QMM1002 Case Study 2 [20%]

Krishno Sarkar

A00246996

Due: April 21, 2022 at 11:59PM

Introduction

In QMM 1001, I started collecting my personalized data on different daily activities which I perform during the day like my zoom meeting time (in classes or group meetings), study time, outings, screen time, sleep time etc. In QMM 1002 this semester (winter 2022), I continued collecting those personalized to understand the pattern of my daily activities, how I plan to spend my daily activities and actually how I end up spending my daily activities.

During my full-time college, my daily activities are more of a routine like getting up early, doing my daily exercises, sit down for studies, attend zoom classes as per schedule, prepare food (when required), complete assignments (as planned), complete everyday routine activities etc. During weekend, my schedule remains almost same except for weekly groceries and occasional outings.

My personalized data is collected where I had recorded my time spent across daily activities from 10th September 2021 till 14th April 2022(191 days). The personalized data is collected across 9 different variables (as listed below) which shows the time spent on major daily activities and my mental well-being. The 10th variable is added to note the difference in my daily activities during Fall 2021 and Winter 2022.

Variable	Description	Type
Date	Date of observation	Identifier
Zoom	Time spent on Zoom Meetings or Classes	Quantitative
Study	Time spent on doing assignments	Quantitative
Sleep	Sleep Time	Quantitative
House	Number of times went out of House	Quantitative
News	Did I watch News Today?	Categorical
Exercise	Did I do Exercise Today?	Categorical
Mood	What is my morning mood today?	Categorical
Screen	Time spent on entertainment today	Quantitative
Semester	My Semester	Categorical

In addition to personalized data set, I will analyze and compare with another data set i.e. combined data set. The data set provides total 7537 observations from September 2019 to February 2022 of students spending time studying in different programs of study (BAPG, CAGC, HAGC). Following table shows the variables in combined data set.

Variable	Description	Type
Date	Date of Observation	Identifier

Variable	Description	Type
Hours Studying	Time spent studying	Quantitative
Semester	Term (F19, W20, F21 or W22)	Categorical
Program	Program enrolled (BAPG, CAGC, HAGC)	Categorical

The summary statistics of both groups are given below.

Summary Statistics

These summary statistics can be used to compare the amount of time I spent studying compared to other students in the program.

Data Set	Mean	Standard Deviation
Krishno	5.2680628	2.4811765
All Students	3.6007841	2.3552081
BAPG	3.5428587	2.3661103
CAGC	4.079602	2.545816
HAGC	3.8118149	2.1271117

The above mean shows that my Average ‘Study Hours’ is more than other students in the program. However, the distribution shows that Crime Analytics students have generally higher average than the other two streams (‘Business’ and ‘Health’ Analytics) students. However, Crime Analytics students Standard Deviation is also more than other stream students. Hence to actually find out if there is any significant difference in the three streams of students, I will be answering the three questions below.

I will use my personalized data set and a combined class data set to answer the following questions: 1. Are there differences in the average study times for students in the different analytics streams? 2. Is the distribution of days studied more than 3.13 hours (the average daily study time for students at McGill) the same for students in the different analytics streams (or in other words, independent of program stream)? 3. How does my personal study time change over time?

I will answer these three questions using one-way ANOVA test, chi-square tests for independence and time series analysis using moving averages and exponential smoothing curve. Finally I will forecast my study time for the next few days.

Data Analysis

Part 1: ANOVA

To answer the first question: Are there differences in the average study times for students in the different analytics streams? I would use the one-way ANOVA test.

Stating the null and alternate hypothesis for the one-way ANOVA test:

H_0 : $\mu_B = \mu_C = \mu_H$ (Note: All three means (BAPG, CAGC, HAGC) are same.)

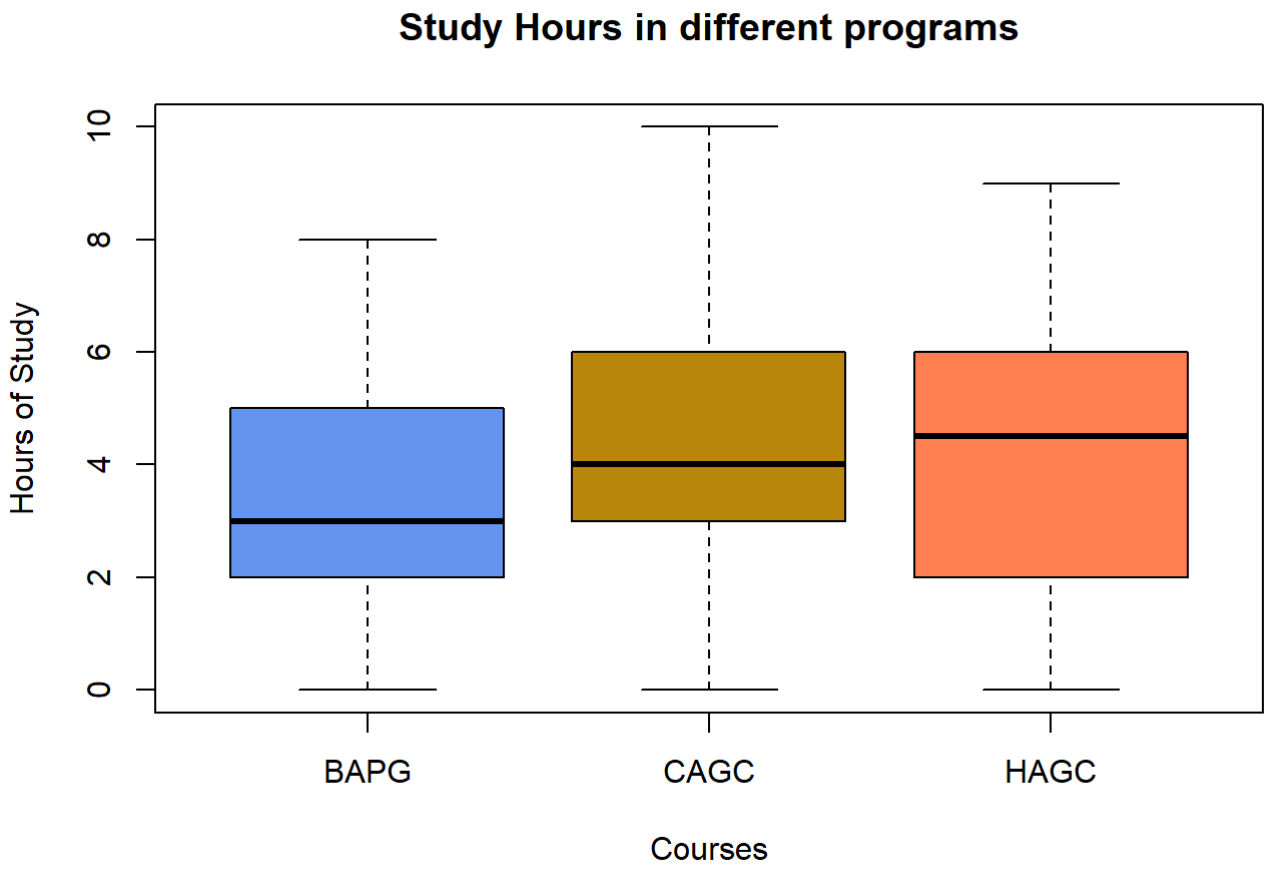
H_A : At least one mean is different

Since the p-value is 0.194 which is more than alpha (= 0.05), so we fail to reject the null hypothesis. Hence all the three courses (BAPG, CAGC, HAGC) study time are same and there is no significant difference in the study times.

Checking the Conditions for ANOVA.

