PROJECT 2

MATH 5900 – ADVANCED DATA ANALYSIS

March 10, 2024

TAIWO JEGEDE

E00755963

PART 1

1

For longitudinal data, we can use the mean, standard deviation and the correllation coefficient to explore the data.

1. Mean: Lets us know the average value of a variable over time, showing trends. It helps in understanding the general level of the variable and how it changes longitudinally
2. Standard deviation: How spread out the data points are at each time point. A higher standard deviation suggests greater variability in the data, while a lower standard deviation indicates more consistency over time.
3. Correlation coefficient: Relationship between two variables at different times. When we understand how variables are related longitudinally, we can get insights into dependencies and interactions within the data set.

These statistics help us understand the data's center, variation, and connections over time.

2

Random effect is important in conditional longitudinal modelling because it accounts for the correlated nature of repeated measurements from the same individual overtime. Lets say for instance, we wanted to get the model trends of weight overtime while controlled by diets. And there is a column for chickID. In R, we would indicate the random factor by (1|ChickID) when fitting our Random Intercept model. So in essence the random effect represents the individual-level variability in the initial level (random intercept) and/or rate of change (random slope) across time.

3

Population-based effects	Subject-specific effects
Used in marginal models, like GEE	Used in conditional models like mixed-effects regression models
Describe the average response over the entire population, treating the correlation within subjects as a nuisance.	Explicitly model the correlation within subjects by including random effects that capture each individual's deviations from the overall population trend.
Population-averaged models provide estimates of the average evolution for the population.	Subject-specific models allow for individual-level trajectories by partitioning the variability into between-subject and within-subject components.

4

Time-independent covariates	Time-dependent covariates
These covariates have values that do not change over time for a given subject.	These covariates can take on different values at each time point for a given subject.
Examples include baseline characteristics like gender, race, treatment group assignment.	Examples include time-varying predictors like medication dose, symptom severity, biomarkers.

PART II

PURPOSE OF THE ANALYSIS

The primary aim of this analysis is to investigate how factors such as the year, PhD-granting status, and institutional type (public or private) influence student enrollment in Nigerian universities over time. Specifically, the analysis seeks to answer the following questions:

- How has student enrollment in Nigerian universities changed over the years, and what trends can be observed?
- Do universities with PhD-granting status tend to have higher or lower student enrollment compared to those without this status?
- Are there significant differences in student enrollment between public and private universities in Nigeria?

By addressing these questions, the analysis will provide insights into the potential impact of these factors on student enrollment patterns in Nigerian universities. This information can be valuable for educational policymakers, university administrators, and stakeholders involved in higher education planning and resource allocation.

DESCRIPTION OF THE DATA

The analysis will utilize a dataset containing information on student enrollment, PhD-granting status, institutional type (public or private), and the corresponding years for selected Nigerian universities. The dataset appears to have been collected from various sources, potentially including university records and educational authorities.

The key variables to be used in the analysis are:

- students5_estimated (Response Variable): This variable represents the estimated student enrollment numbers for each university and year. It is a numerical variable.
- year (Predictor Variable): This variable indicates the specific year for which the enrollment data is provided. It is a numerical variable.
- phd_granting (Predictor Variable): This binary variable (0 or 1) indicates whether a university has PhD-granting status or not.
- private01 (Predictor Variable): This binary variable (0 or 1) represents the institutional type, where 0 indicates a public university and 1 indicates a private university.
- iau_id: These variables provide unique identifiers for the universities, respectively. They will be used for identification and labeling purposes.

The dataset appears to be structured in a long format, with each row representing a specific university, year, and the corresponding enrollment and institutional characteristics. This structure allows for analyzing enrollment trends over time and across different universities.

PROPOSED ANALYSES

To address the research interest of investigating how factors such as year, PhD-granting status, and institutional type (public or private) influence student enrollment in Nigerian universities over time. Here, i will use Graphical representations (line plots, bar charts) to visualize the trends in student enrollment over time for different universities, categorized by their PhD-granting status and institutional type (public or private).

Also, conditional modeling will be employed to account for the hierarchical structure of the data, where student enrollment observations are nested within universities. Specifically, linear mixed-effects models will be fitted using the lmer function from the lme4 package in R.

Random Intercept Model:

The random intercept model will be used to assess the overall effect of the predictor variables (year, PhD-granting status, and institutional type) on student enrollment, while allowing for varying baseline enrollment levels across universities.

