---
title: "Thesis EDA"
author: "Sanjay Satish"
date: "2/21/2022"
output: html_document

```{r setup, include=FALSE}
library(tidyverse)
library(dplyr)
library(flextable)
library(grid)
library(gridExtra)
library(cowplot)
library(stargazer)
library(scales)
library(pROC)
library(patchwork)
library(survminer)
library(survival)
library(kableExtra)
library(ggplot2)
library(broom)
library(GGally)
library(tidyr)
library(knitr)
library(xtable)

students <-
read.table("/Users/Sanscubed/Desktop/Thesis/Thesis_data/STAR_Archive/STAR_Students.tab",
header = T, sep = "\t", fill = TRUE)
schools <- read.table("/Users/Sanscubed/Desktop/Thesis/Thesis_data/STAR_Archive/STAR_K-
3_Schools.tab", header = T, sep = "\t", fill = TRUE)

students$g1tmathss <- as.numeric(students$g1tmathss)
students$g1treadss <- as.numeric(students$g1treadss)
students$g1tlistss <- as.numeric(students$g1tlistss)
students$g1wordskillss <- as.numeric(students$g1wordskillss)

students$gktmathss <- as.numeric(students$gktmathss)
students$gktreadss <- as.numeric(students$gktreadss)
students$gktlistss <- as.numeric(students$gktlistss)
students$gkwordskillss <- as.numeric(students$gkwordskillss)
```

```
students$g2tmathss <- as.numeric(students$g2tmathss)
students$g3tmathss <- as.numeric(students$g3tmathss)

students$g1tmathss[is.na(students$g1tmathss)] = 0
students$g2tmathss[is.na(students$g2tmathss)] = 0
students$g3tmathss[is.na(students$g3tmathss)] = 0
students$gktmathss[is.na(students$gktmathss)] = 0

students$gkfreelunch[is.na(students$gkfreelunch)] = 0
students$g1freelunch[is.na(students$g1freelunch)] = 0
students$g2freelunch[is.na(students$g2freelunch)] = 0
students$g3freelunch[is.na(students$g3freelunch)] = 0
```

Data Manipulation - Merging Schools and Students

```{r merge}
schools<- schools %>%
  rename(schlurbn = var1, grdrange = var2, sch_flagk = var3, sch_flag1 = var4, sch_flag2 = var5,
sch_flag3 = var6) %>%
  select(schid, sch_flagk, sch_flag1, sch_flag2, sch_flag3)

# Need to make individual merges for each grade
#students_and_schools <- merge(students,schools,by="schid")

schools %>%
  filter(sch_flag3 == 0, sch_flagk == 1)
```

Pie Chart Creation

```{r}
students <- students %>%
  rename(flagsgk = var6, flagsg1 = var7, flagsg2 = var8, flagsg3 = var9)

star_students <- students %>%
  filter(flagsgk == 1)

```

Small Classes

```{r}
```

```
# Small Class Pie Chart
small<- star_students %>%
  filter(gkclasstype ==1)

# Total denominator
totalsmall <- nrow(small)

# Remained in experiment, same class type
small_3rd <- small %>%
  filter(g3classtype ==1, flagsg3 == 1)

nsmall3 <- nrow(small_3rd)

sametype_small <- nsmall3/totalsmall

# Remained in experiment, changed class type, ended with same school
notsmall_sameschool_3rd <- small %>%
  filter(g3classtype != 1, flagsg3 == 1, gkschid == g3schid)
nnotsmall3 <- nrow(notsmall_sameschool_3rd)

difftype_small <- nnotsmall3/totalsmall

# Remained in experiment, ended with diff school
diffschool_3rd <- small %>%
  filter(flagsg3 == 1, gkschid != g3schid)
diffschool3 <- nrow(diffschool_3rd)

diffschool_small <- diffschool3/totalsmall

# School left experiment
school_left <- small %>%
  filter(gkschid == 128068 | gkschid == 180344 | gkschid == 205489 | gkschid == 216536 |
gkschid ==244818)
schleft <- nrow(school_left)

schleft_exp <- schleft/totalsmall

# Left to another public school
school_left_pub <- small %>%
  filter(flagsg3 != 1) %>%
  filter(flagg4 == 1 | flagg5 == 1)
lp <- nrow(school_left_pub)
left_public_school <- lp/totalsmall
```

```
# Left Private
pvt <- small %>%
  filter(flagsg3 != 1, gkschid != 128068, gkschid != 180344, gkschid != 205489, gkschid != 216536,
gkschid !=244818) %>%
  filter(flagg4 == 0, flagg5 == 0)
n_pvt <- nrow(pvt)
prop_pvt <- n_pvt/totalsmall
```

```{r}
# Table of Values
tab <- matrix(c(sametype_small, difftype_small, diffschool_small, schleft_exp,
left_public_school, prop_pvt) , ncol=1, byrow=TRUE)
rownames(tab) <- c("Remained in expriment, same class type", "Changed class type within
school", "Changed schools, same class type", "School left expiriment", "Switched to another
public school", "Left public school system")
as.table(tab)

# Pie Chart Creation
slices <- c(sametype_small, difftype_small, diffschool_small, schleft_exp, left_public_school,
prop_pvt)
lbls <- c("Remained in expriment, same class type", "Changed class type within school",
"Changed schools, but remained in expiriment", "School left expiriment", "Switched to non-
participating public school", "Left public school system")
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(slices,labels = lbls, col=rainbow(length(lbls)),
   main="Attrition Profile of Students in Small Classes in Grade K")

```

Regular

```{r}
# Regular Class Pie Chart
regular<- star_students %>%
  filter(gkclasstype ==2)

# Total denominator
total_reg <- nrow(regular)

# Remained in experiment, same class type
reg_3rd <- regular %>%
```

```
  filter(g3classtype ==2, flagsg3 == 1)

nreg3 <- nrow(reg_3rd)

sametype_reg <- nreg3/total_reg

# Remained in experiment, changed class type, ended with same school
notreg_sameschool_3rd <- regular %>%
  filter(g3classtype != 2, flagsg3 == 1, gkschid == g3schid)
nnotreg3 <- nrow(notreg_sameschool_3rd)

difftype_reg <- nnotreg3/total_reg

# Remained in experiment, ended with diff school
diffschool_3rd <- regular %>%
  filter(flagsg3 == 1, gkschid != g3schid)
diffschool3 <- nrow(diffschool_3rd)

diffschool_reg <- diffschool3/total_reg

# School left expiriment
school_left <- regular %>%
  filter(gkschid == 128068 | gkschid == 180344 | gkschid == 205489 | gkschid == 216536 |
gkschid ==244818)
schleft <- nrow(school_left)

schleft_exp <- schleft/total_reg

# Left to another public school
school_left_pub <- regular %>%
  filter(flagsg3 != 1) %>%
  filter(flagg4 == 1 | flagg5 == 1)
lp <- nrow(school_left_pub)
left_public_school <- lp/total_reg

# Left Private
pvt <- regular %>%
  filter(flagsg3 != 1) %>%
  filter(flagg4 == 0, flagg5 == 0)
n_pvt <- nrow(pvt)
prop_pvt <- n_pvt/total_reg
```

```{r}
```

```
# Table of Values
tab <- matrix(c(sametype_reg, difftype_reg, diffschool_reg, schleft_exp, left_public_school, prop_pvt) , ncol=1, byrow=TRUE)
rownames(tab) <- c("Remained in expriment, same class type", "Changed class type within school", "Changed schools, but remained in expiriment", "School left expiriment", "Switched to non-participating public school", "Left public school system")
as.table(tab)

# Pie Chart Creation
slices <- c(sametype_small, difftype_small, diffschool_small, schleft_exp, left_public_school, prop_pvt)
lbls <- c("Remained in expriment, same class type", "Changed class type within school", "Changed schools, but remained in expiriment", "School left expiriment", "Switched to non-participating public school", "Left public school system")
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(slices,labels = lbls, col=rainbow(length(lbls)),
   main="Attrition Profile of Students in Small Classes in Grade K")

```


Regular-Aide

```{r}
# Small Class Pie Chart
reg_aide <- star_students %>%
  filter(gkclasstype ==3)

# Total denominator
totalra <- nrow(reg_aide)

# Remained in experiment, same class typeuk
small_3rd <- reg_aide %>%
  filter(g3classtype ==3, flagsg3 == 1)

nsmall3 <- nrow(small_3rd)

sametype_small <- nsmall3/totalra

# Remained in experiment, changed class type, ended with same school
notsmall_sameschool_3rd <- reg_aide %>%
  filter(g3classtype != 3, flagsg3 == 1, gkschid == g3schid)
nnotsmall3 <- nrow(notsmall_sameschool_3rd)
```

```
difftype_small <- nnotsmall3/totalra

# Remained in experiment, ended with diff school
diffschool_3rd <- reg_aide %>%
  filter(flagsg3 == 1, gkschid != g3schid)
diffschool3 <- nrow(diffschool_3rd)

diffschool_small <- diffschool3/totalra

# School left expiriment
school_left <- reg_aide %>%
  filter(gkschid == 128068 | gkschid == 180344 | gkschid == 205489 | gkschid == 216536 |
gkschid ==244818)
schleft <- nrow(school_left)

schleft_exp <- schleft/totalra

# Left to another public school
school_left_pub <- reg_aide %>%
  filter(flagsg3 != 1) %>%
  filter(flagg4 == 1 | flagg5 == 1)
lp <- nrow(school_left_pub)
left_public_school <- lp/totalra

# Left Private
pvt <- reg_aide %>%
  filter(flagsg3 != 1) %>%
  filter(flagg4 == 0, flagg5 == 0)
n_pvt <- nrow(pvt)
prop_pvt <- n_pvt/totalra
```

```{r}
# Table of Values
tab <- matrix(c(sametype_small, difftype_small, diffschool_small, schleft_exp,
left_public_school, prop_pvt) , ncol=1, byrow=TRUE)
rownames(tab) <- c("Remained in expriment, same class type", "Changed class type within
school", "Changed schools, but remained in expiriment", "School left expiriment", "Switched to
non-participating public school", "Left public school system")
as.table(tab)

# Pie Chart Creation
```

```
slices <- c(sametype_small, difftype_small, diffschool_small, schleft_exp, left_public_school, prop_pvt)
lbls <- c("Remained in expriment, same class type", "Changed class type within school", "Changed schools, but remained in expiriment", "School left expiriment", "Switched to non-participating public school", "Left public school system")
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(slices,labels = lbls, col=rainbow(length(lbls)),
   main="Attrition Profile of Students in Small Classes in Grade K")
```

Survival Analysis

```{r}
students_survival <- students %>%
  filter(yearsstar != 0) %>%
  mutate(averagemathscores_acrossexp = (gktmathss + g1tmathss + g2tmathss + g3tmathss)/(yearsstar), freereduced_average = ((gkfreelunch + g1freelunch + g2freelunch + g3freelunch)/(yearsstar))) %>%
  mutate(high_ability = case_when(averagemathscores_acrossexp >= 579 ~ 1, averagemathscores_acrossexp < 579 ~ 0), freereduced_allyears = case_when(freereduced_average == 1 ~ 1, freereduced_average > 1 & freereduced_average < 2 ~ 2, freereduced_average == 2 ~ 3)) %>%
  select(flagsg3, cmpstype, yearsstar, race, gender, stdntid, averagemathscores_acrossexp, high_ability, freereduced_average, freereduced_allyears)

fit_ct <- survfit(Surv(yearsstar, flagsg3) ~ cmpstype, data = students_survival)

ggsurvplot(fit_ct, data = students_survival, title = "Survival Probabilities by Class Type",
        xlab = "Years in Expiriment", ylab = "Estimated Survival Probability",
        conf.int = T, censor = F,legend.labs = c("Small", "Regular", "Aide"))

fit_freereduced <- survfit(Surv(yearsstar, flagsg3) ~ freereduced_allyears, data = students_survival)

ggsurvplot(fit_freereduced, data = students_survival, title = "Survival Probabilities by Free Lunch Status",
        xlab = "Years in Expiriment", ylab = "Estimated Survival Probability",
        conf.int = T, censor = F,legend.labs = c("Free Lunch all years","Free Lunch some years", "Not on Free Lunch"))
```

```
fit_race <- survfit(Surv(yearsstar, flagsg3) ~ race, data = students_survival)

ggsurvplot(fit_race, data = students_survival, title = "Survival Probabilities by Race",
        xlab = "Years in Expiriment", ylab = "Estimated Survival Probability",
        conf.int = F, censor = F,legend.labs = c("White", "Black", "Asian", "Hispanic", "Native American", "Other"))

fit_gender <- survfit(Surv(yearsstar, flagsg3) ~ gender, data = students_survival)

ggsurvplot(fit_gender, data = students_survival, title = "Survival Probabilities by Gender",
        xlab = "Years in Expiriment", ylab = "Estimated Survival Probability",
        conf.int = T, censor = F,legend.labs = c("Male","Female"))


fit_ability <- survfit(Surv(yearsstar, flagsg3) ~ high_ability, data = students_survival)

ggsurvplot(fit_ability, data = students_survival, title = "Survival Probabilities by Student Ability in Math ",
        xlab = "Years in Expiriment", ylab = "Estimated Survival Probability",
        conf.int = T, censor = F,legend.labs = c("75th percentile or higher","Lower than 75th percentile"))

```


Logits by grade status

```{r}
attrition1 <- star_students %>%
  filter(flaggk == 1) %>%
  mutate(total_score = gktmathss + gktreadss + gktlistss + gkwordskillss, gkclasstype = as.factor(gkclasstype), gktgen = as.factor(gktgen), gksurban = as.factor(gksurban), gktrace = as.factor(gktrace), flagg1 = as.factor(flagg1), gkfreelunch = as.factor(case_when(gkfreelunch == 1 ~ 1,gkfreelunch == 2 ~ 0)), gkspeced = as.factor(case_when(gkspeced == 1 ~ 0,gkspeced == 2 ~ 1)), race = as.factor(race), gender = as.factor(gender)) %>%
  select(total_score, gkclasstype, gktgen, gksurban, gktrace, gktyears, gkclasssize, gkfreelunch, gkspeced, flagg1, race, gender)

# Model of attrition between k to 1st grade, note teacher gender not controlled for as there are only female teachers in gk
model_k1 <- glm(flagg1 ~ total_score + gkclasstype + gksurban + gktrace + gktyears + gkclasssize + gkfreelunch + gkspeced + race + gender, data = attrition1, family = "binomial")

#Model summary
```

```
summary(model_k1)
```

```{r fig:model, fig.cap="\\label{fig:model}", fig.align = 'center'}
#Knitting Model summary into a readable table for view in PDF format
model_k1 %>%
  tidy() %>%
  mutate(
    p.value = scales::pvalue(p.value),
    term = c("Intercept", "Total Score", "Regular Class", "Class w/ Aide", "School Suburban",
"School Rural", "School Urban", "Teacher Race: Black", "Teacher Experience (yrs.)", "Class Size",
"Free Lunch", "Special Education", "Student Race: Black", "Student Race: Asian", "Student Race:
Hispanic", "Student Race: Native American", "Student Race: Other", "Student Gender: Female"),
  ) %>%
  kable(
    caption = "Coefficient-Level Estimates for Model Fitted to Estimate Mortality Risk",
    col.names = c("Covariate", "Coefficient", "Standard Error", "Z-Statistic", "P-Value"),
    digits = c(0, 2, 3, 2, 5)

  )

image <- stargazer(model_k1, title="Coefficient-Level Estimates for Model Fitted to Estimate
Attrition between K and 1st Grade", align=TRUE, type = 'latex', header = FALSE, single.row =
FALSE, column.sep.width = "3pt", font.size = "small", omit.stat=c("f"), model.names = FALSE,
notes.align = "l", covariate.labels = c("Total Score", "Regular Class", "Class w/ Aide", "School
Suburban", "School Rural", "School Urban", "Teacher Race: Black", "Teacher Experience (yrs.)",
"Class Size", "Free Lunch", "Special Education", "Student Race: Black", "Student Race: Asian",
"Student Race: Hispanic", "Student Race: Native American", "Student Race: Other", "Student
Gender: Female"))
writeLines(capture.output(image), "/Users/Sanscubed/Desktop/Example1.tex")