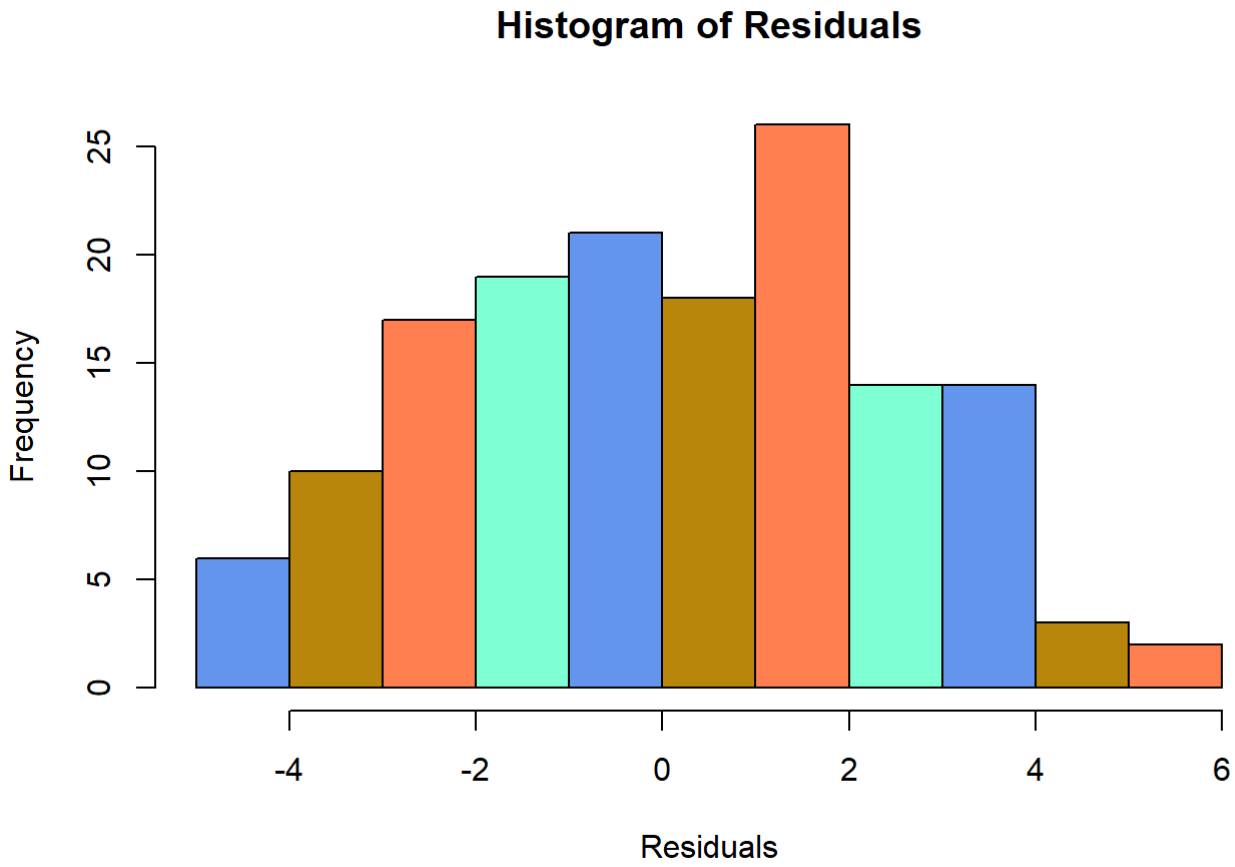
1. Independence Assumption: All the observations in the data set 'combined.csv' for each of the three categories (BAPG, CAGC and HAGC) are randomly selected in a time series data. Hence the randomization condition is met. Moreover, each of the group does not influence the activities of other group. Hence independent group assumption is met.

2. Similar Variance Assumption Creating a box plot of three programs 'BAPG', 'CAGC' and 'HAGC' for comparison of variances.



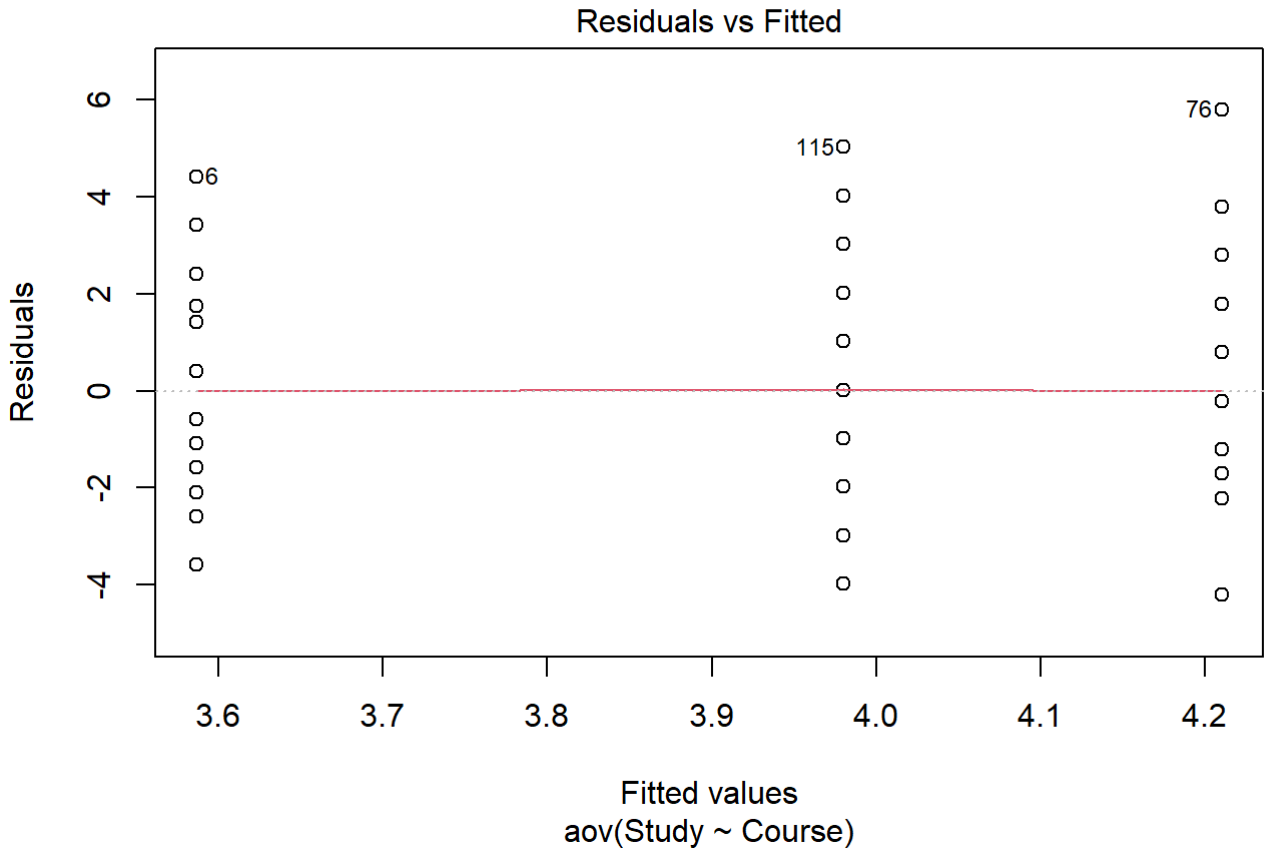
The box plot shows no extreme skews in the data and even though the box of HAGC program looks little bigger than others but it is not significantly bigger. So the the variances in each program can be considered nearly similar.

