UMCH

Cora Boyoung Jung, Enoch Mwesigwa, Jordan Severn

December 08, 2020

Import csv & Change data type

```
UMCHdata <- read_csv("https://raw.githubusercontent.com/Cora-Boyoung-Jung/UMCH/main/data/UMCH.csv",
                     col_types = cols(Birthdate = col_date(format = "%m/%d/%Y"),
                                      Age = col_integer()))
glimpse(UMCHdata)
```

```
## Rows: 36
## Columns: 15
## $ Filename                      <chr> "failed_infant_1", "failed_infant_2",...
## $ Birthdate                     <date> 2020-05-26, 2020-06-03, 2020-03-05, ...
## $ AgeGroup                      <chr> "infant", "infant", "infant", "infant...
## $ Age                           <int> 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1...
## $ Gender                        <chr> "Male", "Female", "Female", "Female",...
## $ PhysicalDevelopment           <dbl> 4.0, 4.0, 13.0, 23.0, 6.0, 24.0, 19.0...
## $ LanguageDevelopment           <dbl> 7.0, 5.0, 11.0, 32.0, 11.0, 36.0, 21....
## $ Adaptive_SelfHelp             <dbl> 3, 3, 4, 6, 3, 11, 8, 8, 8, 6, 7, 8, ...
## $ Adaptive_SocialEmotional      <dbl> 3, 0, 5, 14, 5, 16, 12, 12, 12, 12, 1...
## $ AcademicAndCognitive          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ AcademicAndCognitive_Maths    <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ AcademicAndCognitive_Literacy <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ TotalScore                    <dbl> 17.0, 12.0, 33.0, 75.0, 25.0, 87.0, 6...
## $ Status                        <chr> "failed", "failed", "failed", "passed...
## $ Examiner                      <chr> "Sam McGowen", "Sam McGowen", "Sam Mc...
```

Tidying data

```
UMCH <- UMCHdata %>% mutate(Status = tolower(Status))
neworder <- c("infant","toddler","two_year", "three_year", "four_year")
library(plyr)  ## or dplyr (transform -> mutate)
UMCH <- arrange(transform(UMCH,
             AgeGroup=factor(AgeGroup,levels=neworder)),AgeGroup)
```

Exploring data set

```
head(UMCH)
```

```
##           Filename  Birthdate AgeGroup Age Gender PhysicalDevelopment
## 1  failed_infant_1 2020-05-26   infant   0   Male                   4
## 2  failed_infant_2 2020-06-03   infant   0 Female                   4
## 3 failed_infant_3* 2020-03-05   infant   0 Female                  13
## 4  passed_infant_1 2019-11-13   infant   0 Female                  23
## 5  passed_infant_2 2020-07-01   infant   0   Male                   6
## 6  passed_infant_3 2019-11-30   infant   0   Male                  24
##   LanguageDevelopment Adaptive_SelfHelp Adaptive_SocialEmotional
## 1                   7                 3                        3
## 2                   5                 3                        0
## 3                  11                 4                        5
## 4                  32                 6                       14
## 5                  11                 3                        5
## 6                  36                11                       16
##   AcademicAndCognitive AcademicAndCognitive_Maths AcademicAndCognitive_Literacy
## 1                   NA                         NA                            NA
## 2                   NA                         NA                            NA
## 3                   NA                         NA                            NA
## 4                   NA                         NA                            NA
## 5                   NA                         NA                            NA
## 6                   NA                         NA                            NA
##   TotalScore Status        Examiner
## 1         17 failed     Sam McGowen
## 2         12 failed     Sam McGowen
## 3         33 failed     Sam McGowen
## 4         75 passed Melissa Swanson
## 5         25 passed     Sam McGowen
## 6         87 passed Melissa Swanson
```

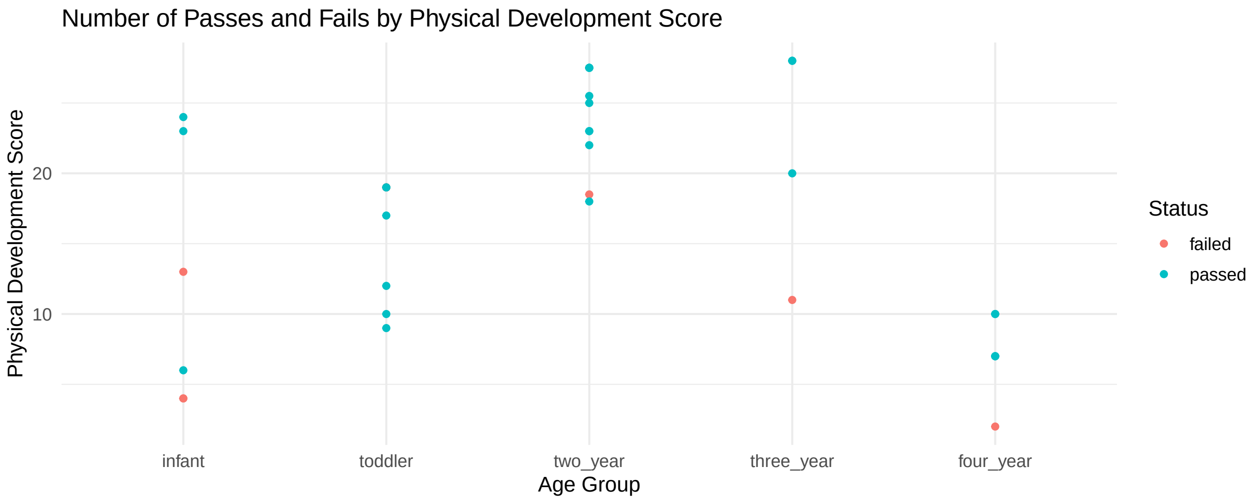
```r
summary(UMCH)
```

```
##    Filename           Birthdate                AgeGroup       Age       
##  Length:36          Min.   :2016-01-21   infant    : 6   Min.   :0.000  
##  Class :character   1st Qu.:2017-06-26   toddler   : 9   1st Qu.:1.000  
##  Mode  :character   Median :2018-04-10   two_year  :11   Median :2.000  
##                     Mean   :2018-04-29   three_year: 4   Mean   :1.861  
##                     3rd Qu.:2019-03-31   four_year : 6   3rd Qu.:3.000  
##                     Max.   :2020-07-01                   Max.   :4.000  
##                                                                         
##     Gender          PhysicalDevelopment LanguageDevelopment Adaptive_SelfHelp
##  Length:36          Min.   : 2.00       Min.   : 5.00       Min.   : 3.0     
##  Class :character   1st Qu.:10.00       1st Qu.:20.75       1st Qu.: 5.0     
##  Mode  :character   Median :19.00       Median :35.75       Median : 8.0     
##                     Mean   :17.00       Mean   :32.00       Mean   : 6.6     
##                     3rd Qu.:23.25       3rd Qu.:43.25       3rd Qu.: 8.0     
##                     Max.   :28.00       Max.   :51.00       Max.   :11.0     
##                                                             NA's   :21       
##  Adaptive_SocialEmotional AcademicAndCognitive AcademicAndCognitive_Maths
##  Min.   : 0.00            Min.   : 2.50        Min.   : 5.000            
##  1st Qu.: 8.00            1st Qu.: 7.50        1st Qu.: 9.375            
##  Median :12.00            Median :12.00        Median :15.250            
##  Mean   :10.07            Mean   :12.95        Mean   :13.000            
##  3rd Qu.:12.00            3rd Qu.:18.00        3rd Qu.:16.625            
##  Max.   :16.00            Max.   :21.50        Max.   :18.000            
##  NA's   :21               NA's   :15           NA's   :30                
##  AcademicAndCognitive_Literacy   TotalScore       Status         
##  Min.   : 0.000                Min.   :12.00   Length:36         
##  1st Qu.: 3.750             1st Qu.:54.00   Class :character
##  Median : 9.500             Median :73.75   Mode  :character
##  Mean   : 6.833             Mean   :66.81
##  3rd Qu.:10.000             3rd Qu.:86.00
##  Max.   :10.000             Max.   :95.50
##  NA's   :30
##    Examiner
##  Length:36
##  Class :character
##  Mode  :character
##
##
##
##
```