```

Peer Effects Try:

```{r}
# Estimating peer effects from attrition between kindergarten and g1

peereffects_k1 <- star_students %>%
  filter(flaggk == 1) %>%
  mutate(total_scorek = gktmathss + gktreadss + gktlistss + gkwordskillss, total_score1 =
gktmathss + gktreadss + gktlistss + gkwordskillss, gkclasstype = as.factor(gkclasstype), gktgen =
as.factor(gktgen), gksurban = as.factor(gksurban), gktrace = as.factor(gktrace), flagg1 =
as.factor(flagg1), gkfreelunch = as.factor(case_when(gkfreelunch == 1 ~ 1,gkfreelunch == 2 ~
```

```
0)), gkspeced = as.factor(case_when(gkspeced == 1 ~ 0,gkspeced == 2 ~ 1)), race =
as.factor(race), gender = as.factor(gender)) %>%
  group_by(gkschid, gktchid) %>%
  mutate(prop_left = 1-(sum(flagsg1==1))/(sum(flagsgk==1)))

peereffects_k1 %>%
  group_by(gkschid, gktchid) %>%
  mutate(leaver_score = case_when(flagsg1 == "1" ~ NA, flagsg1 == "0" ~ total_scorek)) %>%
  mutate(leaver_means = mean(leaver_score))


leave_one_out_mean <- function(x) {
   result <- c()

   for ( i in seq_along(x) ) {
      # note minus-i subsetting is used to subset one observation in each iteration
      # and the na.rm option to handle missing values
      result[i] <- mean(x[-i], na.rm = TRUE)
   }

   return(result)
}

# Use group by but _do not_ pipe the result through summarize()

want <- have %>%
   group_by(CAT) %>%
   mutate(RE = leave_one_out_mean(R),
          IE = leave_one_out_mean(I))
# %>%
#   select(total_scorek, total_score1, gkclasstype, gktgen, gksurban, gktrace, gktyears,
gkclasssize, gkfreelunch, gkspeced, flagg1, race, gender)

```
```
```

---
title: "Thesis_Clustering"
author: "Sanjay Satish"
date: "3/10/2022"
output: html_document

```{r setup, include=FALSE}
library(tidyverse)
library(dplyr)
library(flextable)
library(grid)
library(gridExtra)
library(cowplot)
library(stargazer)
library(scales)
library(pROC)
library(clusterSEs)
library(patchwork)
library(survminer)
library(ggsci)
library(survival)
library(plm)
library(foreign)
library(kableExtra)
library(ggplot2)
library(broom)
library(GGally)
library(tidyr)
library(knitr)
library(xtable)

cluster_data <-
read.csv("/Users/Sanscubed/Desktop/Thesis/Thesis_code/Clustering/PeerEffects_K1_ClusteringData.csv")

cluster_data <- cluster_data %>%
  distinct(stdntid, .keep_all = TRUE)
```

Attempt at K-1 Clustering

Data Wrangling

```{r}
cluster_data_rownames <- cluster_data[,-2]
rownames(cluster_data_rownames) <- cluster_data[,2]

# Data for Clusters based on the following factors (incl things related to income + test scores) :
Race, gender, math score, reading score, free lunch, special ed, special instruction, days absent

cluster_set_all <- cluster_data_rownames %>%
  filter(did_leave == 1) %>%
  mutate(race = as.factor(race), gkfreelunch = as.factor(gkfreelunch), gkspeced =
as.factor(gkspeced), gkspecin = as.factor(gkspecin)) %>%
  select(race, gender, gktmathss, gktreadss, gktlistss, gkwordskillss, gkfreelunch, gkspeced,
gkspecin, gkabsent, gkrepeat)

# Data for clusters NO INCOME/Absences/Special ED
cluster_set_noincome <- cluster_data_rownames %>%
  filter(did_leave == 1) %>%
  mutate(race = as.factor(race), gkfreelunch = as.factor(gkfreelunch), gkspeced =
as.factor(gkspeced), gkspecin = as.factor(gkspecin)) %>%
  select(race, gender, gktmathss, gktreadss, gktlistss, gkwordskillss)

# Data for clusters NO TESTSCORES, but whether or not recommended to repeat kindergarten
cluster_set_noscores <- cluster_data_rownames %>%
  filter(did_leave == 1) %>%
  mutate(race = as.factor(race), gkfreelunch = as.factor(gkfreelunch), gkspeced =
as.factor(gkspeced), gkspecin = as.factor(gkspecin), gkrepeat = as.factor(gkrepeat)) %>%
  select(race, gender, gkfreelunch, gkspeced, gkspecin, gkabsent, gkrepeat)
```

```{r}
library(cluster)
library(factoextra)
library(readr)
library(Rtsne)

cluster_set_noincome <- na.omit(cluster_set_noincome)

gower_dist <- daisy(cluster_set_noincome, metric = "gower")
gower_mat <- as.matrix(gower_dist)

k <- 2
pam_fit <- pam(gower_dist, diss = TRUE, k)

cluster_set_noincome <- cluster_set_noincome %>%
```

```
  mutate(cluster = pam_fit$clustering)

cluster_1 <- cluster_set_noincome %>%
  filter(cluster == 1)

cluster_2 <- cluster_set_noincome %>%
  filter(cluster == 2)

```

```{r}
# C1
sum(cluster_1$race ==1,na.rm = T)/nrow(cluster_1) * 100
sum(cluster_1$race ==2,na.rm = T)/nrow(cluster_1) * 100
sum(cluster_1$race ==3,na.rm = T)/nrow(cluster_1) * 100
sum(cluster_1$race ==4,na.rm = T)/nrow(cluster_1) * 100
sum(cluster_1$race ==5,na.rm = T)/nrow(cluster_1) * 100
sum(cluster_1$race ==6,na.rm = T)/nrow(cluster_1) * 100

mean(cluster_1$gktmathss, na.rm = T)
mean(cluster_1$gktreadss, na.rm = T)
mean(cluster_1$gktlistss, na.rm = T)
mean(cluster_1$gkwordskillss, na.rm = T)

sum(cluster_1$gender ==1,na.rm = T)/nrow(cluster_1) * 100
sum(cluster_1$gkfreelunch ==1,na.rm = T)/nrow(cluster_1) * 100
sum(cluster_1$gkspeced ==1,na.rm = T)/nrow(cluster_1) * 100
sum(cluster_1$gkspecin ==1,na.rm = T)/nrow(cluster_1) * 100

mean(cluster_1$gkabsent, na.rm = T)

# C2

sum(cluster_2$race ==1,na.rm = T)/nrow(cluster_2) * 100
sum(cluster_2$race ==2,na.rm = T)/nrow(cluster_2) * 100
sum(cluster_2$race ==3,na.rm = T)/nrow(cluster_2) * 100
sum(cluster_2$race ==4,na.rm = T)/nrow(cluster_2) * 100
sum(cluster_2$race ==5,na.rm = T)/nrow(cluster_2) * 100
sum(cluster_2$race ==6,na.rm = T)/nrow(cluster_2) * 100

mean(cluster_2$gktmathss, na.rm = T)
mean(cluster_2$gktreadss, na.rm = T)
mean(cluster_2$gktlistss, na.rm = T)
mean(cluster_2$gkwordskillss, na.rm = T)
```

```
sum(cluster_2$gender ==1,na.rm = T)/nrow(cluster_2) * 100
sum(cluster_2$gkfreelunch ==1,na.rm = T)/nrow(cluster_2) * 100
sum(cluster_2$gkspeced ==1,na.rm = T)/nrow(cluster_2) * 100
sum(cluster_2$gkspecin ==1,na.rm = T)/nrow(cluster_2) * 100

mean(cluster_2$gkabsent, na.rm = T)
```

Clustering - Mixed Clusters

All Data

```{r}
library(cluster)
library(factoextra)
library(readr)
library(Rtsne)


# https://towardsdatascience.com/clustering-on-mixed-type-data-8bbd0a2569c3
cluster_set_all <- na.omit(cluster_set_all)

# Data 1
#' Compute Gower distance
gower_dist <- daisy(cluster_set_all, metric = "gower")
gower_mat <- as.matrix(gower_dist)

#' Print most similar students
cluster_set_all[which(gower_mat == min(gower_mat[gower_mat != min(gower_mat)]), arr.ind
= TRUE)[1, ], ]

#' Print most dissimilar students
cluster_set_all[which(gower_mat == max(gower_mat[gower_mat != max(gower_mat)]), arr.ind
= TRUE)[1, ], ]

# Sillouhette
sil_width <- c(NA)
for(i in 2:8){
  pam_fit <- pam(gower_dist, diss = TRUE, k = i)
  sil_width[i] <- pam_fit$silinfo$avg.width
}
```

```
plot(1:8, sil_width,
    xlab = "Number of clusters",
    ylab = "Silhouette Width", main = "Optimal Number of Clusters for Data Including All
Controls")
lines(1:8, sil_width)

# Using k = 6:
k <- 6
pam_fit <- pam(gower_dist, diss = TRUE, k)
pam_results <- cluster_set_all %>%
  mutate(cluster = pam_fit$clustering) %>%
  group_by(cluster) %>%
  do(the_summary = summary(.))
pam_results$the_summary

tsne_obj <- Rtsne(gower_dist, is_distance = TRUE)
tsne_data <- tsne_obj$Y %>%
  data.frame() %>%
  setNames(c("X", "Y")) %>%
  mutate(cluster = factor(pam_fit$clustering))
ggplot(aes(x = X, y = Y), data = tsne_data) +
  geom_point(aes(color = cluster)) + labs(title = "t-SNE Visualization of Clusters in Low-
dimensional Space", subtitle = "For k=6 Clusters - All Controls")

# Using k = 2:

k <- 2
pam_fit <- pam(gower_dist, diss = TRUE, k)
pam_results <- cluster_set_all %>%
  mutate(cluster = pam_fit$clustering) %>%
  group_by(cluster) %>%
  do(the_summary = summary(.))
pam_results$the_summary

tsne_obj <- Rtsne(gower_dist, is_distance = TRUE)
tsne_data <- tsne_obj$Y %>%
  data.frame() %>%
  setNames(c("X", "Y")) %>%
  mutate(cluster = factor(pam_fit$clustering))
ggplot(aes(x = X, y = Y), data = tsne_data) +
  geom_point(aes(color = cluster)) + labs(title = "t-SNE Visualization of Clusters in Low-
dimensional Space", subtitle = "For k=2 Clusters - All Controls")
```

Data w/ no Income

```{r}
cluster_set_noincome <- na.omit(cluster_set_noincome)
# Data 1
#' Compute Gower distance
gower_dist <- daisy(cluster_set_noincome, metric = "gower")
gower_mat <- as.matrix(gower_dist)

#' Print most similar students
cluster_set_noincome[which(gower_mat == min(gower_mat[gower_mat != min(gower_mat)]),
arr.ind = TRUE)[1, ], ]

#' Print most dissimilar students
cluster_set_noincome[which(gower_mat == max(gower_mat[gower_mat != max(gower_mat)]),
arr.ind = TRUE)[1, ], ]

# Sillouhette
sil_width <- c(NA)
for(i in 2:8){
  pam_fit <- pam(gower_dist, diss = TRUE, k = i)
  sil_width[i] <- pam_fit$silinfo$avg.width
}
plot(1:8, sil_width,
     xlab = "Number of clusters",
     ylab = "Silhouette Width", main = "Optimal Number of Clusters for Data Without Income
Controls")
lines(1:8, sil_width)

# Using k = 4:
k <- 4
pam_fit <- pam(gower_dist, diss = TRUE, k)
pam_results <- cluster_set_noincome %>%
  mutate(cluster = pam_fit$clustering) %>%
  group_by(cluster) %>%
  do(the_summary = summary(.))
pam_results$the_summary

tsne_obj <- Rtsne(gower_dist, is_distance = TRUE)
tsne_data <- tsne_obj$Y %>%
  data.frame() %>%
  setNames(c("X", "Y")) %>%
  mutate(cluster = factor(pam_fit$clustering))
ggplot(aes(x = X, y = Y), data = tsne_data) +
```

```
  geom_point(aes(color = cluster)) + labs(title = "t-SNE Visualization of Clusters in Low-dimensional Space", subtitle = "For k=4 Clusters - Data Without Income Controls")

# Using k = 2:

k <- 2
pam_fit <- pam(gower_dist, diss = TRUE, k)
pam_results <- cluster_set_noincome %>%
  mutate(cluster = pam_fit$clustering) %>%
  group_by(cluster) %>%
  do(the_summary = summary(.))
pam_results$the_summary

tsne_obj <- Rtsne(gower_dist, is_distance = TRUE)
tsne_data <- tsne_obj$Y %>%
  data.frame() %>%
  setNames(c("X", "Y")) %>%
  mutate(cluster = factor(pam_fit$clustering))
ggplot(aes(x = X, y = Y), data = tsne_data) +
  geom_point(aes(color = cluster)) + labs(title = "t-SNE Visualization of Clusters in Low-dimensional Space", subtitle = "For k=2 Clusters - Data Without Income Controls")
```

Data w/ no Test Scores

```{r}
cluster_set_noscores <- na.omit(cluster_set_noscores)
# Data 1
#' Compute Gower distance
gower_dist <- daisy(cluster_set_noscores, metric = "gower")
gower_mat <- as.matrix(gower_dist)

#' Print most similar students
cluster_set_noscores[which(gower_mat == min(gower_mat[gower_mat != min(gower_mat)]), arr.ind = TRUE)[1, ], ]

#' Print most dissimilar students
cluster_set_noscores[which(gower_mat == max(gower_mat[gower_mat != max(gower_mat)]), arr.ind = TRUE)[1, ], ]

# Sillouhette
sil_width <- c(NA)
for(i in 2:8){
  pam_fit <- pam(gower_dist, diss = TRUE, k = i)
```

```
  sil_width[i] <- pam_fit$silinfo$avg.width
}
plot(1:8, sil_width,
    xlab = "Number of clusters",
    ylab = "Silhouette Width", main = "Optimal Number of Clusters for Data Without
Achievement Controls")
lines(1:8, sil_width)

# Using k = 6:
k <- 6
pam_fit <- pam(gower_dist, diss = TRUE, k)
pam_results <- cluster_set_noscores %>%
  mutate(cluster = pam_fit$clustering) %>%
  group_by(cluster) %>%
  do(the_summary = summary(.))
pam_results$the_summary

tsne_obj <- Rtsne(gower_dist, is_distance = TRUE)
tsne_data <- tsne_obj$Y %>%
  data.frame() %>%
  setNames(c("X", "Y")) %>%
  mutate(cluster = factor(pam_fit$clustering))
ggplot(aes(x = X, y = Y), data = tsne_data) +
  geom_point(aes(color = cluster)) + labs(title = "t-SNE Visualization of Clusters in Low-
dimensional Space", subtitle = "For k=6 Clusters - Data Without Achievement Controls")

# Using k = 2:

k <- 2
pam_fit <- pam(gower_dist, diss = TRUE, k)
pam_results <- cluster_set_noscores %>%
  mutate(cluster = pam_fit$clustering) %>%
  group_by(cluster) %>%
  do(the_summary = summary(.))
pam_results$the_summary

tsne_obj <- Rtsne(gower_dist, is_distance = TRUE)
tsne_data <- tsne_obj$Y %>%
  data.frame() %>%
  setNames(c("X", "Y")) %>%
  mutate(cluster = factor(pam_fit$clustering))
ggplot(aes(x = X, y = Y), data = tsne_data) +
  geom_point(aes(color = cluster)) + labs(title = "t-SNE Visualization of Clusters in Low-
dimensional Space", subtitle =  "For k=2 Clusters - Data Without Achievement Controls")
```

```

## Regression w/ Fixed Effects on Clusters