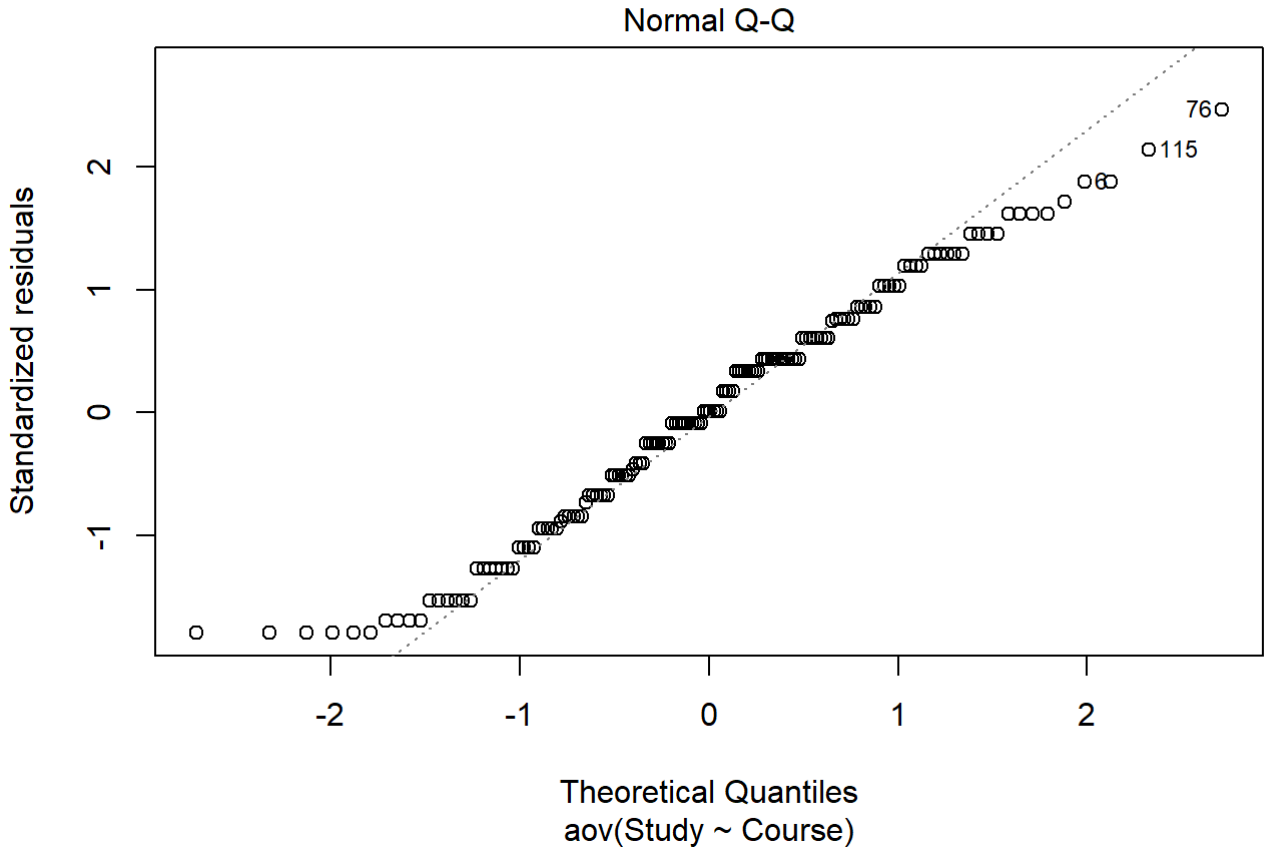
3. Normal Population Assumption Histogram of Residuals Plot

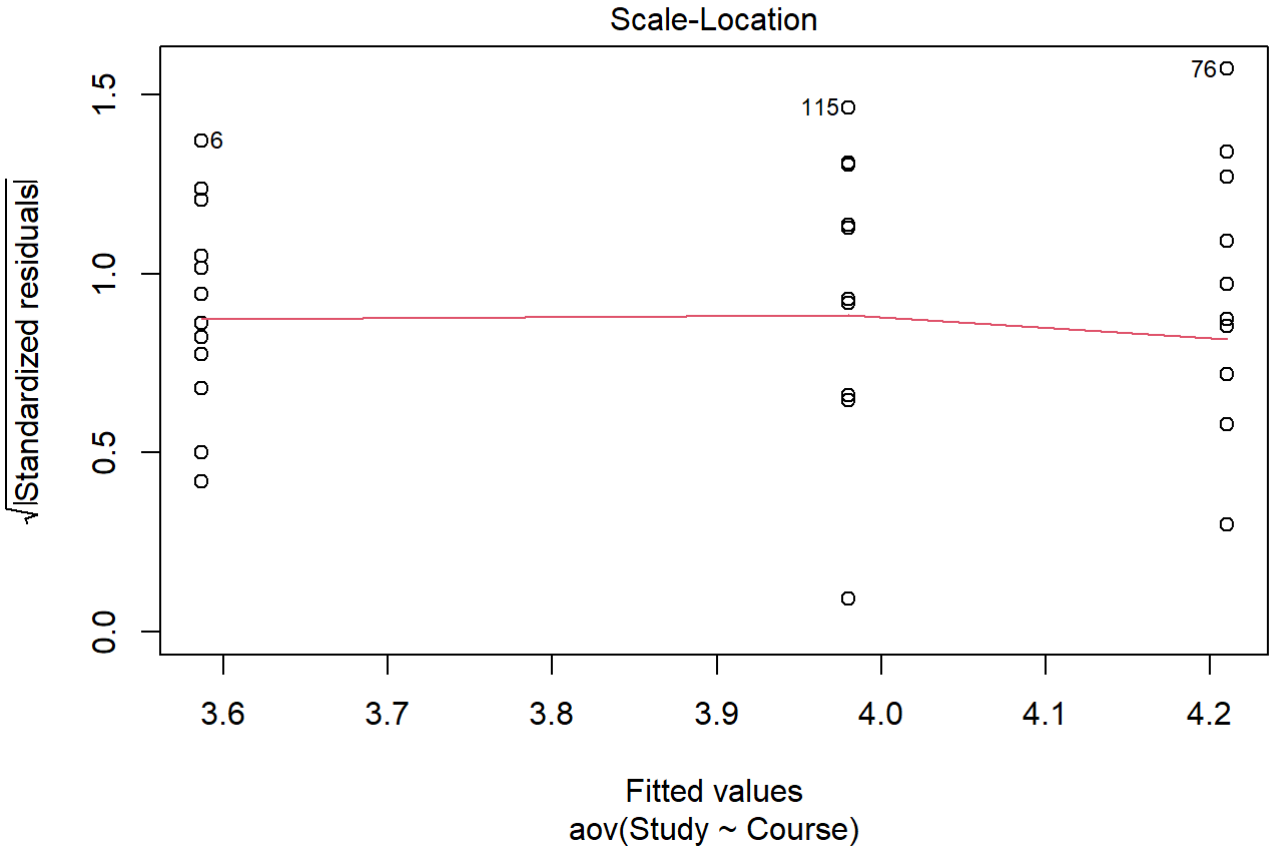


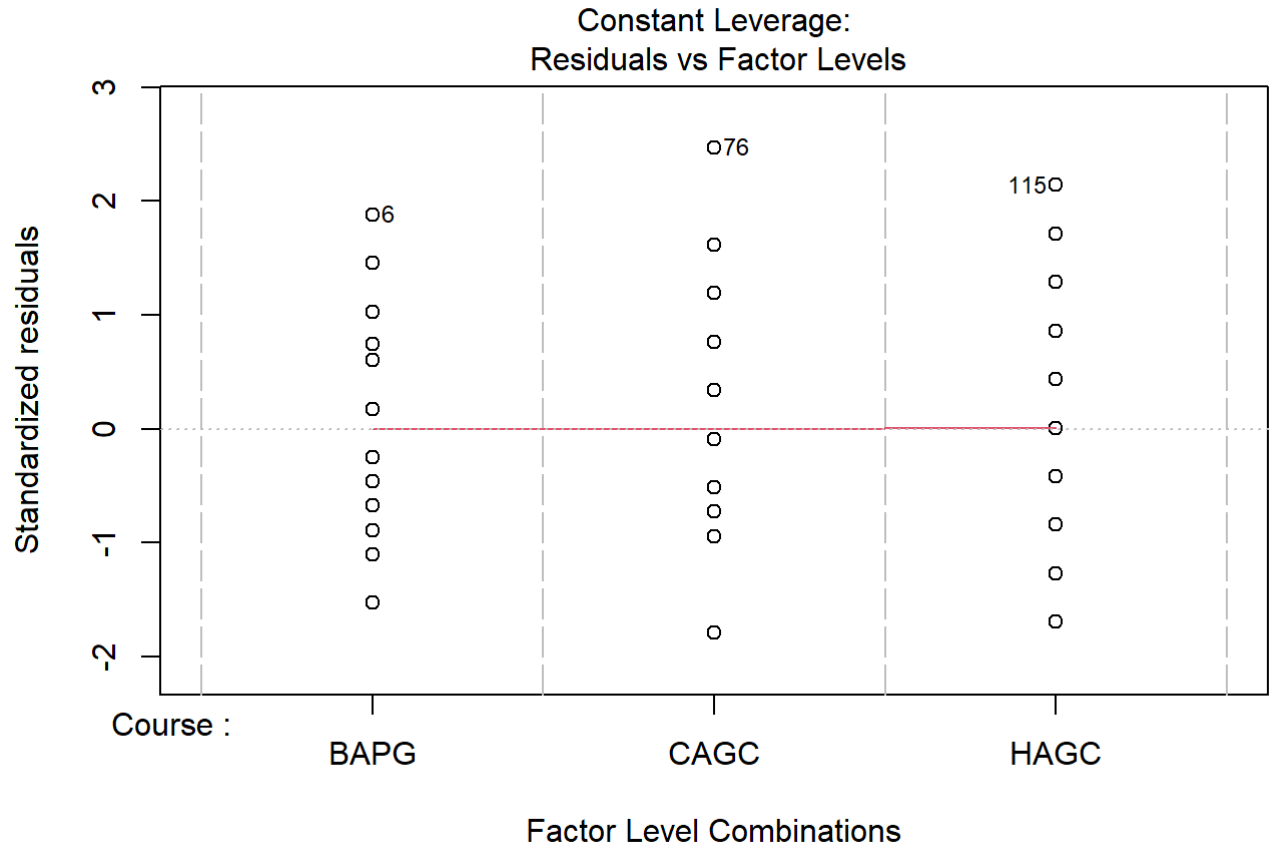
The histogram of residuals show nearly normal with unimodal shape and very low left skewed..

