MET CS 555 Term Project

10 points **(Due December 18, 2022 at 11:59 PM)**

1. Assignment Description

Find a dataset for a research problem of interest, here are some good websites for this

Kaggle Data Science Competitions: http://kaggle.com
UCI Machine Learning repository: http://archive.ics.uci.edu/ml/datasets.php
Google Cloud public data:
https://console.cloud.google.com/marketplace/browse?filter=solution-type:dataset&_ga=2.265202890.490000482.1586060190-1118401016.1586060190&pli=1

Describe a research scenario and specify a research question based on data analytic methods that we learned in class, for example methods like, *one and two sample means, t-test, correlation tests, simple and multiple linear regression, ANOVA and ANCOVA, one and two-Sample Tests for Proportions and logistic regression*.

Clean up your data and reduce it to no more than 500 observations if your data set is large. If your task is to build a prediction model, ensure that you split the data into **training and testing datasets at this stage.**

2. Research Scenario Description (no more than 200 words)

Describe your research scenario in no more than 200 words. This is a general description of the use case. Give relevant background information so that the reader can understand your research question. Assume that the reader does not have any specialized knowledge about your topic. Define all acronyms/abbreviations and try to avoid jargon.

When pursuing a postgraduate education, applicants have to fulfill different requirements for the admission committee of the university to review and decide whether to admit said applicant. In majority cases, a Graduate Record Examination (GRE) exam, Test of English as a Foreign Language (TOEFL) exam, a Statement of Purpose (SOP) essay, two to three Letters of Recommendation (LOR) and an official transcript including their Cumulative Grade Point Average (CGPA) from the university of their undergraduate studies are all required. Additionally, possessing research experience may also serve as an additional benefit to applicants. The purpose of this research is to determine the probability of being admitted after taking into account all the above requirements.

3. Describe the data set (no more than 400 words)

Describe each of the columns of the data set **that are used in your analysis**. Clean up your data before usage (e.g. assess and remove outliers, perform any necessary transformations, ensure assumptions of your analysis are met). Remove unused columns. Describe each step of your data cleaning process and why you did this step. If possible, provide a link to the main data set source.

The dataset contains 400 different applicants applying to different universities with each of the requirements per column and their scores. The university names are not given, but instead a ranking scale out of 5 is given to the universities
Each column in this dataset is as described:
GRE Score: Exam score out of 340
TOEFL Score: Exam score out of 120
University Rating: Random pool of universities ranked out of 5
SOP: Essay graded out of 5 (Assume that the necessary points for a good essay is scaled out of 5)
LOR: Letter is scaled out of 5 (Assume that the necessary points for a good letter is scaled accordingly)
CGPA: Cumulative Grade Point Average out of 10
Research: 1 = Yes, 0 = No
Chances of Admit: A probability score out of 1 with 1 being a guaranteed to being admitted.

Data Cleaning:
This dataset did not require much cleaning except for removing the extra Serial Number column which represented the applicant number. The same serial number can be obtained using the row number. All other columns in the dataset should not be cleaned or removed for potential outliers because in the real-world context, it is very much possible to score full marks for both the GRE and TOEFL exams. Similarly, any university has the potential to score a full 5 out of 5 in the University rating (depending on Ivy-League status, cost of tuition and other factors). The other factors, SOP, LOR and CGPA are all completely possible to obtain a full score. Research is a simple yes or no that should not be removed as some applicants are potentially looking for research opportunities that could potentially stand the applicant out. Finally, the Chances of Admission is the variable that the study will focus more on. The probability score will be studied and observed to see if each of the previously mentioned variables play a vital role in the admission chances.

Attached below is the link to download the original dataset:
https://www.kaggle.com/datasets/akshaydattatraykhare/data-for-admission-in-the-university

4. Research Question (no more than 100 words)

Describe the main research question in one or two sentences. This is similar to the last sentence of our class examples.

Research Question:

To determine if each factor considered when applying for postgraduate education at a university has equal or unequal association with probability of being admitted to a randomly selected university.