```{r}
# Going to run 3 regressions: 3 each for each distance matrix with the 3 different clusters assigned, don't include any of the controls in the real regression as there's endogenity.


```

---
title: "Thesis_Survival"
author: "Sanjay Satish"
date: "3/8/2022"
output: html_document

```{r setup, include=FALSE}
library(tidyverse)
library(dplyr)
library(flextable)
library(grid)

```
library(gridExtra)
library(cowplot)
library(stargazer)
library(scales)
library(pROC)
library(patchwork)
library(survminer)
library(ggsci)
library(survival)
library(foreign)
library(kableExtra)
library(ggplot2)
library(broom)
library(GGally)
library(tidyr)
library(knitr)
library(xtable)

students <-
read.table("/Users/Sanscubed/Desktop/Thesis/Thesis_data/STAR_Archive/STAR_Students.tab",
header = T, sep = "\t", fill = TRUE)
comparison_students <-
read.table("/Users/Sanscubed/Desktop/Thesis/Thesis_data/STAR_Archive/Comparison_Studen
ts.tab", header = T, sep = "\t", fill = TRUE)
schools_spss <-
read.spss("/Users/Sanscubed/Desktop/Thesis/Thesis_data/STAR_Archive/Project
STAR/STAR_K-3_Schools.sav", to.data.frame=TRUE)
schools <- read.table("/Users/Sanscubed/Desktop/Thesis/Thesis_data/STAR_Archive/STAR_K-
3_Schools.tab", header = T, sep = "\t", fill = TRUE)
Converting test scores to numeric types
students$g1tmathss <- as.numeric(students$g1tmathss)
students$g1treadss <- as.numeric(students$g1treadss)
students$g1tlistss <- as.numeric(students$g1tlistss)
students$g1wordskillss <- as.numeric(students$g1wordskillss)

students$gktmathss <- as.numeric(students$gktmathss)
students$gktreadss <- as.numeric(students$gktreadss)
students$gktlistss <- as.numeric(students$gktlistss)
students$gkwordskillss <- as.numeric(students$gkwordskillss)

students$g2tmathss <- as.numeric(students$g2tmathss)
students$g3tmathss <- as.numeric(students$g3tmathss)

```
```
```

Survival Analysis

Data Wrangling:

```{r}
students <- students %>%
  rename(flagsgk = var6, flagsg1 = var7, flagsg2 = var8, flagsg3 = var9)

students_survival <- students %>%
  filter(yearsstar != 0, gkschid != 128068 | gkschid != 180344 | gkschid != 205489 | gkschid != 216536 | gkschid != 244818)
```

Data for Kindergarten Entry Wave

```{r}
survival_k_entry <- students %>%
  filter(flagsgk == 1) %>%
  mutate(did_leave = case_when(yearsstar == 4 ~ 0, yearsstar != 4 ~ 1), when_first_leave = case_when(flagsg1 == 1 & flagsg2 == 1 & flagsg3 == 1 ~ 0,flagsg1 == 0 ~ 1, flagsg2 == 0 & flagsg1 == 1 ~ 2, flagsg1 == 1 & flagsg2 == 1 & flagsg3 == 0 ~ 3)) %>%
  mutate(years_until_first_leave = case_when(when_first_leave == 0 ~ 4, when_first_leave == 1 ~ 1, when_first_leave == 2 ~ 2, when_first_leave == 3 ~ 3))

```

Kaplan-Meier Plot (Only on kids in kindergarten entry wave)

Income (Based on if they're in Free Lunch upon entry)

```{r}
# Using Mine
fit_ct <- survfit(Surv(years_until_first_leave, did_leave) ~ gkfreelunch, data = survival_k_entry)

income_plot <- ggsurvplot(fit_ct, data = survival_k_entry, pval = TRUE, title = "Survival Probabilities by Free Lunch Status in Kindergarten", subtitle = "Among Kindergarten Entry Cohort", risk.table = TRUE,
 tables.height = 0.2, risk.table.title = "Number at risk:",
 tables.theme = theme_cleantable(),
          xlab = "Years in Experiment", ylab = "Estimated Survival Probability",
          conf.int = T, censor = T,legend.labs = c("Free Lunch", "Non-Free Lunch"), palette = "nejm", ggtheme = theme_pubr())
```

```
ggsave("Survival_Income.pdf", plot = ggarrange(income_plot$plot, income_plot$table, nrow=2,
ncol=1, heights=c(3,1)), path = "/Users/Sanscubed/Desktop/Thesis/Thesis_figures/")
```

Ability (Based on Kindergarten Quartile in Math & Reading Combined)

```{r}
survival_k_entry <- survival_k_entry %>%
  mutate(total_math_reading_k = gktmathss + gktreadss + gktlistss + gkwordskillss) %>%
  mutate(gk_ability_quartile = ntile(total_math_reading_k, 100)) %>%
  mutate(gk_ability_category = case_when(gk_ability_quartile <= 25 ~ 1, gk_ability_quartile > 25
& gk_ability_quartile < 75 ~ 2, gk_ability_quartile >= 75 ~ 3))

fit_ct <- survfit(Surv(years_until_first_leave, did_leave) ~ gk_ability_category, data =
survival_k_entry)

ability_plot <- ggsurvplot(fit_ct, data = survival_k_entry, title = "Survival Probabilities by
Kindergarten Test Scores", subtitle = "Among Kindergarten Entry Cohort",
          xlab = "Years in Experiment", ylab = "Estimated Survival Probability",
          conf.int = T, censor = T,legend.labs = c("25th Percentile or Below", "Between 25th-75th
Percentiles", "75th+ Percentile"), risk.table = TRUE,
 tables.height = 0.5, risk.table.title = "Number at risk:",
 tables.theme = theme_cleantable(), palette = "nejm", ggtheme = theme_pubr(), pval = TRUE)

ggsave("Survival_Ability.pdf", plot = ggarrange(ability_plot$plot, ability_plot$table, nrow=2,
ncol=1, heights=c(3,1)), path = "/Users/Sanscubed/Desktop/Thesis/Thesis_figures/")
```

Kindergarten Class Type

```{r}
# Using Mine
fit_ct <- survfit(Surv(years_until_first_leave, did_leave) ~ gkclasstype, data = survival_k_entry)

classtype_plot <- ggsurvplot(fit_ct, data = survival_k_entry, title = "Survival Probabilities by
Kindergarten Class Type", subtitle = "Among Kindergarten Entry Cohort",
          xlab = "Years in Experiment", ylab = "Estimated Survival Probability",
          conf.int = T, censor = T,legend.labs = c("Small", "Regular", "Aide"), risk.table = TRUE,
 tables.height = 0.5, risk.table.title = "Number at risk:",
 tables.theme = theme_cleantable(), palette = "nejm", ggtheme = theme_pubr(), pval = TRUE)

ggsave("Survival_ClassType.pdf", plot = ggarrange(classtype_plot$plot, classtype_plot$table,
nrow=2, ncol=1, heights=c(3,1)), path = "/Users/Sanscubed/Desktop/Thesis/Thesis_figures/")
```

Kindergarten School Rurality

```{r}
fit_ct <- survfit(Surv(years_until_first_leave, did_leave) ~ gksurban, data = survival_k_entry)

schoolurb_plot <- ggsurvplot(fit_ct, data = survival_k_entry, title = "Survival Probabilities by Kindergarten School Urbanicity", subtitle = "Among Kindergarten Entry Cohort",
          xlab = "Years in Experiment", ylab = "Estimated Survival Probability",
          conf.int = T, censor = T,legend.labs = c("Inner City", "Suburban", "Rural", "Urban"),
risk.table = TRUE,
 tables.height = 0.5, risk.table.title = "Number at risk:",
 tables.theme = theme_cleantable(), palette = "nejm", ggtheme = theme_pubr(), pval = TRUE)

ggsave("Survival_SchoolType.pdf", plot = ggarrange(schoolurb_plot$plot, schoolurb_plot$table, nrow=2, ncol=1, heights=c(3,1)), path = "/Users/Sanscubed/Desktop/Thesis/Thesis_figures/")
```

Survival Model for those in Kindergarten Entry Wave

```{r}
# -       Vars of interest:
# switched classes (disaggregate into stayed same, switched from regular to small, switched from small to larger),
#
# -       Controls:
# kindergarten school grade range, exit school grade range, if the school left the experiment, percent students receiving free/reduced lunch in K, percent students bused in K

survival_k_entry <- survival_k_entry %>%
  mutate(switched_schools = case_when(when_first_leave == 0 ~ gkschid!=g3schid, when_first_leave == 1 ~ FALSE, when_first_leave == 2 ~ gkschid!=g1schid, when_first_leave == 3 ~ gkschid!=g2schid), switched_classes = case_when(when_first_leave == 0 ~ gkclasstype!=g3classtype, when_first_leave == 1 ~ FALSE, when_first_leave == 2 ~ gkclasstype!=g1classtype, when_first_leave == 3 ~ gkclasstype!=g1classtype))

schools_k <- schools %>%
  select(schid, var2, var3, var9, var10) %>%
  rename(gkschid = "schid", grdrange = "var2", flag_sk = "var3", gk_fl = "var9", gk_bus = "var10") %>%
  filter(flag_sk == 1)

regression_surv_k <- merge(survival_k_entry,schools_k,by="gkschid")
```

```
# Model of attrition between k to 1st grade, note teacher gender not controlled for as there are only female teachers in gk
weibull_k_entry <- survreg(Surv(years_until_first_leave, did_leave)  ~ as.factor(gkfreelunch) + as.factor(gksurban)  + as.factor(gkclasstype)  + as.factor(gk_ability_category) + as.factor(gender) + as.factor(race) + gktyears + as.factor(gkspeced) + as.factor(gkspecin) + gkabsent + as.factor(grdrange) + gk_fl + gk_bus,
                data=regression_surv_k, dist = "weibull")

#weibull_k_entry$coefficients <- exp(weibull_k_entry$coefficients)
#Model summary
# summary(weibull_k_entry)

# Model for those who were able to switch classes/schools

regression_surv_k2 <- regression_surv_k %>%
  filter(when_first_leave != 1)

weibull_k_entry_nonfirst <- survreg(Surv(years_until_first_leave, did_leave)  ~ as.factor(gkfreelunch) + as.factor(gksurban) + as.factor(switched_schools) + as.factor(gkclasstype) + as.factor(switched_classes) + as.factor(gk_ability_category) + as.factor(gender) + as.factor(race) + gktyears + as.factor(gkspeced) + as.factor(gkspecin) + gkabsent + as.factor(grdrange) + gk_fl + gk_bus,
                data=regression_surv_k2, dist = "weibull")

#summary(weibull_k_entry_nonfirst)
```


Regression Table
```{r fig:model, fig.cap="\\label{fig:model}", fig.align = 'center'}
#Knitting Model summary into a readable table for view in PDF format

complete_table <- stargazer(weibull_k_entry, weibull_k_entry_nonfirst, title="Coefficient-Level Estimates for AFT Model of Attrition for Kindergarten Entry Cohort", align=TRUE, type = 'latex', header = FALSE, single.row = FALSE, column.sep.width = "3pt", font.size = "small", omit.stat=c("f"), model.names = FALSE, notes.align = "l", covariate.labels = c("Non-Free Lunch", "School Suburban", "School Rural", "School Urban", "Switched Schools", "Regular Class", "Regular Class w/ Aide", "Switched Class Types", "Kindergarten Test Score Between 25th-75th Percentile", "Kindergarten Test Score Above 75th Percentile", "Student Gender: Female", "Student Race: Black", "Student Race: Asian", "Student Race: Hispanic", "Student Race: Native American", "Student Race: Other", "Teacher Experience (yrs.)", "Not Special Education", "Not Pulled Out for Special Instruction", "Days Absent", "Kindergarten School K-4", "Kindergarten School K-5", "Kindergarten School K-6", "Kindergarten School K-7", "Kindergarten School K-8",
```

"Kindergarten School K-9", "Percent Kids in Kindergarten School Recieving Free Lunch", "% Kids bused in Kindergarten School"))

Compressed Table

compressed_table <- stargazer(weibull_k_entry, weibull_k_entry_nonfirst, title="Coefficient-Level Estimates for AFT Model of Attrition for Kindergarten Entry Cohort", align=TRUE, type = 'latex', header = FALSE, single.row = FALSE, column.sep.width = "3pt", font.size = "small", omit.stat=c("f"), model.names = FALSE, notes.align = "l", omit = c("gender", "race", "gktyears", "gkabsent", "grdrange", "gk_fl", "gk_bus"), add.lines=list(c("Controls", " Yes", " Yes")), covariate.labels = c("Non-Free Lunch", "School Suburban", "School Rural", "School Urban", "Switched Schools", "Regular Class", "Regular Class w/ Aide", "Switched Class Types", "Kindergarten Test Score Between 25th-75th Percentile", "Kindergarten Test Score Above 75th Percentile", "Not Special Education", "Not Pulled Out for Special Instruction"))

```

### Diagnostics:

```{r}
pdf(file = "/Users/Sanscubed/Desktop/Thesis/Thesis_figures/AFT_Diagnostic_1.pdf")

par(mfrow = c(1,2), cex = .6, mai=c(0.5,0.5,0.5,0.5))
resids <- (log(regression_surv_k$years_until_first_leave) - weibull_k_entry$linear.predictors) / (weibull_k_entry$scale)

rmod <- survfit(Surv(resids, did_leave) ~ 1, data = regression_surv_k)

a <- plot(rmod, main = "Distributional Convergence - AFT Model (1)", xlab = "Residuals for AFT Model of Attrition for Kindergarten Entry Cohort", ylab = "Survival Probability")
exp.x <- seq(min(resids), max(resids), length = 100)
exp.y <- plogis(exp.x, lower.tail = F) # F(t)
lines(exp.x, exp.y, col = "red", lwd = 2)

resids <- (log(regression_surv_k2$years_until_first_leave) - weibull_k_entry_nonfirst$linear.predictors) / (weibull_k_entry_nonfirst$scale)

rmod <- survfit(Surv(resids, did_leave) ~ 1, data = regression_surv_k2)
b <- plot(rmod, main = "Distributional Convergence - AFT Model (2)", xlab = "", ylab = "Survival Probability")
mtext("Residuals for AFT Model of Attrition for Kindergarten Entry Cohort, \n Excluding Students who Left Before 1st Grade", side=1, line=3, cex = .6)
exp.x <- seq(min(resids), max(resids), length = 100)
exp.y <- plogis(exp.x, lower.tail = F) # F(t)
```
```

```
lines(exp.x, exp.y, col = "red", lwd = 2)

dev.off()
```

