What really affects a student's performance during COVID?

Ritik Sharma (1005159187) and Inderjeet Punia (1005102096)

22/08/2021

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Warning: NAs introduced by coercion
```

```
## Warning: NAs introduced by coercion
```

```
## Warning: NAs introduced by coercion
```

SECTION 1: INTRODUCTION

In this report, we are analyzing data collected by STA302 students throughout the semester. In this data set, we have several predictor variables including the grades students received on each quiz, the country they currently reside in, the number of hours they spend studying for this course on a weekly basis and the number of hours students think about COVID-19 weekly. The goal of this project is discover how our predictor variables in our model can affect a student's performance in Quiz 4. In the context of this project, the grade a student receives in Quiz 4 is the dependent variable. The possible predictor variables we have are the independent variables. Furthermore, we aim to develop a model for each region which can be used by individuals to predict their performance on Quiz 4.

As a first step, we will clean the data. This included removing any rows that had a n/a value in it. Then, we will split the data according to regions. Our regions were North America and Asia. There will be two different sets of models based on these regions. As preliminary analysis, we will plot the data using pair plots and do comparisons. By doing this we will be able to tell the distributions of our independent variables. We will also normalize any data that is not following a normal distribution. Once we are sure all the data points are adequate we will build our multi-linear regression model. The main goal of this study is to see which of the predictor variables can actually predict the values of a student's performance in Quiz 4. To achieve this, we will use backward stepwise regression. Furthermore, our models will be verified by checking all of the violations of assumptions so that we make sure we do not have variables that are not needed.

SECTION 2: EXPLORATORY DATA ANALYSIS

SECTION 2.1: Variables of Interest

The dataset we are working with in this project has been renamed to "covid_data" for convenience. The cohort of this dataset were STA302 students enrolled in the 2021 summer semester. There are 227 observations and 14 variables in this data set. From these variables, 11 of them could be used as potential predictors to predict the grade a student receives on quiz 4. But, we would not be using these predictors as is. Instead, we would be aggregating similar variables into mean-averages. We would find the mean study time before quiz 4 for each student and place it in a new "study_avg" variable. Similarly, we would find the mean time

a student thought about covid and place it in a new "covid_hours_avg" variable. Finally, we would also find the mean of the first three quizzes for each student and place it in a new "quiz_avg" variable. This was done because these variables would be a more accurate representation of the initial predictors. These new predictors would also make our model less complex as there are less predictors. We also created new dataframe named "clean_covid_data". This only included the variables we were interested in. These were X1(id), region, quiz_average, covid_hours_avg, study_avg, and Quiz_4_score.

SECTION 2.2: Seperating Data by Region

We split our new dataset (clean_covid_data) into two new dataframes according to the region the student was located in during this semester. The two regions that we included were North America and Asia. We separated the data by region because different regions have different COVID-19 situations. With this information, we intend to develop two different models for the two different regions. We will be using the "asia_covid" and "north_america_covid" dataframes to do so.

SECTION 2.3: Summary Statistics for Variables of Interest

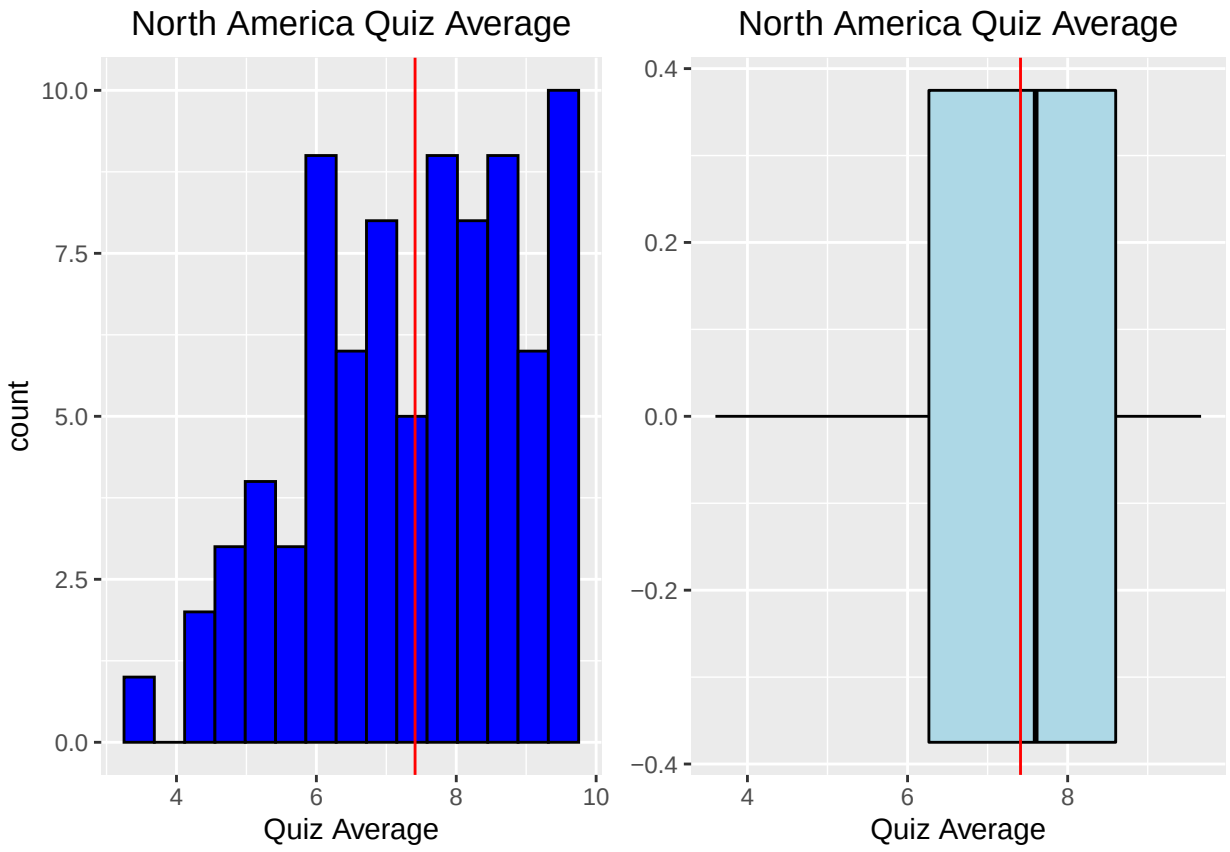
In this section, we intend to display important summary statistics for our variables of interest for both regions.

Quiz Average

North America

```
## Warning: Use of ‘north_america_covid$quiz_avg‘ is discouraged. Use ‘quiz_avg‘
## instead.

