Cava Data Challenge

Jonathan Campbell

December 9, 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

```
# 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.

```
# 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 capture (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 sub setting 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 without downloading the app or downloaded the app without making a purchase.

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

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

```
# 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
```

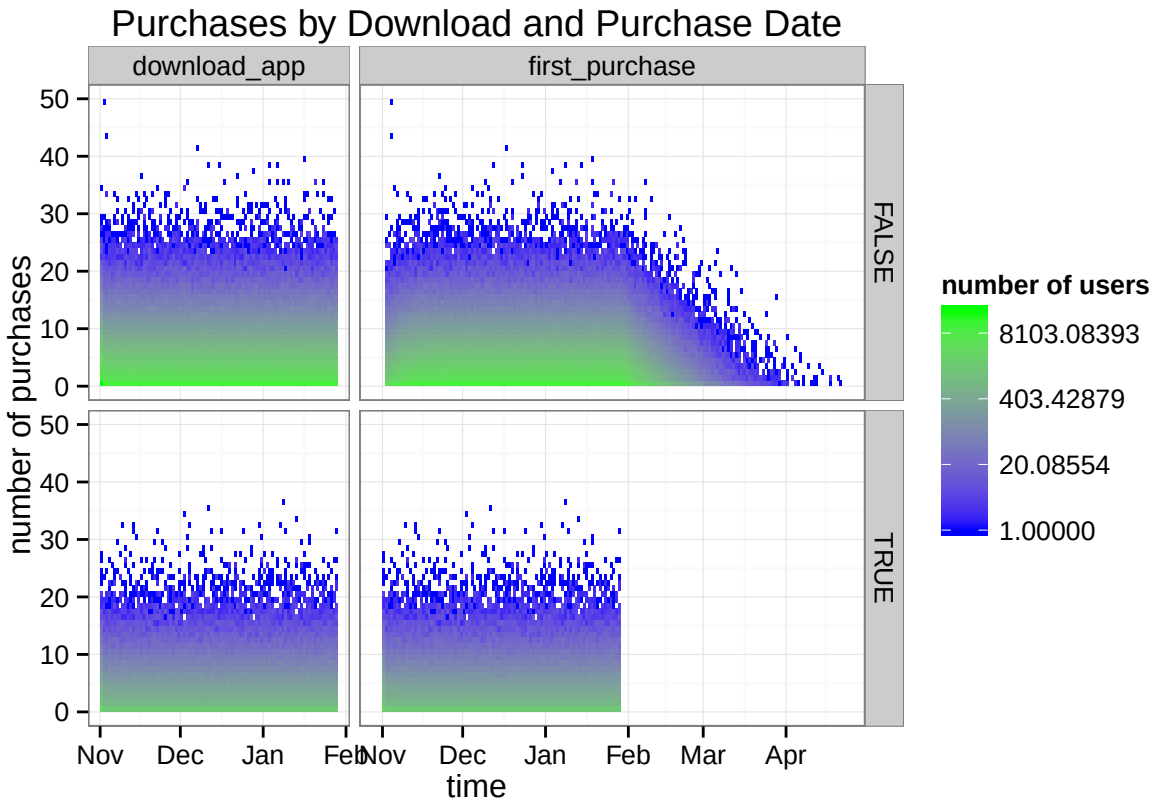
```
# What does the distribution of purchases look like?
summary(df1$purchases)
```

