Stats 101C Midterm Report

Meghal Dubey, Sibo Guo, Humin Zhang, Julius Castro, Qinyi Chen, Luis Ceja Abrica

November 27, 2020

Research Question

Our team chose the diversity dataset provided to us by Prof. Esfandiari to further explore diversity at UCLA. This study has 939 total students and we asked two main questions we were interested in exploring to further understand Diversity at UCLA. In the first question, we are conducting analysis to see if there is any association of different religions (" religion") with the score of feeling respected("divrespectp") and the score of improving the UCLA climate ("impuclaclimate"). We ran two models for this analysis, the first one was conducted using the levels "Christian" and "Spiritual but not associated with a major religion" and the second one using the levels "Christian" and "Not particularly Spiritual."In the second question, we wanted to explore the relationship between feeling respected at UCLA with feeling discriminated against, observed discrimination at UCLA and UCLA climate. We used the variables Uclaclimate, Ucladiscp, and Uclaexclusionaryp as the Predictors, the variable Divrespectp as the Outcome. For the first question, we used a logistic regression model and for the second question, we used an MLR model. We presented our analysis in the form of graphs (bar plots, box plots, scatter plots etc.) and included a summarized version of our findings.

```
                                                Christian
                                                      324
                                        Eastern religion
                                                       13
                                                   Jewish
                                                       15
                                                   Muslim
                                                       10
                                    Not particularly spiritual
                                                      323
Spiritual but not associated with a major religion
                                                      115
```

Exploratory Analysis and Methodology

For question 1, we first made a table of the" religion" variable to see if the data was balanced. From the table above, it is clear that since the data is not perfectly balanced, some data cleaning would be required. We decided to remove Muslim, Jewish and Eastern Religion from our analysis since the numbers are not large enough for us to draw meaningful conclusions. After cleaning out the variables, we checked for outliers using box plots.

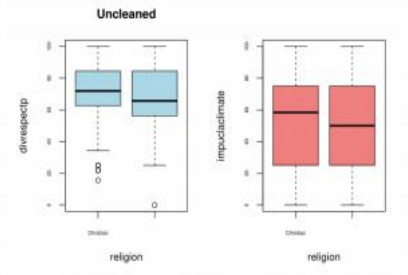


Figure 1: Uncleaned boxplot for divrespectp and impuclimate using christian and spiritual

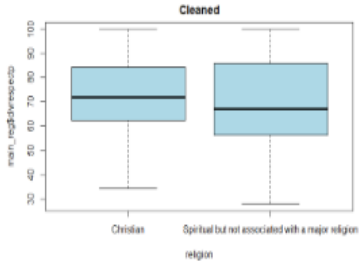


Figure 2: Cleaned Boxplot for our variable "divrespectp" using Christian and Spiritual

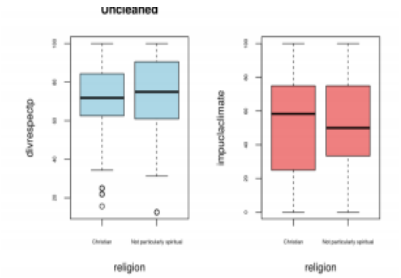


Figure 3: Uncleaned boxplot for divrespecctp and impuclimate using christian and non-spiritua

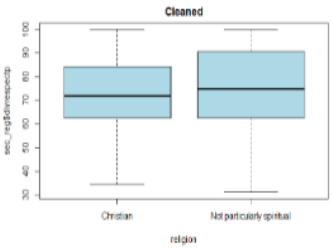


Figure 4: Cleaned Boxplot for our variable "divrespectp" using Christian and non-Spiritual

After we cleaned up our data, we created histograms for the cleaned and uncleaned data to help us visualize the distributions. After we are cleaned the data, we ran the logistic regression models to fit our variables. From the histograms of the final distributions, we can see that while the distribution for the variable" impuclaclimate" has improved significantly, the distribution for" divrespectp" has improved slightly. Lastly, we coded a correlation matrix between the two variables and made a scatter plot to check for any patterns. After observing no visible patterns in the data, we decided to go ahead with our logistic model(s).

