Exploratory Statistical Analysis on Royal Canadian Yacht Club (RCYC) Dataset

with focus on understanding the dining spending habits and the preference of using fitness facilities of RCYC members

Project Group 15: Paul Tang, Jack Duan, Dongfang Yuan

April 1, 2021

Introduction

- Background: We will work with a dataset of 1000 randomly selected RCYC members. The variables in the dataset contain basic information of the members and their RCYC facilities usages. The variables have been jittered (i.e. random noise has been added to them) to anonymize the data. This project aims to identify patterns of how RCYC members use their facilities.
- Outline: First, we will use randomization test to study the difference in median dining spendings between RCYC members who rented a dock and those who did not. Then, we will use linear regression to study the association between RCYC members' spendings at RCYC bars and at RCYC restaurants. Finally, we will use classification tree to predict whether a member used RCYC fitness facilities based on his/her sex and other spendings at RCYC facilities (not counting restaurants and bars).

Research Question 1: Is there a difference between the median spendings in dining (i.e. dollars spent on RCYC's restaurants and bars) of RCYC members who rented a dock at the RCYC in 2017 and the members who didn't?

- Motivation: To compare the dining spendings at RCYC facilities of members who are dock renters and non-dock renters in the hope to learn more about the dining spending habits of RCYC members.

Type of statistical test employed: Randomization Test.

A pair of two hypotheses (called null hypothesis and alternative hypothesis) are formulated based on the research question. In this test, we have

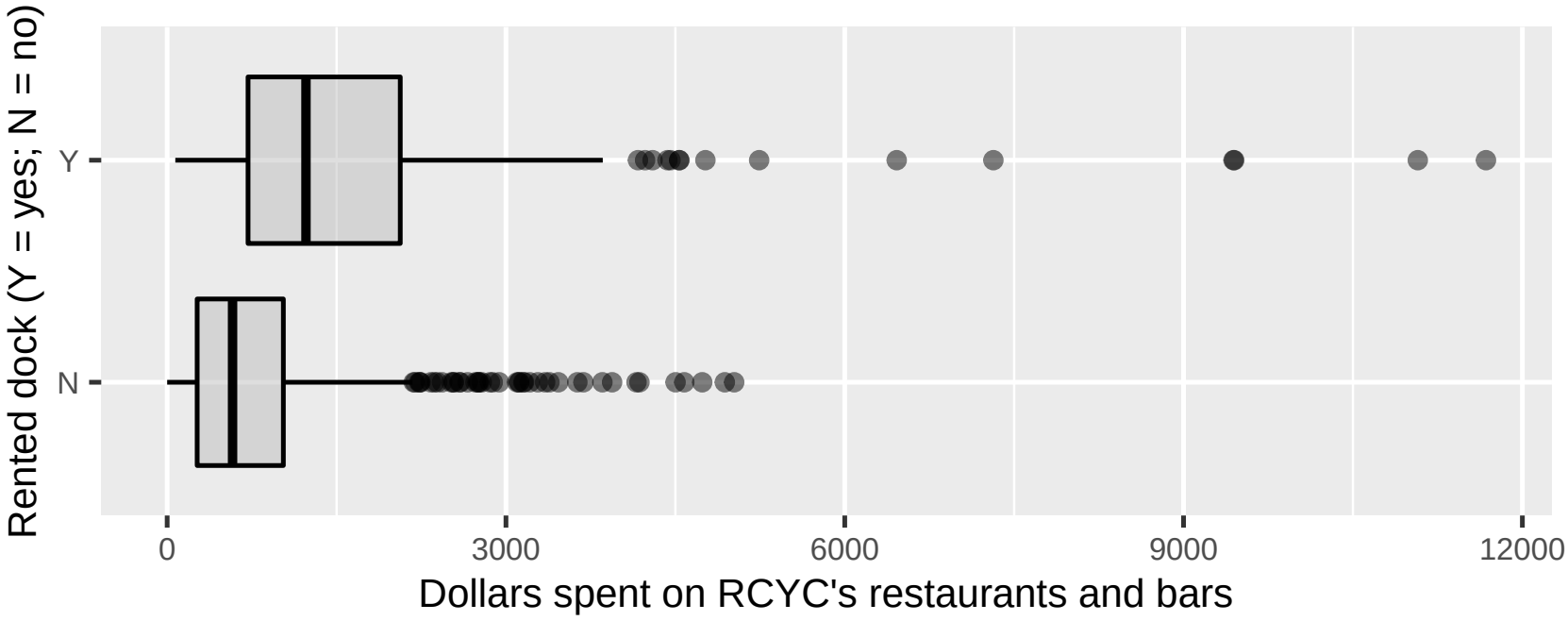
- Null hypothesis: There is no difference in the median spendings in dining at RCYC facilities between members who rented a dock and members who didn't rent a dock (in 2017).
- Alternative hypothesis: There is a difference in the median spendings in dining at RCYC facilities between members who rented a dock and members who didn't rent a dock (in 2017).

