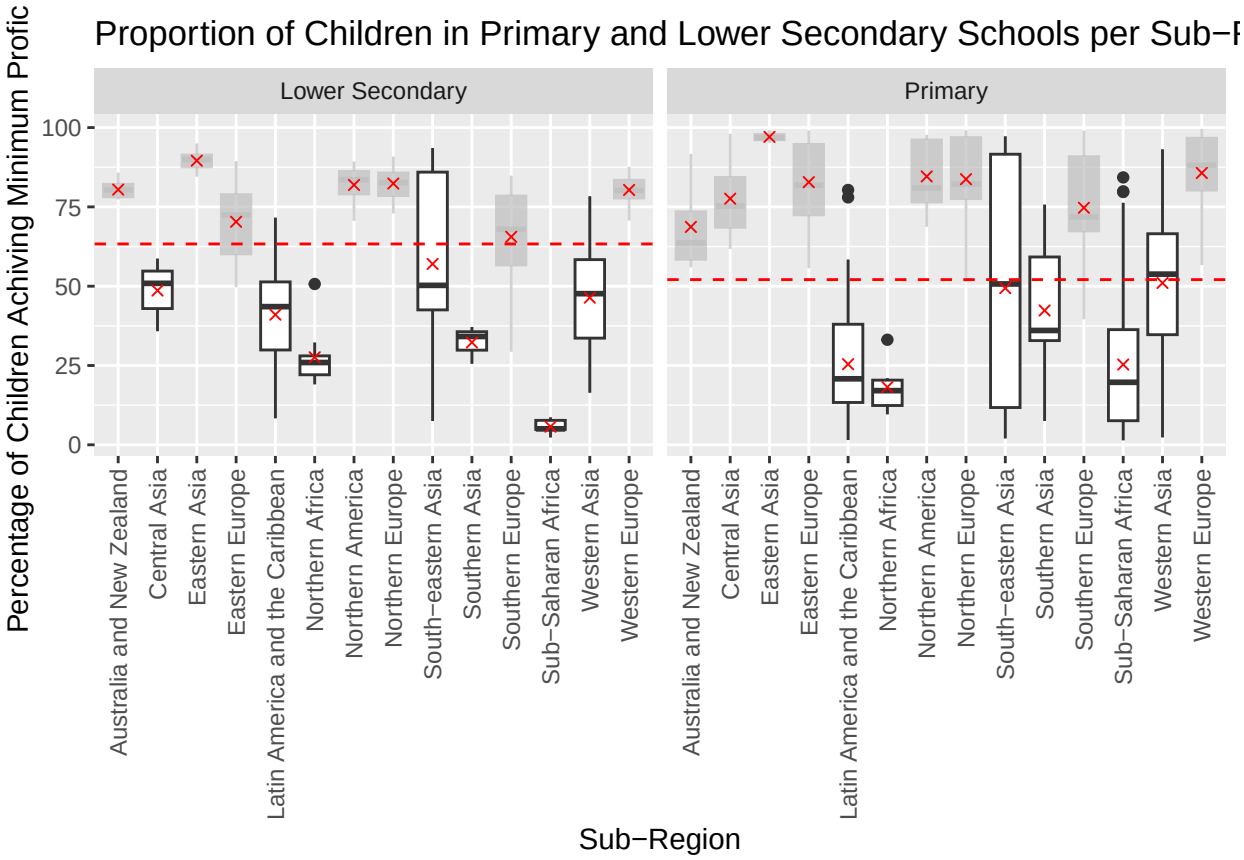
Question 1 EDA

Andrew Dawson

EDA and Analysis of 4.1.1

Comparing Sub-Regions



Within the box plots, the highlighted box plots show each Sub Region's mean being less than the mean for the overall Education Level – for clarity the red crosses are the means for each Sub Region, whilst the black points are the outliers. Within the Lower Secondary section, only one Sub-Region displays an outlier, being Northern Africa. Interestingly, this outlier is still below the grand mean of the section, displaying that Northern Africa is an area of concern. Additionally, South-east Asia is the only Sub-Region where the mean falls below the grand mean, but the third quartile is higher. This shows much more spread across the Sub-Region. Within the Primary section, outlier are shown in Latin America and the Caribbean, Northern Africa, and Sub-Saharan Africa. Additonally, South-east Asia shows an incredible spread of data, with the IQR being NA.