The model can be represented as:

$$Wit = \beta0+\beta1Yearij+\beta2PhDij+\beta3Privateij+ui+\epsilon ij$$

Random Time Slope Model:

The random time slope model will be used to investigate if the effect of year on student enrollment varies across universities, in addition to allowing for varying baseline enrollment levels. The model can be represented as:

$$Wit = \beta0+\beta1Yearij+\beta2PhDij+\beta3Privateij+u0i+u1iYearij+\epsilon ij$$

ANALYSIS RESULT

The line plot depicting student enrollment over time for each university revealed an overall upward trend, with enrollment numbers showing a noticeable increase or spike across the years. This pattern suggests that Nigerian universities have experienced a steady growth in their student populations over the observed time period.

Impact of PhD-Granting Status

The bar chart comparing average student enrollment by PhD-granting status indicated that universities with PhD-granting status tend to have higher enrollment numbers compared to those without this status. This observation highlights the potential influence of advanced degree offerings on attracting a larger student body.

Institutional Type: Public vs. Private

The analysis of average student enrollment by institutional type (public or private) revealed that public universities in Nigeria generally have higher enrollment figures than their private counterparts. This finding suggests that public institutions may be more accessible or preferred by a larger segment of the student population.

After fitting in the model for Random Intercept and the Random (Time) Slope, the RTModel (random time slope model) had a lower AIC value (13381.22) compared to the RIModel (13389.18), suggesting that the RTModel is a better fit for the data. Also, the likelihood ratio test was also used t compare the fit of the two models. The output shows that the RTModel has a significantly better fit than the RIModel (Chi-square = 12.049, df = 2, p-value = 0.002419). The small p-value indicates that the random time slope model (RTModel) provides a significantly better fit to the data compared to the random intercept model (RIModel).

Based on the output for the RTModel (random time slope model), here is the interpretation of the fixed effects and how they address the research question:

Fitted model:

$$y_{ij} = -826372.3 + 417.3\text{Year}_{ij} + 6242.6\text{PhD}_{ij} - 11287.8\text{Private}_{ij} + u_{0i} + u_{1i}\text{Year}_{ij} + \epsilon\text{ij}$$

Interpretation of Fixed Effects:

(Intercept): -826372.3 represents the expected student enrollment when all predictor variables are zero. However, since the year variable is centered, this intercept may not have a meaningful interpretation in the context of the data.

year: 417.3 indicates that, on average, student enrollment increases by approximately 417 students per year across universities, holding other variables constant.

This positive coefficient suggests an overall upward trend in student enrollment over time in Nigerian universities.

phd_granting: 6242.6 shows that universities with PhD-granting status tend to have, on average, 6242 more students enrolled compared to universities without PhD-granting status, controlling for other factors.

This positive coefficient highlights the potential impact of offering advanced degree programs on attracting a larger student body.

private01: -11287.8, the negative coefficient for the private01 variable indicates that, on average, private universities have 11287 fewer students enrolled compared to public universities, holding other variables constant. This finding suggests that public universities in Nigeria generally have higher student enrollment than private institutions.

Addressing the Research Question:

The research question I wanted to answer was that, "How do factors such as year, PhD-granting status, and institutional type (public or private) influence student enrollment in Nigerian universities over time?"

Based on the fixed effects interpretations, I can say that student enrollment in Nigerian universities has shown an overall increasing trend over time, with an average increase of approximately 417 students per year across universities. Universities with PhD-granting status tend to have significantly higher student enrollment compared to those without this status,

suggesting that offering advanced degree programs may attract a larger student population. Public universities in Nigeria generally have higher student enrollment compared to private institutions, indicating that public institutions may be more accessible or preferred by a larger segment of the student population.