The randomization test answers if there is enough statistical evidence to reject the null hypothesis.

Data summary

- Variables used for this question:
  - *city_dining*: Yearly amount spent on dining at RCYC's restaurants in the city of Toronto (mainland) for 2017.
  - *island_dining*: Yearly amount spent on dining at RCYC's restaurants on the Toronto Islands for 2017.
  - *bar_spending*: Yearly amount spent in the RCYC's bars for 2017.
  - *dock*: Whether the member rent a dock at RCYC in 2017.
- Data wrangling (preparing the data for doing the statistical test):
  1. Removed all observations in the dataset whose entry for *dock* is NA (i.e. no entry for *dock*).
  2. Replaced the *city_dining* value of all observations in the dataset whose such value is NA to 0 (I decided to not remove observations whose *city_dining* is NA such as to not discard too much data).
  3. Same as 2 but for *island_dining* and *bar_spending* values.
  4. Created a new variable *dining_spendings* in the dataset that represents the sum of *city_dining*, *island_dining*, and *bar_spending*.

Visualization (box plot)



- In this dataset, the median spendings in dining at RCYC facilities of members who rented a dock is 1229 dollars, which is higher than that of members who didn't rent a dock, which is 579 dollars (visually, the line inside the respective gray "boxes" indicate the median spendings).
- There are few dock renters who spent much more in dining at RCYC facilities than others (i.e. spending exceeds 6000 dollars).

Randomization test result

- A metric used for determining whether there is enough statistical evidence to reject the null hypothesis is the non-negative number called p-value (the smaller the p-value, the more evidence we have to reject the null hypothesis).
- The p-value for this test is 0. This means if the null hypothesis is true, then it is highly unlikely (about 0%) that we will get the different median spendings in RCYC dining facilities between members who rented a dock and members who didn't that we see in this dataset. In short, the 0 p-value is a very strong evidence against the null hypothesis.
- Therefore, I reject the null hypothesis in favour of the alternative hypothesis. Thus, I conclude that it is very likely that there is a difference in the median spendings in dining at RCYC facilities between members who rented a dock and members who didn't rent a dock (in 2017) (in particular, the median spendings in RCYC dining facilities of members who rented a dock is very likely *higher* than that of members who didn't rent a dock).

Limitations of test result

Limitations

- Even though the Randomization Test result suggests that there is a difference in the median spendings in dining at RCYC facilities between members who rented a dock and members who didn't rent a dock (in 2017), the test cannot guarantee that this conclusion is necessarily true (further testing would be needed to establish this).
- Half of the data in the original dataset has to be discarded since they don't have a *dock* value. This smaller data size could contribute to an increased inaccuracy in the calculated p-value, thus influencing the result; however, this increased inaccuracy is likely to be insignificant to affect the result of the test by any notable amount.

Research Question 2: Is there a linear association between members' spendings at RCYC bars and at RCYC restaurants (in 2017)?

- Motivation: To explore the association between members' spendings at RCYC bars and restaurants in the hope to understand more about the dining spending habits of RCYC members.

Type of statistical test employed: Simple Linear Regression.

