R Term Project

Mutasim Mahmud, 0813161

2023-12-07

Introduction

Students are the breathing future of a prospering country. So, it is up-most important to understand and make sure that our generation of students can get the best out of their education. So, we need to make sure they get the best environment from their institution. But is that the only case? What about the internal aspects of the students themselves? Does that also affect their studies? Could there a way we can statistically measure what aspects of their life impacts their journey other than the environment of the institution they study in? Here at VanderStel (2014) , they explained how the demographic of a student can affect their study. They focused primarily on Chugiak High school and researched on the students of the institute. They explained in detail how the students demographic characteristics such as Family, socioeconomic status, culture, community, ethnicity can impact the their education. At Rubright, Jodoin, & Barone (2019) , the researchers conducted study on the student of United States Medical Licensing Examination. They analyzed the previously studeid modern USMLE step format.They attemped to use hierarchical linear modeling on the US and Canadian medical graduates. Their result showed significant differences by gender, race, citizenship, age. But it should also be noted that the demographic differences were tempered by previous exam performance and undergrad performance. On Bilal, Omar, Anwar, Bokhari, & Choi (2022) , The authors described The role of demographic and academic features in a student performance prediction. The study used Python's SciKit learn and Pandas libraries to analyze the dataset. The data contained demographic features,High School Certificate marks, Higher Secondary School Certificate (HSSC) subjects marks, and first semester SGPA of DVM program. The partitioned the data into 15-crossfold validation 85% training and 15% testing datasets. They used Five surpervised classification algorithms and evaluated the performance using Precision, Recall, and Accuracy as metrics. At the end they compared which one had the best performance. According to them,support vector machine machine had the best performance with an accuracy of 92%. on Johnson (2015) , the author conducted research on on-campus and full-online students, while Comparing Demographics, Digital Technology Use and Learning Characteristics. the data contained a sample of 185 students. they used various quires on the two specific student groups and campared the data by applying independent sample t-test. they didnt find significance diffirance in gender, program, employment and economic needs but found that online students were older and are mostly native speakers. No difference in reading stetigies, learing belief but on campus students higheg level of motivation than other group. here at Coldwell, Craig, Paterson, & Mustard (2008) , they conducted study on the relationship between Participation, Demographics and Academic Performance on online students. The dataset had a sample of 500 online students and analyzed using descriptc and inferential statistics.They studied how participation in course activities, demographic characteristics, and academic performance are related. They found out that age, gender, ethnicity had no significance quizes, assignment and discussion forums were correlated with final grade. on Fitchett & Heafner (2017) , the authors used student demographics and teacher characteristcs to predict history knowledge of

elementary-age students. they used the dataset from 2010 National Assessment for Education Progress (NAEP) U.S. History test. they use mutilevel modeling to analyze the dataset.Results suggest that the teachers' subject matter background coupled with instructional decision-making was associated with increased academic outcomes. at Lord, Layton, Ohland, Brawner, & Long (2014) , the authors studied the demographics and outcome of chemical engineering students.They sampled a dataset with 137649 FTIC students in engineering and 39354 transfer students in engineering. later on they modified the data and focused on 11899 FTIC students and 2370 transfer students who had ethnecity of Asian, black, Hispanic, White, status of CE major and six year graduation rates from 1987-2010. They found out that the trajectories of the students differ by ethnicity. Women graduates had higher rate than men and same ethnicity. here Pilotte, Ohland, Lord, Layton, & Orr (2017) , they conducted study on Student Demographics, Pathways, and Outcomes in Industrial Engineering. This study focuses on the 9278 students(FTIC) and 1716 transfer students.They compared the rates of choosing,graduating,persisting students by ethnecity and stickiness the students of industrial engineering. They ethnicity out that women choose Industrial Engineering more than men and out all the ethnicity, hispanic students had the highest graduation rate. on Colorado & Eberle (2012) , The authors studied on both students demographic and success in online environment. Their sample contained 170 graduate students enrolled in online classes during the time period of 2005 spring and summer semesters. the dataset had demographic data of the students age, enrollment status, work status, GPA, number of degrees and last enrollment status. They applied one-way analyses of variance to compare the variable with academic performance. They also used Independent sample t-tests for variables using two categories. They found strong relationship between age and elobaration, age and critical thinking, age and cognitive regulation. this finding helped to identify students with a successfull profile. at Okpala (2002) , Here the authors conducted a study to find any relation between educational resources, student demographic and achivement scores. they used education production function because that is the dominant paradigm for analyzing impact of educational resources on student achievement. They used students socioeconomic status, community environment, class size, school size as input factors and achievement as the outcome. They Used Pearsons correlationa and regression to analyze the data. They found that student demographic was significant in explaining their achievement score. teaching experience, education level and school spending did not have much significant effect. But students socioeconomic status contributed significantly on their achievement score.

