STATS 112 Final Project Report

Aidan Gresham (Senior), Brittney Trinh (Senior),
Edward Halim (Senior), Maneh Begi (Senior),
Min Jung Kim (Senior), and Yaquan Wang (Senior)

Group 8, Discussion 1B

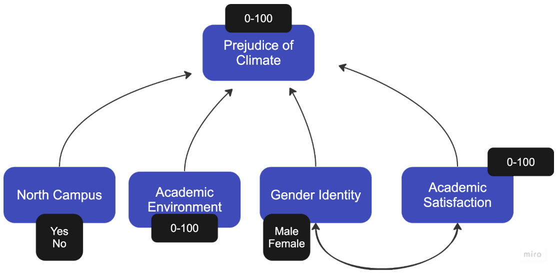
Contents

1 Abstract

This study focuses on understanding the dynamics of campus climate at the University of California, Los Angeles (UCLA), particularly exploring how various factors influence the perceived prejudice within the university environment. This study aims to predict the level of prejudice against various groups at UCLA based on several key variables: satisfaction with academic involvement and performance, the overall academic environment, enrollment in "North Campus" (non-STEM) programs, and the student's gender identity. Furthermore, our research will examine the interaction effect between academic satisfaction and gender identity on perceived prejudice. Utilizing the Multiple Linear Regression model, this study seeks to provide insights into how these factors collectively shape the respectfulness and inclusivity of the UCLA campus climate, offering a comprehensive view of the academic experience for undergraduate students. From our study, we can conclude that there is a statistically significant association between the prejudice of the climate against various groups at UCLA and the satisfaction with academic involvement and performance at UCLA, the academic environment at UCLA, whether a student is enrolled in a traditionally "North Campus" (non-STEM) program, the student's gender identity, and the interaction between academic satisfaction and the student's gender identity. Some potential shortcomings of the study are missing values, not using many predictors, having many leverage points, and more. Some recommendations for further research include applying transformations to reduce the amount of leverage points, splitting the data set into training and testing data in order to avoid overfitting, and applying cross-validation.

2 Statement of the Problem

1. Can the prejudice of the climate against various groups at UCLA be predicted from the satisfaction with academic involvement and performance at UCLA, the academic environment at UCLA, whether a student is enrolled in a traditionally "North Campus" (non-STEM) program, the student's gender identity, and the interaction between academic satisfaction and the student's gender identity?



2. How powerful is the model in terms of explaining the outcome? (The

coefficient of determination R^2)

3. What can we conclude from the model? Provide some interpretations in context.

3 Data Description / Variables

3.1 Variables

Variables	Variable Types	Measurement
Prejudice of the Climate (prejudiceenvp)	Outcome	Numerical (0-100)
Academic Satisfaction (academicsp)	Predictor	Numerical (0-100)
Academic Environment (academicenvp)	Predictor	Numerical (0-100)
North Campus (NorthCampus)	Predictor	Categorical (Yes or No)
Gender Identity (new_sex)	Predictor	Categorical (Female or Male)

3.2 Data Description

1. prejudiceenvp: Respectful climate for various groups at UCLA; 0-100 scale, high scores indicate high degrees of respect for many groups.

2. academicsp: Satisfaction with academic involvement and performance at UCLA; 0-100 scale, high scores indicate high satisfaction.

3. academicenvp: I feel valued/respected by peers and faculty, have academic opportunities, feel free to express self, etc.; 0-100 scale, high scores indicate strong degree of agreement.

4. NorthCampus: Whether respondent is enrolled in a traditionally "North Campus" (non-STEM) program; binary.

5. new_sex: Student's gender identity; only binary identifications supplied. Note: The 14 observations that reported gender identity as "Unknown/Other" were removed from the data set before data analysis.

4 Exploratory Data Analysis (EDA)

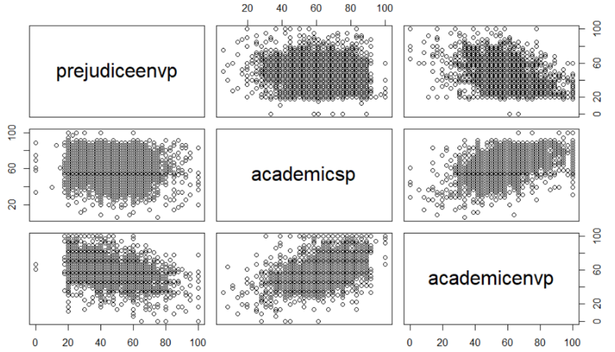
4.1 Means and Standard Deviations of Numerical Variables

	prejudiceenvp	academicsp	academicenvp
Mean	43.95678	63.97168	61.01964
Standard Deviation	16.24346	14.15091	16.23058

4.2 Summary Statistics of Numerical Variables

	prejudiceenvp	academicsp	academicenvp
Minimum	0.00	5.556	0.00
1st quartile	32.50	55.556	50.00
Median	42.50	63.889	61.11
3rd quartile	57.50	72.222	66.67
Maximum	100.00	100.000	100.00

4.3 Scatterplot Matrix and Correlation Matrix



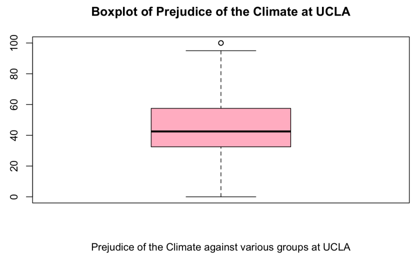
- Here is a scatterplot matrix for our 3 numerical variables: prejudice of the climate against various groups at UCLA, academic satisfaction at UCLA, and academic environment at UCLA. As we can see, all numerical variables have relatively low correlation with each other.

