Midterm Report

Your Midterm Report should be written in a style similar to a report that you would write for an employer, that is, it should be well organized and well-written in complete sentences. More marks will be awarded to submissions that are consistent with the style of a written report. As a reminder, you should not include the questions in your report, nor should you answer in a 'question and answer' format. Your report should be structured in paragraphs and sentences.

Your report must be typed. You may create your document in Word, Google Docs, LaTeX or any other word processor. Numerical and graphical summaries must be obtained using R.

Uploading your Midterm Report

(1) You must upload your Midterm Report according to the instructions given on Crowdmark to facilitate marking.
(2) You must upload your Midterm Report to the appropriate LEARN Dropbox as **one pdf file** so that your submission may be evaluated using plagiarism detection software.
(3) You must upload your dataset, in csv format, to the appropriate LEARN Dropbox
(4) You must upload a file of the R code (including comments) that you used to do the report in the appropriate LEARN Dropbox.

Penalties will be applied if the above instructions are not followed. Reports which are not typed will not be marked.

What to Do

For your Midterm Report you will be responding to a hypothetical scenario based on a new dataset of tweets. Instructions on how to download your Midterm dataset, and the variates it contains, are summarized in the document Midterm Dataset Information which is posted in the Midterm Report folder in LEARN. **You will need to read this entire document**, including the descriptions of the variates, to complete the Midterm Report.

Scenario: The Canadian government has hired you to advise their social media team on the matter of provincial health communication. To do so, you have gathered data from several provincial health agency Twitter accounts. Your report will help the government formulate a social media strategy for the period from 1 January 2022 to 31 March 2022. Your report is therefore expected to help the government anticipate the Twitter activity of provincial health agencies during this three-month period. It is hoped that your report will provide insights as to when, how often, and what material to tweet.

Your manager, who took STAT 231 several years ago and thinks it's super great, has provided you with an outline of what you should include in your report. Your report will follow the PPDAC structure introduced in Chapter 3 of the Course Notes.

Important: Your report should be structured with sections for Problem, Plan, Data, Analysis, and Conclusion. Each section should contain the items specified below. These items represent the minimum set of details and information that your report should include. As a reminder, you should not include the questions or your R code in your report. Instructions in Crowdmark indicate how the steps are to be uploaded to Crowdmark for marking purposes.

Problem

Clearly define the target population for the study. Be sure that you indicate what the units are in the target population.

In this study you are asked to analyse the following variates: `health`, `covid`, `vaccine`, `media`, `first.tweet`, `retweets`, `time.of.day`, `likes`, `username`, `is.retweet`. For each variate specify its type and give an attribute in the target population for this variate.

Example:
The variate `retweets` is a ___ type of variate. An attribute of interest for this variate is ___.

The Problem step includes the motivating questions for the study which are stated in terms of the attributes of the **target population**.

Motivating Questions

(a) In the target population what proportion of tweets contain a particular keyword (health, covid or vaccine)?

(b) Suppose that 50 tweets are drawn at random from the target population. What is the probability that at least half of these tweets will contain the keyword?

(c) In terms of media items, is there a difference between first tweets of the day compared to not first tweets of the day in the target population?

(d) What is the mean number of retweets in the target population?

(e) Does the Exponential model fit a transformation of the variate retweets in the target population?

(f) Is the distribution of the number of retweets in the target population different depending on whether the tweet contains a particular keyword or not?

(g) In the target population is the mean number of likes received by tweets which are tweeted during the time period $9:00-12:00$ different compared to the mean number of likes received by tweets which are tweeted during the time period $12:00-15:00$?

(h) Is there a difference among the provincial health agencies with respect to how often they use information retweeted from other accounts versus their own original tweets?

Specify the type of problem (descriptive, causative, predictive) for each of these motivating questions by completing the following sentence and including it in your report:

Motivating questions (),...,() are descriptive problems, (),...,() are causative problems, and (),...,() are predictive problems.

where you should place the relevant letters in the parentheses (). Note that this list contains at least two different types of problems.

Note: Because your manager gave you this list of motivating questions, it is not necessary to include these as part of your report. You may reference them with the letters (a), (b), etc.

Plan

What type of study is this and why?

Clearly define the study population or process for this study.

Describe at least two possible sources of study error, including justification for why you believe they are sources of study error. Be sure to reference attributes when discussing possible.

Describe the sampling protocol for this study in as much detail as possible. Be sure to indicate the sample size for your dataset.

Clearly identify one possible source of measurement error.

Data

In $1-2$ sentences, discuss any concerns or issues that you have in obtaining your sample (dataset).

Analysis

Complete the analyses given below as well as any other analysis you deem necessary. You should provide as much explanation as you think necessary for your manager to understand your report. Be sure to write in complete sentences. Since this is a report, your analyses should not contain any detailed mathematical derivations or calculations. All tables and plots require titles/labels as appropriate.