Linear regression uses the value of a variable (called predictor) to predict the value of another variable (called response). In this test, we have:

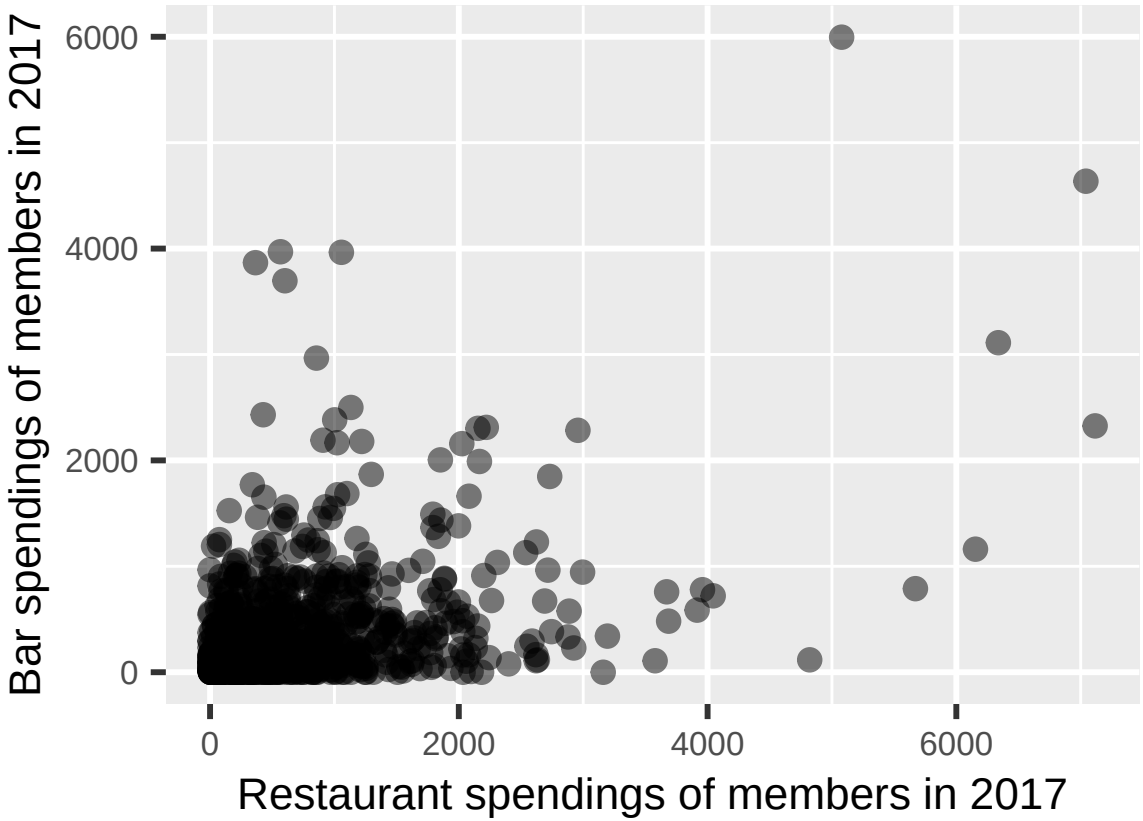
- Predictor: A member's spendings at RCYC restaurants.
- Response: A member's spendings at RCYC bars.

By assessing the accuracy of the predictions made by the linear regression model, we can determine the strength of the linear association between members' spendings in RCYC bars and restaurants.

Data summary

- Variables used for this question:
  - *city_dining*: Yearly amount spent on dining at RCYC's restaurants in the city of Toronto (mainland) for 2017.
  - *island_dining*: Yearly amount spent on dining at RCYC's restaurants on the Toronto Islands for 2017.
  - *bar_spending*: Yearly amount spent in the RCYC's bars for 2017.
- Data wrangling:
  1. Removed variables that are not *city_dining*,*island_dining*, or *bar_spending* from the dataset.
  2. Removed all observations in the dataset whose entry for *city_dining*, *island_dining*, or *bar_spending* is NA.
  3. Created a new variable *restaurant_spendings* in the dataset that represents the sum of *city_dining* add *island_dining*.
  4. Splited the remaining data to training (80%) and testing(20%).