Based on these studies we can create a hypothesis. We can hypothesize to determine whether or not student demographic has any association with students Grade performance. We will use Pearson's correlation and Multiple Linear Regression as our inferential statistics, and Cohen's (1988) conventions as our effect size.

Method

The dataset we collected is from *Students Performance* (n.d.) . the dataset contains 10 personal queries,6 family quires. Other question are from educational habits. It has a sample of 145 students and 33 attributes. The columns are as follows;

- Student ID(Character) which has Unique identity characters.
- Student Age(Numeric) where 1: 18-21, 2: 22-25, 3: above 26.
- Sex(Numeric) where 1: female, 2: male.

- Graduated high-school type(Numeric) where 1: private, 2: state, 3: other.
- Scholarship type(Numeric) where 1: None, 2: 25%, 3: 50%, 4: 75%, 5: Full.
- Additional work(Numeric) where 1: Yes, 2: No .
- Regular artistic or sports activity (Numeric) where 1: Yes, 2: No .
- Do you have a partner (Numeric) where 1: Yes, 2: No .
- Total salary if available (Numeric) where 1: USD 135-200, 2: USD 201-270, 3: USD 271-340, 4: USD 341-410, 5: above 410 .
- Transportation to the university (Numeric) where 1: Bus, 2: Private car/taxi, 3: bicycle, 4: Other .
- Accommodation type in Cyprus (Numeric) where 1: rental, 2: dormitory, 3: with family, 4: Other .
- Mother's Education (Numeric) where 1: primary school, 2: secondary school, 3: high school, 4: university, 5: MSc., 6: Ph.D.

- Father's Education (Numeric) where 1: primary school, 2: secondary school, 3: high school, 4: university, 5: MSc., 6: Ph.D.
- Number of sisters/brothers (Numeric) where 1: 1, 2:, 2, 3: 3, 4: 4, 5: 5 or above .
- Parental status (Numeric) where 1: married, 2: divorced, 3: died - one of them or both .
- Mother's Occupation (Numeric) 1: retired, 2: housewife, 3: government officer, 4: private sector employee, 5: self-employment, 6: other .
- Father's Occupation(Numeric) 1: retired, 2: government officer, 3: private sector employee, 4: self-employment, 5: other .
- Weekly study hours (Numeric) 1: None, 2: <5 hours, 3: 6-10 hours, 4: 11-20 hours, 5: more than 20 hours .
- Reading frequency (non-scientific books/journals) (Numeric) 1: None, 2: Sometimes, 3: Often
- Reading frequency (scientific books/journals) (Numeric) 1: None, 2: Sometimes, 3: Often
- Attendance to the seminars/conferences related to the department (Numeric) 1: Yes, 2: No
- Impact of your projects/activities on your success (Numeric) 1: positive, 2: negative, 3: neutral
- Attendance to classes (Numeric) 1: always, 2: sometimes, 3: never
- Preparation to midterm exams 1 (Numeric) 1: alone, 2: with friends, 3: not applicable
- Preparation to midterm exams 2 (Numeric) 1: closest date to the exam, 2: regularly during the semester, 3: never
- Taking notes in classes (Numeric) 1: never, 2: sometimes, 3: always
- Listening in classes (Numeric) 1: never, 2: sometimes, 3: always
- Discussion improves my interest and success in the course(Numeric) 1: never, 2: sometimes, 3: always
- Flip-classroom (Numeric) 1: not useful, 2: useful, 3: not applicable
- Cumulative grade point average in the last semester (/4.00) (Numeric) 1: <2.00, 2: 2.00-2.49, 3: 2.50-2.99, 4: 3.00-3.49, 5: above 3.49 applicable *Expected Cumulative grade