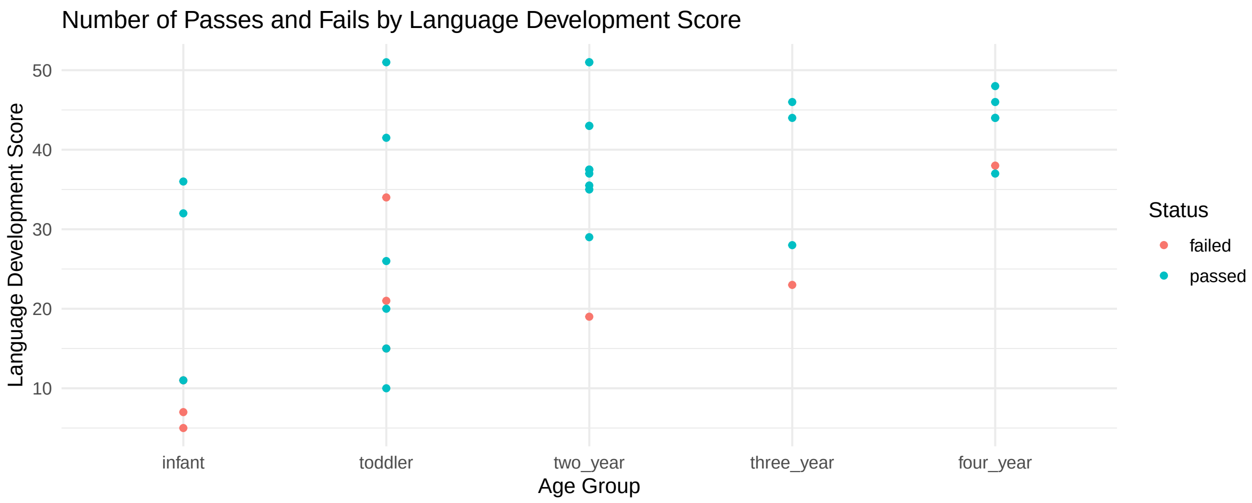
Graphics

- Which domain in the areas of development is scored the lowest and highest in which age group and overall?

```
ggplot(UMCH, aes(x = AgeGroup,  y = PhysicalDevelopment, color = Status)) +
  geom_point() +
  labs(x = "Age Group",
       y = "Physical Development Score",
       title = "Number of Passes and Fails by Physical Development Score")
```

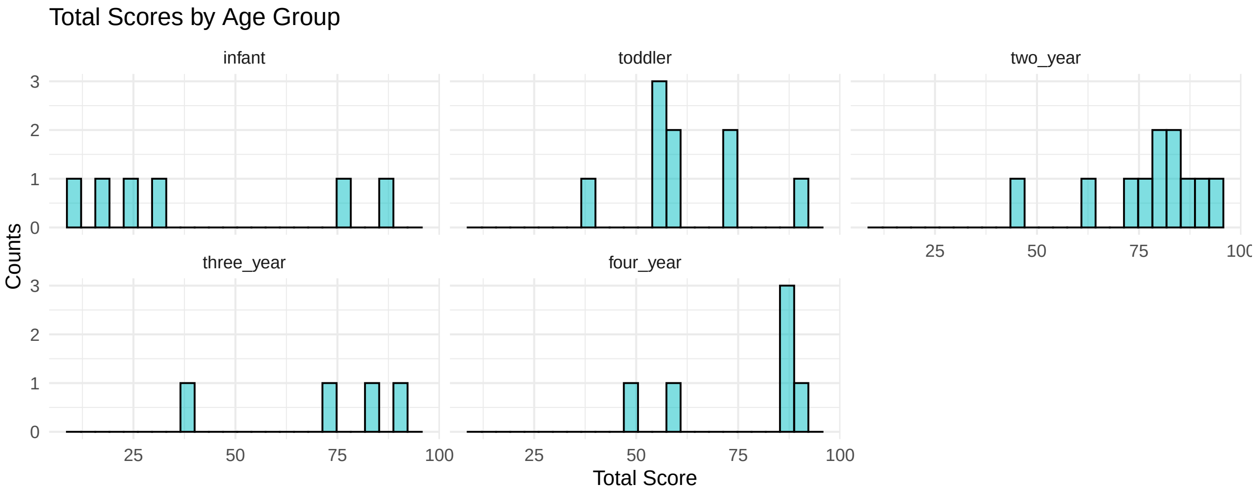


```
ggplot(UMCH, aes(x = AgeGroup,  y = LanguageDevelopment, color = Status)) +
  geom_point() +
  labs(x = "Age Group",
       y = "Language Development Score",
       title = "Number of Passes and Fails by Language Development Score")
```