Checking other residual plots









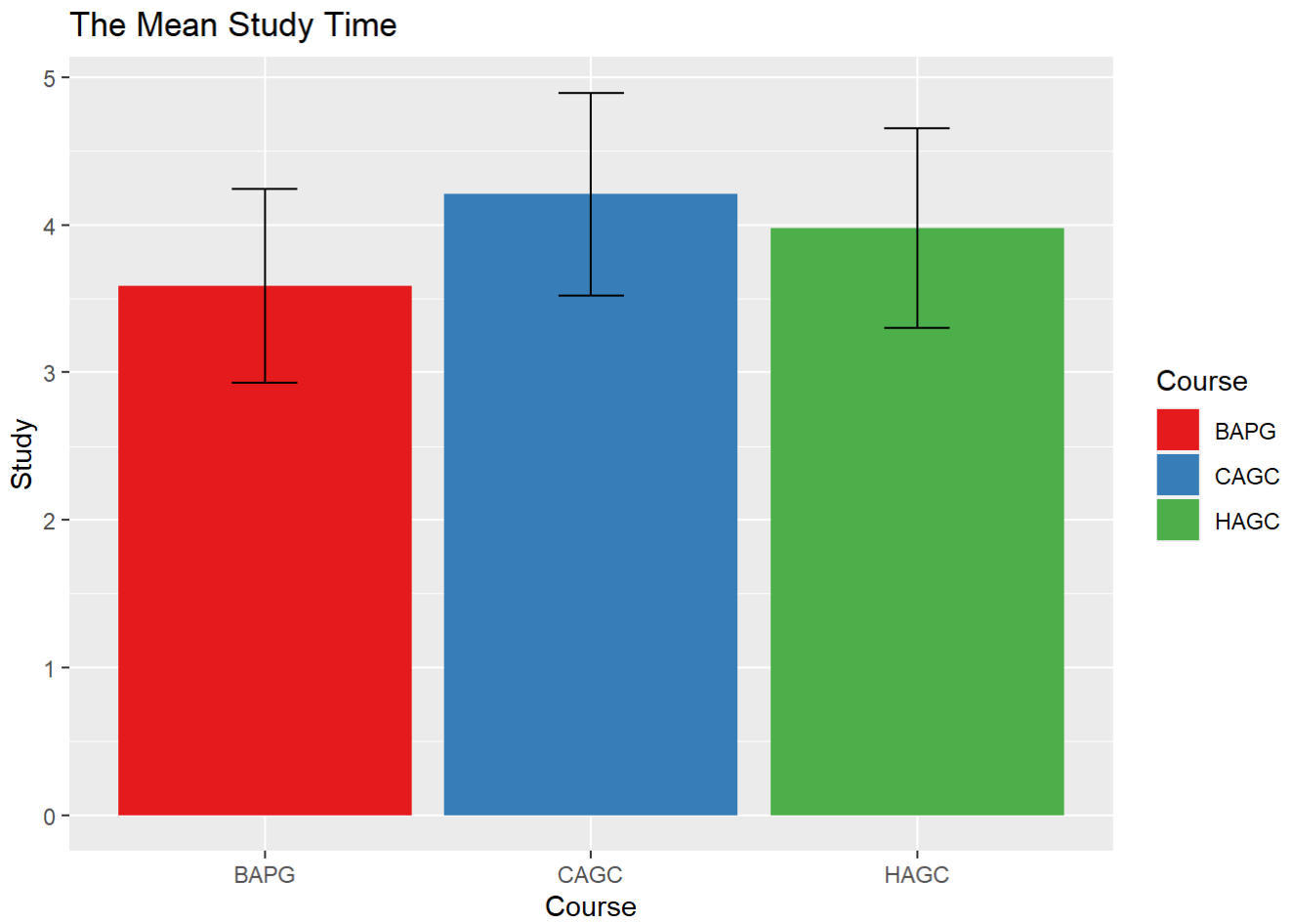
The residual plot shows nearly straight line. Since the assumptions and conditions meet the criteria for ANOVA test and so ANOVA test results are perfect for this case. Now checking the error plot of the ‘Mean Study Time’ for three programs.

Using Tukey's HSD Using Tukey's HSD to determine if the mean study times are different for each of the programs.

Interpreting the results of Tukey's HSD. The p-value difference between CAGC-BAPG: p-value = 0.1912921 which is more than α = 0.05. Hence there is no significant difference between the mean study time of program CAGC with BAPG. Similarly, HAGC-BAPG: p-value = 0.3886784 which is more than α = 0.05. Hence there is no significant difference between the mean study time of program HAGC with BAPG. Similarly, HAGC-CAGG: p-value = 0.9025809 which is more than α = 0.05. Hence there is no significant difference between the mean study time of program HAGC with CAGC.

The Tukey's HSD also shows that there are not significant difference between the average study hours per day for the program 'BAPG', 'CAGC' and 'HAGC'. This confirms the Anova test results.

The Error Bar Plot



The error bar plots shows almost same size for all the three programs. Though 'CAGC' is error plot is slightly bigger than HAGC error plot followed by BAPG error plot. The difference in error bar plot are not significant enough to consider that the means are different.

The mean study times of the three programs have no significant statistical differences because the nature of programs are primarily same for all the three courses. All the three courses relates to data analytics in different fields like business, crime and health, so there are many modules which are same for all the three courses and hence the study times for completing the assignments are same.

Part 2: Chi-Square Tests

In this section I am going to answer the questions 'Is the distribution of days studied more than 3.13 hours (the average daily study time for students at McGill) the same for students in the different analytics streams (or in other words, independent of program stream)?' For this I will use Chi-Square test for Independence because I

will be checking if the distribution of study hours (above or below) are same or different across various programs like Business Analytics (BAPG), Crime Analytics (CAGC) or Health Analytics (HAGC).

For doing the chi-square test for independence, I will do the following steps one after another 1. Prepare the data set to select 50 random samples from my personalized data set. 2. State the null and alternate Hypothesis. 3. Check the assumptions and conditions. 4. Calculate the test statistics. 5. Make a decision using p-values.

** Preparing the data set** Continuing to use the data set that contains the 150 random days selected for students in the three programs and selecting 50 random values for each stream. And then adding a categorical variable to the data set with two categories: o Above - if the hours studied are greater than 3.13 hours o Below - if the hours studied are less than or equal to 3.13 hours

Please note that 3.13 hours is the amount of time spent studying per day by students of McGill University and is purportedly the most in Canada.

From the table above we note that in Business Analytics (BAPG) program, there are 22 instances of days where students studied more than 3.13 hours (on average) on that day, while 28 instances days when the average was less than 3.13 hours. For crime analytics (CAGC) it is 26 and 24 days respectively while for health analytics (HAGC), it is 29 and 21 days respectively. To know whether this distribution of days across programs are statistically same or different, we will use chi-square test for independence. We will set up the Hypothesis for the statistics test.

Stating the null and alternate hypothesis. H_0: The study distribution of ‘Above’ and ‘Below’ is same across the programs. H_A: The study distribution of ‘Above’ and ‘Below’ is NOT same across the programs.

p = 0.3726 > 0.05 = alpha. The p-value is more than alpha. The result indicates we fail to reject the null hypothesis. Which means that the study distribution of ‘Above’ and ‘Below’ is same across the three programs.There is no significant difference in study distribution across different programs. This makes sense because all the three programs share many of the modules together. Hence the work load would be same across three programs. Hence the result. Now I will check the assumptions and conditions for the test suitability and accuracy.

