Main task: helping the Royal Canadian Yacht Club (RCYC) gain insights into **how various subsets of their members use the facilities**. They are looking for your help to identify patterns in usage across demographic groups, so that they can better adapt to the needs of their members.

Resource: only data dictionary(Table2)

Caution:

1. You should make sure that your final insights are specific and tied to the data; i.e. not general recommendations like, "the RCYC should renovate their fitness facilities" which are not tied to these data. You may consider searching for additional sources of data to combine with this data, if you wish, but you are expected to use the data provided.
2. The above objective is very broad, so in your proposal, you will need to formulate specific questions to help the RCYC better understand how different groups of members use their facilities over time so that they can better adapt to members' needs.
3. You are NOT required to use all of the statistical methods covered in this class in your project. Rather, you should only use methods which make sense to answer the specific questions you want to address.

Content of the proposal:

1. 1-2 pages long.
2. Ideas for 3 interesting research questions
3. For each question include the following information:

(i) A clear statement of the specific research question you've formulated

- Make sure your questions are answerable based on the data you have
- Make sure your questions are answerable based on the methods we have covered so far in the course, or that we will cover later in the course
- Make sure your questions are interesting and meaningful (i.e. would the answer to this question help a manager at the RCYC better understand the demographics of club members and how the club can adapt to members' interests?)

(ii) Describe the population you are trying to make inference for in this specific question.

(iii) 2–3 sentences describing a visualization which you think might be interesting to explore this question (i.e. what type of plot would be useful to explore this question) and what variables you would use. You should also include a brief justification of why this type of visualization is appropriate and what information it will allow your

audience (the managers at the RCYC) to learn. *Hint: Think about a visualization that would be interesting and meaningful for managers at the RCYC, who aren't statisticians.*

(iv) 2–3 sentences describing the variables you plan on using to investigate this question (see Table 2 below) and the statistical method you will use.

4. Checklist for each question:

Would the RCYC be interested in knowing the answer to this question and / or would it help them understand RCYC club members and/or their interests better?

Do you have the data required to answer this question? (e.g. do we currently have the necessary variables or can we find them somewhere else?)

Have you learned/will you learn an appropriate statistical method which will allow you to answer this question?

Table 1 contents:

Week	Method	Type of variable(s) or hypotheses	Example questions
Week 4	Hypothesis test for one proportion	H_0: $p = p_0$ H_1: $p \neq p_0$ (be sure to define the parameters)	- Is the proportion of male students in this class 50%?
Week 5	Hypothesis test (Randomization test) comparing the means or proportions between two groups	Examples: H_0: $p_1 = p_2$; H_1: $p_1 \neq p_2$ *or* H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 \neq \mu_2$ (be sure to define the parameters)	- Is the proportion of male students in this class the same as Canada? (In case you're interested, the current national sex ratio is 49.63%!) - Is the mean time spent commuting to class similar between students who live with their family or those who live elsewhere (e.g. in a shared house, on campus, etc.)?
Week 6 & 7	Bootstrap confidence intervals	Focuses on a population parameter; e.g. population mean, median, proportion, etc.	- What is a range of plausible (i.e., reasonable) values for the average post-graduation salary of somebody with a UofT undergrad degree in statistics? - What is a range of plausible values for the average treatment effect of a new weight loss drug?
Weeks 8 & 9	Linear regression	- Response: numerical - Predictor(s): numerical and/or categorical - We will also learn how to make scatter plots (visualisations for two numeric variables)	- Based on some information that we know about a person (e.g. past grades, time spent studying, etc.), can we predict their grade on the STA130 midterm? - Is there an association between the number of tutorials attended and final STA130 grade?
Week 10	Classification trees	- Response: categorical - Predictors: categorical and/or numerical	- Based on information that we know about somebody (e.g. based on their age, gender, etc.), can we predict their favorite type of music? (e.g. pop, rock, rap, etc.)

Table 1: method generalization

Method 1: hypothesis test on 1 or 2 populations. Find the parameter of a population or compare the parameters of 2 populations. Better be numerical data.

Method 2: range method. Totally numerical data. Find the range of the parameter for a population.

Method 3: prediction method. Predict the parameter for a population based on the information we have.

Table 2

Variable name	Variable type (and sample values)	Description
MemberID	Categorical	Random Member ID for each member of the RCYC
Sex	Categorical ("M", "F")	Member's sex (coded as "M", "F")
YearJoined*	Numerical (integer)	Year that the member joined the RCYC
year	Numerical (integer)	Year for which the status, age, facility usage (uses_fitness, plays_racquet, and dock) and spending variables are measured. The data you will receive only contains data for 2017 and 2020
Status	Categorical ("ACTIVE", "HOLDOVER")	Takes the value "ACTIVE" for members active in a particular year, and "HOLDOVER" for members who put their membership on hold for that year (i.e. inactive members)
Age*	Numerical (integer)	Member's age on January 1st of a particular year.
fitness	Categorical ("Y", "N")	"Y" if the member uses RCYC fitness facilities, "N" otherwise
racquets	Categorical ("Y", "N")	"Y" if the member plays racquet sports at the RCYC, "N" otherwise
dock	Categorical ("Y", "N")	Does the member rent a dock at the RCYC? (this is only available for 2017.)
city_dining*	Numerical (integer)	Yearly amount spent on dining at the RCYC's restaurants in the city of Toronto (mainland); this is only available for 2017.
island_dining*	Numerical (integer)	Yearly amount spent on dining at the RCYC's restaurants on the Toronto Islands; this is only available for 2017.
bar_spending*	Numerical (integer)	Yearly amount spent in the RCYC's bars; this is only available for 2017.
other_spending*	Numerical (integer)	Other spending at RCYC facilities; this is only available for 2017.

Table 2: variables generalization

Sex(categorical): can be used for hypothesis.

Year joined(numerical): can be used for all.

Age(numerical): can be used for all.

City_dining(numerical): can be used for all.

Bar_spending(numerical): can be used for all.