## Warning: Use of ‘north_america_covid$quiz_avg‘ is discouraged. Use ‘quiz_avg‘
## instead.
```



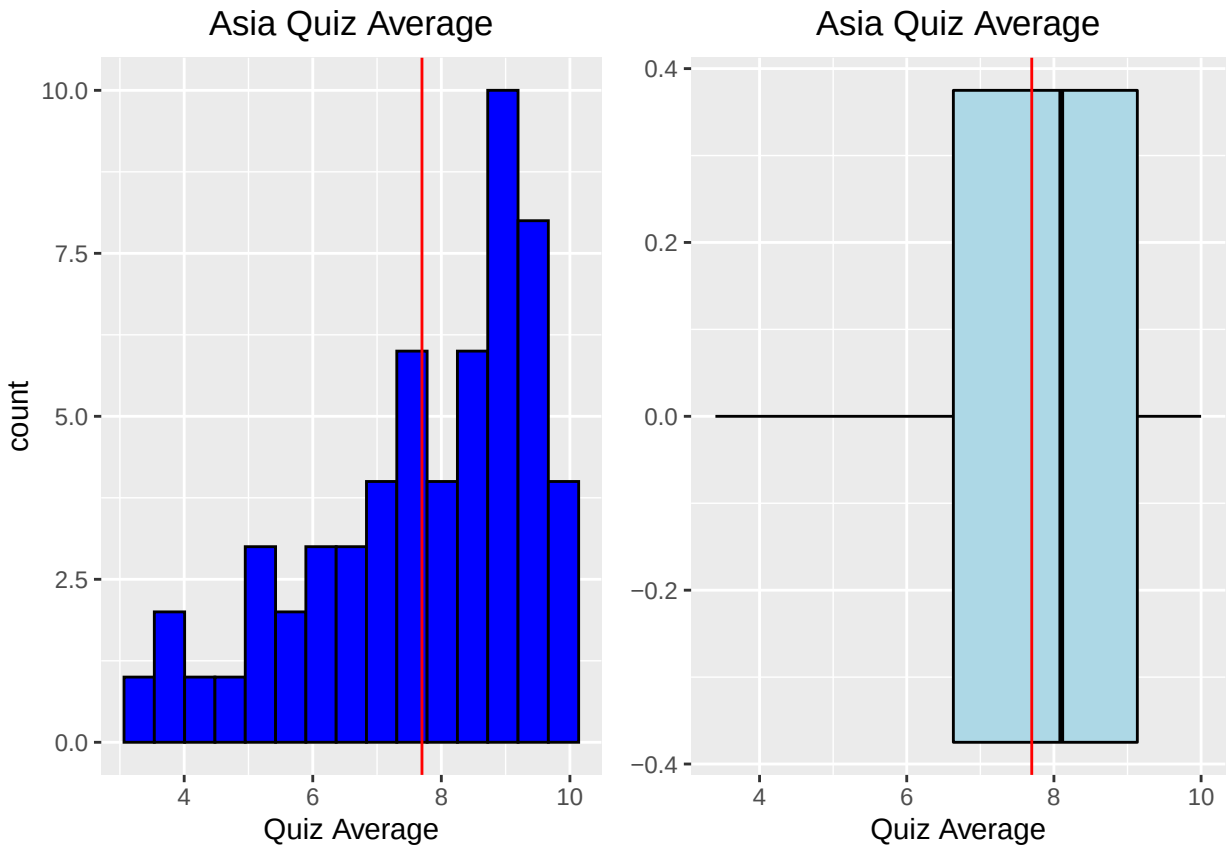
These visualizations for North America's Quiz average tell us the following:

- The mean is **7.411245**.
- The median is **7.6**.
- The minimum is **3.6**.
- The maximum is **9.6666667**.
- The first quartile is **6.2666667**.
- The third quartile is **8.6**.
- The interquartile range is **2.3333333**.

Asia

```
## Warning: Use of ‘asia_covid$quiz_avg‘ is discouraged. Use ‘quiz_avg‘ instead.