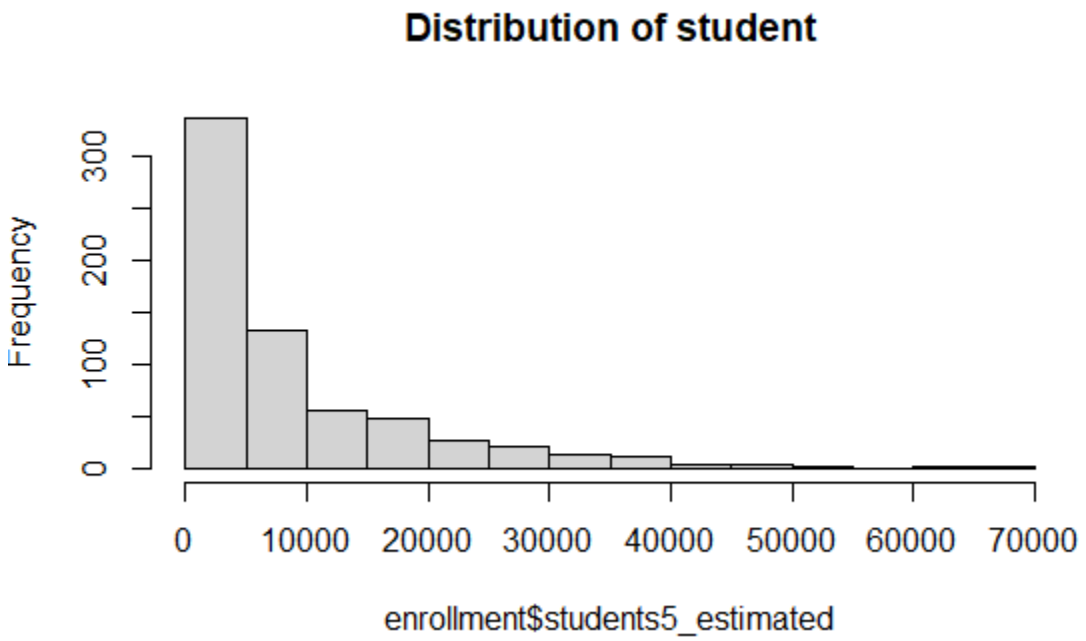
CONCLUSION

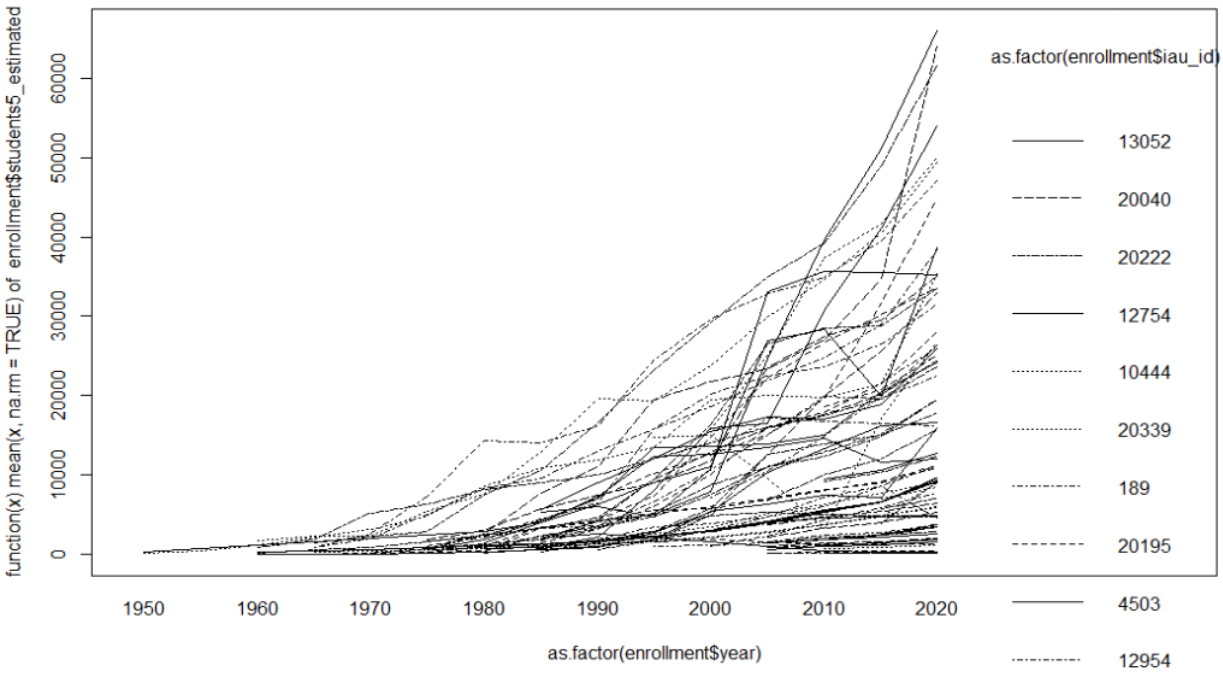
In summary, the analysis revealed that factors such as the passage of time, PhD-granting status, and institutional type (public or private) play a significant role in shaping student enrollment patterns in Nigerian universities. These findings can inform educational policymakers, university administrators, and stakeholders in making informed decisions regarding resource allocation, program offerings, and strategies to meet the growing demand for higher education in Nigeria.