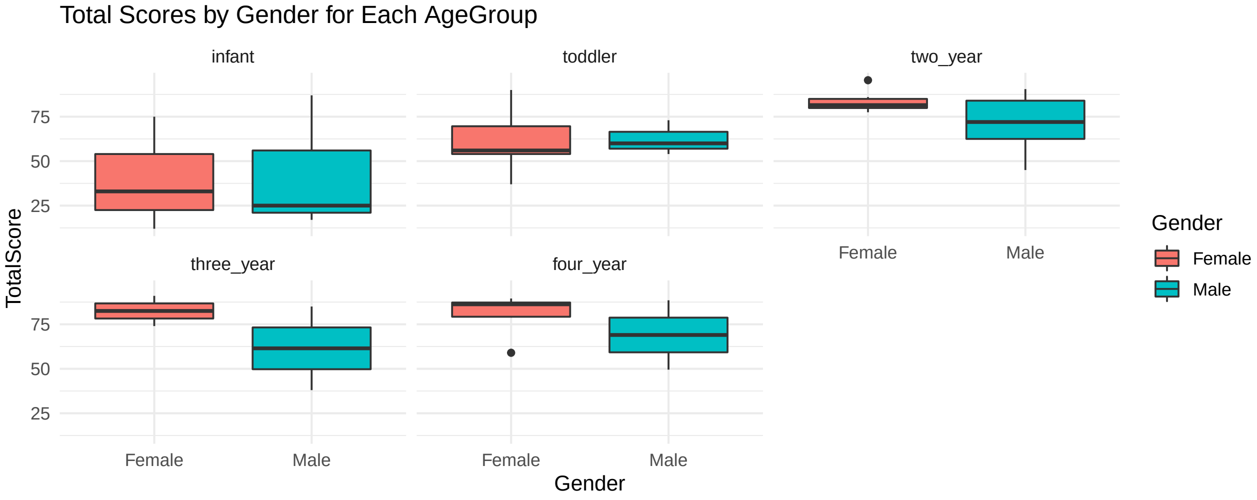
- How does score vary by age group?

```
gf_histogram(~TotalScore, data=UMCH, fill='#00BFC4', color='black') %>%
  gf_labs(title="Total Scores by Age Group", x="Total Score", y="Counts") + facet_wrap(~AgeGroup)
```



- How does score vary by gender?

```
ggplot(UMCH, aes(x = Gender,  y = TotalScore, fill=Gender)) +
  geom_boxplot() +
  facet_wrap(~AgeGroup) +
  labs(title="Total Scores by Gender for Each AgeGroup", x="Gender")
```



Statistics in Raw Score

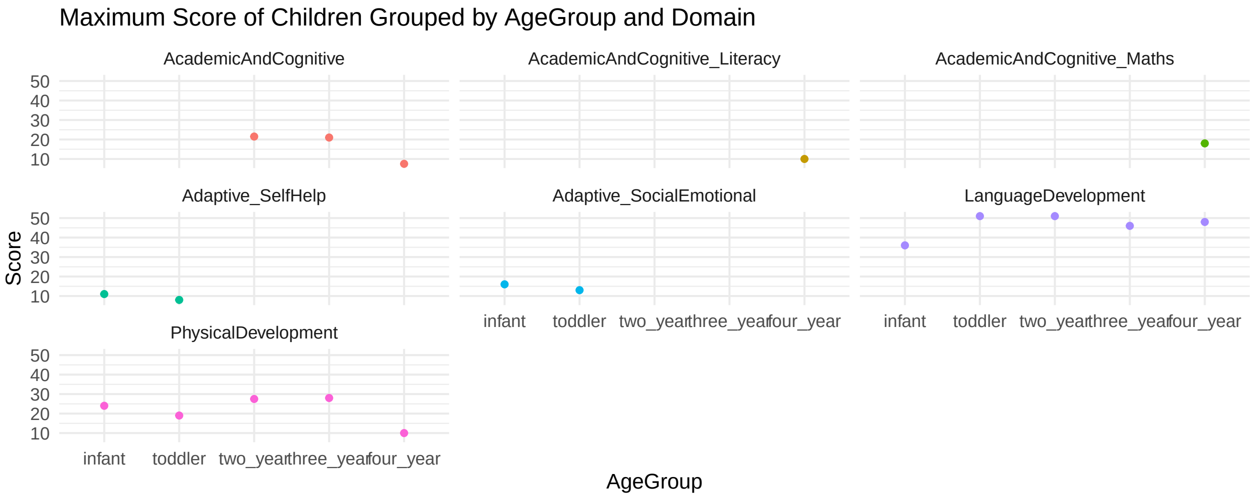
```r
UMCH_long <- UMCH %>%
  subset(select = -c(Filename, Birthdate, Age, Gender, TotalScore, Status, Examiner)) %>%
  pivot_longer(!AgeGroup, names_to = "Domain", values_to = "Score") %>%
  na.omit() %>%
  mutate(AgeGroup = as.factor(AgeGroup),
         Domain = as.factor(Domain),
         Score = as.numeric(Score))

UMCH_stat <- UMCH_long %>%
  dplyr::group_by(AgeGroup,Domain) %>%
  dplyr::summarise(Min=min(Score),
                   Max=max(Score),
                   Mean=mean(Score),
                   Median=median(Score))
```

```r
UMCH_stat %>%
  knitr::kable()
```