Survival Model for those in 1st grade entry wave

```{r}
survival_1_entry <- students %>%
  filter(flagsgk == 0 && flagsg1 == 1) %>%
  mutate(did_leave = case_when(yearsstar == 3 ~ 0, yearsstar != 3 ~ 1), when_first_leave =
case_when(flagsg2 == 1 & flagsg3 == 1 ~ 0, flagsg2 == 0 ~ 1, flagsg2 == 1 & flagsg3 == 0 ~ 2))
%>%
  mutate(years_until_first_leave = case_when(when_first_leave == 0 ~ 3, when_first_leave == 1
~ 1, when_first_leave == 2 ~ 2)) %>%
  mutate(total_math_reading_1 = g1tmathss + g1treadss + g1tlistss + g1wordskillss) %>%
  mutate(g1_ability_quartile = ntile(total_math_reading_1, 100)) %>%
  mutate(g1_ability_category = case_when(g1_ability_quartile <= 25 ~ 1, g1_ability_quartile >
25 & g1_ability_quartile < 75 ~ 2, g1_ability_quartile >= 75 ~ 3)) %>%
  mutate(switched_schools = case_when(when_first_leave == 0 ~ g1schid!=g3schid,
when_first_leave == 1 ~ FALSE, when_first_leave == 2 ~ g1schid!=g2schid), switched_classes =
case_when(when_first_leave == 0 ~ g1classtype!=g3classtype, when_first_leave == 1 ~ FALSE,
when_first_leave == 2 ~ g1classtype!=g2classtype))

schools_1 <- schools %>%
  select(schid, var2, var4, var21, var22) %>%
  rename(g1schid = "schid", grdrange = "var2", flag_s1 = "var4", g1_fl = "var21", g1_bus =
"var22") %>%
  filter(flag_s1 == 1)

regression_surv_1 <- merge(survival_1_entry,schools_1,by="g1schid")

# Model for 1st grade entry cohort
weibull_1_entry <- survreg(Surv(years_until_first_leave, did_leave)  ~ as.factor(g1freelunch) +
as.factor(g1surban)  + as.factor(g1classtype)  + as.factor(g1_ability_category) +
as.factor(gender) + as.factor(race) + g1tyears + as.factor(g1speced) + as.factor(g1specin) +
g1absent + as.factor(grdrange) + g1_fl + g1_bus,
                data=regression_surv_1, dist = "weibull")

# Model for those who stayed past 1st grade
regression_surv_f2 <- regression_surv_1 %>%
  filter(when_first_leave != 1)
```

```
weibull_1_entry2 <- survreg(Surv(years_until_first_leave, did_leave)  ~ as.factor(g1freelunch) + as.factor(g1surban) + as.factor(switched_schools) + as.factor(g1classtype) + as.factor(switched_classes) + as.factor(g1_ability_category) + as.factor(gender) + as.factor(race) + g1tyears + as.factor(g1speced) + as.factor(g1specin) + g1absent + as.factor(grdrange) + g1_fl + g1_bus,
                data=regression_surv_f2, dist = "weibull")



#weibull_k_entry$coefficients <- exp(weibull_k_entry$coefficients)
#Model summary
#summary(weibull_1_entry)


#Model summary
#summary(weibull_1_entry2)
```

Regression Table
```{r fig:model, fig.cap="\\label{fig:model}", fig.align = 'center'}
#Knitting Model summary into a readable table for view in PDF format

acomplete_table_first <- stargazer(weibull_1_entry, weibull_1_entry2, title="Coefficient-Level Estimates for AFT Model of Attrition for 1st Grade Entry Cohort", align=TRUE, type = 'latex', header = FALSE, single.row = FALSE, column.sep.width = "3pt", font.size = "small", omit.stat=c("f"), model.names = FALSE, notes.align = "l", covariate.labels = c("Non-Free Lunch", "School Suburban", "School Rural", "School Urban", "Switched Schools", "Regular Class", "Regular Class w/ Aide", "Switched Class Types", "1st Grade Test Score Between 25th-75th Percentile", "1st Grade Test Score Above 75th Percentile", "Student Gender: Female", "Student Race: Black", "Student Race: Asian", "Student Race: Hispanic", "Student Race: Native American", "Student Race: Other", "Teacher Experience (yrs.)", "Not Special Education", "Not Pulled Out for Special Instruction", "Days Absent", "1st Grade School K-4", "1st Grade School K-5", "1st Grade School K-6", "1st Grade School K-7", "1st Grade School K-8", "1st Grade School K-9", "Percent Kids in 1st Grade School Recieving Free Lunch", "% Kids bused in 1st Grade School"))

# Compressed Table w/ Exponentiated coefficients

compressed_table_first <- stargazer(weibull_1_entry, weibull_1_entry2, title="Coefficient-Level Estimates for AFT Model of Attrition for 1st Grade Entry Cohort", align=TRUE, type = 'latex', header = FALSE, single.row = FALSE, column.sep.width = "3pt", font.size = "small", omit.stat=c("f"), model.names = FALSE, notes.align = "l", omit = c("gender", "race", "g1tyears", "g1absent", "grdrange", "g1_fl", "g1_bus"), add.lines=list(c("Controls", "  Yes", "  Yes")), covariate.labels = c("Non-Free Lunch", "School Suburban", "School Rural", "School Urban",
```

"Switched Schools", "Regular Class", "Regular Class w/ Aide", "Switched Class Types", "1st Grade Test Score Between 25th-75th Percentile", "1st Grade Test Score Above 75th Percentile", "Not Special Education", "Not Pulled Out for Special Instruction"))

```

### Diagnostics:

```{r}

pdf(file = "/Users/Sanscubed/Desktop/Thesis/Thesis_figures/AFT_Diagnostic_2.pdf")

par(mfrow = c(1,2), cex = .6, mai=c(0.5,0.5,0.5,0.5))
resids <- (log(regression_surv_1$years_until_first_leave) - weibull_1_entry$linear.predictors) / (weibull_1_entry$scale)

rmod <- survfit(Surv(resids, did_leave) ~ 1, data = regression_surv_1)

a <- plot(rmod, main = "Distributional Convergence - AFT Model (1)", xlab = "Residuals for AFT Model of Attrition for 1st Grade Entry Cohort", ylab = "Survival Probability")
exp.x <- seq(min(resids), max(resids), length = 100)
exp.y <- plogis(exp.x, lower.tail = F) # F(t)
lines(exp.x, exp.y, col = "red", lwd = 2)

resids <- (log(regression_surv_f2$years_until_first_leave) - weibull_1_entry2$linear.predictors) / (weibull_1_entry2$scale)

rmod <- survfit(Surv(resids, did_leave) ~ 1, data = regression_surv_f2)
b <- plot(rmod, main = "Distributional Convergence - AFT Model (2)", xlab = "", ylab = "Survival Probability")
mtext("Residuals for AFT Model of Attrition for 1st Grade Entry Cohort, \n Excluding Students who Left Before 2nd Grade", side=1, line=3, cex = .6)
exp.x <- seq(min(resids), max(resids), length = 100)
exp.y <- plogis(exp.x, lower.tail = F) # F(t)
lines(exp.x, exp.y, col = "red", lwd = 2)

dev.off()

```
```

---
title: "Thesis_PeerEffects"
author: "Sanjay Satish"
date: "3/9/2022"
output: html_document

```{r setup, include=FALSE}
library(tidyverse)
library(dplyr)
library(flextable)
library(grid)
library(gridExtra)
library(cowplot)
library(stargazer)
library(scales)
library(pROC)
library(lfe)
library(clusterSEs)
library(patchwork)
library(survminer)
library(ggsci)
library(survival)
library(plm)
library(foreign)
library(kableExtra)
library(ggplot2)
library(broom)
library(GGally)
library(tidyr)
library(knitr)
library(xtable)

students <-
read.table("/Users/Sanscubed/Desktop/Thesis/Thesis_data/STAR_Archive/STAR_Students.tab",
header = T, sep = "\t", fill = TRUE)
comparison_students <-
read.table("/Users/Sanscubed/Desktop/Thesis/Thesis_data/STAR_Archive/Comparison_Studen
ts.tab", header = T, sep = "\t", fill = TRUE)
schools_spss <-
read.spss("/Users/Sanscubed/Desktop/Thesis/Thesis_data/STAR_Archive/Project
STAR/STAR_K-3_Schools.sav", to.data.frame=TRUE)
schools <- read.table("/Users/Sanscubed/Desktop/Thesis/Thesis_data/STAR_Archive/STAR_K-
3_Schools.tab", header = T, sep = "\t", fill = TRUE)
```

```
# Converting test scores to numeric types
students$g1tmathss <- as.numeric(students$g1tmathss)
students$g1treadss <- as.numeric(students$g1treadss)
students$g1tlistss <- as.numeric(students$g1tlistss)
students$g1wordskillss <- as.numeric(students$g1wordskillss)

students$gktmathss <- as.numeric(students$gktmathss)
students$gktreadss <- as.numeric(students$gktreadss)
students$gktlistss <- as.numeric(students$gktlistss)
students$gkwordskillss <- as.numeric(students$gkwordskillss)

students$g2tmathss <- as.numeric(students$g2tmathss)
students$g3tmathss <- as.numeric(students$g3tmathss)
```

General Peer Effects

Data Wrangling:

```{r}
students <- students %>%
  rename(flagsgk = var6, flagsg1 = var7, flagsg2 = var8, flagsg3 = var9)
```

Kindergarten to First Grade

```{r}
students_peereffects_k1 <- students %>%
  filter(yearsstar != 0,  flagsgk ==1) %>%
  mutate(did_leave = case_when(flagsg1 == 0 ~ 1, flagsg1 == 1 ~ 0)) %>%
  mutate(total_math_reading_k = gktmathss + gktreadss + gktlistss + gkwordskillss,
total_math_reading_1 = g1tmathss + g1treadss + g1tlistss + g1wordskillss, switched_schools =
case_when(gkschid == g1schid ~ 0, gkschid != g1schid ~ 1), gkt_masters =
case_when(gkthighdegree == 2 | gkthighdegree == 1 ~ 0, gkthighdegree >= 3 ~ 1), g1t_masters
= case_when(g1thighdegree == 2 | g1thighdegree == 1 ~ 0, g1thighdegree >= 3 ~ 1))


# Prop Leavers
students_peereffects_k1 <- students_peereffects_k1 %>%
  group_by(gkschid, gktchid) %>%
  mutate(prop_leavers = (sum(did_leave)/gkclasssize)*100) %>%
  mutate(log_prop_leavers = log(prop_leavers))
```

```
# Leaver Means & Variances - Brute Force Method
students_peereffects_k1 <- students_peereffects_k1 %>%
  group_by(gkschid, gktchid) %>%
  mutate(leaver_scores = case_when(did_leave == 1 ~ total_math_reading_k, did_leave == 0 ~
0), leaver_score_exists = case_when(did_leave == 1 & total_math_reading_k >= 0 ~ 1, did_leave
== 0 ~ 0)) %>%
  mutate(leaver_mean = sum(leaver_scores, na.rm = T)/sum(leaver_score_exists, na.rm = T),
leaver_variance = var(leaver_scores[leaver_scores!=0], na.rm=T))
```

```{r}
leave_dta <- students_peereffects_k1 %>%
  filter(flagsg1 == 0)

dep_abl <- ggplot(data=leave_dta[!(is.na(leave_dta$gkfreelunch)), ],
aes(x=total_math_reading_k, group=as.factor(gkfreelunch), fill=as.factor(gkfreelunch))) +
    geom_density(adjust=1.5, alpha=.4) + labs(title = "Distribution of Total Kindergarten Grade
Test Scores by Free Lunch Status", subtitle = "Among Departed Students") + xlab("Total
Kindergarten Grade Test Score") + ylab("Density") + theme_pubr() + scale_fill_nejm(name =
"Free Lunch Status", labels = c("Recieves Free Lunch", "Non Free Lunch")) + xlim(1500, 2750) +
theme(plot.title = element_text(size=13))

ggsave("Departed_Ability.pdf", plot = dep_abl, path =
"/Users/Sanscubed/Desktop/Thesis/Thesis_figures/")

leave_dta[!(is.na(leave_dta$gkfreelunch)), ] %>%
  filter(gkfreelunch == 1) %>%
  select(total_math_reading_k) %>%
  summary(total_math_reading_k)

leave_dta[!(is.na(leave_dta$gkfreelunch)), ] %>%
  filter(gkfreelunch == 2) %>%
  select(total_math_reading_k) %>%
  summary(total_math_reading_k)

```



```{r}
# Leave one out mean for k
students_peereffects_k1 <- students_peereffects_k1 %>%
  group_by(gkschid, gktchid) %>%
  mutate(staying_score_exists = case_when(did_leave == 0 & total_math_reading_k >= 0 ~ 1,
did_leave == 1 ~ 0)) %>%
```

```
  mutate(leave_one_out_avgk = (sum(total_math_reading_k[staying_score_exists==1], na.rm =
T)-total_math_reading_k) / (sum(staying_score_exists, na.rm = T) -1))


# Leave one out mean for g1
# need to get average of all kids so no filtering for k cohort

leave_o1 <- students %>%
  mutate(total_math_reading_1 = g1tmathss + g1treadss + g1tlistss + g1wordskillss) %>%
  group_by(g1schid, g1tchid) %>%
  mutate(leave_one_out_avg1 = (sum(total_math_reading_1, na.rm = T)-
total_math_reading_1)/(length(total_math_reading_1[!is.na(total_math_reading_1)])-1)) %>%
  select(stdntid, leave_one_out_avg1, g1tchid, g1schid)

students_peereffects_k1 <- merge(students_peereffects_k1,leave_o1,by=c("stdntid", "g1tchid",
"g1schid"))



# School Characteristics
schools_1 <- schools %>%
  select(schid, var4, var21, var22) %>%
  rename(g1schid = "schid", flag_s1 = "var4", g1_fl = "var21", g1_bus = "var22") %>%
  filter(flag_s1 == 1)

schools_k <- schools %>%
  select(schid, var2, var3, var9, var10) %>%
  rename(gkschid = "schid", grdrange = "var2", flag_sk = "var3", gk_fl = "var9", gk_bus = "var10")
%>%
  filter(flag_sk == 1)

students_peereffects_k1 <- merge(students_peereffects_k1,schools_k,by="gkschid")

write.csv(students_peereffects_k1,"/Users/Sanscubed/Desktop/Thesis/Thesis_code/Clustering/
PeerEffects_K1_ClusteringData.csv", row.names = FALSE)

students_peereffects_k1 <- merge(students_peereffects_k1,schools_1,by="g1schid")
```


EDA Plots Income & Ability:

```{r}
```

```
fl_test_k <-
ggplot(data=students_peereffects_k1[!(is.na(students_peereffects_k1$gkfreelunch)), ],
aes(x=total_math_reading_k, group=as.factor(gkfreelunch), fill=as.factor(gkfreelunch))) +
   geom_density(adjust=1.5, alpha=.4) + labs(title = "Distribution of Total Kindergarten \nTest
Scores by Free Lunch Status") + xlab("Total Kindergarten Test Score") + ylab("Density") +
theme_pubr() + scale_fill_nejm(name = "Free Lunch Status in Kindergarten", labels =
c("Recieves Free Lunch", "Non Free Lunch")) + xlim(1500, 2750) + theme(plot.title =
element_text(size=13))

fl_test_1 <-
ggplot(data=students_peereffects_k1[!(is.na(students_peereffects_k1$g1freelunch)), ],
aes(x=total_math_reading_1, group=as.factor(g1freelunch), fill=as.factor(g1freelunch))) +
   geom_density(adjust=1.5, alpha=.4) + labs(title = "Distribution of Total First Grade \nTest
Scores by Free Lunch Status") + xlab("Total First Grade Test Score") + ylab("Density") +
theme_pubr() + scale_fill_nejm(name = "Free Lunch Status in First Grade", labels = c("Recieves
Free Lunch", "Non Free Lunch")) + xlim(1500, 2750) + theme(plot.title = element_text(size=13))

arranged <- ggarrange(fl_test_k, fl_test_1, ncol=2, nrow=1, common.legend = TRUE,
legend="bottom")
ggsave("FL_Ability.pdf", plot = arranged, path =
"/Users/Sanscubed/Desktop/Thesis/Thesis_figures/")
```

