Cava Data Challenge

Jonathan Campbell

December 3, 2015

Data Analysis:

You should have received a file named "data.csv" along with these instructions (if not, you can download from https://goo.gl/vI9p39). There are 3,000,000 rows; each row represents a consumer and a summary of his/her purchasing information. The first column is the day the person heard about Cava, the second is the number of days between hearing about Cava and first purchasing (a 0 means they heard and bought on the same day), and the third column is the total number of purchases the consumer has made at the time the data file was created1.
\ Your task is to analyze the data and describe any meaningful insights you find (this task is open ended on purpose). Your analysis should be as thorough as possible. Feel free to be creative in your exploration and to document the process you used in your investigation. Pretty graphs are certainly welcome, but make sure there is sufficient textual analysis as well.

Data Input

```r
# Looking at the structure of the data before loading it in
readLines("data.csv", 5)
```

```
## [1] "\"download_app\",\"first_purchase\",\"purchases\""
## [2] "2014-11-22,2014-12-11,4"
## [3] "2015-01-19,2015-02-07,1"
## [4] "2014-12-26,2015-01-08,3"
## [5] "2014-12-10,2014-12-22,8"
```

An initial look at the data shows us that our variables are in date and integer format. One piece of information this is missing from the data is a unique identifier per customer. With no clear formatting issues, I'll load in the data, add in an identifier, and format as necessary.

```r
# Read in data
df1 <- data.table::fread("data.csv", data.table = FALSE)
df1$download_app <- as.Date(df1$download_app)
df1$first_purchase <- as.Date(df1$first_purchase)

# addition of id and additional variables
df1$id <- 1:nrow(df1)
df1$difftime <- as.numeric(df1$first_purchase - df1$download_app)
df1$same_day <- as.factor(ifelse(df1$difftime==0,TRUE,FALSE))
```

Initial Exploration

In addition to adding an ID variable, I added two more variables as well to caputure (1) the number of days between the date of download and a customer's first purchase and (2) an indicator as to whether a customer purchased the same day as downloading the app. The reasoning behind adding in a same-day indicator is that the difference in immediate action versus delayed action might imply a difference in the distribution and the indicator will make subsetting between the two groups easier.

With this extra information added, I'd like to gain a bit more information about the distribution of downloads and purchases. The first concern is whether there are any missing values or are there individuals in the set who have either purchased withough downloading the app or downloaded the app without making a purchase.

```r
# Are there missing values?
table(is.na(df1[,1:3]))
```

```
##
##   FALSE
## 9000000
```

```r
# Table of purchasing
table(Same_Day = df1$same_day,
      Delayed_Download = df1$download_app < df1$first_purchase)
```

```
##         Delayed_Download
## Same_Day   FALSE    TRUE
##    FALSE       0 2601383
##    TRUE   398617       0
```

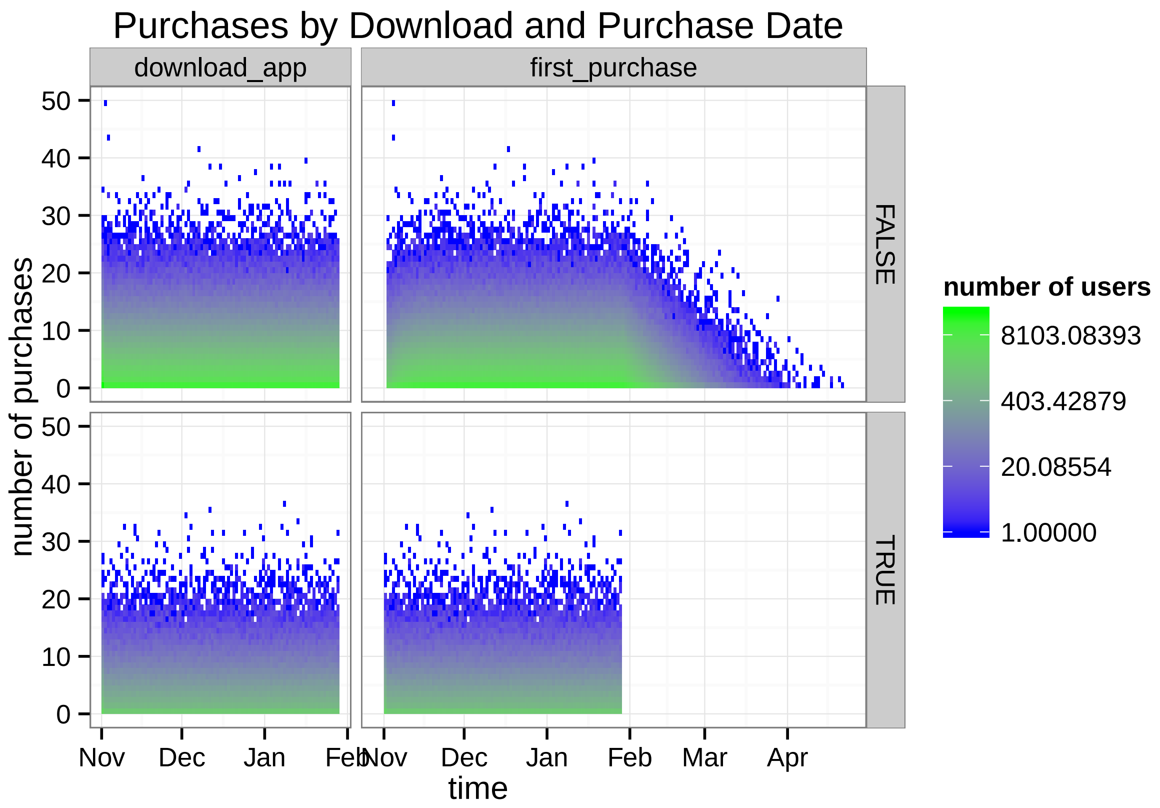
```r
# What does the distribution of purchases look like?
summary(df1$purchases)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   2.000   2.986   4.000  50.000
```