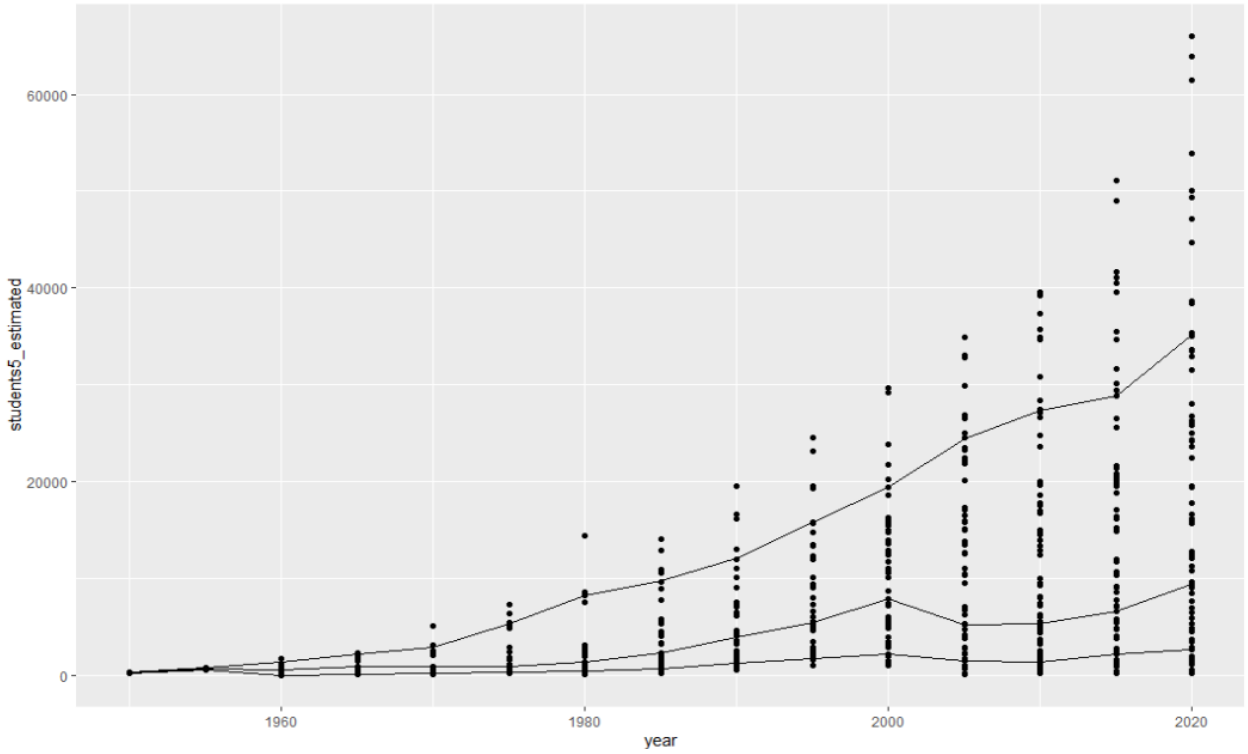
REFERENCES

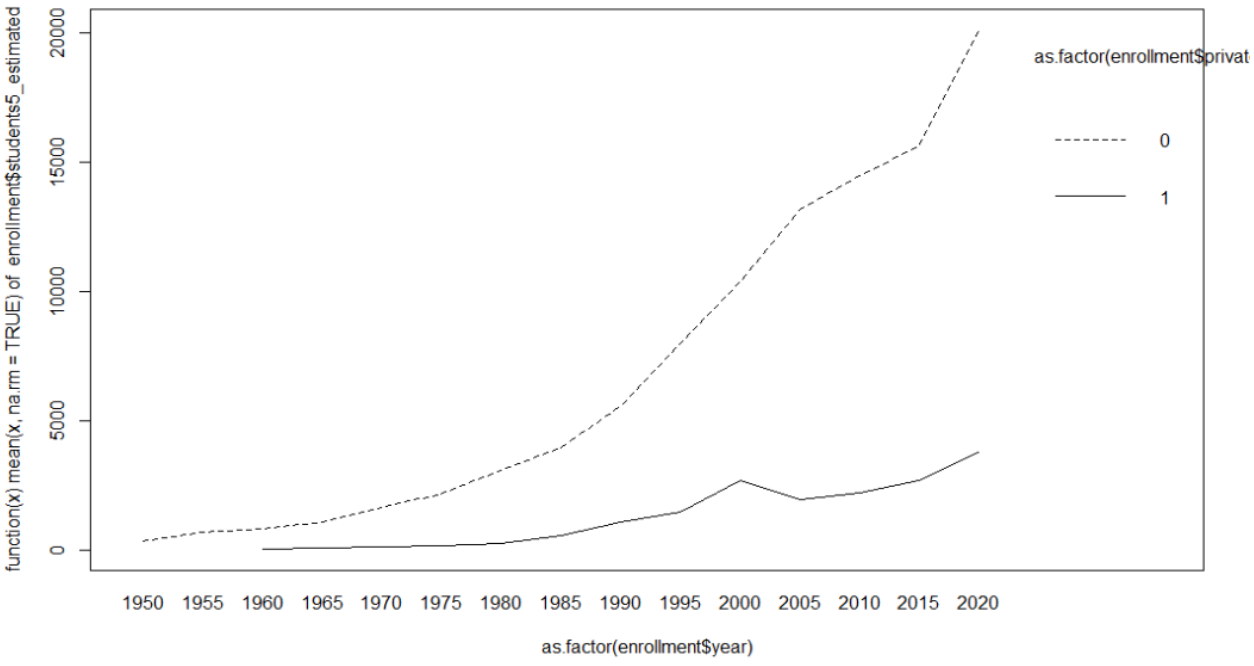
Hedeker, D. (2012). *Longitudinal Data Analysis.* John Wiley And Sons.