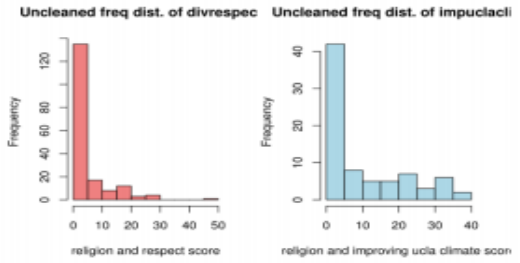


Figure 5: uncleaned distribution

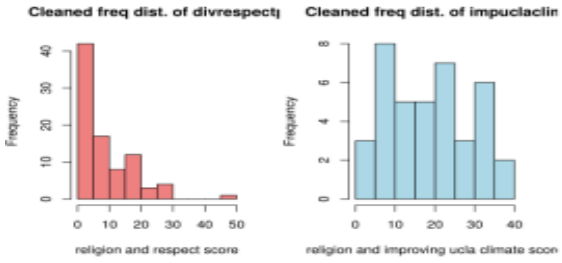


Figure 6: Cleaned distribution

Correlation matrix

```
             [,1]        [,2]
[1,] 1.000000000 0.006048362
[2,] 0.006048362 1.000000000
```

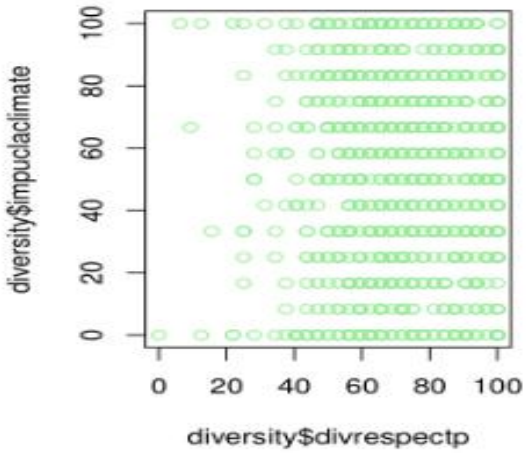


Figure 7: Scatter plot of Predictor Variables

Since the correlation between the two numerical variables is lesser than 0.1 and there is no pattern observed in their scatterplot, we can conclude that there is no correlation between predictors in the model.

For question 2, We observed that there was 1 missing observation in the predictor variable Uclaclimate. Since we have 939 obsevations, we removed this missing observation from our dataset. Additionally, since there were only 7 observations in" Very uncomfortable" group, we combined them with the" Uncomfortable" group to give us a total of 46 observations in the" Uncomfortable" group. Then, we ordered the levels of Uclaclimate from" Uncomfortable","

Somewhat comfortable”,” Comfortable”, to” Very comfortable” so that it would make more sense for us when we make the barplot of Uclaclimate.

```
     Comfortable Somewhat comfortable       Uncomfortable
             448                   242                  46
Very comfortable
             202
```

Figure 9: Table of ”uclaclimate” after cleaning the data

After cleaning up the” uclaclimate” variable and depicting a plot of its proportion table, we also depicted a boxplot of” uclaclimate” with our outcome variable, ”divrespectp.” Since we are only removing 1 observation in a 939-variable dataset, (as opposed to a larger amount of observations in question 1) we do not remove the outliers since it will not significantly impact our model.Then, we plotted the distribution of our outcome variable, ”divrespectp” followed by the scatter plots of ”divrespectp” and ”ucladiscp” as well as ”divrespectp” and ”uclaexclusionaryp” to better visualize the data. Lastly, we made a correlation matrix and scatterplot to check if there was any correlation between the two numerical variables. After seeing no patterns in the data that might impact our analysis, we decided to go ahead with our MLR model.

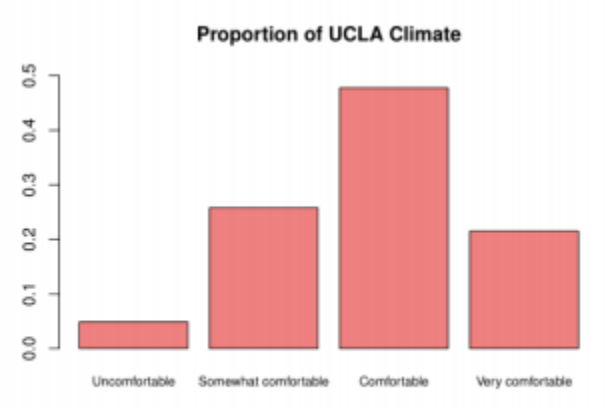


Figure 10: Barplot of ”ucla climate” after cleaning the data”

The bar plot of the proportion of UCLA climate shows that 47.8% of the student feel comfortable with the UCLA climate, 25.8% of the student feel somewhat comfortable, and 21.5% of the student feel very comfortable. Only 4.9% of the student feel uncomfortable. It indicates that UCLA is doing good to provide an equitable, diverse, and inclusive environment for the students.

