**UNIVERSITY-WIDE PREDICTORS OF ACADEMIC PERFORMANCE AT CAPE PENINSULA UNIVERSITY OF TECHNOLOGY**

A Technical Report

Presented to

Fundani Institutional Planning CPUT

by

Chumani Platyi

STUDENT NO. 219203261

As an Assessment for the Module

MATHEMATICAL SCIENCES PROJECT 4 (MSP470S)  
Within the Qualification

ADVANCED DIPLOMA IN MATHEMATICAL SCIENCES

ACADEMIC SUPERVISOR: Dr. T. Farrar

Cape Peninsula University of Technology

November, 2023

# DECLARATION

I, Chumani Platyi, declare that the contents of this research report represent my own work. I know that plagiarism is wrong. Plagiarism is to use another’s work and pretend that it is one’s own. Each contribution to, and quotation in, this report from the work(s) of other people has been attributed and has been cited and referenced.

I have not allowed and will not allow anyone to copy my work with the intention of passing it off as his or her own work.

Click on the box to confirm your agreement:

Date of Declaration: 29 November 2023

# ACKNOWLEDGEMENTS

I would like to thank my academic supervisor Dr. T. Farrar for giving me feedback and advice on my project.

**TABLE OF CONTENTS**

[DECLARATION ii](#_Toc110358305)

[ACKNOWLEDGEMENTS iii](#_Toc110358306)

[LIST OF TABLES v](#_Toc110358307)

[LIST OF FIGURES vi](#_Toc110358308)

[LIST OF SYMBOLS AND ABBREVIATIONS vii](#_Toc110358309)

[EXECUTIVE SUMMARY ix](#_Toc110358311)

[1. INTRODUCTION 1](#_Toc110358312)

[2. METHODOLOGY 3](#_Toc110358313)

[3. RESULTS AND DISCUSSION 4](#_Toc110358314)

[4. CONCLUSION AND RECOMMENDATIONS 5](#_Toc110358315)

[APPENDIX A. R CODE 6](#_Toc110358316)

[REFERENCES 7](#_Toc110358317)

# LIST OF TABLES

|  |  |  |  |
| --- | --- | --- | --- |
| **Number** | **Variable Name** | **Types** | **Units Measurement** |
| 1 | IDEN | Integer | Student Identification number. |
| 2 | FACULTY\_NAME | Categorical | A group of university departments. |
| 3 | DEPARTMENT\_NAME | Categorical | A division of courses. |
| 4 | PERIOD\_OF\_STUDY | Continuous | The time it takes for a student to successful complete a degree program. |
| 5 | SUBJECT\_NAME | Categorical | Name of each subject students registered for. |
| 6 | EXAM\_MONTH | Continuous | First and second semester (6 & 11) |
| 7 | FULL\_PERIOD\_MARK | Integer | The final mark you got after all your assessment. |
| 8 | FINAL\_MARK | Integer | Final marks of each student. |
| 9 | PASS\_FAIL | Categorical | You pass or fail an exam. |
| 10 | TRANSACTION\_DATE | Continuous | The date a student was registered. |
| 11 | YEAR\_APPLIED\_FOR | Continuous | The academic year you apply for. |
| 12 | ADMIT\_Y\_N | Categorical | Admitted or accepted. |
| 13 | ENGLISH | Integer | NSC English results. |
| 14 | ENGLISH\_TYPE | Categorical | Home Language or First Additional Language. |
| 15 | QUALIFICATION | Categorical | Name of qualification of each student. |
| 16 | RESIDENCE\_NAME | Categorical | Name of each student accommodation. |
| 17 | ACADEMIC\_YEAR | Continuous | Year a student registered. |
| 19 | FTEN\_STATUS\_THIS\_YEAR | Categorical | Current registered in 2023 academic year. |
| 20 | MATRIC\_DATE | Continuous | Year you did your secondary school. |
| 21 | CANCELLATION\_DATE | Continuous | Cancellation of subjects. |
| 22 | FINAL\_YEAR\_Y\_N | Continuous | Year you currently doing. |
| 23 | FIRST\_YEAR\_OF\_ENTRY | Continuous | The first you enroll in university |
| 24 | PROVINCE | Categorical | Province, you did secondary school. |
| 25  26  27 | Results  STAYED\_IN\_RESYes  RegDelayWeeks | Integer  Categorical  Integer | The final mark you got.  YES or NO.  Weeks from 1 to 15 |

# LIST OF FIGURES

**Figure 5**

A graph of a number of blue and pink bars

Description automatically generated

**Figure 6**

A graph of a number of students

Description automatically generated

# LIST OF SYMBOLS AND ABBREVIATIONS

STAYED\_IN\_RESYes: students who stayed in residence.