AgeGroup	Domain	Min	Max	Mean	Median
infant	Adaptive_SelfHelp	3.0	11.0	5.000000	3.50
infant	Adaptive_SocialEmotional	0.0	16.0	7.166667	5.00
infant	LanguageDevelopment	5.0	36.0	17.000000	11.00
infant	PhysicalDevelopment	4.0	24.0	12.333333	9.50
toddler	Adaptive_SelfHelp	6.0	8.0	7.666667	8.00
toddler	Adaptive_SocialEmotional	11.0	13.0	12.000000	12.00
toddler	LanguageDevelopment	10.0	51.0	25.944444	21.00
toddler	PhysicalDevelopment	9.0	19.0	15.888889	19.00
two_year	AcademicAndCognitive	7.5	21.5	15.636364	14.50
two_year	LanguageDevelopment	19.0	51.0	38.045454	37.50
two_year	PhysicalDevelopment	18.0	27.5	24.090909	25.00
three_year	AcademicAndCognitive	4.0	21.0	15.000000	17.50
three_year	LanguageDevelopment	23.0	46.0	35.250000	36.00
three_year	PhysicalDevelopment	11.0	28.0	21.750000	24.00
four_year	AcademicAndCognitive	2.5	7.5	6.666667	7.50

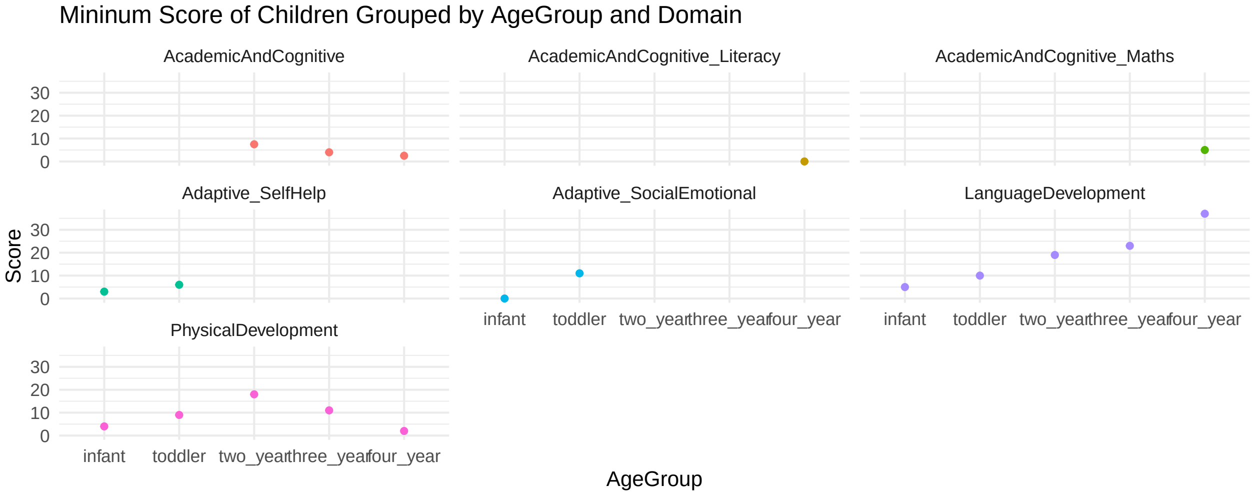
AgeGroup	Domain	Min	Max	Mean	Median
four_year	AcademicAndCognitive_Literacy	0.0	10.0	6.833333	9.50
four_year	AcademicAndCognitive_Maths	5.0	18.0	13.000000	15.25
four_year	LanguageDevelopment	37.0	48.0	42.833333	44.00
four_year	PhysicalDevelopment	2.0	10.0	7.166667	7.00

```r
ggplot(UMCH_stat, aes(x = AgeGroup,  y = Max, color=Domain)) +
  geom_point() +
  facet_wrap(~Domain) +
  labs(title="Maximum Score of Children Grouped by AgeGroup and Domain", y="Score") +
  theme(legend.position = "none")
```

