BA Homework 2

Devarshi Pancholi

9/13/2019

PROBLEM 1: Citibike Analysis

Analytics Questions:

1. Compute summary statistics for tripduration

```
library(readr)
Citi <- read_csv(file= "/Users/devarshipancholi/Downloads/JC-201709-citibike-tripdata.csv")
```

```
## Parsed with column specification:
## cols(
##   tripduration = col_double(),
##   starttime = col_character(),
##   stoptime = col_character(),
##   `start station id` = col_double(),
##   `start station name` = col_character(),
##   `start station latitude` = col_double(),
##   `start station longitude` = col_double(),
##   `end station id` = col_double(),
##   `end station name` = col_character(),
##   `end station latitude` = col_double(),
##   `end station longitude` = col_double(),
##   bikeid = col_double(),
##   usertype = col_character(),
##   `birth year` = col_character(),
##   gender = col_double()
## )
```

```
summary(Citi$tripduration)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##      61.0     238.0     355.0     756.9     610.0 2181628.0
```

```
mean(Citi$tripduration)
```

```
## [1] 756.902
```

```
var(Citi$tripduration)
```

```
## [1] 159480876
```

```
range(Citi$tripduration)
```

```
## [1]      61 2181628
```

2. Compute summary statistics for age

```
Current <- 2019
Citi$'Age' <- Current - as.numeric(Citi$'birth year')
CitiC <- na.omit(Citi)
head(CitiC,10)
```

```
## # A tibble: 10 x 16
##    tripduration starttime stoptime `start station ~ `start station ~
##           <dbl> <chr>     <chr>               <dbl> <chr>
##  1          364 9/1/17 0~ 9/1/17 ~             3183 Exchange Place
##  2          357 9/1/17 0~ 9/1/17 ~             3187 Warren St
##  3          432 9/1/17 0~ 9/1/17 ~             3195 Sip Ave
##  4          934 9/1/17 0~ 9/1/17 ~             3272 Jersey & 3rd
##  5          932 9/1/17 0~ 9/1/17 ~             3272 Jersey & 3rd
##  6          625 9/1/17 0~ 9/1/17 ~             3194 McGinley Square
##  7          178 9/1/17 0~ 9/1/17 ~             3183 Exchange Place
##  8          557 9/1/17 0~ 9/1/17 ~             3183 Exchange Place
##  9          220 9/1/17 0~ 9/1/17 ~             3187 Warren St
## 10          153 9/1/17 1~ 9/1/17 ~             3272 Jersey & 3rd
## # ... with 11 more variables: `start station latitude` <dbl>, `start
## #   station longitude` <dbl>, `end station id` <dbl>, `end station
## #   name` <chr>, `end station latitude` <dbl>, `end station
## #   longitude` <dbl>, bikeid <dbl>, usertype <chr>, `birth year` <chr>,
## #   gender <dbl>, Age <dbl>
```

```
summary(CitiC$'Age')
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   32.00   36.00   38.87   44.00  132.00
```

```
mean(CitiC$'Age')
```

```
## [1] 38.87496
```

```
var(CitiC$'Age')
```

```
## [1] 100.9908
```

```
range(CitiC$'Age')
```

```
## [1]  18 132
```

3. Compute summary statistics for tripduration in minutes

```
Minutes <- 60
Citi$'tripMin' <- as.numeric(Citi$'tripduration')/Minutes
summary(Citi$tripMin)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     1.02     3.97     5.92    12.62    10.17 36360.47
```

```r
sd(Citi$tripMin)
```

```
## [1] 210.4762
```

4. Compute the correlation between age and tripduration

```r
cor(CitiC$Age, CitiC$tripduration)
```

```
## [1] 0.007055148
```

Business Questions:

1. What is the total revenue assuming all users riding bikes from 0 to 45 minutes pay $3 per ride and user exceeding 45 minutes pay an additional $2 per ride.

```r
x=0
y=0
for(a in 1:nrow(Citi)){
  Citi$Cost[a]= 0
  if(Citi$tripMin[a]<46){
    Citi$Cost[a]= 3
  }
  else{
    Citi$Cost[a]= 5
  }
}
```

```
## Warning: Unknown or uninitialised column: 'Cost'.
```

```r
sum(Citi$Cost)
```

```
## [1] 100601
```

Hence the total revenue of all the users present in the dataset is $100,601.

2. Looking at tripduration in minutes, what can you say about the variance in the data.

```r
var(Citi$tripMin)
```

```
## [1] 44300.24
```

The variance here is 44,300 minutes approximately. This means that the data is spread out. This is because we have a very huge maximum in the datase of about 36,360 minutes.

