R PROJEKT MA0009: Graduate Admissions

Wie man eine Bewerbung stärken kann

Mikel Funes Morfín, Maxim Alexander Baumgärtel

2024-02-10

Contents

1 Einleitung

Als heranwachsende Studenten aus internationalen Hintergründen interessiert uns die Optimierung von unseren Bewerbungen für ein weiteres Studium und auch die Stärke einer solchen Bewerbung. Wir beziehen uns auf den Datensatz von [https://www.kaggle.com/datasets/akshaydattatraykhare/data-for-admission-in-the-university]. Weiter wollen wir schauen wie sich die Durchschnittsergebnisse aus den standardisierten Tests verändert anhand des Ratings der Universität.

1.1 Erste Übersicht

Wir erstellen eine intiale Übersicht mittels der *summary()* - Funktion

```
dataset %>%
  summary()
```

```
##    Serial No.      GRE Score      TOEFL Score    University Rating
##  Min.   :  1.0   Min.   :290.0   Min.   : 92.0   Min.   :1.000
##  1st Qu.:100.8   1st Qu.:308.0   1st Qu.:103.0   1st Qu.:2.000
##  Median :200.5   Median :317.0   Median :107.0   Median :3.000
##  Mean   :200.5   Mean   :316.8   Mean   :107.4   Mean   :3.087
##  3rd Qu.:300.2   3rd Qu.:325.0   3rd Qu.:112.0   3rd Qu.:4.000
##  Max.   :400.0   Max.   :340.0   Max.   :120.0   Max.   :5.000
##       SOP             LOR             CGPA          Research
##  Min.   :1.0   Min.   :1.000   Min.   :6.800   Min.   :0.0000
##  1st Qu.:2.5   1st Qu.:3.000   1st Qu.:8.170   1st Qu.:0.0000
##  Median :3.5   Median :3.500   Median :8.610   Median :1.0000
##  Mean   :3.4   Mean   :3.453   Mean   :8.599   Mean   :0.5475
##  3rd Qu.:4.0   3rd Qu.:4.000   3rd Qu.:9.062   3rd Qu.:1.0000
##  Max.   :5.0   Max.   :5.000   Max.   :9.920   Max.   :1.0000
##  Chance of Admit
##  Min.   :0.3400
##  1st Qu.:0.6400
##  Median :0.7300
##  Mean   :0.7244
##  3rd Qu.:0.8300
##  Max.   :0.9700
```

Wir haben 9 verschiedene Spalten: Serial No., GRE Score, TOEFL Score, University Ranking, SOP, LOR, CGPA, Research und Chance of Admit.

1.1.1 Aufbau eines Dateneintrags

In diesem Abschnitt wollen wir zeigen, wie die Daten zustande gekommen sind und wo mögliche Korrelationen enstehen könnten zwischen den Variablen und einer Bewewerbung.

Angenommen wir sind ein Student im Bachelor und wollen uns auf ein Masterprogramm bewerben. Wir wollen unsere 'Chance of Admit' maximieren, das entspricht unserer Wahrscheinlichkeit in das Programm angenommen zu werden. Hier müssen wir kurz klarstellen, dass ein maximaler 'Chance of Admit' Eintrag nicht direkt aussagt, dass man angenommen wurde. Abhängig von der Universität die wir besuchen wollen, könnten sich die Aufnahme Kriterien verändern, je nach Ranking. Jetzt kommen wir zu den Kriterien die wir als Student erfüllen könnten.

Alle folgenden Unterüberschriften sind bezogen auf die Variablen innerhalb unserer Datenanalyse.

1.1.2 Was ist GRE Score?

Der GRE (Graduate Record Examination) ist ein standardisierter Test, der darauf abzielt, die Fähigkeiten eines Kandidaten in den Bereichen analytisches Schreiben, abstraktes Denken, Mathematik und allgemeines Vokabular zu bewerten. Für amerikanische Graduate Schools ist der GRE ein Schlüsselfaktor, um die Eignung eines Kandidaten für ihre Masterprogramme zu beurteilen. Der Test besteht aus drei Teilen: quantitativ, verbal und schriftlich. In jedem der beiden ersten Bereiche können maximal 170 Punkte und im schriftlichen Teil bis zu 6 Punkte erreicht werden.

1.1.3 Was ist TOEFL Score?

Der TOEFL (Test of English as a Foreign Language) ist eine standardisierte Prüfung, die speziell dafür entwickelt wurde, um die Englischkenntnisse von Nicht-Muttersprachlern zu beurteilen. Dieser Test ist besonders relevant für Studierende, die sich auf amerikanische Masterprogramme bewerben. Der TOEFL überprüft die Fähigkeiten in vier Bereichen: Lesen, Hören, Sprechen und Schreiben. Dabei werden das Verständnis und die Analyse von akademischen Texten, das Verständnis der englischen Sprache durch Hörbeispiele, die Fähigkeit, auf Situationen oder Themen zu reagieren und Meinungen zu äußern, sowie die Kompetenz, klar und logisch in Englisch zu schreiben, bewertet. Die maximale Punktzahl im TOEFL Test beträgt 120 Punkte, wobei jeder Bereich bis zu 30 Punkte beitragen kann.

1.1.4 Was ist University Ranking?

Das University Ranking ist eine Variable im Datensatz, die das Ranking der Universität darstellt, wobei keine weiteren Informationen über die Bewertungsmethodik vorliegen. Es wird angenommen, dass eine 5 das bestmögliche Ranking repräsentiert.