## Warning: Use of ‘asia_covid$quiz_avg‘ is discouraged. Use ‘quiz_avg‘ instead.
```



These visualizations for Asia's Quiz average tell us the following:

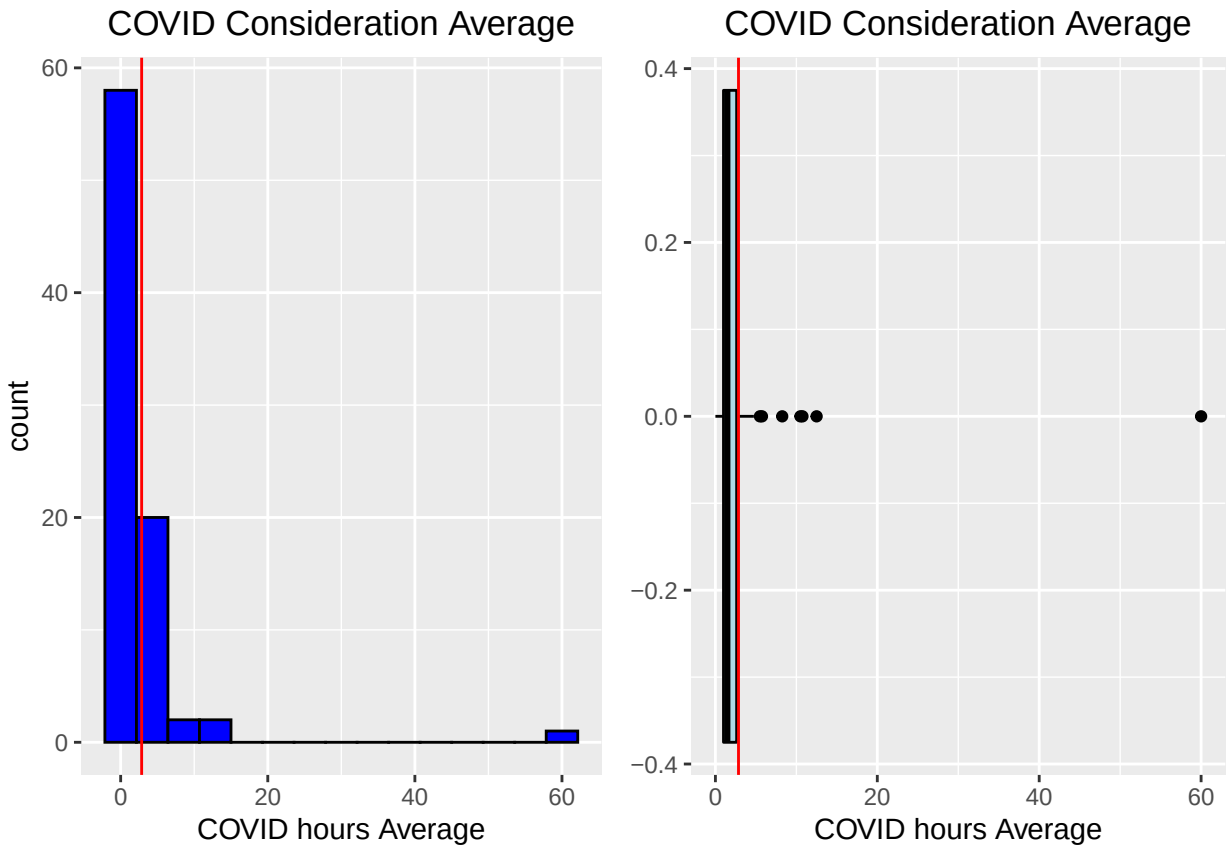
- The mean is **7.6988506**.
- The median is **8.1**.
- The minimum is **3.4**.
- The maximum is **10**.
- The first quartile is **6.6333333**.
- The third quartile is **9.1333333**.
- The interquartile range is **2.5**.

Covid Hours Average

North America

```
## Warning: Use of ‘north_america_covid$covid_hours_avg‘ is discouraged. Use
## ‘covid_hours_avg‘ instead.
```

```
## Warning: Use of ‘north_america_covid$covid_hours_avg‘ is discouraged. Use
## ‘covid_hours_avg‘ instead.
```



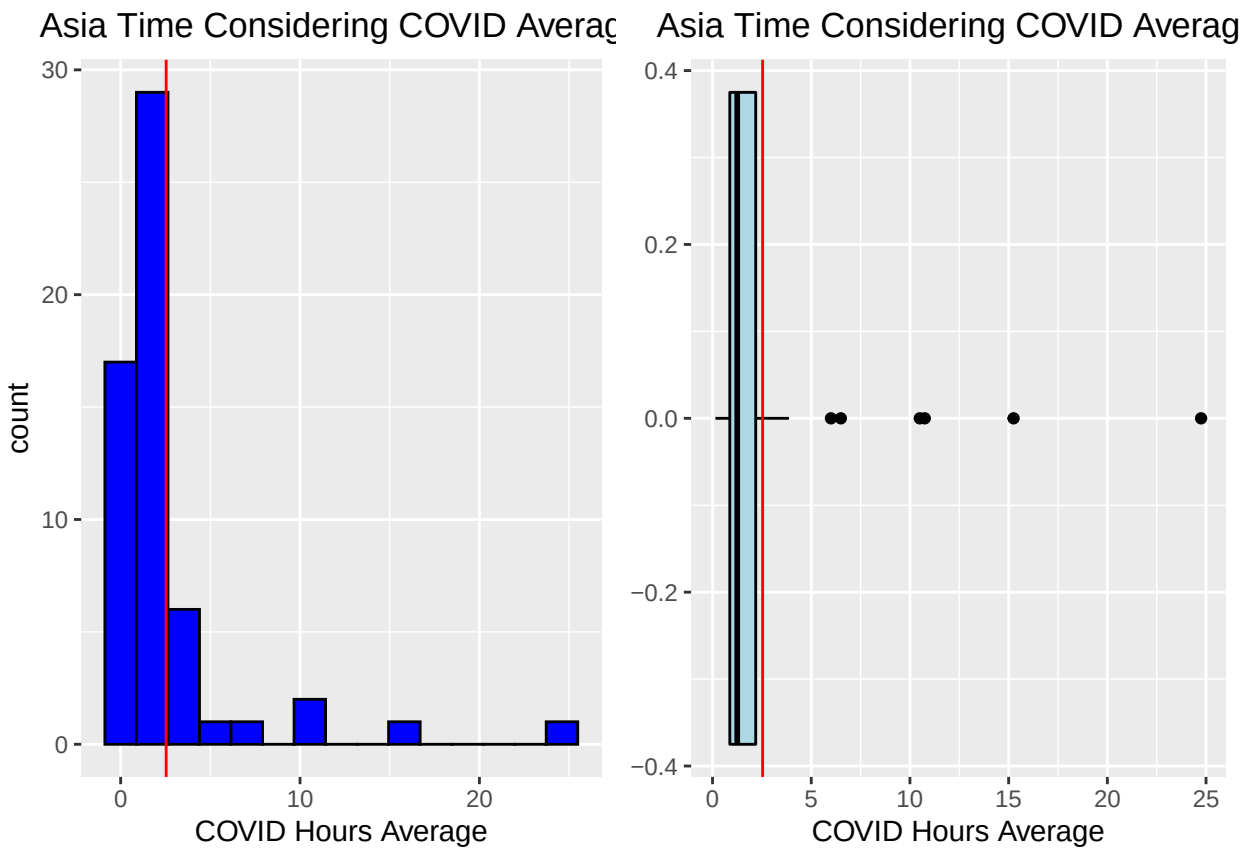
These visualizations for North America's time considering COVID average tell us the following:

- The mean is **2.8523202**.
- The median is **1.5**.
- The minimum is **0**.
- The maximum is **60**.
- The first quartile is **1**.
- The third quartile is **2.625**.
- The interquartile range is **1.625**.

Asia

```
## Warning: Use of ‘asia_covid$covid_hours_avg‘ is discouraged. Use
## ‘covid_hours_avg‘ instead.

