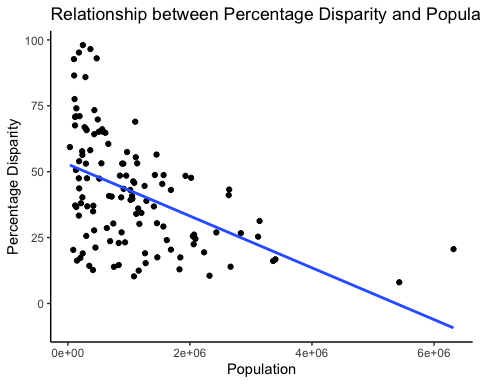
## percentage\_disparity population  
## percentage\_disparity 1.0000000 -0.4654723  
## population -0.4654723 1.0000000  
## school\_infrastructure\_score -0.5084837 0.6245886  
## schools\_per\_area -0.3093819 0.2288219  
## learning\_score -0.6789601 0.4779587  
## enrolment\_score -0.7336718 0.3131249  
## school\_infrastructure\_score schools\_per\_area  
## percentage\_disparity -0.5084837 -0.3093819  
## population 0.6245886 0.2288219  
## school\_infrastructure\_score 1.0000000 0.3022317  
## schools\_per\_area 0.3022317 1.0000000  
## learning\_score 0.5749361 0.2735892  
## enrolment\_score 0.5172622 0.3231786  
## learning\_score enrolment\_score  
## percentage\_disparity -0.6789601 -0.7336718  
## population 0.4779587 0.3131249  
## school\_infrastructure\_score 0.5749361 0.5172622  
## schools\_per\_area 0.2735892 0.3231786  
## learning\_score 1.0000000 0.5789248  
## enrolment\_score 0.5789248 1.0000000

#Check linearity  
ggplot(PakEdu.data, aes(population, percentage\_disparity)) +  
 geom\_point() + geom\_smooth(method = "lm", se = FALSE) + theme\_classic() +  
 labs(y = "Percentage Disparity", x = "Population", title = "Relationship between Percentage Disparity and Population")

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

## Warning: Removed 24 rows containing non-finite values (stat\_smooth).

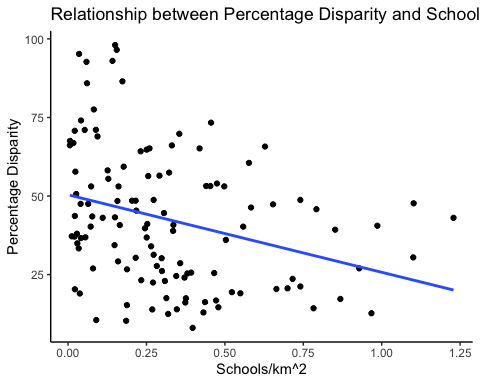
## Warning: Removed 24 rows containing missing values (geom\_point).



ggplot(PakEdu.data, aes(schools\_per\_area, percentage\_disparity)) +  
 geom\_point() + geom\_smooth(method = "lm", se = FALSE) + theme\_classic() +  
 labs(y = "Percentage Disparity", x = "Schools/km^2", title = "Relationship between Percentage Disparity and Schools/ km^2")

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

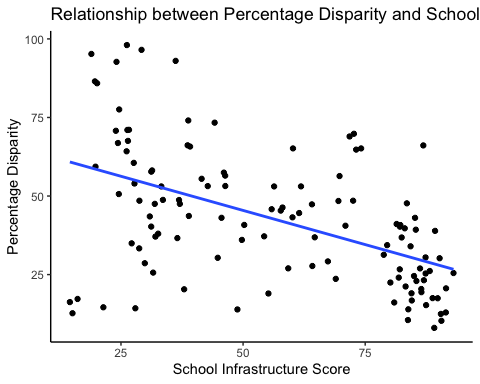
## Warning: Removed 24 rows containing non-finite values (stat\_smooth).  
  
## Warning: Removed 24 rows containing missing values (geom\_point).



ggplot(PakEdu.data, aes(school\_infrastructure\_score, percentage\_disparity)) +  
 geom\_point() + geom\_smooth(method = "lm", se = FALSE) + theme\_classic() +  
 labs(y = "Percentage Disparity", x = "School Infrastructure Score", title = "Relationship between Percentage Disparity and School Infrastructure scores")

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

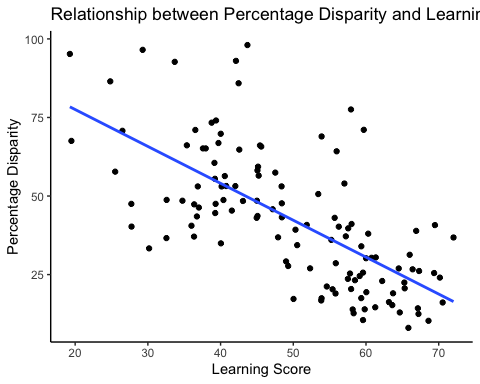
## Warning: Removed 24 rows containing non-finite values (stat\_smooth).  
  