Regression : Kindergarten to 1st Grade

```{r}
# Regression Data

peereffect_k1_reg_data <- students_peereffects_k1 %>%
  filter(flagsgk == 1 & flagsg1 == 1) %>%
  filter(gkschid == g1schid) %>%
  filter(gkfreelunch == g1freelunch) %>%
  mutate(is_smallk = case_when(gkclasstype == 1 ~ 1, TRUE ~ 0), is_regulark =
case_when(gkclasstype == 2 ~ 1, TRUE ~ 0), is_rak = case_when(gkclasstype == 3 ~ 1, TRUE ~ 0),
is_small1 = case_when(g1classtype == 1 ~ 1, TRUE ~ 0), is_regular1 = case_when(g1classtype ==
2 ~ 1, TRUE ~ 0), is_ra1 = case_when(g1classtype == 3 ~ 1, TRUE ~ 0))

peereffects_no_fe <- felm(total_math_reading_1 ~ total_math_reading_k + prop_leavers +
leaver_mean + leaver_variance + prop_leavers * leaver_mean + prop_leavers * leaver_variance
+ as.factor(is_smallk) + as.factor(is_small1) + leave_one_out_avgk + as.factor(race) +
as.factor(gender) + as.factor(gkfreelunch) + gkabsent + g1absent + as.factor(gkspeced) +
as.factor(gkspecin) + as.factor(g1speced) + as.factor(g1specin) + as.factor(gktrace) +
as.factor(gkt_masters) + gktyears + as.factor(g1trace) + as.factor(g1t_masters) + g1tyears +
```

```
as.factor(gksurban) + as.factor(grdrange) + gk_fl + gk_bus + g1_fl + g1_bus| 0 | 0 |gktchid, data = peereffect_k1_reg_data)

peereffects_with_fe <- felm(total_math_reading_1 ~ total_math_reading_k + prop_leavers + leaver_mean + leaver_variance + prop_leavers * leaver_mean + prop_leavers * leaver_variance + as.factor(is_smallk) + as.factor(is_small1) + leave_one_out_avgk + as.factor(race) + as.factor(gender) + as.factor(gkfreelunch) + gkabsent + g1absent + as.factor(gkspeced) + as.factor(gkspecin) + as.factor(g1speced) + as.factor(g1specin) + as.factor(gktrace) + as.factor(gkt_masters) + gktyears + as.factor(g1trace) + as.factor(g1t_masters) + g1tyears| gkschid | 0 |gktchid, data = peereffect_k1_reg_data)

peereffects_with_fe_noprop <- felm(total_math_reading_1 ~ total_math_reading_k + leaver_mean + as.factor(is_smallk) + as.factor(is_small1) + leave_one_out_avgk + as.factor(race) + as.factor(gender) + as.factor(gkfreelunch) + gkabsent + g1absent + as.factor(gkspeced) + as.factor(gkspecin) + as.factor(g1speced) + as.factor(g1specin) + as.factor(gktrace) + as.factor(gkt_masters) + gktyears + as.factor(g1trace) + as.factor(g1t_masters) + g1tyears| gkschid | 0 |gktchid, data = peereffect_k1_reg_data)

peereffects_with_fe_onlyprop <- felm(total_math_reading_1 ~ total_math_reading_k + prop_leavers + leaver_mean + leaver_variance + prop_leavers * leaver_mean + prop_leavers * leaver_variance + as.factor(is_smallk) + as.factor(is_small1) + leave_one_out_avgk + as.factor(race) + as.factor(gender) + as.factor(gkfreelunch) + gkabsent + g1absent + as.factor(gkspeced) + as.factor(gkspecin) + as.factor(g1speced) + as.factor(g1specin) + as.factor(gktrace) + as.factor(gkt_masters) + gktyears + as.factor(g1trace) + as.factor(g1t_masters) + g1tyears| gkschid | 0 |gktchid, data = peereffect_k1_reg_data)

```

More EDA:

```{r}
ggplot(peereffect_k1_reg_data, aes(x=leaver_mean, y=total_math_reading_1)) +
  geom_point() +
  geom_smooth(method=lm , color="red", fill="#69b3a2", se=TRUE) + labs(title = "Own Test Score vs. Leaver Mean Test Score") + theme_pubr() + ggpubr::stat_cor(method="pearson")

ggplot(peereffect_k1_reg_data, aes(x=leave_one_out_avgk, y=total_math_reading_1)) +
  geom_point() +
  geom_smooth(method=lm , color="red", fill="#69b3a2", se=TRUE) + labs(title = "Own Test Score vs. Kindergarten Peer Mean Test Score")  + theme_pubr() +
ggpubr::stat_cor(method="pearson")

ggplot(peereffect_k1_reg_data, aes(x=total_math_reading_k, y=total_math_reading_1)) +
  geom_point() +
```

```
  geom_smooth(method=lm , color="red", fill="#69b3a2", se=TRUE) + labs(title = "Own Test Score vs. Own Kindergarten Test Score") + theme_pubr() +
ggpubr::stat_cor(method="pearson")

ggplot(peereffect_k1_reg_data, aes(x=prop_leavers, y=leaver_mean)) +
  geom_point() +
  geom_smooth(method=lm , color="red", fill="#69b3a2", se=TRUE) + labs(title = "Proportion of Class who Left vs. Mean Test Score of Leavers") + theme_pubr() +
ggpubr::stat_cor(method="pearson")

leave_k <- ggplot(peereffect_k1_reg_data, aes(x=prop_leavers, y=total_math_reading_1)) +
  geom_point() +
  geom_smooth(method=lm , color="red", fill="#69b3a2", se=TRUE) + labs(title = "Proportion of Class who Left \nvs. Own First Grade Test Score") + theme_pubr() +
ggpubr::stat_cor(method="pearson" , label.x = 30, label.y = 2650) + xlab("Proportion of Class \nWho Left in Kindergarten") + ylab("Total First Grade Test Score") + ylim(1600, 2650)

leave_1 <- ggplot(peereffect_k1_reg_data, aes(x=prop_leavers, y=total_math_reading_k)) +
  geom_point() +
  geom_smooth(method=lm , color="red", fill="#69b3a2", se=TRUE) + labs(title = "Proportion of Class who Left vs. \nOwn Kindergarten Test Score ") + theme_pubr() +
ggpubr::stat_cor(method="pearson", label.x = 30, label.y = 2650) + ylim(1600, 2650) +
xlab("Proportion of Class \nWho Left in Kindergarten") + ylab("Total Kindergarten Grade Test Score")

arranged <- ggarrange(leave_k, leave_1, ncol=2, nrow=1)
ggsave("PropLeave.pdf", plot = arranged, path =
"/Users/Sanscubed/Desktop/Thesis/Thesis_figures/")

 leave_k <- ggplot(peereffect_k1_reg_data, aes(x=leave_one_out_avgk,
y=total_math_reading_k)) +
  geom_point() +
  geom_smooth(method=lm , color="red", fill="#69b3a2", se=TRUE) + labs(title = "Leave-One-Out Average of \nPeers Kindergarten Test Scores \nvs. Own Test Score") + theme_pubr() +
ggpubr::stat_cor(method="pearson", label.x = 1800, label.y = 2600) + ylim(1600, 2600) +
xlim(1600, 2200) + xlab("Leave-One-Out Average of Peers \nKindergarten Test Scores") +
ylab("Total Own Kindergarten Test Score")

leave_1 <- ggplot(peereffect_k1_reg_data, aes(x=leaver_mean, y=total_math_reading_k)) +
  geom_point() +
  geom_smooth(method=lm , color="red", fill="#69b3a2", se=TRUE) + labs(title = "Own Test Score \nvs. Leaver Mean Test Score") + theme_pubr() + ggpubr::stat_cor(method="pearson", label.x = 1800, label.y = 2600) + ylim(1600, 2600) + xlab("Leaver's Mean \nKindergarten Test Score") + ylab("Total Own Kindergarten Test Score") + xlim(1600, 2200)
```

```
arranged <- ggarrange(leave_k, leave_1, ncol=2, nrow=1)
ggsave("XCorr.pdf", plot = arranged, path =
"/Users/Sanscubed/Desktop/Thesis/Thesis_figures/")

```


```{r}
fulltable <- stargazer(peereffects_no_fe, peereffects_with_fe, peereffects_with_fe_noprop,
peereffects_with_fe_onlyprop, title="Coefficient-Level Estimates for Peer Effects among
Kindergarten Entry Cohort", align=TRUE, type = 'latex', header = FALSE, single.row = FALSE,
column.sep.width = "3pt", font.size = "small", omit.stat=c("f"), model.names = FALSE,
notes.align = "l", omit = c("gender", "race", "gktyears", "gkabsent", "grdrange", "gk_fl",
"gk_bus", "gkfreelunch", "g1absent", "gkspeced", "gkspecin", 'g1speced', "g1specin", "gktrace",
"gkt_masters", "g1trace",  "g1t_masters",  "g1tyears",  "gksurban", "g1_fl", "g1_bus"),
add.lines=list(c("Mean pm SD of Response",""),c("Fixed Effects", "No", "Yes", "Yes", "Yes")))

```


Regression Tables

```{r}
compressed_table <- stargazer(peereffects_no_fe, peereffects_with_fe,
peereffects_with_fe_noprop, peereffects_with_fe_onlyprop, title="Coefficient-Level Estimates
for Peer Effects among Kindergarten Entry Cohort", align=TRUE, type = 'latex', header = FALSE,
single.row = FALSE, column.sep.width = "3pt", font.size = "small", omit.stat=c("f"),
model.names = FALSE, notes.align = "l", omit = c("total_math_reading_k" , "prop_leavers" ,
"leaver_mean" , "leaver_variance", "is_smallk", "is_small1" , "leave_one_out_avgk",
"gkfreelunch" , "gkabsent" , "g1absent" , "gkspeced" , "gkspecin" , "g1speced" , "g1specin" ,
"gktrace" , "gkt_masters" , "gktyears" , "g1trace" , "g1t_masters"), add.lines=list(c("Mean pm
SD of Response",""),c("Fixed Effects", "No", "Yes", "Yes", "Yes")))

m1 <- felm(formula = total_math_reading_1 ~ total_math_reading_k| gkschid | 0 |gktchid,
data = peereffect_k1_reg_data)
m2 <- felm(formula = total_math_reading_1 ~ leave_one_out_avgk| gkschid | 0 |gktchid, data
= peereffect_k1_reg_data)
m3 <- felm(formula = total_math_reading_1 ~ leaver_mean| gkschid | 0 |gktchid, data =
peereffect_k1_reg_data)
m4 <- felm(formula = total_math_reading_1 ~ total_math_reading_k + leave_one_out_avgk|
gkschid | 0 |gktchid, data = peereffect_k1_reg_data)
```

```
m5 <- felm(formula = total_math_reading_1 ~ total_math_reading_k + leave_one_out_avgk + leaver_mean + total_math_reading_k*leave_one_out_avgk + total_math_reading_k*leaver_mean| gkschid | 0 |gktchid, data = peereffect_k1_reg_data)

bias_table <- stargazer(m1, m2, m3, m4, m5, title="Coefficient-Level Estimates for First Grade Test Score on Lagged Achievement Measures", align=TRUE, type = 'latex', header = FALSE, single.row = FALSE, column.sep.width = "3pt", font.size = "small", omit.stat=c("f"), model.names = FALSE, notes.align = "l", add.lines=list(c("Mean pm SD of Response", "2163 pm 169")))
```

Diagnostics

```{r}
# Create Robustness model for measuring levergae
fe_model_robust <- lm(total_math_reading_1 ~ total_math_reading_k + prop_leavers + leaver_mean + leaver_variance + prop_leavers * leaver_mean + prop_leavers * leaver_variance + as.factor(gkclasstype) + as.factor(g1classtype) + leave_one_out_avgk + leave_one_out_avg1 + as.factor(race) + as.factor(gender) + as.factor(gkfreelunch) + as.factor(g1freelunch) + gkabsent + g1absent + as.factor(gkspeced) + as.factor(gkspecin) + as.factor(g1speced) + as.factor(g1specin) + as.factor(gktrace) + as.factor(gkt_masters) + gktyears + as.factor(g1trace) + as.factor(g1t_masters) + g1tyears + as.factor(gksurban) + as.factor(grdrange) + gk_fl + gk_bus + g1_fl + g1_bus + gkschid - 1, data = peereffect_k1_reg_data)

# Augment dataset for plotting leverage and residuals
panel_output <- augment(fe_model_robust) %>%
  mutate(obs_num = row_number())

# Add residuals and standardize
panel_output <- panel_output %>%
  mutate(predicted = predict.lm(fe_model_robust), resid = residuals(fe_model_robust))

# Create Leverage threshold
leverage_threshold <- 2*(32+1)/(nrow(panel_output))

# Plot and save Figure 4
ggplot(data = panel_output ,aes(x = obs_num, y = .hat)) +
  geom_point(alpha = 0.7) +
  geom_hline(yintercept = leverage_threshold,color = "red")+
  labs(x = "Observation Number", y = "Leverage",title = "Scatterplot of Observations with Respect to Leverage for Fixed Effects Model", caption = "Leverage Threshold in Red")


# Plot Figure 5 and save to device
```

```
ggplot(data = panel_output, aes(x = obs_num, y = .cooksd)) +
  geom_point(alpha = 0.7) +
  geom_hline(yintercept=1,color = "red")+
  labs(x= "Observation Number",y = "Cook's Distance",title = "Scatterplot of Observations with
Respect to Cook's Distance for Fixed Effects Model", caption = "Cook's Distance Threshold in
Red") +
  geom_text(aes(label = ifelse(.hat>1,as.character(obs_num),"")))


# Plot Figure 6 and save to device
ggplot(data = panel_output, aes(x = .fitted,y = .std.resid)) +
  geom_point(alpha = 0.7) +
  geom_hline(yintercept = 0,color = "red") +
  geom_hline(yintercept = -2,color = "red",linetype = "dotted") +
  geom_hline(yintercept = 2,color = "red",linetype = "dotted") +
  labs(x ="Predicted Value",y ="Standardized Residuals",title = "Standardized Residuals (Normal
~ (0,1)) vs. Predicted Values for ", caption = "Observation Numbers Labeled for those Outside
+/- 2 Standard Deviations From the Mean") +
  geom_text(aes(label = ifelse(abs(.std.resid) >2,as.character(obs_num),"")), nudge_x = 0.08)


# Pull Observations outside +/- 2 SD of residuals
highresiduals <- panel_output %>% filter(abs(.std.resid) > 2)

# Table 4

# refitting model
refit_data <- peereffect_k1_reg_data[-c(6,   14,   116,  137,  169,  185,  361,  362,  363,  370,
376,  378,  380,  399,  466,  495,  535,  538,  576,  600,  750,
763,  850,  869,  923,  925,  958,  998,  1002, 1020, 1027, 1036, 1041, 1073, 1255, 1266, 1395,
1609, 1631, 1692, 1701, 1713,
1862, 1936, 1982, 1989, 1995, 1999, 2013, 2029, 2163, 2269, 2294, 2295, 2297, 2457, 2486,
2555, 2558, 2626, 2671, 2687, 2690,
 2706, 2716, 2735, 2811, 2847, 2914, 2920, 2962, 2981, 3035, 3149, 3151, 3229, 3244, 3246,
3248, 3258, 3264, 3303, 3351, 3354,
 3361, 3377, 3393, 3403, 3425, 3434, 3452, 3460, 3481, 3483, 3636, 3670, 3679, 3698, 3720,
3743, 3829, 3843, 3859, 3860, 3949,
4021, 4029, 4054, 4060, 4143, 4175, 4228, 4238), ]

refit_with_fe <- plm(total_math_reading_1 ~ total_math_reading_k + prop_leavers +
leaver_mean + leaver_variance + prop_leavers * leaver_mean + prop_leavers * leaver_variance
+ as.factor(gkclasstype) + as.factor(g1classtype) + leave_one_out_avgk + leave_one_out_avg1 +
as.factor(race) + as.factor(gender) + as.factor(gkfreelunch) + as.factor(g1freelunch) + gkabsent
+ g1absent + as.factor(gkspeced) + as.factor(gkspecin) + as.factor(g1speced) +
```

```
as.factor(g1specin) + as.factor(gktrace) + as.factor(gkt_masters) + gktyears + as.factor(g1trace)
+ as.factor(g1t_masters) + g1tyears + as.factor(gksurban) + as.factor(grdrange) + gk_fl + gk_bus
+ g1_fl + g1_bus, data = refit_data,
                index = c("gkschid"),
                model = "within")