point average in the graduation (/4.00)(Numeric) 1: <2.00, 2: 2.00-2.49, 3: 2.50-2.99, 4: 3.00-3.49, 5:

- Course ID (Numeric) unique number for each course
- Grade (Numeric) 0: Fail, 1: DD, 2: DC, 3: CC, 4: CB, 5: BB, 6: BA, 7: AA

Now we will explain step by step what we did in coding phase,

First we loaded the file "StudentsPerformance_with_headers.csv" using read_csv() function.

```
# loading dataset

library(readr)
Rdata <- read_csv("C:/Local Disk(A)/Trent University/Data Analytics with R/Assignment/Assignment 4/StudentsPerformance_with_headers.csv")
head(Rdata)
```

Then we checked the structure of the data using str() function for one last time before moving to data pre-processing phase to see if they are ok or not.

```
#checking data structure
str(Rdata)
```

We checked the name of the columns to see if they have any unnecessary naming scheme.

```
# checking column names
colnames(Rdata)
```

As it was already mentioned on the table above, most of them had big names so we had to change them to make them short and appropriate.

```
# Need to change the column name to shorter names.
colnames(Rdata)[colnames(Rdata)=="Graduated high-school type"] <- "School Type"
colnames(Rdata)[colnames(Rdata)=="Regular artistic or sports activity"] <- "Activity"
colnames(Rdata)[colnames(Rdata)=="Do you have a partner"] <- "partner stat"
colnames(Rdata)[colnames(Rdata)=="Total salary if available"] <- "salary"
colnames(Rdata)[colnames(Rdata)=="Transportation to the university"] <- "transport type"
colnames(Rdata)[colnames(Rdata)=="Accommodation type in Cyprus"] <- "Accommodation"
colnames(Rdata)[colnames(Rdata)=="Number of sisters/brothers"] <- "sibligs count"
colnames(Rdata)[colnames(Rdata)=="Reading frequency...19"] <- "reading non science"
colnames(Rdata)[colnames(Rdata)=="Reading frequency...20"] <- "reading science"
colnames(Rdata)[colnames(Rdata)=="Attendance to the seminars/conferences related to the department"] <- "attending seminars"
colnames(Rdata)[colnames(Rdata)=="Impact of your projects/activities on your success"] <- "impact of activities"
colnames(Rdata)[colnames(Rdata)=="Discussion improves my interest and success in the course"] <- "impact of discussion"
colnames(Rdata)[colnames(Rdata)=="Cumulative grade point average in the last semester (/4.00)"] <- "last grade"
colnames(Rdata)[colnames(Rdata)=="Expected Cumulative grade point average in the
```

```
graduation (/4.00)"] <- "expected grade"

colnames(Rdata)
```

