Restaurant Data Analysis and Future Profit Prediction

Andres Brett

Fall 2020

Data Preparation

Here we will need to take the following steps to successfully prepare the client's restaurant information.

1)Load necessary library's 2)Create a variable to house the client's information 3)Randomly sample the client's information into testing and training groups 4)Export the data to csv for future use and comment out the random sample syntax for consistency 5)Import the static data set for training and testing

```
#1)Load necessary library's
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------- tidyv

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v stringr 1.4.0
## v tidyr   1.1.2     v forcats 0.5.0
## v readr   1.3.1

## -- Conflicts ------------------------------------------------------------------------- tidyverse_c
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 4.0.3
```

```
library(tinytex)

#2)Create a variable to house the client's information
restaurants <- read.csv("restaurants.csv", header = TRUE)
restaurants <- as_tibble(restaurants)



#3)Randomly sample the client's information into testing and training groups
#rt.training <- sample_n(restaurants, 170)
#rt.testing <- filter(restaurants, !(ID %in%rt.training$ID))

#4)Export the data to csv for future use and comment out the random sample syntax for consistency
#write.csv(rt.training,"rt.training.csv", row.names = TRUE)
#write.csv(rt.testing,"rt.testing.csv", row.names = TRUE)

#5)Import the static data set for training and testing
rt.training <- read.csv("rt.training.csv", header = TRUE)
rt.testing <- read.csv("rt.testing.csv", header = TRUE)
```

Data Exploration

Problem 1

This problem asks how many restaurants market their food as traditional? And how many market their food as creative? Is there a significant difference in table turns between restaurants whose food is marketed as traditional versus creative? Also use a side-by-side box plot to present the table turns of the two types of restaurants.

We collected the restaurant's dataset mean, standard deviation, and boxplot graph to addresses this problem.

The next code chunk will do the following. 1)Create a new object to group restaurants by cuisine type. 2)Summarize the different cuisine type using mean, standard deviation, and count of restaurants that fall in each category. 3) Present the numerical results in a nicely formatted table

```
#1)Create a new object to group restaurants by cuisine type.
TableTurns <- group_by(restaurants, Cuisine)
#2)Summarize the different cuisine type using mean, standard deviation, and count of restaurants that f
TableTurnsGroup <- summarize(TableTurns,
                             Table_Turn_Mean = mean(Table.Turns),
                             Table_Turn_Sd = sd(Table.Turns),
                             Table_Turn_Size = length(Table.Turns)
                             )
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
#3) Present the numerical results in a nicely formatted table
kable(TableTurnsGroup)
```

Cuisine	Table_Turn_Mean	Table_Turn_Sd	Table_Turn_Size
0	202.3171	73.89179	82
1	186.8644	84.83971	118

With the table above we are able to determine that restaurants with traditional cuisine have a significantly higher weekly average table turn rate. Additionally, we are able to see that each type of cuisine does deviate from the mean in nearly the amount.