1.1.5 Was ist SOP?

Das SOP (Statement of Purpose) ist vergleichbar mit einem Motivationsschreiben, das in Bewerbungen für Master- oder Doktorandenprogramme verwendet wird. Es zielt darauf ab, die Persönlichkeit des Bewerbers, seine Karriereziele, Inspirationen und Qualifikationen zu präsentieren. Wesentliche Aspekte des SOP umfassen die Darstellung der Motivation und der Gründe für das Interesse an einem bestimmten Studienfeld oder Programm, die Beschreibung der Karriereziele und wie das Programm dabei helfen kann, diese zu erreichen, die Hervorhebung relevanter akademischer und beruflicher Erfahrungen sowie Einblicke in die Persönlichkeit des Bewerbers und deren Beitrag zum akademischen und beruflichen Werdegang. Die Bewertung des SOP erfolgt auf einer Skala von 0 bis 5, wobei 5 die höchste Punktzahl darstellt.

1.1.6 Was ist LOR?

Ein LOR (Letter of Recommendation) ist ein wichtiges Dokument, das üblicherweise bei Bewerbungen für höhere Bildungsprogramme eingereicht wird. Es dient dazu, die Qualifikationen, Fähigkeiten und Charaktereigenschaften des Bewerbers aus der Perspektive einer vertrauenswürdigen und meist hierarchisch übergeordneten Person darzustellen. Ein effektiver LOR bestätigt die Qualifikationen des Bewerbers, bietet persönliche Einsichten in seine Persönlichkeit und Arbeitsweise und stärkt die Bewerbung durch die Bestätigung der Eignung für das Programm oder die Position aus einer glaubwürdigen Quelle.

1.1.7 Was ist CGPA?

Der “Current GPA” (aktuelle GPA) bezieht sich auf den aktuellen Notendurchschnitt eines Studierenden. “GPA” steht für “Grade Point Average” (Notendurchschnitt).

1.1.8 Was ist Research?

Die Variable “Research” wird in den Daten lediglich als Wert 1 oder 0 angegeben, wobei eine 1 vorhandene Forschungserfahrung repräsentiert und eine 0 das Fehlen von Forschungserfahrung anzeigt.

2 Daten

Die Einträge aus dem Datensatz enthalten Leerzeichen innerhalb der Nomenklatur. Dies könnte zu möglichen Fehlern innerhalb der Bearbeitung führen. Um das zu umgehen haben wir der Sauberkeithalber die Nomenklatur manuell ersetzt. In dem neuen Datensatz sind an Stelle der Leerzeichen nun "." Wir erkennen auch, dass der Eintrag 'Serial No.' keine Aussagekraft für unsere Analyse besitzt. Somit können wir diese Variable rauslassen. Das machen wir durch mit "dplyr."

```
dataset1 <- dataset1 %>% select(-Serial.No.)
```

Innerhalb dieses Abschnittes wollen wir die Verteilungen der Variablen analysieren und veranschaulichen. Dazu betraachten wir in erster Linie Boxplots der Daten.

Wir erhalten eine erste Übersicht der Daten durch die *head()*-Funktion.

```
# Zusammenfassung der Daten
head(dataset1)
```

```
## # A tibble: 6 x 8
##   GRE.Score TOEFL.Score University.Rating   SOP   LOR  CGPA Research
##       <dbl>       <dbl>             <dbl> <dbl> <dbl> <dbl>    <dbl>
## 1       337         118                 4   4.5   4.5  9.65        1
## 2       324         107                 4   4     4.5  8.87        1
## 3       316         104                 3   3     3.5  8           1
## 4       322         110                 3   3.5   2.5  8.67        1
## 5       314         103                 2   2     3    8.21        0
## 6       330         115                 5   4.5   3    9.34        1
## # i 1 more variable: Chance.of.Admit <dbl>
```

```
# Struktur der Daten
str(dataset1)
```

```
## tibble [400 x 8] (S3: tbl_df/tbl/data.frame)
##  $ GRE.Score        : num [1:400] 337 324 316 322 314 330 321 308 302 323 ...
##  $ TOEFL.Score      : num [1:400] 118 107 104 110 103 115 109 101 102 108 ...
##  $ University.Rating: num [1:400] 4 4 3 3 2 5 3 2 1 3 ...
##  $ SOP              : num [1:400] 4.5 4 3 3.5 2 4.5 3 3 2 3.5 ...
##  $ LOR              : num [1:400] 4.5 4.5 3.5 2.5 3 3 4 4 1.5 3 ...
##  $ CGPA             : num [1:400] 9.65 8.87 8 8.67 8.21 9.34 8.2 7.9 8 8.6 ...
##  $ Research         : num [1:400] 1 1 1 1 0 1 1 0 0 0 ...
##  $ Chance.of.Admit  : num [1:400] 0.92 0.76 0.72 0.8 0.65 0.9 0.75 0.68 0.5 0.45 ...
```

2.1 Visualisierung der Daten

Wir wollen die Verteilung der Beobachtungen visualisieren, das machen wir durch Histogramme:

2.1.1 CGPA

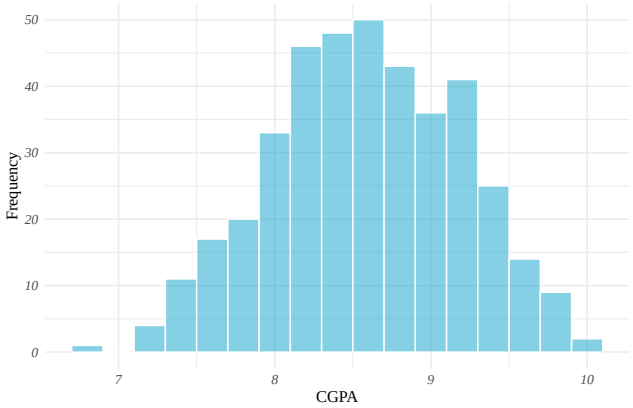


Figure 1: CGPA Histogramm

Hier können wir erkennen, dass der Median und der Mittelwert bei ungefähr 8.6 liegen. Dies weißt daraufhin, dass die Durchschnittsnote der Bewerbungen relativ hochliegt. Weiter können wir sehen, dass der minimale Notenschnitt bei 6.8 liegt und der maximale bei ungefähr 9.92 liegt. Wir müssen hier anmerken, dass die Durchschnittsnoten nicht mit Notensystem übereinstimmen, was an der TUM verbreitet ist.