Note: The numbering below is only used to help with instructions for uploading the parts to Crowdmark for ease of marking.

(1) Your dataset contains three keyword variates (`covid`, `health`, and `vaccine`) that indicate whether or not each tweet contains a keyword.

You should decide which of these keywords you think is most important for the purposes of this study. You should then analyze the corresponding variate as follows:

Why did you choose this word?

Provide an estimate of the proportion of tweets in the study population that contain your chosen keyword. What model and method have you used to obtain this estimate? Explain why the model you used is reasonable.

Provide a 15% likelihood interval estimate for the proportion of tweets in the study population that contain your chosen keyword.

Suppose 50 tweets are chosen at random from the study population. Give an estimate of the probability that at least half of these tweets will contain your chosen keyword. Explain clearly how you obtained this estimate.

(2) Your dataset contains the variates `media` and `first.tweet`.

Complete the following table which summarizes the number of media items used for first tweets of the day and tweets which are not first tweets of the day.

	0	1	2	3	4	Total
First tweet						
Not first tweet						
Total						

What is the sample mode for the variate `media` for first tweets of the day? What is the sample mode for the variate media for tweets which are not first tweets of the day?

What is the sample mean for the variate `media` for first tweets of the day? What is the sample mean for the variate `media` for tweets which are not first tweets of the day?

Propose a model which could be used for modeling the variate `media` for first tweets of the day and justify your choice. What is/are the unknown parameter(s) in this model? Give the maximum likelihood estimate(s) for the unknown parameter(s).

(3) Your dataset contains the variate retweets. For this analysis you will need to create the transformed variate retweets.log = log(retweets + 1)

Give the five number summary and sample skewness for `retweets.log`.For ease of marking please put these in a table. Values may be rounded to two decimal places.

Assume that `retweets.log` follows an Exponential(λ) distribution. Provide the maximum likelihood estimate of λ. Indicate how the parameter λ relates to the attribute of interest in the study population.

Provide the plot of the relative frequency histogram of `retweets.log` with a superimposed Exponential($\hat{\lambda}$) probability distribution function.

(4) For this analysis you will be examining the variate `retweets.log` for the keyword you chose in part (1) of the Analysis step.

Provide the five-number summary for the number of retweets for tweets containing your chosen keyword. For ease of marking please put these in a table. Values may be rounded to two decimal places.

Provide the five-number summary for the number of retweets that do not contain your chosen keyword. For ease of marking please put these in a table. Values may be rounded to two decimal places.

Assume that `retweets.log` follows an Exponential(λ_0) distribution for tweets that **do not** contain your chosen keyword. Assume that `retweets.log` follows an Exponential(λ_1) distribution for tweets that **do** contain your chosen keyword. Provide the maximum likelihood estimates of λ_0 and λ_1.

Provide approximate 95% confidence intervals for λ_0 and λ_1 based on the appropriate asymptotic Gaussian pivotal quantity.

Provide a side-by-side boxplot of `retweets.log` for tweets not containing your word and tweets that contain your keyword.

(5) Your dataset contains the variate `likes` and `time.of.day`. For this analysis you will need to create a new variate from the `time.of.day` variate which is the time of day converted to hours. Use a 24 hour clock.

Provide the five-number summary for the number of `likes` received by tweets which were tweeted during the time period $9:00-12:00$. For ease of marking please put these in a table. Values may be rounded to two decimal places.

Provide the five-number summary for the number of `likes` received by tweets which were tweeted during the time period $12:00-15:00$. For ease of marking please put these in a table. Values may be rounded to two decimal places.

Assume a $G(\mu_{am}, \sigma_{am})$ model for the number of `likes` received by tweets which were tweeted during the time period $9:00-12:00$. Assume a $G(\mu_{pm}, \sigma_{pm})$ model for the number of likes received by tweets which were tweeted during the time period $12:00-15:00$.

Note: a Gaussian model may not be appropriate for these data. However, your manager has insisted you use a Gaussian model, so you should answer these questions assuming one is suitable.

Give the maximum likelihood estimates of μ_{am} and μ_{pm}. Give a 95% confidence interval for both of these parameters.

Give the maximum likelihood estimates of σ_{am} and σ_{pm}. Give a 90% confidence interval for both of these parameters.

(6) Your dataset contains the variate `username` and `is.retweet`

You have been asked to examine how often different provincial health agencies use information retweeted from other other accounts compared to their own original tweets.

Your analysis should include a proposed model, a discussion of whether the model is reasonable, and estimates (including interval estimates) of any relevant parameters.

Conclusion

The conclusions you make should be based on the Analysis step and should address the motivating questions $(a)-(h)$ in the Problem step.

Conclusions can only be made in relation to the **study population**.

Your conclusion should address any statistical uncertainty in your answers to the motivating questions.

Your conclusion should also include a discussion of at least three limitations of the study. Each limitation should be clearly described, with an explanation as to why and how what you identify is a limitation to the study.