Prediction Tree Market Analysis
Joshua Kotzker

Background:

The analysis complete in this report is a prediction tree analysis of several categorical variables. Revenue will be predicted for AirBnB houses in San Bernardino County, California. For a more detailed walkthrough of the variables being observed, see appendix.

Prior to the report the data was cleaned. Detailed steps of cleaning are provided in the appendix. Initial observations found potential correlation between the predictors of zip code and host types. A graph demonstrating this initial observation can be observed within the appendix.

The goal of this study was to attempt to link the categorical variables to actionable predictions of revenue for certain conditions of AirBnb rentals. A prediction tree based off of an ANOVA model provides a clean and neat display of prediction for this project.

Methodology:

The predictive model of choice for this project is a prediction tree. The motivation behind this decision was that due to the fact that the dataset contained several predetermined classification groups.

The initial code for the first model was done below:

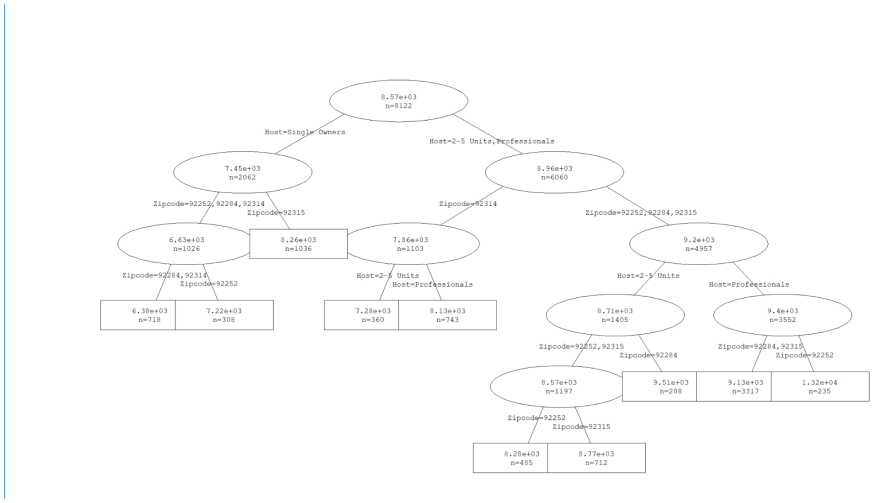
```
install.packages('rpart')
install.packages(c("zoo","xts","quantmod","abind"))
install.packages( "C:\\Users\\joshk\\Downloads\\DMwR_0.4.1.tar.gz", repos=NULL,
                  type="source" )
library(rpart)
library(DMwR)

#Creating training and test data
set.seed(555)
index <- sample(2, nrow(marketM), replace=TRUE, prob=c(0.8,.2))
treemarketMTrain <- marketM[index==1, 1:3]
treemarketMTest <- marketM[index==2, 1:3]


#Tree is made using the train data
ptree<-rpart(Revenue~Zipcode+Host,data=treemarketMTrain,method="anova",control =rpart.control(minsplit =1,minbucket=1, cp=0))
prettyTree(ptree)
```

The necessary packages were initialized, a test and train set were made.

A prediction tree was then created using no specific minimum number of node or bucket sizes.

Because the tree is limited to only a few categorical variables, there are only a few splits that can be made.

The resulting prediction tree is as follows:



The nodes and buckets of the tree represent their values in thousands. For example, this model calculates the predicted revenue of a single owner residence in the 92315 zip code to be $8260.

One key observation of this model is that it gives a clear distinction in predictive values for

single owner vs professional/multi home owner revenue. This is validating towards our initial assumptions of a discrepancy existing.

Before getting too excited, it is important to calculate the reliability of our model.

To calculate how good this model is, we can use the RMSE.