## Warning: Use of ‘asia_covid$covid_hours_avg‘ is discouraged. Use
## ‘covid_hours_avg‘ instead.
```



These visualizations for Asia's time considering COVID average tell us the following:

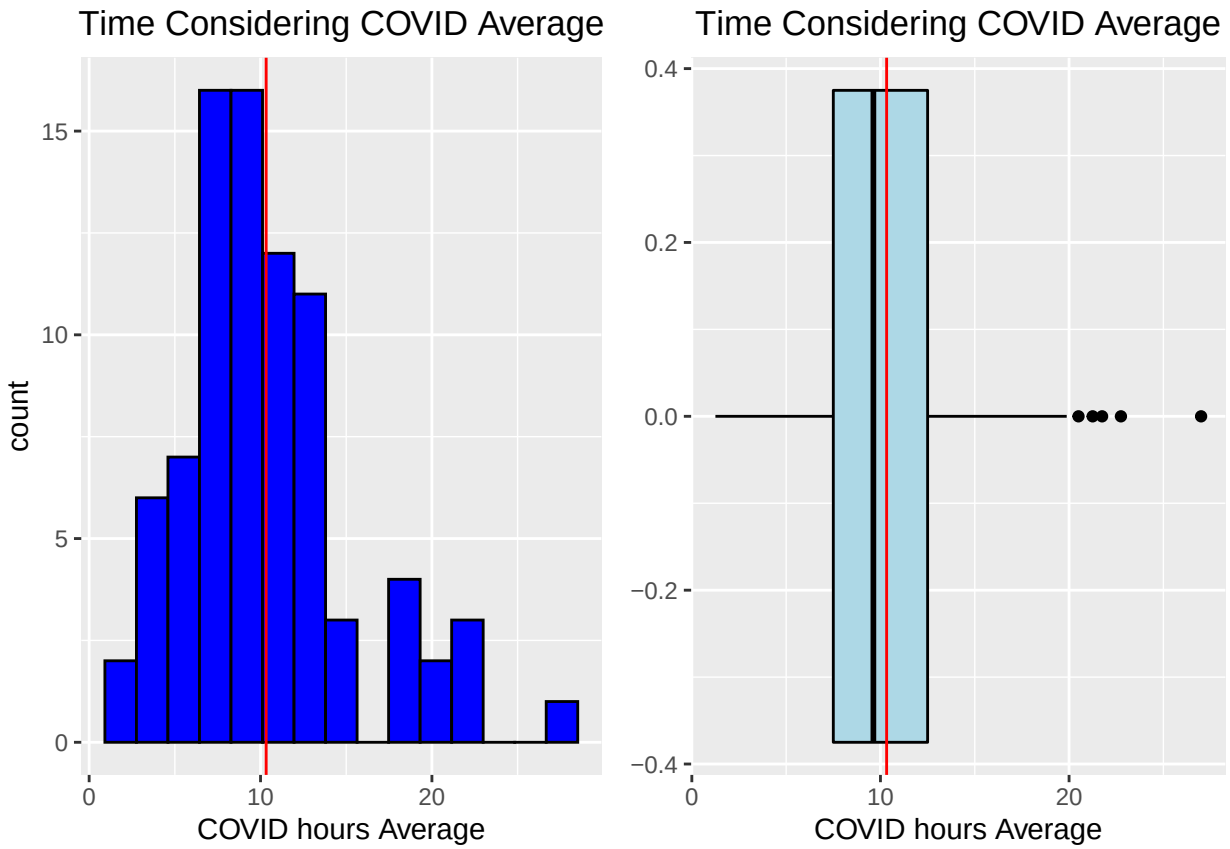
- The mean is **2.5360616**.
- The median is **1.25**.
- The minimum is **0.14575**.
- The maximum is **24.75**.
- The first quartile is **0.875**.
- The third quartile is **2.1875**.
- The interquartile range is **1.3125**.

Study Hours Average

North America

```
## Warning: Use of ‘north_america_covid$study_avg‘ is discouraged. Use ‘study_avg‘
## instead.