2.1.2 GRE Score

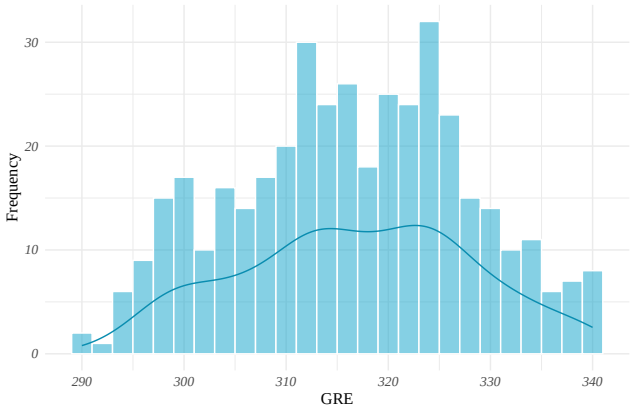


Figure 2: GRE Score Histogramm

Bei den GRE Scores der Bewerber wird ein ähnliches Bild gespiegelt, wie in der vorigen Variable. Der Median und Durchschnitt liegen bei ungefähr 317 Punkten. Hier wäre es auch interessant zu sehen, wie die Punkte aufgebaut sind anhand der Resultate aus dem verbalen und dem quantitativen Teil. Für weitere Infos zu den Verteilungen der Resultate (Blog 2021). Der maximale Wert aus dem Datensatz ist 340.

2.1.3 TOEFL Score

Durch das Histogramm können wir erkennen, dass der Durchschnitt und Median nahe bei einanderliegen, mit einem ungefähren Wert von 107. Das Maximum beträgt hier 120 Punkte und das Minimum sind 92 Punkte.

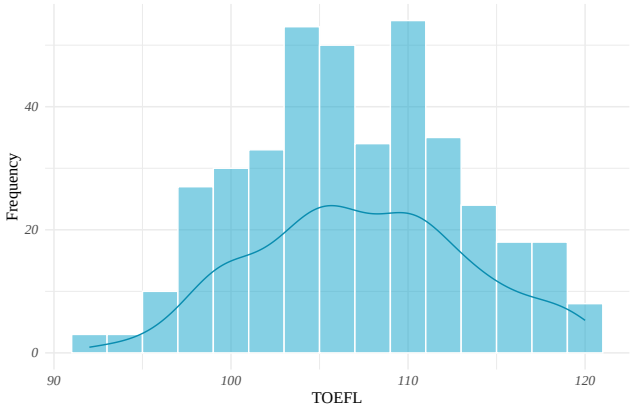


Figure 3: TOEFL Score Histogramm

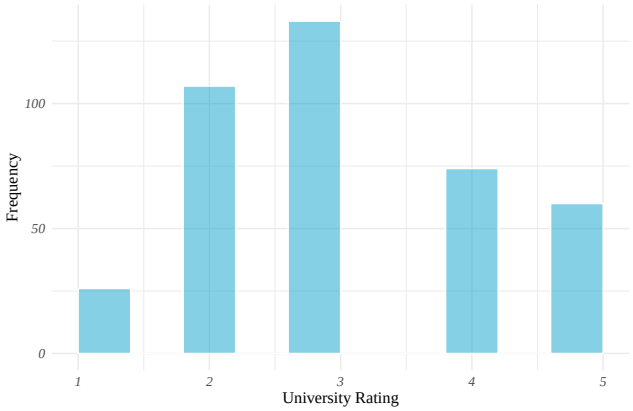


Figure 4: University Rating Histogramm

2.1.4 University Rating

Die Daten zum Hochschulrating geben Auskunft über die Bewertung der Hochschulen mit einem Minimum von 1 und einem Maximum von 5. Der Median von 3 zeigt an, dass die Hälfte der Hochschulen ein gleiches oder niedrigeres Rating hat, während die andere Hälfte ein gleiches oder höheres Rating hat. Der Mittelwert, der mit 3,087 etwas höher als der Median ist, spiegelt den Durchschnitt der Bewertungen der Hochschulen im Datensatz wider und gibt einen Überblick über die Verteilung der Hochschulbewertungen.

2.1.5 Statement of Purpose

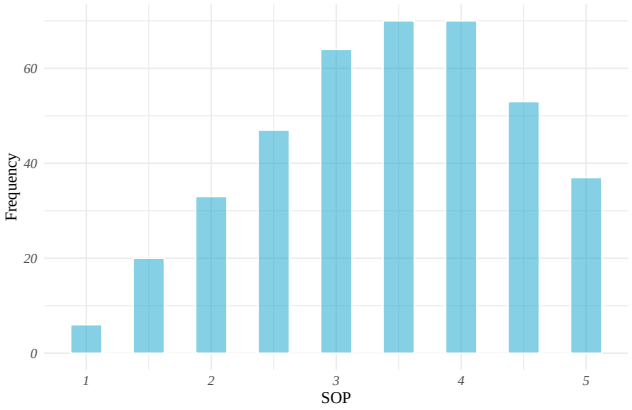


Figure 5: SOP Histogramm

Die Variable “SOP” (Statement of Purpose) reicht von 1 bis 5. Der Median von 3,5 zeigt an, dass die Hälfte der SOPs gleich oder niedriger bewertet wird, während die andere Hälfte gleich oder höher bewertet wird. Der Mittelwert, der mit 3,4 etwas niedriger als der Median ist, spiegelt den Durchschnitt der Bewertungen wider, die den SOPs im Datensatz zugewiesen wurden, und liefert eine Gesamtschätzung des wahrgenommenen Qualitätsniveaus der bewerteten Absichtserklärungen.