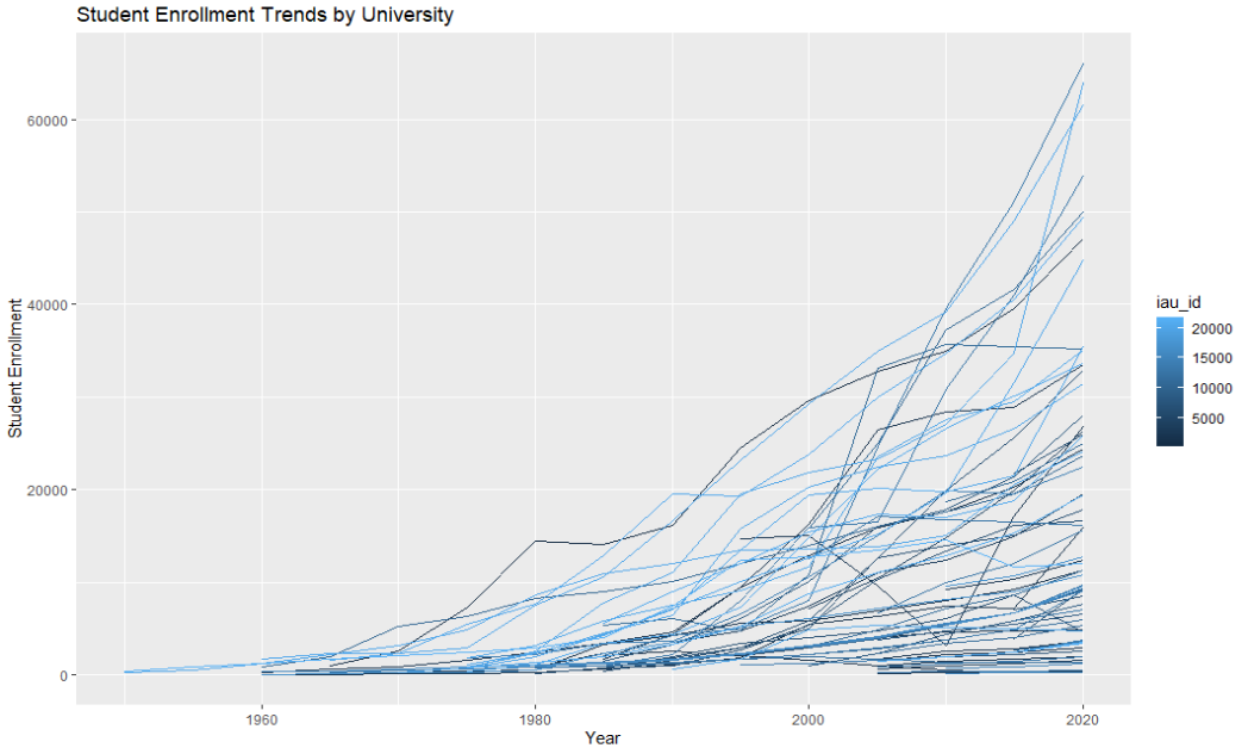
APPENDIX

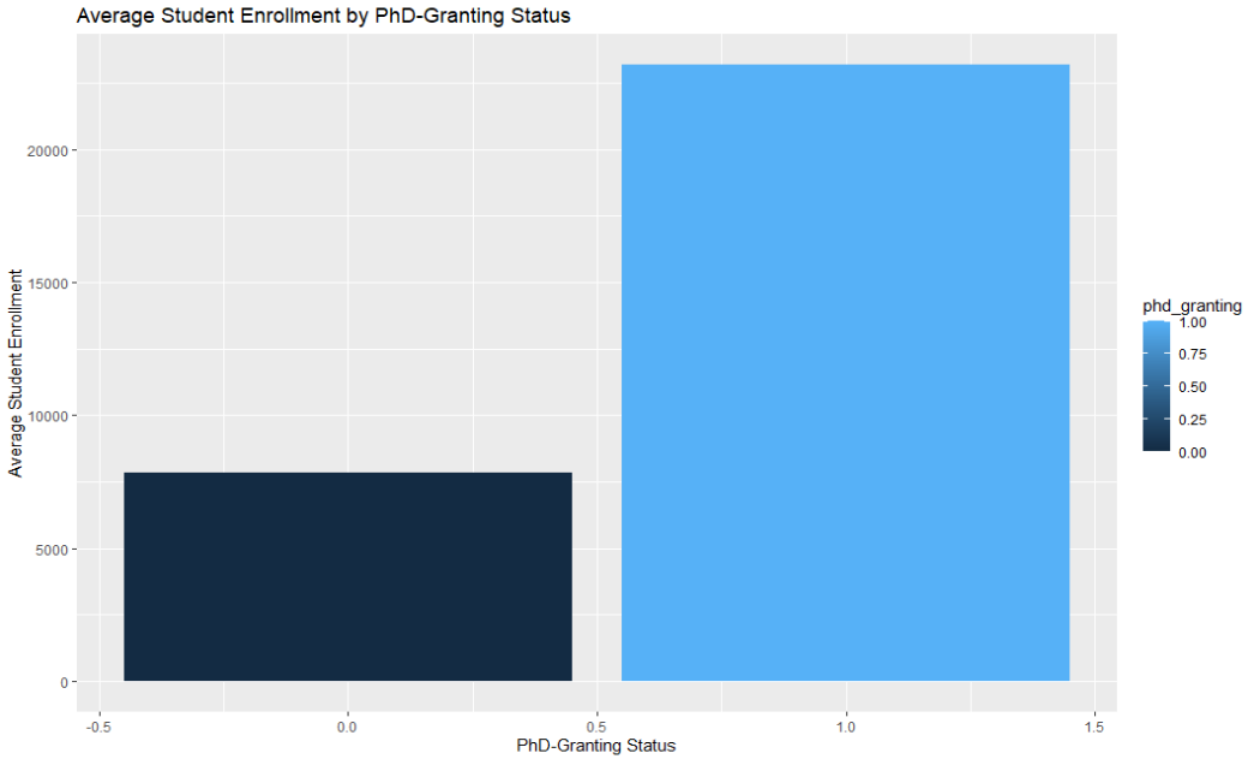

Distribution of student
Frequency
0
100
200
300
0
10000
20000
30000
40000
50000
60000
70000
enrollment$students5_estimated

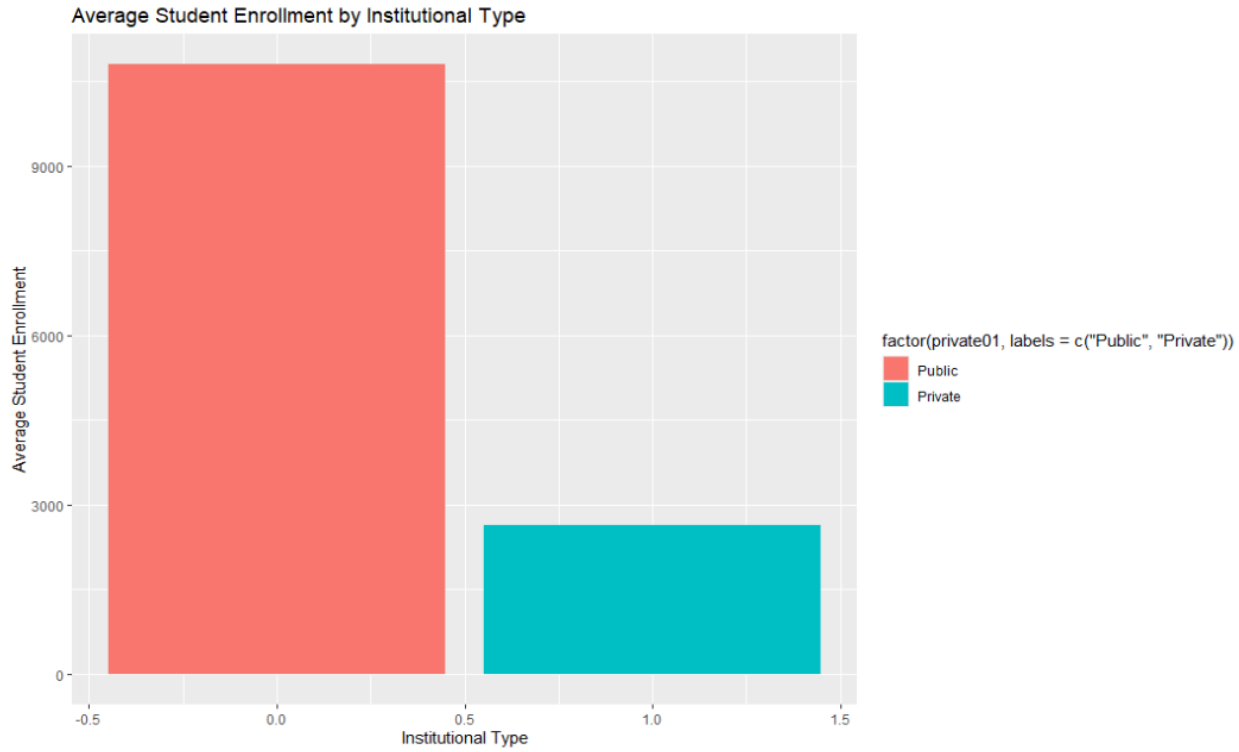

function(x) mean(x, na.rm = TRUE) of enrollment$students5_estimated
0
10000
20000
30000
40000
50000
60000
1950
1960
1970
1980
1990
2000
2010
2020
as.factor(enrollment$year)
as.factor(enrollment$iau_id)
13052
20040
20222
12754
10444
20339
189
20195
4503
12954