Then we removed the following columns from the data : STUDENT ID, COURSE ID, expected grade, Additional work,Mother's education,Father's education,reading non science,reading science,impact of activities,Flip-classroom,last grade. We are removing student id because it is just a unique character, expected grade because we already have a column for previous grade, additional work because we have column which shows the impact of these additional work, Mother's and Father's education because it is not necessary because we already have other columns were it shows the occupation of the parents which can be more useful according to out literature review, reading science and non science had no particular reason to remove but it was just to decrease variables. Same reason for other varibles, they are not that necessary.

```
# We don't need all columns.
Rdata <- subset(Rdata, select = -c(`STUDENT ID`, `COURSE ID`, `expected grade`,
`Additional work`,`Mother's education`,`Father's education`,`reading non science`,`reading
science`,`impact of activities`,`Flip-classroom`,`last grade`))
View(Rdata)
```

Then we checked for null value on the dataset,

```
# Checking for Null values
is.na(Rdata)
```

Fortunately we did not observe any null in the dataset.

Now we are ready to move on to test our hypothesis. We will use the followings.

- Inferential Statistics
  - Correlation
  - Multiple Linear Regression
- Effect Size
  - Cohen's (1988) conventions(Multiple R^2)

We will first check the descriptive statistics of our data.

```
library(readr)
library(psych)
library(rstatix)
library(corrplot)

core.data <- Rdata
head(core.data)
# Descriptive Statistics
describe(core.data, fast = TRUE)
```

Before we go ahead with the test, we need to test some assumptions,

1. identifying **extreme outliers**
- Checking if there are any extreme outliers in the dataset.
2. Normality test
- We will use *Shapiro Wallis* test to see if the variables are **normally distriburted**.
3. linearity test
- We will check if the data in the variables are **linear** or not.

1)Identifying outliers:

```
# Testing some assumptions :

# 1) identifying extreme outliers.
identify_outliers(core.data, 'Student Age')
identify_outliers(core.data, Sex)
identify_outliers(core.data, 'School Type')
identify_outliers(core.data, 'Scholarship type')
identify_outliers(core.data, Activity)
identify_outliers(core.data, 'partner stat')
identify_outliers(core.data, salary)
identify_outliers(core.data, 'transport type')
identify_outliers(core.data, Accommodation)
identify_outliers(core.data, 'sibligs count')
identify_outliers(core.data, 'Parental status')
identify_outliers(core.data, 'Mother’s occupation')
identify_outliers(core.data, 'Father’s occupation')
identify_outliers(core.data, 'Weekly study hours')
identify_outliers(core.data, 'attending seminars')
identify_outliers(core.data, 'Attendance to classes')
identify_outliers(core.data, 'Preparation to midterm exams 1')
identify_outliers(core.data, 'Preparation to midterm exams 2')
identify_outliers(core.data, 'Taking notes in classes')
identify_outliers(core.data, 'Listening in classes')
identify_outliers(core.data, 'impact of discussion')
```

The following variables had extreme outliers : School type, Parental status, Mother’s occupation, attending seminars, Attendance to classes, Preparation to midterm exams 2 .

2)Normality test:

```
# 2)Normality test.
shapiro_test(core.data, vars = c("Student Age","Sex","School Type","Scholarship type","Activity","partner stat","salary","transport type","Accommodation","sibligs count","Parental status","Mother’s occupation","Father’s occupation","Weekly study hours","attending seminars","Attendance to classes","Preparation to midterm exams 1","Preparation to midterm exams 2","Taking notes in classes","Listening in classes","impact of discussion"))
```

In all cases p-value is less than alpha(0.05), thus the data is not normal.

