The RCYC research plan: possible improvement suggested by data analysis

Yi Kuang

April 1, 2021

Introduction

In this project, we want to provide Royal Canadian Yacht Club (RCYC) with statistical approaches to better fit the members' needs and to optimize the overall service. The population we are going to investigate is all the members in RCYC and we will do our research based on the information of random 1,000 RCYC members. Hopefully, by showing three research cases, RCYC may gain certain insights.

Research questions & Objectives

- **Is there an association between the island dining spending and the bar spending?**
  Objective: It helps the RCYC better manipulate the financial distribution plan on the service for the island dining and the bar.
- **Is the mean island dining spending similar for those who rent a dock at RCYC and for those who do not?**
  Objective: The RCYC manager can take this as a reference to promote discounts or any sort of activities to satisfy more needs of the members and, in the meantime, discover more business opportunities.
- **Given whether the member plays racquet sports at the RCYC and the city dining spending for the member, can we predict if the member uses RCYC facilities?**
  Objective: It can help promote the RCYC facility maintenance scheme and related financial plans.

Data Summary

Table 1: Data used (The first two variables are categorical, and the rest is numerical)

Variable	Description
fitness	"Y" if the member uses RCYC fitness facilities, "N" otherwise
racquets	"Y" if the member plays racquet sports at the RCYC, "N" otherwise
city_dining	Yearly amount spent on dining at the RCYC's restaurants in the city of Toronto (mainland); this is only available for 2017.
island_dining	Yearly amount spent on dining at the RCYC's restaurants on the Toronto Islands; this is only available for 2017.
bar_spending	Yearly amount spent in the RCYC's bars; this is only available for 2017.

Research Question 1: Is there an association between the island dining spending and the bar spending?

Finding out the association between the spending for the island dining and the bar enables us to plan on whether there can be a possible transition between these two activities, so to make the members much convenient to attend another activity and, more importantly, maximize the RCYC's overall profits. For example, the RCYC's restaurants on the island can offer the member a discount after spending certain money or there can set a ride service allowing the members to commute between the bar and the restaurant.

To further explore this question, let us apply the linear regression model which enables us to gain insights into the statistics between these two variables, island_dining, and bar_spending.