Checking the assumptions and conditions

1. **Counted data condition:** The data counts the number of ‘Above’ and ‘Below’ in each program. Hence counts for the categorical variable is fulfilled.
2. **Independence Assumption:** The data is collected on a everyday basis as the average study time in each program. Hence each data is independent of each other. Moreover, each day study hours is the average of all the students in the program and so there is minimal influence of one day study hours over the other day.
3. **Randomization Condition:** The data collected seem to be random sample of students studying in the program. Hence the randomization condition is met.
4. **Expected cell frequency/ Sample size condition:** The sample size is 7537 observations (BAPG: 6314, CAGC: 402, HAGC: 821). All these observations are more than 5 random samples. Hence this condition is met.

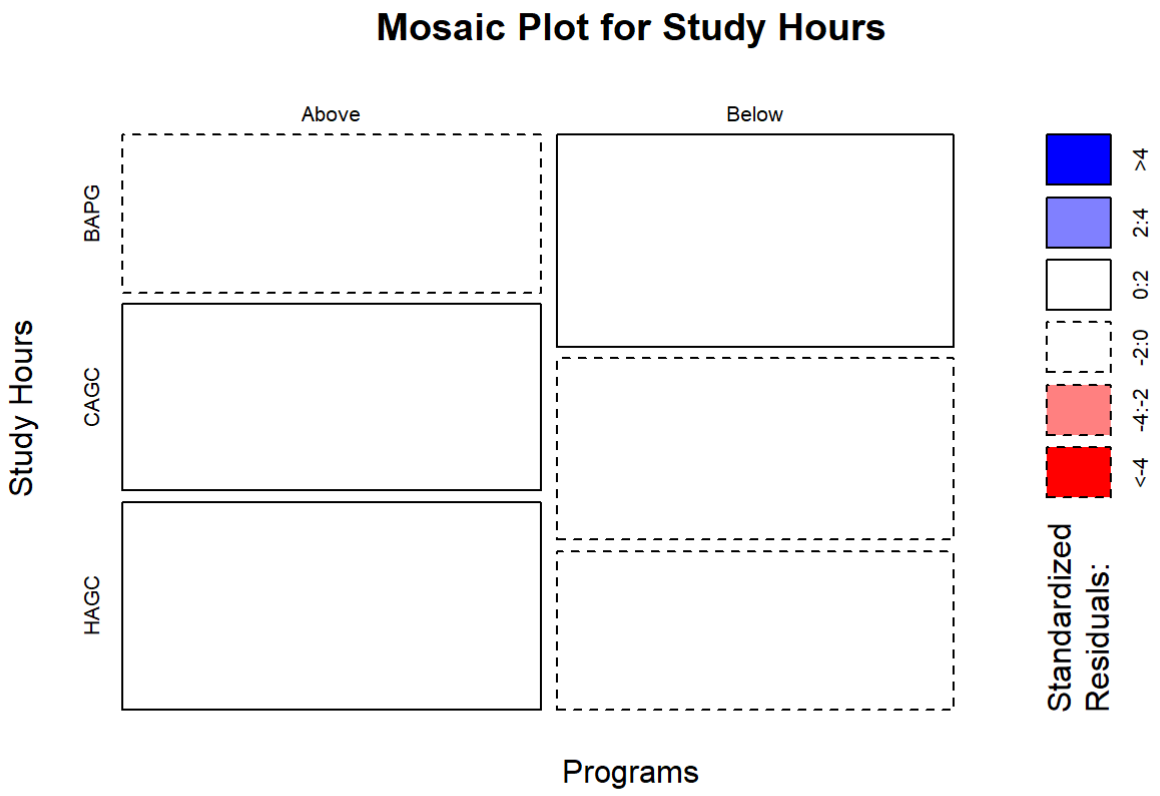
Hence all the assumptions and conditions are met for the chi-square test and its suitability. So the accuracy of the chi-square test is good.

Checking the residuals plot

```
##
##                BAPG        CAGC        HAGC
##   Above -0.72374686  0.06579517  0.65795169
##   Below  0.74331112 -0.06757374 -0.67573738
```

All residuals are within -3 and +3 range. Hence it supports the decision 'fail to reject'.

Creating Mosaic plot



The mosaic plot shows no boxes are colored which means all residuals are between -2 and 2. No residuals are unusual because they are not outside of the middle 95% of the data (according to the 68-95-99.7% rule).The boxes for the 'Above' part shows HAGC program is bigger than BAPG, while the 'Below' part shows the opposite. The BAPG box is slightly wider than HAGC.

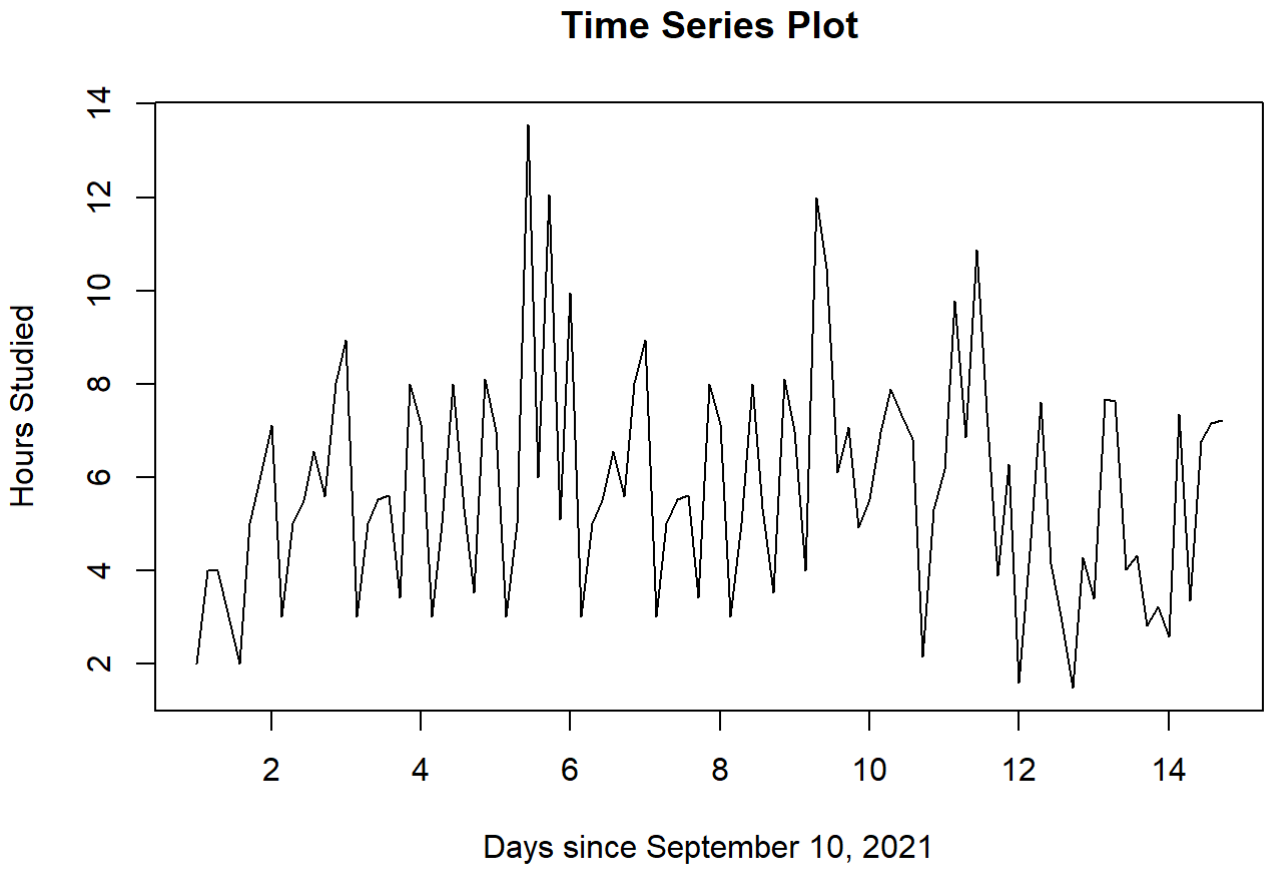
The mosaic chart shows that the variances (in average study times) in 'Above' category is more for the programs 'HAGC' and 'CAGC' as compared to 'BAPG'. While the variances (in average study times) in 'Below' category is less for the programs 'HAGC' and 'CAGC' as compared to 'BAPG'.