5. Statistical Analysis

Give at least one main visualization that supports your conclusion. Be sure to use correct axis labels and avoid common mistakes when generating visualizations. **State all assumptions of the statistical technique(s) that you are using and give evidence as to whether or not these assumptions are met.**

1. Run the cor() function and display the correlation pair graphs. We want to look at each factor compared against the admission chances. From the cor() function, we can see that each factor has different correlation values that are significantly enough to be different.
2. Calculate the R^2 to determine how the proportion of the variation in the admission chances is affected by the MLR model containing all the independent variables.
3. Calculate the Least Squares Regression Equation that predicts Admission Chances based on all the above factors. We get the equation below:
   Admission Chances = -1.2594325 + 0.0017374GRE + 0.0029196TOEFL + 0.0057167UniRank - 0.0033052SOP + 0.0223531LOR + 0.1189395CGPA + 0.0245251Research
4. Formally test if these factors are associated with admission chances at α = 0.01
   Using the 5-step method, we set up the hypothesis, choose the appropriate test statistic, state the decision rule, compute the test statistic and conclude.
   H0: βGRE = βTOEFL = βUniRank = βSOP = βLOR = βCGPA = βResearch = 0
   H1: Any of the β predictors are not equal to 0
   Using Global F-Test: F = MS Reg / MS Res
   Decision Rule: Using qf(0.95, df1 = 7, df2 = 392) = 2.685086, We reject H0 if F >= 2.685086, else FAIL to reject
   From the summary() function of the model, the F-statistic value is 228.9. So, we REJECT H0 since 228.9 > 2.685086
5. Since we rejected the H0, the overall model is significant. So we run an inference t-test for each of the parameters to determine the relative contribution of each factor. Here we test at α = 0.05 and determine the decision rule to reject H0 if |t| >= 1.965927, else FAIL to reject.
   For each variable, below is the calculated t statistic:
   GRE = 0.0017374 / 0.0005979 = 2.905837097 (Reject H0)
   TOEFL = 0.0029196 / 0.0010895 = 2.67976135842 (Reject H0)
   University Rating = 0.0057167 / 0.0047704 = 1.19836910951 (FAIL to Reject H0)

SOP = -0.0033052 / 0.0055616 = -0.59428941311 (FAIL to Reject H0)
LOR = 0.0223531 / 0.0055415 = 4.03376342146 (Reject H0)
CGPA = 0.1189395 / 0.0122194 = 9.73366122723 (Reject H0)
Research = 0.0245251 / 0.0079598 = 3.08112012865 (Reject H0)

6. Calculate the Confidence Interval at the 95% level:

GRE.Score	0.0005619253	0.002912898
TOEFL.Score	0.0007775204	0.005061633
University.Rating	-0.0036621610	0.015095477
SOP	-0.0142395495	0.007629211
LOR	0.0114583790	0.033247876
CGPA	0.0949156272	0.142963280
Research	0.0088759540	0.040174259

7. Check the 4 assumptions of the Multiple Linear Regression via plotting the Model (See below at number 7 for each plot used to test assumptions):
    a. Linearity
    We can confirm Linearity by referring to the Residuals vs Fitted plot. The line of best fit forms a straight line. Here, it is not a perfect line, a subtle rise but still within acceptable range to be considered linear.
    b. Independence
    They are independent due to the residuals vs fitted plot being scattered. This means the result of an observation is not affected by the results of the previous datapoint.
    c. Homoscedasticity
    Using the Scale-Location plot, we can see the datapoints form a funnel shape. This indicates a violation of the homoscedasticity.
    d. Normality
    Use the Normal Q-Q plot. Based on the plot, an estimated +-40% of the 400 data points are not on or near the dotted line, all of which reside at the lower end of the plot. Due to this deviation, the Normality assumption is also considered to be violated.

