COMP3162 Project 2

620118149, 620110644

3/30/2021

Project Part 02 – Teamwork

(2 persons per team)

[Weighted 15% of course marks]

DUE: April 07, 2021

1. Tweet Data Analysis [20]

a. Merge the tweets collected from project 01 by each person in the team. [1]

```
tweet_beer <- rbind(phillip_beer, subset(annabelle_beer, select = -c(X)) )
tweet_beverage <- rbind(phillip_beverage, subset(annabelle_beverage, select = -c(X)))
tweet_concert <- rbind(phillip_concert, subset(annabelle_concert, select = -c(X)))
tweet_party <- rbind(phillip_party, subset(annabelle_party, select = -c(X)))
```

- clean each dataframe again

```
# clean the text

tweet_beer.clean <- tweet_beer
tweet_beer.clean$text <- gsub("https.*","", tweet_beer.clean$text)
tweet_beer.clean$text <- gsub("http.*","", tweet_beer.clean$text)
tweet_beer.clean$text <- gsub("#.*","", tweet_beer.clean$text)
tweet_beer.clean$text <- gsub("@.*","", tweet_beer.clean$text)
tweet_beer.clean$text <- gsub("[^[:alnum:][:blank:]?&/\\-]","", tweet_beer.clean$text)
tweet_beer.clean$text <- gsub("U00..","", tweet_beer.clean$text)
tweet_beer.clean$text <- gsub("[^\x20-\x7E]","", tweet_beer.clean$text)

tweet_beverage.clean <- tweet_beverage
tweet_beverage.clean$text <- gsub("https.*","", tweet_beverage.clean$text)
tweet_beverage.clean$text <- gsub("http.*","", tweet_beverage.clean$text)
tweet_beverage.clean$text <- gsub("#.*","", tweet_beverage.clean$text)
tweet_beverage.clean$text <- gsub("@.*","", tweet_beverage.clean$text)
tweet_beverage.clean$text <- gsub("[^[:alnum:][:blank:]?&/\\-]","", tweet_beverage.clean$text)
tweet_beverage.clean$text <- gsub("U00..","", tweet_beverage.clean$text)
tweet_beverage.clean$text <- gsub("[^\x20-\x7E]","", tweet_beverage.clean$text)

tweet_concert.clean <- tweet_concert
```

```
tweet_concert.clean$text <- gsub("https.*","", tweet_concert.clean$text)
tweet_concert.clean$text <- gsub("http.*","", tweet_concert.clean$text)
tweet_concert.clean$text <- gsub("#.*","", tweet_concert.clean$text)
tweet_concert.clean$text <- gsub("@.*","", tweet_concert.clean$text)
tweet_concert.clean$text <- gsub("[^[:alnum:][:blank:]?&/\\-]","", tweet_concert.clean$text)
tweet_concert.clean$text <- gsub("U00..","", tweet_concert.clean$text)
tweet_concert.clean$text <- gsub("[^\x20-\x7E]","", tweet_concert.clean$text)

tweet_party.clean <- tweet_party
tweet_party.clean$text <- gsub("https.*","", tweet_party.clean$text)
tweet_party.clean$text <- gsub("http.*","", tweet_party.clean$text)
tweet_party.clean$text <- gsub("#.*","", tweet_party.clean$text)
tweet_party.clean$text <- gsub("@.*","", tweet_party.clean$text)
tweet_party.clean$text <- gsub("[^[:alnum:][:blank:]?&/\\-]","", tweet_party.clean$text)
tweet_party.clean$text <- gsub("U00..","", tweet_party.clean$text)
tweet_party.clean$text <- gsub("[^\x20-\x7E]","", tweet_party.clean$text)

# clean the locations

tweet_beer.clean$location <- gsub("https.*","", tweet_beer.clean$location)
tweet_beer.clean$location <- gsub("http.*","", tweet_beer.clean$location)
tweet_beer.clean$location <- gsub("#.*","", tweet_beer.clean$location)
tweet_beer.clean$location <- gsub("@.*","", tweet_beer.clean$location)
tweet_beer.clean$location <- gsub("[^[:alnum:][:blank:]?&/\\-]","", tweet_beer.clean$location)
tweet_beer.clean$location <- gsub("U00..","", tweet_beer.clean$location)
tweet_beer.clean$location <- gsub("[^\x20-\x7E]","", tweet_beer.clean$location)

tweet_beverage.clean$location <- gsub("https.*","", tweet_beverage.clean$location)
tweet_beverage.clean$location <- gsub("http.*","", tweet_beverage.clean$location)
tweet_beverage.clean$location <- gsub("#.*","", tweet_beverage.clean$location)
tweet_beverage.clean$location <- gsub("@.*","", tweet_beverage.clean$location)
tweet_beverage.clean$location <- gsub("[^[:alnum:][:blank:]?&/\\-]","", tweet_beverage.clean$location)
tweet_beverage.clean$location <- gsub("U00..","", tweet_beverage.clean$location)
tweet_beverage.clean$location <- gsub("[^\x20-\x7E]","", tweet_beverage.clean$location)

tweet_concert.clean$location <- gsub("https.*","", tweet_concert.clean$location)
tweet_concert.clean$location <- gsub("http.*","", tweet_concert.clean$location)
tweet_concert.clean$location <- gsub("#.*","", tweet_concert.clean$location)
tweet_concert.clean$location <- gsub("@.*","", tweet_concert.clean$location)
tweet_concert.clean$location <- gsub("[^[:alnum:][:blank:]?&/\\-]","", tweet_concert.clean$location)
tweet_concert.clean$location <- gsub("U00..","", tweet_concert.clean$location)
tweet_concert.clean$location <- gsub("[^\x20-\x7E]","", tweet_concert.clean$location)

tweet_party.clean$location <- gsub("https.*","", tweet_party.clean$location)
tweet_party.clean$location <- gsub("http.*","", tweet_party.clean$location)
tweet_party.clean$location <- gsub("#.*","", tweet_party.clean$location)
tweet_party.clean$location <- gsub("@.*","", tweet_party.clean$location)
tweet_party.clean$location <- gsub("[^[:alnum:][:blank:]?&/\\-]","", tweet_party.clean$location)
tweet_party.clean$location <- gsub("U00..","", tweet_party.clean$location)
tweet_party.clean$location <- gsub("[^\x20-\x7E]","", tweet_party.clean$location)
```