RegDelayWeeks: Registration date in weeks (the number of weeks).

# EXECUTIVE SUMMARY

This study describes the predictors of academic performance among all undergraduate Diploma and Degree programmes at Cape Peninsula University of Technology. This study predicts students’ performance to identify who these students are likely to fail so that the institution can help students who are facing academic difficulties. The aim of this project is to aggregate data from all undergraduate Diploma and Degree programmes (“first qualifications”) across the institution to obtain a “big picture” sense of the relationship between certain key variables and academic success, which could be turned into talking points for institutional decision makers. Predicting students’ performance will help the institute management to make strategy and decision making related to improving student performance. In this study R and Microsoft Power BI were used for analysis and one algorithm known as logistic regression.

This study finds that students who stayed in residence are likely to pass the subjects compared to those who are not stayed in residence. Also, this study shows that students who registered earlier in weeks 1 and 2 are likely to pass the subjects compared to those who registered late in week 4 to 15. Furthermore, it is found that the delay in registration has a negative impact on students’ performance. For every additional 1% in NSC English mark, odds of passing a CPUT subject are estimated to increase by 1.26%. Odds of passing are 40.8% higher for a student staying in residence, compared to a student not staying in residence and odds of passing decrease by 11.7% for every additional week of delay in registering.

Develop and implement specific academic support programs that address the identified predictors of academic performance. Engage with Student Organizations. Collaborate with student organizations to promote the importance of timely registration.

# INTRODUCTION

## Background

Performance Dashboards are a key component of a performance management system. Identifying and monitoring key performance metrics is crucial for university administration. The basic function of performance metrics is to assist in determining how well a particular university or department/faculty has achieved its respective goals. The performance indicators must be related to the objectives and strategies in universities and therefore this performance information must be presented to managers in a concise, intuitive format to support the university management processes. University can use dashboards to manage student, staff, department and research performance by setting metrics and managing those indicators over time through data visualization (Bologa, Florea, Muntean and Surcel, 2010).

## Business Problem

The Fundani Centre for Higher Education Development (CHED) is a strategic unit which, under the direction of CPUT vision and Mission as well as the Strategic Plan, serves the university as the site that initiates and facilitates higher education development in alignment with relevant international and national imperatives.

The Cape Peninsula University of Technology (CPUT) is facing challenges in optimizing academic performance across its entire university system. To enhance student success and educational outcomes, there is a need to identify and understand the key predictors of academic performance that are applicable university wide. By gaining insights into these predictors, the university aims to implement targeted interventions and strategies to improve overall student achievement.

## Chosen Solution(s)

So, we understand one solution to this dropping out problem is to identify as early as possible identity who these students who are likely to fail. To identify the lower performance, first we must predict future student performance.A Power BI dashboard that visualizes pass rate per subject in terms of the number of students enrolled, the registration date, and dashboard showing the relationship between a student’s in-residence and the outcome per subject they will be helpful in the problem of students’ performance.

* 1. **Objectives**

The goal of this project is to combine data from all Diploma and Degree programs across the institution to get a "big" picture of the relationship between specific priorities and academic success. Also, to answer the following three research questions:

1. What is the relationship between a student’s residence status (in res; not in res) and the outcome per subject (passes; fails)?

2. What is the relationship between a student’s registration date and the outcome per subject (passes; fails)?

3. What is the relationship between a student’s residence status (in res; not in res) and the throughput outcome (graduates in minimum time; graduates but not in minimum time; does not graduate)?

* 1. **Deliverables**

In this project R and Microsoft Power BI were used for analysis and one algorithm known as logistic regression. A Power BI dashboard showing the relationship between certain key variables and academic success will be created.

* 1. **Benefits**

The identification of predictors of academic performance allows the university to implement targeted interventions, leading to improved overall student achievement. This contributes to higher graduation rates and the academic achievement of students. Faculty members can benefit from insights into the predictors of academic success. Understanding these factors can guide professional development programs for educators, helping them tailor their teaching methods to better meet the needs of students.