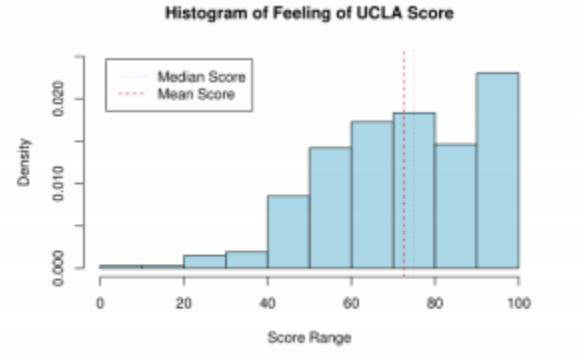


Figure 11: Histogram of ”divrespectp”

The histogram plot of the feeling respect score is skewed left, meaning that most scores are in the high range of values and a few are very low. It indicates that most students are feeling comfortable about UCLA. Also, the median and mean scores are 75 and 72.6.

```
1.0000000 0.2940275
0.2940275 1.0000000
```

Figure 14: Correlation matrix of numerical variables "ucladiscp" and "uclaexclusionaryp"

Since the correlation between the two numerical variables is lesser than 0.3 and there is no pattern observed in their scatterplot, we can conclude that there is no multicollinearity in the model.

Statistical Analysis

Question 1: Association of different religions with the score of feeling respected and the score of improving the UCLA climate. Model 1: Is there any association of religion levels "Christian" and "Spiritual but not associated with a major religion" with the feeling respected at UCLA (divrespectp) and UCLA climate (impuclaclimate)?

```
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.4373     0.0955  4.5810   0.0000
main_reg$divrespectp -0.0024     0.0012 -1.9603   0.0506
main_reg$impuclaclimate -0.0001  0.0007 -0.1075   0.9145
```

Figure 16: Table of Coefficients and their corresponding p-values

```
(Intercept)    main_reg$divrespectp main_reg$impuclaclimate
  1.5484653               0.9975885              0.9999279
```

Figure 17: Table of log-odds

```
main_reg$divrespectp main_reg$impuclaclimate
            1.003232               1.003232
```

Figure 18: VIF table for model1

Our models have two predictors, the score of feeling respected and the improving ucla climate score with our outcome variable being religion (with two levels, Christian and Spiritual). For the first model we are looking at the relationship between Christian and spirtual with score on feeling respected and ucla climate score. Based on the result obtained from the model, we can conclude that feeling respected score and improving UCLA climate score has no strong associations with the religion Christian and spiritual since their p values are greater than our significance level of 0.05. This value is also confirmed when we run AIC on our model. For this model, we get the odd ratios for feeling respected score as 0.9975885 and 0.9999279 for improving ucla climate score. This implies that 1 unit increase in respect score will increase the odds of religion being Christian and spiritual by 0.9975885 and 1 unit increase in improving ucla climate score will increase the odds of religion being christian and spiritual by 0.9999279. Since the VIF for our variables is less than 5, we can say that there is no multicollinearity in the model. From the outlier analysis above, we can see that while there are 6 leverage points, none of them are bad leverage points (StudRes < |2|)

```
      StudRes         Hat        CookD
275 -0.815930 0.018551803 0.004198002
367  1.862178 0.009594858 0.011133972
388  1.457795 0.022420092 0.016203793
412  1.854957 0.015212689 0.017617341
413  1.854957 0.015212689 0.017617341
419  1.867180 0.011926490 0.013946302
```

Figure 19: leverage values

```
Model:
main_reg$religion1 ~ main_reg$divrespectp
                        Df Deviance    AIC F value Pr(>F)
<none>                       82.151 514.72
main_reg$impuclaclimate  1   82.149 516.71  0.0115 0.9145
```

Figure 20: AIC model and their respective p-values

```
Model:
main_reg$religion1 ~ main_reg$impuclaclimate
                      Df Deviance    AIC F value  Pr(>F)
<none>                     82.886 518.56
main_reg$divrespectp  1   82.149 516.71  3.8428 0.05061 .
```

Figure 21: AIC model and their respective p-values

Model2: Is there any association of religion levels "Christian" and "Non-Spiritual" with the feeling respected at UCLA (divrespectp) and UCLA climate (impuclaclimate)?

Model result

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.2860274866	0.0946151070	3.0230636	0.002602828
sec_reg$divrespectp	0.0025041589	0.0011658222	2.1479768	0.032092092
sec_reg$impuclaclimate	0.0005646872	0.0006610354	0.8542466	0.393289522

Our second looks at the relationship between Christian and non-spiritual with the score of feeling respected and the improving ucla climate score. Based on the result, we observe that the score on feeling respected is significant for religions of Christian and non-spiritual since the p value is 0.0321 which is less than 0.05. Since the p-value for impuclaclimate is above our significance level of 0.05, we fail to reject the null and conclude that there is no significant association between impuclaclimate and the religions Christian and non-spiritual. This result is also confirmed when we run AIC on the model.