b. Remove all duplicate tweets in the newly merged set of tweets. A tweet is a duplicate if the text is exactly the same as the text in another tweet. In removing the duplicate tweets, it might be useful to

keep the one that has the highest retweet count. [2]

```
tweet_beer.clean.unique <- subset(tweet_beer.clean, !duplicated(tweet_beer.clean$text))
tweet_beverage.clean.unique <- subset(tweet_beverage.clean, !duplicated(tweet_beverage.clean$text))
tweet_concert.clean.unique <- subset(tweet_concert.clean, !duplicated(tweet_concert.clean$text))
tweet_party.clean.unique <- subset(tweet_party.clean, !duplicated(tweet_party.clean$text))
```

c. Explore the merged tweets and provide descriptive statistics. [3]

- Beer descriptive statistics

```
beer_count = nrow(tweet_beer.clean.unique)

beer_user_most_retweet = unname( tweet_beer.clean.unique[
                max(tweet_beer.clean.unique$followers_count)==tweet_beer.clean.unique$followers_count
                ,"screen_name"
                ]
              )

beer_most_retweet = unname(
              tweet_beer.clean.unique[
                max(tweet_beer.clean.unique$retweet_count)==tweet_beer.clean.unique$retweet_count
                ,"text"
                ]
              )

beer_location_count <- table(tweet_beer.clean.unique$location)
beer_location_count <- as.data.frame(beer_location_count)

# remove empty location row
beer_location_count <- subset(beer_location_count, beer_location_count$Var1 != "")


beer_location_most_tweet = unname(
              beer_location_count[
                max(beer_location_count$Freq)==beer_location_count$Freq
                ,"Var1"
                ]
              )
```

```
## [1] "--------------------------------------------------"

## [1] "                    Beer"

## [1] "--------------------------------------------------"

## [1] "The number of Twets retrieved are:  5667"

## [1] "The user with the most followers:  ABC"

## [1] "The tweets with the most retweet:  its the weekend baby youknow what that means  its time to dr
```

```
## [1] "The location with the most tweets:  Los Angeles CA"
```