## Warning: Use of ‘north_america_covid$study_avg‘ is discouraged. Use ‘study_avg‘
## instead.
```



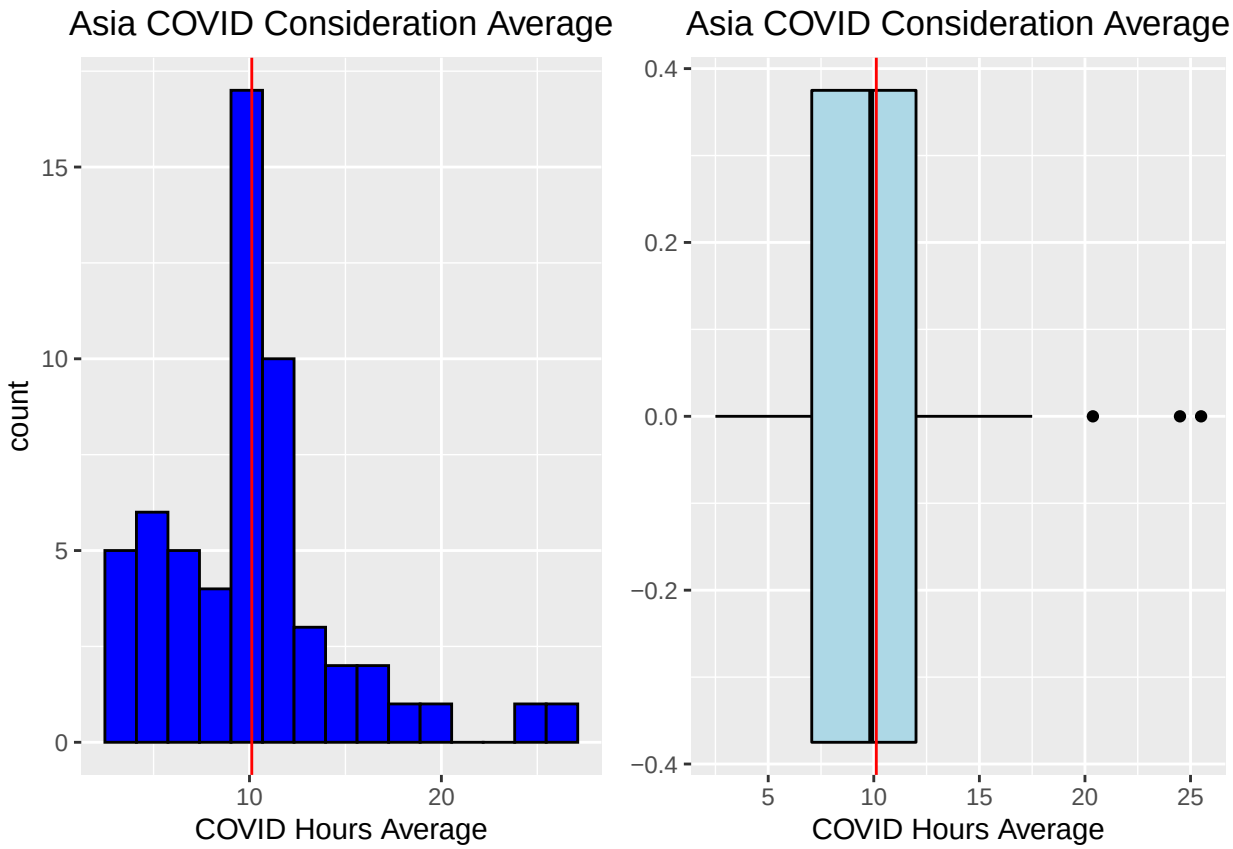
These visualizations for North America's study hours average tell us the following:

- The mean is **10.3283133**.
- The median is **9.625**.
- The minimum is **1.25**.
- The maximum is **27**.
- The first quartile is **7.5**.
- The third quartile is **12.5**.
- The interquartile range is **5**.

Asia

```
## Warning: Use of ‘asia_covid$study_avg‘ is discouraged. Use ‘study_avg‘ instead.

## Warning: Use of ‘asia_covid$study_avg‘ is discouraged. Use ‘study_avg‘ instead.
```



These visualizations for Asia's study hours average tell us the following:

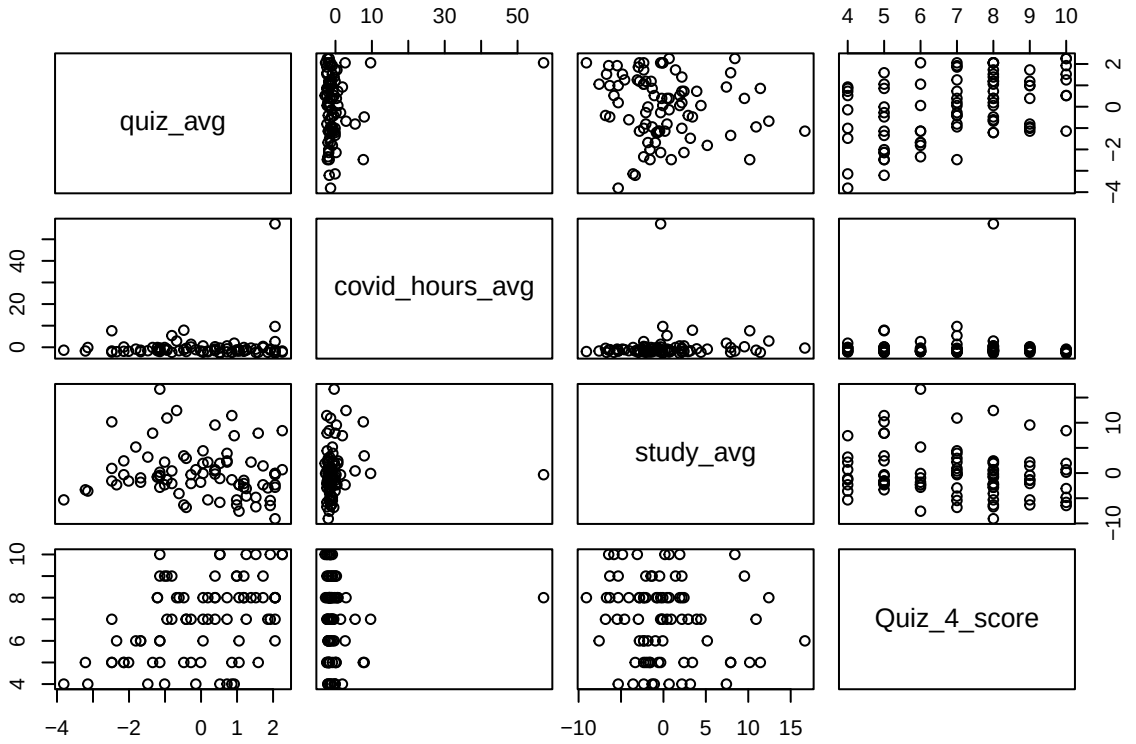