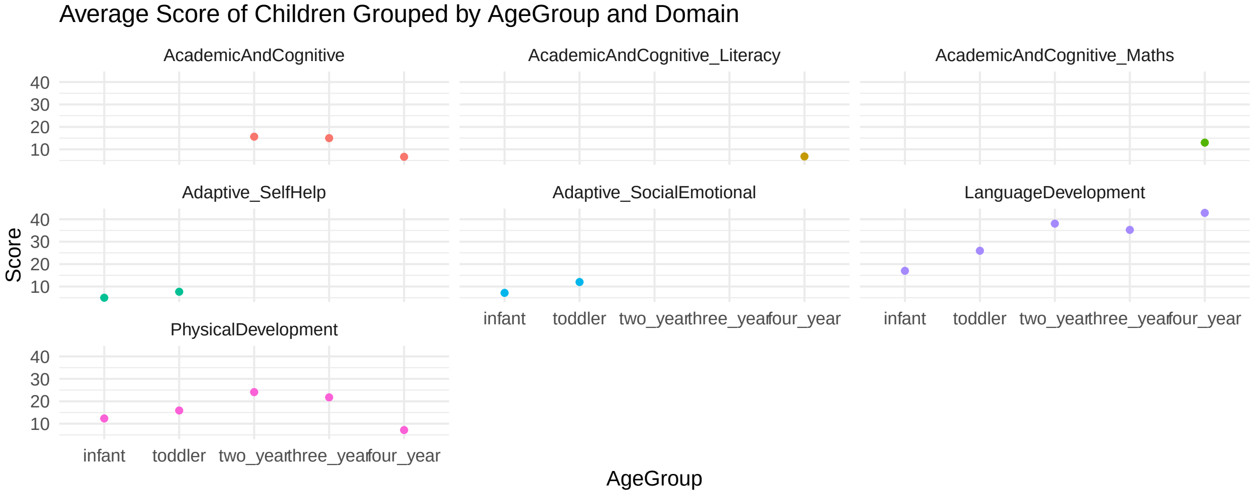


‘

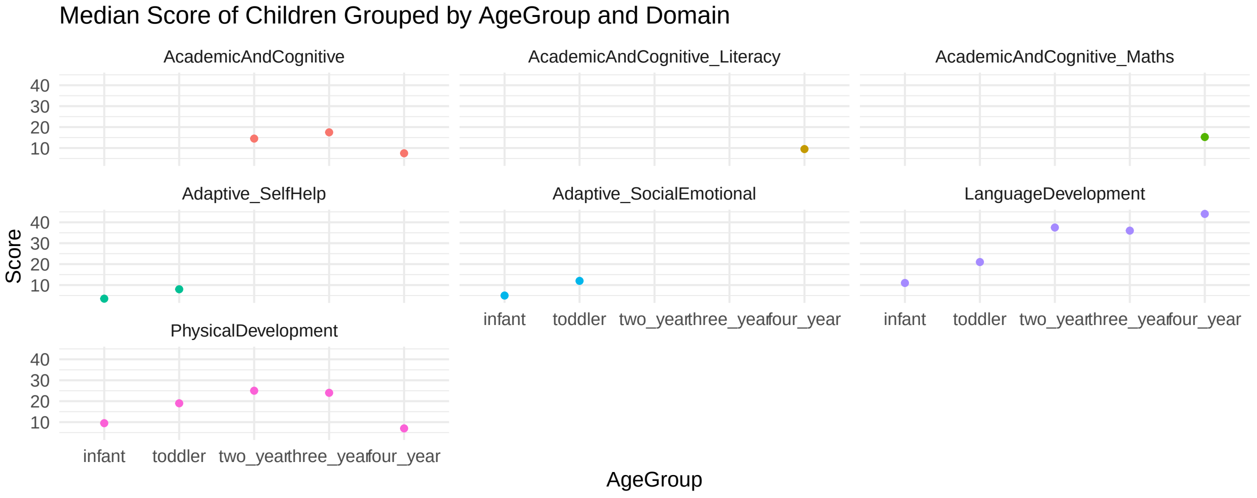
```r
ggplot(UMCH_stat, aes(x = AgeGroup,  y = Min, color=Domain)) +
  geom_point() +
  facet_wrap(~Domain) +
  labs(title="Mininum Score of Children Grouped by AgeGroup and Domain", y="Score") +
  theme(legend.position = "none")
```



```r
ggplot(UMCH_stat, aes(x = AgeGroup,  y = Mean, color=Domain)) +
  geom_point() +
  facet_wrap(~Domain) +
  labs(title="Average Score of Children Grouped by AgeGroup and Domain", y="Score") +
  theme(legend.position = "none")
```



```
ggplot(UMCH_stat, aes(x = AgeGroup,  y = Median, color=Domain)) +
  geom_point() +
  facet_wrap(~Domain) +
  labs(title="Median Score of Children Grouped by AgeGroup and Domain", y="Score") +
  theme(legend.position = "none")
```



Statistics in Percentage

```
# group and get median of scores in each age group
UMCH_grouped <- UMCHdata %>%
    select(AgeGroup, PhysicalDevelopment,LanguageDevelopment, Adaptive_SelfHelp,
           Adaptive_SocialEmotional, AcademicAndCognitive, AcademicAndCognitive_Maths,
          AcademicAndCognitive_Literacy) %>%
  group_by(AgeGroup) %>%
  summarise_each(funs(round(mean(., na.rm = TRUE), digits=2)))
```

```
## Warning: `summarise_each_()` is deprecated as of dplyr 0.7.0.
## Please use `across()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
## Warning: `funs()` is deprecated as of dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```r
UMCH_grouped <- arrange(transform(UMCH_grouped,
             AgeGroup=factor(AgeGroup,levels=neworder)),AgeGroup)

View(UMCH_grouped)
```

```
## Warning in View(UMCH_grouped): unable to open display
```

```
## Error in .External2(C_dataviewer, x, title): unable to start data viewer
```

```r
#make data with max scores
AgeGroup <- c('infant', 'toddler', 'two_year', 'three_year', 'four_year')
UMCH_max <-  as.data.frame(rbind(c('infant', 28,44,12, 16, NA, NA, NA),
                                 c('toddler', 19,59,9, 13, NA, NA, NA),
                                 c('two_year', 27.5,51,NA, NA, 21.5, NA, NA),
                                 c('three_year', 31,48,NA, NA, 21, NA, NA),
                                 c('four_year', 15, 48,NA, NA, 10, 17, 10 )))

#rename columns
names(UMCH_max)[1] <- "AgeGroup"
names(UMCH_max)[2] <- "PhysicalDevelopment"
names(UMCH_max)[3] <- "LanguageDevelopment"
names(UMCH_max)[4] <- "Adaptive_SelfHelp"
names(UMCH_max)[5] <- "Adaptive_SocialEmotional"
names(UMCH_max)[6] <- "AcademicAndCognitive"
names(UMCH_max)[7] <- "AcademicAndCognitive_Maths"
names(UMCH_max)[8] <- "AcademicAndCognitive_Literacy"

#convert to characters
UMCH_max$PhysicalDevelopment <- as.numeric(UMCH_max$PhysicalDevelopment)
UMCH_max$LanguageDevelopment <- as.numeric(UMCH_max$LanguageDevelopment)
UMCH_max$Adaptive_SelfHelp <- as.numeric(UMCH_max$Adaptive_SelfHelp)
UMCH_max$Adaptive_SocialEmotional <- as.numeric(UMCH_max$Adaptive_SocialEmotional)
UMCH_max$AcademicAndCognitive <-  as.numeric(UMCH_max$AcademicAndCognitive)
UMCH_max$AcademicAndCognitive_Maths <- as.numeric(UMCH_max$AcademicAndCognitive_Maths)
UMCH_max$AcademicAndCognitive_Literacy <- as.numeric(UMCH_max$AcademicAndCognitive_Literacy)
View(UMCH_max)
```

```
## Warning in View(UMCH_max): unable to open display
## Error in .External2(C_dataviewer, x, title): unable to start data viewer
```

```
getPercent <- function(list1, list2){
  newList <- vector(mode = "list", length = 5)
  for (i in 1:length(list1)) {
    if(is.na(list1[i]))
    {
      newList[[i]] <- NA
    }
    else
    {
      newList[[i]] <- round( (list1[i] / list2[i] * 100), digits = 2)
    }
  }
  newList
}
```

```
UMCH_percent <- UMCH_grouped %>%
    mutate(PhysicalDevelopment= getPercent(UMCH_grouped$PhysicalDevelopment,UMCH_max$PhysicalDevelopmen
           LanguageDevelopment= getPercent(UMCH_grouped$LanguageDevelopment,UMCH_max$LanguageDevelopmen
           Adaptive_SelfHelp= getPercent(UMCH_grouped$Adaptive_SelfHelp,UMCH_max$Adaptive_SelfHelp),
           Adaptive_SocialEmotional= getPercent(UMCH_grouped$Adaptive_SocialEmotional,UMCH_max$Adaptive_
           AcademicAndCognitive= getPercent(UMCH_grouped$AcademicAndCognitive,UMCH_max$AcademicAndCogni
           AcademicAndCognitive_Maths= getPercent(UMCH_grouped$AcademicAndCognitive_Maths,UMCH_max$Acad
           AcademicAndCognitive_Literacy= getPercent(UMCH_grouped$AcademicAndCognitive_Literacy,UMCH_ma

reactable(UMCH_percent)
```

```
## Error in reactable(UMCH_percent): could not find function "reactable"
```

```
#install.packages("reactable") may need to run once
library(reactable)
```