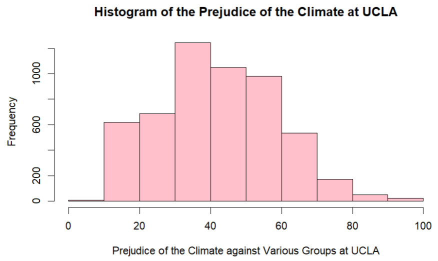
	prejudiceenvp	academicsp	academicenvp
prejudiceenvp	1.00000000	-0.24139632	0.01349082
academicsp	-0.24139632	1.00000000	0.04961439
academicenvp	0.01349082	0.04961439	1.00000000

- The correlation matrix confirms that all numerical variables have relatively low correlation with each other.

4.4 Boxplot and Histogram of Numerical Variables

- **Boxplot and Histogram of the Prejudice of the Climate at UCLA (Prejudiceenvp)**

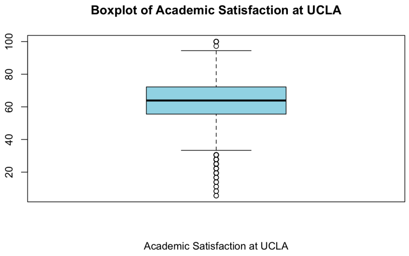


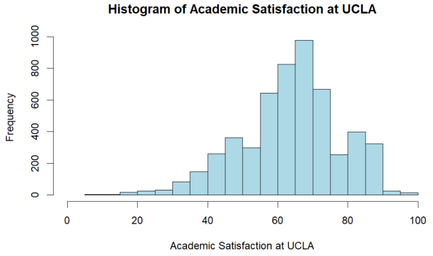


Boxplot Analysis: The boxplot summarizes the distribution of the prejudice of the climate against various groups at UCLA. The line in the middle of the box represents the data's median value, which is 42.50. This indicates that half of the participants rated the Prejudice of the Climate below this value and half above. The box itself shows the middle 50% of the data, stretching from 32.5 to 57.50. This range indicates where the central half of the Prejudice of the Climate data lies. The whiskers reach the smallest and largest values within 1.5 times the Interquartile Range (IQR) from the lower and upper quantiles. Anything beyond the whiskers is considered potential outliers. It is shown that there is 1 outlier.

Histogram Analysis: The histogram shows the frequency distribution of the prejudice of the climate against various groups at UCLA. The distribution is slightly right-skewed. However, the distribution is approximately symmetric since the skewness is around 0.29, within (-0.5, 0.5). This indicates that the data is generally symmetric but with a slight tendency towards lower prejudice scores. The skewness within the -0.5 to 0.5 range reinforces that the distribution is not far from symmetric but with a noticeable minority of higher prejudice scores. Furthermore, the tallest bar represents the most common range of values, approximately 30 to 40. The y-axis represents the frequency of observations. The peak frequency is around 1200, which occurs in the 30-40 range.

- **Boxplot and Histogram of Academic Satisfaction (Academicsp)**

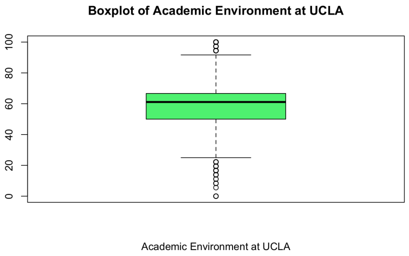


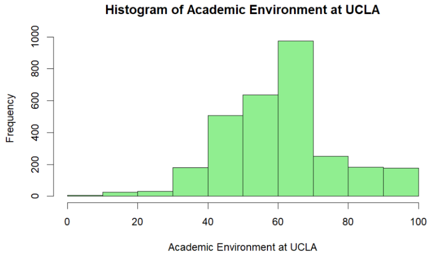


Boxplot Analysis: The boxplot summarizes the distribution of Academic Satisfaction at UCLA. The line in the middle of the box represents the data's median value, approximately 63.89. This indicates that half of the participants rated their Academic Satisfaction below this value and half above. The box shows the middle 50% of the data, stretching from 55.56 to 72.22. This range indicates where the central half of the participants' rates of academic satisfaction lies. It is shown that there are 12 outliers.

Histogram Analysis: The histogram shows the frequency distribution of Academic Satisfaction at UCLA. The distribution is slightly left skewed. However, since the skewness is around -0.44, within (-0.5, 0.5), the distribution is approximately symmetric. This indicates that the Academic Satisfaction data is generally symmetric but with a slight tendency towards higher satisfaction scores. The skewness within the -0.5 to 0.5 range reinforces the idea that the distribution is not far from symmetric but with a noticeable minority of lower satisfaction scores. Furthermore, the tallest bar represents the most common range of values. The y-axis represents the frequency of observations, with a peak frequency of around 1000.

- **Boxplot and Histogram of Academic Environment (academicenvp)**


Boxplot of Academic Environment at UCLA

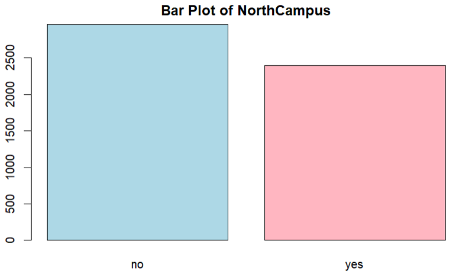

Histogram of Academic Environment at UCLA

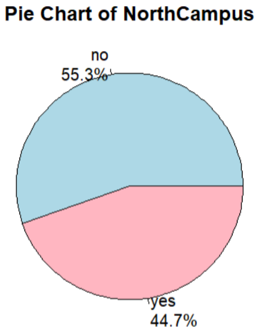
Boxplot Analysis: The boxplot summarizes the distribution of the Academic Environment at UCLA. The line in the middle of the box represents the median value of the data, which is 61.11. This indicates that half of the participants rated Academic Environment below this value, and half are above. The box itself shows the middle 50% of the data, stretching from 50 to 66.67. This range indicates where the central half of the Academic Environment data lies. It is shown that there are 11 outliers, which indicates some significantly different values in the data set.