Part 3: Time Series Analysis

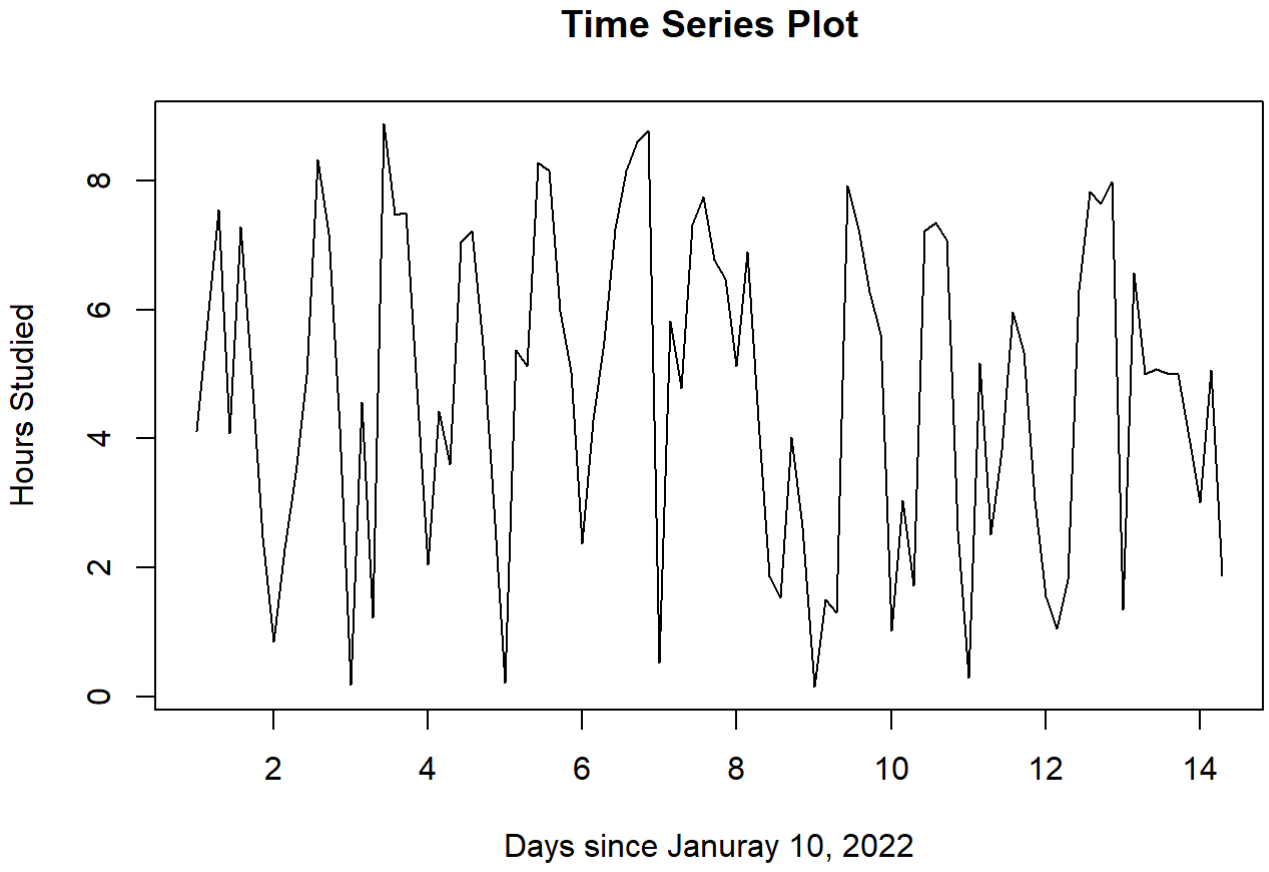
In this section I am going to answer the questions 'How does my personal study time change over time?' For this I will use Time Series Analysis which will help me to check my study time across days and I will be able to predict my study time in future. The time series analysis is divided into two parts 1. Fall 2021 2. Winter 2022

The time series plots are given for both the groups separately since there is a time gap between December 15, 2021 till January 10, 2022. However, I will do all the further analysis, the moving average model, the forecast and error estimation on winter 2022 semester.

The Time Series plot for Fall 2021 Semester



Time Series plot for Winter 2022 Semester

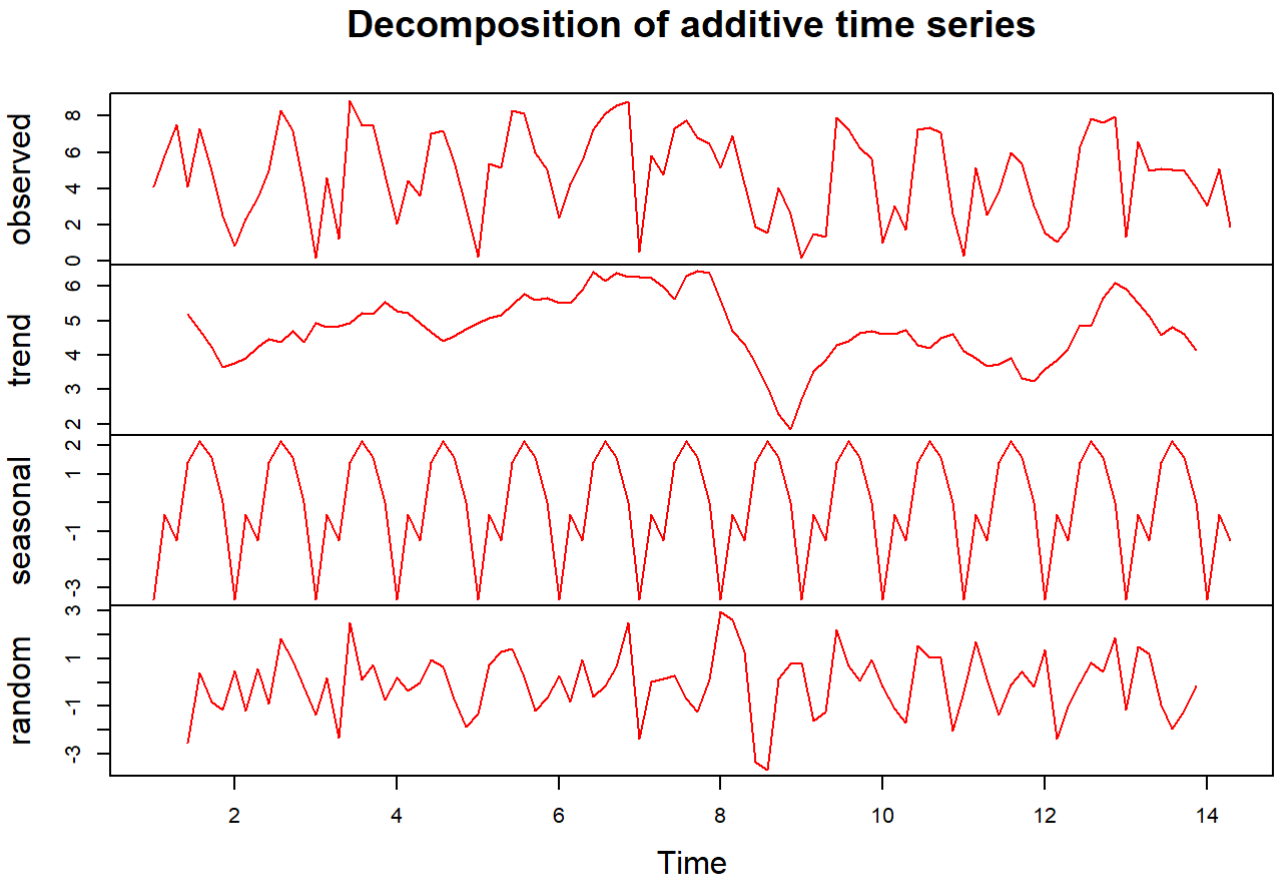


The time series plot of my personalized study time shows no particular trend component but there is seems to be strong cyclical component which goes up and down every week. The plot looks stationary with no particular trend which shows that my average study time almost remained same throughout the fall and winter semesters. There is drop in average study time from the fall semester (5.77 hours) to the winter semester (4.75 hours). This is because the day light hours are more in Fall semester than winter semester and so the average study hours are more in fall semester than winter semester.