- The mean is **10.1185345**.
- The median is **9.875**.
- The minimum is **2.5**.
- The maximum is **25.5**.
- The first quartile is **7.0625**.
- The third quartile is **12**.
- The interquartile range is **4.9375**.

SECTION 2.4: Transformation of Data

We decided to center each of our variables of interest around its respective mean. This did two things for us. First, we a new interpretation of the intercept. The new intercept is the mean of the response when all of the predictors have a value of 0. Secondly, the slope between the predictors and the response did not change. It still has the same value as it had before the mean-center transformation was applied.

SECTION 2.5: Correlation

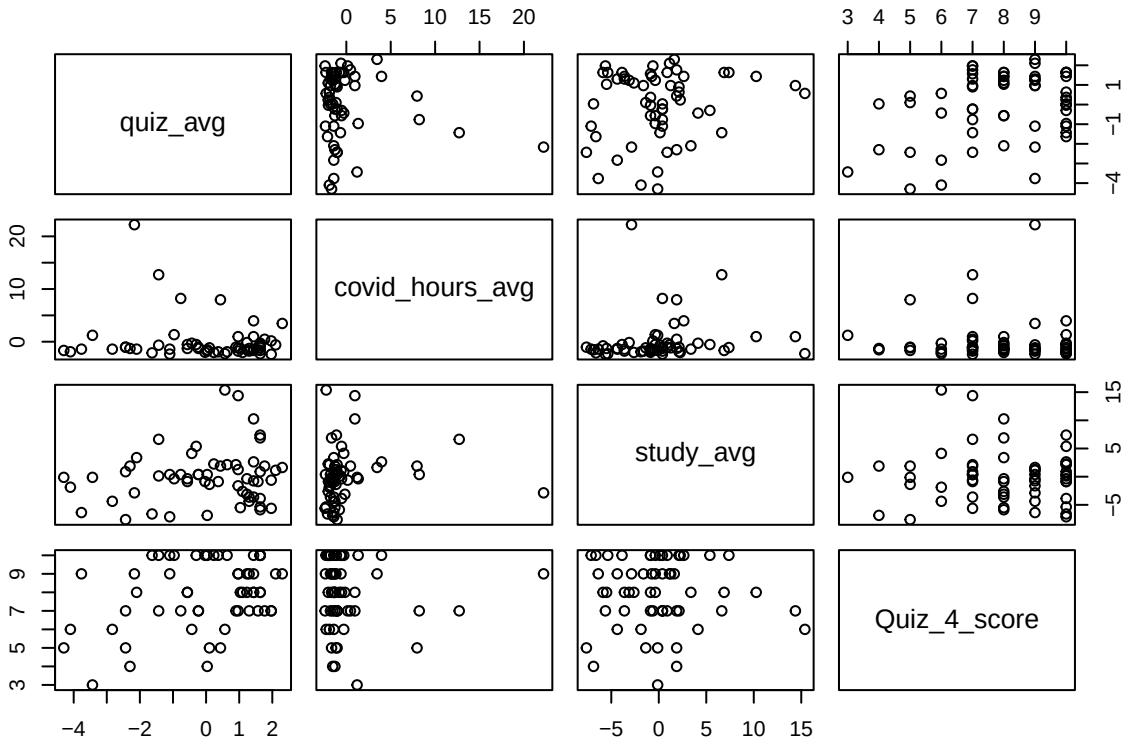
North America