Histogram Analysis: The histogram shows the frequency distribution of Academic Environment at UCLA. The distribution appears slightly right-skewed, as the right tail is longer than the left one. This skewness is also reflected in the boxplot with more outliers on the upper end. However, it is important to note that the skewness is not too extreme. Since the skewness is around 0.17, within (-0.5, 0.5), the distribution is approximately symmetric. This indicates that the data is generally symmetric but with a slight tendency towards lower academic environment ratings. The skewness within the -0.5 to 0.5 range reinforces the idea that the distribution is not far from symmetric but with a noticeable minority of higher academic environment ratings. Furthermore, the tallest bar represents the most common range of values, approximately 60 to 70. The y-axis represents the frequency of observations. The peak frequency is around 1000, which occurs in the 60-70 range.

4.5 Barplots and Table of Relative Frequencies for Categorical Predictors

- **Barplot and Pie Chart of NorthCampus**



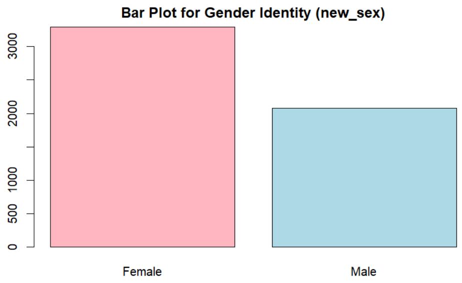


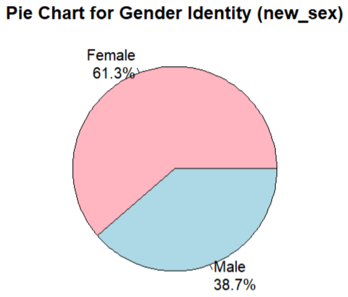
- **Table of Relative Frequency for North Campus**

Yes	No
0.4474323	0.5525677

Analysis: The bar plot comparing students in North and South Campus programs illustrates the disparity in the number of individuals enrolled in each. Specifically, there are 2396 participants enrolled in North Campus programs and 2959 participants enrolled in South Campus programs. When observing the pie chart and relative frequency table for North Campus, it reveals that 44.74% of participants are enrolled in North Campus programs while 55.26% are enrolled in South Campus programs. Both show that more students from South Campus participate in campus climate research, but it is not unbalanced.

- **Barplot and Pie Chart for Gender Identity (new_sex)**





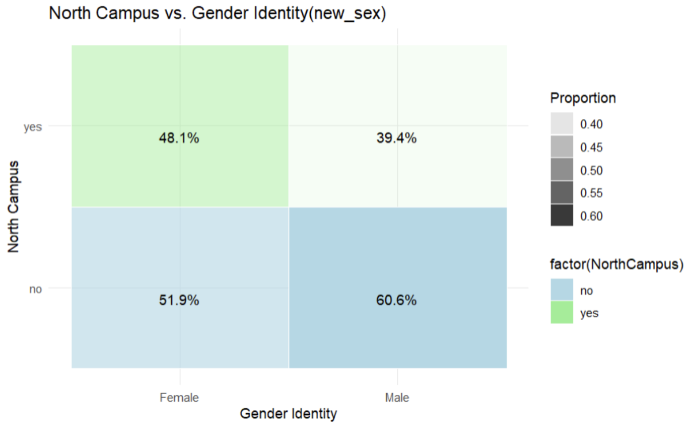
- **Table of Relative Frequency for Gender Identity**

Analysis: The bar plot comparing female and male gender identities illustrates the disparity in the number of individuals in each gender category.

Female	Male
0.6132638	0.3867362

Specifically, there are 3292 female participants and 2076 male participants. Observing the pie chart and relative frequency table for Gender Identity reveals that approximately 61.33% of participants are female while approximately 38.67% are male. Both show that more participants of the campus climate study are female, but it is not majorly imbalanced.

- **Table of Relative Frequency for North Campus vs. Gender Identity**



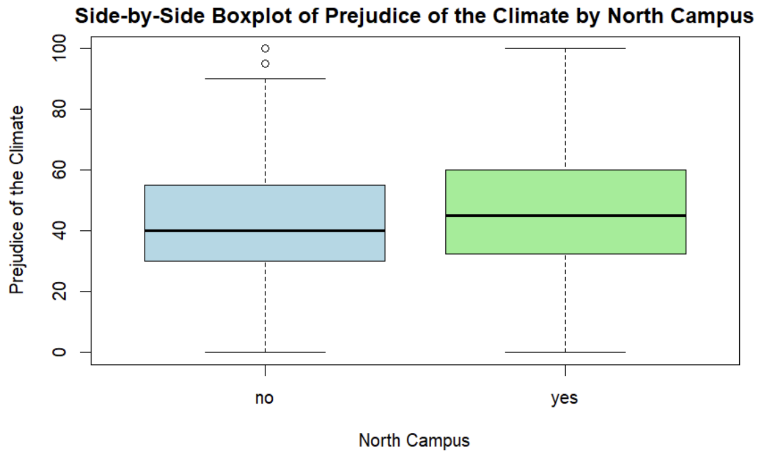
- Table of Frequency for North Campus vs. Gender Identity

	Female	Male
North Campus	0.4812671	0.3938224
South Campus	0.5187329	0.6061776

Analysis: As indicated by the proportion plot and relative frequency table for North Campus versus Gender Identity, it is apparent that 60.6% of male students are enrolled in South Campus programs, while 39.4% of male students are enrolled in North Campus programs. Meanwhile, 51.9% of female students are enrolled in South Campus programs and 48.1% of female students are enrolled in North Campus programs. The proportion of male participants in South Campus majors is higher than that of female participants. The proportion of female participants in North Campus majors is higher than that of male participants. In addition, the proportion difference between females enrolled in North and South Campus programs appears to be smaller than that of males.