```

---
title: "Thesis_PeerEffects_K2"
author: "Sanjay Satish"
date: "3/10/2022"
output: html_document

```{r setup, include=FALSE}
library(tidyverse)
library(dplyr)
library(flextable)
library(grid)
library(gridExtra)
library(cowplot)
library(stargazer)
library(scales)
library(pROC)
library(clusterSEs)
library(patchwork)
library(survminer)
library(ggsci)
library(survival)
library(plm)
library(foreign)
library(kableExtra)
library(ggplot2)
library("lmtest")
library("sandwich")
library(lfe)
library(broom)
library(GGally)
library(tidyr)
library(knitr)
library(xtable)

students <-
read.table("/Users/Sanscubed/Desktop/Thesis/Thesis_data/STAR_Archive/STAR_Students.tab",
header = T, sep = "\t", fill = TRUE)
comparison_students <-
read.table("/Users/Sanscubed/Desktop/Thesis/Thesis_data/STAR_Archive/Comparison_Studen
ts.tab", header = T, sep = "\t", fill = TRUE)
schools_spss <-
read.spss("/Users/Sanscubed/Desktop/Thesis/Thesis_data/STAR_Archive/Project
STAR/STAR_K-3_Schools.sav", to.data.frame=TRUE)
```

```
schools <- read.table("/Users/Sanscubed/Desktop/Thesis/Thesis_data/STAR_Archive/STAR_K-3_Schools.tab", header = T, sep = "\t", fill = TRUE)
# Converting test scores to numeric types
students$g1tmathss <- as.numeric(students$g1tmathss)
students$g1treadss <- as.numeric(students$g1treadss)
students$g1tlistss <- as.numeric(students$g1tlistss)
students$g1wordskillss <- as.numeric(students$g1wordskillss)

students$gktmathss <- as.numeric(students$gktmathss)
students$gktreadss <- as.numeric(students$gktreadss)
students$gktlistss <- as.numeric(students$gktlistss)
students$gkwordskillss <- as.numeric(students$gkwordskillss)

students$g2tmathss <- as.numeric(students$g2tmathss)
students$g2treadss <- as.numeric(students$g2treadss)
students$g2tlistss <- as.numeric(students$g2tlistss)
students$g2wordskillss <- as.numeric(students$g2wordskillss)
```

General Peer Effects

Data Wrangling:

```{r}
students <- students %>%
  rename(flagsgk = var6, flagsg1 = var7, flagsg2 = var8, flagsg3 = var9)
```

First to Second Grade

```{r}
students_peereffects_12 <- students %>%
  filter(yearsstar != 0,  flagsg1 ==1) %>%
  mutate(did_leave = case_when(flagsg2 == 0 ~ 1, flagsg2 == 1 ~ 0)) %>%
  mutate(total_math_reading_2 = g2tmathss + g2treadss + g2tlistss + g2wordskillss,
total_math_reading_1 = g1tmathss + g1treadss + g1tlistss + g1wordskillss, switched_schools =
case_when(g2schid == g1schid ~ 0, g2schid != g1schid ~ 1), g2t_masters =
case_when(g2thighdegree == 2 | g2thighdegree == 1 ~ 0, g2thighdegree >= 3 ~ 1), g1t_masters
= case_when(g1thighdegree == 2 | g1thighdegree == 1 ~ 0, g1thighdegree >= 3 ~ 1))


# Prop Leavers
students_peereffects_12 <- students_peereffects_12 %>%
```

```
  group_by(g1schid, g1tchid) %>%
  mutate(prop_leavers = (sum(did_leave)/g1classsize)*100) %>%
  mutate(log_prop_leavers = log(prop_leavers))

# Leaver Means & Variances - Brute Force Method
students_peereffects_12 <- students_peereffects_12 %>%
  group_by(g1schid, g1tchid) %>%
  mutate(leaver_scores = case_when(did_leave == 1 ~ total_math_reading_1, did_leave == 0 ~
0), leaver_score_exists = case_when(did_leave == 1 & total_math_reading_1 >= 0 ~ 1,
did_leave == 0 ~ 0)) %>%
  mutate(leaver_mean = sum(leaver_scores, na.rm = T)/sum(leaver_score_exists, na.rm = T),
leaver_variance = var(leaver_scores[leaver_scores!=0], na.rm=T))
```

```{r}
# Leave one out mean for 1
students_peereffects_12 <- students_peereffects_12 %>%
  group_by(g1schid, g1tchid) %>%
  mutate(staying_score_exists = case_when(did_leave == 0 & total_math_reading_1 >= 0 ~ 1,
did_leave == 1 ~ 0)) %>%
  mutate(leave_one_out_avg1 = (sum(total_math_reading_1[staying_score_exists==1], na.rm =
T)-total_math_reading_1) / (sum(staying_score_exists, na.rm = T) -1)) %>%
  filter(is.finite(leave_one_out_avg1 == T))


# Leave one out mean for g2
# need to get average of all kids so no filtering for 1 cohort

leave_o1 <- students %>%
  group_by(g2schid, g2tchid) %>%
  mutate(total_math_reading_2 = g2tmathss + g2treadss + g2tlistss + g2wordskillss) %>%
  mutate(leave_one_out_avg2 = (sum(total_math_reading_2, na.rm = T)-
total_math_reading_2)/(length(total_math_reading_2[!is.na(total_math_reading_2)])-1)) %>%
  select(stdntid, leave_one_out_avg2, g2tchid, g2schid)%>%
  filter(is.finite(leave_one_out_avg2 == T))


students_peereffects_12 <- merge(students_peereffects_12,leave_o1,by=c("stdntid",
"g2tchid", "g2schid"))

# School Characteristics
schools_1 <- schools %>%
  select(schid, var2, var4, var21, var22) %>%
```

```
  rename(g1schid = "schid",grdrange = "var2", flag_s1 = "var4", g1_fl = "var21", g1_bus =
"var22") %>%
  filter(flag_s1 == 1)

schools_2 <- schools %>%
  select(schid, var5, var33, var34) %>%
  rename(g2schid = "schid", flag_s2 = "var5", g2_fl = "var33", g2_bus = "var34") %>%
  filter(flag_s2 == 1)

students_peereffects_12 <- merge(students_peereffects_12,schools_1,by="g1schid")

students_peereffects_12 <- merge(students_peereffects_12,schools_2,by="g2schid")


```

Regression : Kindergarten to 1st Grade

```{r}
# Regression Data

peereffect_12_reg_data <- students_peereffects_12 %>%
  filter(flagsg1 == 1 & flagsg2 == 1) %>%
  filter(g1schid == g2schid) %>%
  mutate(is_small2 = case_when(g2classtype == 1 ~ 1, TRUE ~ 0), is_regular2 =
case_when(g2classtype == 2 ~ 1, TRUE ~ 0), is_rak = case_when(g2classtype == 3 ~ 1, TRUE ~ 0),
is_small1 = case_when(g1classtype == 1 ~ 1, TRUE ~ 0), is_regular1 = case_when(g1classtype ==
2 ~ 1, TRUE ~ 0), is_ra1 = case_when(g1classtype == 3 ~ 1, TRUE ~ 0))


peereffects_no_fe <- felm(total_math_reading_2 ~ total_math_reading_1 + prop_leavers +
leaver_mean + leaver_variance + prop_leavers * leaver_mean + prop_leavers * leaver_variance
+ as.factor(is_small1) + as.factor(is_small2) + leave_one_out_avg1 + as.factor(race) +
as.factor(gender) + as.factor(g1freelunch) + g1absent + as.factor(g1speced) +
as.factor(g1specin) + as.factor(g2trace) + as.factor(g2t_masters) + g2tyears + as.factor(g1trace)
+ as.factor(g1t_masters) + g1tyears + as.factor(g2surban) + as.factor(grdrange) + g2_fl +
g2_bus| 0 | 0 |g1tchid, data = peereffect_12_reg_data)

test <- peereffect_12_reg_data %>%
  select(total_math_reading_2, total_math_reading_1, prop_leavers, is_small1 , is_small2 ,
leave_one_out_avg1 , race , gender , g1freelunch , g1absent , g1speced , g1specin , g2trace ,
g2t_masters , g2tyears , g1trace , g1t_masters , g1tyears ) %>%
  na.omit()
```

```
mean(test$total_math_reading_2)
sd(test$total_math_reading_2)

peereffects_with_fe <- felm(total_math_reading_2 ~ total_math_reading_1 + prop_leavers +
leaver_mean + leaver_variance + prop_leavers * leaver_mean + prop_leavers * leaver_variance
+ as.factor(is_small1) + as.factor(is_small2) + leave_one_out_avg1 + as.factor(race) +
as.factor(gender) + as.factor(g1freelunch) + g1absent + as.factor(g1speced) +
as.factor(g1specin) + as.factor(g2trace) + as.factor(g2t_masters) + g2tyears + as.factor(g1trace)
+ as.factor(g1t_masters) + g1tyears| g1schid | 0 |g1tchid, data = peereffect_12_reg_data)

peereffects_with_fe_noprop <- felm(total_math_reading_2 ~ total_math_reading_1 +
prop_leavers + leaver_mean + as.factor(is_small1) + as.factor(is_small2) + leave_one_out_avg1
+ as.factor(race) + as.factor(gender) + as.factor(g1freelunch) + g1absent + as.factor(g1speced) +
as.factor(g1specin) + as.factor(g2trace) + as.factor(g2t_masters) + g2tyears + as.factor(g1trace)
+ as.factor(g1t_masters) + g1tyears| g1schid | 0 |g1tchid, data = peereffect_12_reg_data)

peereffects_with_fe_onlyprop <- felm(total_math_reading_2 ~ total_math_reading_1 +
prop_leavers + as.factor(is_small1) + as.factor(is_small2) + leave_one_out_avg1 +
as.factor(race) + as.factor(gender) + as.factor(g1freelunch) + g1absent + as.factor(g1speced) +
as.factor(g1specin) + as.factor(g2trace) + as.factor(g2t_masters) + g2tyears + as.factor(g1trace)
+ as.factor(g1t_masters) + g1tyears| g1schid | 0 |g1tchid, data = peereffect_12_reg_data)
```