The above plot also shows that:

1. My average study time during Winter Semester Weekends was 6.58 hours which is more than my winter semester average of 4.75 hours. This is mainly because on Weekends, I do not attend zoom class times and that time is also used personal studies and assignments.
2. On 2nd, 3rd and 4th April 2022, my average study time was 7.98 hours (much higher than winter average of 4.75 hours). This is because I had two big assignments submission, Assignment 11 and 12 of BTA 1016 on 5th April. Hence the study time increased on those days.
3. From 28th March till 31st March, my personal study time was lower (average 1.47 hours) as compared to winter semester average (4.75 hours). This is because on those days my zoom class time has increased with almost full day classes on every Tuesday (8 to 9 hours) and considerably higher amount of classes on Wednesdays (4 to 5 hours) and Thursdays (5 to 6 hours). Also on Thursday I spend two hours on tutoring, which reduces my personal study time.
4. The Fall semester study hours average was 5.77 hours while the Winter Semester study time average was 4.75 hours. The study hours are higher during the pandemic time because there are no commuting time. Hence there is a time saving of one to two hours each day due to commute to and fro college and its allied activities.
5. On 9th and 10th March 2022, my average study time was 1.40 hours as against winter semester average (4.77 hours). This is because on these days I binged watched a TV series on YouTube channel which I liked very much. I watched the TV series on an average of 4.79 hours each day against W22 average (1.81 hours).
6. During the study week, in end of March 2022, my average study time hours rose to 6.38 hours per day against W22 average (4.77 hours per day) because there was no zoom class meetings and entire time was used for assignments completion and catching up on lost lectures. Though I also went out on a group party with friends, yet the average study time was better than other weeks.

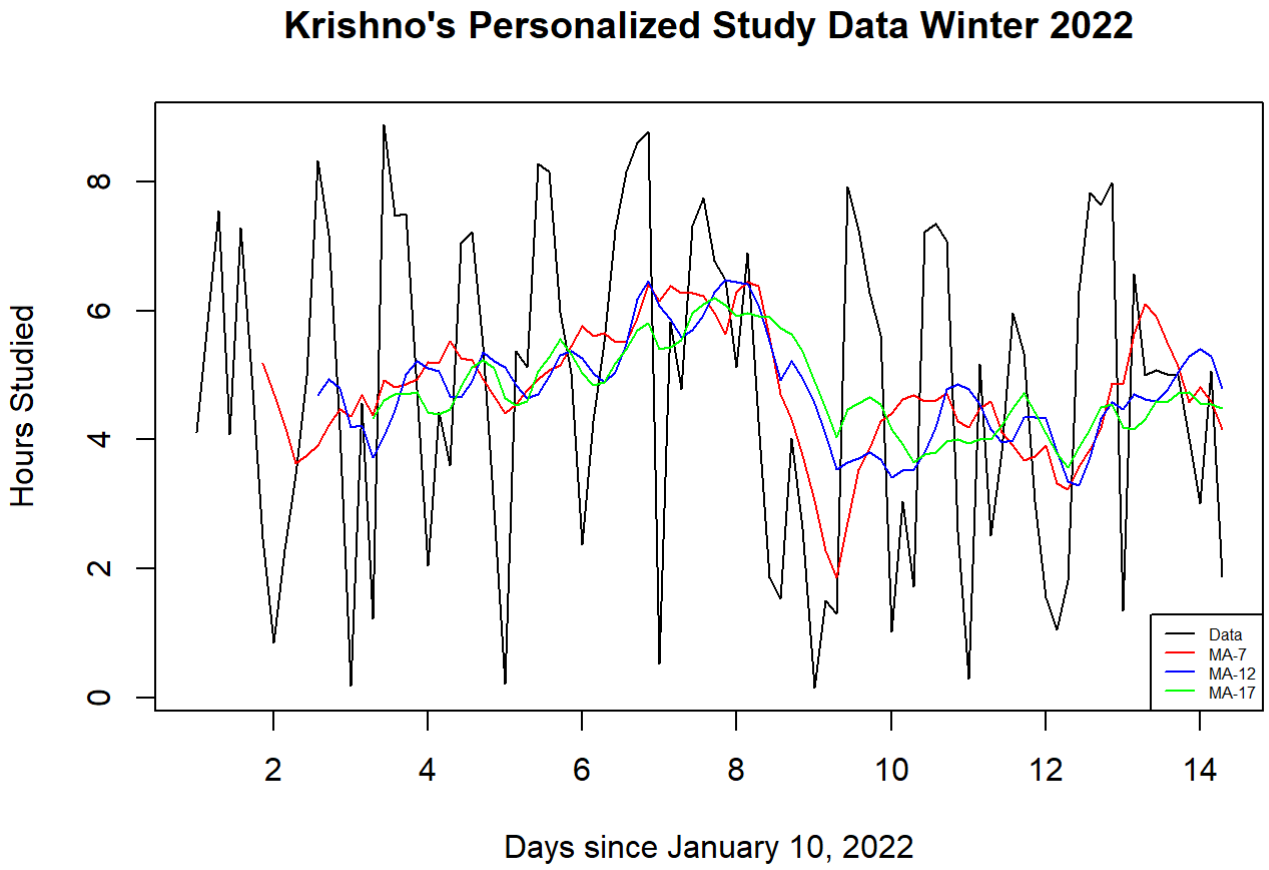
Plotting the decomposition of the time series for hours studied for winter 2022 semester



The decomposition curve for **winter 2022 semester** shows:

1. **Trend**: There is no particular trend in the plot. However the trend line shows slow increase in the number of study hours from the beginning of the semester to the middle (because as semester progressed assignments came in more and so more study time) and then there is a sudden drop in the study hours between 8 to 10 period (this is because of the study week) and later it steadied again at around 4 hours which increased later at the end of semester (again due to end of semester final projects submission).
2. **Seasonality**: There looks to be a strong seasonality in the graphs particularly on every weekends the number of study hours increases as compared to weekdays. Apart from that there seems to be a small peak too before every big peak, which is on Thursday when the zoom meeting time was low.
3. **Random**: There seems to not much irregularity in the curve expect few. The irregularity during the 8th period is noted due to the study week period where the study hours reduced in general. There are few more peaks in the curve, those are mainly due to the submission of BTA 1016 (Connecting Data) assignment submission.

Fitting the Moving Average Models and plotting the curves to Winter 2022 semester



The above smoothing curve shows that the green line seems to match better with the data. The green line corresponds to moving average length 17. We will check the error margin of the curves and confirm accordingly.

Forecasting the smoothing models

The result shows, using moving average models: MA-7: 4.15 Hours MA-12: 4.80 Hours MA-17: 4.48 Hours

All three predictions from smoothing model shows a if the next semester continues in the same way then the average study hours would be more than 4 hours per day.