Figure 1

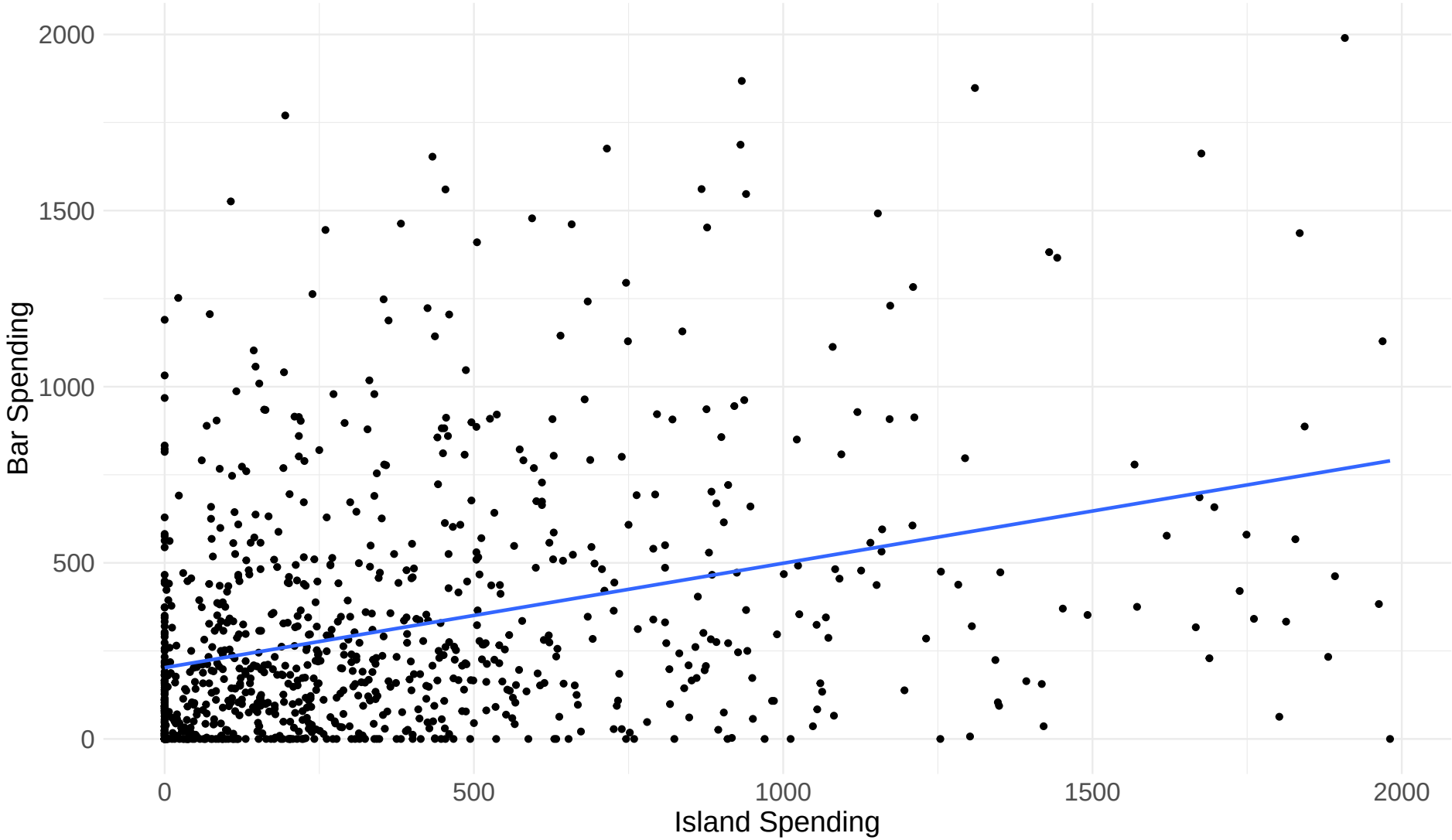

Figure 1 A linear regression model that illustrates the spending for the members dining at the RCYC's restaurants on the Toronto Islands and for the members in the RCYC's bar (Data for 2017)

Interpretation regarding the assocaition

Table 1

Table 1 Test statistically whether there is an association between the spending for the members dining at the RCYC's restaurants on the Toronto Islands and for the members in the RCYC's bar

```
##                  Estimate  Std. Error  t value     Pr(>|t|)
## (Intercept)   202.393299 14.61855174 13.84496 9.953125e-40
## island_dining   0.296429  0.02736763 10.83138 8.230681e-26
```

To figure out if there is an association, we may want to introduce a null hypothesis that there is no association between these two variables, and an alternative hypothesis that there is an association. In this case, we would like to apply the hypothesis test (p-test), a method we apply to test the validity of the hypothesis. Since the p-value, according to table 1, is 8.230681e-26 which is far less than 0.001, we have very strong evidence against the null hypothesis and, to a large extent, the alternative hypothesis is true. Thus, there is an association between the spending for the members dining at the RCYC's restaurants on the Toronto Islands and for the members in the RCYC's bar.

Interpretation regarding the association

After verifying if there is an association, we may want to check more details about this association.

The linear regression model attempts to explain the relationship between the island dining spending and bar spending. As the graph (see Figure 1) shown, the dots scattered on the plot represent the island and bar spending of an individual, and the blue line shown on the graph explains the relationship between these two variables for the whole population. Evidently, the slope of the line is positive, so we can infer that there is a positive association between these two variables. That is, generally, those who spend more money on the island service tend to spend more money on the bar service.

- To briefly summarize, according to the statistics, there is a positive association between island dining spending and bar spending.

Research Question 2: Is the island dining spending similar between those who rent a dock at RCYC and those who do not?

Since the island dining spending may be easily overlooked, this finding may help RCYC raise awareness towards this potential business opportunity or potential membership improvement. Evidently, finding out the difference allows RCYC to determine if there can be any favorable policy for those who rent or do not rent a dock (depending on the result), as this can help encourage the members to make more consumptions and ultimately promote the total benefit of RCYC. For example, those who enjoyed a meal at the island dining can be offered a maintenance ticket for the dock. Also, those who rent a dock can get a discount ticket at the island RCYC restaurant.

Thus, we may want to apply hypothesis test again, where it is an appropriate way to examine the difference. Yet, to make the data much approachable, the boxplots can help illustrate the data, and the data involved for this research question are island_dining and dock.