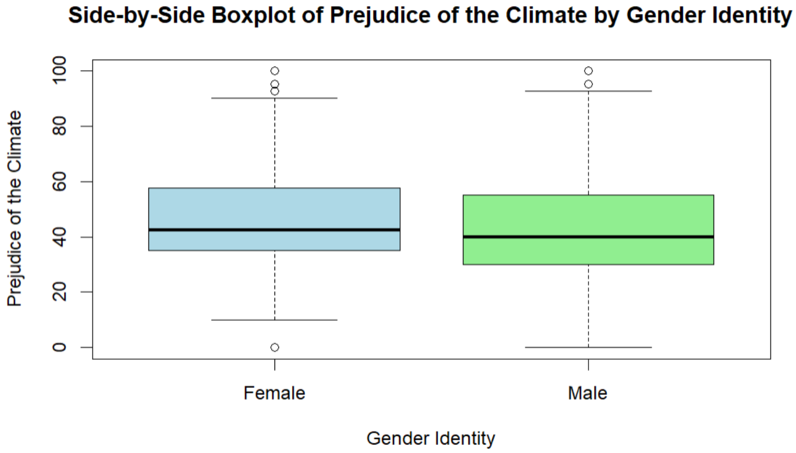
4.6 Boxplots Comparison for Numerical and Categorical Variable

- **Side by Side Boxplot: Prejudice of the Climate (prejudiceenvp) v. North Campus**



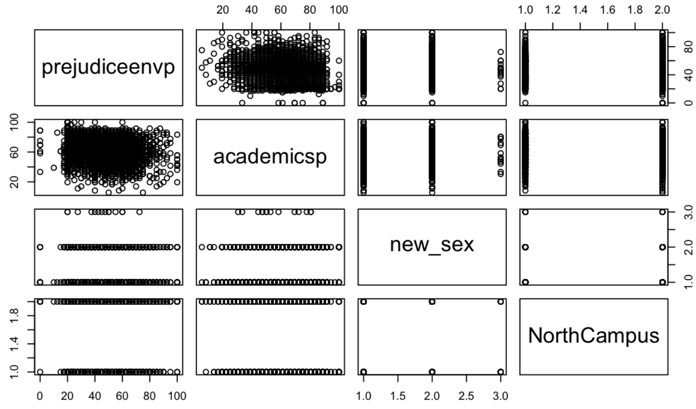
The boxplot illustrates the distribution of the scores of perceived prejudice against various groups at UCLA among participants from North and South Campus programs. Observing the plot, it is evident that the 1st quartile, median, and 3rd quartile of prejudice scores for South Campus students are lower than those of their North Campus counterparts. Notably, both South and North Campus majors display scores ranging from 0 to 100. For South Campus majors, the 1st quartile, median, and 3rd quartile scores stand at 30, 40, and 55, respectively. In contrast, North Campus students showcase slightly higher scores, with corresponding quartiles of 32.5, 45, and 60.

- **Side by Side Boxplot: Prejudice of the Climate (prejudiceenvp) vs. Gender Identity (new_sex)**



The boxplot illustrates the distribution of scores of perceived prejudice against various groups at UCLA among participants who are male or female. According to the plot, we can see that the 1st quartile, median, and 3rd quartile of prejudice scores for male students are lower than those of female students. Both female and male students have scores ranging from 0 to 100. For male students, the 1st quartile, median, and 3rd quartile prejudice scores are at 30, 40, and 55, respectively. In contrast, female students have slightly higher scores, with corresponding quartiles of 35, 42.5, and 57.5.

4.7 Checking For Multicollinearity



- There appear to be no significant patterns in the numerical scatterplots, so we can assume that the variables are not highly correlated to one another.
- For categorical variables, it is hard to say whether they are correlated or not since there are no distinct patterns generated.

4.8 Correlation Coefficient

A correlation coefficient is used to calculate the strength and direction of a linear relationship between variables. It aids in understanding associations and dependencies between different factors.

Based on the output of the cor() function between 2 numerical variables, which in our case are prejudice of the climate against various groups at UCLA and academic satisfaction at UCLA have an output of -0.240684. This suggests a weak negative linear relationship between the variables. While it provides some insight into the association between the variables, it is not a strong or decisive indication of their relationship.

Since we could not calculate the correlation coefficient of categorical variables using the cor() function, we could still determine their value by using Cramer V. The output shows that the correlation value between gender identity and North Campus is 0.06043.

It is not possible to calculate the correlation coefficient between a numerical variable and a categorical variable. However, we can use the Kruskal-Wallis rank sum test. This test does not really show the value of the correlation coefficient, but from this test, it can show how significant the correlation between one numerical and one categorical variable is based on the p-value that is shown. From the output above, we can conclude that the correlation between each numerical and categorical variables that we used for this model is statistically significant since all of the p-values from the given output are less than 0.05.