Regression Tables

```{r}

full_table <- stargazer(peereffects_no_fe, peereffects_with_fe, peereffects_with_fe_noprop,
peereffects_with_fe_onlyprop, title="Coefficient-Level Estimates for Peer Effects among
Kindergarten Entry Cohort", align=TRUE, type = 'latex', header = FALSE, single.row = FALSE,
column.sep.width = "3pt", font.size = "small", omit.stat=c("f"), model.names = FALSE,
notes.align = "l", omit = c("total_math_reading_1", "prop_leavers" ,  "leaver_mean",
"is_small1", "is_small2", "leave_one_out_avg1", "g1freelunch"), add.lines=list(c("Fixed Effects",
"No", "Yes", "Yes", "Yes")))

compressed_table <- stargazer(peereffects_no_fe, peereffects_with_fe,
peereffects_with_fe_noprop, peereffects_with_fe_onlyprop, title="Coefficient-Level Estimates
for Peer Effects among Kindergarten Entry Cohort", align=TRUE, type = 'latex', header = FALSE,
single.row = FALSE, column.sep.width = "3pt", font.size = "small", omit.stat=c("f"),
model.names = FALSE, notes.align = "l", omit = c("gender", "race", "g2tyears", "g2absent",
"grdrange", "g2_fl", "g2_bus", "g1absent", "g2speced", "g2specin", 'g1speced', "g1specin",
"g2trace", "g2t_masters", "g1trace",  "g1t_masters",  "g1tyears",  "g2surban", "g1_fl",
"g1_bus"), add.lines=list(c("Fixed Effects", "No", "Yes", "Yes", "Yes")))
```

```
library(tidyverse)
library(dplyr)
library(flextable)
library(grid)
library(gridExtra)
library(cowplot)
library(stargazer)
library(scales)
library(patchwork)
library(survminer)
library(survival)
library(kableExtra)
library(ggplot2)
library(broom)
library(tidyr)
library(knitr)
library(xtable)

students <-
read.table("/Users/Sanscubed/Desktop/Thesis/Thesis_data/STAR_Archive/STAR_Students.tab",
header = T, sep = "\t", fill = TRUE)
schools <- read.table("/Users/Sanscubed/Desktop/Thesis/Thesis_data/STAR_Archive/STAR_K-
3_Schools.tab", header = T, sep = "\t", fill = TRUE)

students$g1tmathss <- as.numeric(students$g1tmathss)
students$g1treadss <- as.numeric(students$g1treadss)
students$g1tlistss <- as.numeric(students$g1tlistss)
students$g1wordskillss <- as.numeric(students$g1wordskillss)

students$gktmathss <- as.numeric(students$gktmathss)
students$gktreadss <- as.numeric(students$gktreadss)
students$gktlistss <- as.numeric(students$gktlistss)
students$gkwordskillss <- as.numeric(students$gkwordskillss)

# Grade 1 Summary Table
small<- students %>%
  filter(g1classtype == 1)

regular <- students %>%
  filter(g1classtype == 2)

regularaide <- students %>%
  filter(g1classtype == 3)
```

```
n.small <- nrow(small)
n.reg <- nrow(regular)
n.rega <- nrow(regularaide)

# Building Table
nyears <- c(paste0(format(round(mean(small$yearsstar), digits = 1),nsmall = 1), "±",
format(round(sd(small$yearsstar), digits =1), nsmall = 1)),
paste0(format(round(mean(regular$yearsstar), digits = 1),nsmall = 1), "±",
format(round(sd(regular$yearsstar), digits =1), nsmall = 1)),
paste0(format(round(mean(regularaide$yearsstar), digits = 1),nsmall = 1), "±",
format(round(sd(regularaide$yearsstar), digits =1), nsmall = 1)))
nyears_small <- c(paste0(format(round(mean(small$yearssmall), digits = 1),nsmall = 1), "±",
format(round(sd(small$yearssmall), digits =1), nsmall = 1)),
paste0(format(round(mean(regular$yearssmall), digits = 1),nsmall = 1), "±",
format(round(sd(regular$yearssmall), digits =1), nsmall = 1)),
paste0(format(round(mean(regularaide$yearssmall), digits = 1),nsmall = 1), "±",
format(round(sd(regularaide$yearssmall), digits =1), nsmall = 1)))

  # School Urbanicity
sch_inner <- c(paste0(format(round((sum(small$g1surban ==1, na.rm = TRUE )/nrow(small) *
100), digits = 1), nsmall = 1), "\\%"), paste0(format(round((sum(regular$g1surban ==1, na.rm =
TRUE )/nrow(regular) * 100), digits = 1), nsmall = 1), "\\%"),
paste0(format(round((sum(regularaide$g1surban ==1, na.rm = TRUE )/nrow(regularaide) *
100), digits = 1), nsmall = 1), "\\%"))
sch_suburban <- c(paste0(format(round((sum(small$g1surban ==2, na.rm = TRUE )/nrow(small)
* 100), digits = 1), nsmall = 1), "\\%"), paste0(format(round((sum(regular$g1surban ==2, na.rm
= TRUE )/nrow(regular) * 100), digits = 1), nsmall = 1), "\\%"),
paste0(format(round((sum(regularaide$g1surban ==2, na.rm = TRUE )/nrow(regularaide) *
100), digits = 1), nsmall = 1), "\\%"))
sch_rural <- c(paste0(format(round((sum(small$g1surban ==3, na.rm = TRUE )/nrow(small) *
100), digits = 1), nsmall = 1), "\\%"), paste0(format(round((sum(regular$g1surban ==3, na.rm =
TRUE )/nrow(regular) * 100), digits = 1), nsmall = 1), "\\%"),
paste0(format(round((sum(regularaide$g1surban ==3, na.rm = TRUE )/nrow(regularaide) *
100), digits = 1), nsmall = 1), "\\%"))
sch_urb <- c(paste0(format(round((sum(small$g1surban ==4, na.rm = TRUE )/nrow(small) *
100), digits = 1), nsmall = 1), "\\%"), paste0(format(round((sum(regular$g1surban ==4, na.rm =
TRUE )/nrow(regular) * 100), digits = 1), nsmall = 1), "\\%"),
paste0(format(round((sum(regularaide$g1surban ==4, na.rm = TRUE )/nrow(regularaide) *
100), digits = 1), nsmall = 1), "\\%"))

  #Teacher Characteristics
```

```
teach_f <- c(paste0(format(round((sum(small$g1tgen ==2, na.rm = TRUE )/nrow(small) * 100),
digits = 1), nsmall = 1), "\\%"), paste0(format(round((sum(regular$g1tgen ==2, na.rm = TRUE
)/nrow(regular) * 100), digits = 1), nsmall = 1), "\\%"),
paste0(format(round((sum(regularaide$g1tgen ==2, na.rm = TRUE )/nrow(regularaide) * 100),
digits = 1), nsmall = 1), "\\%"))
teach_white <- c(paste0(format(round((sum(small$g1trace ==1, na.rm = TRUE )/nrow(small) *
100), digits = 1), nsmall = 1), "\\%"), paste0(format(round((sum(regular$g1trace ==1, na.rm =
TRUE )/nrow(regular) * 100), digits = 1), nsmall = 1), "\\%"),
paste0(format(round((sum(regularaide$g1trace ==1, na.rm = TRUE )/nrow(regularaide) * 100),
digits = 1), nsmall = 1), "\\%"))
teach_yoe <- c(paste0(format(round(mean(small$g1tyears, na.rm=TRUE), digits = 1),nsmall = 1),
"±", format(round(sd(small$g1tyears, na.rm=TRUE), digits =1), nsmall = 1)),
paste0(format(round(mean(regular$g1tyears), digits = 1),nsmall = 1), "±",
format(round(sd(regular$g1tyears), digits =1), nsmall = 1)),
paste0(format(round(mean(regularaide$g1tyears), digits = 1),nsmall = 1), "±",
format(round(sd(regularaide$g1tyears), digits =1), nsmall = 1)))

  #Class and Student Characteristics
class_s<- c(paste0(format(round(mean(small$g1classsize), digits = 1),nsmall = 1), "±",
format(round(sd(small$g1classsize), digits =1), nsmall = 1)),
paste0(format(round(mean(regular$g1classsize), digits = 1),nsmall = 1), "±",
format(round(sd(regular$g1classsize), digits =1), nsmall = 1)),
paste0(format(round(mean(regularaide$g1classsize), digits = 1),nsmall = 1), "±",
format(round(sd(regularaide$g1classsize), digits =1), nsmall = 1)))
freelunch <- c(paste0(format(round((sum(small$g1freelunch ==1, na.rm = TRUE )/nrow(small) *
100), digits = 1), nsmall = 1), "\\%"), paste0(format(round((sum(regular$g1freelunch ==1, na.rm
= TRUE )/nrow(regular) * 100), digits = 1), nsmall = 1), "\\%"),
paste0(format(round((sum(regularaide$g1freelunch ==1, na.rm = TRUE )/nrow(regularaide) *
100), digits = 1), nsmall = 1), "\\%"))
speced <- c(paste0(format(round((sum(small$g1speced ==1, na.rm = TRUE )/nrow(small) *
100), digits = 1), nsmall = 1), "\\%"), paste0(format(round((sum(regular$g1speced ==1, na.rm =
TRUE )/nrow(regular) * 100), digits = 1), nsmall = 1), "\\%"),
paste0(format(round((sum(regularaide$g1speced ==1, na.rm = TRUE )/nrow(regularaide) *
100), digits = 1), nsmall = 1), "\\%"))
math <- c(paste0(format(round(mean(small$g1tmathss, na.rm=T), digits = 1),nsmall = 1), "±",
format(round(sd(small$g1tmathss, na.rm = T), digits =1), nsmall = 1)),
paste0(format(round(mean(regular$g1tmathss, na.rm = T), digits = 1),nsmall = 1), "±",
format(round(sd(regular$g1tmathss, na.rm = T), digits =1), nsmall = 1)),
paste0(format(round(mean(regularaide$g1tmathss, na.rm = T), digits = 1),nsmall = 1), "±",
format(round(sd(regularaide$g1tmathss, na.rm = T), digits =1), nsmall = 1)))
reading <- c(paste0(format(round(mean(small$g1treadss, na.rm = T), digits = 1),nsmall = 1), "±",
format(round(sd(small$g1treadss, na.rm = T), digits =1), nsmall = 1)),
paste0(format(round(mean(regular$g1treadss, na.rm = T), digits = 1),nsmall = 1), "±",
format(round(sd(regular$g1treadss, na.rm = T), digits =1), nsmall = 1)),
```

```
paste0(format(round(mean(regularaide$g1treadss, na.rm = T), digits = 1),nsmall = 1), "±",
format(round(sd(regularaide$g1treadss, na.rm = T), digits =1), nsmall = 1)))
listening <- c(paste0(format(round(mean(small$g1tlistss, na.rm = T), digits = 1),nsmall = 1), "±",
format(round(sd(small$g1tlistss, na.rm = T), digits =1), nsmall = 1)),
paste0(format(round(mean(regular$g1tlistss, na.rm = T), digits = 1),nsmall = 1), "±",
format(round(sd(regular$g1tlistss, na.rm = T), digits =1), nsmall = 1)),
paste0(format(round(mean(regularaide$g1tlistss, na.rm = T), digits = 1),nsmall = 1), "±",
format(round(sd(regularaide$g1tlistss, na.rm = T), digits =1), nsmall = 1)))
wordstudy <- c(paste0(format(round(mean(small$g1wordskillss, na.rm = T), digits = 1),nsmall =
1), "±", format(round(sd(small$g1wordskillss, na.rm = T), digits =1), nsmall = 1)),
paste0(format(round(mean(regular$g1wordskillss, na.rm = T), digits = 1),nsmall = 1), "±",
format(round(sd(regular$g1wordskillss, na.rm = T), digits =1), nsmall = 1)),
paste0(format(round(mean(regularaide$g1wordskillss, na.rm = T), digits = 1),nsmall = 1), "±",
format(round(sd(regularaide$g1wordskillss, na.rm = T), digits =1), nsmall = 1)))


#Table construction
table <- matrix(c(nyears, nyears_small, sch_inner, sch_suburban, sch_rural, sch_urb, teach_f,
teach_white, teach_yoe, class_s, freelunch, speced, math, reading, listening, wordstudy), ncol
=3, byrow = TRUE)
colnames(table) <- c("Small Class", "Regular Class", "Regular Class with Aide" )
rownames(table) <- c("Number of Years in STAR", "Number of Years in Small Classes", "Inner
City", "Suburban", "Rural", "Urban", "Female", "White", "Years of Experience", "Class Size",
"Recieves Free Lunch", "Special Education", "Math SAT Scaled Score", "Reading SAT Scaled
Score", "Listening SAT Scaled Score", "Word Study Skills SAT Scaled Score")

table <- kable(table, booktabs = T, "latex") %>%
  pack_rows("Expiriment Characteristics", 1, 2, latex_gap_space = "0.25em") %>%
  pack_rows("School Urbanicity", 3, 6, latex_gap_space = "0.25em") %>%
  pack_rows("Teacher Characteristics", 7, 10, latex_gap_space = "0.25em")

#Attrition Summary Table
star_students <- students %>%
  filter(flaggk == 1)

attrition_1st <- star_students %>%
  filter(flagg1 == 0, is.na(g1classtype) == TRUE)

# Grade K Summary Table
small<- attrition_1st %>%
  filter(gkclasstype == 1)

regular <- attrition_1st %>%
  filter(gkclasstype == 2)
```

```
regularaide <- attrition_1st %>%
  filter(gkclasstype == 3)


# Attrition Statistics
n.small <- nrow(small)
n.reg <- nrow(regular)
n.rega <- nrow(regularaide)

# Building Table
nyears <- c(paste0(format(round(mean(small$yearsstar), digits = 1),nsmall = 1), "±",
format(round(sd(small$yearsstar), digits =1), nsmall = 1)),
paste0(format(round(mean(regular$yearsstar), digits = 1),nsmall = 1), "±",
format(round(sd(regular$yearsstar), digits =1), nsmall = 1)),
paste0(format(round(mean(regularaide$yearsstar), digits = 1),nsmall = 1), "±",
format(round(sd(regularaide$yearsstar), digits =1), nsmall = 1)))
nyears_small <- c(paste0(format(round(mean(small$yearssmall), digits = 1),nsmall = 1), "±",
format(round(sd(small$yearssmall), digits =1), nsmall = 1)),
paste0(format(round(mean(regular$yearssmall), digits = 1),nsmall = 1), "±",
format(round(sd(regular$yearssmall), digits =1), nsmall = 1)),
paste0(format(round(mean(regularaide$yearssmall), digits = 1),nsmall = 1), "±",
format(round(sd(regularaide$yearssmall), digits =1), nsmall = 1)))

# School Urbanicity
sch_inner <- c(paste0(format(round((sum(small$gksurban ==1, na.rm = TRUE )/nrow(small) *
100), digits = 1), nsmall = 1), "\\%"), paste0(format(round((sum(regular$gksurban ==1, na.rm =
TRUE )/nrow(regular) * 100), digits = 1), nsmall = 1), "\\%"),
paste0(format(round((sum(regularaide$gksurban ==1, na.rm = TRUE )/nrow(regularaide) *
100), digits = 1), nsmall = 1), "\\%"))
sch_suburban <- c(paste0(format(round((sum(small$gksurban ==2, na.rm = TRUE )/nrow(small)
* 100), digits = 1), nsmall = 1), "\\%"), paste0(format(round((sum(regular$gksurban ==2, na.rm
= TRUE )/nrow(regular) * 100), digits = 1), nsmall = 1), "\\%"),
paste0(format(round((sum(regularaide$gksurban ==2, na.rm = TRUE )/nrow(regularaide) *
100), digits = 1), nsmall = 1), "\\%"))
sch_rural <- c(paste0(format(round((sum(small$gksurban ==3, na.rm = TRUE )/nrow(small) *
100), digits = 1), nsmall = 1), "\\%"), paste0(format(round((sum(regular$gksurban ==3, na.rm =
TRUE )/nrow(regular) * 100), digits = 1), nsmall = 1), "\\%"),
paste0(format(round((sum(regularaide$gksurban ==3, na.rm = TRUE )/nrow(regularaide) *
100), digits = 1), nsmall = 1), "\\%"))
sch_urb <- c(paste0(format(round((sum(small$gksurban ==4, na.rm = TRUE )/nrow(small) *
100), digits = 1), nsmall = 1), "\\%"), paste0(format(round((sum(regular$gksurban ==4, na.rm =
TRUE )/nrow(regular) * 100), digits = 1), nsmall = 1), "\\%"),
```

```
paste0(format(round((sum(regularaide$gksurban ==4, na.rm = TRUE )/nrow(regularaide) *
100), digits = 1), nsmall = 1), "\\%"))

#Teacher Characteristics
teach_f <- c(paste0(format(round((sum(small$gktgen ==2, na.rm = TRUE )/nrow(small) * 100),
digits = 1), nsmall = 1), "\\%"), paste0(format(round((sum(regular$gktgen ==2, na.rm = TRUE
)/nrow(regular) * 100), digits = 1), nsmall = 1), "\\%"),
paste0(format(round((sum(regularaide$gktgen ==2, na.rm = TRUE )/nrow(regularaide) * 100),