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

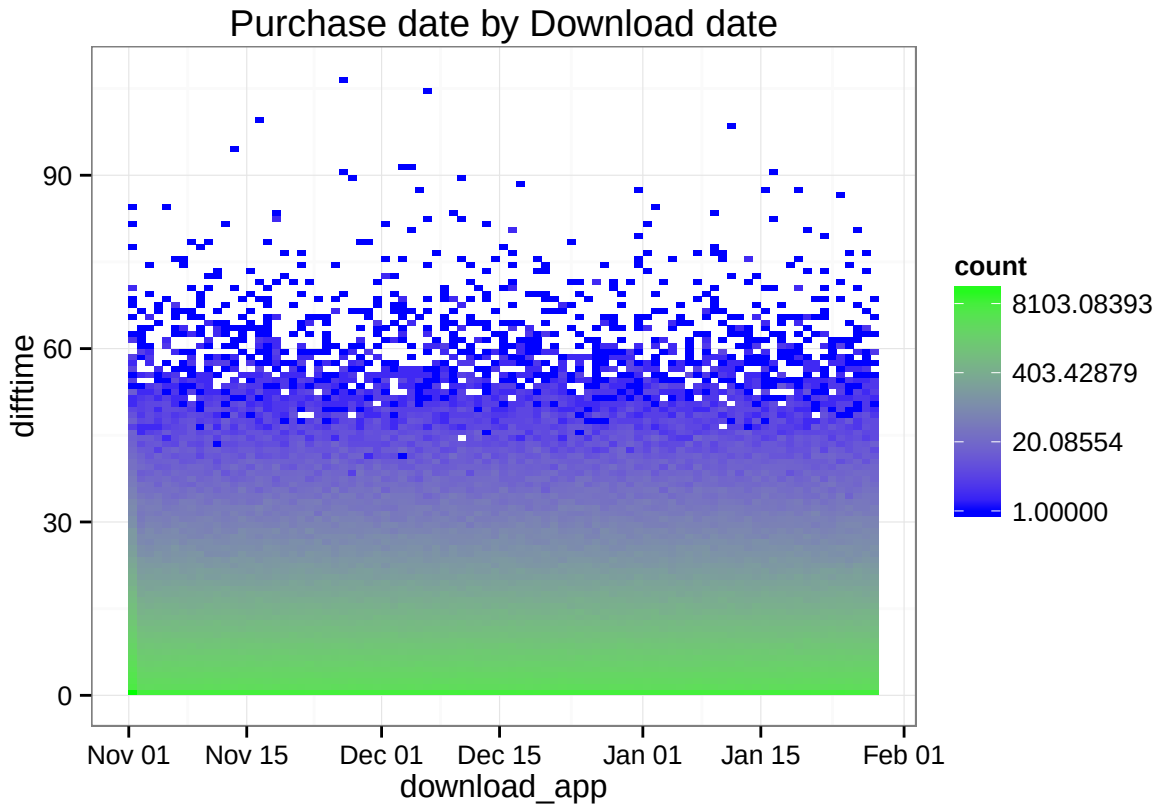
From tables and summary above we can note that there are no missing observations. Additionally, customers in this set of data either purchased same day or purchased after downloading the app (as there are no observations in the negative for both same day purchases and delayed download). Lastly, there 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.