2.1.6 Letter of Recommendation

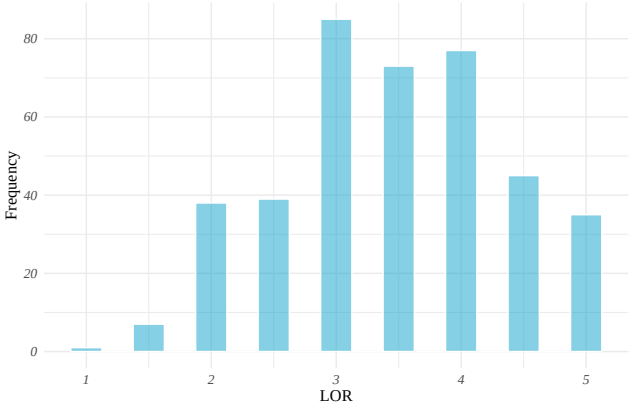


Figure 6: LOR Histogramm

Für die Variable “LOR” (Letter of Recommendation) geben die Daten Aufschluss über die wahrgenommene Qualität der Empfehlungsschreiben, mit einer Mindestbewertung von 1 und einer Höchstbewertung von 5. Der Median von 3,5 zeigt an, dass die Hälfte der Empfehlungsschreiben gleich oder niedriger bewertet wird, während die andere Hälfte gleich oder höher bewertet wird. Der Mittelwert, der mit 4 höher ist als der Median, spiegelt die durchschnittliche Qualität der Empfehlungsschreiben im Datensatz wider und gibt einen Überblick über das Niveau der abgegebenen Empfehlungen.

2.1.7 Research

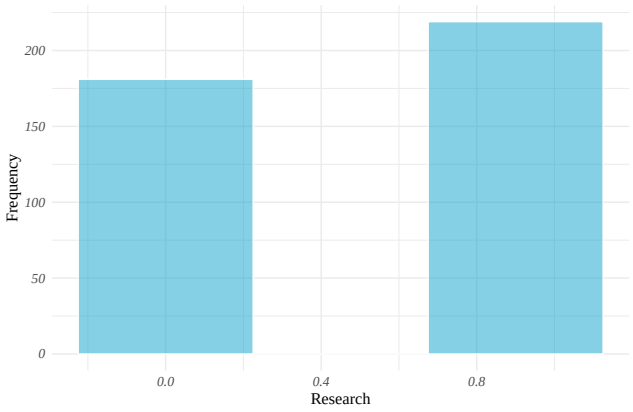


Figure 7: Research

Für die Variable “Forschung,” die angibt, ob Personen Forschung betrieben haben oder nicht, zeigen die Daten eine Verteilung, bei der der Median bei 1 liegt, was bedeutet, dass mindestens die Hälfte der Personen Forschung betrieben hat. Der Mittelwert, der mit 0,5475 niedriger als der Median ist, zeigt einen gewichteten Durchschnitt des Vorhandenseins von Forschung im Datensatz an, wobei diejenigen berücksichtigt werden, die keine Forschung betrieben haben. Der Minimalwert von 0 und der Maximalwert von 1 spiegeln den binären Charakter dieser Variable wider, wobei 0 für das Fehlen und 1 für das Vorhandensein von Forschung steht.

3 Methoden unserer Analyse

In diesem Abschnitt wollen wir die Methoden vorstellen die wir nutzen werden.

3.1 Lineare Regression

Die lineare Regressionsmethode ist ein Modell, das zur Annäherung an die Abhängigkeitsbeziehung zwischen einer abhängigen und einer unabhängigen Variable verwendet wird. Mit dieser Methode wird versucht, die Summe der quadratischen Fehler zwischen den beobachteten Werten und den durch das Modell vorhergesagten Werten zu minimieren. Sie wird in der Statistik und in verschiedenen Bereichen wie der Ökonometrie, dem Ingenieurwesen und der Datenwissenschaft häufig verwendet, um Vorhersagen zu treffen und Beziehungen zwischen Variablen zu analysieren. Es ist jedoch wichtig, ihre Annahmen wie Linearität und Fehlerunabhängigkeit zu berücksichtigen, um die Ergebnisse richtig zu interpretieren und Verzerrungen in den Schlussfolgerungen zu vermeiden. Darüber hinaus gibt es Varianten wie die multiple lineare Regression, die es ermöglicht, die Beziehung zwischen einer abhängigen Variablen und mehreren unabhängigen Variablen zu modellieren, sowie Regularisierungstechniken, um eine Überanpassung des Modells an die Trainingsdaten zu behandeln.

In unserer Analyse nutzen wir die Lineare Regression um eine Korrelation zwischen den Variablen und der 'Chance of Admit' Variable darzustellen. Hier sind alle Variablen, die nicht 'Chance of Admit' sind unsere unabhängigen Variablen und 'Chance of Admit' die abhängige Variable. Die Lineare Regression ist am besten geignet für diese Untersuchung und wird somit in dem nächsten Abschnitt genutzt.

3.2 t-Test

In unserer Analyse nutzen wir die Lineare Regression um eine Korrelation zwischen den Variablen und der 'Chance of Admit' Variable darzustellen. Hier sind alle Variablen, die nicht 'Chance of Admit' sind unsere unabhängigen Variablen und 'Chance of Admit' die abhängige Variable. Die Lineare Regression ist am besten geignet für diese Untersuchung und wird somit in dem nächsten Abschnitt genutzt.