Visualization (scatter plot)



- By the scatter plot, we can see that there is a weak to moderate positive (i.e. proportional) linear relationship between the restaurant spendings of members and the bar spendings of members.
- Most members in the dataset spend less than 2000$ at RCYC restautant and bars in 2017, respectively.
- The plot seems cone-shaped (see Limitations for implication).

Linear regression result

- The linear regression result suggests that on average, 100 dollar increase in a member's spending at RCYC restaurants correlates to 27 dollars increase in his spending at RCYC bars.
- The p-value (i.e. a metric used to determine if there is enough statistical evidence to establish a linear relationship between the predictor and response) is around 5.15e-38. Such a small p-value may indicate that we have very strong evidence that there is a linear association between the members' spendings at RCYC restaurants and at RCYC bars; however, this result may be invalid, see Limitations.
- The RMSE (i.e. a metric used to determine the prediction accuracy of the linear regression model) is around 505.89$. The large RSME indicates the accuracy of the linear regression model is not great despite there being a (plausible) linear association between the predictor and the response.
- All things considered, there does not seem to be a meaningful association between members' spendings at RCYC restaurants and at RCYC bars. However, future studies may try to include more predictors to achieve a better predictive model.

Limitations of test result

Limitations

- The scatter plot we obtained is cone-shaped. This fact violates one of the four assumptions that needs to be met in order for the p-value of the linear regression model to be valid. Therefore, the p-value we obtained may be invalid, and there may not be a linear association between the members' spendings at RCYC restaurants and at RCYC bars (this aspect is reflected by the scatterplot as well).
- Slightly more than half of the data in the original dataset has to be discarded since they don't have *city_dining*, *island_dining*, or *bar_spending* values. This smaller data size could contribute to an increased inaccuracy of the linear regression model, thus leading to a potentially higher RSME value. However, this increased inaccuracy is likely to be insignificant to affect the result of the test by any notable amount.

Research Question 3: Can we predict whether a member used RCYC fitness facilities in 2017 based on his/her sex and other spendings at RCYC facilities (not counting restaurants and bars)?

- Motivation: To explore whether a member's sex and his/her total spendings on RCYC facilities reflects his/her preference to use RCYC fitness facilities in the hope to understand more about what type of members prefer to use fitness facilities.

Type of statistical test: Classification Tree.

A classification tree uses the value of one or more variables (called predictors) to predict the (categorical) value of another variable (called response). In this test, we have:

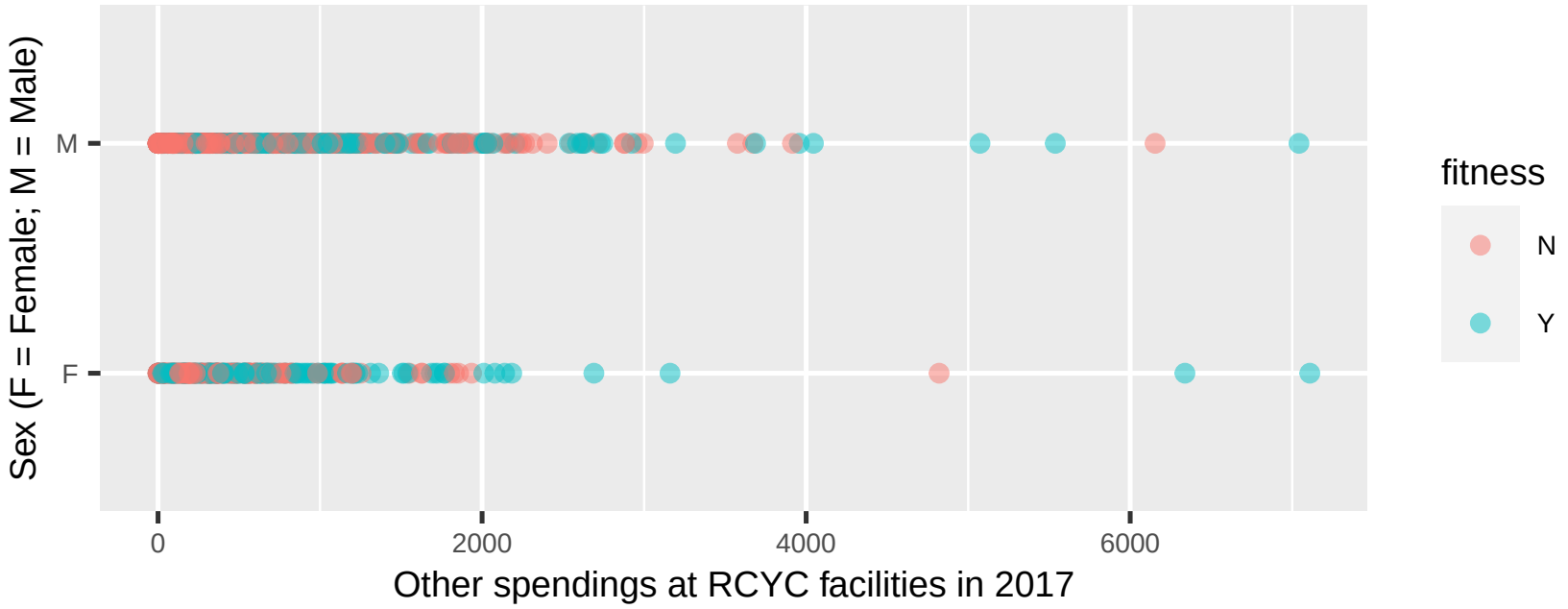
- Predictors: A member's sex; a member's spendings at RCYC facilities.
- Response: Whether the member used RCYC fitness facilities.