6. State Your Conclusion (no more than 100 words)

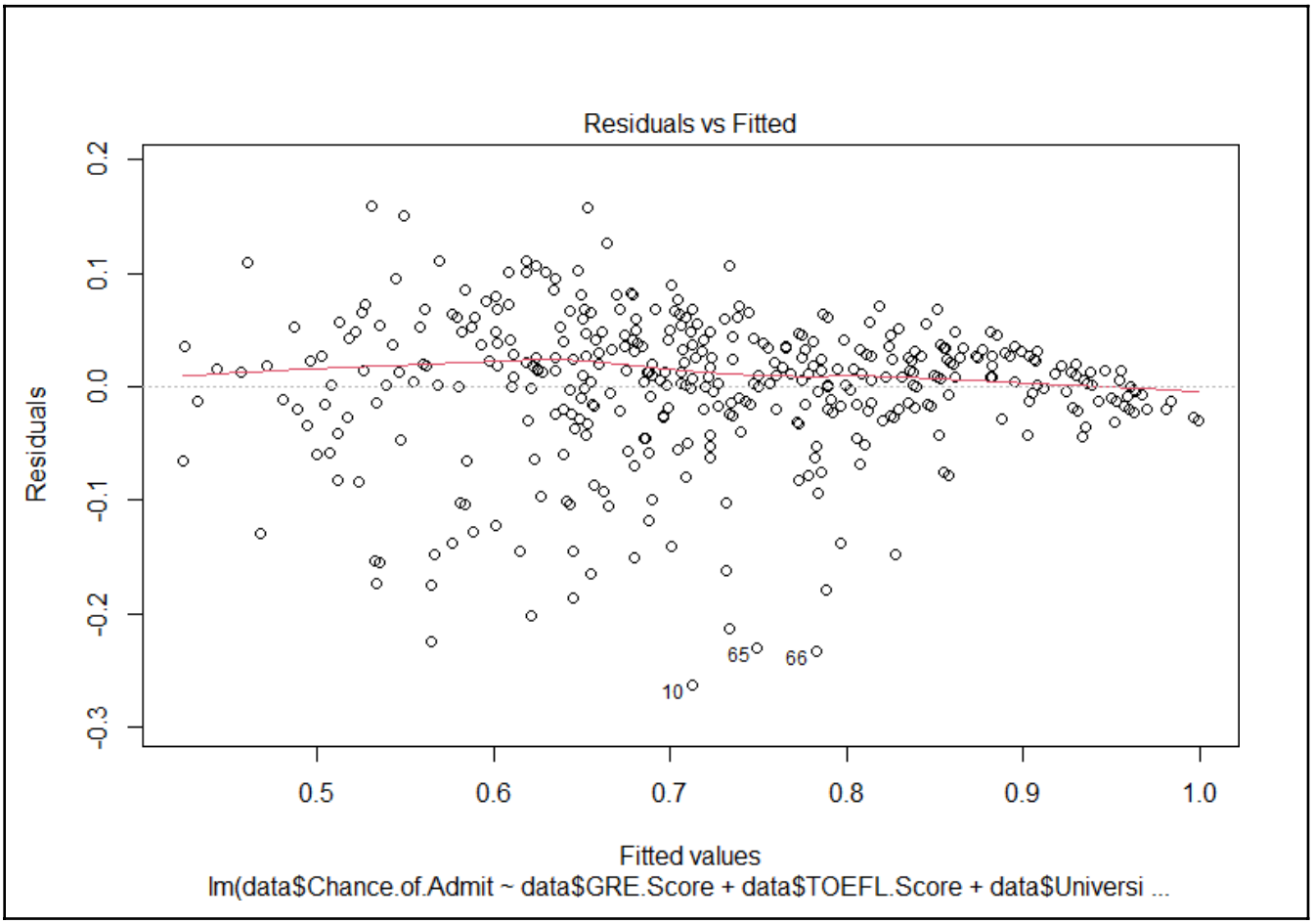
State the conclusion(s) of your analysis in a way so that a non-statistician can understand.