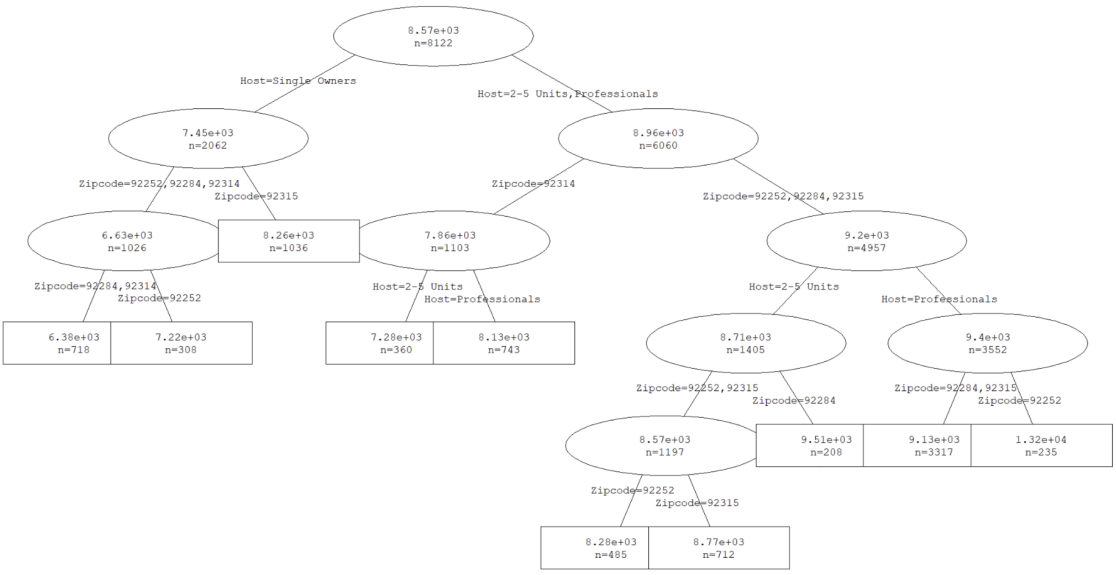
```
> #Assessing results
>
> yhat<-predict(ptree,treemarketMTest)
>
> y<-treemarketMTest$Revenue
>
> sse<-sum((y-yhat)^2)
>
> #RMSE
> sqrt(sse/nrow(treemarketMTest))
[1] 6986.726
```

Given our RMSE of 6986.726 compared to the mean of the dependent variable Revenue at around 8548.211, it is clear that this model is far from perfect. In fact, there is quite a bit of error in the model.

By changing the minimum split size and minimum bucket size, a lower RMSE was acquired. This was the only pruning that resulted in a lower RMSE.

```
> #Assessing results
>
> yhat<-predict(ptree,treemarketMTest)
>
> y<-treemarketMTest$Revenue
>
> sse<-sum((y-yhat)^2)
>
> #RMSE
> sqrt(sse/nrow(treemarketMTest))
[1] 6983.719
```

While this is a minimal change, it proves that further optimization was still possible.



Despite this model being far from perfect, there is some really interesting information to gather from it.

Conclusions:

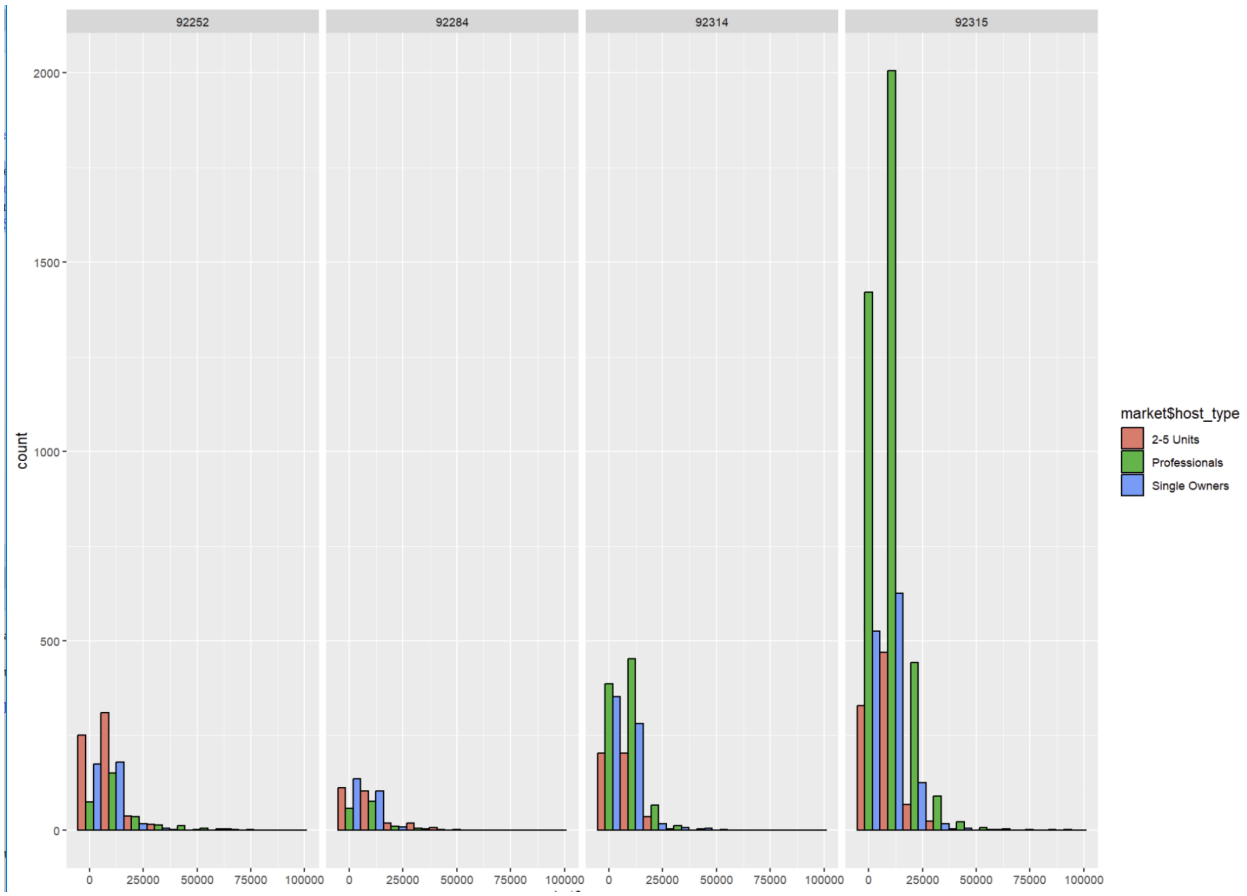
Firstly, single owners almost universally will make less revenue than 2-5 unit owners and professionals for every single county. This is completely plausible in the real world. Professional business men and women are fine tuned on what will make them profit and use homeowning as a means to make more money. Single owners may be renting out their primary residence or secondary residence simply for some side cash.

Secondly, the error rate of this model is high. This is plausible as the housing market can be sporadic and highly unpredictable. Calculating any model using a simple analysis tool like a prediction tree is wildly difficult. A more sophisticated model should be run on this data in the future.

Appendix:

Histograms collected in pre findings serve as motivation for the project.

A strong favor of revenue for renting professionals can be seen in 92315. Comparatively slighter favoritism exists within the other zip codes. Notably, single owners only barely take the edge in some instances within the 92284. Otherwise, they are typically at a disadvantage.



Below are my cleaning procedures that I committed prior to analyses.

Cleaning procedures:
1. On markeAnalysisFull many of the numbers substitute a "." for a ",". These should be switched. EX: marketAnalysisFull[1,12] outputs -> `449,9799957`

This should be $449.
(revenue, occupancy, nightly.rate,lead.time,length.stay)
2. guest column variables on marketAnalysisFull are characters currently. "15+" should just be "15"

3. Remove “AIR” from unified_id column on both amenities and geolocation and then convert the data type to double.
4. Horizontal merge amenities and geolocation location to marketAnalysisFull using unified_id variable and rename it something like “finalMarketAnalysis”
###5. It will probably be worth splitting the month column into month and year columns. Easiest way to do this would be just to create a new column called year, loop through the month column and grab the 20XX string. Paste it into the new column and delete it from the old. Then convert both to integers.