By assessing the accuracy and error rate of the predictions made by the classification tree, we can determine if a member's sex and his/her total spendings on RCYC facilities reflects his/her preference to use RCYC fitness facilities.

Data summary

- Variables used for this question:
  - *Sex*: The gender of members('M' for male and 'F' for female).
  - *Fitness*: "Y" if the member used RCYC fitness facilities in 2017, "N" otherwise.
  - *other_spending*: Other spendings at RCYC facilities in 2017.
- Data wrangling:
  1. Removed variables that are not *Sex*, *Fitness*, or *other_spending* in the dataset.
  2. Removed all observations in the dataset whose entry for *Sex* is NA.
  3. Splited the remaining data to training (80%) and testing (20%).

Visualization



- By the scatter plot, there is not a clear relationship between members' sex, their other spendings at RCYC facilities, and whether they used RCYC fitness facilities.
- It seems that many women who spent over 400$ at RCYC facilities used RCYC fitness facilities.
- It seems that a large porpotion of man did not use RCYC fitness facilities.

Classification tree result

- The accruacy of the classification tree (on testing data) is around 65%, which is not high. This suggests a member's sex and his/her total spendings on RCYC facilities (not counting restaurants and bars) are not meaningful predictors for knowing whether the member used RCYC fitness facilities or not.
- The classification tree predicted, with 30% error rate (i.e. 70% of the predictions are correct), that members who spent less than 493.5$ at RCYC facilities (not counting restaurants and bars) did not use RCYC fitness facilities.
- All things considered, a member's sex and his/her total spendings on RCYC facilities (not counting restaurants and bars) does not reflect accurately on his/her preference to use RCYC fitness facilities. However, future studies may try to include more meaningful predictors such as a member's age to achieve a better predictive accuracy.

Limitations of test result

Limitations

- Since we used 20% data for testing, only 80% data are used to train the classfication tree. This smaller data size could contribute to an increased inaccuracy of the classification tree. However, this increased inaccuracy is likely to be insignificant to affect the result of the test by any notable amount.

Conclusion

- Summary:
  1. The median spendings in RCYC dining facilities of members who rented a dock is very likely *higher* than that of members who didn't rent a dock.
  2. There does not seem to be a meaningful association between members' spendings at RCYC restaurants and at RCYC bars.
  3. A member's sex and his/her total spendings on RCYC facilities (not counting restaurants and bars) does not reflect accurately on his/her preference to use RCYC fitness facilities.
- Next steps:
  - It is recommended for futre studies to explore whether it is true that the median spendings in RCYC dining facilities of members who rented a dock is higher than that of members who didn't rent a dock.