Figure 2

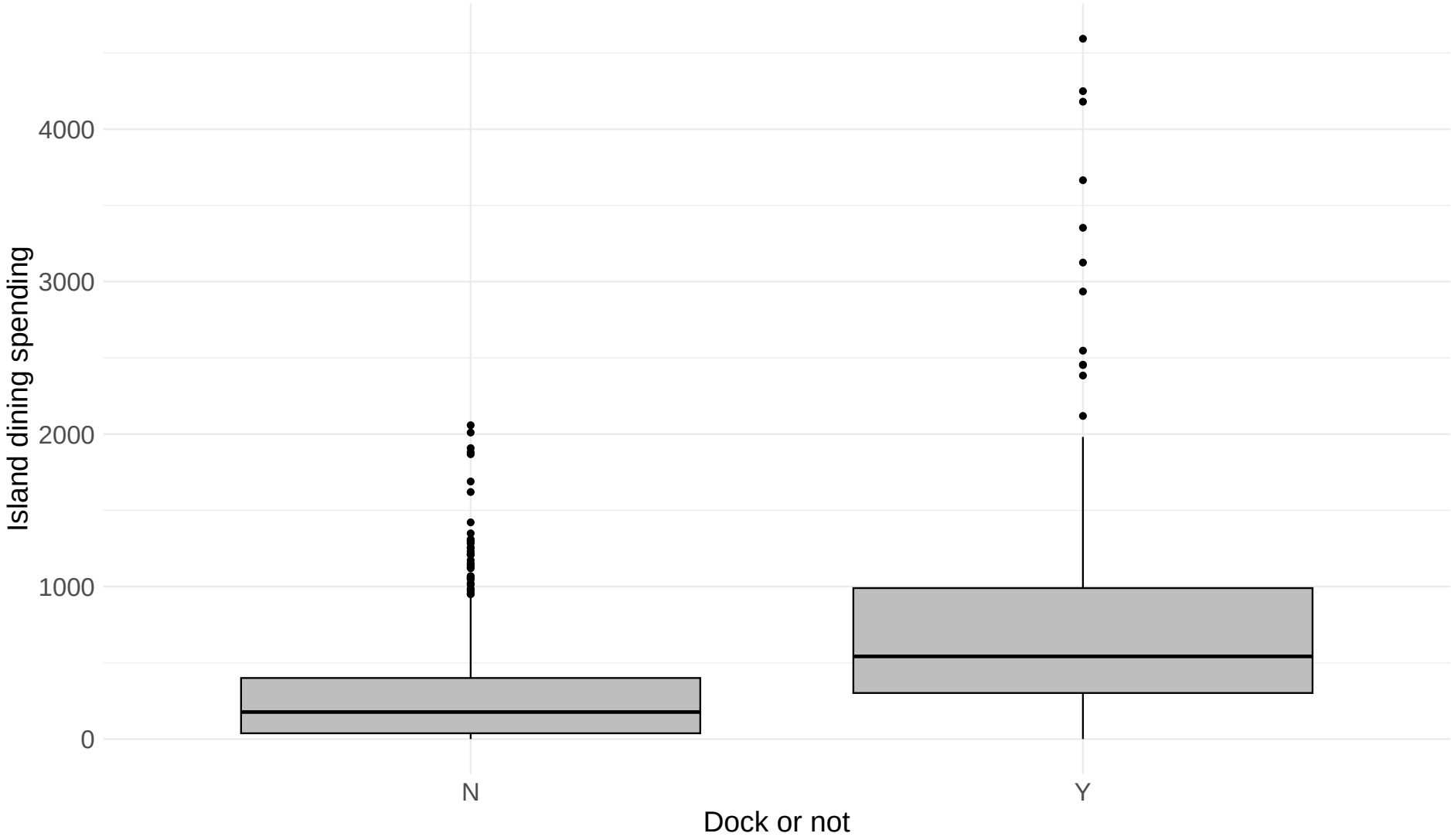

Figure 2 The boxplots that illustrate the island dining spending for those who rent and who do not, with those who rent on the right and who do not rent on the left (Data for 2017)

Interpretation regarding the boxplot

As can be seen from the figure (Figure 2), the ceiling for those who rent a dock is much higher than the counterpart, so does the median which is approximately 300 more than those who do not rent a dock. Also, the interquartile range of dining spending is wider among those who rent a dock than those who do not. In other words, there are more people who rent a dock spend more on island dining than those who do not.

