**Introduction Section**

1. **Provide details regarding the question you want to address.**

* Goal of this report is to study the relationship between interim STA302H1 quiz scores (quiz 1 – 3 scores), study time (weeks 1 – 4), and COVID contemplation time (week 1- 4), vs. final STA302H1 quiz score (quiz 4 score).
* Studies show individual relationship effects
* TODO: Studies show positive correlation b/w consecutive weeks of study and quiz scores
* TODO: Studies show times of high anxiety (e.g., COVID) and illness, quiz scores decrease
* My paper studies all these covariates together to examine collective effect on quiz performance AND effects of covariates on each other (hence the interaction terms)
* Population:
  + Summer 2021 STA302H1 cohort
  + n = 227 students at the beginning of the course
  + n = 198 students as of August 13, 2021
* Professor surveyed students (all short answer) on Quercus for the first 4 weeks of STA302H1:
  + End of Week 1 (July 5 – July 9)
  + End of Week 2 (July 12 – July 16)
  + End of Week 3 (July 19 – July 23)
  + End of Week 4 (July 26 – July 30)

1. **What is the purpose of developing this model?**

* We want to determine the strongest predictors of quiz 4 score:
  + interim STA302H1 quiz scores,
  + study time,
  + COVID contemplation time,
  + country, or
  + some other combination of terms

1. **Explain how you’ll develop the model (without stating any model results).**

For example, you could:

* 1. Take mean of 3 quiz scores, weeks 1 – 4 COVID, weeks 1 – 4 STA302H1 study
  2. Take median of 3 quiz scores, weeks 1 – 4 COVID, weeks 1 – 4 STA302H1 study
  3. Analyze similar countries separately or add country as a predictor variable and group similar countries together to reduce number of categories.
* Reformat data.
  + Remove “hours” by casting to double
  + Remove non-Unicode characters “<UTF-098>”
  + Capitalize “canada” and “china” so they’re treated as the same country as “Canada” and “China”
  + Rearrange columns, group COVID times, study times, and quiz scores together
* Handle missing data.
  + Identify the number of quizzes students miss.
  + 1 – 2 NAs for quiz scores is OK, 3 – 4 NAs indicates the student dropped STA302H1
    - Perhaps STA302H1 students join late from the waitlist.
    - The new “best 3 of 5” quiz marking scheme allows students to miss at most 2 quizzes without penalty.
    - Students who chronically skip quizzes are usually not committed to (e.g., given up on) completing STA302H1 and are highly likely to drop STA302H1 since they have a slim chance of succeeding.
  + Missing the country is OK, just mark country as unknown
  + Missing any number of COVID and STA302H1 hours is OK, those students either forgot or abstained from sharing because they felt uncomfortable sharing it.
  + Missing COVID or STA302H1 hours alone is no reason to exclude them from the dataset. They might still write quizzes though.
    - Hawthorne effect: students have an incentive to abstain or be dishonesty to appear socially desirable 🡨 this goes in limitations
* Create descriptive statistics
  + Create histograms to analyze frequencies of COVID time intervals, study times and quiz scores, not grouped by country, analyzed data as a whole
  + Create boxplots separated by country to visualize the 5 number summary (Q1, median, Q3, min, max)
  + Create summary statistics to calculate median and mean
* Create pair scatterplot to come up with possible relationships to examine more closely – try to find a discernable relationship (e.g., linear, quadratic, logarithmic, sqrt, etc.)
  + Look at quiz scores alone 🡨 looks more (negative) quadratic than linear
  + Look at COVID times alone 🡨 looks more (positive) quadratic than linear
  + Look at study times alone 🡨 looks linear? Or no relationship?
* Create correlation matrix to find correlation between two variables
  + Find correlation between two covariates
  + Find correlation between covariate and response variable
* Develop initial model
  + Decide what model to use
    - Linear 🡨 keep your model simple
    - Quadratic, Cubic
    - Logarithm, Square Root
  + We could include country (but country turns out not to be significant)
  + Take median of 3 quiz scores
  + Take median of weeks 1 – 4 COVID hours
  + Take median of weeks 1 – 4 STA302H1 study hours
    - I can grab means and medians from the boxplots.
    - Median is less prone to skew than mean.
* Refine original model
  + Use backwards selection on original model (state original model terms and why you chose them)
  + Run AIC to find which subset model contains lowest AIC (state results in the model development section, add reference to appendix)
* Claim it’s a linear model
  + State the 4 linear model assumptions, and how you’ll test them
    - Linearity
      * Examine scatterplot itself
    - Independence of errors
      * For all predictor variables, examine scatterplot of residuals vs. fits, show no discernable relation
    - Homoscedasticity (constant variance)
      * Examine scatterplot of residuals vs. fits – show there’s no megaphone effect or bowtie effect
    - Normality of errors
      * Show points follow line in qqplot
      * Show histogram of residuals for approximate normality or use CLT.
  + Put residual plots, res. vs. fit in appendix.
* Find influential outliers
  + Use VIC method
  + Use Cook’s distance?
  + Use DDFITS, DDBETA? Leverage points?
  + define cook's distance, leverage points
  + define homoscedasticity (constant variance)
* Validate model
  + Check if mean = 0?
  + Perform t-tests?
  + Find 95% confidence intervals?
  + Use confidence intervals for each (numerical) predictor variable
  + TODO: Use VIF to fine influential outliers?
  + TODO: Cook’s distance, what it says about influential outliers