## Warning: Removed 24 rows containing missing values (geom\_point).



ggplot(PakEdu.data, aes(learning\_score, percentage\_disparity)) +  
 geom\_point() + geom\_smooth(method = "lm", se = FALSE) + theme\_classic() +  
 labs(y = "Percentage Disparity", x = "Learning Score", title = "Relationship between Percentage Disparity and Learning Scores")

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

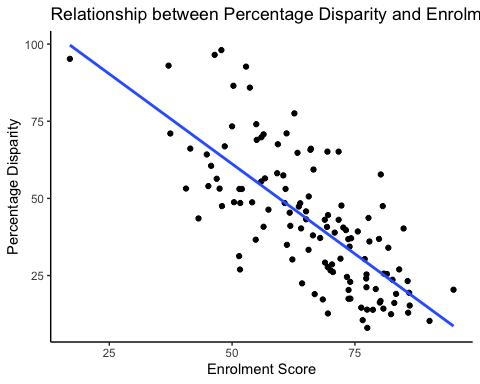
## Warning: Removed 24 rows containing non-finite values (stat\_smooth).  
  
## Warning: Removed 24 rows containing missing values (geom\_point).



ggplot(PakEdu.data, aes(enrolment\_score, percentage\_disparity)) +  
 geom\_point() + geom\_smooth(method = "lm", se = FALSE) + theme\_classic() +  
 labs(y = "Percentage Disparity", x = "Enrolment Score", title = "Relationship between Percentage Disparity and Enrolment Scores")

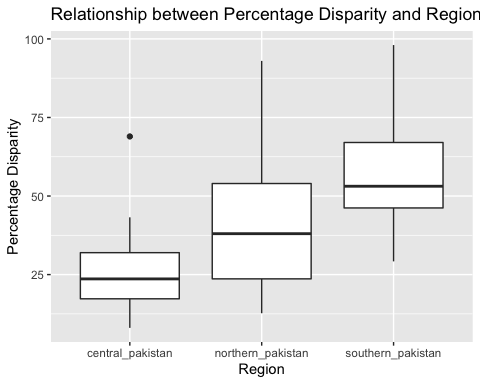
## `geom\_smooth()` using formula 'y ~ x'

## Warning: Removed 24 rows containing non-finite values (stat\_smooth).  
  
## Warning: Removed 24 rows containing missing values (geom\_point).



ggplot(PakEdu.data, aes(x = reorder(region, percentage\_disparity, na.rm = TRUE), y = percentage\_disparity)) +  
 geom\_boxplot() +  
 labs(y = "Percentage Disparity", x = "Region", title = "Relationship between Percentage Disparity and Region")

## Warning: Removed 24 rows containing non-finite values (stat\_boxplot).

 Findings:

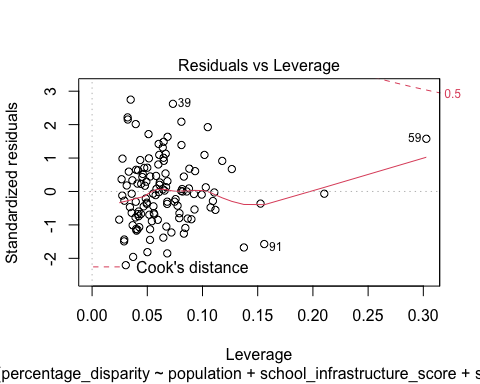
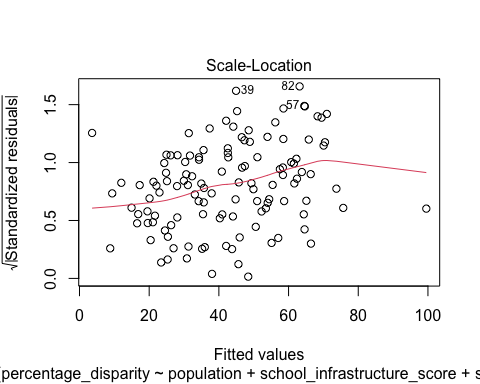
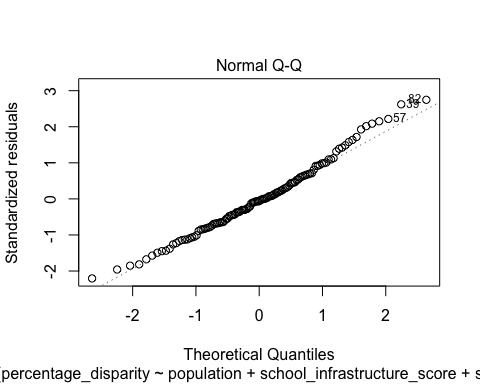
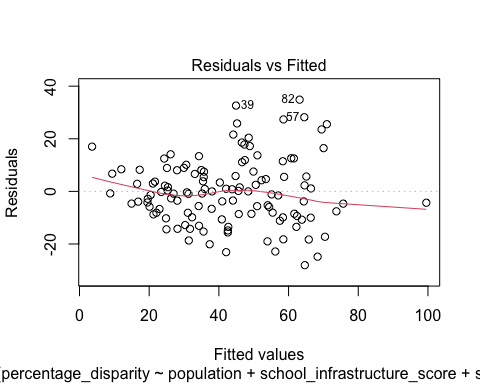
-There is evidence of low/medium/strong linear relationships between percentage disparity and the predictor variables -No evidence of multicollinearity between predictor variables