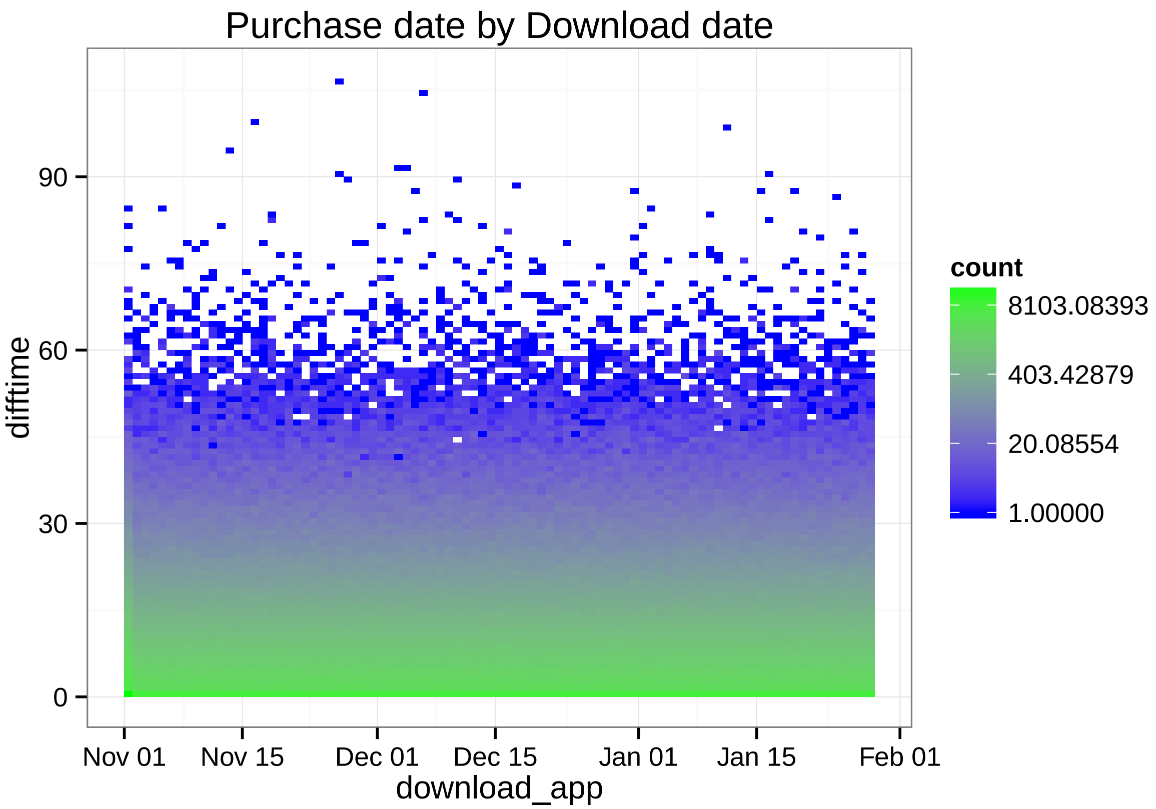
From tables and summary above we can note that there are no missing observations. Additionally, customers in this set of data either purchased same day or purchased after downloading the app (as there are no observations in the negative for both same day purchases and delayed download). Lastly, there are cases in which the field for purchases is 0. Seeing as there are no missing values, I'll assume that the field stands for subsequent purchases. The summary information also tells us that the distribution of subsequent purchases is right tailed due to the mean being greater than the median value.

With this summary information in hand, we can now look at some initial plots of the data to look for any trends. Due to the size of the data, I'll be using bin plots.