3)linearity test:

```
plot(core.data$GRADE, core.data$`Student Age`)
plot(core.data$GRADE, core.data$Sex)
plot(core.data$GRADE, core.data$`School Type`)
plot(core.data$GRADE, core.data$`Scholarship type`)
plot(core.data$GRADE, core.data$Activity)
plot(core.data$GRADE, core.data$`partner stat`)
plot(core.data$GRADE, core.data$salary)
plot(core.data$GRADE, core.data$`transport type`)
plot(core.data$GRADE, core.data$Accommodation)
plot(core.data$GRADE, core.data$`sibligs count`)
plot(core.data$GRADE, core.data$`Parental status`)
plot(core.data$GRADE, core.data$`Mother’s occupation`)
plot(core.data$GRADE, core.data$`Father’s occupation`)
plot(core.data$GRADE, core.data$`Weekly study hours`)
plot(core.data$GRADE, core.data$`attending seminars`)
plot(core.data$GRADE, core.data$`Attendance to classes`)
plot(core.data$GRADE, core.data$`Preparation to midterm exams 1`)
plot(core.data$GRADE, core.data$`Preparation to midterm exams 2`)
plot(core.data$GRADE, core.data$`Taking notes in classes`)
plot(core.data$GRADE, core.data$`Listening in classes`)
plot(core.data$GRADE, core.data$`impact of discussion`)
```

The following data are not linear : sex, activity, partner stat, attending seminars, and attendance to classes. Other than these exception, all other varibles showed linearity.

No we can go and apply our inferential statistics. first we will apply Pearson’s Correlation Test. corr.test() is because we want to use all the varibles at once. We are including method= “pearson” because this is Pearson’s correlation test and adjust=“none” because we have too many variables. If we dont inclue adjust=“none” then we dont see the associated p-values. Here we are try to see correlation between Grade and other variables.

```
# Correlation test:
R <-corr.test(core.data[,c("Student Age","Sex","School Type","Scholarship
type","Activity","partner stat","salary","transport type","Accommodation","sibligs
count","Parental status","Mother’s occupation","Father’s occupation","Weekly study
hours","attending seminars","Attendance to classes","Preparation to midterm exams
1","Preparation to midterm exams 2","Taking notes in classes","Listening in classes","impact of
discussion","GRADE")],method = "pearson", adjust = "none")
R
```

the details of this result can be seen on results section.

Finally we can apply our Linear Regression. We are using LM function to use our model.Here we are using GRADE as our outcome and all other variables as predictors. To indicate all other variables we are using “~”. Here we are essentially trying to predict the grades using other varibles.

```
# Multiple linear regression
LM <- lm(`GRADE` ~ . , data = core.data)
summary(LM)
```

Just like correlation, the result of regression will be on the result section.

Here we visualized the correlations of the variables using carrplot library.

```
library(corrplot)
correlation_matrix <- cor(core.data[c("Student Age","Sex","School Type","Scholarship type","Activity","partner stat","salary","transport type","Accommodation","sibligs count","Parental status","Mother's occupation","Father's occupation","Weekly study hours","attending seminars","Attendance to classes","Preparation to midterm exams 1","Preparation to midterm exams 2","Taking notes in classes","Listening in classes","impact of discussion","GRADE")])

corrplot(correlation_matrix, method = "color", type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45, tl.cex = 0.6)
```

Results

Column Name	R Value	P-Value
Sex	0.34	p < 0.001
attending seminars	-0.18	p < 0.001
Student Age	-0.10	0.25
School type	0.34	0.21
Scholarship type	0.02	0.77
activity	-0.06	0.45
partner stat	-0.05	0.54
salary	-0.17	0.05
transport type	-0.16	0.06
accommodation	0.02	0.78
sibling count	0.08	0.31
parental status	0.07	0.43
mothers occupation	-0.03	0.71
fathers occupation	-0.04	0.60
weekly study hours	-0.03	0.69
attendance to classes	-0.14	0.09
preparation to mid exam 1	0.01	0.86
preparation to mid exam 2	0.07	0.38
taking notes in classes	0.04	0.59
listening in classes	0.09	0.31
impact of discussion	0.15	0.08