# METHODOLOGY

**Data Source**

This study uses a dataset that consists of academic records and application data for all CPUT students studying undergraduate qualification from 2018 to 2022. This study also uses residence, curriculum and enrollment dataset. The dataset includes for each student the enrollment date for the diploma or degree, every single course they enrolled in. In this study R was used for analysis and Microsoft Power BI for data visualization to do dashboards. Academic records data have 1746490 observations and 26 variables, residence data have 48728 observations and 11 variables, and Application data 1348770 observations of 40 variables. The following table describes the variables of interest that we used in analysis.

academic records data variables: VIDEN, FACULTY\_NAME, DEPARTMENT\_NAME, YEAR, QUALIFICATION, QUALIFICATION\_DESCRIPTION, BLOCK\_CODE, OFFERING\_TYPE, PERIOD\_OF\_STUDY, SUBJECT, SUBJECT\_NAME, ACADEMIC\_BLOCK\_CODE, ACADEMIC\_BLOCK\_NAME\_DESC, OFFERING\_TYPE\_CODE, CAMPUS\_NAME, PERIOD\_OF\_STUDY\_1, SUBJECT\_TYPE\_CODE, STATS\_CREDIT, EXAM\_MONTH, EXAM\_TYPE, EXAM\_TYPE\_CODE, FULL\_PERIOD\_MARK,

FINAL\_MARK, RESULT\_DESCRIPTION,

PASS\_FAIL, TRANSACTION\_DATE.

application data variables: IDEN, ETHNIC\_GROUP, GENDER, COUNTRY, FACULTY\_NAME, DEPARTMENT\_NAME, YEAR\_APPLIED\_FOR, QUALIFICATION, QUALIFICATION\_DESCRIPTION, BLOCK\_CODE, OFFERING\_TYPE, CAMPUS\_NAME, PERIOD\_OF\_STUDY, CHOICE, ADMIT\_Y\_N, ADMIT\_STATUS\_NAME, MATRIC\_TYPE\_DESCRIPTION, SECONDARY\_SCHOOL\_QUINTILE, TRANSACTION\_DATE, MATRIC\_SUBJECT\_LEVEL, SCHOOL\_PROVINCE, ENGLISH, ENGLISH\_TYPE, MATHS, MATHS\_TYPE, LIFE ORIENTATION, LIFE SCIENCES, PHYSICAL SCIENCES, GEOGRAPHY, BUSINESS STUDIES, HISTORY, ECONOMICS, ACCOUNTING, TOURISM, AGRICULTURAL SCIENCE, COMPUTER APPLIC TECHN, CONSUMER STUDIES, ENG GRAPHICS & DESIGN, OTHER, and RSA LANGUAGE.

Residence data variables: IDEN, QUALIFICATION, QUALIFICATION\_DESCRIPTION, BLOCK\_CODE, PERIOD\_OF\_STUDY, OFFERING\_TYPE, CAMPUS\_NAME, RESIDENCE\_NAME, DATE\_IN, DATE\_OUT, and EXIT\_CODE\_DESCCRIPTION.

**Transformation of Data**

The academic records and residence data were joined using left join and the primary keys are IDEN and YEAR. New variables were created STAYED\_IN\_RES, RegDelayWeeks and RegDelayDays was created after joining the academic records data with the residence data.

STAYED\_IN\_RES, if a student had a “Date in “in the residence data in a particular year, they were considered a “Yes’ (stayed in residence) and, if they did not have a “Date in “in the residence data in a particular year, they were considered a “No’ (did not stayed in residence).

RegDelayWeeks variable, the number of days (or weeks) was counted between the earliest possible registration date for that period of study and the date when the student registered, for that period of study.

Also, the academic records and application data were joined using left join and primary key is IDEN. The Period of Study was also filtered to be 4 or less, meaning that undergraduate students only, since Period of Study 5 and above are postgraduate students.

**Data analysis**

In this study R and Microsoft Power BI were used for analysis. This study uses the following software packages: tidyverse (Chang,Miller, and Muller, 2019) and broom (Couch , Hayes, and Robinson, 2023).