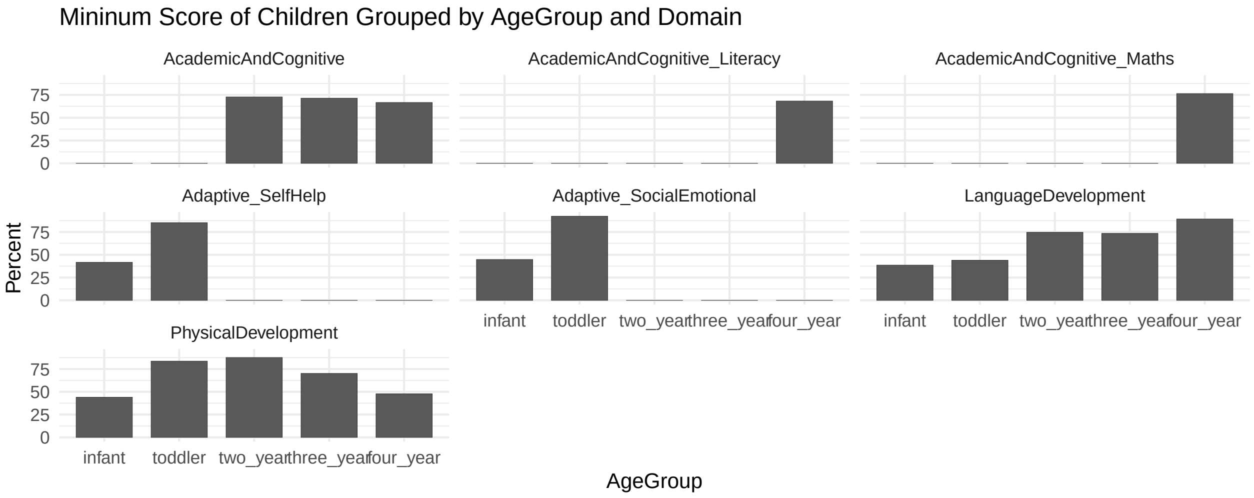
```
UMCH_percent_long <- UMCH_percent %>%
  pivot_longer(!AgeGroup, names_to = "Domain", values_to = "Percentage") %>%
  na.omit()

UMCH_percent_long %>%
  knitr::kable()
```

AgeGroup	Domain	Percentage
infant	PhysicalDevelopment	44.04
infant	LanguageDevelopment	38.64
infant	Adaptive_SelfHelp	41.67
infant	Adaptive_SocialEmotional	44.81
infant	AcademicAndCognitive	NA
infant	AcademicAndCognitive_Maths	NA
infant	AcademicAndCognitive_Literacy	NA
toddler	PhysicalDevelopment	83.63
toddler	LanguageDevelopment	43.97
toddler	Adaptive_SelfHelp	85.22
toddler	Adaptive_SocialEmotional	92.31
toddler	AcademicAndCognitive	NA
toddler	AcademicAndCognitive_Maths	NA
toddler	AcademicAndCognitive_Literacy	NA

AgeGroup	Domain	Percentage
two_year	PhysicalDevelopment	87.6
two_year	LanguageDevelopment	74.61
two_year	Adaptive_SelfHelp	NA
two_year	Adaptive_SocialEmotional	NA
two_year	AcademicAndCognitive	72.74
two_year	AcademicAndCognitive_Maths	NA
two_year	AcademicAndCognitive_Literacy	NA
three_year	PhysicalDevelopment	70.16
three_year	LanguageDevelopment	73.44
three_year	Adaptive_SelfHelp	NA
three_year	Adaptive_SocialEmotional	NA
three_year	AcademicAndCognitive	71.43
three_year	AcademicAndCognitive_Maths	NA
three_year	AcademicAndCognitive_Literacy	NA
four_year	PhysicalDevelopment	47.8
four_year	LanguageDevelopment	89.23
four_year	Adaptive_SelfHelp	NA
four_year	Adaptive_SocialEmotional	NA
four_year	AcademicAndCognitive	66.7
four_year	AcademicAndCognitive_Maths	76.47
four_year	AcademicAndCognitive_Literacy	68.3

```r
#View(UMCH_percent)
```

```r
#barplot(UMCH_percent_long, main="Car Distribution",
 #  xlab="Number of Gears") + facet_wrap(~Domain)

ggplot(UMCH_percent_long, aes(x = AgeGroup,  y = Percentage)) +
  geom_bar(stat = "identity", width = 0.75, color = "#2b2b2b", size = 0.05 ) +
  facet_wrap(~Domain, ncol= 3) +
  labs(title="Mininum Score of Children Grouped by AgeGroup and Domain", y="Percent")  +
  theme(
    strip.text.x = element_text(margin = margin(5, 5, 5, 5))
  )
```