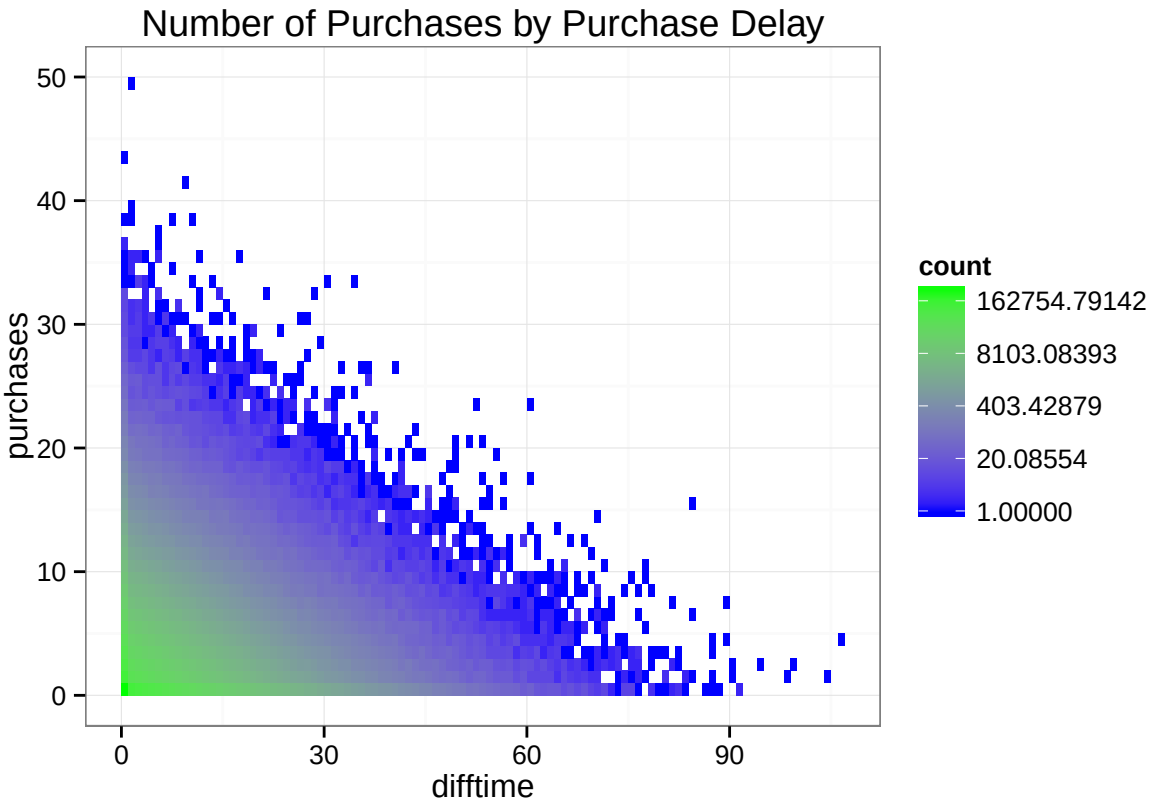
```
# Plot of Number of Purchases by Download and Purchase Date
df_plot <- reshape2::melt(df1, id=c("id", "purchases", "difftime", "same_day"))
ggplot(df_plot, aes(x=value, y=purchases)) +
  stat_bin2d(binwidth=c(1,1)) +
  facet_grid(same_day ~ variable, space = "free", scales = "free_x") +
  theme_bw() +
  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()
```




```
# 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 based 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 whether you purchased same day or later. This is fairly interesting because 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 shows a clear negative relationship such that a smaller purchase delay corresponds 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
```