- Beverage descriptive statistics

```
beverage_count = nrow(tweet_beverage.clean.unique)

beverage_user_most_retweet = unname( tweet_beverage.clean.unique[
                max(tweet_beverage.clean.unique$followers_count)==tweet_beverage.clean.unique$followers
                ,"screen_name"
                ]
              )

beverage_most_retweet = unname(
              tweet_beverage.clean.unique[
                max(tweet_beverage.clean.unique$retweet_count)==tweet_beverage.clean.unique$retweet_cou
                ,"text"
                ]
              )

beverage_location_count <- table(tweet_beverage.clean.unique$location)
beverage_location_count <- as.data.frame(beverage_location_count)

# remove empty location row
beverage_location_count <- subset(beverage_location_count, beverage_location_count$Var1 != "")


beverage_location_most_tweet = unname(
              beverage_location_count[
                max(beverage_location_count$Freq)==beverage_location_count$Freq
                ,"Var1"
                ]
              )
```

```
## [1] "-------------------------------------------------"

## [1] "                    beverage"

## [1] "-------------------------------------------------"

## [1] "The number of Twets retrieved are:  12863"

## [1] "The user with the most followers:  timesofindia"

## [1] "The tweets with the most retweet:  Erica Nlewedim bags new endorsement deal with beverage compa

## [1] "The location with the most tweets:  United States"
```

- party descriptive statistics

```
party_count = nrow(tweet_party.clean.unique)

party_user_most_retweet = unname( tweet_party.clean.unique[
                max(tweet_party.clean.unique$followers_count)==tweet_party.clean.unique$followers_count
                ,"screen_name"
                ]
              )

party_most_retweet = unname(
              tweet_party.clean.unique[
                max(tweet_party.clean.unique$retweet_count)==tweet_party.clean.unique$retweet_count
                ,"text"
                ]
              )

party_location_count <- table(tweet_party.clean.unique$location)
party_location_count <- as.data.frame(party_location_count)

# remove empty location row
party_location_count <- subset(party_location_count, party_location_count$Var1 != "")


party_location_most_tweet = unname(
              party_location_count[
                max(party_location_count$Freq)==party_location_count$Freq
                ,"Var1"
                ]
              )
```

```
## [1] "-------------------------------------------------"

## [1] "                    party"

## [1] "-------------------------------------------------"

## [1] "The number of Twets retrieved are:  11297"

## [1] "The user with the most followers:  XHNews"

## [1] "The tweets with the most retweet:  Family welcome Mark back home party F389 "

## [1] "The location with the most tweets:  United States"
```

- concert descriptive statistics

```
concert_count = nrow(tweet_concert.clean.unique)

concert_user_most_retweet = unname( tweet_concert.clean.unique[
                max(tweet_concert.clean.unique$followers_count)==tweet_concert.clean.unique$followers_c
                ,"screen_name"
                ]
```

```
              )

concert_most_retweet = unname(
              tweet_concert.clean.unique[
                max(tweet_concert.clean.unique$retweet_count)==tweet_concert.clean.unique$retweet_count
                ,"text"
                ]
              )

concert_location_count <- table(tweet_concert.clean.unique$location)
concert_location_count <- as.data.frame(concert_location_count)

# remove empty location row
concert_location_count <- subset(concert_location_count, concert_location_count$Var1 != "")


concert_location_most_tweet = unname(
              concert_location_count[
                max(concert_location_count$Freq)==concert_location_count$Freq
                ,"Var1"
                ]
              )
```

```
## [1] "--------------------------------------------------"

## [1] "                    concert"

## [1] "--------------------------------------------------"

## [1] "The number of Twets retrieved are:  9802"

## [1] "The user with the most followers:  cnni"

## [1] "The tweets with the most retweet:  keep wearing your mask so we can scream what a shame shes fu

## [1] "The location with the most tweets:  she/her"
```