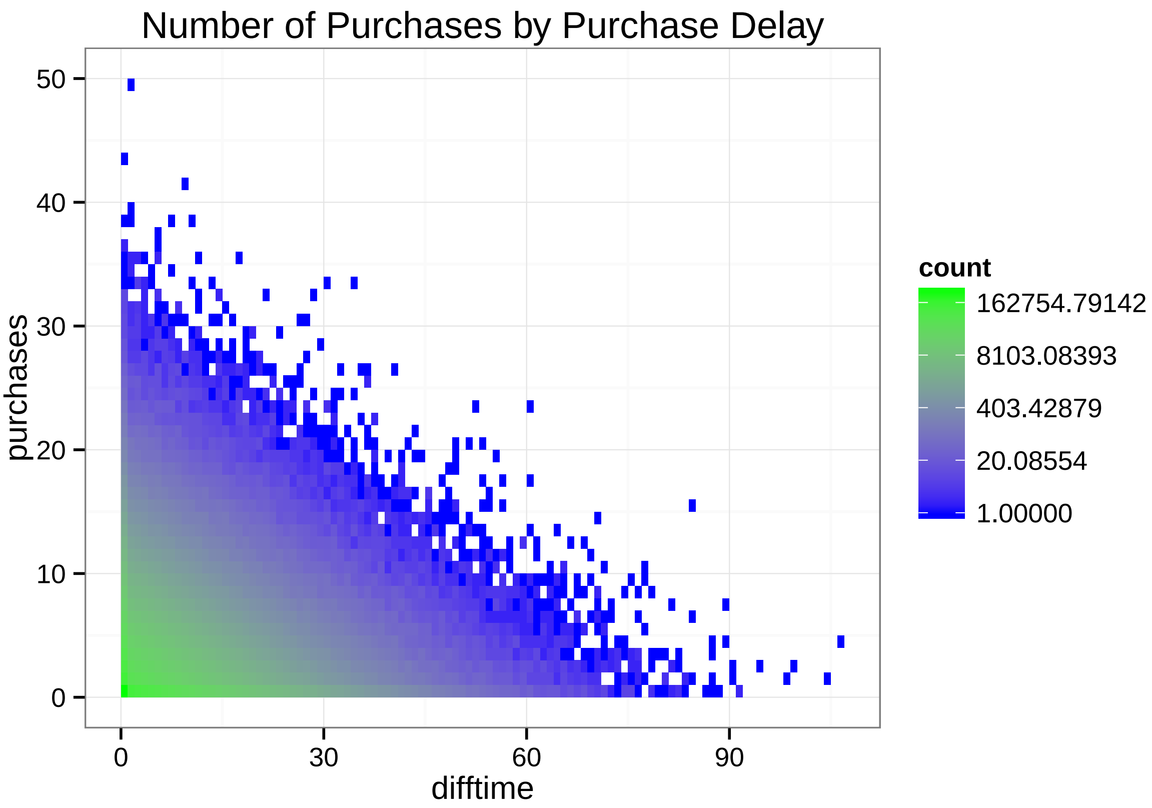
```r
# Plot of Number of Purchases by Download and Purchase Date
df_plot <- reshape2::melt(df1, id=c("id", "purchases", "difftime", "same_day"))
ggplot(df_plot, aes(x=value, y=purchases)) +
  stat_bin2d(binwidth=c(1,1)) +
  facet_grid(same_day ~ variable, space = "free", scales = "free_x") +
  theme_bw() +
  scale_x_date(labels = scales::date_format("%b"), breaks = scales::date_breaks(width = "1 month")) +
  scale_fill_continuous(trans="log", low="blue", high="green") +
  ggtitle("Purchases by Download and Purchase Date") +
  xlab("time") + ylab("number of purchases") + labs(fill="number of users")
```



```
# Plot of Purchase date by Download Date
ggplot(df1, aes(x=download_app, y=difftime)) +
  ggtitle("Purchase date by Download date") +
  geom_bin2d(binwidth=c(1,1)) +
  scale_fill_continuous(trans="log", low="blue", high="green") +
  theme_bw()
```



```r
# Plot of Purchase date by Download Date
ggplot(df1, aes(x=difftime, y=purchases)) +
  ggtitle("Number of Purchases by Purchase Delay") +
  geom_bin2d(binwidth=c(1,1)) +
  scale_fill_continuous(trans="log", low="blue", high="green") +
  theme_bw()
```