-1 high-leverage outliers (Population outlier of city Karachi was a high-leverage outlier as confirmed by leverage analysis and thus was removed previously)

* Percentage Disparity is negatively correlated with Population, School Infrastructure Score, Learning Score, Enrollment Score, which confirms out hypothesis. We also found Pecentage Disparity to increase as region changes from Central to Northern to Southern Pakistan. it is interesting to see that Southern Pakistan has higher gender disparity than Northern Pakistan considering the demographics of the regions.

**Step 5: Fitting the Data with lm() and Checking the Assumptions of “Linear” models**

#Ajusting baseline for Region  
PakEdu.data$region <- relevel(PakEdu.data$region, ref="northern\_pakistan")  
fit\_1 <- lm(percentage\_disparity ~ population + school\_infrastructure\_score + schools\_per\_area + learning\_score + enrolment\_score + region, data = PakEdu.data)  
plot(fit\_1)

 Findings:

#Residual vs Fitted:

Mean of residuals is approximately zero, and the residuals are scattered randomly around the horizontal line. This is a good indicator we dont have a non-linear relationship

#Normal Q-Q:

Residuals line up well on straight dashed line. Thus, we can assume residuals are normally distributed.

#Scale-Location:

While there isn’t a completely horizontal line, no apparent pattern of residuals are seen. We will thus assume equal variance.

#Residuals vs Leverage:

No residuals lay outside of Cooks distance so no outliers are influential to the regression results.

**Step 6: Evaluating the model.**

summary(fit\_1)

##   
## Call:  
## lm(formula = percentage\_disparity ~ population + school\_infrastructure\_score +   
## schools\_per\_area + learning\_score + enrolment\_score + region,   
## data = PakEdu.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -28.051 -8.542 -0.774 7.531 34.876   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.180e+02 1.006e+01 11.731 < 2e-16 \*\*\*  
## population -3.893e-06 1.710e-06 -2.277 0.02469 \*   
## school\_infrastructure\_score 1.028e-01 8.413e-02 1.222 0.22423   
## schools\_per\_area -2.774e+00 5.541e+00 -0.501 0.61764   
## learning\_score -4.792e-01 1.545e-01 -3.101 0.00243 \*\*   
## enrolment\_score -8.054e-01 1.139e-01 -7.075 1.34e-10 \*\*\*  
## regioncentral\_pakistan -2.148e+00 5.075e+00 -0.423 0.67282   
## regionsouthern\_pakistan 3.279e+00 3.816e+00 0.859 0.39200   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 12.92 on 113 degrees of freedom  
## (24 observations deleted due to missingness)  
## Multiple R-squared: 0.6608, Adjusted R-squared: 0.6398   
## F-statistic: 31.45 on 7 and 113 DF, p-value: < 2.2e-16

Findings:

Model equation:

Percentage-Disparity-predicted = (1.180e+02) + -3.893e-06(Population) + 1.028e-01(school infrastructure score) + -2.774e+00(schools/km^2) + -4.792e-01(learning score) + -8.054e-01(enrollment score) + -2.148e+00(if\_in\_central\_pakistan) + 3.279e+00(if\_in\_southern\_pakistan)

1.This model is a good fits as it significantly predicts the outcome variable.