Innerhalb unseres Bachelor Studiums, kommen wir eher selten auf die Möglichkeit Forschung durchzuführen, mit Ausnahme von diesem Projekt. Aufgrunddessen, nutzen wir einen t-Test um den Einfluss von Forschung (Research) in einer Bewerbung auf die 'Chance of Admit' Variable zu messen und zu erkennen ob diese Eigenschaft stochastisch relevant ist. Durch weitere Recherche, erkennen wir die Möglichkeit auf einen Welch t-Test, (Kubinger, Rasch, and Moder 2009) und (Rolles 2024). Wir führen so einen Test durch, da wir nicht Aussagen können ob die Varianzen der Population mit Forschungserfahrung und ohne gleich sind.

4 Auswertung unserer Methoden

Fassen wir zusammen, was wir bis jetzt gesehen haben:

Es gibt viele verschiedene Faktoren die die Wahrscheinlichkeit einer Zusage für einen Master oder ein PhD Programm beeinflussen können. Viele von diesen Faktoren sind schwer zu messen, wie die Aussagekraft eines 'Statement of Purpose' oder einem 'Letter of Recommendation.'

Wir haben auch erkannt, dass Universitäten, die ein höheres Ranking erhalten, auch höhere Durchschnittsresultate in Bewerbungen erhalten. Kurzgefasst, sind hier die Universitäten mit Rating 1 und 5, mit den Durchschnitts GRE-Scores im Vergleich:

```
dataset1 %>%
  filter(University.Rating %in% c(1)) %>%
  summarise(mean = mean(GRE.Score))
```

```
## # A tibble: 1 x 1
##    mean
##   <dbl>
## 1  303.
```

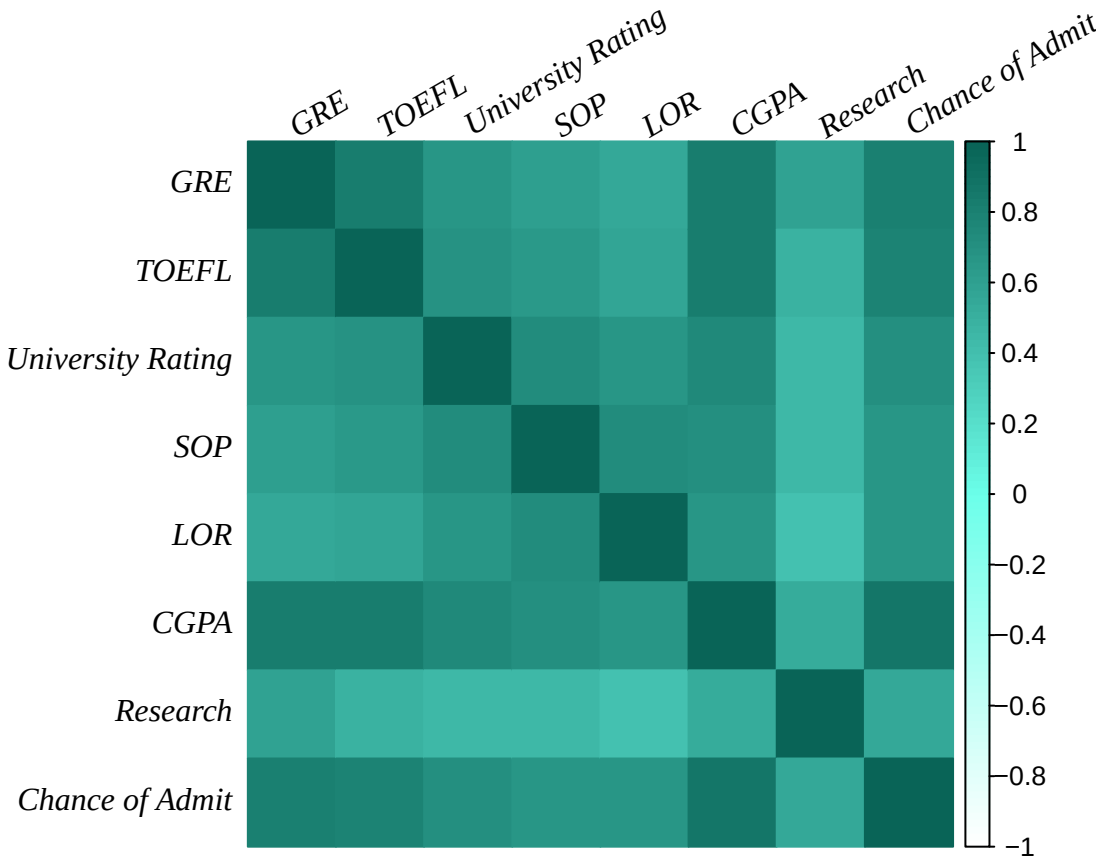
```
dataset1 %>%
  filter(University.Rating %in% c(5)) %>%
  summarise(mean = mean(GRE.Score))
```

```
## # A tibble: 1 x 1
##    mean
##   <dbl>
## 1  328.
```

4.1 Korrelation der Variablen

Wir wollen visualisieren welche Variablen stärker zusammenhängen. Das machen durch eine Färbung abhängig von dem Wert der Korrelation.

```
## corrplot 0.92 loaded
```