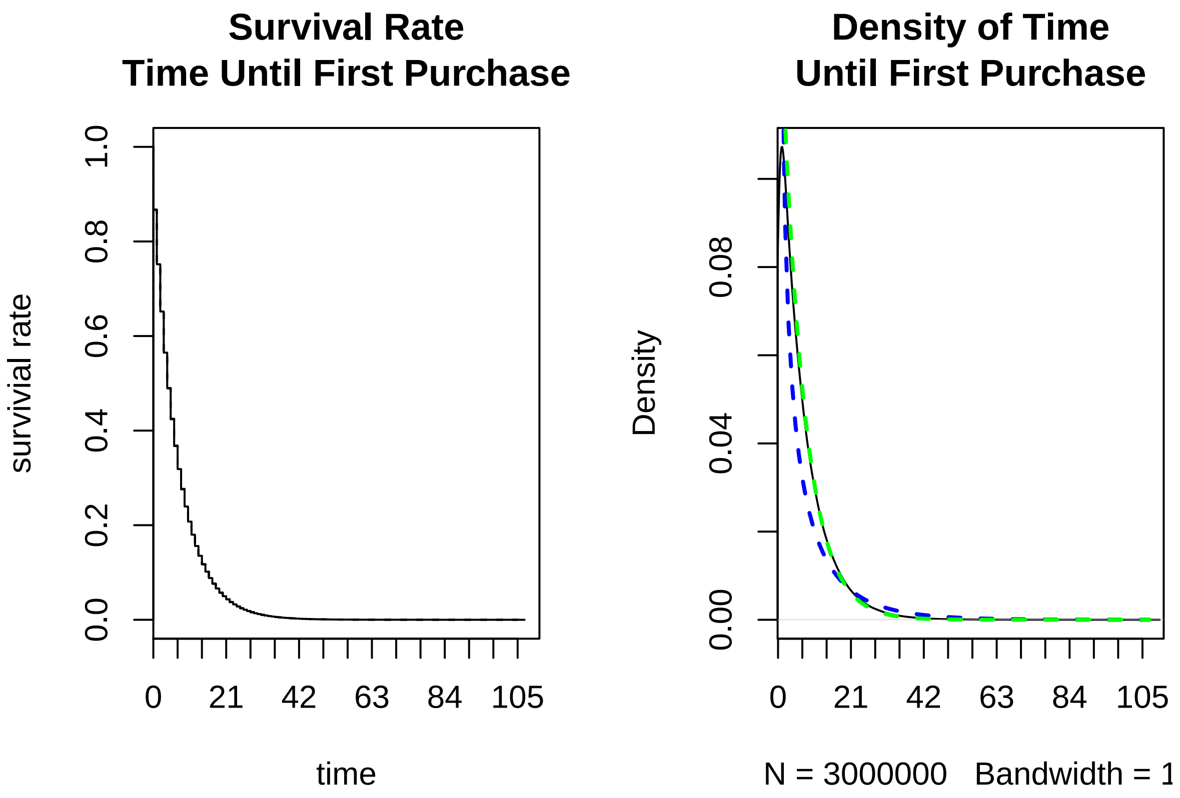


The facet plot above shows us the density of points for subsequent purchases both boased on download date and first purchase date. From this first plot, we can see that the distribution of subsequent purchases seems to be steady across time, regardless of download date or date of first purchase as well as whenther you purchased same day or later. This is fairly interesting becasue it indicates that there isn't much of a relationship between the number of subsequent purchases and your download/purchase date.
The second plot shows the distribution of days until first purchase across time. Once again we see a steady relationship across time regarding the habits of users to purchase for the first time. These lack of changing trends will make the analysis easier as it seems we can exclude time as a factor.
Finally, seeing no need to control for time, the last plot looks at the number of subsequent purchases by purchase delay. This graph shouls a clear negative relationship such that a smaller purchase delay correspons with a larger range of subsequent purchases. While strength of the trend is amplified by time limitations, it does display that users are more likely to purchase more over shorter periods of time with a shorter purchase delay.
With this relationship in mind, I plan to run a survival analysis to look at the propensity of a customer to purchase over time by using the initial download date and the first purchase date.

Survival Rate

```
# Survival Analysis
df1$start_time <- 0
df1$difftime2 <- df1$difftime + 1/(24*60) # Added a minute to all times in order to use a gamme MLE
df1$event <- 1

library(survival)
s2firstBuy <- survfit(formula=Surv(time=df1$start_time, time2 = df1$difftime2,event=df1$event) ~ 1)

dist2 <- MASS::fitdistr(df1$difftime2, densfun="gamma")
```

```
dist3 <- MASS::fitdistr(df1$difftime, densfun="exponential")

par(mfrow=c(1,2))
plot(s2firstBuy,
     main = "Survival Rate\nTime Until First Purchase",
     xlab = "time", xaxt="n",
     ylab = "survivial rate")
axis(1,at=seq(0,107, by=7))

plot(density(df1$difftime, bw=1),
     xlim=c(4,max(df1$difftime)), xaxt="n",
     main = "Density of Time\nUntil First Purchase")
axis(1,at=seq(0,107, by=7))
lines( sort(df1$difftime) , # GAMMA 2
       y = dgamma(
         sort(df1$difftime2) ,
         shape = dist2$estimate[1],
         rate = dist2$estimate[2]),
       col = "blue" , lty = 2 , lwd = 2 )
lines( sort(df1$difftime) , # EXPOENTIAL
       y = dexp(
         sort(df1$difftime) ,
         rate = dist3$estimate[1]),
       col = "green" , lty = 2 , lwd = 2 )
```