fig.1 : Pearson Correlation Output Output.

```
##
## Call:
## lm(formula = GRADE ~ ., data = core.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3904 -1.3274 -0.3318  1.4993  4.1948
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.14312    2.56322   0.446   0.6564
## `Student Age`                  -0.83631    0.34598  -2.417   0.0171 *
## Sex                             1.66937    0.40370   4.135 6.52e-05 ***
## `School Type`                   0.34489    0.34523   0.999   0.3197
## `Scholarship type`              0.02035    0.23828   0.085   0.9321
## Activity                       -0.33626    0.38192  -0.880   0.3803
## `partner stat`                  0.19485    0.38031   0.512   0.6093
## salary                         -0.31332    0.18178  -1.724   0.0873 .
## `transport type`               -0.34036    0.18040  -1.887   0.0616 .
## Accommodation                   0.24522    0.26516   0.925   0.3569
## `sibligs count`                 0.01824    0.14332   0.127   0.8989
## `Parental status`               0.63490    0.40554   1.566   0.1200
## `Mother's occupation`           0.01116    0.25202   0.044   0.9648
## `Father's occupation`          -0.15837    0.13852  -1.143   0.2551
## `Weekly study hours`           -0.09668    0.22080  -0.438   0.6623
## `attending seminars`           -0.58235    0.49818  -1.169   0.2447
## `Attendance to classes`        -0.28765    0.43195  -0.666   0.5067
## `Preparation to midterm exams 1` -0.04617    0.31408  -0.147   0.8834
## `Preparation to midterm exams 2`  0.21848    0.51222   0.427   0.6705
## `Taking notes in classes`      -0.16323    0.34278  -0.476   0.6348
## `Listening in classes`          0.58635    0.28429   2.062   0.0413 *
## `impact of discussion`          0.33829    0.30724   1.101   0.2730
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.011 on 123 degrees of freedom
## Multiple R-squared:  0.2846, Adjusted R-squared:  0.1625
## F-statistic: 2.331 on 21 and 123 DF,  p-value: 0.002171
```

fig.2 : Multiple Linear Regression Output.

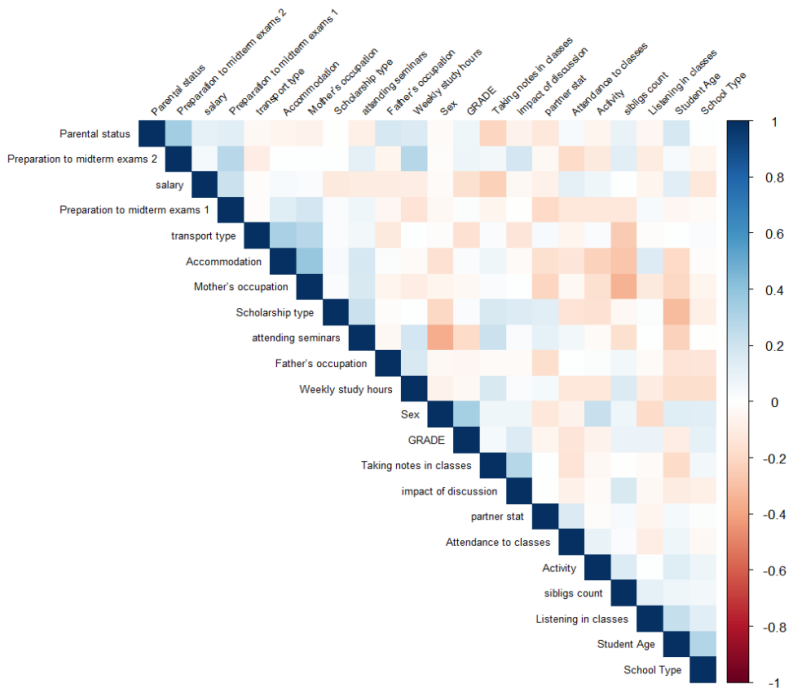

Fig.3 : Visualization of correlation between variables.