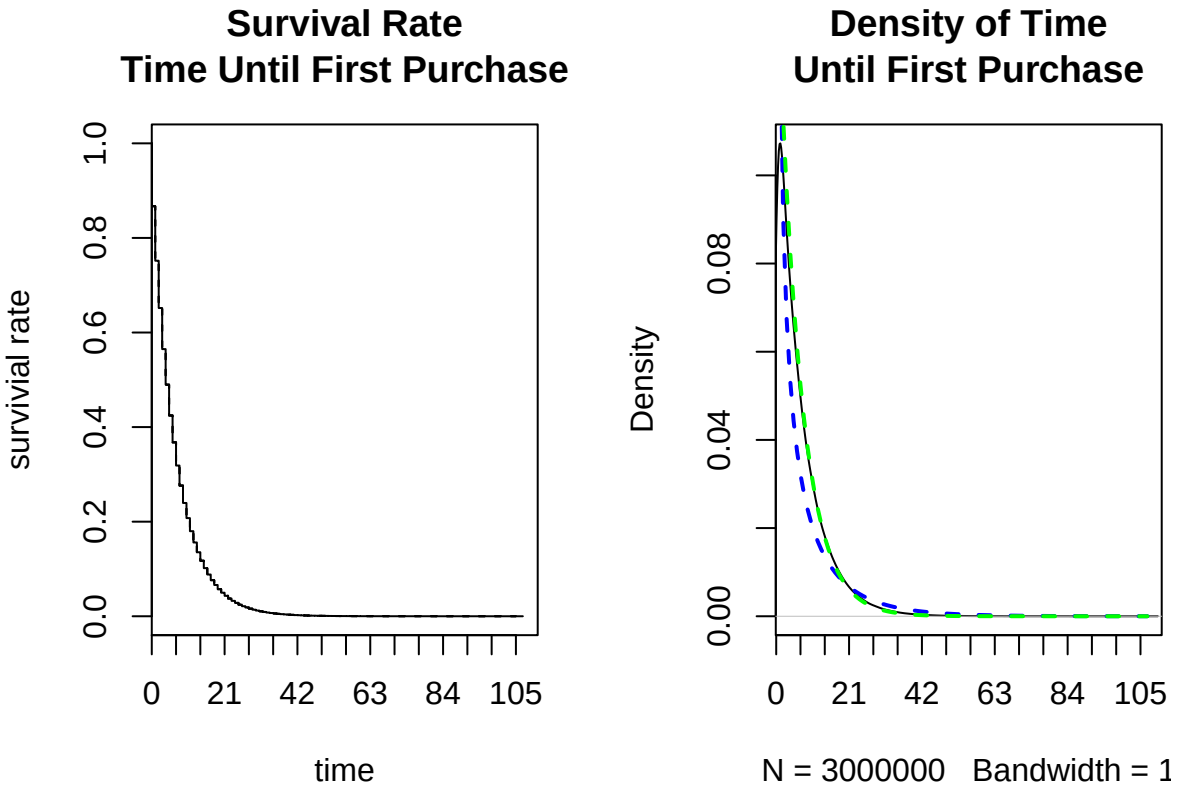
```r
df1$event <- 1

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 their 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 exponential 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 data set, each row of data represents one item in a larger order. Aggregating this data so that we can get more information about individual order will require some data transformation.

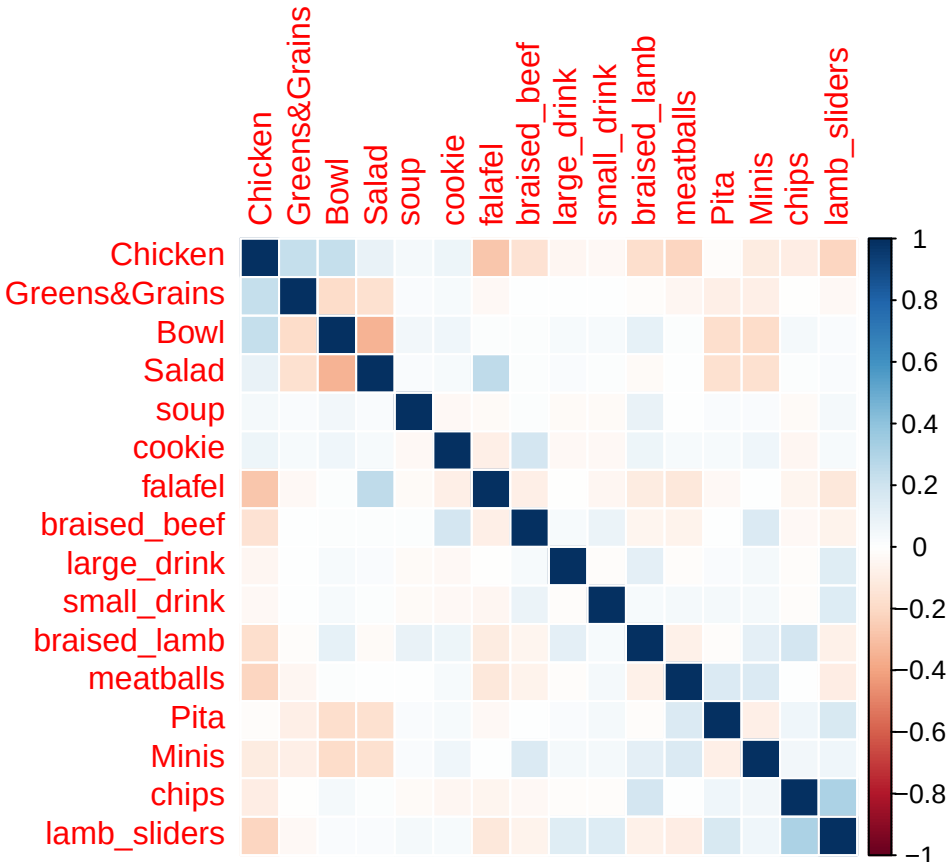
```
# Read in Data
df2 <- data.table::fread("cavaitemssold.csv", data.table = FALSE)