1. **How does the model meet the purpose of your model?**
   * Current professors can identify possible weak topics, reflect on things they did/did not help students, and tailor their future lessons that way
   * Teaching stream professors can devote resources to developing carefully curated course content for material students find most challenging (based on quiz scores)
   * Future professors can establish reasonable course expectations, and identify common pitfalls so they can prepare extra well for those formative lectures
   * Current students can focus on key material to getting high grades on hard quizzes
   * Future students can establish reasonable expectations about workload and develop strategies to maximize their time and success in STA302H1 with available resources

**Exploratory Data Analysis**

1. **Give a detailed description of the variables in the dataset.**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Meaning** | **Type of Variable** |
| Quiz\_1\_Score | Student’s quiz 1 score, out of 10 | Numerical ratio, continuous |
| Quiz\_2\_Score | Student’s quiz 2 score, out of 10 | Numerical ratio, continuous |
| Quiz\_3\_Score | Student’s quiz 3 score, out of 10 | Numerical ratio, continuous |
| Quiz\_4\_Score | Student’s quiz 4 score, out of 10 | Numerical ratio, continuous |
| COVID..hours.W1 | Time student spent thinking about COVID-19 during Week 1, in hrs | Continuous numeric |
| COVID..hours.W2 | Time student spent thinking about COVID-19 during Week 2, in hrs | Continuous numeric |
| COVID..hours.W3 | Time student spent thinking about COVID-19 during Week 3, in hrs | Continuous numeric |
| COVID..hours.W4 | Time student spent thinking about COVID-19 during Week 4, in hrs | Continuous numeric |
| STA302..hours.W1 | Time student spent studying for STA302H1 during Week 1 (can include lecture time, review time, quiz time, or assignment time) | Continuous numeric |
| STA302..hours.W2 | Time student spent studying for STA302H1 during Week 2 (can include lecture time, review time, quiz time, or assignment time) | Continuous numeric |
| STA302..hours.W3 | Time student spent studying for STA302H1 during Week 3 (can include lecture time, review time, quiz time, or assignment time) | Continuous numeric |
| STA302..hours.W4 | Time student spent studying for STA302H1 during Week 4 (can include lecture time, review time, quiz time, or assignment time) | Continuous numeric |
| Country | Student’s country of origin | Categorical/nominal |

* TODO: Did I get everything below?
* Quiz scores are out of 10
* STA302H1 study time and COVID contemplation time are measured in hours.
* Response variable:
  + Quiz 4 score (out of 10) 🡨 performance on quiz 4, numerical ratio variable
* Predictor variables:
  + Quiz\_1\_Score, …, Quiz\_4\_Score = grades on quiz 1 – 3 scores (out of 10) 🡨 numerical ratio, not necessarily numeric
  + covid1, …, covid4 = Weeks 1 – 4 STA302H1 study time (hours) 🡨 continuous numeric
  + Weeks 1 – 4 COVID-19 contemplation time (hours) 🡨 continuous numeric
  + Country = country of origin – categorical/nominal variable
* e.g., quiz 1 – 3 are integers, quiz 4 is doubles. – quantitative: numeric ratio, ordinal

1. **Display appropriate tables or figures that highlight certain characteristics of your variables that you deem important to mention. (Do not discuss model results)**

* TODO: This would be identifying the most significant variables, right?
* TODO: Show gtsummary of lm() output, showing significant and non-significant variables at the 5% significance level.
* Add descriptive statistics (i.e., histograms, boxplots, scatterplots)
* Display histograms of all predictor variables against quiz 4 score
* Display boxplots of all predictor variables against quiz 4 score
* Display pair scatterplots of relationships between quiz 4 score and one of the predictor variables
* Describe each descriptive statistic – “this histogram/boxplot/scatterplot displays relationship between X and Y”
* Don’t discuss relationship results though
* See figure X in appendix
* TODO: Consult 3 – 4 external sources to confirm your findings.