3. What does this mean for the pricing strategy?

From question 3 above we have seen that 3rd quartile in the dataset is 10.17 minutes. This means 75% of the data is we have about our tripduration in minutes is falls within those 10 minutes. Our mean for tripduration in minutes is 12.62 minutes. This means that majority of the customers hire the bikes for about 15 minutes. According to the current pricing strategy, we charge $3 for first 45 minutes. But in this case, it would make more sense if we reduce those minutes for the flat rate.

So according to my understanding charging a flat rate of USD 3 for 25 minutes will be more appropriate as most users who go above that time period will have to pay an additional of USD 2. This will help out in increasing revenue for the company without much difference for most of the exisisting users. This will be benifical for customer retention. Here is the total revenue calculation according to our new strategy:

```
x=0
y=0
for(a in 1:nrow(Citi)){
  Citi$Cost[a]= 0
  if(Citi$tripMin[a]<26){
    Citi$Cost[a]= 3
  }
  else{
    Citi$Cost[a]= 5
  }
}
sum(Citi$Cost)
```

```
## [1] 103195
```

As visible above, we have already increased the revenue by $3,195(approximately). This can be tweaked a little to maximise profits

4. What does this mean for inventory availability?

As majority of our customers spend about 15 minutes on a ride and with the new pricing strategy which limits the user for about 25 minutes for the initial charge of $3, it would be pretty safe to say that Citi wont need as much inventory as before. Most of the bikes would be rotating between different stations within those 25 minutes. Citi can cut some of the inventory costs too in order to increase their revenues.

PROBLEM 2: Zagat Descriptive Analytics

Analytics Questions:

1. What can you say about the central tendency of the ratings? 2. What can you say about the spread and dispersion of the ratings?

```
library(readr)
Zag <- read_csv(file= "/Users/devarshipancholi/Downloads/zagat.csv")
```

```
## Parsed with column specification:
## cols(
##   Name = col_character(),
##   Food = col_double(),
##   Decor = col_double(),
##   Service = col_double(),
##   Price = col_double()
## )
```

```
head(Zag,10)
```

```
## # A tibble: 10 x 5
##    Name                 Food Decor Service Price
##    <chr>               <dbl> <dbl>   <dbl> <dbl>
```

```
##  1 107 West               16    13      16    26
##  2 2nd Street cafe        14    13      15    21
##  3 44 & Hell's kitchen    22    19      19    42
##  4 55 wall                21    22      21    54
##  5 55 wall street         21    22      21    54
##  6 92                     15    15      15    43
##  7 Angelica kitchen       20    14      15    22
##  8 Angelo's               21    11      14    22
##  9 Avenue                 18    14      14    36
## 10 Avra estiatorio        24    21      20    50
```

```
library(psych)
ZagC <- Zag[-1]
describe(ZagC)
```

```
##         vars   n  mean    sd median trimmed   mad min max range  skew
## Food       1 300 19.39  3.69     19   19.39  4.45   9  28    19 -0.09
## Decor      2 300 15.72  4.95     16   15.82  4.45   3  27    24 -0.18
## Service    3 300 16.90  3.57     16   16.72  4.45   8  26    18  0.39
## Price      4 300 36.55 14.88     35   35.80 16.31   8  80    72  0.45
##         kurtosis   se
## Food       -0.41 0.21
## Decor      -0.28 0.29
## Service    -0.37 0.21
## Price      -0.39 0.86
```

```
harmonic.mean(Zag$Food)
```

```
## [1] 18.62024
```

```
harmonic.mean(Zag$Decor)
```

```
## [1] 13.4959
```

```
harmonic.mean(Zag$Service)
```

```
## [1] 16.15621
```

```
harmonic.mean(Zag$Price)
```

```
## [1] 30.12219
```

From Central tendencies of the rating, it is safe to say that the data is evenly distributed. Price has the maximum range of 72 as well as mean of 36.55. All other rating dimensions are comparatively low. So i decided to go for harmonic mean as it gives equal weights to all data points and the difference wasn't noteworthy.

As for the spread and the dispersion, Price has the higest spread of 72.