Two logistic Regression Equations:

Logit(P)= -0.412 +0.0126ENGLISH​+0.369​ENGLISH\_TYPEHOME ​+0.0151LIFE ORIENTATION

Logit(P) = 1.77+0.34STAYED\_IN\_RESYes ​-0.111RegDelayWeeks​

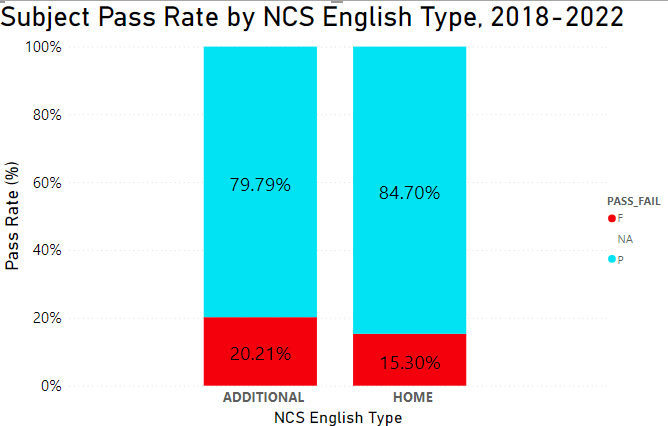
# RESULTS AND DISCUSSION

**Figure 1**

|  |
| --- |
| term estimate std.error statistic p.value |
| *<chr>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* |
| 1 (Intercept) -0.412 0.0266 -15.5 4.11e- 54 |
| 2 ENGLISH 0.0126 0.000392 32.1 2.00e-225 |
| 3 ENGLISH\_TYPEHOME 0.369 0.00711 51.9 0 |
| 4 `LIFE ORIENTATION` 0.0151 0.000337 44.8 0 |

ENGLISH = 1.012635. For every additional 1% in NSC English mark, odds of passing a CPUT subject are estimated to increase by 1.26%. LIFE ORIENTATION = 1.01523. For every additional 1% in NSC Life Orientation mark, odds of passing a CPUT subject are estimated to increase by 1.52%. ENGLISH\_TYPEHOME = 1.446236. Odds of passing a CPUT subject are 44.6% higher for a student who did English Home Language in NSC, then for a student who did English First Additional Language in NSC.

**Figure 2**



This study shows that students who hold a National Senior Certificate (NSC) with English as home language has more chances to pass a CPUT subject compared to those who did English as a first additional language. This study shows that about 84.70% of students who pass English as home language and 79.79% of students who pass English as a first additional language. Also, it is found that students who did English as home language are more likely to pass compared to those who did English as a first additional language. About 20.21% of student who fail English as a first additional language and only 15.30% of those who fail English as home language.

**Figure 3**

A graph of blue and pink bars

Description automatically generated

This study shows that NSC English has an impact in subject pass rate at CPUT. It is found that students who got NSC English mark between 0 to 60 are likely to fail their CPUT subjects compared to those who got 60 to 70 and 70 to 80. This study shows that students who pass their NSC English mark with 0 to 60 marks are at risk of failing CPUT subjects. This study found that students who got 70 and above in NSC English mark has more chances to pass CPUT subjects compared to those who got in NSC English mark less than 70. Furthermore, this study shows that NSC English mark has an impact with the subject pass rate at CPUT. It has a negative impact to students who got NSC English mark less than 70, that clearly shows us that the logistics regression model and this graph give us the same results. From the logistics regression model above, it is found that for every additional 1% in NSC English mark, odds of passing a CPUT subject are estimated to increase by 1.26%.

**Figure 4**

A graph of a number of students

Description automatically generated

This study shows that the histogram is Skewed left. Therefore, the mean is less than the median. The distribution has many more data points on one side of a graph than the other. Students who wrote NSC English many of them pass with 50% and above.

The APS Score is calculated using three methos at CPUT namely method 1, method 2, and method 3. NSC English (HL or FAL) is the required subject to calculate APS score with all three methods. Almost all programmes at CPUT have an admission requirement of at least code 4 (50%) in English (HL or FAL). This is why there are very few students in the histogram with less than 50 tend not to be accepted into programmes at CPUT.

**Figure 5**

In figure 5 the study shows that students who pass their NSC Life Orientation Mark between 0 to 40, 40 to 50, and 50 to 60 are likely to fail CPUT subjects. Also, it is found that students who got NSC Life Orientation Mark between 60 to70 and above are likely to pass CPUT subjects. This study found that the two NSC variables English Mark and Life Orientation Mark shows similar results. We can conclude that NSC English and NSC Life Orientation Mark are directly proportional to the CPUT subject pass rate.