Analysis on emotions

```
t.beer <- tweet_beer.clean.unique
t.beverage <- tweet_beverage.clean.unique
t.party <- tweet_party.clean.unique
t.concert <- tweet_concert.clean.unique

#extract emotions
t.beer.emot <- get_nrc_sentiment(t.beer$text)
t.beverage.emot <- get_nrc_sentiment(t.beverage$text)
t.party.emot <- get_nrc_sentiment(t.party$text)
t.concert.emot <- get_nrc_sentiment(t.concert$text)
```

```
t.beer <- cbind(t.beer, t.beer.emot)
t.beverage <- cbind(t.beverage, t.beverage.emot)
t.party <- cbind(t.party, t.party.emot)
t.concert <- cbind(t.concert, t.concert.emot)

#extract sentiments
t.beer.sent <-  get_sentiment(t.beer$text)
t.beverage.sent <-  get_sentiment(t.beverage$text)
t.party.sent <-  get_sentiment(t.party$text)
t.concert.sent <-  get_sentiment(t.concert$text)

t.beer <- cbind(t.beer, t.beer.sent)
t.beverage <- cbind(t.beverage, t.beverage.sent)
t.party <- cbind(t.party, t.party.sent)
t.concert <- cbind(t.concert, t.concert.sent)
```

d. What are the dominant emotions associated with beverages in any two locations? [4]

```
# The two location that were chosen is the two top locations for beverages
us_data <- subset(t.beverage, t.beverage$location=="United States",  select = names(t.beer.emot) )
us_data.sum <- colSums(us_data)

ny_data <- subset(t.beverage, t.beverage$location=="New York, NY",  select = names(t.beer.emot) )
ny_data.sum <- colSums(ny_data)

# get the dominant emotions

us_dom_emot <- names(us_data.sum)[match(max(us_data.sum),us_data.sum)]

ny_dom_emot = names(ny_data.sum)[match(max(ny_data.sum),ny_data.sum)]
```

```
## [1] "The dominant emotions are: "

## [1] "United States :  positive"

## [1] "New York, NY :  anger"
```

e. What are the dominant emotions in the overall dataset? [2]

```
t.beer.emot.sum <- colSums(t.beer.emot)
t.beer.emot.dom <- names(t.beer.emot.sum)[match(max(t.beer.emot.sum),t.beer.emot.sum)]

t.beverage.emot.sum <- colSums(t.beverage.emot)
t.beverage.emot.dom <- names(t.beverage.emot.sum)[match(max(t.beverage.emot.sum),t.beverage.emot.sum)]

t.party.emot.sum <- colSums(t.party.emot)
t.party.emot.dom <- names(t.party.emot.sum)[match(max(t.party.emot.sum),t.party.emot.sum)]

t.concert.emot.sum <- colSums(t.concert.emot)
t.concert.emot.dom <- names(t.concert.emot.sum)[match(max(t.concert.emot.sum),t.concert.emot.sum)]
```

```
## [1] "The dominant emotions in the oeral data set are: "
```

```
## [1] "Beers :  positive"
```

```
## [1] "Beverages :  positive"
```

```
## [1] "Parties :  positive"
```

```
## [1] "Concerts :  positive"
```

f. What is the overall sentiment in tweets regarding "beverages" and "party or concert" (separately)?[4]

```
beer_overall_sent <- sum(t.beer.sent)
bev_overall_sent <- sum(t.beverage.sent)
party_overall_sent <- sum(t.party.sent)
concert_overall_sent <- sum(t.concert.sent)
```

```
## [1] "The overall sentiments are: "
```

```
## [1] "Beer :  1674.05 ,  Postive"
```

```
## [1] "Beverage :  9222.8 ,  Postive"
```

```
## [1] "Party :  6755.95 ,  Postive"
```

```
## [1] "Concert :  3769.85 ,  Postive"
```