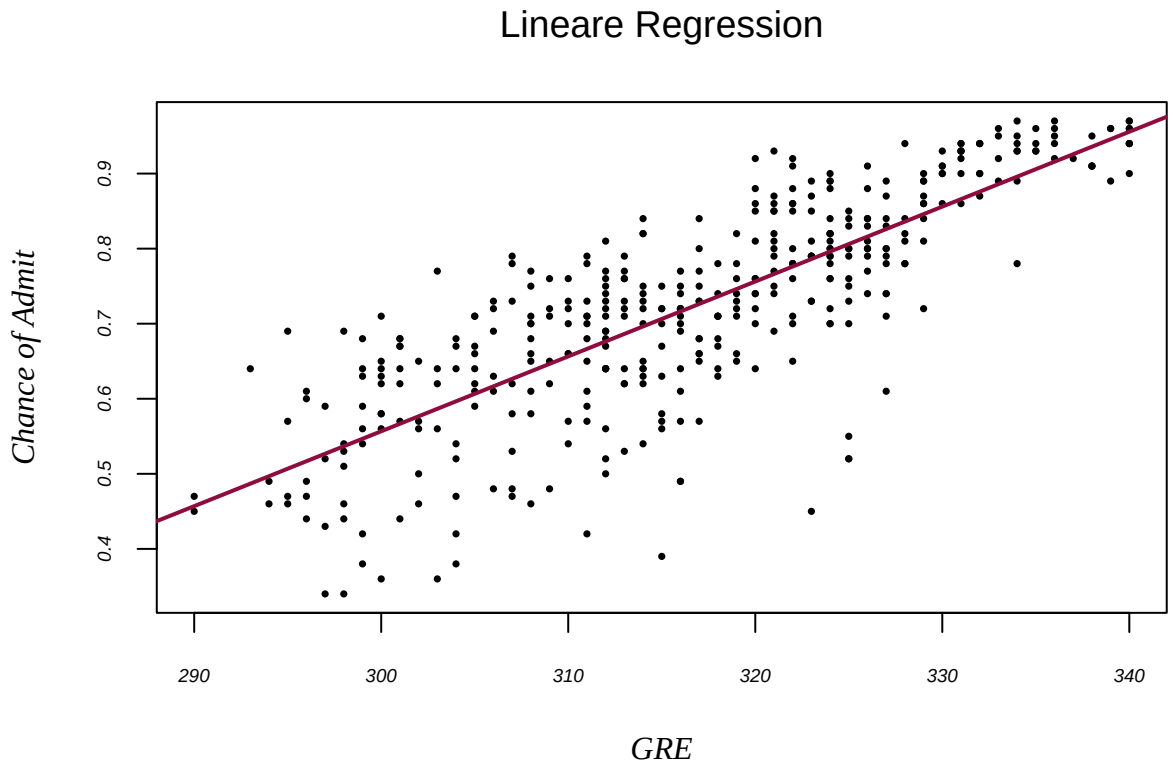
Hier können wir erkennen, dass es einen positiven Zusammenhang zwischen den Variablen: CGPA und GRE Score und der Variable Chance of Admit. Weiter wollen wir die Korrelationen zwischen den anderen Variablen visualisieren.

4.2 Lineare Regression: GRE Score v Chance of Admit

Die Graduate Record Examination (GRE) ist für Bewerber, die sich um die Zulassung zu Masterstudiengängen verschiedener Fachrichtungen bewerben, von großer Bedeutung. Als standardisierter Test dient das GRE als allgemeiner Maßstab, um die Eignung und Vorbereitung der Bewerber für ein Studium auf Graduiertenebene zu bewerten. Universitäten verwenden GRE-Ergebnisse oft als eines von mehreren Kriterien, um das akademische Potenzial und die Bereitschaft der Bewerber für ein weiterführendes Studium zu beurteilen. In vielen Fällen, insbesondere in Bereichen wie den Ingenieur-, Natur- und Sozialwissenschaften, können gute GRE-Ergebnisse ein wichtiger Faktor sein, um Bewerber in wettbewerbsorientierten Zulassungsverfahren zu unterscheiden. Darüber hinaus können GRE-Ergebnisse den Zulassungsausschüssen wertvolle Einblicke in das quantitative Denken, das sprachliche Denken und die analytischen Schreibfähigkeiten der Bewerber geben, die für den Erfolg in den Kursen und der Forschung auf Graduiertenebene unerlässlich sind. Der GRE ist zwar nur eine Komponente des Bewerbungspakets, seine Bedeutung liegt jedoch in seiner Fähigkeit, einen standardisierten Maßstab für die Bewertung der akademischen Fähigkeiten der Bewerber und ihres potenziellen Beitrags zum gewählten Studienfach zu bieten. Wir werden versuchen, einen Zusammenhang zwischen der Chance auf eine Zulassung und dem GRE-Score zu finden, um zu zeigen, dass dies der Fall ist. Dafür verwenden wir lineare Regression:

```
dataset1 <- read.csv("../Daten/adm_data1.csv")
modell <- lm(`Chance.of.Admit` ~ `GRE.Score`, data = dataset1)

par(font.main = 8, cex.axis = 0.6, font.axis = 3)

plot(dataset1$`GRE.Score`, dataset1$`Chance.of.Admit`,
     main = "Lineare Regression",
```

```
     xlab = "",
     ylab = "",
     col = "black",
     pch = 16,
     cex = 0.5,
     )

abline(modell, col = "#900C3F", lwd = 2)

mtext("GRE", side = 1, line = 3, col = "black", cex = 1, font=3, family = "serif")

mtext("Chance of Admit", side = 2, line = 3, col = "black", cex = 1, font=3, family = "serif")
```




```
dataset1 <- read.csv("../Daten/adm_data1.csv")
modell <- lm(`Chance.of.Admit` ~ `GRE.Score`, data = dataset1)

x <-  dataset1$`GRE.Score`
y <- dataset1$`Chance.of.Admit`

linearmodell <- lm(x ~ y, data = dataset1)

mean_x <- mean(x)
mean_y <- mean(y)

var_x <- var(x)
var_y <- var(y)

cov_xy <- cov(x, y)
```

```r
r_xy <- cor(x ,y , method ="pearson")

lin_coef <- coef(linearmodell)

residuen <- residuals(linearmodell)

sum_res_squared <- sum(residuen^2)
```

```
## [1] "Empirische Mittelwert der GRE: 316.8075"
```

```
## [1] "Empirische Mittelwert der Aufnahmechance: 0.72435"
```

```
## [1] "Empirische Varianz der GRE: 131.644555137845"
```

```
## [1] "Empirische Varianz der Aufnahmechance: 0.0203374210526316"
```

```
## [1] "Empirischer Kovarianz: 1.31327055137845"
```

```
## [1] "Empirischer Korrelationskoeffizient: 0.80261045959035"
```