```
## NULL
```

From the pair-plot of the North American data, we can see that no two predictors are highly correlated with each other. This is a good as there is no multicollinearity among our predictors. Furthermore, we can see there is significant correlation between "Quiz_4_score" and "quiz_avg". This tells us that "Quiz_4_score" and "quiz_avg" are significantly related.

Asia




```
## NULL
```

From the pair-plot of the Asian data, we can see that no two predictors are siginificantly correlated with each other. This tells is that there is no multicollinearity among the predictors. We also noted there is significant correlation between "Quiz_4_score" and "quiz_avg". This tells us that "Quiz_4_score" and "quiz_avg" are significantly related.

SECTION 2.6: Finding Influential Points

In this section we aim to find influential points in our datasets. We will continue to analyze the Asian dataset and the North American dataset separately. The influential points would be removed if necessary. We are doing this to increase the stability of our model as influential points affect the model output the most. Influential points are not the same as outliers. We are removing influential points because they have higher leverage than outliers. We will be using Cook's distance to identify influential points. Points that have a Cook's Distance greater than 3 times the mean of Cook's distances are considered influential.

North America

As a starting point, we will make a linear model including all of our predictors. This will be called "base_north_america_model". We aim to remove any data points that align with Cook's Distance greater than 3 times the mean of Cook's distances criteria. Upon inspection, only 1 data point fit this criteria. It was removed from our dataset. From now on, we will be working with a new dataframe called

"north_america_covid_clean". This dataframe does not include that point. We realized that is an influential point (X1 = 137) and should be removed as the input for COVID Hours (W1) was 160 hours which is cretainly miseading data.

Asia

As a starting point, we will make a linear model including all of our predictors. This will be called "base_asia_model". We will remove any data point that has a Cook's Distance greater than 3 times the mean value of Cook's distances. Upon inspection, there were 3 data points that met this criteria. These were removed. This can possibly be accredited to a mistake in data entry or the point just being an outlier. Taking a closer look, we realize that the average time spent considering COVID for one of these points was abnormally high. We realized that it was high because one of the student's input for COVID Hours (W4) was 60 hours which is misleading data. The data point X= 86 was removed because the student's average study time was 3.25 hours and this can't be true as every student spent at least 6 hours studying which watching the course lectures.

SECTION 3: MODEL DEVELOPMENT

SECTION 3.1: Backward Stepwise Regression

In our analysis, we used backward stepwise regression to come up with our final model. We started with a model that had all three predictor variables. These three predictor variables were the average time a student spent studying during weeks 1-4, the average time a student spent thinking about COVID, and the average of the first three quizzes. To check the goodness of the fit of our model, we used Akaike information criterion (AIC). We used backward stepwise regression separately on both the North American dataset and the Asian dataset. We chose to use backward stepwise regression because we wanted to see which variables were significant. The easiest way to do this was to check the significance level in the summary table for the each model.

North America

Using backward stepwise regression, three models were made for the cleaned North America dataset. The first model included all three predictors namely quiz_avg, covid_hours_avg, study_avg. Upon calling the summary of this model, we realized that covid_hours_avg and study_avg were not significant variables. The adjusted R-squared value for this model was 0.1872. The AIC value for this model was 86.5034. Since covid_hours_avg and study_avg were not significant, we decided to remove study_hours from our model. Now, our second model only included covid_hours_avg and quiz_avg as predictors. The adjusted R-Squared value for this multivariate model was 0.1911. This was higher than the previous model so this meant that we were moving in the right direction. The AIC value also decreased to 85.1500. This meant that the goodness of fit of our model was improving. Finally, we decided to remove covid_hours_avg as a predictor variable as well. We did this because in the summary statistics for our model, this variable was insignificant. Our final model only included quiz_avg as a predictor variable. This model had a R-Squared (not adjusted R-Squared as this is a bivariate model) value of 0.1937. The AIC value for this model was 84.9385. The R-Squared value increased with this model and the AIC value decreased. Both of these were good signs and we knew we had the best possible model for our North America Data. The table below summarizes our findings for all three models. The code for model development can be found in the Appendix.

The final model for North America was:

$$Quiz4Grade = 6.977 + 0.547 * studyaverage + \epsilon$$

Below is a summary table for the chosen North America linear regression model.