-We have a high-adjusted R^2 (0.6398). This is a large effect size according to Cohen’s guide. It suggests that 64% of the variability in the city’s percentage disparity can be explained (predicted) by the predictor variables in this model (Population, School Infrastructure Score, Schools/km^2, Learning Score, Enrollment Score, Region) -The F-statistic also has a p value of essentially 0 i.e. F(7, 113) = 31.45, p < 0.01.

1. Enrollment scores, Learning Score and Population are significant predictors in our model.

3.Enrollment score has a larger predicted than Learning Score and Population

4.Using Model (fit\_1):

What would be the predicted percentage disparity for city Abbottabad?

province\_abb <- c('northern\_pakistan')  
  
new <- data.frame(population = 880666, school\_infrastructure\_score = 59.24411, schools\_per\_area = 0.9283172, learning\_score = 52.31015, enrolment\_score = 84.00923, region = province\_abb)  
new

## population school\_infrastructure\_score schools\_per\_area learning\_score  
## 1 880666 59.24411 0.9283172 52.31015  
## enrolment\_score region  
## 1 84.00923 northern\_pakistan

predict(fit\_1, new, interval = "predict")

## fit lwr upr  
## 1 25.37504 -1.070763 51.82084

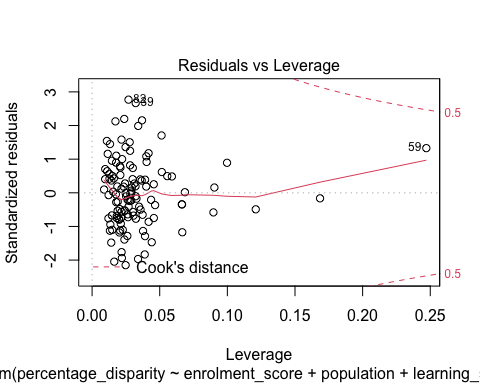
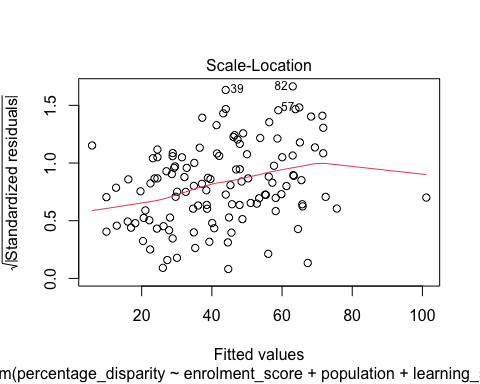
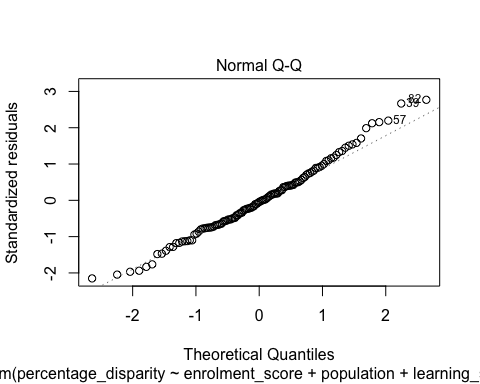
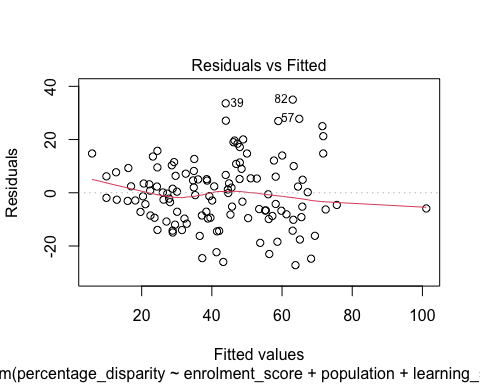
Predicted percentage gender disparity in Abbottabad: 25.37504

Actual percentage gender disparity in Abbottabad: 26.9875

**Step 7:Testing a Revised Model: CONTROL MODEL**

For this model, remove school\_infrastructure\_score, schools\_per\_area and region since these aren’t significant.

#We will try another model with the significant predictors i.e. enrolment\_score, learning score and population  
fit\_2 <- lm(percentage\_disparity ~ enrolment\_score + population + learning\_score, data = PakEdu.data)  
plot(fit\_2)

 Findings:

#Residual vs Fitted:

Mean of residuals is approximately zero, and the residuals are scattered randomly around the horizontal line. This is a good indicator we dont have a non-linear relationship

#Normal Q-Q:

Residuals line up well on straight dashed line. Thus, we can assume residuals are normally distributed.

#Scale-Location:

While there isn’t a completely horizontal line, no apparent pattern of residuals are seen. We will thus assume equal variance.