Current study south to determine whether or not the grade performance of a student has any assocoation with a students demographic characteristics. The study used a sampled data of 145 students. before applying any test we tested assumptions. The data contained extreme outliers in the variables of School type, Parental status, Mother's occupation, attending seminars, Attendance to classes and Preparation to midterm exams 2 . We applied Shapiro-Wilks test which showed significant p-values on all cases, meaning all instances the variables were not normally distributed. finally we used scatter plot to check the linearity of the data and The variables that were not linear were the following; sex, activity, partner stat, attending seminars, and attendance to classes. Other variables were linear. Nonetheless, Pearson's correlations were performed. Results of the correlational analysis showed that there were significant associations between Grade and Sex , r= 0.34, $p < 0.001$, Grade and attending seminars r= -0.18, $p < 0.001$. Other variables did not show significant associations. There was no significant association between Grade and Student Age r= -0.10, $p < 0.25$, Grade and School type r= 0.34 , $p < 0.21$, Grade and Scholarship type r= 0.02, $p < 0.77$, grade and activity r= -0.06, $p < 0.45$, grade and partner stat r= -0.05, $p < 0.54$, grade and salary r= -0.17, $p < 0.05$, grade and transport type r= -0.16, $p < 0.06$, grade and accommodation r= 0.02, $p < 0.78$, grade and sibling count r= 0.08, $p < 0.31$, grade and parental status r= 0.07, $p < 0.43$, grade and mothers occupation r= -0.03, $p < 0.71$, grade and fathers occupation r= -0.04, $p < 0.60$, grade and weekly study hours r= -0.03, $p < 0.69$, grade and attendance to classes r= -0.14, $p < 0.09$, grade and preparation to mid exam 1 r=0.01 , $p < 0.86$, grade and preparation to mid exam 2 r= 0.07, $p < 0.38$, grade and taking notes in classes r= 0.04, $p < 0.59$, grade and listening in classes r=0.09 , $p < 0.31$, grade and impact of

discussion r= 0.15, p < 0.08. Additionally, a multiple linear regression model was tested with Grade variable as outcome and Student Age,Sex,School Type,Scholarship type,Activity,partner stat,salary,transport type,Accommodation,sibligs count,Parental status,Mother's occupation,Father's occupation,Weekly study hours,attending seminars,Attendance to classes,Preparation to midterm exams 1,Preparation to midterm exams 2,Taking notes in classes,Listening in classes,impact of discussion as predictors. The overall regression model was significant; F(21,123) = 2.331, p<0.002171 , R^2=0.2846 . According to Cohen's (1988) conventions, the overall model explained a Large proportion of variability in Grade performance prediction. However only three predictors were significant in this model; Student Age t= -2.417,p<0.0171 , Sex t= 4.135,p<0.001, Listening in Classes t= 2.062,p<0.0413 . When one unit of Student Age increases, Grade performance decreases. However when One unit of sex and listening in classes increases, Grade performance also increases. The predictors that were not significant were the following; School Type t= 0.999,p< 0.3197, Scholarship type t= 0.085,p<0.9321 , Activity t= -0.880,p< 0.3803, partnar stat t= 0.512,p< 0.6093 , salary t= -1.724,p<0.0873 , transport type t= -1.887,p<0.0616 , Accommodation t= 0.925,p< 0.3569, siblings count t= 0.127,p<0.8989 , parental status t= 1.566,p< 0.1200, Mother's Occupation t= 0.044,p<0.9648 , Father's Occupation t= -1.143,p< 0.2551, Weekly study hour t= -0.438,p<0.6623 , attending seminar t= =1.169,p<0.2447 , attendance to classes t= -0.666,p<0.5067 , preparation to mid exam 1 t= -0.147,p<0.8834 , preparation to mid exam 2 t= 0.427,p< 0.6705, Taking notes in classes t= -0.476,p<0.6348 and impact of discussin t= 1.101,p<0.2730 .