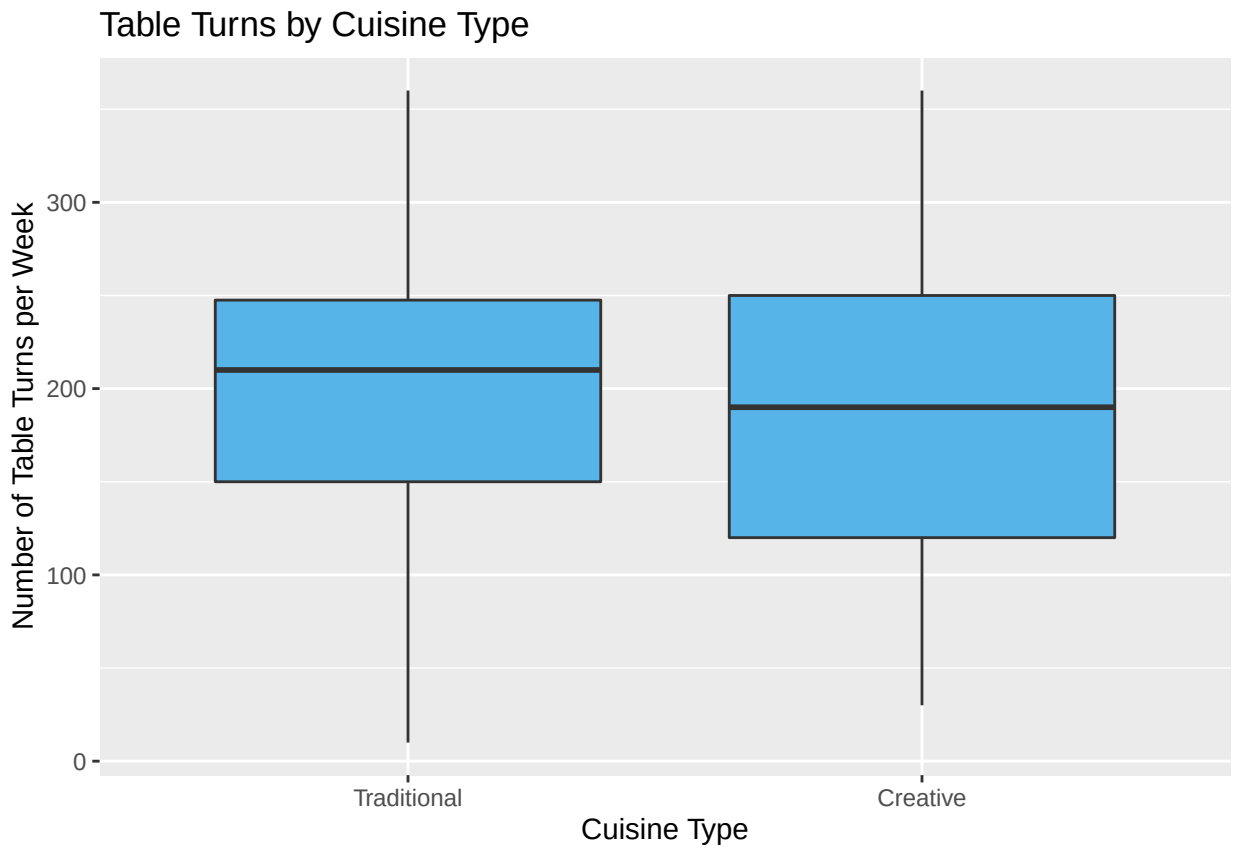
To better understand the data in our table, below we will be using a side-by-side boxplot to aide with visual comprehension.

The steps we take below are as follows... 1)Transform the Cuisine variable into observations that are simpler to understand (traditional & creative versus 0 & 1) 2)Create the Table Turns boxplot object 3)Present the side-by-side boxplot

```
#1)Transform the Cuisine variable into observations that are simpler
#to understand (traditional & creative versus 0 & 1)
restaurants$Cuisine <- as_factor(restaurants$Cuisine)
levels(restaurants$Cuisine) <- c("Traditional", "Creative")

#2)Create the Table Turns boxplot object
TableTurnsBoxPlot <- ggplot(restaurants, aes(x = Cuisine, y = Table.Turns)) +
  geom_boxplot(fill = "#56B4E9") +
  labs(title = "Table Turns by Cuisine Type",
       x = "Cuisine Type",
       y = "Number of Table Turns per Week")

#3)Present the side-by-side boxplot
TableTurnsBoxPlot
```



In the above side-by-side boxplot we can clearly see the traditional cuisine type with a higher average table turn rate. This is significant as it indicates a higher volume of customers. Also, we notice a tighter standard

deviation from the mean of traditional cuisines providing the business a more reliable source of income from customers.

Problem 2

Create a new variable "age" in the restaurants data set, which is the age of the restaurants in 2019. Report the mean, standard deviation, maximum, and minimum of the age variable.

Below we determine the restaurants age in 2019 and report out the coinciding average, standard deviation, maximum, and minimum ages.