Above, there are two charts that display the results of applying a survival function to the data to determine the distribution of delay times for customer purchased. Unsurprisingly, we see a fairly smooth survival rate considering the distributions seen before. The plot on the left shows the Kaplan-Meier estimate for the survival rate. From this plot, we can estimate a 50% probability that a customer hasn't made thier first purchase after the first week that probability decreases to about 12.5% after two weeks.

This survival curve very closely resembles the the density curve for the distribution of wait times. In order to generalize what we believe to be this distribution, I attempted to fit two probability models: the gamma distribution (blue) and the expenential distribution (green), both of which are commonly used to model waiting times. The result of this fit can be seen on the right, where we can see an almost perfect fit with an exponential distribution with the parameterization $\lambda = 0.1535173$.

Data Analysis 2:

You should have received a file named "cavaitemssold.csv" along with these instructions (if not, you can download from goo.gl/4tVcWk). There are an unknown amount of rows; each row represents an item sold and a corresponding checkid. The first column is the item sold and the second is the checkid2. The question you are trying to answer is if people are more inclined to get certain items once they have selected other items. \ Your task is to analyze the data and describe any meaningful insights you find (this task is open ended on purpose). Your analysis should be as thorough as possible. Feel free to be creative in your exploration and to document the process you used in your investigation. Pretty graphs are certainly welcome, but make sure there is sufficient textual analysis as well.

Data Input

```
# Look at Data Structure
readLines("cavaitemssold.csv", 5)
```

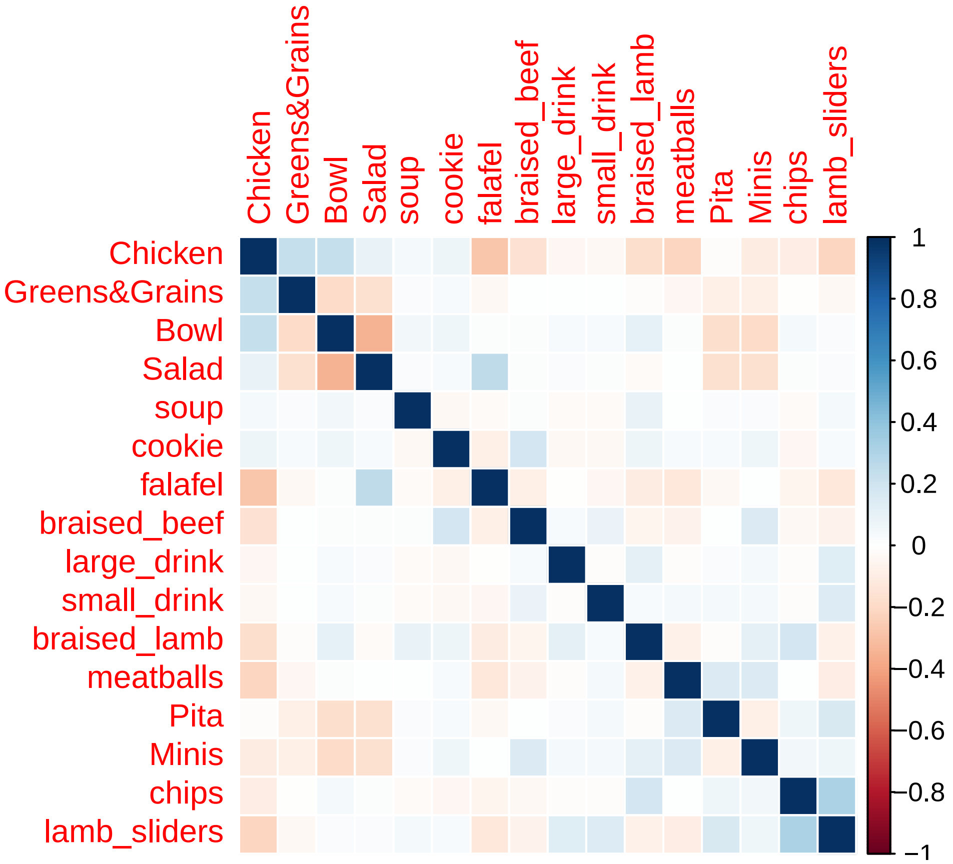
```
## [1] "item,transactionid"
## [2] "Bowl,14549363-7413-45df-906c-cdc0e896aeb1"
## [3] "falafel,14549363-7413-45df-906c-cdc0e896aeb1"
## [4] "Pita,80dd482d-b377-4aab-ae78-3cd1b5415615"
## [5] "Chicken,80dd482d-b377-4aab-ae78-3cd1b5415615"
```