**Model Development Selection**

1. Give a detailed description of the process used to come up with the final model.

Display results

* Tentative models:
  + Quiz 4 ~ Quizzes 1 – 3
  + Quiz 4 ~ Weeks 1 – 4 COVID-19
  + Quiz 4 ~ Weeks 1 – 4 STA302H1
* Select type of model
  + First order model?
    - Linear model?
    - Quadratic model?
    - Cubic model?
  + Second-order model?
    - Think about interaction terms & multicollinearity
    - Correct multicollinearity with recentering, and perform analysis based on recentered model – undo transformation to reveal true relationship
    - Correct skewness with variable transformation, and perform analysis based on transformed model
* Find correlation matrix
  + use highly correlated combinations of predictor variables as a heuristic for determining most significant terms in model
* Use model selection criterion to identify best model
  + For each of the three models, there should be about 8 (2^3) – 16 (2^4) possible models to choose from
  + Use R^2, adjusted R^2, C, AIC, forward selection, backwards selection criterion
  + TODO: In the event of multiple best models, use testing set to break ties. Usually, one model tends to perform better
* used correlation matrix and visual inspected scatterplots to derive an initial model (state original variables, and why)
* used backwards selection (AIC) to refine model to reduce the model to a simpler model with a smaller AIC value (quote AIC model).
* include AIC values in appendix.
* explain why variables are significant/insignificant, with p-values and t-values -- "for every unit increase in ... " (quote coefficients)
* state r^2 values, what they mean ("x% of the total variation in quiz4 can be explained by all variables in the model")
* adjusted r^2 = r^2 adjusted for predictor values.

1. **Add statistical and empirical justifications for your model.**

* Show results of 4 assumptions for linear model:
* model obeys A1: model has only linear terms and interaction terms, and
* we showed that all predictor variables show randomness when regressed with residuals.
* model obeys A2: random sample = assume errors are independent, and see 1st plot (residual vs. fitted plot)
* model obeys A3: approximately homoscedastic, from scale location plot, we see random scattered points about red horizontal line (mean?)
* model obeys A4: approximately normal -- slightly more left skewed (qqplot) than right skewed
* no influential outliers -- no points outside of cook's distance in upper right and lower left quadrants of plot, so no points to remove.
  1. Statistical:
     + Checking Gauss-Markov assumptions
     + Using prior knowledge (e.g., quiz grades tend to be left skewed because few STA302H1 students fail quizzes)
     + Statistical theory from STS302H1 or STA248H1
  2. Empirical:
* Showing residual plots
* Showing skewed histograms
* Showing well fitted qqplots

In results section

* Histograms of grades are left skewed (mean < median),
* Histograms of study times and COVID contemplations are right skewed (mean > median)
* r^2 value, r^2 adjusted, what they mean
* State significant and insignificant variables w.r.t. significance level
* no leverage points/influential points, so no points removed

1. **Add in-depth diagnostics to illustrate the goodness of the model.**

* TODO: Anything else, other than residual plot and qqplot to assess goodness of fit?

**Conclusion**

1. **Explain why model is useful in the context of the data.**

TODO: Figure out how to rephrase purpose of model from introduction section.

1. **Interpret final model in non-technical language (i.e., explain how the variables work, discuss predictions)**

* quiz 3 score: quiz 3 is closer to quiz 4 in difficulty because students are used to online quiz format, and study more to better prepare for them
* (sta302week1, sta302week2), (sta302week2, sta302week3), (sta302week3, sta302week4) covariates: increase in study times between consecutive weeks (week 1 -> week 2 -> week 3 -> week 4)
* Interpret coefficients: “With unit increase in X, y increase in Y…”

1. **Discuss any limitations/problems remaining with the model, and how they might impact its use in the real world.**