We take the following steps: 1)Calculate and append the age (as of 2019) variable to a new object 2)Perform statistical calculations on the age variable

```
#1)Calculate and append the age (as of 2019) variable to a new object
restaurants_w_age <- mutate(restaurants,
       age = 2019 - restaurants$Year)

#2)Perform statistical calculations on the age variable
Mean_Age <- mean(restaurants_w_age$age)
Sd_Age <- sd(restaurants_w_age$age)
Max_Age <- max(restaurants_w_age$age)
Min_Age <- min(restaurants_w_age$age)

Mean_Age
```

```
## [1] 12.96
```

```
Sd_Age
```

```
## [1] 8.446604
```

```
Max_Age
```

```
## [1] 31
```

```
Min_Age
```

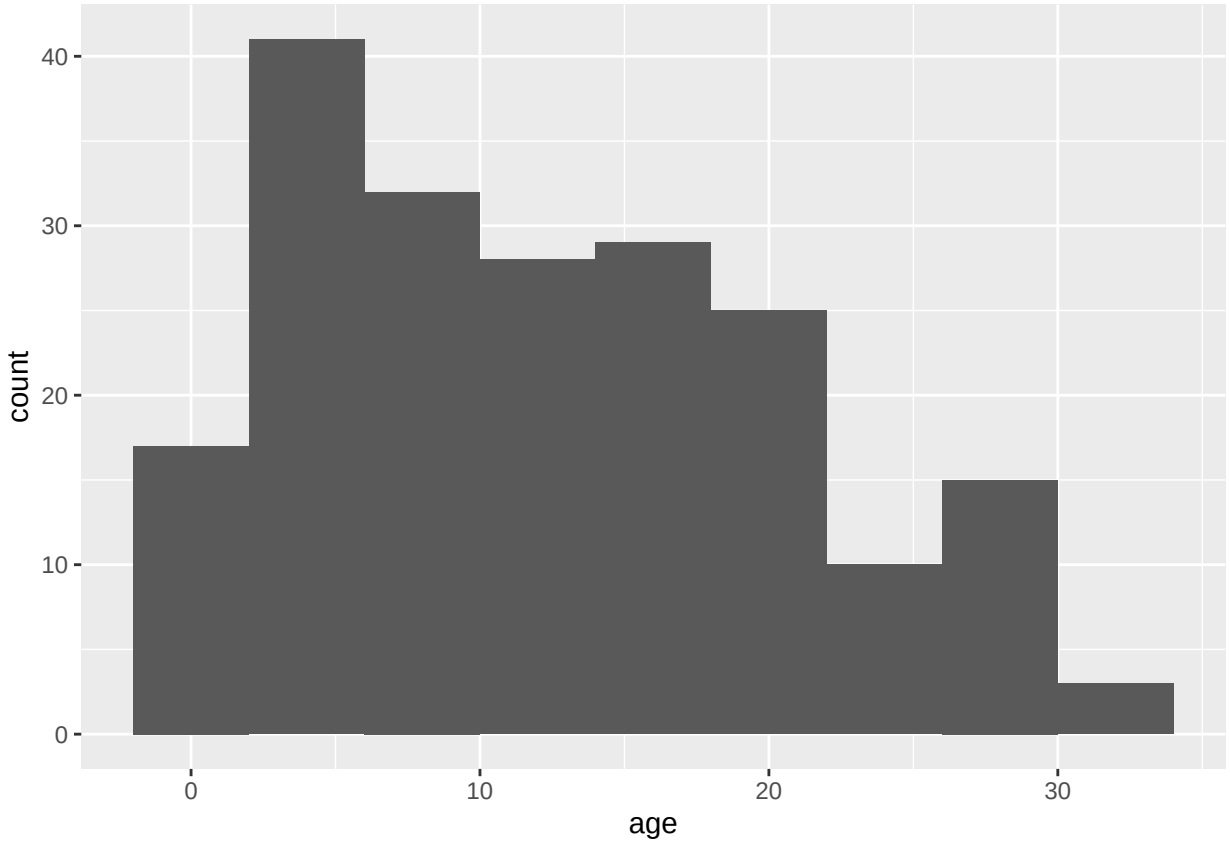
```
## [1] 0
```

The average age of our restaurants sample is 12.96, the standard deviation of our sample is 8.4466044, the max age of our sample is 31, and the minimum age of our sample is 0.

With these data points we can determine that a standard life cycle of restaurants is between 4.5133956 and 21.4066044... taking for granted there are outliers in both directions.

In addition, we present a histogram of the age variable. 3)Create a show a histogram of the age variable

```
#3)Create a show a histogram of the age variable
ggplot(restaurants_w_age, aes(x = age)) +
  geom_histogram(binwidth = 4)
```



The age histogram above was binned in intervals of 4 which displays a right tailed histogram. This shows that the number of restaurants open decreases as the years go on.