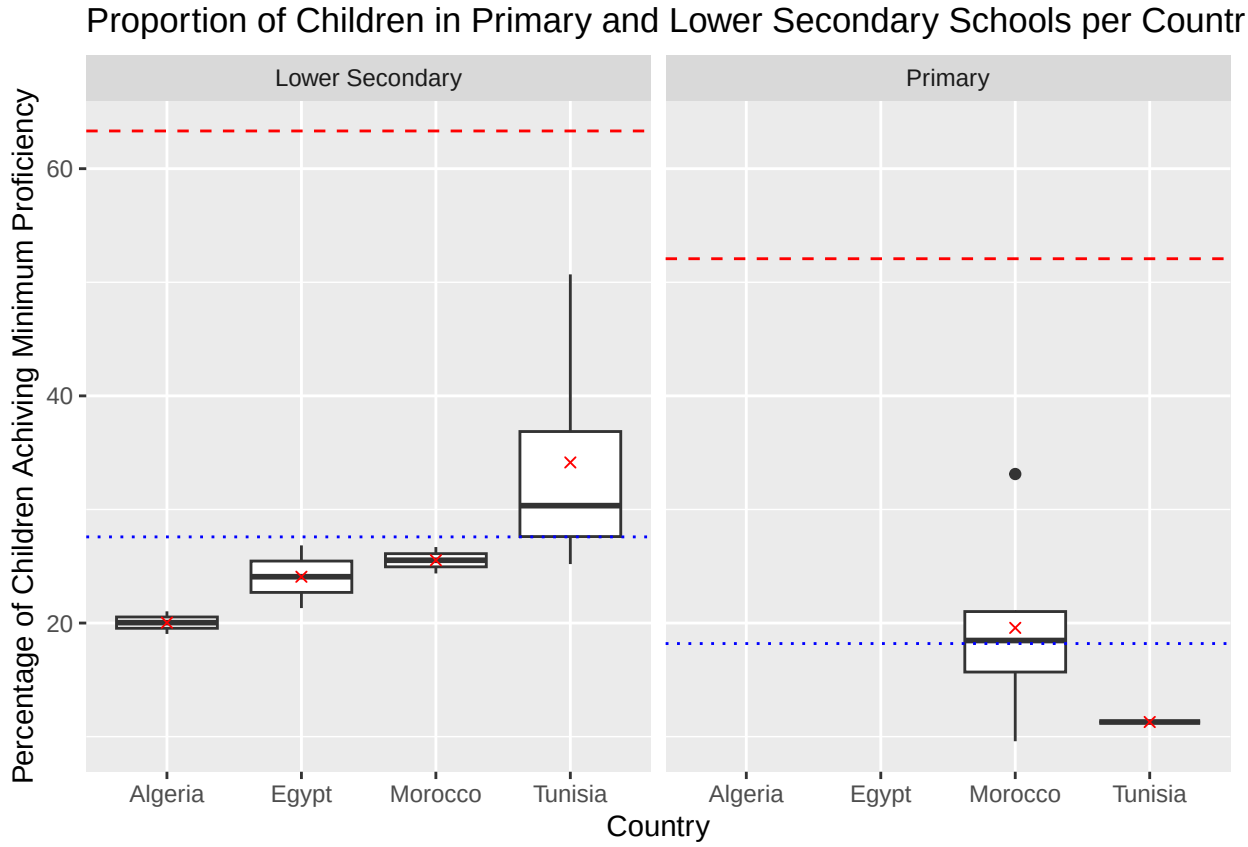
Whilst the Lower Secondary grand mean, being 63.32, is higher than the Primary's grand mean, being 52.07, it is important to consider the two means are not independent of another. In other words, in order to get

into secondary school one must pass primary, thus the grand mean of the Primary schools would affect the mean of the Lower Secondary schools.

South-east Asia also demonstrates the most spread across the sub-regions. Contextually, this aligns with the general developmental growth of the region. For instance, the education within Singapore is usually shown as the world's best standard for education, whilst a nation like Laos or Thailand, lacks behind, which all three countries are within the same sub-region.

Highlighting a Concerning Sub-Region

Northern Africa is one of the sub-regions that shown a lot of concern when looking at the proportion of children achieving minimum proficiency. As highlighted before, the most extreme part of the ranges for both Primary and Secondary education does not reach the Grand Mean.