3. What are the correlations between rating dimensions?

```
ZagC <- Zag[-1]
cor(ZagC)
```

```
##              Food     Decor   Service     Price
## Food    1.0000000 0.3626698 0.7097296 0.5378984
## Decor   0.3626698 1.0000000 0.7263683 0.7789019
## Service 0.7097296 0.7263683 1.0000000 0.8487140
## Price   0.5378984 0.7789019 0.8487140 1.0000000
```

4. Using the information in 1-3, design a weighted average (index) that computes scores for each restaurant. Your index needs to reflect which ratings (decor vs food vs service vs price) you wish to amplify with loads/weights

```
for(a in 1:nrow(Zag)){
  Zag$WFood[a] <- (Zag$Food[a]*2)/300
  Zag$WDecor[a] <- (Zag$Decor[a]*2)/300
  Zag$WService[a] <- (Zag$Service[a]*3)/300
  Zag$WPrice[a] <- (Zag$Price[a]*3)/300
}
head(Zag,10)
```

```
## # A tibble: 10 x 9
##    Name              Food Decor Service Price  WFood WDecor WService WPrice
##    <chr>            <dbl> <dbl>   <dbl> <dbl>  <dbl>  <dbl>    <dbl>  <dbl>
##  1 107 West            16    13      16    26 0.107  0.0867     0.16   0.26
##  2 2nd Street cafe     14    13      15    21 0.0933 0.0867     0.15   0.21
##  3 44 & Hell's kit~    22    19      19    42 0.147  0.127      0.19   0.42
##  4 55 wall             21    22      21    54 0.14   0.147      0.21   0.54
##  5 55 wall street      21    22      21    54 0.14   0.147      0.21   0.54
##  6 92                  15    15      15    43 0.1    0.1        0.15   0.43
##  7 Angelica kitchen    20    14      15    22 0.133  0.0933     0.15   0.22
##  8 Angelo's            21    11      14    22 0.14   0.0733     0.14   0.22
##  9 Avenue              18    14      14    36 0.12   0.0933     0.14   0.36
## 10 Avra estiatorio     24    21      20    50 0.16   0.14       0.2    0.5
```

```
for(a in 1:nrow(Zag)){
  Zag$Score[a] <- (Zag$WFood[a] + Zag$WDecor[a] + Zag$WService + Zag$WPrice)
}
head(Zag,10)
```

```
## # A tibble: 10 x 10
##    Name        Food Decor Service Price  WFood WDecor WService WPrice Score
##    <chr>      <dbl> <dbl>   <dbl> <dbl>  <dbl>  <dbl>    <dbl>  <dbl> <dbl>
##  1 107 West      16    13      16    26 0.107  0.0867     0.16   0.26 0.613
##  2 2nd Stree~    14    13      15    21 0.0933 0.0867     0.15   0.21 0.6
##  3 44 & Hell~    22    19      19    42 0.147  0.127      0.19   0.42 0.693
##  4 55 wall       21    22      21    54 0.14   0.147      0.21   0.54 0.707
##  5 55 wall s~    21    22      21    54 0.14   0.147      0.21   0.54 0.707
##  6 92            15    15      15    43 0.1    0.1        0.15   0.43 0.62
##  7 Angelica ~    20    14      15    22 0.133  0.0933     0.15   0.22 0.647
##  8 Angelo's      21    11      14    22 0.14   0.0733     0.14   0.22 0.633
##  9 Avenue        18    14      14    36 0.12   0.0933     0.14   0.36 0.633
## 10 Avra esti~    24    21      20    50 0.16   0.14       0.2    0.5  0.72
```

Seeing the correlations between the rating dimensions, I have dedided to give weights as following: Food: 2 Decor: 2 Service: 3 Price: 3

Taking the mean with those weights and summing up all four rating dimensions, I have came up with a score system in a range of 0 to 1. This means the first restaurant in our dataset has a rating of 0.613/1.000.

Business Questions:

1. What makes a business more profitable?

Profit = Revenue - Expenses. So the more the revenue generation the better for the business. On the other hand, expenses should be as minimum as possible to increase the profit margin. In our case, the restaurant owners have to be aware of what rating dimension is most important for their restaurant. If they focus on maintaining/developing those dimensions, revenues shall be increased.

2. If you were hired to advise a new restaurant operator, what would you recommend in terms of the balance & trade-offs between food, decor, service, and price?

First of all, I need to find out which rating dimensions are critical for the restaurant based on the setting and the location of the place. Suppose a restaurant is a simple food cart along the sidewalk, "Decor" and "Service" is not at all important for such restaurant. As a result of this, "Food" and "Price" becomes increasingly important to the food cart. Good quick food at relatively cheap price is the goal for such restaurants.

On the other hand, if i am hired by a michelin star restaurant, I need to leave the "Food" dimension upto the chef there. As the food will be priced high, a lot of focus needs to be put on the "Service" as we have seen almost 85% correlation between "Price" and "Service" in our analysis. "Decor" needs to be on point too.

Generalizing all of these, as per my correlation analysis, "Price" and "Service" goes hand in hand. There is no direct dependency between "Food" and "Decor". This means you can sell good food from a food truck as long as it is cheap and do not have to worry about "Service" and "Decor".