Regression Modeling and Interpretations

Problem 3

Fit a regression model to explain restaurant table turns. 1)Review the regression model using Table Turns as the response variable and all other variables as explanatory. 2)Determine if any variable in the data are not significant (p-value > .05) 3)Reevaluate regression model to only include significant variables. 4)Comment out unnecessary regression models. 5)Summarize results

```
#1)Review the regression model using Table Turns as the response variable and all other variables as ex
#2)Determine if any variable in the data are not significant (p-value > .05)
#4)Comment out unnecessary regression models.

#lm.restaurants <- lm(Table.Turns ~ ID + Advertising + Adsplay + Year + Days + Price + Parking + Rating
#lm.results <- summary(lm.restaurants)
#lm.results

#3)Reevaluate regression model to only include significant variables.
lm.restaurants2 <- lm(Table.Turns ~ Advertising + Adsplay + Rating + Price, data = restaurants)

#5)Summarize results
lm.results2 <- summary(lm.restaurants2)
```

Problem 4

Interpret the regression results. You should follow the 5 steps covered in class to analyze the results. The 5th step – prediction – is conducted in the next question. When interpreting the results, make sure you address your client's concern regarding what can be done to improve table turns. For example, your client is interested in knowing if advertising is worth it. What would you recommend your client to do?

Step 1

Interpret the overall model

```
lm.results2
```

```
##
## Call:
## lm(formula = Table.Turns ~ Advertising + Adsplay + Rating + Price,
##     data = restaurants)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -116.594  -30.372    2.326   30.106  139.877
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.026e+02  3.568e+01  -2.875  0.00449 **
## Advertising  8.466e-02  6.839e-03  12.379  < 2e-16 ***
## Adsplay      3.380e+00  2.744e-01  12.317  < 2e-16 ***
## Rating       1.045e+01  2.422e+00   4.312 2.56e-05 ***
## Price        1.325e+00  5.461e-01   2.427  0.01613 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 46.51 on 195 degrees of freedom
## Multiple R-squared:  0.6745, Adjusted R-squared:  0.6678
## F-statistic:   101 on 4 and 195 DF,  p-value: < 2.2e-16
```

In our example, it can be seen that p-value of the F-statistic is $<$ 2.2e-16, which is highly significant. We say the regression model overall is significant. This means that, at least, one of the explanatory variables is significantly related to the response variable.

Step 2

Here we will interpret the regression (beta) coefficients.

```
lm.results2.coef <- lm.results2$coefficients
lm.results2.coef
```

```
##                  Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept) -102.55946816 35.676801262 -2.874682 4.493375e-03
## Advertising    0.08465569  0.006838878 12.378594 2.389662e-26
## Adsplay        3.38011705  0.274418614 12.317375 3.662317e-26
## Rating        10.44610748  2.422396387  4.312303 2.563589e-05
## Price          1.32537627  0.546076885  2.427087 1.612948e-02
```

```
kable(lm.results2.coef, digits = c(3, 3, 3, 3))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-102.559	35.677	-2.875	0.004
Advertising	0.085	0.007	12.379	0.000
Adsplay	3.380	0.274	12.317	0.000
Rating	10.446	2.422	4.312	0.000
Price	1.325	0.546	2.427	0.016

```
View(restaurants)
```

The estimated coefficients are 0.0846557 for advertising, 3.380117 for Adsplay, 10.4461075 for Rating and 1.3253763 for Price.

These coefficients indicate that when all else is constant, as the yearly advertising expenditure increases by $1,000, the number of tables turned is increased by 0.0846557.

As the number of commercials play per week on web pages increase by 1 unit, the number of tables turned is increased by 3.380117.

As the overall service rating is increased by 1 unit, the number of tables turned is increased by 10.4461075.

As the the average spend per person is increased by $1, the number of tables turned is increased by 1.3253763.

Overall, advertising is not the best thing to invest in for additional customer volume from table turns. Instead the client should work to increase their overall service rating as this provides the highest return on investment.

Step 3

We determine the regression equation with the explanatory variable(s) identified in the previous step. The equation is written below.

Estimated Table Turns = -102.559 + 0.085 * Advertising + 3.380 * Adsplay + 10.446 * Rating + 1.325 * Price

Step 4

Next we assess the model using R-Square 1)Determine R Squared 2)Determine the Adjusted R Squared