Similar to the first plot, the red line indicates the grand mean of the world per education level; however, the blue line indicates the grand mean of the sub-region, Northern Africa. No country's was able to achieve the grand mean within either education level. In addition, certain bits of data were missing for the Primary section, being Algeria and Egypt.

Table (Not Sure)

GeoAreaName	sub-region	TimePeriod	Value	Education level	Type of skill
Niger	Sub-Saharan Africa	2014	1.40000	PRIMAR	SKILL_MATH
Dominican Republic	Latin America and the Caribbean	2013	1.50000	PRIMAR	SKILL_MATH

GeoAreaName	sub-region	TimePeriod	Value	Education level	Type of skill
Chad	Sub-Saharan Africa	2019	1.80000	PRIMAR	SKILL_MATH
Zambia	Sub-Saharan Africa	2016	1.80000	PRIMAR	SKILL_READ
Lao People's Democratic Republic	South-eastern Asia	2019	2.00000	PRIMAR	SKILL_READ
Dominican Republic	Latin America and the Caribbean	2019	2.10000	PRIMAR	SKILL_MATH
Niger	Sub-Saharan Africa	2014	2.10000	PRIMAR	SKILL_READ
Zambia	Sub-Saharan Africa	2017	2.29846	LOWSEC	SKILL_MATH
Yemen	Western Asia	2011	2.32531	PRIMAR	SKILL_MATH
Côte d'Ivoire	Sub-Saharan Africa	2019	2.60000	PRIMAR	SKILL_MATH

Recomendation

For Sub-Regions which have it's maximum underneath the grand mean, more allocation of resources is needed there, whilst for Sub-Regions where its maximum does go above the grand mean but the Sub-Regions' mean is below the grand mean, further analysis per Sub-Region is needed to ensure proper allocation of resources, i.e. the Singaporean example within South-east Asia.

References

Duncalfe, Luke. 2024. "ISO-3166 Country and Dependent Territories Lists with UN Regional Codes." github.com/lukes/ISO-3166-Countries-with-Regional-Codes?tab=readme-ov-file.

United Nations. 2023. "SGD Indicators Database." https://unstats.un.org/sdgs/dataportal/database.

Appendix

```
#for future plots
tiltXText <- theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))

#filter data
DataPro <- Data %>% filter(Indicator == "4.1.1", Units == "PERCENT", Sex == "BOTHSEX", `sub-region` != 

#set up data for plot
lines <- DataPro %>% group_by(`Education level`) %>% summarise(grandMean = mean(Value))
points <- DataPro %>% group_by(`sub-region`, `Education level`) %>% summarise(mean = mean(Value))

#merge for highlighted plot
DataPro <- left_join(DataPro, lines, by = "Education level")

#plot filtered data
DataPro %>% ggplot(aes(`sub-region`, Value))+
  geom_boxplot()+
  facet_wrap(~`Education level`, labeller = labeller(`Education level` = c("LOWSEC" = "Lower Secondary"
  gghighlight(mean(Value) < mean(grandMean), calculate_per_facet = T)+
  tiltXText+
  geom_hline(aes(yintercept = grandMean), lty = 2, colour = "red", data=lines)+
  geom_point(data = points, aes(y=mean), pch = 4, colour="red")+
  labs(x = "Sub-Region", y = "Percentage of Children Achiving Minimum Proficiency", title = "Proportion
#filter just for N Afr
```

```r
DataAfr <- DataPro %>% filter(`sub-region` == "Northern Africa")
points2 <- DataAfr %>% group_by(GeoAreaName, `Education level`) %>% summarise(mean = mean(Value))

#plot
DataAfr %>% ggplot(aes(GeoAreaName, Value))+
  geom_boxplot()+
  facet_wrap(~`Education level`, labeller = labeller(`Education level` = c("LOWSEC" = "Lower Secondary"
  geom_hline(aes(yintercept = grandMean), lty=2,colour="red",data=lines)+
  geom_hline(aes(yintercept = mean), lty=3, colour="blue", data=points[points$`sub-region` == "Northern
  geom_point(data = points2, aes(y=mean), pch = 4, colour="red")+
  labs(x = "Country", y="Percentage of Children Achiving Minimum Proficiency", title = "Proportion of C
#table
tab <- DataPro %>%
  arrange(Value) %>%
  select(GeoAreaName, `sub-region`, TimePeriod, Value, `Education level`, `Type of skill`) %>%
  slice_head(n=10)

kable(tab)
```