TotalScore Model

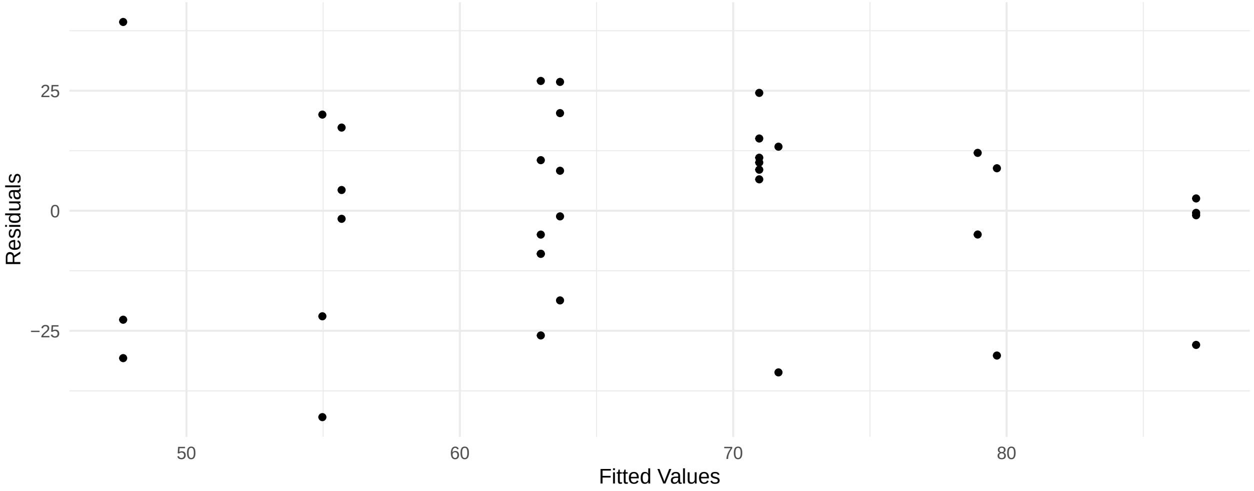
We used linear regression model because it is used to show/predict the relationship between variables where the response variable is continuous. We can use up to 2 predictors given the size of the data (36/15). We will not add any interaction or random effect due to the size of the dataset. Our predictors will be Age and Gender because all other variables contributes to the Total Score which does not make logical sense to include those.
Response variable: TotalScore Predictor(s): Age, Gender
Regression model: Linear regression

```
mod_total <- lm(TotalScore ~ Age + Gender,
          data = UMCH)
summary(mod_total)
```

```
##
## Call:
## lm(formula = TotalScore ~ Age + Gender, data = UMCH)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.971 -11.387   3.447  12.380  39.316
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   54.971      6.700   8.205 1.79e-09 ***
## Age            7.990      2.629   3.039  0.00461 **
## GenderMale    -7.287      6.902  -1.056  0.29872
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.4 on 33 degrees of freedom
## Multiple R-squared:  0.2436, Adjusted R-squared:  0.1977
## F-statistic: 5.313 on 2 and 33 DF,  p-value: 0.009995
```

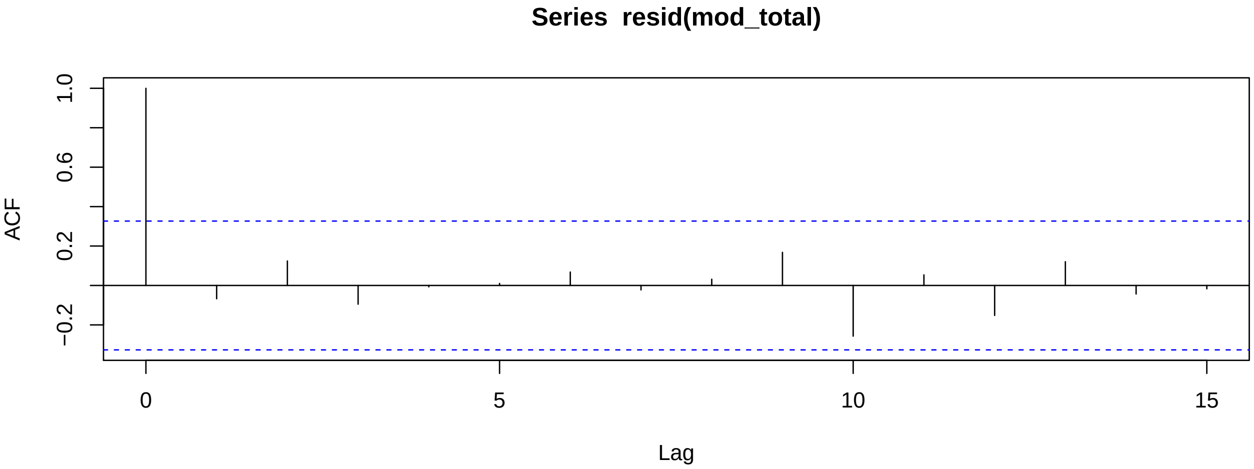
Model Assessment

```
gf_point(resid(mod_total) ~ fitted(mod_total)) %>%
  gf_labs(x = 'Fitted Values', y = 'Residuals')
```



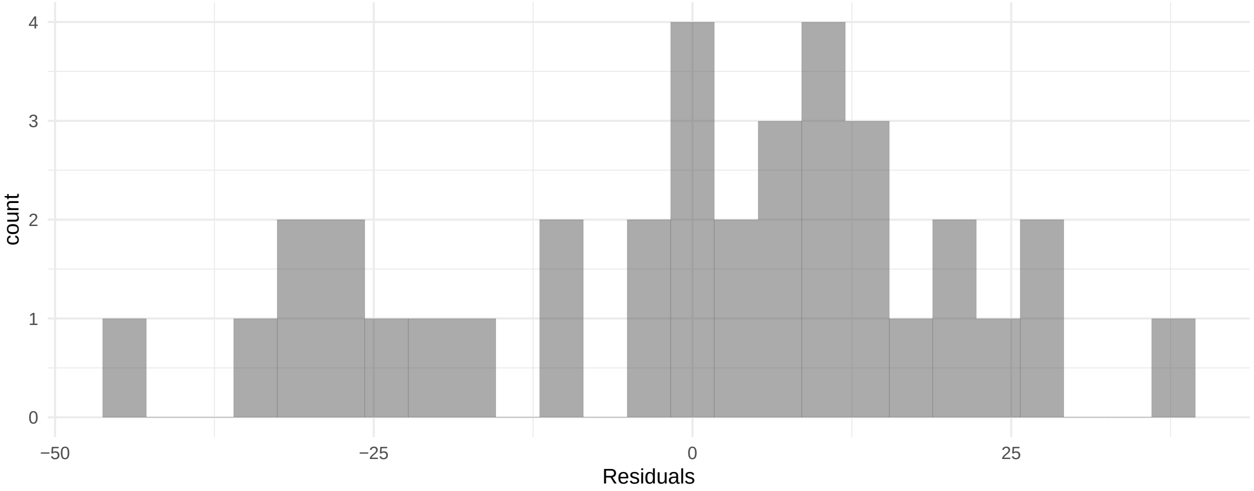
The fitted values vs. the residual plot is used for linearity and error variance. There is no trumpeting present so it passes the error variance condition. However, there is some clustering present and the data points make vertical lines so linearity is a little more iffy on whether or not it passes.

```
acf(resid(mod_total))
```