Once again, we can take an initial look at the data which shows us that our variables are in menu items and transaction ids. Different from the first dataset, each row of data represents one item in a larger order. Aggregating this data so that we can get more information about individial order will require some data transformation.

```
# Read in Data
df2 <- data.table::fread("cavaitemssold.csv", data.table = FALSE)

protein <- c("braised_beef","meatballs","braised_lamb","Chicken","falafel", "Greens&Grains")
meal <- c("Salad","Bowl","Minis","Pita","lamb_sliders")
side <- c("large_drink","small_drink","chips","cookie","soup")

df2_orders <- left_join(
  df2 %>%
    group_by(transactionid) %>%
    summarize(tot_items=n(),
              proteins=sum(item %in% protein),
              meals=sum(item %in% meal),
              sides=sum(item %in% side)),
  reshape2::dcast(df2, transactionid ~ item, fun.aggregate = length))
```

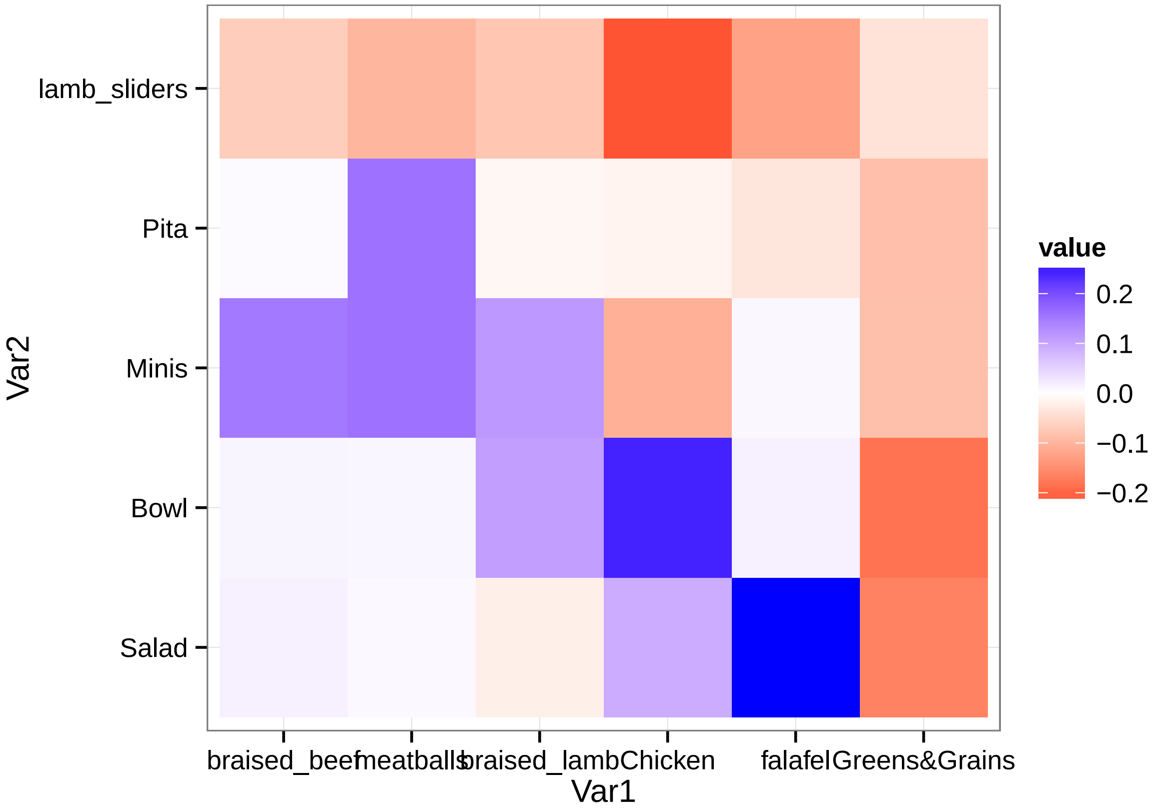
```
## Using transactionid as value column: use value.var to override.
## Joining by: "transactionid"
```

```
library(corrplot)
itemcorr <- cor(df2_orders[,-1:-5])
corrplot(itemcorr, method="color", order = "FPC")
```



