**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).
* 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, or
  + country

1. Explain how you’ll develop the model.

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.
* Identify the number of quizzes students miss.
  + 1 – 2 NAs for quiz scores is OK, 3 – 4 NAs indicates the student dropped STA302H1 ~~Missing both quiz 2 and quiz 3 marks indicates dropped course~~
    - 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 or ~~TODO: categorize it as “Other.”~~
* Missing any number of COVID and STA302H1 hours is OK, those students either forgot or abstained.
  + A missing COVID or STA302H1 study time might mean that the student doesn’t feel comfortable sharing it.
  + These transgressions alone are no reason to exclude them from the dataset. They might still write quizzes though.
  + These variables have negligible effect on quiz 4 grades anyways

~~ORIGINAL IDEA: Personally, I think that analyzing countries separately is easier. This way, you can group results by continent too (e.g., America, Europe, Asia, etc.)~~

* + ~~Also, having to convert categorical variables into quantitative variables is tricky, and sometimes ambiguous.~~

~~EDIT: all countries except for Canada, China, and Unknown have enough data points. Other countries only have 1 – 3 data points so makes more sense to analyze data across all countries rather than by individual country.~~

~~When I tried to filter by country, it turns out all countries but Canada, China, and Unknown contain only 1 – 3 entries. When I tried to construct the correlation matrix, I obtained a bunch of NAs or r = +/- 1.~~

* + ~~Therefore, it doesn’t make sense to analyze each country separately. Instead, I’ve decided not to group data by country.~~
* ~~ANOTHER IDEA: Separate countries into three groups using factors~~
  + ~~TODO: Canada, China, and other (including unknown)~~
* You 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.
  + Grades are left skewed (mean < median)
  + Study times and COVID contemplations are right skewed (mean > median)
* Examine scatterplots to come up with possible relationships to try out
  + # quiz4 and quiz2, quiz3 = (negative) quadratic > linear
  + # quiz4 and covid2, covid3 = (positive) quadratic > linear
  + # quiz4 and study1 , ..,. linear?
* don't try sqrt() or log() -- too hard
  + Use correlation matrix
  + Look at scatterplot, guess relationship

1. How does the model meet the purpose mentioned earlier?

* Developing this model will serve the following purposes:
  + Helps current professors identify possible weak topics, reflect on what things they tried to do that helped students and things that didn’t help, and tailor their future lessons that way
  + Guides future professors establish course expectations, and common pitfalls so they have a chance to prepare extra well for those formative lectures
  + Help teaching stream professors develop carefully curated course content tailored to material where students score lowest.
  + Helps current students focus on key components to getting good Quiz 4 grades
  + Helps future students establish reasonable expectations and develop strategies to maximize their time and success in STA302H1

**Exploratory Data Analysis**

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

* 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.

* TODO: This would be identifying the most significant variables, right?
* Add descriptive statistics (i.e., histograms, boxplots, scatterplots)
  + TODO: insert scatterplots of strong relationships between quiz 4 score and one of the predictor variables
  + TODO: insert histograms of noteworthy relationships between quiz 4 score and one of the predictor variables
  + TODO: insert boxplots
* Display histograms, boxplots, pair scatterplots
* Describe each histogram – displays relationship between X and Y
* don’t discuss relationship results
* 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.

* Identify tentative model. 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
  + ~~TODO: Try a logarithmic transformation~~ TOO hard!
  + ~~TODO: Try a sqrt() transformation~~ Too hard!
* Compute 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

1. Add statistical and empirical justifications for your model.
   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

1. Add in-depth diagnostics to illustrate the goodness of the model
   1. TODO: Use residual plot and qqplot to assess goodness of fit

* 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.

- 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 one 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.

- test 4 assumptions for linear model:

- define cook's distance, leverage points -- no leverage points/influential points, so no points removed

- define homoscedasticity (constant variance)

**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)

- interpretation of final model

- quiz 3 scores -- closer to quiz 4 in difficulty, students better prepared and primed

- increase in study times between consecutive weeks (week 1 -> week 2 -> week 3 -> week 4)

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)?