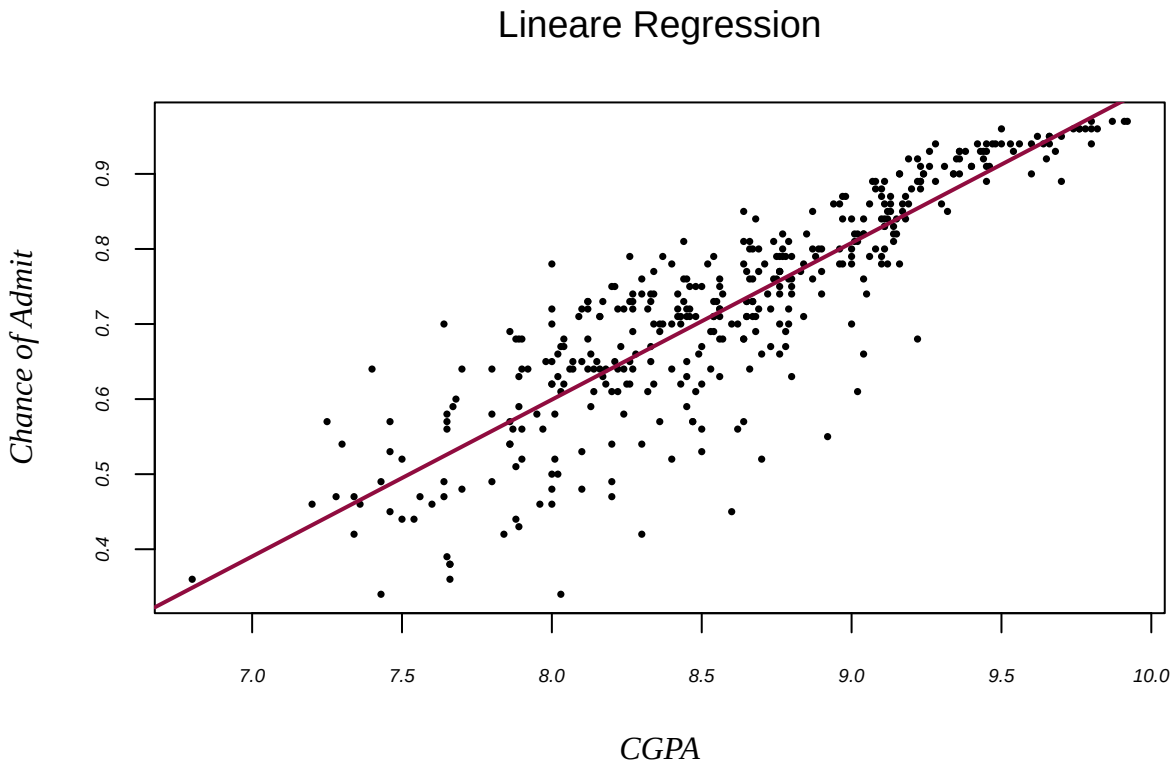
```
## [1] "Koefizienten der Lineare Regression: 270.033254562037"
## [2] "Koefizienten der Lineare Regression: 64.5740946199526"
```

```
## [1] "Summe der Residuen zum quadrat: 18689.6780183227"
```

Basierend auf der linearen Regression zwischen den Graduate Record Examination (GRE) Ergebnissen und der Chance auf Zulassung zeigt sich ein positiver Zusammenhang zwischen diesen beiden Variablen. Der empirische Korrelationskoeffizient von 0.80 deutet auf eine starke lineare Beziehung hin, was bedeutet, dass höhere GRE-Werte tendenziell mit einer höheren Zulassungschance korrelieren. Die Parameter der linearen Regression zeigen, dass jeder zusätzliche Punkt im GRE-Ergebnis im Durchschnitt die Chance auf Zulassung um etwa 0.01 erhöht. Die Residuen, die quadratische Summe der Abweichungen zwischen den beobachteten und vorhergesagten Werten, sind mit 18689.68 hoch, was darauf hinweisen könnte, dass das Modell möglicherweise nicht perfekt ist und noch Raum für Verbesserungen bietet.

4.3 Lineare Regression: CGPA v Chance of Admit

Der kumulative Notendurchschnitt (CGPA) spielt vermutlich eine entscheidende Rolle bei der Bewertung von Bewerbern für Hochschulzulassungen. Er fungiert als Maßstab für die akademische Leistung und das Engagement eines Bewerbers während seines Studiums. Ein hoher CGPA deutet auf konsistente Exzellenz in den Kursen hin, was wiederum die Fähigkeit des Bewerbers demonstriert, akademische Anforderungen zu erfüllen und erfolgreich im Studium zu sein. Im Hinblick auf die Zulassungsentscheidungen werden wir versuchen, eine Korrelation zwischen dem akademischen Durchschnitt und den Zulassungschancen aufzuzeigen, die die Bedeutung des akademischen Hintergrunds für die Entscheidungsfindung des Zulassungsausschusses verdeutlichen wurde. Ein solides CGPA kann ein Bewerber sein, der sich durch Beharrlichkeit, Disziplin und Fähigkeit zur Leistung auszeichnet, was ihn zu einem attraktiven Kandidaten für eine Zulassung macht. Dies wird, wie beim GRE, durch eine lineare Regressionsanalyse erreicht. Wir nutzen den gleichen Code wie oben und lassen ihn deshalb an dieser Stelle weg.




```
## [1] "Empirische Mittelwert der CGPA: 8.598925"
## [1] "Empirische Mittelwert der Aufnahmechance: 0.72435"
## [1] "Empirische Varianz der CGPA: 0.355594079573935"
## [1] "Empirische Varianz der Aufnahmechance: 0.0203374210526316"
## [1] "Empirischer Kovarianz: 0.0742648383458647"
## [1] "Empirischer Korrelationskoeffizient: 0.8732890993553"
## [1] "Koefizienten der Lineare Regression: 5.95386319414893"
## [2] "Koefizienten der Lineare Regression: 3.65163499116596"
## [1] "Summe der Residuen zum quadrat: 33.6779929054999"
```