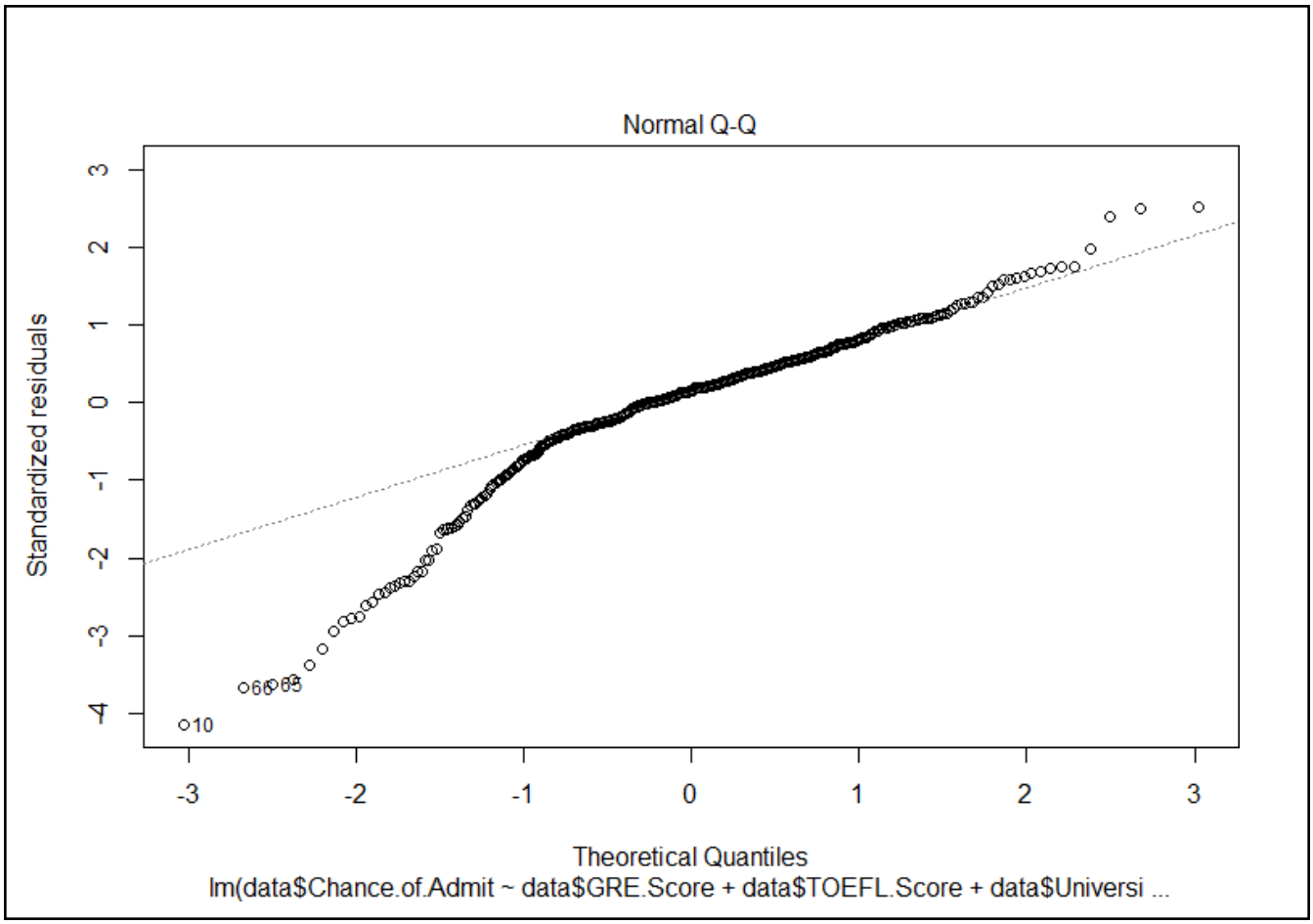
Based on the analysis, 2 out of the 4 assumptions for Multiple Linear Regression have failed (Homoscedasticity and Normality). These violations potentially damage the interpretations about whether each factor considered when applying for postgraduate education at a university has equal or