* Many factors can affect quiz 4 score:
* Internal factors within STA302H1
  + Studying
  + Attending lectures
  + Attending office hours
  + Completing assignments
  + Completing readings
* External factors outside of STA302H1
  + Prerequisite knowledge
  + Current work ethic, discipline, diligence
  + TODO: Do grade adjustments count?
* Online offerings are different from equivalent in-person offerings of STA302H1
  + More discipline necessary to succeed
  + More technological literacy required to succeed, privileging students who have access to computers or internet – some families are too poor to afford computers and internet, and must share or use public computers
  + Performance has more room to fluctuate – some people may prefer online courses over in person lectures due to long commute times, and vice versa
  + Harder to create boundaries with work/rest, and work environments vs. play/sleep environment
* Events outside of one’s time in STA302H1 (studying for quizzes, attending lectures, and doing assignments)
  + Mental health
    - Better mental health = more resilient individuals, more positive outlook on life, more altruistic
  + Commute times
    - Increased commute times increases students stress levels because it leaves less time available for STA302H1 and other courses, and affects one’s disposition towards class format (in-person vs. online)
* Social connections
  + Staying connected helps stave off negative COVID-19 thoughts and promote community
  + Fewer opportunities for study partners
* Number of hours slept, quality of sleep
  + Well-rested brains tend to respond quicker, can better regulate their mood, have an easier time focusing during lecture, and make fewer mistakes on quizzes and assignment (or notice them more easily and readily)
* Anxiety levels?
  + Anxiety may influence quiz scores and assignment performance because it affects cognitive performance – brain may hyperfocus (fixation) or hypofocus (distraction) – fight or flight, focus on survival (not dying of COVID-19) rather than thriving (succeeding in STA302H1)
* Number of extra-curricular activities, professional (e.g., preparing for job interviews) or recreational (e.g., going outside, biking, etc.)
  + Alternatives for STA302H1 studying, effects are debatable depending on how related they are to STA302H1 or statistics in general
* Level of physical activity (exercise helps brain learn, which may improve performance)
  + Exercise and physical activity make you more alert, less prone to illness and injury, promote a positive attitude, and improves relationships with other people
* Caring for family
  + Family responsibilities may distract or interfere with one’s progress on STA302H1 assignments and consume STA302H1 study time
  + Family members may also provide a nurturing environment for your studies through quality family time, strong family values, and family work connections
* Discipline
  + Exercising discipline allows one to consistency interact with STA302H1 material and increase their chances of retaining the material and performing well in STA302H1 assignments and quizzes.
* Prior background – direct (past stats/ML courses) or indirect (quantitative background, programming background)
  + Those who are more familiar with prerequisite statistics material and do well in prior courses have an easier time studying for STA302H1 and are more likely to succeed – ceteris paribus.
* Attitude towards school (keen vs. disengaged)
  + Those who are keen in school tend to perform well, regardless of class format, and vice versa
* Thoughts about COVID
  + Positive thoughts can help tune out some negative COVID-19 thoughts and focus on what you have control over. Persisting COVID-19 thoughts can cause one to feel despair and pessimism
  + COVID case counts and current COVID restrictions in country of origin? 🡨 can be inferred from Country of origin
  + COVID-heavy countries tend to require more social distancing and public has greater anxiety over contracting COVID-19
* Time zone 🡨 can be inferred from Country of origin
  + Time zone may influence sleep schedule, may be trickier to coordinate group projects and multi-person assignments
  + One upside with working with teammates in opposite time zones is that someone is always working on the final project
* Chronotypes
  + Students in other time zones may alter their chronotypes to accommodate STA302H1, which may offset adverse effects of sleep loss on performance.
* Time spent in STA302H1 during weeks 3 – 4 includes completing the mini assignments
  + This explains why Weeks 3 – 4 STA302H1 study times may be slightly inflated
* Sample size is roughly n = 200 people
  + Excludes students who dropped STA302H1
  + Blank entries and imputations for missing survey responses and missing quiz scores due to some students skipping quizzes

- r2 and adjusted r2 are smalls value. could improve this model to increase them

- missing data: presence of NA values + remove 28 dropped students, missing other survey data (country, covid hours, study hours)

Possible model improvements

* Covid = high stress
* Hours of sleep = predict stress + quiz
* Add country back to predict time zone = sleep schedule = hours of sleep (in original data, Mongolia was only significant country)
* Social time = might increase study time, reduces covid time,
* Physical activity = predict stress + increase study time + reduce covid time (cite studies)
  + Read title scholastic paper and cite, don’t read whole thing, introductions
* ~~Large sample size?~~ n = 199 is a good enough sample size
* Generalizability of model (by country)? – not really (mainly b/c online, data 3rd year students tend to study more and fail less than 1st student)?