Looking at the acf plot, the plot passes the independence condition. All the lines are within the dotted lines and the lag at 0 is normal.

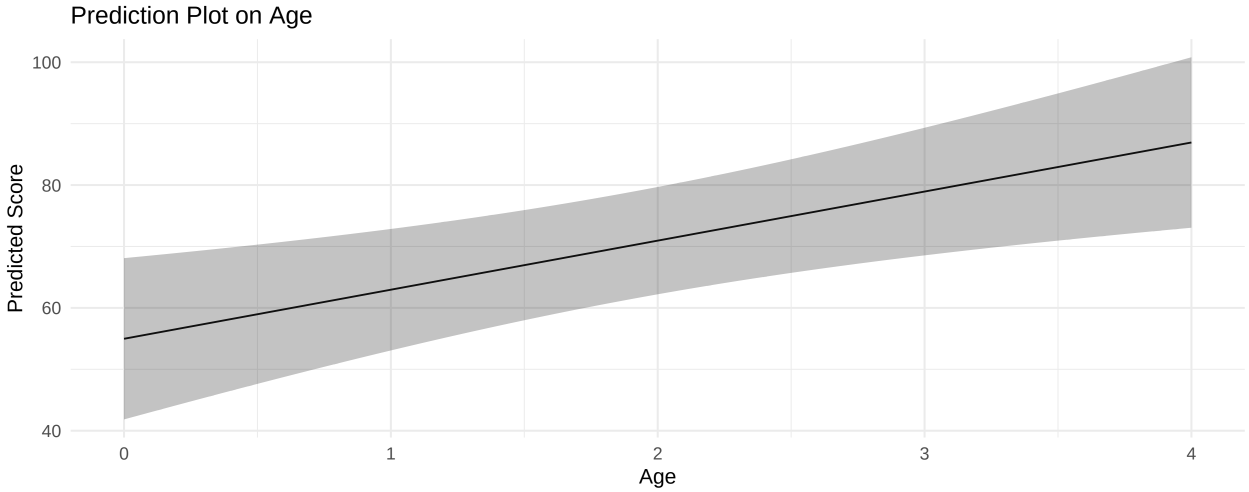
```
gf_histogram(~resid(mod_total)) %>%
  gf_labs(x = "Residuals")
```



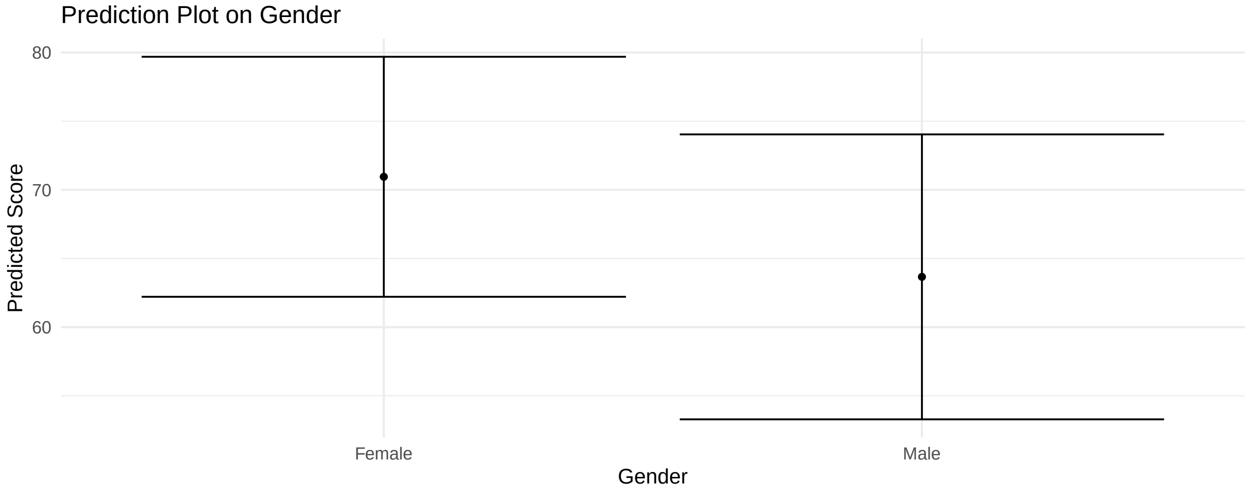
Since the data is limited, the normality of the histogram seems off as the counts of each bin only go up to the highest of 4. It is hard to tell whether the histogram passes as we are unsure if it is caused by the limited data or something else. Also when we made exploratory graphs, all the histograms were skewed which may play a role in the normality of the histogram.

Prediction Plot

```
pred_plot(mod_total, 'Age') %>%
  gf_labs(y = 'Predicted Score',
          title = 'Prediction Plot on Age')
```



```
pred_plot(mod_total, 'Gender') %>%
  gf_labs(y = 'Predicted Score',
          title = 'Prediction Plot on Gender')
```



According to the prediction plot for Age, there seems to be a positive trend between the score and age where as age increases, the score also increases. This makes sense because the "pass" score for each age group differs according to the age with lower age having the lowest pass score.

According to the prediction plot for Gender, it looks like the total score is a little higher for female compared to male. However, the overlapping range is high and the number of children is not the same (21 females and 15 males), so it brings down the reliability score of this data but it is worth noting.