Odd ratio:

(Intercept)	sec_reg$divrespectp	sec_reg$impuclaclimate
1.331129	1.002507	1.000565

For the second model, we get the odd ratios for feeling respected score as 1.002507 and 1.000565 for improving ucla climate score. This implies that 1 unit increase in respect score will increase the odds of religion being Christian and spiritual by 1.002507 and 1 unit increase in improving ucla climate score will increase the odds of religion being Christian and spiritual by 1.000565. However, as we can see from the above results the odd ratios are all very close to one for both models which indicates the change are close to nothing or very small.

sec_reg$divrespectp	sec_reg$impuclaclimate
1.000486	1.000486

Figure 23: VIF table for Model 2

	StudRes	Hat	CookD
275	-0.7509242	0.01472757	0.002811533
393	1.2411051	0.01208666	0.006276463
464	1.1704790	0.01607193	0.007455184
530	1.2349037	0.01225563	0.006301996

Figure 24: Leverage Values

Since the VIF for our variables is less than 5, we can say that there is no multicollinearity in the model. From the outlier analysis above, we can see that while there are 4 leverage points, none of them are bad leverage points (StudRes < |2|)

Model:
sec_reg$religion2 ~ sec_reg$divrespectp

	Df	Deviance	AIC	F value	Pr(>F)
<none>		158.83	930.31		
sec_reg$impuclaclimate	1	158.65	931.58	0.7297	0.3933

Figure 25: AIC

Model:
sec_reg$religion2 ~ sec_reg$impuclaclimate

	Df	Deviance	AIC	F value	Pr(>F)
<none>		159.80	934.20		
sec_reg$divrespectp	1	158.65	931.58	4.6138	0.03209 *

Figure 26: AIC

Based on the two models we can clearly see that the score on feeling respected will be impactful when there are non-spiritual people in the group. From the results we can conclude that feeling respected score will be a differentiating factor when facing with different religion groups whereas the improving ucla climate score seems to be meaning less when used as a predictor for different religions.

Question 2: Is there any relationship between feeling respected at UCLA with feeling discriminated against, observed exclusion at UCLA and UCLA climate?