5 Multiple Linear Regression Model

5.1 Multiple Linear Regression Overview

The utilization of multiple linear regression within this project is instrumental in comprehending the dynamics influencing the overall climate's respectfulness for diverse groups at UCLA. This statistical method serves as a robust analytical tool for exploring the relationships between various predictors and the level of respect observed within the campus environment.

Linear regression, at its core, is a statistical technique employed to model the association between a dependent variable (in this case, the perceived prejudice of the climate against various groups at UCLA) and several independent variables (predictors). By employing a linear equation, this method aims to elucidate how changes in these predictors are linked to variations in the dependent variable.

In our study, numerous numerical and categorical predictors have been identified from the data on UCLA campus climate data set, ranging from individual experiences such as satisfaction with academic involvement and instances of exclusionary behavior to broader perceptions of friendliness, prejudice, and welcoming environments. Additionally, demographic characteristics, including ethnicity, family income, first-generation college status, gender identity, disability status, sexual orientation, political views, and enrollment specifics, serve as categorical predictors.

The multiple linear regression model allows us to assess how these diverse predictors collectively influence the perceived prejudice of the climate against various groups at UCLA. By fitting a linear equation to the data, the model aims to determine the relationships between these predictors and the level of respect observed across various groups on campus.

This approach will enable us to identify which predictors significantly impact the perceived prejudice of the climate against various groups and ascertain the extent of their influence. Furthermore, the resulting regression model may offer predictive capabilities, providing insights into how changes in these predictors might affect the the perceived prejudice of the climate for different groups within the UCLA campus environment.

By leveraging multiple linear regression, this study seeks to unravel the intricate interplay between the identified predictors and the overall climate's respectfulness, ultimately offering valuable insights that could inform strategies to enhance inclusivity and respect for diverse groups within the academic setting.

5.2 Multiple Linear Regression (MLR) Model

Below is a multiple linear regression model with the response variable prejudice of the climate against various groups at UCLA, and predictors academic satisfaction, academic environment, North Campus, and gender identity.

```
model <- lm(prejudiceenvp ~ academicsp + academicenvp +
NorthCampus + new_sex, data = cc_new)
summary(model)

```

5.3 Summary of Multiple Linear Regression

Predictors	Estimate	Std. Error	T-value	Pr($>|t|$)
(Intercept)	76.34766	1.34678	56.689	$<$ 2e-16 ***
academicsp	-0.08578	0.02137	-4.014	6.13e-05 ***
academicenvp	-0.44735	0.1910	-23.423	$<$ 2e-16 ***
NorthCampus	3.35113	0.54990	6.094	1.24e-09 ***
new_sex	-2.02627	0.55358	-3.660	0.000256 ***

Adjusted R^2: 0.2327
Interpretation of Numerical Predictors in Context:

- Holding all else constant, a one unit increase in academic satisfaction results in a 0.08578 decrease in perception of prejudice of the climate against various groups at UCLA. There is a statistically significant relationship between academic satisfaction and prejudice of the climate against various groups at UCLA.

- Holding all else constant, a one unit increase in the academic environment at UCLA results in a 0.44735 decrease in perception of prejudice of the climate against various groups at UCLA.

Interpretation of Categorical Predictors in Context:

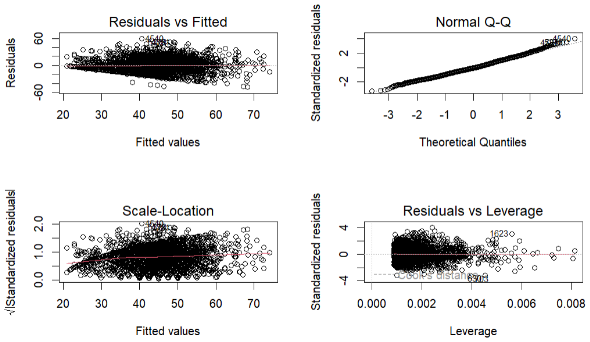
- The NorthCampus variable has 2 levels, where yes represents students enrolled in a traditionally “North Campus” (non-STEM) program, and no represents students not enrolled in a traditionally “North Campus” program. Students not enrolled in a traditionally “North Campus” program are the base.

  ∗ Holding all else constant, compared to those who are not enrolled in a traditionally “North Campus” program, students who are enrolled in a traditionally “North Campus” program report prejudice of the climate against various groups at UCLA approximately 3.35113 more.

- The new_sex variable has 2 levels, where Female represents students who reported female gender identity, and Male represents students who reported male identity. Students with a female gender identity are the base.

  ∗ Holding all else constant, compared to female students, male students report prejudice of the climate against various groups at UCLA approximately 2.02627 less.

5.4 Linear Model Assumptions



Assumptions of the Linear Model:

- **Independence:** We can assume that each observation is independent of each other, and that the errors are independent.

- **Linearity:** The Residuals vs Fitted Plot indicates no significant pattern with scattered points. Therefore, we can assume that there is a linear relationship.

- **Constant Variance:** The Residuals vs. Fitted and Scale-Location plots have a straight horizontal line across indicating that the variance is constant.

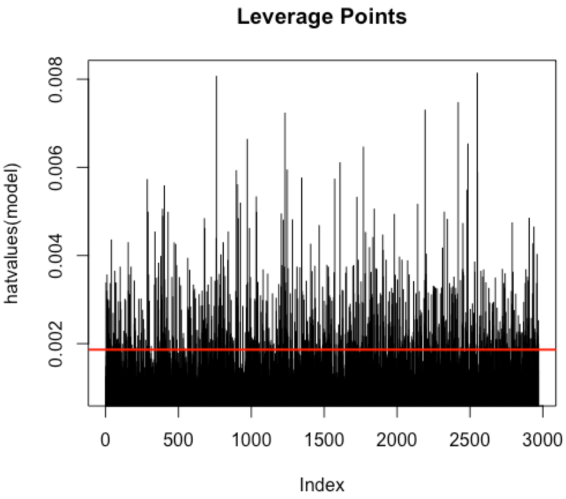
- **Normality:** The Normal Q-Q Plot demonstrates that the majority of the points are following a straight line, with few deviations towards the ends. This shows that the data follows an approximately normal distribution.