```
#1)Determine R Squared
lm.results2$r.squared
```

```
## [1] 0.6745007
```

```
#2)Determine the Adjusted R Squared
lm.results2$adj.r.squared
```

```
## [1] 0.6678238
```

The R-square of the regression model is 0.6745007. It tells us that the regression model explains about 67.4500692% of the variance in the regression model.

Prediction and Validation

Problem 5

Based on the regression results, we predict which 10 restaurants in the testing set ranked highest in table turns. Clearly state your answer in the report (i.e., the 10 restaurant IDs).

We determined this by doing the following: 1)Predict estimated table turns using the regression model against the testing data set. 2)Append the predictions to the testing object 3)Sort the testing data set by highest estimated table turn and assign to new object 4)Create object and display the estimated top ten restaurant ID's (based on prediction)

```
#1)Predict estimated table turns using the regression model against the testing data set.
#2)Append the predictions to the testing object
rt.testing <- mutate(rt.testing,
       Estimated_Table_Turn = predict(lm.restaurants2, data.frame(rt.testing)))

#3)Sort the testing data set by highest estimated table turn and assign to new object
Table.Turn.Estimate.Sort <- arrange(rt.testing, desc(Estimated_Table_Turn))

#4)Create object and display the estimated top ten restaurant ID's (based on prediction)
Table.Turn.Sort.Top.10.Estimate <- head(Table.Turn.Estimate.Sort, 10)
Table.Turn.Sort.Top.10.Estimate$ID
```

```
##  [1]  43  62  11  93   3  42 148 141  19  50
```

Based on our results the top ten restaurants with table turns are 43, 62, 11, 93, 3, 42, 148, 141, 19, 50 (displayed with the highest table turn restaurant ID - far left, to the lowest - far right).

Problem 6

Here we validate our predictions using the real table turns values in the testing set: Are there any top 10 actual restaurants in our predicted top 10. 1)Sort the testing data by highest Table Turns and assign to a new object. 2)Create object and display the top ten restaurant ID's (based on prediction) 3)Compare prediction versus actuals

```
#1)Sort the testing data by highest actual Table Turns and assign to a new object.
Table.Turn.Sort <- arrange(rt.testing, desc(Table.Turns))

#2)Create object and display the top ten restaurant ID's (based on prediction)
Table.Turn.Sort.Top.10 <- head(Table.Turn.Sort, 10)
Table.Turn.Sort.Top.10
```

```
##     X  ID Advertising Adsplay Year Days Price Parking Rating Cuisine
## 1   1   3    1445.563      35 2006    6    52       0      7       1
## 2  25 148    1294.099      38 1989    5    50       0      7       1
## 3  10  43    2000.000      31 2016    5    66       0      7       1
## 4  11  50     689.547      46 2010    5    62       0      7       1
## 5  12  62    1500.000      38 2014    5    65       0      8       0
## 6   9  42      50.000      63 2012    3    66       0      7       1
## 7   3  10     174.093      40 2009    5    56       1      7       1
## 8  17  93    1600.000      24 2010    5    65       0      9       0
```

```
## 9   4  11    1720.806      32 2004    5    63       1      7       1
## 10 20 105     305.268      54 1996    3    64       0      6       1
##    Table.Turns Estimated_Table_Turn
## 1          360             280.1621
## 2          360             274.8294
## 3          340             332.1331
## 4          340             266.5961
## 5          340             322.5868
## 6          310             275.2183
## 7          300             194.7270
## 8          300             294.1769
## 9          290             307.9018
## 10         290             253.3103
```

```
#3)Compare prediction versus actuals
View(Table.Turn.Sort.Top.10.Estimate)
View(Table.Turn.Sort.Top.10)
```

Based upon our prediction results, 8 of the top 10 predicted table turn restaurants are in the top 10 actual table turn restaurants. This means that we can strongly rely on the prediction and regression models that have been developed.

The data points provided from the client can be used to to help determine what can be done to improve table turns. That is, to improve Table Turns is to increase the clients investment on overall service ratings and Adsplay. For future predictions, we would provide and use the following estimated regression equation there Estimated Table Turns = -102.559 + 0.085 * Advertising + 3.380 * Adsplay + 10.446 * Rating + 1.325 * Price.