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       81.37797    0.93479  87.055  < 2e-16 ***
ucladiscp                         -0.46255    0.04597 -10.062  < 2e-16 ***
uclaexclusionaryp                 -0.15024    0.02754  -5.455 6.28e-08 ***
uclaclimateSomewhat comfortable -6.66994    1.28383  -5.195 2.51e-07 ***
uclaclimateUncomfortable        -6.40532    2.53578  -2.526   0.0117 *
uclaclimateVery comfortable      9.07868    1.31203   6.920 8.42e-12 ***
```

Figure 27: Model summary

The summary result shows that the two quantitative predictors experienced discrimination and observed exclusion at UCLA have a negative relationship with the feeling respected at UCLA. They are all significant predictors that have a p-value < 0.05. The qualitative predictor Comfortable, somewhat comfortable, and Uncomfortable, in UCLA climate is also a significant predictor because they have a p-value < 0.05. Therefore, we reject the null hypothesis, and conclude that there is a relationship between feeling about UCLA and experienced discrimination at UCLA, observed exclusion, and UCLA Climate.

Base on the Multiple Linear Regression: $y_i = \beta_0 + \beta_1 xi_1 + \beta_2 xi_2 + \beta_3 xi_3 + \beta_4 xi_4 + \beta_5 xi_5$

$y_i = \beta_0 + \beta_1 xi_1 + \beta_2 xi_2$ if UCLA climate is Comfortable.

$y_i = \beta_0 + \beta_1 xi_1 + \beta_2 xi_2 + \beta_3 xi_3$ if UCLA climate is Somewhat comfortable.

$y_i = \beta_0 + \beta_1 xi_1 + \beta_2 xi_2 + \beta_4 xi_4$ if UCLA climate is Uncomfortable.

$y_i = \beta_0 + \beta_1 xi_1 + \beta_2 xi_2 + \beta_5 xi_5$ if UCLA climate is Very comfortable.

The Null hypothesis: $H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$.

The Alternative Hypothesis: at least one $\beta \neq 0$.

Interpretations:

Keeping all other predictor variables fixed, for one unit of increase in ucladiscp (feeling dscriminated against), on average, divrespectp (feeling respected at ucla) decreases by 0.046.

Keeping all other predictor variables fixed, for one unit of increase in uclaexlusionaryp (feeling excluded), on average, divrespectp (feeling respected at ucla) decreases by 0.15.

Since the minimum value for both the numerical variables interpreted above was equal to 0, these interpretations make sense.

Keeping all other predictor variables fixed, the students who feel somewhat comfortable, on average, feel 6.6 points less respected at UCLA as compared to those who feel comfortable.

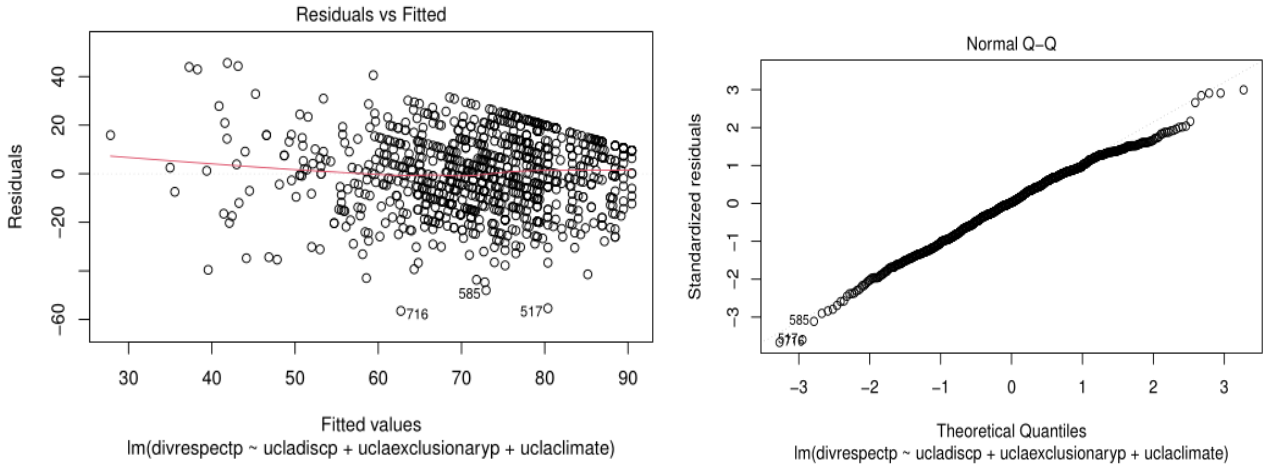
Keeping all other predictor variables fixed, the students who feel uncomfortable, on average, feel 6.4 points less respected at UCLA as compared to those who feel comfortable.

Keeping all other predictor variables fixed, the students who feel very comfortable, on average, feel 9 points more respected at UCLA as compared to those who feel comfortable.

Intuitively, our interpretations from our model make a lot of sense. If a student was feeling more discriminated on campus, they might feel less respected. Similarly, if they were feeling more excluded, they might feel less respected.

Compared to a student who feels comfortable, it is expected that students who feel somewhat comfortable or uncomfortable will feel less respected on campus and those who feel very comofrtable will feel more respected on campus.

From the summary table of the model, we can see that the value of R^2 is equal to 31.54%. This implies that 31.54% of variation in "divrespectp" is explained by our model (which uses the predictors "uclaexclusionaryp","ucladiscp" and "uclaclimate")



We are meeting the assumption of equality of error variance as there is no pattern in the Residual Plot. The Normal Probability Plot also looks good. Most of the points are on the qq line leading us to the conclusion that we are meeting the normality assumption. Also, there is no multicollinearity between the predictors since all the VIF < 5. There are 5 leverage points in the data, 673 is a good leverage point (since the absolute value of "StudRes" is less than 2) and 363, 517, 716, and 857 are bad leverage points.

	GVIF	Df	GVIF^(1/(2*Df))
ucladiscp	1.296500	1	1.138639
uclaexclusionaryp	1.103667	1	1.050556
uclaclimate	1.225921	3	1.034532

Figure 28: VIF table

After conducting our initial analysis, we used best subset as well as forward and backward selection to check the relationship between the outcome and predictor variables. We then compared the three methods to see if there was any disparity between the three methods.

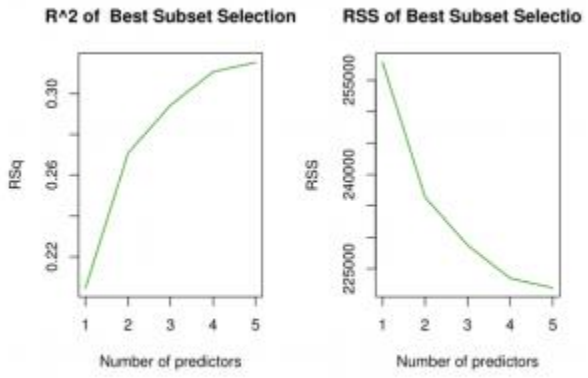


Figure 30: Best Subset Selection

We used a blue line for best subset selection, the red line for forward selection and the green line for backward selection to make two plots to compare the different between R^2 and RSS in the three methods. Since those 3 lines perfectly overlap each other, we can conclude that there is no significant difference between them, and our model selection is accurate.

Conclusion: There is a relationship between feeling about UCLA and experienced discrimination at UCLA, observed exclusion, and UCLA Climate. Most students are feeling comfortable about UCLA, and UCLA is doing good to provide an equitable, diverse, and inclusive environment for the students.

Shortcomings and Practical Significance

In question 1, we had originally thought that there would be enough observations in religions such as Jewish, Muslim and Eastern religions for us to investigate the relationship between different religions with respect to the respect score and ucla climate score. Since those religions were way too small for us to analyze, we had to remove them from our analysis. This caused a problem for us because we were only looking at the majority religions which can be potentially different from our original idea since we know religions such as Jewish, Muslim and Eastern religions tend to be the religions that face more stereotypes or discrimination issues which is highly related to the two predictors that we are using.If we had a more balanced number for each religion, we would have been able to see the association between outcome variable and predictors much better by running a linear model for each religion and comparing all of them. Since our current model has Christian as the majority base and only spiritual and non-spiritual levels, our conclusion might be slightly biased.

From the model that we currently have, it was very interesting to learn that the score on improving the UCLA climate (impuclaclimate) was not significant for both our models. This shows that for the levels "Christian", "Spiritual" and "NonSpiritual," there is no association with improving UCLA climate. Additionally, since the score on feeling respected (divrespectp) was significant in model 2 and not significant in model 1, it was a differentiating factor between the two. Practically speaking, this information can be helpful since we know that feeling respected on campus impacts those who are nonspiritual.

In question 2, our largest shortcoming is that we could have used a few more predictors to make our analysis stronger. While we chose three predictors that we thought would be interesting to explore, further looking into predictors such as sexualorient, ethnicity and socioeco would have helped us in further analyzing how these may impact feeling respected on campus. Additionally, even though we have a large amount of observations, we could have modified the model by removing the bad leverage point to see how it could have impacted our analysis.

Conclusion

In our analysis, we used the diversity data set, with a total number of 939 students/observations. We found some interesting conclusions using procedures such multiple linear regression and logistic regression. We discussed some of today's most relevant topics and wanted to use Statistical Analysis to depict our findings. We wanted to explore topics such as discrimination at UCLA, exclusion on campus and UCLA climate to see how they impacted the level of respect students feel on campus. We also wanted to see if religion was correlated with observed discrimination on campus and UCLA climate.

We conducted exploratory analysis, depicting barplots, frequency and proportion tables to get a better understanding of the data and clean up the data if there were not enough observations to draw conclusions. We also acknowledge that some of the data was not entirely balanced, which may have caused some of our conclusions to be skewed and that additional analysis needs to be conducted to draw concrete results.

There is a relationship between feeling about UCLA and experienced discrimination at UCLA, observed exclusion, and UCLA Climate. There is no strong association between religions with feeling respected and improving UCLA climate score This project was an extremely exciting experience for us since we could take a deeper dive into some of the most pressing issues that UCLA students face and learn a lot more about our campus in the process!

6 Appendix - Code