Model Parameters	R-Squared	Adjusted R-Squared	AIC
study_avg, covid_hours_avg, quiz_avg	0.2173	0.1872	86.5034
covid_hours_avg, quiz_avg	0.2111	0.1911	85.1500
quiz_avg	0.1937	0.1836	84.9385

Figure 1: North America Table

Characteristic	Beta	95% CI	p-value
quiz_avg	0.55	0.30, 0.79	<0.001

Asia

Similar to the North American models, backward stepwise regression was used to develop three models for the cleaned Asia dataset. The first model included all three predictors namely quiz_avg, covid_hours_avg, study_avg. Upon calling the summary of this model, we realized that covid_hours_avg and study_avg were not significant variables. The adjusted R-squared value for this model was 0.0984. The AIC value for this model was 53.9105. Since covid_hours_avg and study_avg were not significant, we decided to remove study_hours from our model. Now, our second model only included covid_hours_avg and quiz_avg as the independent variables. The adjusted R-Squared value for this multivariate model was 0.1011. This was higher than the previous model so this meant that we were moving in the right direction. The AIC value also decreased to 52.8181. This meant that the goodness of fit of our model was improving. Finally, we decided to remove covid_hours_avg as an independent variable as well. This was done because in the summary statistics for our model, this variable was insignificant. Our final model only included quiz_avg as an independent variable. This model had a R-Squared (not adjusted R-Squared as this is a bivariate model) value of 0.1057. The AIC value for this model was 52.6070. The R-Squared value increased with this model and the AIC value decreased. Both of these were good signs and we knew we had the best possible model for our Asia Data. The table below summarizes our findings for all three models. The code for model development can be found in the Appendix.

Model Parameters	R-Squared	Adjusted R-Squared	AIC
study_avg, covid_hours_avg, quiz_avg	0.1485	0.0984	53.9105
covid_hours_avg, quiz_avg	0.1344	0.1011	52.8181
quiz_avg	0.1057	0.0889	52.6070

Figure 2: Asia Table

The final model for Asia was :

$$Quiz4Grade = 8.019 + 0.323 * studyaverage + \epsilon$$

Below is a summary table for the chosen Asia linear regression model.

Characteristic	Beta	95% CI	p-value
quiz_avg	0.32	0.07, 0.58	0.015

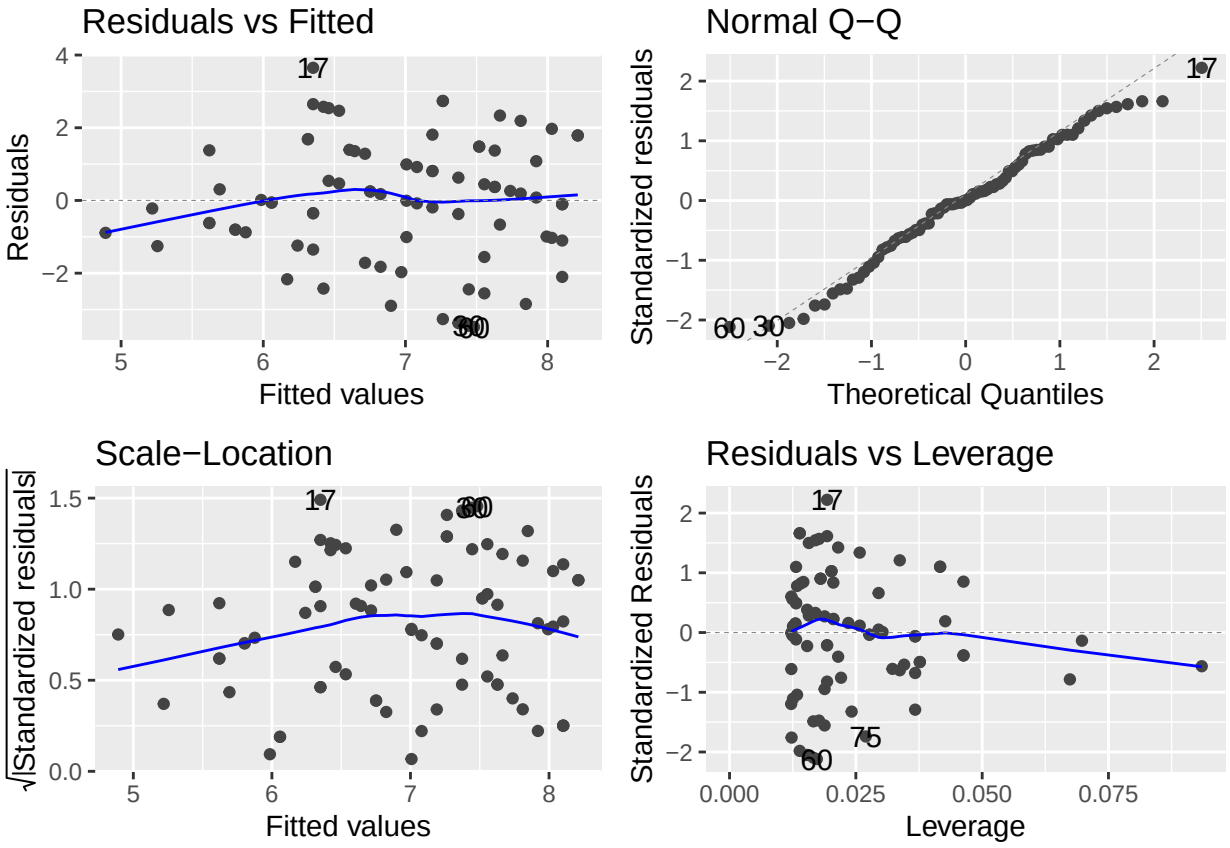
4. The slope estimate term for avg_study was 0.547. This means when avg_study goes up by 1 hour, the grade a student receives on Quiz 4 goes up by 0.547.

Asia

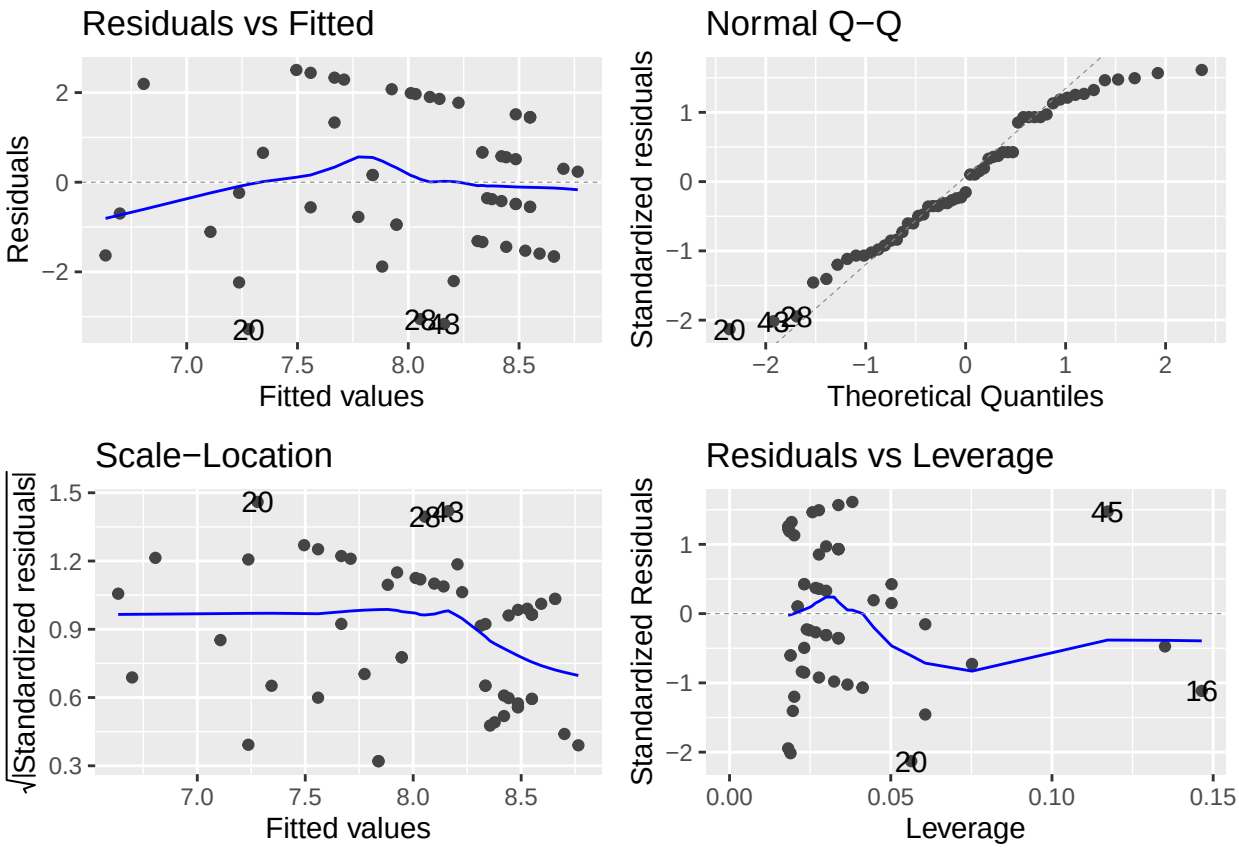
The slope intercept for the final Asia model was 8.019. Our data was mean centered. So, this means when the average hours studied is zero, a student is expected to receive a grade of 8.019 on Quiz 4. The slope estimate term for avg_study was 0.323 This means when avg_study goes up by 1 hour, the grade a student receives on Quiz 4 goes up by 0.323.

SECTION 3.3: Assumptions of Linear Regression

North America



Asia



- Talk about the assumptions of linear regression models. These include linearity (check using scatterplot, if it is not linear then we need to transform the data), homoscedasticity, no multicollinearity (check for this the way we did in GGR276. Use VIF (Variance Inflation Factor)), no autocorrelation (check using Durbin-Watson Statistic), normality of residuals(this can be checked using the goodness of git test. if data is not normally distributed then apply log transformation)

SECTION 4: CONCLUSION

The goal of this project was too determine the impact our predictor variables had on our response variable. Our predictor variables were quiz averages, the average amount of studying a student did prior to quiz 4, and the average time a student spent thinking about COVID 19. Our response variable was the grade a student received on quiz four. We aimed to identify a relation or determine the impact of these three variables on quiz 4 grades, varying by region.

We created various models for the two separate regions(Asia and North America). In the first region of North America, three models were created using a total of three variables. The reliability of these three variables was determined from tests run on the models. In all of the models the p value for quiz_average was always the lowest meaning that this predictor was important in our models. Secondly, the p value for study_avg was low, but not low enough that we can say with full confidence that it had a major impact on the grades of quiz four. The covid_hours_avg predictor was shown to be insignificant in many different models due to the p value being very large. The results were similar in the Asian modelling as well.

Based on our analysis, we were able to conclude that the only predictor necessary to predict a student's performance on Quiz 4 was the average of the previous three quizzes. This was quite surprising as we were

expecting COVID to have an influence in at least one of the regions but that was not the case. However, the two regions had different models. The slope intercept for each was different as was the coefficient for average_quiz.

Work Breakdown

- Ritik Sharma worked on the introduction, Data Cleaning, Exploratory Data Analysis, and the Interpretation of final model (part of Model Development), and Model Development.
- Inderjeet Punia worked on Model Development and the Conclusion
- Note: Model Development was done together as a team.

APPENDIX

Code for Model Development