#Residuals vs Leverage:

No residuals lay outside of Cooks distance so no outliers are influential to the regression results.

summary(fit\_2)

##   
## Call:  
## lm(formula = percentage\_disparity ~ enrolment\_score + population +   
## learning\_score, data = PakEdu.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -27.208 -8.550 -0.575 7.160 34.995   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.257e+02 6.206e+00 20.253 < 2e-16 \*\*\*  
## enrolment\_score -8.005e-01 1.060e-01 -7.551 1.04e-11 \*\*\*  
## population -3.355e-06 1.307e-06 -2.567 0.0115 \*   
## learning\_score -5.388e-01 1.247e-01 -4.319 3.30e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 12.81 on 117 degrees of freedom  
## (24 observations deleted due to missingness)  
## Multiple R-squared: 0.6549, Adjusted R-squared: 0.6461   
## F-statistic: 74.02 on 3 and 117 DF, p-value: < 2.2e-16

Findings:

Model equation:

Percentage-Disparity-predicted = (1.257e+02) + -3.355e-06(Population) + -5.388e-01(learning score) + -8.005e-01(enrollment score)

1.This model is a good fits as it significantly predicts the outcome variable.

-We have a high-adjusted R^2 (0.6461). This is a large effect size according to Cohen’s guide. It suggests that 65% of the variability in the city’s percentage disparity can be explained (predicted) by the predictor variables in this model (Population, Learning Score, Enrollment Score)

-The F-statistic has a p value of essentially 0 i.e. F(3, 117) = 74.02, p < 0.01.

2.Enrollment scores, Learning Score and Population are significant predictors in this model as well.

3.Enrollment score has a larger predicted impact than any of the provinces.

4.Using Model (fit\_2):

What would be the predicted percentage disparity for city Abbottabad?

province\_d <- c('KP')  
  
new <- data.frame(population = 880666, school\_infrastructure\_score = 59.24411, schools\_per\_area = 0.9283172, learning\_score = 52.31015, enrolment\_score = 84.00923, province = province\_d)  
new

## population school\_infrastructure\_score schools\_per\_area learning\_score  
## 1 880666 59.24411 0.9283172 52.31015  
## enrolment\_score province  
## 1 84.00923 KP

predict(fit\_2, new, interval = "predict")

## fit lwr upr  
## 1 27.30702 1.5812 53.03285

(P1)Predicted percentage gender disparity in Abbottabad using fit\_1: 25.37504

(P2)Predicted percentage gender disparity in Abbottabad using fit\_2: 27.30702

(A)Actual percentage gender disparity in Abbottabad: 26.9875

A - P1 = 26.9875 - 25.37504 = 1.61246

A - P2 = 26.9875 - 27.30702 = -0.31952

Thus, fit\_2 gave better prediction and it has less predictor variables.

#Finally, we can conclusively compare the two models based on AIC.  
AIC(fit\_1)

## [1] 972.3839

AIC(fit\_2)

## [1] 966.4765

Since 966.4765 < 972.3839, fit\_2 is a better model.

**Conclusion** Thoughts:

For this project, I wanted to explore the relationship of Percentage Disparity and the predictor variables Population, School Infrastructure Score, Learning Score, Enrollment Score, Total schools/ km^2 and Region. I found negative correlations between Percentage Disparity and Population, School Infrastructure Score, Learning Score, Enrollment Score and Total schools/km^2. We also found that Percentage Disparity varies across different regions in Pakistan.

I hypothesized that Total schools/ km^2, Learning score, Enrollment score and School Infrastructure Score would be more strongly related to Percentage Disparity. Through my analysis, I found that population, learning score and enrollment score were the most significant predictor variables. It was interesting to note that Infrastructure Score, which is an idicator of the facilities provided by schools, did not affect percentage disparity significantly. I was also suprised to find that the proximity of schools (Total schools/ km^2) did not significant impact disparity in my model.

Results:

For this project, we developed a model that effectively predicts percentage disparity in schools across cities in Pakistan using Multiple Regression in R. We found the significant predictors to be population, learning score and enrollment score.

The model equation is as follows:

Percentage-Disparity-predicted = (1.257e+02) + -3.355e-06(Population) + -5.388e-01(learning score) + -8.005e-01(enrollment score)

-As Population increases, Percentage-Disparity is predicted to decrease

-As Learning Score increases, Percentage-Disparity is predicted to decrease

-As Enrollment score increases, Percentage-Disparity is predicted to decrease

This model significantly predicts the outcome variable (R^2 = 0.6461; F(3, 117) = 74.02, p < 0.01).

Thus, if we have the population, (mean) learning score, and (mean) enrollment score of a city in Pakistan, we can effectively predict the percentage disparity in that city using our model.