In order to help Royal Canadian Yacht Club(RCYC) to improve its efficiency and economic profit, I am planning to establish three research questions on the composition of its members and their consumption patterns. These questions will be based on the sample data containing information about 1000 random members of the club.

My first question is sex-related: “Is the proportion of males in members with “ACTIVE” status (members active in a particular year) in the RCYC 50%?” The population I will make inference of in this question is all members of the RCYC with “ACTIVE” status from 2017 to 2020, whose parameter is the actual percentage of male in “ACTIVE” members. Since the sex variable is categorical, I will take it as the x-axis to generate a barplot to visualize the distribution of the sample data, in which there will be one bar for each category (in this case two) whose height represents the number of values in the corresponding category (in this case the number of male and female members). This plot will provide a clear and direct illustration of the number, proportion and difference of each sex in “ACTIVE” members of RCYC, and thus help deciding whether to focus more on men or women for advertising. I will utilize a hypothesis test for one proportion for this research question, taking variables “Sex” and “Status”. I will take “the proportion of males in members with “ACTIVE” status is equal to 50%” as my null hypothesis(H0), and “the proportion of males in members with “ACTIVE” status is not equal to 50%” as my alternative hypothesis(H1). I will figure out the extremeness of the test statistic (the proportion of males in “ACTIVE” members in the sample data) by calculating the p-value(the probability of observing data that are at least as extreme as the sample data, assuming my null hypothesis is true) and thus determining whether there are enough evidence against my null hypothesis.

The second question I am going to investigate is: “What is a range of plausible values for the average age of RCYC members who spent more than the median expenditure on the RCYC’s restaurants in Toronto?” I am inferring about both the population, which is all members of RCYC in 2017(including those who didn’t spend at all), and its parameter(or true value), the true average age of RCYC members who consumed more than the median value. I am going to use a histogram to visualize the test statistic(the average age of members who spent more than the median expenditure in city dining in this sample) not only because the age variable is numeric, but also due to the features of the histogram. Same as a barplot, a histogram reveals the statistics legibly by forming a list of bars, each assigned with the corresponding number of values(in this case the number of people who spent more than the median consumption). Moreover, since numeric data(ages) are continuous and ordered, a histogram shows the skewness, modality, center and variance of its distribution, delivering more specific message to the viewer. The approach I will apply to address this question is the bootstrap method, taking variables “Age” and “city_dining”. Starting with the sample data, I will generate 5000 bootstrap samples with replacement and calculate the statistics (in this case the average age of members who spent more than the median expenditure in city dining) for each one. Afterwards, I am going to determine a confidence level α, and use percentiles(the pth percentile is the smallest value that is larger or equal to p% of all the values) to find the corresponding confidence interval [a, b](meaning that if I repeat the bootstrapping process above many times, α% of their confidence levels would include the true value) for my population parameter, thus eventually draw the conclusion that I am α% confident that the mean age of RCYC members who spent more than the median expenditure in city dining is between a and b.

For my last question, I will focus on the usage of facilities in RCYC. The question will be: "What is a range of plausible values for the median age of RCYC members who use both fitness and racquet facilities?" This question makes inferences about members who used fitness and racquet facilities in RCYC from 2017 to 2020, taking the actual median age of fitness and racquet users as parameter(or true value). The plot I will apply is a boxplot, using the age variable to label the horizontal axis. Although this plot doesn't visualize the shape and spread of the distribution, it clearly represents the median value (age in this case), which is the exact the parameter of this question. In addition, it shows the first and third quartile, along with the interquartile range of the distribution, which are important data and reference for the bootstrap test method. I will investigate this question with the help of the bootstrap method, as I did with my second question. I will take "Age", "fitness" and "racquet" as variables and re-sample with replacement from the original sample to establish a sampling distribution, from which I will decide a confidence level β, and use quantiles(same as percentiles) to compute the confidence interval(defined exactly the same as in my second question) and conclude that I am β% confident that the actual median age of RCYC members who use both fitness and racquet facilities falls within the confidence interval.