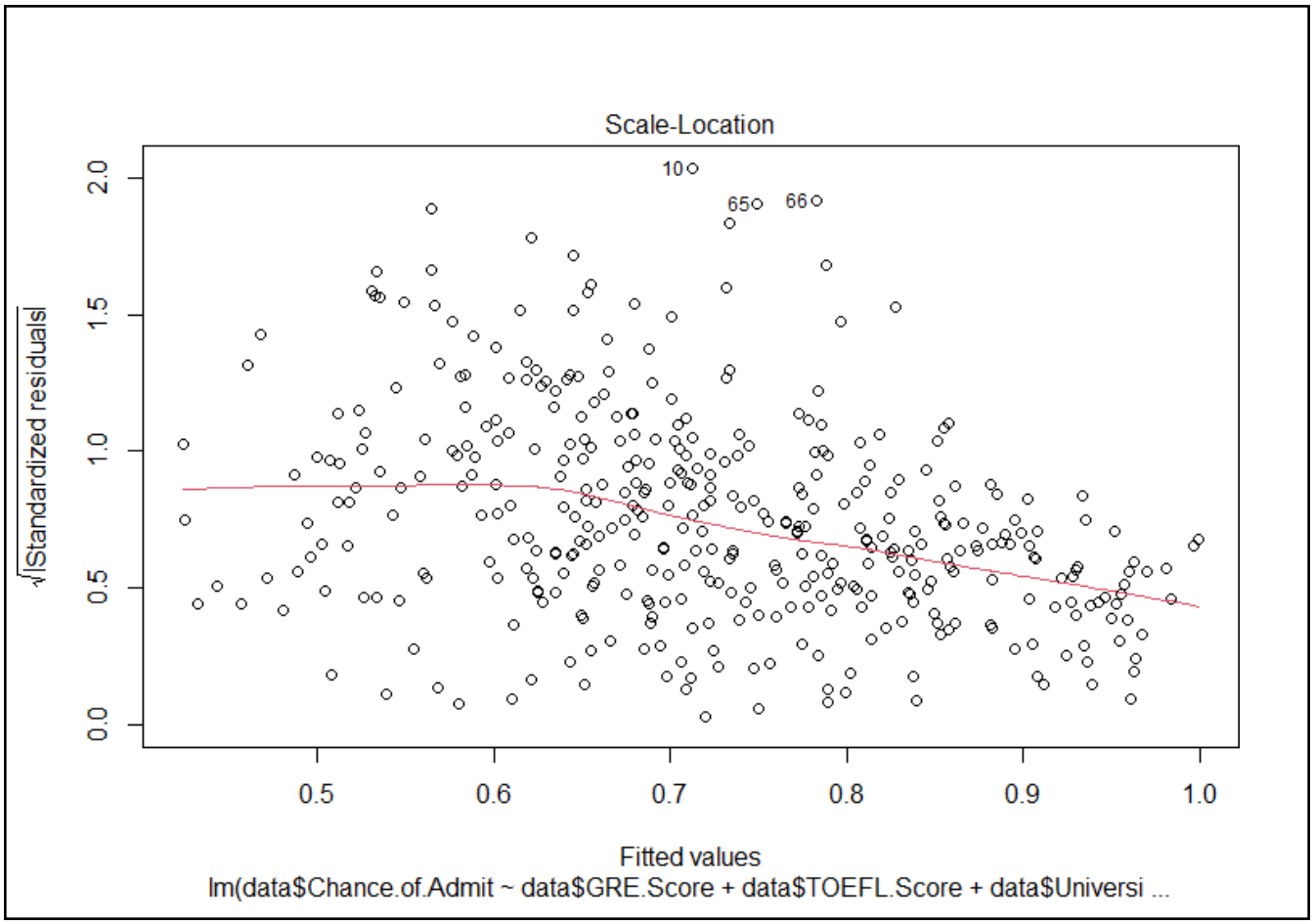
unequal association with probability of being admitted to a randomly selected university. It is highly recommended to consider alternative approaches to overcome the violations, including but not limited to transforming the attributes or using a more robust regression method that are less sensitive to violations of assumptions.