g. Conduct ONE additional analysis of your choice to discover any further useful insights.[4]

```
# beer

beer_location_sent <- beer_location_count
beer_location_sent$sentiment <- c(0)
cnt <- NROW(beer_location_sent$Var1)

for(i in 1:cnt){
  data_sub <- subset(t.beer, t.beer$location == toString(beer_location_sent$Var1[i]))
  beer_location_sent$sentiment[i] <- sum(data_sub$t.beer.sent)
}

beer_loc_most_pos_sent <- beer_location_sent$Var1[match(max(beer_location_sent$sentiment),beer_location

# beverage

beverage_location_sent <- beverage_location_count
beverage_location_sent$sentiment <- c(0)
cnt <- NROW(beverage_location_sent$Var1)

for(i in 1:cnt){
  data_sub <- subset(t.beverage, t.beverage$location == toString(beverage_location_sent$Var1[i]))
  beverage_location_sent$sentiment[i] <- sum(data_sub$t.beverage.sent)
}
```

```
beverage_loc_most_pos_sent <- beverage_location_sent$Var1[match(max(beverage_location_sent$sentiment),be

# party

party_location_sent <- party_location_count
party_location_sent$sentiment <- c(0)
cnt <- NROW(party_location_sent$Var1)

for(i in 1:cnt){
  data_sub <- subset(t.party, t.party$location == toString(party_location_sent$Var1[i]))
  party_location_sent$sentiment[i] <- sum(data_sub$t.party.sent)
}

party_loc_most_pos_sent <- party_location_sent$Var1[match(max(party_location_sent$sentiment),party_loca

# concert

concert_location_sent <- concert_location_count
concert_location_sent$sentiment <- c(0)
cnt <- NROW(concert_location_sent$Var1)

for(i in 1:cnt){
  data_sub <- subset(t.concert, t.concert$location == toString(concert_location_sent$Var1[i]))
  concert_location_sent$sentiment[i] <- sum(data_sub$t.concert.sent)
}

concert_loc_most_pos_sent <- concert_location_sent$Var1[match(max(concert_location_sent$sentiment),conc
```

```
## [1] "The location with the most positive sentiment are: "
```

```
## [1] "Beer :  Atlanta GA"
```

```
## [1] "Beverage :  United States"
```

```
## [1] "Party :  London England"
```

```
## [1] "Concert :  Chicago IL"
```

2. Collect, Explore, Prepare Structured Data [20 marks]

a. Download the datafile consumer_pt02_2021.csv from OurVLE

b. Explore the data and provide details on all fields retrieved. You should ensure all features in the dataset (each column) are reviewed and summarized to verify things such as value ranges, missing values etc. Be sure to generate relevant graphical representations where necessary to demonstrate your review and decision making. [7]

c. Fix noise, outlier and any other issues discovered (example: na values). You must provide discussion / explanation of all activities done and why each decision has been made. [8]

d. Format/reformat the data as necessary. Please note that as you proceed through the project, you may need to do additional formatting to enable your analysis. [5]

3. Structured Data Analysis/Modeling [35]

Write code to conduct analysis that will answer the questions below. You are encouraged to use tables/graphs where necessary to visualize results. Additionally, your code should be shown along with each question, the result and notes that explain the results. a. What is the average spend on beverages in each country? [3]

b. Which country has the highest spending on beverages? [2]

c. Which country consumes the most beverages? [2]

d. What is the average profit from the sale of beverages in each country? [3]

e. What has been the total revenue from beverages for each year since 2014? [5]

f. Plot a time series graph showing change in overall revenues from beverages for the last six months (in the dataset). [4]

g. What is the dominant sales channel for beverages?[2]

h. Determine whether beverages units sold is above the overall average for units sold for all other products. [3]

i. In which season (Spring, Summer, Autumn, Winter) does persons spend the most on beverages? [6]

j. Is there a correlation between the season and the units sold for beverages? Explain the result. [5]

4. Recommendation:

a. Based on your analysis of both the tweet data and structured data, what would you recommend to Hard Knocks and why?

5. BONUS – 10 marks

a. Which features in the dataset can be used to predict the units sold for beverages?