5.5 Multicollinearity Assumption using Variance Inflation Factor

Predictors	academicsp	academicenvp	NorthCampus	new_sex
VIF	1.317324	1.305796	1.016131	1.008942

To check for multicollinearity between the predictors, we calculate the Variance Inflation Factor (VIF). A VIF value greater than 5 indicates that the predictors may be too correlated with one another. Since all the VIF values are less than 5, we do not need to investigate multicollinearity further, as the associated regression coefficients are not poorly estimated due to multicollinearity.

5.6 Leverage Points



- There are n = 5368 total observations in our data set, and p = 4 predictors in our model, so we can calculate 2 * (p+1) / n = 0.001862891.
- Leverage Points for Multiple Linear Regression (MLR) are defined when leverage is greater than 0.001862891.
- Approximately 27.3% of the hatvalues indicate potential leverage points.

5.7 Interaction Effect

Below is a linear regression model with the response variable prejudice of the climate against various groups at UCLA and the predictors academic satisfaction, gender identity, and the interaction effect between both.

```
model <- lm(prejudiceenvp ~ academicsp * new_sex , data =cc_new)
summary(model)
```

Predictors	Estimate	Std. Error	T-value	Pr(> $\|t\|$)
(Intercept)	60.59328	1.26646	47.845	< 2e-16 ***
academicsp	-0.24765	0.01947	-12.722	< 2e-16 ***
new_sexMale	2.34669	2.04660	1.147	0.2516
academicsp:new_sexMale	-0.06799	0.03111	-2.185	0.0289 *

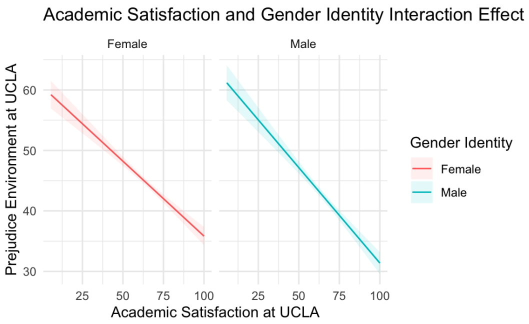
- Adjusted R-squared: 0.0625

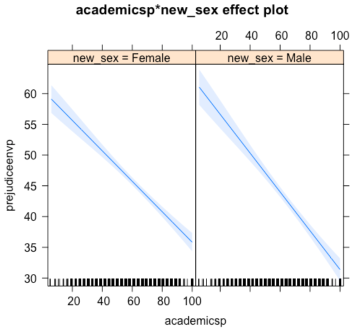
Interpretation of Interaction Effects

- There is a statistically significant association between academic satisfaction and the prejudice of the climate against various groups at UCLA.
- The interaction effect of academic satisfaction and gender identity is statistically significant, which suggests that the effect of gender identity being male as opposed to female on academic satisfaction is associated with the prejudice of the climate against various groups at UCLA.

5.8 Plot of Interaction Effect

- **Plot of Interaction Effect of Academic Satisfaction and Gender Identity**





Interpretation of Interaction Effect
The plot of interaction effect shows a decreasing trend for prejudice of the climate against various groups at UCLA as academic satisfaction increases, which

is consistent for both female and male students. However, male students have a steeper slope compared to female students. The fact that both slopes for females and males are negative does not negate the significance of the interaction. The difference in steepness of these slopes demonstrates that the interaction effect is significant. As prejudice of the climate against various groups at UCLA increases, academic satisfaction drops more for male than female participants. Therefore, the interaction effect of academic satisfaction and gender identity is statistically significant, as seen in the regression model.

5.9 Multiple Linear Regression Model with Interaction Effect

Response Variable: prejudiceenvp
Predictor Variable(s): academicsp, academicenvp, NorthCampus, new_sex, academicsp*new_sex

Below is a multiple linear regression model with the response variable prejudice of the climate against various groups at UCLA, and predictors academic satisfaction, academic environment, North Campus, gender identity, and the interaction between academic satisfaction and gender identity.

```
model <- lm(prejudiceenvp ~ academicsp + academicenvp +
NorthCampus + new_sex + academicsp*new_sex, data = cc_new)
summary(model)
```

Predictors	Estimate	Std. Error	T-value	Pr(> $\|t\|$)
(Intercept)	75.58445	1.67713	45.068	< 2e-16 ***
academicsp	-0.07372	0.02657	-2.774	0.00557 **
academicenvp	-0.44739	0.01910	-23.423	< 2e-16 ***
NorthCampusyes	3.33750	0.55023	6.066	1.48e-09 ***
new_sexMale	-0.15336	2.51411	-0.061	0.95136
academicsp:new_sexMale	-0.02891	0.03786	-0.764	0.44511

- Adjusted R-squared: 0.2325
- The Adjusted R-squared value (coefficient of determination) from the model with interaction effect is less than that of the model without interaction effect. This leads us to believe that the model without interaction effect may be more powerful in terms of explaining the outcome.

5.10 Variable Selection

We perform variable selection with the multiple linear regression model containing all 4 predictors and the interaction effect between academic satisfaction and gender identity.

Both backwards selection using AIC and BIC result in our multiple linear regression model being the best-suited model.