Die Analyse zwischen dem kumulativen Notendurchschnitt (CGPA) und der Chance auf Zulassung zeigt einen noch stärkeren Zusammenhang als bei der vorherigen Analyse mit GRE. Der hohe empirische Korrelationskoeffizient von 0.87 deutet auf eine sehr starke positive lineare Beziehung hin, was bedeutet, dass höhere CGPA-Werte tendenziell mit einer höheren Zulassungschance korrelieren. Die Parameter der linearen Regression zeigen, dass jeder zusätzliche Punkt im CGPA im Durchschnitt die Chance auf Zulassung um etwa 0.01 erhöht, ähnlich wie bei der vorherigen Analyse. Die Residuen, die quadratische Summe der Abweichungen zwischen den beobachteten und vorhergesagten Werten, sind mit 33.68 relativ niedrig, was darauf hindeutet, dass das Modell die Daten gut erklären kann und nur eine geringe Streuung aufweist. Dies deutet darauf hin, dass das Modell eine gute Passform für die Daten aufweist und die Vorhersagen akkurat sind.

4.4 t-Test

Um den Einfluss der Forschungserfahrung (Research) auf die Zulassungschancen (Chance.of.Admit) zu untersuchen, können wir einen t-Test für unabhängige Stichproben in R verwenden. Dieser Test vergleicht die Mittelwerte der Zulassungschancen zwischen zwei Gruppen: Kandidaten mit Forschungserfahrung (Research = 1) und Kandidaten ohne Forschungserfahrung (Research = 0).

Ich werde nun die Daten in R laden und den entsprechenden t-Test durchführen. Lassen Sie uns die Ergebnisse anschauen und daraus Schlussfolgerungen ziehen. Wir nutzen den t-Test aus (Unwin 2013)

```
library(dplyr)
dataset1 <- read.csv("../Daten/adm_data1.csv")
# Trennung der Daten in zwei Gruppen basierend auf der Forschungserfahrung
group_with_research <- filter(dataset1, Research == 1)$`Chance.of.Admit`
group_without_research <- filter(dataset1, Research == 0)$`Chance.of.Admit`

# Durchführung eines t-Tests für unabhängige Stichproben
t_test_result <- t.test(group_with_research, group_without_research, var.equal = FALSE)

# Ausgabe des Testergebnisses
t_test_result
```

```
##
##  Welch Two Sample t-test
##
## data:  group_with_research and group_without_research
## t = 13.347, df = 392.98, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1349844 0.1816199
## sample estimates:
## mean of x mean of y
## 0.7959817 0.6376796
```

Der durchgeführte t-Test für unabhängige Stichproben ergab einen t-Wert von etwa 13.35 und einen sehr kleinen p-Wert (2.2e-16). Dieser p-Wert ist deutlich kleiner als das übliche Signifikanzniveau von 0.05, was darauf hindeutet, dass es einen statistisch signifikanten Unterschied in den Zulassungschancen zwischen den Gruppen mit und ohne Forschungserfahrung gibt.

Die Forschungserfahrung hat einen signifikanten Einfluss auf die Chance auf Zulassung. Kandidaten mit Forschungserfahrung haben statistisch signifikant höhere Chancen auf Zulassung im Vergleich zu Kandidaten ohne Forschungserfahrung.

Zusammenfassend können wir erkennen, dass binnen unseres Datensatzes eine Universität oder Hochshule die ein höheres Rating besitzt, auch qualitativ höhere Bewerbungen haben wird. Dies kann man sich durch viele verschiedene Erklärungen und Begründungen erarbeiten. Eine Möglichkeit ist, dass man intuitiv meint bessere Noten für eine höher angesehene Universität zu brauchen. Eine andere Perspektive ist, dass man sein gesamtes Bewerbungsportfolio stärken möchte für eine höher angesehene Universität und dadurch auch der Durchschnitt aller beobachteten Variablen steigt. Wir nehmen aus unserem Projekt zur Kenntnis, dass es keine eine Variable gibt die eine Aufnahme sichert. Es ist ein viel holistischer Ansatz und das “Gesamtpaket” wird betrachtet einer Bewerbung. Man könnte den Datensatz erweitern mit Perspektive auf überfachliche Arbeit oder spezifische Eigeninitiative, weiter könnten auch Interviews und die Performance innerhalb eines Interviews gemessen werden um den Datensatz zu erweitern.

Literatur

Blog, Magoosh. 2021. “GRE Score Percentiles.” https://magoosh.com/gre/gre-score-percentiles/.

Kubinger, Klaus D., Dieter Rasch, and Karl Moder. 2009. “Zur Legende Der Voraussetzungen Des t-Tests für Unabhängige Stichproben.” Doi: 10.1026/0033-3042.60.1.26. *Psychologische Rundschau* 60 (1): 26–27. https://doi.org/10.1026/0033-3042.60.1.26.

Rolles, Silke. 2024. “Einführung in Die Wahrscheinlichkeits Theorie Und Statistik.”

Unwin, Antony. 2013. “Discovering Statistics Using r by Andy Field, Jeremy Miles, Zoë Field.” *International Statistical Review* 81 (April). https://doi.org/10.1111/insr.12011_21.