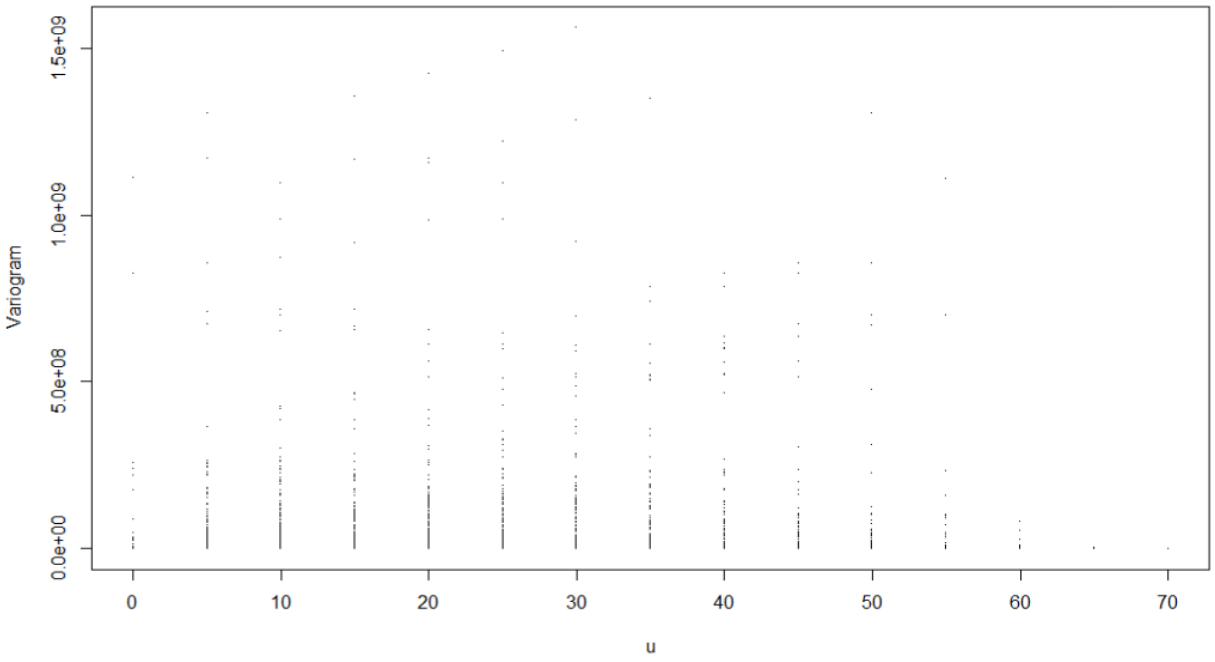

60000
40000
20000
0
students5_estimated
1960
1980
2000
2020
year


function(x) mean(x, na.rm = TRUE) of enrollment$students5_estimated
0
5000
10000
15000
20000
1950 1955 1960 1965 1970 1975 1980 1985 1990 1995 2000 2005 2010 2015 2020
as.factor(enrollment$year)
as.factor(enrollment$privat
0
1


Student Enrollment Trends by University
Student Enrollment
60000
40000
20000
0
1960
1980
2000
2020
Year
iau_id
20000
15000
10000
5000


Average Student Enrollment by PhD-Granting Status
Average Student Enrollment
20000
15000
10000
5000
0
-0.5
0.0
0.5
1.0
1.5
PhD-Granting Status
phd_granting
1.00
0.75
0.50
0.25
0.00


Average Student Enrollment by Institutional Type
Average Student Enrollment
9000
6000
3000
0
-0.5
0.0
0.5
1.0
1.5
Institutional Type
factor(private01, labels = c("Public", "Private"))
Public
Private


Variogram
0.0e+00
5.0e+08
1.0e+09
1.5e+09
0
10
20
30
40
50
60
70
u

```
> AIC(RIModel,RTModel)
        df      AIC
RIModel  6 13320.85
RTModel  8 13314.16
> BIC(RIModel,RTModel)
        df      BIC
RIModel  6 13347.76
RTModel  8 13350.05
> anova(RIModel,RTModel)
refitting model(s) with ML (instead of REML)
Data: enrollment
Models:
RIModel: students5_estimated ~ year + phd_granting + private01 + (1 | iau_id)
RTModel: students5_estimated ~ year + phd_granting + private01 + (1 + year | iau_id)
        npar   AIC   BIC  logLik deviance  Chisq Df Pr(>Chisq)
RIModel    6 13375 13402 -6681.4    13363
RTModel    8 13368 13404 -6676.0    13352 10.805  2   0.004505 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

R-Code

