Project Report - Analyzing Educational Data

Daniel Posmik, Aristofanis Rontogiannis, Jizhou Tian

2024-10-24

## Table of contents

Introduction (Daniel)

For this project, we will be analyzing educational data. We are interested in understanding the predictors of student performance as measured by exam scores. We will be using a publicly available dataset from Kaggle that contains information about students and their exam scores.

Table 1: Variable Summary for the Educational Data

Variable Name	Variable Type	Description
Hours_Studied	numeric	Hours Studied
Attendance	numeric	Attendance
Parental_Involvement	factor	Parental Involvement
Access_to_Resources	factor	Access to Resources
Extracurricular_Activities	factor	Extracurricular Activities
Sleep_Hours	numeric	Sleep Hours
Previous_Scores	numeric	Previous Scores
Motivation_Level	factor	Motivation Level
Internet_Access	factor	Internet Access
Tutoring_Sessions	numeric	Tutoring Sessions

Variable Name	Variable Type	Description
Family_Income	factor	Family Income
Teacher_Quality	factor	Teacher Quality
School_Type	factor	School Type
Peer_Influence	factor	Peer Influence
Physical_Activity	numeric	Physical Activity
Learning_Disabilities	factor	Learning Disability
Parental_Education_Level	factor	Parental Education Level
Distance_from_Home	factor	Distance from Home
Gender	factor	Gender
Exam_Score	numeric	Exam Score

Now, we can also take a look at how the explanatory variables relate to each other.
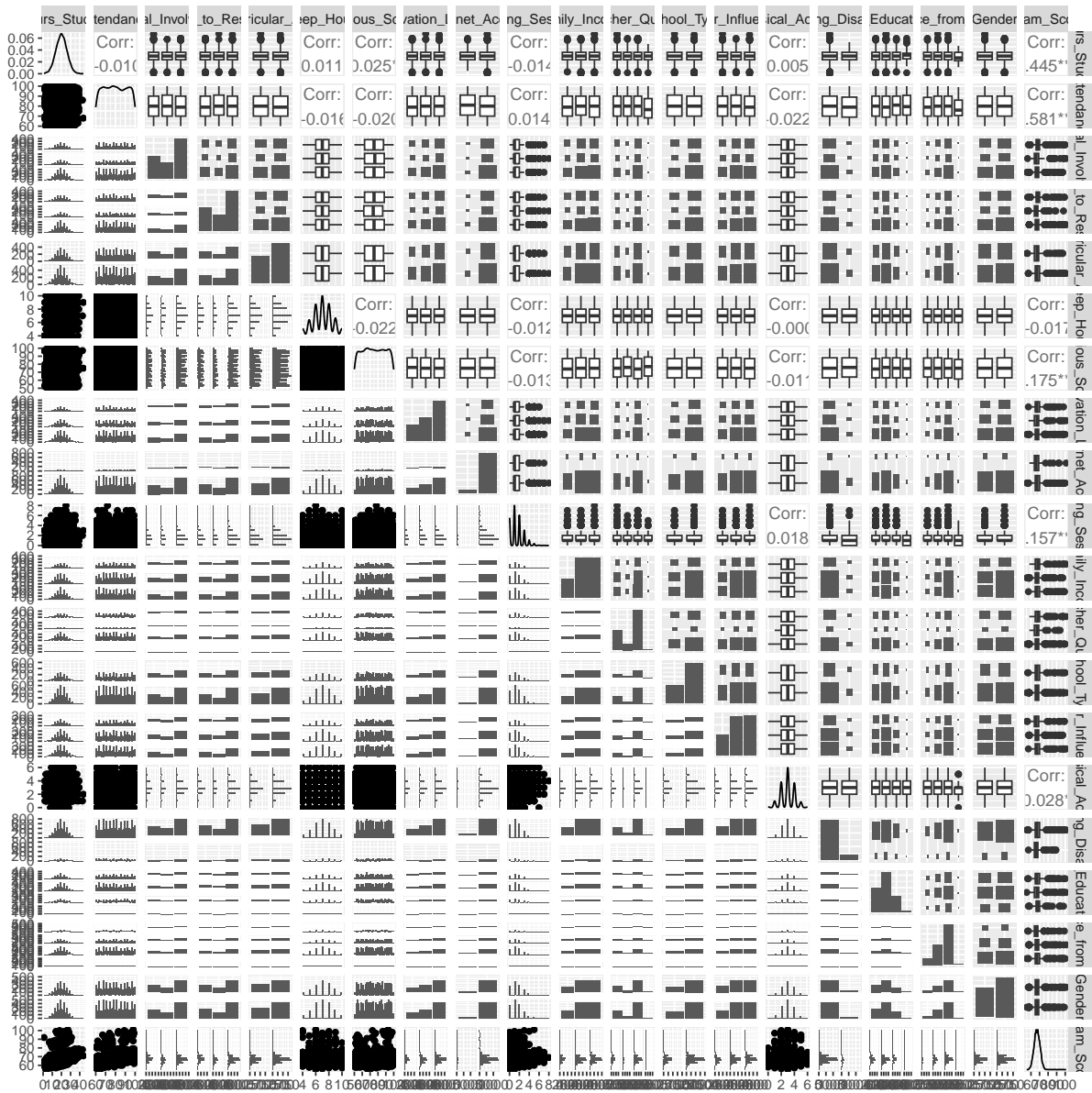
Part 1: Linear Regression Analysis (Daniel)