**Figure 6**

In figures 4 and 6 there is not much difference, the two histograms look the same. This graph shows that the histogram is Skewed left. Therefore, the mean is less than the median. The distribution has many more data points on one side of a graph than the other. Students who wrote NSC Life Orientation many of them pass with 50% and above. Calculating the APS Score at CPUT Life Orientation cannot be used for any of the calculations. However, for predictions we can use NSC Life Orientation since all students who enroll at CPUT did NSC Life Orientation. From the logistics regression model above, it is found that for every additional 1% in NSC Life Orientation mark, odds of passing a CPUT subject are estimated to increase by 1.52%.

**Figure 7**

|  |
| --- |
| term estimate std.error statistic p.value |
| *<chr>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* |
| 1 (Intercept) 1.77 0.00448 397. 0 |
| 2 STAYED\_IN\_RESYes 0.342 0.00690 49.6 0 |
| 3 RegDelayWeeks -0.111 0.000947 -117. 0 |

STAYED\_IN\_RESYes = 1.408232. Odds of passing are 40.8% higher for a student staying in residence, compared to a student not staying in residence. RegDelayWeeks = 1.11731 odds of passing decrease by 11.7% for every additional week of delay in registering.

**Figure 8**

A graph of a number of people with blue and red squares

Description automatically generated with medium confidence

This study shows that about 85.86% of students who stayed in residence pass their subjects and 80.30% of students do not stay in residence pass their subjects. The chances of passing are 5.56% higher for a student staying in residence, compared to a student not staying in residence. This study finds that students who stayed in residence are likely to pass the subjects compared to those who are not stayed in residence. Students who are not in residence 19.70% of them fail their subjects while students who stay in residence only 14.13% who fail their subjects.

**Figure 9**

A graph of a number of blue and red bars

Description automatically generated

This study shows that students who registered earlier in weeks 1 and 2 are likely to pass the subjects compared to those who registered late in week 3 to 15. Furthermore, it is found that the delay in registration has a negative impact on students’ performance. The cumulative effects of limited course choices, scheduling conflicts, missed assignments, and increased stress can lead to lower academic performance. Late registrants may struggle to achieve the same level of success as those who registered on time. That’s why students who registered later in weeks 10 to 15 have higher percentage of fail compared to those who registered earlier in weeks 1 and 2. It may happen they missed some assessments in week 1 and 2.

# CONCLUSION AND RECOMMENDATIONS

This study finds that students who stayed in residence are likely to pass the subjects compared to those who are not stayed in residence. Also, this study shows that students who registered earlier in week 1 and 2 are likely to pass the subjects compared to those who registered late in week 4 to 15. Furthermore, it is found that the delay in registration has a negative impact on students’ performance. Develop and implement specific academic support programs that address the identified predictors of academic performance. This could include tutoring services, study groups, or workshops tailored to the needs of students in different disciplines. Engage with Student Organizations. Collaborate with student organizations to promote the importance of timely registration. Student leaders can play a key role in encouraging their peers to register early and can serve as advocates for the benefits of timely registration. Monitor and Analyze Registration Data, regularly monitor registration data to identify trends and patterns. Analyzing the data can provide insights into the reasons behind late registrations, allowing the university to tailor interventions to specific student needs. Enhance communication channels to ensure that students receive clear and timely information about registration deadlines, processes, and requirements. Utilize multiple communication platforms, including the university website, social media, emails, and posters, to reach a broad audience.

# APPENDIX A. r code

setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

install.packages("tidyverse")

library("tidyverse")

load("academic\_records\_2023-09-23.RData")

load("enrollments\_residence\_curriculum\_2023-09-23.RData")

load("appdat.RData")

gc()

names(acadat)

install.packages("dplyr")

library("dplyr")

resdat2 <- resdat %>%

mutate(DATE\_IN = as.Date(DATE\_IN, format = "%d/%B/%y"),

YEAR = year(DATE\_IN),

STAYED\_IN\_RES = "Yes") %>%

select(c(IDEN, YEAR, STAYED\_IN\_RES)) %>%

distinct()

# First semester subjects only

acajoin <- left\_join(x = acadat,

y = resdat2,

by = join\_by(IDEN == IDEN,

YEAR == YEAR)) %>%

mutate(STAYED\_IN\_RES = if\_else(is.na(STAYED\_IN\_RES), "No", "Yes")) %>%

filter(YEAR >= 2018,

PERIOD\_OF\_STUDY <= 4)

#Replace NA with zero

na.omit(acajoin)