```
library(dplyr)
div<-read.csv("div.oct.17.csv")
diversity <- read.csv("div.oct.17.csv")
mydata <- select(div,divrespectp,impuclaclimate,religion)

mydata<-mydata %>% filter(religion != "Other")
#any(is.na(mydata)) #check to see if any missing value exist
#attach(mydata)



t<-table(mydata$religion)
t

#cleaning
a <-mydata %>% filter(religion == "Christian")
b<-mydata %>% filter(religion == "Spiritual but not associated with a major religion")
c <-mydata %>% filter(religion == "Not particularly spiritual")
main_reg<-rbind(a,b)
sec_reg<- rbind(a,c)
new1<-rbind(main_reg,c)



par(mfrow=c(1,2))
boxplot(main_reg$divrespectp~main_reg$religion,
        cex.axis=0.5,xlab = "religion", ylab = "divrespectp", main = "Uncleaned", col = "lightblue") #u
boxplot(main_reg$impuclaclimate~main_reg$religion,
        xlab = "religion", cex.axis = 0.5, ylab = "impuclaclimate", col = "lightcoral")

out1<-boxplot(main_reg$divrespectp~main_reg$religion, plot=FALSE)$out
main_reg<- main_reg[-which(main_reg$divrespectp %in% out1),]
boxplot(main_reg$divrespectp~main_reg$religion, main = "Cleaned", xlab ="religion",col = "lightblue")
#remove outliers

par(mfrow=c(1,2))
boxplot(sec_reg$divrespectp~sec_reg$religion,
        cex.axis=0.5,xlab = "religion", ylab = "divrespectp", col = "lightblue",main = "Uncleaned" )#we
boxplot(sec_reg$impuclaclimate~sec_reg$religion,
        xlab ="religion", cex.axis = 0.5, ylab = "impuclaclimate", col = "lightcoral")

out2<-boxplot(sec_reg$divrespectp~sec_reg$religion, plot=FALSE)$out
sec_reg<- sec_reg[-which(sec_reg$divrespectp %in% out2),]
boxplot(sec_reg$divrespectp~sec_reg$religion, main = "Cleaned", xlab ="religion", col = "lightblue")