digits = 1), nsmall = 1), "\\%"))
teach_white <- c(paste0(format(round((sum(small$gktrace ==1, na.rm = TRUE )/nrow(small) *
100), digits = 1), nsmall = 1), "\\%"), paste0(format(round((sum(regular$gktrace ==1, na.rm =
TRUE )/nrow(regular) * 100), digits = 1), nsmall = 1), "\\%"),
paste0(format(round((sum(regularaide$gktrace ==1, na.rm = TRUE )/nrow(regularaide) * 100),
digits = 1), nsmall = 1), "\\%"))
teach_yoe <- c(paste0(format(round(mean(small$gktyears, na.rm=TRUE), digits = 1),nsmall = 1),
"±", format(round(sd(small$gktyears, na.rm=TRUE), digits =1), nsmall = 1)),
paste0(format(round(mean(regular$gktyears), digits = 1),nsmall = 1), "±",
format(round(sd(regular$gktyears), digits =1), nsmall = 1)),
paste0(format(round(mean(regularaide$gktyears, na.rm=T), digits = 1),nsmall = 1), "±",
format(round(sd(regularaide$gktyears, na.rm=T), digits =1), nsmall = 1)))

#Class and Student Characteristics
class_s<- c(paste0(format(round(mean(small$gkclasssize), digits = 1),nsmall = 1), "±",
format(round(sd(small$gkclasssize), digits =1), nsmall = 1)),
paste0(format(round(mean(regular$gkclasssize), digits = 1),nsmall = 1), "±",
format(round(sd(regular$gkclasssize), digits =1), nsmall = 1)),
paste0(format(round(mean(regularaide$gkclasssize), digits = 1),nsmall = 1), "±",
format(round(sd(regularaide$gkclasssize), digits =1), nsmall = 1)))
freelunch <- c(paste0(format(round((sum(small$gkfreelunch ==1, na.rm = TRUE )/nrow(small) *
100), digits = 1), nsmall = 1), "\\%"), paste0(format(round((sum(regular$gkfreelunch ==1, na.rm
= TRUE )/nrow(regular) * 100), digits = 1), nsmall = 1), "\\%"),
paste0(format(round((sum(regularaide$gkfreelunch ==1, na.rm = TRUE )/nrow(regularaide) *
100), digits = 1), nsmall = 1), "\\%"))
speced <- c(paste0(format(round((sum(small$gkspeced ==1, na.rm = TRUE )/nrow(small) * 100),
digits = 1), nsmall = 1), "\\%"), paste0(format(round((sum(regular$gkspeced ==1, na.rm = TRUE
)/nrow(regular) * 100), digits = 1), nsmall = 1), "\\%"),
paste0(format(round((sum(regularaide$gkspeced ==1, na.rm = TRUE )/nrow(regularaide) *
100), digits = 1), nsmall = 1), "\\%"))
math <- c(paste0(format(round(mean(small$gktmathss, na.rm=T), digits = 1),nsmall = 1), "±",
format(round(sd(small$gktmathss, na.rm = T), digits =1), nsmall = 1)),
paste0(format(round(mean(regular$gktmathss, na.rm = T), digits = 1),nsmall = 1), "±",
format(round(sd(regular$gktmathss, na.rm = T), digits =1), nsmall = 1)),
paste0(format(round(mean(regularaide$gktmathss, na.rm = T), digits = 1),nsmall = 1), "±",
format(round(sd(regularaide$gktmathss, na.rm = T), digits =1), nsmall = 1)))
```

```
reading <- c(paste0(format(round(mean(small$gktreadss, na.rm = T), digits = 1),nsmall = 1), "±",
format(round(sd(small$gktreadss, na.rm = T), digits =1), nsmall = 1)),
paste0(format(round(mean(regular$gktreadss, na.rm = T), digits = 1),nsmall = 1), "±",
format(round(sd(regular$gktreadss, na.rm = T), digits =1), nsmall = 1)),
paste0(format(round(mean(regularaide$gktreadss, na.rm = T), digits = 1),nsmall = 1), "±",
format(round(sd(regularaide$gktreadss, na.rm = T), digits =1), nsmall = 1)))
listening <- c(paste0(format(round(mean(small$gktlistss, na.rm = T), digits = 1),nsmall = 1), "±",
format(round(sd(small$gktlistss, na.rm = T), digits =1), nsmall = 1)),
paste0(format(round(mean(regular$gktlistss, na.rm = T), digits = 1),nsmall = 1), "±",
format(round(sd(regular$gktlistss, na.rm = T), digits =1), nsmall = 1)),
paste0(format(round(mean(regularaide$gktlistss, na.rm = T), digits = 1),nsmall = 1), "±",
format(round(sd(regularaide$gktlistss, na.rm = T), digits =1), nsmall = 1)))
wordstudy <- c(paste0(format(round(mean(small$gkwordskillss, na.rm = T), digits = 1),nsmall =
1), "±", format(round(sd(small$gkwordskillss, na.rm = T), digits =1), nsmall = 1)),
paste0(format(round(mean(regular$gkwordskillss, na.rm = T), digits = 1),nsmall = 1), "±",
format(round(sd(regular$gkwordskillss, na.rm = T), digits =1), nsmall = 1)),
paste0(format(round(mean(regularaide$gkwordskillss, na.rm = T), digits = 1),nsmall = 1), "±",
format(round(sd(regularaide$gkwordskillss, na.rm = T), digits =1), nsmall = 1)))


#Table construction
table_katrit <- matrix(c(nyears, nyears_small, sch_inner, sch_suburban, sch_rural, sch_urb,
teach_f, teach_white, teach_yoe, class_s, freelunch, speced, math, reading, listening,
wordstudy), ncol =3, byrow = TRUE)
colnames(table_katrit) <- c("Small Class (N=453)", "Regular Class (N=603)", "Regular Class with
Aide (N=580)" )
rownames(table_katrit) <- c("Number of Years in STAR", "Number of Years in Small Classes",
"Inner City", "Suburban", "Rural", "Urban", "Female", "White", "Years of Experience", "Class
Size", "Recieves Free Lunch", "Special Education", "Math SAT Scaled Score", "Reading SAT
Scaled Score", "Listening SAT Scaled Score", "Word Study Skills SAT Scaled Score")

table_katrit <- kable(table_katrit, booktabs = T, "latex") %>%
  pack_rows("Expiriment Characteristics", 1, 2, latex_gap_space = "0.25em") %>%
  pack_rows("Kindergarten School Urbanicity", 3, 6, latex_gap_space = "0.25em") %>%
  pack_rows("Kindergarten Teacher Characteristics", 7, 10, latex_gap_space = "0.25em")

#Table for students who stayed in k=1 and g=1

#Attrition Summary Table
stayed_1st <- students %>%
  filter(flaggk == 1, flagg1 == 1)

# Grade K Summary Table
small<- stayed_1st %>%
```

```
  filter(gkclasstype == 1)

regular <- stayed_1st %>%
  filter(gkclasstype == 2)

regularaide <- stayed_1st %>%
  filter(gkclasstype == 3)


# Attrition Statistics
n.small <- nrow(small)
n.reg <- nrow(regular)
n.rega <- nrow(regularaide)

# Building Table
nyears <- c(paste0(format(round(mean(small$yearsstar), digits = 1),nsmall = 1), "±",
format(round(sd(small$yearsstar), digits =1), nsmall = 1)),
paste0(format(round(mean(regular$yearsstar), digits = 1),nsmall = 1), "±",
format(round(sd(regular$yearsstar), digits =1), nsmall = 1)),
paste0(format(round(mean(regularaide$yearsstar), digits = 1),nsmall = 1), "±",
format(round(sd(regularaide$yearsstar), digits =1), nsmall = 1)))
nyears_small <- c(paste0(format(round(mean(small$yearssmall), digits = 1),nsmall = 1), "±",
format(round(sd(small$yearssmall), digits =1), nsmall = 1)),
paste0(format(round(mean(regular$yearssmall), digits = 1),nsmall = 1), "±",
format(round(sd(regular$yearssmall), digits =1), nsmall = 1)),
paste0(format(round(mean(regularaide$yearssmall), digits = 1),nsmall = 1), "±",
format(round(sd(regularaide$yearssmall), digits =1), nsmall = 1)))

# School Urbanicity
sch_inner <- c(paste0(format(round((sum(small$gksurban ==1, na.rm = TRUE )/nrow(small) *
100), digits = 1), nsmall = 1), "\\%"), paste0(format(round((sum(regular$gksurban ==1, na.rm =
TRUE )/nrow(regular) * 100), digits = 1), nsmall = 1), "\\%"),
paste0(format(round((sum(regularaide$gksurban ==1, na.rm = TRUE )/nrow(regularaide) *
100), digits = 1), nsmall = 1), "\\%"))
sch_suburban <- c(paste0(format(round((sum(small$gksurban ==2, na.rm = TRUE )/nrow(small)
* 100), digits = 1), nsmall = 1), "\\%"), paste0(format(round((sum(regular$gksurban ==2, na.rm
= TRUE )/nrow(regular) * 100), digits = 1), nsmall = 1), "\\%"),
paste0(format(round((sum(regularaide$gksurban ==2, na.rm = TRUE )/nrow(regularaide) *
100), digits = 1), nsmall = 1), "\\%"))
sch_rural <- c(paste0(format(round((sum(small$gksurban ==3, na.rm = TRUE )/nrow(small) *
100), digits = 1), nsmall = 1), "\\%"), paste0(format(round((sum(regular$gksurban ==3, na.rm =
TRUE )/nrow(regular) * 100), digits = 1), nsmall = 1), "\\%"),
paste0(format(round((sum(regularaide$gksurban ==3, na.rm = TRUE )/nrow(regularaide) *
100), digits = 1), nsmall = 1), "\\%"))
```

```
sch_urb <- c(paste0(format(round((sum(small$gksurban ==4, na.rm = TRUE )/nrow(small) *
100), digits = 1), nsmall = 1), "\\%"), paste0(format(round((sum(regular$gksurban ==4, na.rm =
TRUE )/nrow(regular) * 100), digits = 1), nsmall = 1), "\\%"),
paste0(format(round((sum(regularaide$gksurban ==4, na.rm = TRUE )/nrow(regularaide) *
100), digits = 1), nsmall = 1), "\\%"))

#Teacher Characteristics
teach_f <- c(paste0(format(round((sum(small$gktgen ==2, na.rm = TRUE )/nrow(small) * 100),
digits = 1), nsmall = 1), "\\%"), paste0(format(round((sum(regular$gktgen ==2, na.rm = TRUE
)/nrow(regular) * 100), digits = 1), nsmall = 1), "\\%"),
paste0(format(round((sum(regularaide$gktgen ==2, na.rm = TRUE )/nrow(regularaide) * 100),
digits = 1), nsmall = 1), "\\%"))
teach_white <- c(paste0(format(round((sum(small$gktrace ==1, na.rm = TRUE )/nrow(small) *
100), digits = 1), nsmall = 1), "\\%"), paste0(format(round((sum(regular$gktrace ==1, na.rm =
TRUE )/nrow(regular) * 100), digits = 1), nsmall = 1), "\\%"),
paste0(format(round((sum(regularaide$gktrace ==1, na.rm = TRUE )/nrow(regularaide) * 100),
digits = 1), nsmall = 1), "\\%"))
teach_yoe <- c(paste0(format(round(mean(small$gktyears, na.rm=TRUE), digits = 1),nsmall = 1),
"±", format(round(sd(small$gktyears, na.rm=TRUE), digits =1), nsmall = 1)),
paste0(format(round(mean(regular$gktyears), digits = 1),nsmall = 1), "±",
format(round(sd(regular$gktyears), digits =1), nsmall = 1)),
paste0(format(round(mean(regularaide$gktyears, na.rm=T), digits = 1),nsmall = 1), "±",
format(round(sd(regularaide$gktyears, na.rm=T), digits =1), nsmall = 1)))

#Class and Student Characteristics
class_s<- c(paste0(format(round(mean(small$gkclasssize), digits = 1),nsmall = 1), "±",
format(round(sd(small$gkclasssize), digits =1), nsmall = 1)),
paste0(format(round(mean(regular$gkclasssize), digits = 1),nsmall = 1), "±",
format(round(sd(regular$gkclasssize), digits =1), nsmall = 1)),
paste0(format(round(mean(regularaide$gkclasssize), digits = 1),nsmall = 1), "±",
format(round(sd(regularaide$gkclasssize), digits =1), nsmall = 1)))
freelunch <- c(paste0(format(round((sum(small$gkfreelunch ==1, na.rm = TRUE )/nrow(small) *
100), digits = 1), nsmall = 1), "\\%"), paste0(format(round((sum(regular$gkfreelunch ==1, na.rm
= TRUE )/nrow(regular) * 100), digits = 1), nsmall = 1), "\\%"),
paste0(format(round((sum(regularaide$gkfreelunch ==1, na.rm = TRUE )/nrow(regularaide) *
100), digits = 1), nsmall = 1), "\\%"))
speced <- c(paste0(format(round((sum(small$gkspeced ==1, na.rm = TRUE )/nrow(small) * 100),
digits = 1), nsmall = 1), "\\%"), paste0(format(round((sum(regular$gkspeced ==1, na.rm = TRUE
)/nrow(regular) * 100), digits = 1), nsmall = 1), "\\%"),
paste0(format(round((sum(regularaide$gkspeced ==1, na.rm = TRUE )/nrow(regularaide) *
100), digits = 1), nsmall = 1), "\\%"))
math <- c(paste0(format(round(mean(small$gktmathss, na.rm=T), digits = 1),nsmall = 1), "±",
format(round(sd(small$gktmathss, na.rm = T), digits =1), nsmall = 1)),
paste0(format(round(mean(regular$gktmathss, na.rm = T), digits = 1),nsmall = 1), "±",
```

```
format(round(sd(regular$gktmathss, na.rm = T), digits =1), nsmall = 1)),
paste0(format(round(mean(regularaide$gktmathss, na.rm = T), digits = 1),nsmall = 1), "±",
format(round(sd(regularaide$gktmathss, na.rm = T), digits =1), nsmall = 1)))
reading <- c(paste0(format(round(mean(small$gktreadss, na.rm = T), digits = 1),nsmall = 1), "±",
format(round(sd(small$gktreadss, na.rm = T), digits =1), nsmall = 1)),
paste0(format(round(mean(regular$gktreadss, na.rm = T), digits = 1),nsmall = 1), "±",
format(round(sd(regular$gktreadss, na.rm = T), digits =1), nsmall = 1)),
paste0(format(round(mean(regularaide$gktreadss, na.rm = T), digits = 1),nsmall = 1), "±",
format(round(sd(regularaide$gktreadss, na.rm = T), digits =1), nsmall = 1)))
listening <- c(paste0(format(round(mean(small$gktlistss, na.rm = T), digits = 1),nsmall = 1), "±",
format(round(sd(small$gktlistss, na.rm = T), digits =1), nsmall = 1)),
paste0(format(round(mean(regular$gktlistss, na.rm = T), digits = 1),nsmall = 1), "±",
format(round(sd(regular$gktlistss, na.rm = T), digits =1), nsmall = 1)),
paste0(format(round(mean(regularaide$gktlistss, na.rm = T), digits = 1),nsmall = 1), "±",
format(round(sd(regularaide$gktlistss, na.rm = T), digits =1), nsmall = 1)))
wordstudy <- c(paste0(format(round(mean(small$gkwordskillss, na.rm = T), digits = 1),nsmall =
1), "±", format(round(sd(small$gkwordskillss, na.rm = T), digits =1), nsmall = 1)),
paste0(format(round(mean(regular$gkwordskillss, na.rm = T), digits = 1),nsmall = 1), "±",
format(round(sd(regular$gkwordskillss, na.rm = T), digits =1), nsmall = 1)),
paste0(format(round(mean(regularaide$gkwordskillss, na.rm = T), digits = 1),nsmall = 1), "±",
format(round(sd(regularaide$gkwordskillss, na.rm = T), digits =1), nsmall = 1)))


#Table construction
table_stayed1 <- matrix(c(nyears, nyears_small, sch_inner, sch_suburban, sch_rural, sch_urb,
teach_f, teach_white, teach_yoe, class_s, freelunch, speced, math, reading, listening,
wordstudy), ncol =3, byrow = TRUE)
colnames(table_stayed1) <- c("Small Class (N=1303)", "Regular Class (N=1425)", "Regular Class
with Aide (N=1490)")
rownames(table_stayed1) <- c("Number of Years in STAR", "Number of Years in Small Classes",
"Inner City", "Suburban", "Rural", "Urban", "Female", "White", "Years of Experience", "Class
Size", "Recieves Free Lunch", "Special Education", "Math SAT Scaled Score", "Reading SAT
Scaled Score", "Listening SAT Scaled Score", "Word Study Skills SAT Scaled Score")

table_stayed1 <- kable(table_stayed1, booktabs = T, "latex") %>%
  pack_rows("Expiriment Characteristics", 1, 2, latex_gap_space = "0.25em") %>%
  pack_rows("Kindergarten School Urbanicity", 3, 6, latex_gap_space = "0.25em") %>%
  pack_rows("Kindergarten Teacher Characteristics", 7, 10, latex_gap_space = "0.25em")

# Pie Chart
small<- star_students %>%
  filter(gkclasstype ==1)
nrow(small)
small_3rd <- small %>%
```

```
  filter(g3classtype ==1)
nrow(small_3rd)
```