6 Conclusion

Response Variable: prejudiceenvp
Predictor Variable(s): academicsp, academicenvp, NorthCampus, new_sex

Final Model:

```
model <- lm(prejudiceenvp ~ academicsp + academicenvp +
NorthCampus + new_sex, data = cc_new)
```

Our best linear model is our original model including all 4 predictors (academic satisfaction, academic environment, North Campus, and gender identity), without interaction effect.

Conclusion

- Adjusted R-squared: 0.2327
- Models with an R-squared value of around 20% are normal for research in the social sciences, where what we are examining may be abstract.

6.1 Shortcomings

- There was an abundance of missing values in variables, which can be resolved by imputing, feature removal and filtering.
- We did not include all of the predictors in the model. This could reduce the bias and inaccurate inferences of the model.
- We did not apply further transformations to reduce the amount of leverage points.
- We did not split the data set into training and testing to avoid overfitting, and see how well our model performs on unforeseen data.
- We did not explore other nonlinear or more flexible machine learning models such as random forest or neural networks.
- Furthermore, we did not apply cross validation (such as K-fold) to provide confidence in performance and evaluate its predictive capabilities.

6.2 Recommendations

- Use tidymodels to explore more machine learning models and different engines since it has the capability to stack the models and evaluate each of their performances respectively.

- Imputing, filtering, mutating, and transformations can all be done using different types of step() function in a recipe.

- Perform initial_split() function to split the data set into training and testing. Usually training consists of 75% of the data and testing consists of 25% of the data.

- Perform hyperparameters tuning such as random grid or grid search to select an optimal set of adjusted parameters for a given learning algorithm.