```
library(tidyverse)
library(lme4)
library(lmerTest)

# Load the dataset
setwd('C:/Users/student75/Desktop/Assignements and Project/MATH5900/Project 2')
file_path <- 'enrollments1.csv'
enrollment <- read.csv(file_path)

# Load required packages
library(ggplot2)
library(dplyr)
library(plotly)
library(tidyr)
# Summary statistics for numerical variables
summary(enrollment[, c("private01", "phd_granting", "students5_estimated")])
```

```
# Histograms for selected variables
hist(enrollment$students5_estimated, main = "Distribution of student")
interaction.plot(x.factor=as.factor(enrollment$year),
                 trace.factor=as.factor(enrollment$iau_id),response=enrollment$students5_estimated,
                 fun=function(x)mean(x,na.rm=TRUE))
## SPECIFIC PERCENTILES ##
ggplot(data=enrollment,aes(x=year,y=students5_estimated,group=iau_id)) +
  geom_point() +
  stat_summary(aes(group=1),geom="line",
fun.y=function(x){quantile(x,probs=c(0.10),na.rm=TRUE)}) +
  stat_summary(aes(group=1),geom="line",
fun.y=function(x){quantile(x,probs=c(0.50),na.rm=TRUE)}) +
  stat_summary(aes(group=1),geom="line",
fun.y=function(x){quantile(x,probs=c(0.90),na.rm=TRUE)})

## INTERACTION PLOT ##
interaction.plot(x.factor=as.factor(enrollment$year),trace.factor=as.factor(enrollment$private01),
                 response=enrollment$students5_estimated,fun=function(x)mean(x,na.rm=TRUE))

library(joineR)
# Assuming 'enrollment' is your dataframe
enrollment$iau_id <- as.numeric(enrollment$iau_id)
# Assuming 'enrollment' is your dataframe
enrollment$iau_id <- as.numeric(enrollment$iau_id, na.rm = TRUE)
# Line plot of student enrollment over time by university
ggplot(enrollment, aes(x = year, y = students5_estimated, color = iau_id, group = iau_id)) +
  geom_line() +
  labs(title = "Student Enrollment Trends by University",
       x = "Year",
       y = "Student Enrollment")
```

```
# Bar chart of student enrollment by PhD-granting status
ggplot(enrollment, aes(x = phd_granting, y = students5_estimated, fill = phd_granting)) +
  geom_bar(stat = "summary", fun = "mean", position = "dodge") +
  labs(title = "Average Student Enrollment by PhD-Granting Status",
       x = "PhD-Granting Status",
       y = "Average Student Enrollment")

# Bar chart of student enrollment by institutional type
ggplot(enrollment, aes(x = private01, y = students5_estimated, fill = factor(private01, labels =
c("Public", "Private")))) +
  geom_bar(stat = "summary", fun = "mean", position = "dodge") +
  labs(title = "Average Student Enrollment by Institutional Type",
       x = "Institutional Type",
       y = "Average Student Enrollment")

# Overall summary statistics
summary(enrollment$students5_estimated)

# Summary statistics by year
enrollment %>%
  group_by(year) %>%
  summarize(mean_enrollment = mean(students5_estimated),
            median_enrollment = median(students5_estimated),
            sd_enrollment = sd(students5_estimated))

## VARIOGRAM ##
data = lm(students5_estimated~year+phd_granting+private01 ,data=enrollment)
resids = residuals(data)
chickVG = variogram(enrollment$iau_id,enrollment$year,resids)
plot(chickVG)
```

```
library(lme4)
# RANDOM INTERCEPT MODEL #
RIModel = lmer(students5_estimated~year+phd_granting+private01+(1|iau_id),
        data=enrollment,REML=TRUE)
RIModel
summary(RIModel)

# RANDOM TIME (SLOPE) MODEL #
RTModel = lmer(students5_estimated~year+phd_granting+private01+(1+year|iau_id),
        data=enrollment,REML=TRUE)
summary(RTModel)

AIC(RIModel,RTModel)
BIC(RIModel,RTModel)

anova(RIModel,RTModel)
RTModel
RIResid = residuals(RIModel)
RIPred = predict(RIModel)

## RESIDUALS VERSUS PREDICTED VALUES ##
qplot(RIPred,RIResid)
## RESIDUALS VERSUS PREDICTED TIME ##
qplot(enrollment$year,RIResid)
```