Part 2: Principal Component Analysis (Aristofanis)

Part 3: Non-linear Regression Analysis (Jizhou)

Conclusion

Code Appendix

```r
# Set up knit environment
knitr::opts_chunk$set(echo = F)
knitr::opts_chunk$set(error = F)
knitr::opts_chunk$set(warning = F)
knitr::opts_chunk$set(message = F)

# Load necessary packages
library(tidyverse)
library(kableExtra)
library(knitr)
library(ggplot2)
library(naniar)
library(gtsummary)
library(GGally)

# Load the data
educ_dta <- read_csv("student_performance.csv") %>%
  mutate(
    Parental_Involvement = as.factor(Parental_Involvement),
    Access_to_Resources = as.factor(Access_to_Resources),
    Extracurricular_Activities = as.factor(Extracurricular_Activities),
    Motivation_Level = as.factor(Motivation_Level),
    Internet_Access = as.factor(Internet_Access),
    Family_Income = as.factor(Family_Income),
    Teacher_Quality = as.factor(Teacher_Quality),
    School_Type = as.factor(School_Type),
    Peer_Influence = as.factor(Peer_Influence),
    Learning_Disabilities = as.factor(Learning_Disabilities),
    Parental_Education_Level = as.factor(Parental_Education_Level),
    Distance_from_Home = as.factor(Distance_from_Home),
    Gender = as.factor(Gender)
  )

# Summary table
table_summary <- tibble(
  "Variable Name" = colnames(educ_dta),
  "Variable Type" = sapply(educ_dta, class),
  "Description" = c(
    "Hours Studied", #1
    "Attendance", #2
```

```
    "Parental Involvement", #3
    "Access to Resources", #4
    "Extracurricular Activities", #5
    "Sleep Hours", #6
    "Previous Scores", #7
    "Motivation Level", #8
    "Internet Access", #9
    "Tutoring Sessions", #10
    "Family Income", #11
    "Teacher Quality", #12
    "School Type", #13
    "Peer Influence", #14
    "Physical Activity", #15
    "Learning Disability", #16
    "Parental Education Level", #17
    "Distance from Home", #18
    "Gender", #19
    "Exam Score" #20
  )
)

# Display the table
knitr::kable(table_summary, caption = "Variable Summary for the Educational Data")

GGally::ggpairs(educ_dta)
```