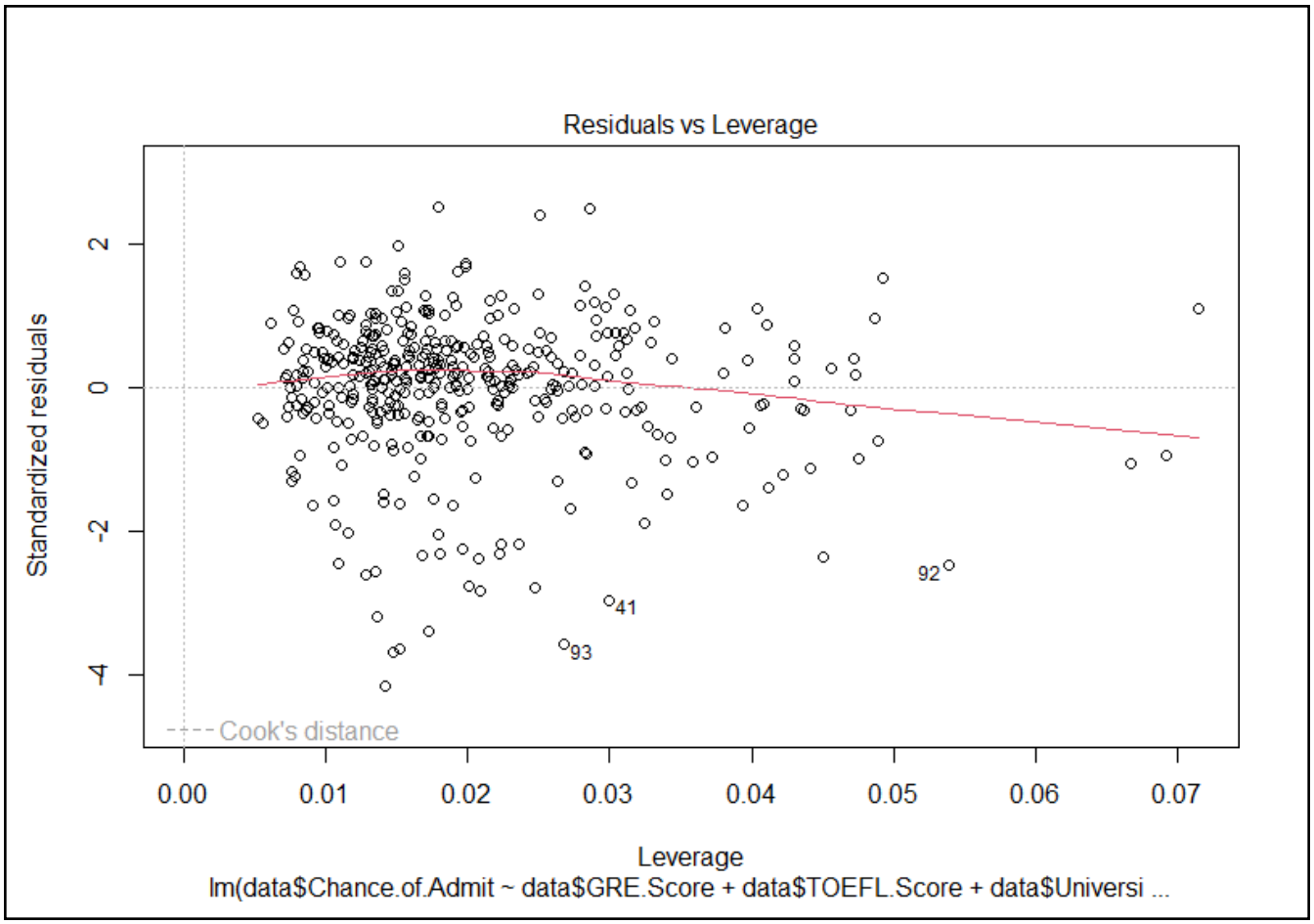
7. Plots Used

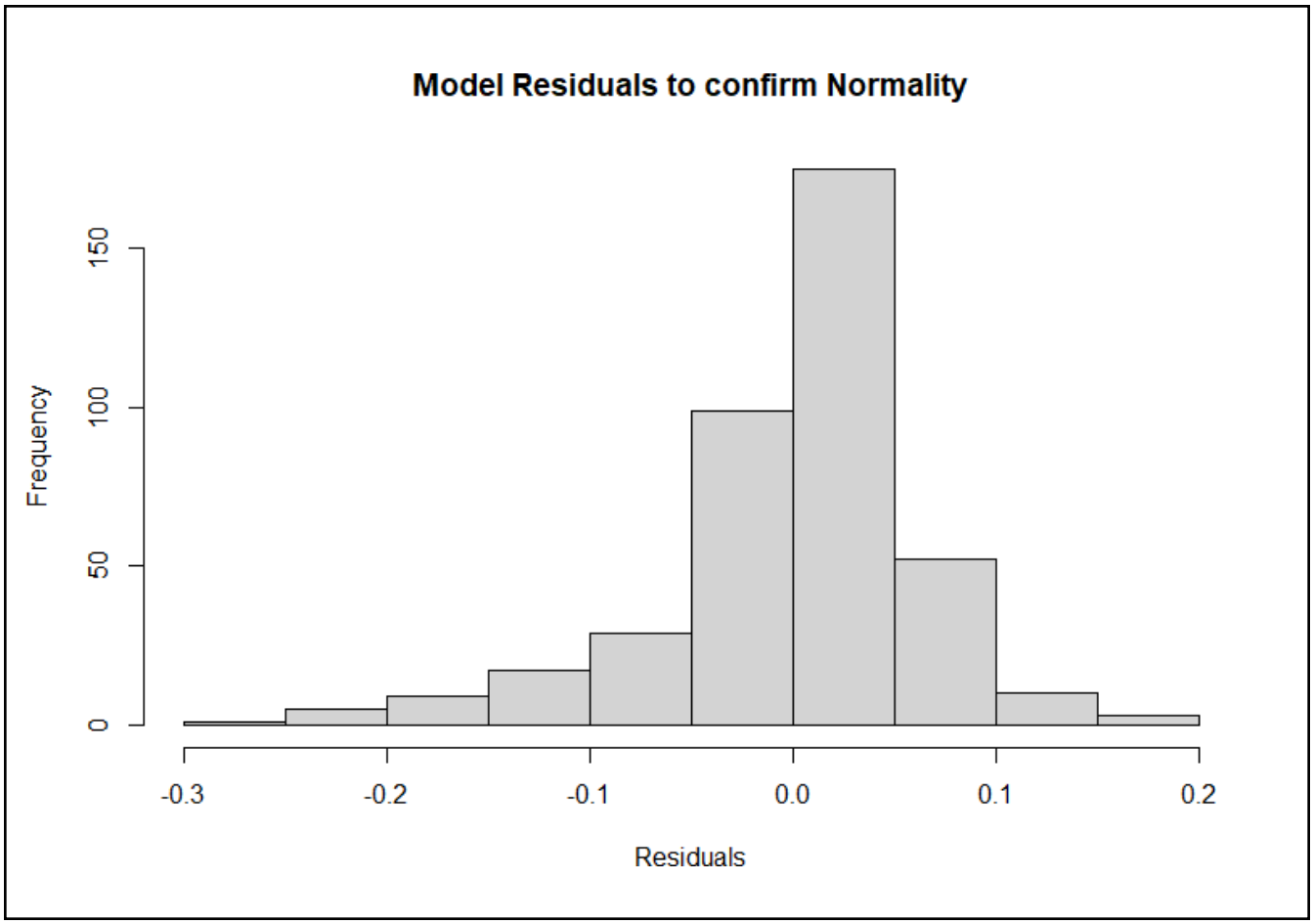
Below are all the plots used for analysis




Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(data$Chance.of.Admit ~ data$GRE.Score + data$TOEFL.Score + data$Universi ...
10


Scale-Location
10
65
66
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0.5
0.6
0.7
0.8
0.9
1.0
Fitted values
lm(data$Chance.of.Admit ~ data$GRE.Score + data$TOEFL.Score + data$Universi ...


Residuals vs Leverage
Standardized residuals
2
0
-2
-4
Cook's distance
92
41
93
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
Leverage
lm(data$Chance.of.Admit ~ data$GRE.Score + data$TOEFL.Score + data$Universi ...



Solution Submission

1. **Fill up this word file and upload it.**
2. **Upload your data set. This is the data set after cleaning (a small CSV file)**
3. **Submit R code as either a well-commented .R file or as a .Rmd file with associated .html file.**

Grading will be done based on

1. **Originality of selected data set and data analysis approach**
2. **Data Preparation set and cleanup**
3. **General Correctness of data analysis**

4. **Quality of your R code and output results**
5. **Correct final conclusion and useful visualization**