#Convert acajoin to csv file

write.csv(acajoin, file = "acajoin.csv")

acajoin %>%

filter(ACADEMIC\_BLOCK\_CODE %in% c("0", "1")) %>%

mutate(FirstRegDate = case\_when(

YEAR == 2018 ~ as.Date("2018-01-08"),

YEAR == 2019 & (TRANSACTION\_DATE < "2019-01-14" |

PERIOD\_OF\_STUDY > 1) ~ as.Date("2019-01-07"),

YEAR == 2019 ~ as.Date("2019-01-14"),

YEAR == 2020 & (TRANSACTION\_DATE < "2020-01-13" |

PERIOD\_OF\_STUDY > 1) ~ as.Date("2020-01-06"),

YEAR == 2020 ~ as.Date("2020-01-13"),

YEAR == 2022 & (TRANSACTION\_DATE < "2022-01-31" |

PERIOD\_OF\_STUDY > 1) ~ as.Date("2022-01-10"),

YEAR == 2022 ~ as.Date("2022-01-31"),

YEAR == 2021 & (TRANSACTION\_DATE < "2022-03-08" |

PERIOD\_OF\_STUDY > 1) ~ as.Date("2021-01-11"),

YEAR == 2021 ~ as.Date("2021-03-08")),

RegDelayWeeks = interval(start = ymd(FirstRegDate),

end = ymd(TRANSACTION\_DATE)) %/% weeks(1),

RegDelayDays = interval(start = ymd(FirstRegDate),

end = ymd(TRANSACTION\_DATE)) %/% days(1)) %>%

filter(between(RegDelayDays, 0, 105)) ->

acajoin1

#Replace NA with zero

na.omit(acajoin1)

#Convert acajoin1 to csv file

write.csv(acajoin1, file = "acajoin1.csv")

acajoin1 %>%

drop\_na(PASS\_FAIL) %>%

ggplot(mapping = aes(x = RegDelayWeeks, fill = PASS\_FAIL)) +

geom\_bar(stat = "count", position = "fill") +

scale\_y\_continuous(labels = scales::percent) +

labs(fill = "Subject Outcome",

x = "Delay in Registering (Weeks)",

y = "Pass Rate (Percent)",

title = "Subject Pass Rate by Registration Delay, 2018-2022")

# Check for correlation between registration delay and final mark

cor.test(acajoin1$FINAL\_MARK,

acajoin1$RegDelayDays, na.rm = TRUE)

acajoin %>%

drop\_na(PASS\_FAIL) %>%

ggplot(mapping = aes(x = STAYED\_IN\_RES, fill = PASS\_FAIL)) +

geom\_bar(stat = "count", position = "fill") +

scale\_y\_continuous(labels = scales::percent) +

labs(x = "Stayed in Residence",

y = "Pass Rate (%)",

title = "Subject Pass Rate by Residence Status, 2018-2022")

acajoin.lm <- acajoin1 %>%

drop\_na(PASS\_FAIL) %>%

mutate(PASS\_FAIL = factor(PASS\_FAIL) %>% relevel(ref = "F"))

mylogistic <- glm(PASS\_FAIL ~ STAYED\_IN\_RES + RegDelayWeeks,

data = acajoin.lm,

family = binomial(link = "logit"))

install.packages("broom")

library("broom")

results <- tidy(mylogistic)

exp(results$estimate[2])

# 1.408232

# Odds of passing are 40.8% higher for a student staying in residence

# compared to a student not staying in residence

1 / exp(results$estimate[3])

# 1.11731

# odds of passing decrease by 11.7% for every additional week

# of delay in registering

appdat2 <- appdat %>%

drop\_na(ENGLISH, `LIFE ORIENTATION`, ENGLISH\_TYPE) %>%

select(IDEN, ENGLISH, `LIFE ORIENTATION`, ENGLISH\_TYPE) %>%

arrange(IDEN, desc(ENGLISH), desc(`LIFE ORIENTATION`)) %>%

distinct(IDEN, .keep\_all = TRUE)

appjoin <- left\_join(x = acadat,

y = appdat2,

by = join\_by(IDEN == IDEN)) %>%

filter(YEAR >= 2018, PERIOD\_OF\_STUDY <= 4)

#Replace NA with zero

na.omit(appjoin)