Thus, it is easy to tell that renting a dock at RCYC can imply a greater island dining spending, so there is a difference between those who rent a dock and those who do not in terms of the island dining spending. However, we can not conclude it is the case yet. We have to testify the validity using hypothesis test (p-test).

Evidence regarding the interpretation from the boxplot

In this case, we may want to make a null hypothesis, H_0, that there is no difference between those who rent a dock and who do not rent a dock in terms of island dining spending, and an alternative hypothesis, H_1 that there is a difference.

The result of p-test is shown as follows:

```
## [1] 0
```

As can be seen, the p-value, an indicator that represents the strength of evidence against H_0, is 0 which is far less than 0.001 and we have very very strong evidence against the null hypothesis. To note, the smaller the p-value, the more 'evidence' we have against H_0.

- To briefly summarize, there is a difference between those who rent a dock and those who do not rent a dock in terms of island dining spending.

Research Question 3: Given whether the member plays racquet sports at the RCYC and the spending of city dining of the member, can we predict if the member uses RCYC facilities?

From this question, we may gain insights into how might the member playing racquet sports at the RCYC and the spending on the city dining would influence if the member uses RCYC facilities. This is the prediction we make about the status of the member using RCYC facilities. By doing so, RCYC may have the ability to pre-allocate resources on RCYC facilities such as staff distribution, financial support, and corresponding policies, so that the members can get facility supports in a much responsive way.

To make such a prediction, we need to implement a classification tree, a model that comes up with a prediction based on the input data. To note, the data involved in this case are fitness, racquets, and city_dining.