Code

```r
cc <- read.csv("campusclimate-5.csv")

# eliminate observations for which new_sex is "Other/Unknown"
cc_new <- cc[-which(cc$new_sex == "Other/Unknown"), ]

# Keep only the variables we are interested in
cc_new <- cc_new[, c(10, 14, 32, 172, 135)]

## To create the clean data campusclimate-5updated.csv: write.csv(cc_new, "C:/Users/aidan/OneDrive/Documents/campusclimate-5updated.csv", row.names = TRUE)

# Exploratory Data Analysis (EDA)

## Means and Standard Deviations of Numerical Variables

### Prejudiceenvp

mean(cc_new$prejudiceenvp)
sd(cc_new$prejudiceenvp)

### Academicsp

mean(cc_new$academicsp)
sd(cc_new$academicsp)

### Academicenvp

mean(cc_new$academicenvp, na.rm = T)
sd(cc_new$academicenvp, na.rm = T)

## Summary Statistics of Numerical Variables

summary(cc_new$prejudiceenvp)
summary(cc_new$academicsp)
summary(cc_new$academicenvp)

## Scatterplot Matrix and Correlation Matrix

### Scatterplot Matrix
my_matrix <- data.frame(prejudiceenvp = cc_new$prejudiceenvp, academicsp = cc_new$academicsp, academicenvp = cc_new$academicenvp)
pairs(my_matrix)

## Boxplot and Histogram of Numerical Variables

### Prejudiceenvp

# Boxplot
boxplot(cc_new$prejudiceenvp, main = "Boxplot of the Prejudice of the Climate at UCLA", xlab = "Prejudice of the Climate against Various Groups at UCLA", col  = "pink")

# Histogram
hist(cc_new$prejudiceenvp, main = "Histogram of the Prejudice of
    the Climate at UCLA", xlab = "Prejudice of the Climate against
    Various Groups at UCLA", col  = "pink")

# Calculate skewness
library(moments)
skewness(cc_new$prejudiceenvp)

### Academicsp

# Boxplot
boxplot(cc_new$academicsp, main = "Boxplot of Academic Satisfaction
     at UCLA", xlab = "Academic Satisfaction at UCLA", col  = "
    lightblue")

# Histogram
hist(cc_new$academicsp, main = "Histogram of Academic Satisfaction
    at UCLA", xlim = c(0, 100), xlab = "Academic Satisfaction at
    UCLA", col  = "lightblue")

# Calculate skewness
skewness(cc_new$academicsp)

### Academicevnp

boxplot(cc_new$academicenvp, main = "Boxplot of Academic
    Environment at UCLA", xlab = "Academic Environment at UCLA",
    col  = "lightgreen")

# Histogram
hist(cc_new$academicenvp, main = "Histogram of Academic Environment
     at UCLA", xlab = "Academic Environment at UCLA", col  = "
    lightgreen")

# Calculate skewness
library(e1071)
skewness(cc_new$academicenvp, na.rm = TRUE)

# Barplots and Tables of Relative Frequencies for Categorical
    Predictors

## Barplot and Pie Chart of NorthCampus

# Remove any observations with NAs in NorthCampus
cc_new_new <- cc_new[complete.cases(cc_new$NorthCampus), ]

# Convert NorthCampus to a factor
cc_new_new$NorthCampus <- as.factor(cc_new_new$NorthCampus)

# Create the bar plot and pie chart for NorthCampus
plot(cc_new_new$NorthCampus,
   main = "Bar Plot of NorthCampus",
   col = c("lightblue", "lightpink"))
table_data <- table(cc_new_new$NorthCampus)
percentages <- round(prop.table(table_data) * 100, 1)
pie(table_data, main = "Pie Chart of NorthCampus", col = c("
    lightblue", "lightpink"), labels = paste(names(table_data), "\n
    ", percentages, "%", sep = ""))
```

```r

## Counts for NorthCampus
table(cc_new$NorthCampus)

## Table of Frequency for North Campus
prop.table(table(cc_new$NorthCampus))

## Barplot and Pie Chart of Gender Identity
table_data <- table(as.character(cc_new$new_sex))
barplot(table_data, col = c("lightpink", "lightblue"),
        main = "Bar Plot for Gender Identity (new_sex)")
percentages <- round(prop.table(table_data) * 100, 1)
pie(table_data, main = "Pie Chart for Gender Identity (new_sex)", col = c("lightpink", "lightblue"), labels = paste(names(table_data), "\n", percentages, "%", sep = ""))

## Counts for Gender Identity
table(cc_new$new_sex)

## Table of Frequency for Gender Identity
prop.table(table(cc_new$new_sex))

## Table of Relative Frequency for North Campus vs. Gender Identity
library(dplyr)
library(ggplot2)
cc_new$new_sex <- as.factor(as.character(cc_new$new_sex))
cc_new2 <- cc_new[-which(is.na(cc_new$NorthCampus)), ]
proportion_data <- cc_new2 %>%
  group_by(new_sex, NorthCampus) %>%
  summarize(count = n()) %>%
  mutate(prop = count / sum(count))
ggplot(data = proportion_data, aes(x = new_sex, y = NorthCampus, fill = factor(NorthCampus))) +
  geom_tile(aes(alpha = prop), color = "white") +
  geom_text(aes(label = scales::percent(prop)),
            vjust = 1.5, color = "black") +
  labs(x = "Gender Identity", y = "North Campus", alpha = "Proportion", title = "North Campus vs. Gender Identity(new_sex)") +
  scale_fill_manual(values = c("lightblue", "lightgreen")) +
  theme_minimal()

# Boxplots Comparison for Numerical and Categorical Predictors

## Side by Side Boxplot: Prejudiceenvp v. North Campus
boxplot(prejudiceenvp ~ NorthCampus, data = cc_new2,
  col = c("lightblue", "lightgreen"),
  main = "Side-by-Side Boxplot of Prejudice of the Climate by North Campus",
  xlab = "North Campus",
  ylab = "Prejudice of the Climate")

## Side by Side Boxplot: Prejudiceenvp vs. Gender Identity (new sex)
boxplot(prejudiceenvp ~ as.character(new_sex), data = cc_new,
  col = c("lightblue", "lightgreen"),
  main = "Side-by-Side Boxplot of Prejudice of the Climate by
```

```
    Gender Identity",
  xlab = "Gender Identity",
  ylab = "Prejudice of the Climate")

# Checking for Multicollinearity
my_matrix <- data.frame(prejudiceenvp = cc$prejudiceenvp, academicsp =cc$academicsp, new_sex = cc$new_sex, NorthCampus = cc$NorthCampus)

## Full MLR Model (without interaction)
model <- lm(prejudiceenvp ~ academicsp + academicenvp + NorthCampus + new_sex, data = cc_new)

## Table of VIF (Variance Inflation Factor)
library(regclass)
VIF(model)

# Correlation Coefficient
library(rcompanion)
cor(my_matrix$prejudiceenvp, my_matrix$academicsp)
cramerV(my_matrix$new_sex, my_matrix$NorthCampus)

plot(my_matrix)

# Multiple Linear Regression Model
model <- lm(prejudiceenvp ~ academicsp + academicenvp + NorthCampus + new_sex, data = cc_new)

# Summary of MLR Model
summary(model)

# Linear Model Assumptions
par(mfrow = c(2, 2))
plot(model)

# Leverage Points

## Leverage Points for Multiple Linear Regression (MLR) are defined when leverage is greater than
n <- nrow(cc_new)
p <- 4
number <- 2 * ((p + 1) / n)

# Plot of leverage points
plot(hatvalues(model), type = ’h’)
leverage_points <- which(hatvalues(model) > number)
abline(h = number, col = "red")

# What proportion of observations indicate potential leverage points
mean(hatvalues(model) > number)

# Interaction Effect between Academicsp & new_sex
model <- lm(prejudiceenvp ~ academicsp*new_sex, data = cc_new)

# Summary of model
summary(model)
```

```r

## Plots of interaction effect
library(car)
library(effects)

### First plot of interaction effect
cc_new$new_sex <- as.factor(cc_new$new_sex)
newdata <- expand.grid(academicsp = seq(min(cc_new$academicsp), max(cc_new$academicsp), length.out = 100), new_sex = levels(cc_new$new_sex))
pred <- predict(model, newdata, interval = "confidence")
newdata$prejudiceenvp <- pred[, "fit"]
newdata$lower_CI <- pred[,"lwr"]
newdata$upper_CI <- pred[,"upr"]

library(ggplot2)
ggplot(data = newdata, aes(x = academicsp, y = prejudiceenvp, color = new_sex, fill = new_sex)) +
  geom_line() +
  geom_ribbon(aes(ymin = lower_CI, ymax = upper_CI), alpha = 0.1, color = NA) +
  facet_wrap(~new_sex) +
  labs(title = "Academic Satisfaction and Gender Identity Interaction Effect Plot", x = "Academic Satisfaction at UCLA", y = "Prejudice Environment at UCLA", color = "Gender Identity",fill = "Gender Identity" ) +
  theme_minimal()

### Second plot of interaction effect
plot(allEffects(model), ask=FALSE)

# Variable Selection for MLR Model with Interaction Effect

## MLR Model with Interaction Effect
model <- lm(prejudiceenvp ~ academicsp + academicenvp + NorthCampus + new_sex + academicsp*new_sex, data = cc_new)

## Summary of model
summary(model)

### Backwards elimination using AIC.
backAIC <- step(model, direction = "backward", data = cc_new)
backAIC

### Backwards elimination using BIC.
backBIC <- step(model, direction = "backward", data = cc_new, k = log(nrow(cc_new)))
backBIC
```