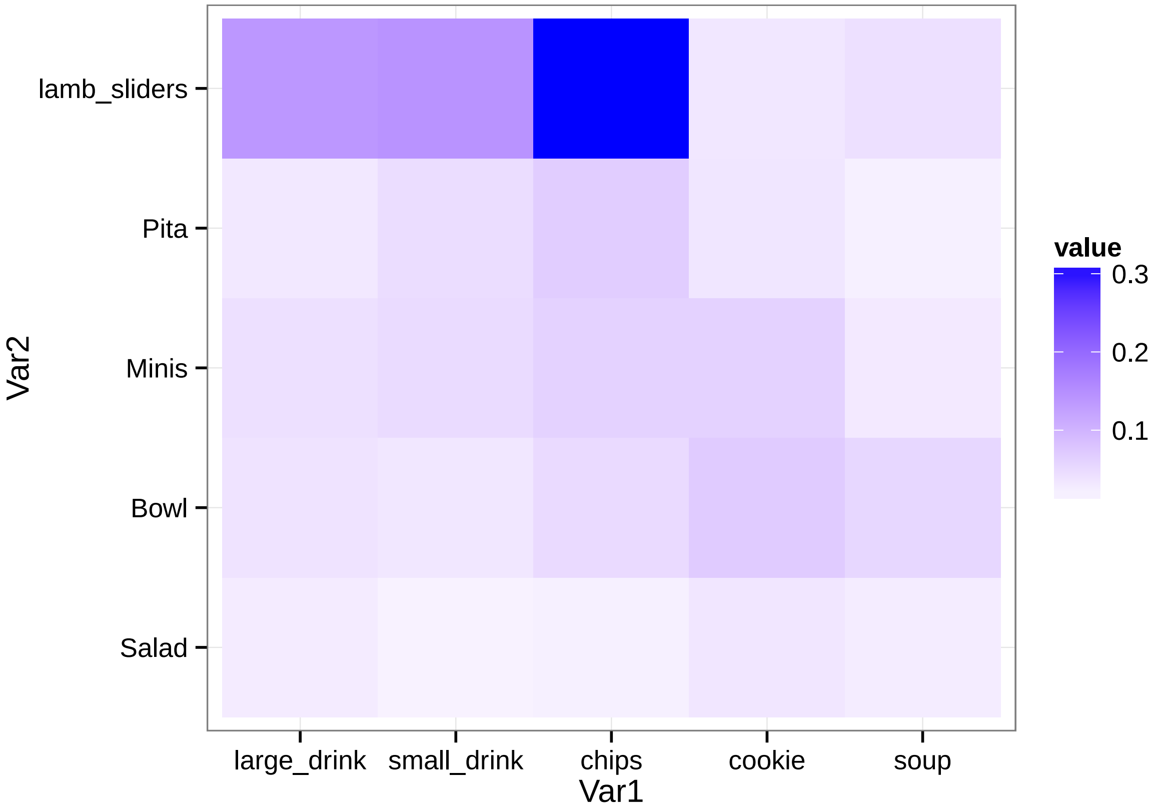
```
mealprotein <- reshape2::melt(cor(df2_orders[,protein],df2_orders[,meal]))
ggplot(mealprotein, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() +
  scale_fill_gradient2(low = "red", high = "blue", mid = "white") +
  theme_bw()
```

```
## Warning: Non Lab interpolation is deprecated
```