Figure 3

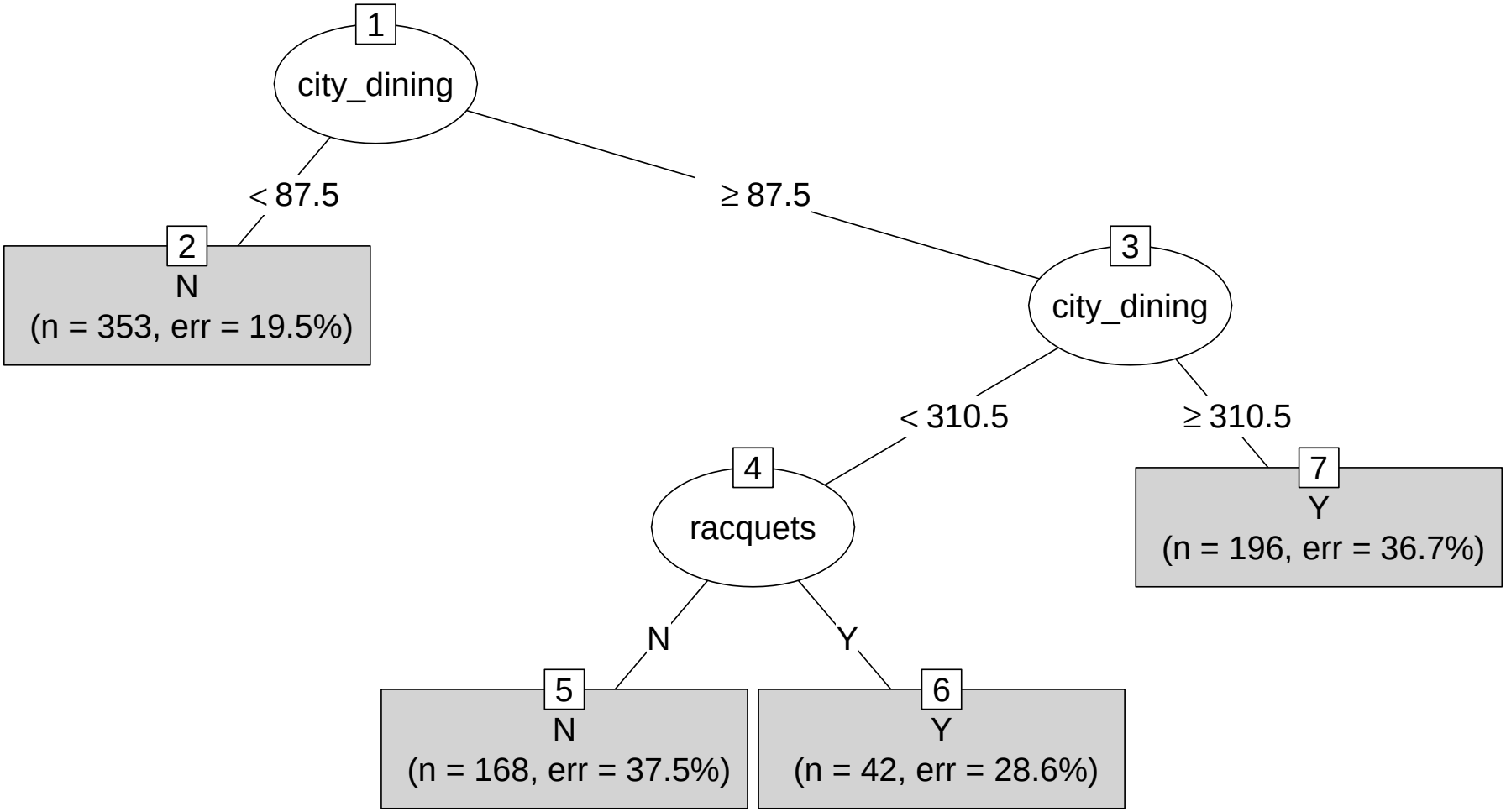

Figure 3 The classification tree that illustrates the prediction of the status for the member using RCYC fitness facility based on the data of the status of racquets and the city dining spending (Data for 2017)

Interpretation regarding the classification tree

As can be seen from the plot(Figure 3), we can predict the outcome by tracing the path along the tree. Though with a little bit of error, we can still hit the true outcome to a large extent.

For example, if a person spends 90 on the city dining and does not play racquets sports, this person is very likely that he/she uses a fitness facility in RCYC, according to the classification tree(Figure 3), but it is 37.5% likely that the prediction is incorrect. Take another example. If a person spends 10 on the city dining, regardless of providing the status of playing racquets sports, this person is likely that he/she does not use a fitness facility, with 19.5% of the chance that the prediction is incorrect.

Overall, it is a strong model that makes correct predictions most of the time. Yet, we still want to test this model(Figure 3) to see its prediction performance.

Table 2

Table 2 a confusion matrix that shows accuracy, sensitivity and specificity of the model and it shows how many are correctly/ incorrectly classified (Figure 2).

```
##
## tree_pred  N  Y
##          N 84 43
##          Y 26 37
```

As can be seen from Table 2, the accuracy which presents the proportion of true results as predicted is 121/190 (64%), the sensitivity which shows the proportion of those who use the facilities as predicted is 84/110 (76%), and the specificity which shows the proportion of those who do not use the facilities as predicted is 37/80 (46%).

- To briefly sumarize, we can predict whether if the member uses facilities or not with 64% of the predictions made being correct.

Limitation

- The data used is limited in terms of the year that the data is available. Since most of the data used are from 2017, we do not know what happens after 2017 and the data may not fit the current situation.
- The sample collected may not be representative, so the analysis may contain bias and the results may be accordingly inaccurate.
- For the hypothesis test (p-test as mentioned before), there may exist the possibility that we fail to reject the hypothesis that is actually false or rejects the hypothesis that is actually true.
- For the linear regression model we constructed, we are confident to figure out the association, but, if we want to predict the outcome using that model, the outcome yielded may be different from the true value.
- For the classification tree we built in the third section, the outcome may not always be correct and true, as can be seen from the confusion matrix.

Conclusion

Through exploring three aspects of the data from RCYC, we are glad that there are relationships between different variables which can be utilized and applied to better fit the RCYC members' needs and expand our business scope. At last, findings are concluded so as to provide RCYC with an overview of what we have brought to light using statistics. Besides, more details can be checked from the previous demonstration.

- There is a positive association between the spending for the mean island dining and the bar.
- There is a difference in terms of the island dining spending between those who rent a dock at RCYC and those who do not, and those who rent a dock tend to spend more on the island dining.
- We can predict if one uses fitness facilities in RCYC based on his/her city dining spending and the status for playing racquet sports in RCYC.

Hopefully, our data analysis brings the RCYC some valuable inspirations and we are looking forward to working in close collaboration with the RCYC in the future.