#unclean data distribution
par(mfrow=c(1,2))
divr<-table(mydata$divrespectp,mydata$religion)
hist(divr,main="Uncleaned freq dist. of divrespectp" ,
     xlab="religion and respect score", col = "lightcoral")
div_imp<-table(mydata$impuclaclimate,mydata$religion)
hist(div_imp,main="Uncleaned freq dist. of impuclaclim",
```

```
     xlab="religion and improving ucla climate score", col = "lightblue")



#clean data distribution
par(mfrow=c(1,2))

tt22<-table(new1$divrespectp,new1$religion)
hist(tt22,main ="Cleaned freq dist. of divrespectp",
     xlab="religion and respect score", col = "lightcoral")
tt33<-table(new1$impuclaclimate,new1$religion)
hist(tt33,main="Cleaned freq dist. of impuclaclim",
     xlab="religion and improving ucla climate score", col = "lightblue")


df25 <- cbind(diversity$divrespectp, diversity$impuclaclimate)

cor(df25)
plot(diversity$divrespectp, diversity$impuclaclimate, col = "lightgreen")

diversity <- read.csv("div.oct.17.csv")
na_id <- which(is.na(diversity$uclaclimate))
#na_id

diversity <- diversity[-na_id, ]
#dim(diversity)

table(diversity$uclaclimate)

# Figure: Table of "uclaclimate" before cleaning the data


diversity$uclaclimate <- as.factor(diversity$uclaclimate)
levels(diversity$uclaclimate)[5] <- c("Uncomfortable")
table(diversity$uclaclimate)
# Figure: Table of "uclaclimate" after cleaning the data
prop.table(table(diversity$uclaclimate))

ucla_levels <- c("Uncomfortable", "Somewhat comfortable",
                 "Comfortable", "Very comfortable")

diversity$uclaclimate <- factor(diversity$uclaclimate, levels = ucla_levels,
                                ordered = TRUE)
# Figure: Proportion table of "ucla climate" after cleaning the data"
barplot(prop.table(table(diversity$uclaclimate)), col = "lightcoral", cex.names = 0.75,
        ylim = c(0, 0.5), main = "Proportion of UCLA Climate")
# Figure: Barplot of "ucla climate" after cleaning the data"

hist(diversity$divrespectp, freq = FALSE, col = "lightblue", ylim = c(0, 0.025),
     xlab = "Score Range", main = "Histogram of Feeling of UCLA Score")

abline(v = c(median(diversity$divrespectp), mean(diversity$divrespectp)),
       lty = c(3, 2), col = c("blue", "red"))
```

```r
legend("topleft", c("Median Score", "Mean Score"), col = c("blue", "red"),
       lty = c(3, 2), inset = 0.05)

boxplot(diversity$divrespectp ~ diversity$uclaclimate, col = "lightcoral")

# Figure: Boxplot of "divrespectp" vs "diversityfuclaclimate"


plot(diversity$divrespectp, diversity$ucladiscp, xlab = "Score of experienced discrimination",
     ylab = "Score of feeling about UCLA", col = "red",
     main = "Score of feeling about UCLA and experienced discrimination")

# Figure: Scatterplot of variables "divrespectp" and "ucladiscp"

plot(diversity$divrespectp, diversity$uclaexclusionaryp, xlab = "Score of observed exclusion",
     ylab = "Score of feeling about UCLA", col = "green",
     main = "Score of feeling about UCLA and observed exclusion")
# Figure: Scatterplot of variables "divrespectp" and "uclaexclusionaryp"

df3 <- cbind(diversity$ucladiscp, diversity$uclaexclusionaryp)

cor(df3)

# Figure: Correlation matrix of numerical variables "ucladiscp" and "uclaexclusionaryp"

plot(diversity$ucladiscp, diversity$uclaexclusionaryp, col = "yellow")