Equation,

GRADE = 1.143 - 0.836(Student Age) + 1.669(Sex) + 0.586(Listening in classes)

Discussion

In conclusion, The study aimed to determine whether or not student demographic had any association with students Grade performance. We focused on Student Age,Sex,School Type,Scholarship type,Activity,partner stat,salary,transport type,Accommodation,siblings count,Parental status,Mother's occupation,Father's occupation,Weekly study hours,attending seminars,Attendance to classes,Preparation to midterm exams 1,Preparation to midterm exams 2,Taking notes in classes,Listening in classes and impact of discussion. We first pre-processed the dataset and then applied Pearson's Correlation test and Multiple Linear Regression on the data. In the end we did find some insights in what influences a students overall grade performance. From Pearson's Correlation we found out the significant correlations between grade and Sex, grade and attending seminars. From our Linear model we found significance in the predictors of Student age, sex and listening in classes. Overall the whole model showed significant p-value(less than 0.05). According to Cohen's (1988) conventions, the overall model explained a Large proportion of variability in Grade performance prediction. which is a good sign. According to our model, as student get older their grades get lower. As more time a student spends listening to class lecture, their chance of performing well also increases. Also according to our model male student tend to do better than female students. All these information can be very useful to the school authority. Because with this they know exactly what to change and what will effectively increase the performance of the students. They can encourage students to spend more time on class lectures to increase their performance. But there are a lot of draw backs in the model too. Its very important to discuss it and this is where our limitations comes. Before

we tested our inferential statistics, we had to test some assumptions. School type, Parental status, Mother's occupation, attending seminars, Attendance to classes, Preparation to midterm exams 2 showed extreme outliers in their data. None of the variables showed any sort of normality. sex, activity, partner stat, attending seminars, and attendance to classes failed to show any linearity. The first violation of assumption are serious. if our data has extreme outliers then that means our regression line was most likely influenced by those outliers. As none of the data are not normal, we simply cant use the data for normal distribution. This means we simply can not just trust the output of the model. For future project we need to be more cautious of analyzing the data. Need to take necessary to prevent them from violating any assumptions. To summarize, our study found significant association on the performance of students grade and students demographics with a large proportion of variability. But even with this valuable insight, we also need to acknowledge the limitation of our study.

Reference

Bilal, M., Omar, M., Anwar, W., Bokhari, R. H., & Choi, G. S. (2022). The role of demographic and academic features in a student performance prediction. *Scientific Reports*, *12*(1), 12508.

Coldwell, J., Craig, A., Paterson, T., & Mustard, J. (2008). Online students: Relationships between participation, demographics and academic performance. *Electronic Journal of e-Learning*, *6*(1), pp19–28.

Colorado, J. T., & Eberle, J. (2012). *Student demographics and success in online learning environments.*

Fitchett, P. G., & Heafner, T. L. (2017). Student demographics and teacher characteristics as predictors of elementary-age students' history knowledge: Implications for teacher education and practice. *Teaching and Teacher Education*, *67*, 79–92.

Johnson, G. M. (2015). On-campus and fully-online university students: Comparing demographics, digital technology use and learning characteristics. *Journal of University Teaching & Learning Practice*, *12*(1), 4.

Lord, S., Layton, R., Ohland, M., Brawner, C., & Long, R. (2014). A multi-institution study of student demographics and outcomes in chemical engineering. *Chemical Engineering Education*, *48*(4), 231–238.

Okpala, C. O. (2002). Educational resources, student demographics and achievement scores. *Journal of Education Finance*, *27*(3), 885–907.

Pilotte, M., Ohland, M. W., Lord, S. M., Layton, R. A., & Orr, M. K. (2017). Student demographics, pathways, and outcomes in industrial engineering. *International Journal of Engineering Education*, *33*(2), 506–518.

Rubright, J. D., Jodoin, M., & Barone, M. A. (2019). Examining demographics, prior academic performance, and united states medical licensing examination scores. *Academic Medicine*, *94*(3), 364–370.

Students Performance. (n.d.). Retrieved from https://www.kaggle.com/datasets/joebeachcapital/students-performance

VanderStel, A. (2014). *The impact of demographics in education*.