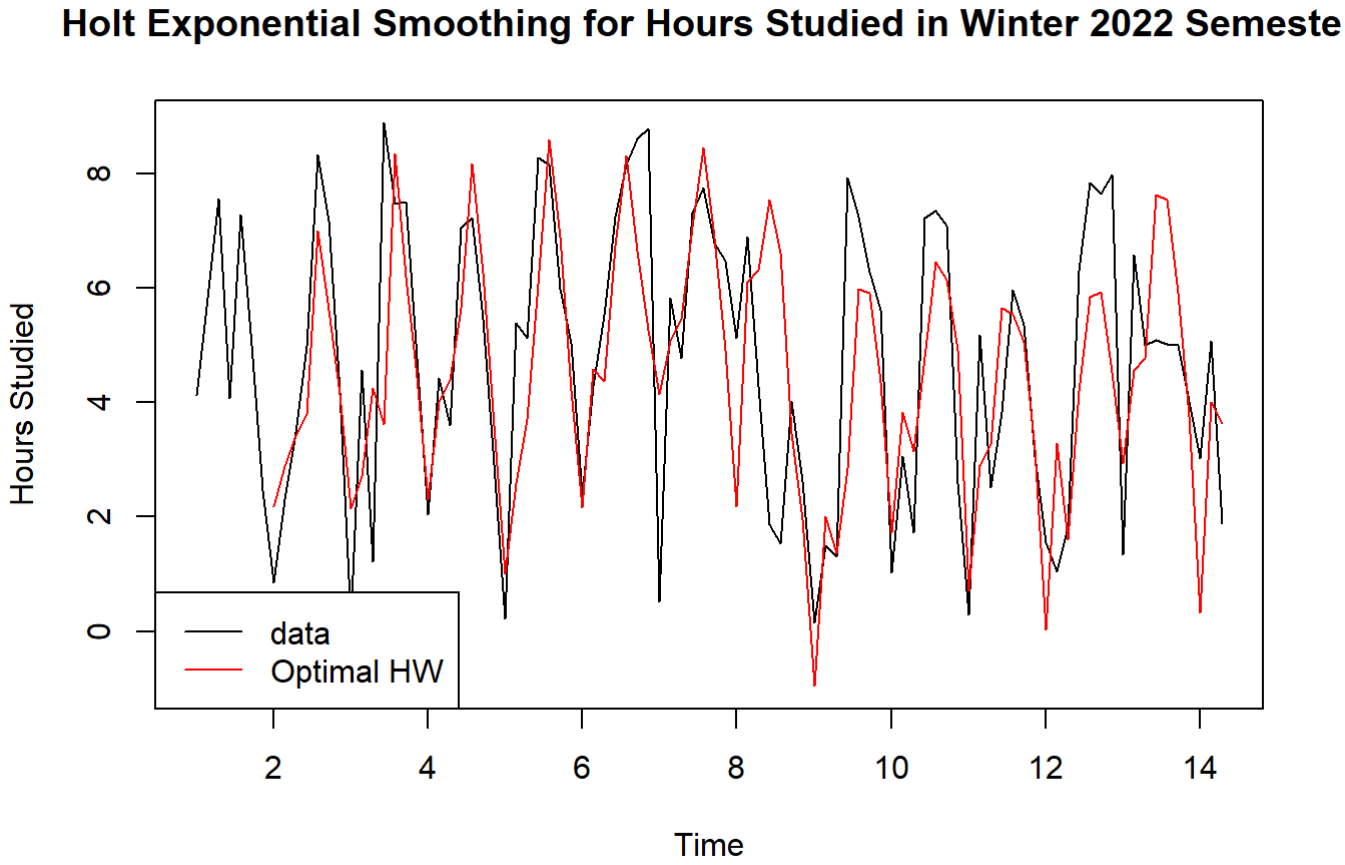
Getting the Error Forecasts Getting the error forecasts to choose the right model.

The results are as follows:

Error	MA-7	MA-12	MA-17
MSE	006.59	007.08	006.45
MAD	002.13	002.23	002.14
MAPE	167.07	189.36	161.77

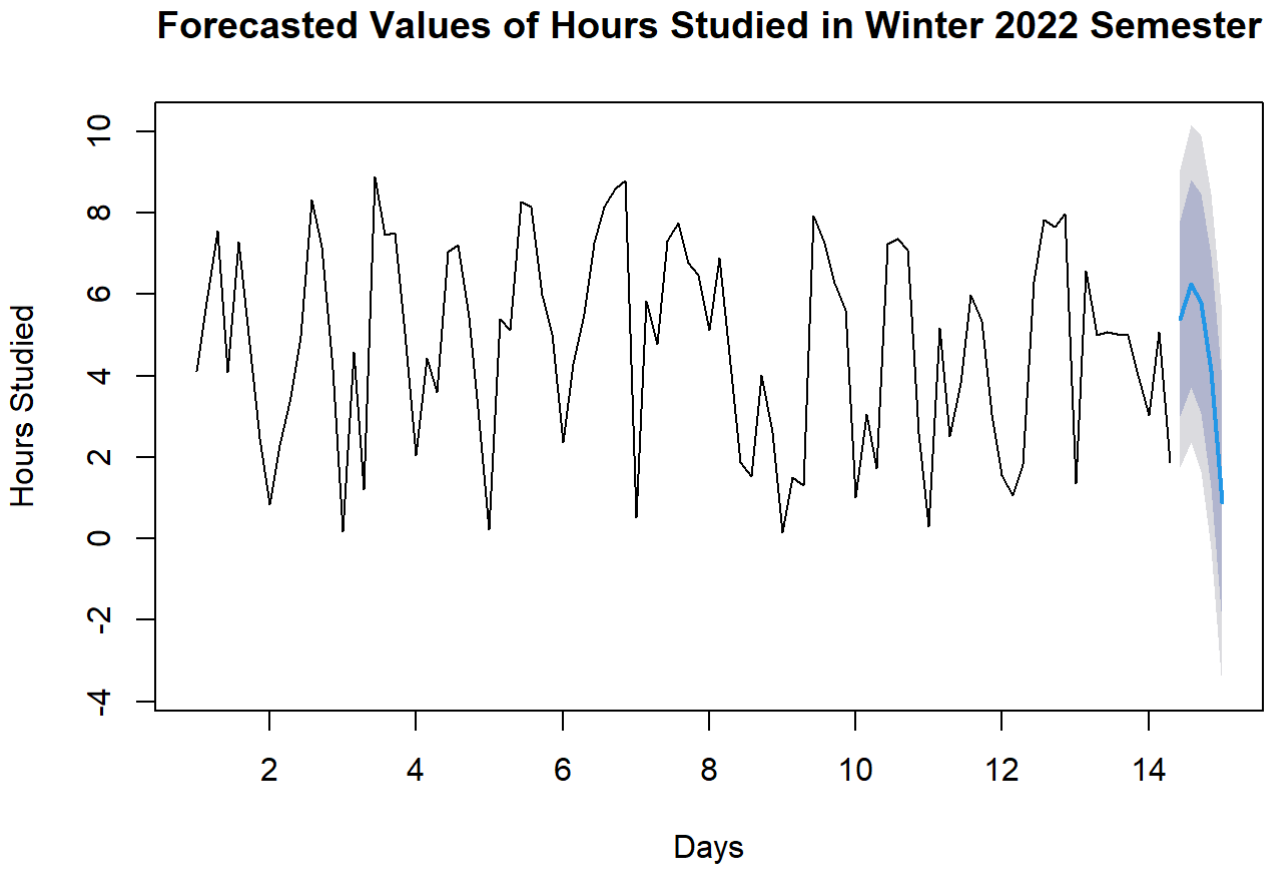
The model with the least error seems to be MA-17 (Moving Average 17). Also the trend shows the Moving average Length 17 matches well with the curve and gives proper prediction.

Fitting the best exponential smoothing curve and Plotting for next 5 days forecast Since my time series curve shows both trend, randomness and seasonality, so I will use the Holt Winters exponential smoothing with trends and and seasonality



Predicting the average Hours studied for the next 5 days

```
##          Point Forecast     Lo 80    Hi 80      Lo 95     Hi 95
## 14.42857      5.4025506  3.013892 7.791209  1.7494128  9.055688
## 14.57143      6.2573726  3.708783 8.805962  2.3596413 10.155104
## 14.71429      5.7735197  3.073398 8.473642  1.6440394  9.903000
## 14.85714      4.1055865  1.260985 6.950188 -0.2448568  8.456030
## 15.00000      0.8933525 -2.089705 3.876410 -3.6688408  5.455546
```



As per the forecast model, the ‘Hours Studied’ is going to decline over the next 5 days.

Conclusion

Based on my analysis of the data, I conclude that:

1. There is no significant difference in the study time of students from different streams like ‘Business’, ‘Crime’ and ‘Health’.
2. Moreover, the distribution of ‘Above’ (study hours more than 3.13 hours) and ‘Below’ (study hours less than 3.13 hours) across the three streams (‘Business’, ‘Crime’ and ‘Health’) have no significant difference.
3. My personal study time increased slowly as the semester progressed because the assignment workload increased and also as the modules became more complex, it was taking more time to complete. Hence my study time increased as the semester progressed. However, during the study week, the study time came down because I relaxed few days during those times. Again as the semester progressed, my study time slowly increased till the end of the semester due to increased load of assignments. Also it is seen that during weekends my study time goes up because there is less zoom meetings.
4. The daily collection of personalized data is a good tool to be aware of how I spend my daily activities, what activities are productive and what are not that much productive. The collection of personalized data helps me to track my actual time vs. my plan time. Also it will help me to prepare a realistic plan. This is a good time management tool. I am planning to continue to collect my personal data. I feel indebted to my professor for this useful and insightful project work.