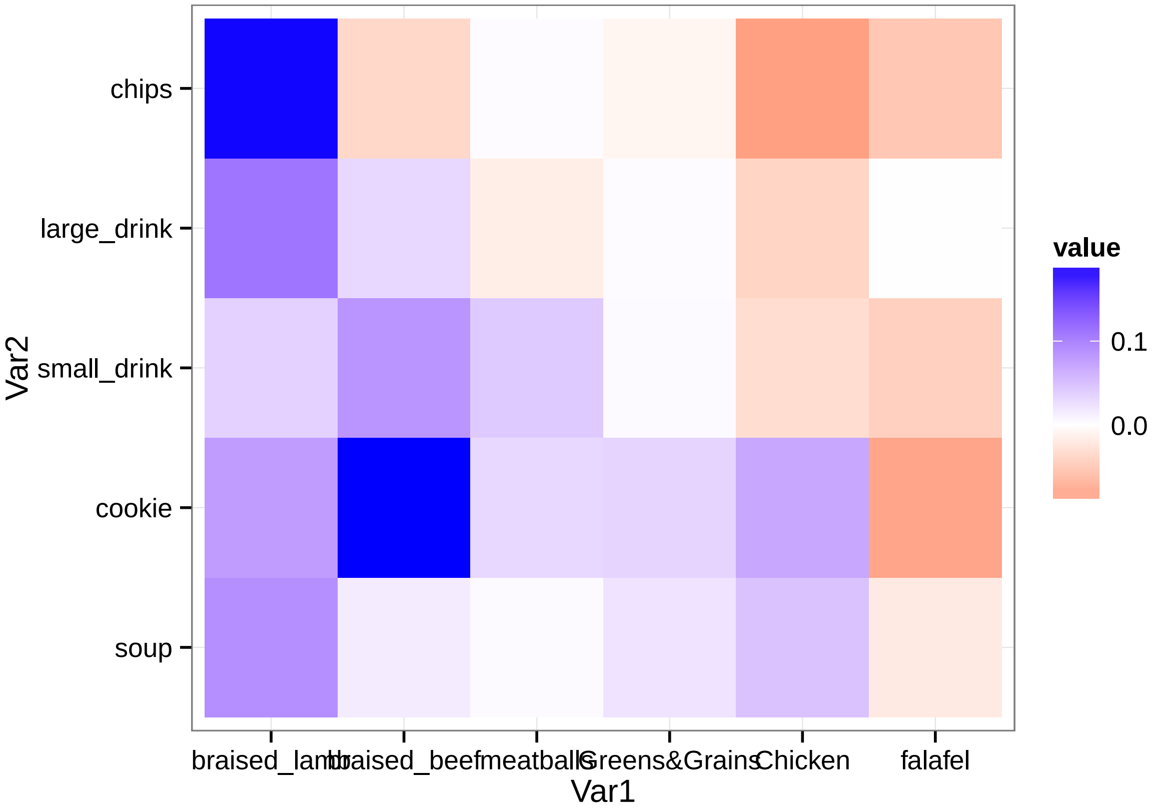
```
mealside <- reshape2::melt(cor(df2_orders[,side],df2_orders[,meal]))
ggplot(mealside, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() +
  scale_fill_gradient2(low = "red", high = "blue", mid = "white") +
  theme_bw()
```

```
## Warning: Non Lab interpolation is deprecated
```



```
protein <- c("braised_lamb","braised_beef","meatballs","Greens&Grains","Chicken","falafel")
side <- c("soup","cookie","small_drink","large_drink","chips")
proteinside <- reshape2::melt(cor(df2_orders[,protein],df2_orders[,side]))
ggplot(proteinside, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() +
  scale_fill_gradient2(low = "red", high = "blue", mid = "white") +
  theme_bw()
```

```
## Warning: Non Lab interpolation is deprecated
```



```
itemcorr[upper.tri(itemcorr, diag=T)] <- NA
itemcorr <- na.omit(arrange(reshape2::melt(itemcorr, value.name="corr"), -corr))
itemcorr <- itemcorr[itemcorr$corr!=1,]
head(itemcorr,10)
```

```
##              Var1         Var2      corr
## 1   lamb_sliders        chips 0.3103896
## 2          Salad      falafel 0.2585556
## 3        Chicken         Bowl 0.2395038
## 4  Greens&Grains      Chicken 0.2352955
## 5         cookie braised_beef 0.1890028
## 6          chips braised_lamb 0.1874487
## 7           Pita lamb_sliders 0.1675422
## 8           Pita    meatballs 0.1575689
## 9          Minis    meatballs 0.1575307
## 10         Minis braised_beef 0.1504099
```

Analysis

Visualized above, we have multiple correlation plots for the menu items contained within the dataset. The first plot displays the correlations among all of the different menu items. First off from the graph, we can see that there tends to be either negative correlation or slight positive correlation between items of the same category. The subsequent plots are breakdowns of correlations for different types of menu items.
The first subset, looks at the correlations between protein choices and meal options. We can see that there tends to be a large correlation with chicken bowls and also with falafel salads.
The second subset, looks at the correlations between meal options and side choices. While all combinations

share a positive correlation, the lamb sliders tend to have the largies correlation with different sides, and chips tends to have the highest correlations with different meals.

Finally, the last subset shows the correlations between protein choices and different sides. This plot shows the strongest progressive trend across both axes. While the strongest correlations exist between chips and braised lamb and cookies and braised beef, we see that in general braised lamb has positive correlations with all sides and falafel has negative correlations with all sides. We might imagine there correlations to be in line with the tendancy for customers to make choices consistent with being health conscious.

Lastly, I've listed the 10 pairs of items that tend to be grouped together most frequently.

Study Design:

One challenge we face at Cava Grill is in trying to find the best methods to motivate non-purchasing consumers to make their first purchase. One tactic is to simply give away a free bag of chips or a free entrée. Before giving away any money for food, though, we would like to determine the results we can expect by running a test first.
Your second task is to design such a test. The goal of the test is to determine whether or not giving a free pita card to consumers who have never purchased can successfully motivate them to make a first purchase. You may use any data you would reasonably expect a company like Cava Grill to have in preparing and executing your test. The test should be written as a business proposal and should include a section on the methods you intend on using for the evaluation of the test. A sample of some of the criteria you will be evaluated on include:

1. *The clarity of the stated aims, hypotheses, and expected results*
2. *The appropriate generation of your consumer sample*
3. *Your test evaluation methods and statistical soundness*
4. *The extent to which you thoroughly investigate possible outcomes and conclusions*

Please be as specific as possible in your design. For instance, make sure that each of your design decisions have associated explicit reasoning.