#model:
religion1<- rep(NA,nrow(main_reg)) #new col for religion and change to num in order to fit in glm
main_reg<- cbind(main_reg,religion1)

main_reg$religion1[main_reg$religion == "Christian"]<-0
main_reg$religion1[main_reg$religion == "Spiritual but not associated with a major religion"]<-1
# class(main_reg$religion1)
lr <- glm(main_reg$religion1~main_reg$divrespectp+main_reg$impuclaclimate,data=main_reg)
round(summary(lr)$coef, 4)

#Figure: Table of Coefficients and their corresponding p-values

christian_and_spirtual <- exp(coef(lr))#odd ratios
christian_and_spirtual

# Figure: Table of log-odds
library(car)
vif(lr)

# Figure: VIF table for Model 1

infl2 <- influencePlot(lr)
infl2

# Figure: Influence Plot and Leverage Values
```

```r
nnn<-nrow(main_reg)
#AIC:
lr.sse0 <- sum(resid(lr) ^2)
b <- nnn + nnn*log(2*pi) + nnn * log(lr.sse0 / nnn) + 2 * (1+1)
a <- AIC(lr,k=2)

v <- nnn + nnn * log(2*pi) + nnn*log(lr.sse0/nnn) + log(nnn)*(1+1)
d<-  AIC(lr,k=log(nnn))

add1(glm(main_reg$religion1~main_reg$divrespectp),main_reg$religion1~main_reg$divrespectp+main_reg$impu
add1(glm(main_reg$religion1~main_reg$impuclaclimate),main_reg$religion1~main_reg$divrespectp+main_reg$i

religion2<- rep(NA,nrow(sec_reg)) #new col for religion and change to num in order to fit in glm
sec_reg<- cbind(sec_reg,religion2)

sec_reg$religion2[sec_reg$religion == "Christian"]<-0
sec_reg$religion2[sec_reg$religion == "Not particularly spiritual"]<-1
lr1 <- glm(sec_reg$religion2~sec_reg$divrespectp+sec_reg$impuclaclimate,data=sec_reg)
summary(lr1)$coef
#table for odd ratios
christian_and_nonspirtual <-exp(coef(lr1))#odd ratios
christian_and_nonspirtual


# VIF
library(car)
vif(lr1)

# Figure: VIF table for Model 2

infl1 <- influencePlot(lr1)
infl1

# Figure: Influence Plot and Leverage Values

#AIC:
lr1.sse0 <- sum(resid(lr1) ^2)
b1 <- nnn + nnn*log(2*pi) + nnn * log(lr1.sse0 / nnn) + 2 * (1+1)
c1 <- AIC(lr,k=2)

d1 <- nnn + nnn * log(2*pi) + nnn*log(lr1.sse0/nnn) + log(nnn)*(1+1)
a1 <- AIC(lr1,k=log(nnn))

add1(glm(sec_reg$religion2~sec_reg$divrespectp),sec_reg$religion2~sec_reg$divrespectp+sec_reg$impuclacl
add1(glm(sec_reg$religion2~sec_reg$impuclaclimate),sec_reg$religion2~sec_reg$divrespectp+sec_reg$impucl


diversity$uclaclimate <- as.character(diversity$uclaclimate)
lm <- lm(divrespectp ~ ucladiscp + uclaexclusionaryp + uclaclimate, data = diversity)
summary(lm)
plot(lm)
library(car)
vif(lm)
```

```r
infl <- influencePlot(lm)
infl

library(leaps)

best_subset <- regsubsets(divrespectp ~ ucladiscp + uclaexclusionaryp + uclaclimate,
                          data = diversity, nbest = 1, nvmax = 6,
                          intercept = TRUE, method = "exhaustive",
                          really.big = FALSE)
sumBS <- summary(best_subset)
sumBS

forward_sel <- regsubsets(divrespectp ~ ucladiscp + uclaexclusionaryp + uclaclimate,
                          data = diversity, nbest = 1, nvmax = 6,
                          intercept = TRUE, method = "forward",
                          really.big = FALSE)
sumF <- summary(forward_sel)
sumF

backward_sel <- regsubsets(divrespectp ~ ucladiscp + uclaexclusionaryp + uclaclimate,
                           data = diversity, nbest = 1, nvmax = 6,
                           intercept = TRUE, method = "backward",
                           really.big = FALSE)
sumB <- summary(backward_sel)
sumB

par(mfrow = c(1, 2))

plot(sumBS$rsq, xlab = "Number of predictors", ylab = "RSq",
     type = "l", col = "blue", main = 'R^2 of  Best Subset Selection')
lines(sumF$rsq, col = "red")
lines(sumB$rsq, col = "green")

plot(sumBS$rss, xlab = "Number of predictors", ylab = "RSS",
     type = "l", col = "blue", main = 'RSS of Best Subset Selection')
lines(sumF$rss, col = "red")
lines(sumB$rss, col = "green")
```