#Convert appjoin to csv file

write.csv(appjoin, file = "appjoin.csv")

appjoin %>%

distinct(IDEN, .keep\_all = TRUE) %>%

ggplot(mapping = aes(x = ENGLISH)) +

geom\_histogram(colour = "black", fill = "cyan", binwidth = 5) +

labs(x = "NSC English Mark",

y = "Number of Students",

title = "NSC English Marks of Undergraduate Students, 2018-2022")

appjoin %>%

distinct(IDEN, .keep\_all = TRUE) %>%

ggplot(mapping = aes(x = `LIFE ORIENTATION`)) +

geom\_histogram(colour = "black", fill = "cyan", binwidth = 5) +

labs(x = "NSC Life Orientation Mark",

y = "Number of Students",

title = "NSC Life Orientation Marks of Undergraduate Students, 2018-2022")

appjoin %>%

drop\_na(PASS\_FAIL, ENGLISH) %>%

ggplot(mapping = aes(x = cut(ENGLISH,

breaks = c(0, seq(50, 100, 10)),

include.lowest = TRUE),

fill = PASS\_FAIL)) +

geom\_bar(stat = "count", position = "fill") +

scale\_y\_continuous(labels = scales::percent) +

labs(x = "NSC English Mark",

y = "Pass Rate (%)",

title = "Subject Pass Rate by NSC English Mark, 2018-2022")

appjoin %>%

drop\_na(PASS\_FAIL, ENGLISH\_TYPE) %>%

ggplot(mapping = aes(x = ENGLISH\_TYPE,

fill = PASS\_FAIL)) +

geom\_bar(stat = "count", position = "fill") +

scale\_y\_continuous(labels = scales::percent) +

labs(x = "NSC English Type",

y = "Pass Rate (%)",

title = "Subject Pass Rate by NSC English Type, 2018-2022")

appjoin %>%

drop\_na(PASS\_FAIL, `LIFE ORIENTATION`) %>%

ggplot(mapping = aes(x = cut(`LIFE ORIENTATION`,

breaks = c(0, seq(40, 100, 10)),

include.lowest = TRUE),

fill = PASS\_FAIL)) +

geom\_bar(stat = "count", position = "fill") +

scale\_y\_continuous(labels = scales::percent) +

labs(x = "NSC Life Orientation Mark",

y = "Pass Rate (%)",

title = "Subject Pass Rate by NSC Life Orientation Mark, 2018-2022")

appjoin.lm <- appjoin %>%

drop\_na(PASS\_FAIL) %>%

mutate(PASS\_FAIL = factor(PASS\_FAIL) %>% relevel(ref = "F"))

mylogistic2 <- glm(PASS\_FAIL ~ ENGLISH + ENGLISH\_TYPE + `LIFE ORIENTATION`,

data = appjoin.lm,

family = binomial(link = "logit"))

library(broom)

results2 <- tidy(mylogistic2)

exp(results2$estimate[2])

# 1.012635

# For every additional 1% in NSC English mark,

# odds of passing a CPUT subject are estimated to increase by 1.26%

exp(results2$estimate[4])

# 1.01523

# For every additional 1% in NSC Life Orientation mark,

# odds of passing a CPUT subject are estimated to increase by 1.52%

exp(results2$estimate[3])

# 1.446236

# Odds of passing a CPUT subject are 44.6% higher for a student

# who did English Home Language in NSC, than for a student

# who did English First Additional Language in NSC

# REFERENCES

*Cape Peninsula University of Technology*. Available at: https://www.cput.ac.za/about/history (Accessed: 29 September 2023).

Fox J, Weisberg S (2019). \_An R Companion to Applied Regression\_, Third edition. Sage, Thousand Oaks CA. <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>.

*Fundani - Cape Peninsula University of Technology*. Available at: https://sda.cput.ac.za/wp-content/uploads/2022/06/FUNDANI-update.pdf (Accessed: 29 September 2023).

Huang, S. and Fang, N., 2013. Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers & Education*, *61*, pp.133-145.

Muntean, M.I.H.A.E.L.A., Sabau, G., Bologa, A., Surcel, T.R.A.I.A.N. and Florea, A.L.E.X.A.N.D.R.A., 2010, September. Performance dashboards for universities. In *Proceedings of the 2nd international conference on manufacturing engineering, quality and production systems* (pp. 206-211).

Wagner, K., Merceron, A. and Sauer, P., 2020. Accuracy of a cross-program model for dropout prediction in higher education. In *Companion Proceedings of the 10th International Learning Analytics & Knowledge Conference (LAK 2020)* (pp. 744-749).