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))
```

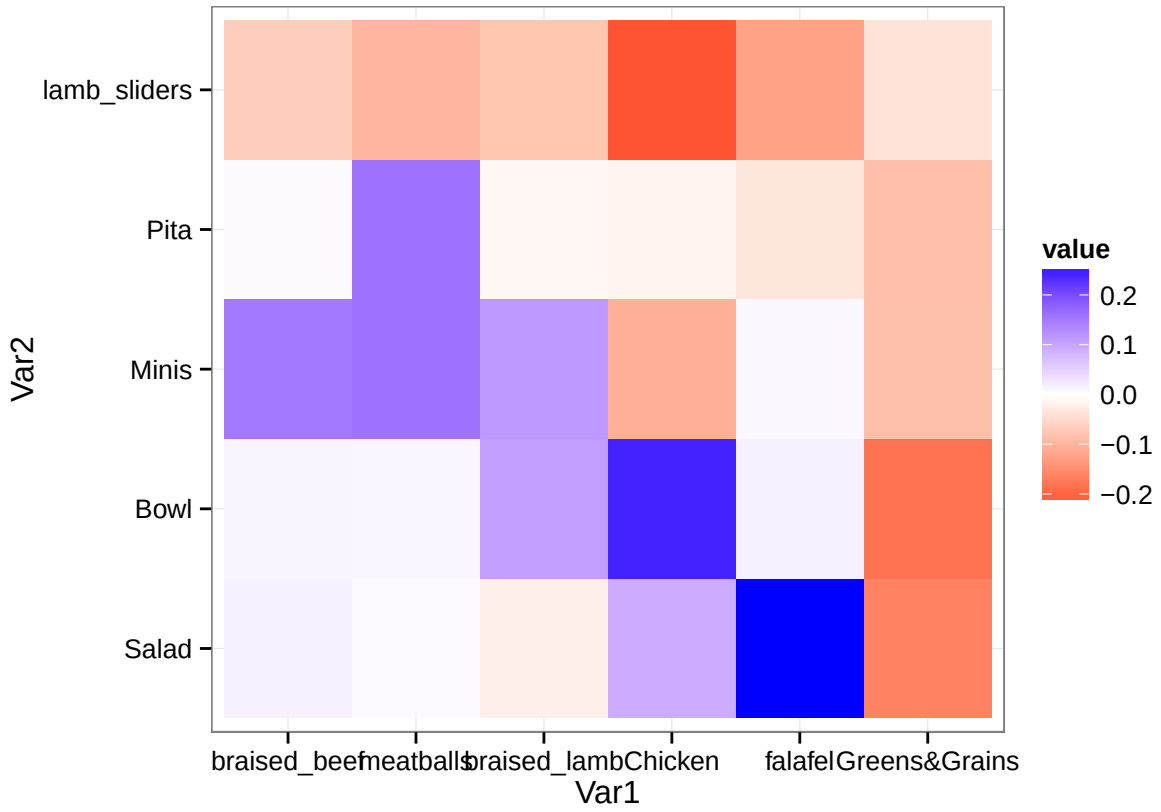
```
## Using transactionid as value column: use value.var to override.
## Joining by: "transactionid"
```

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



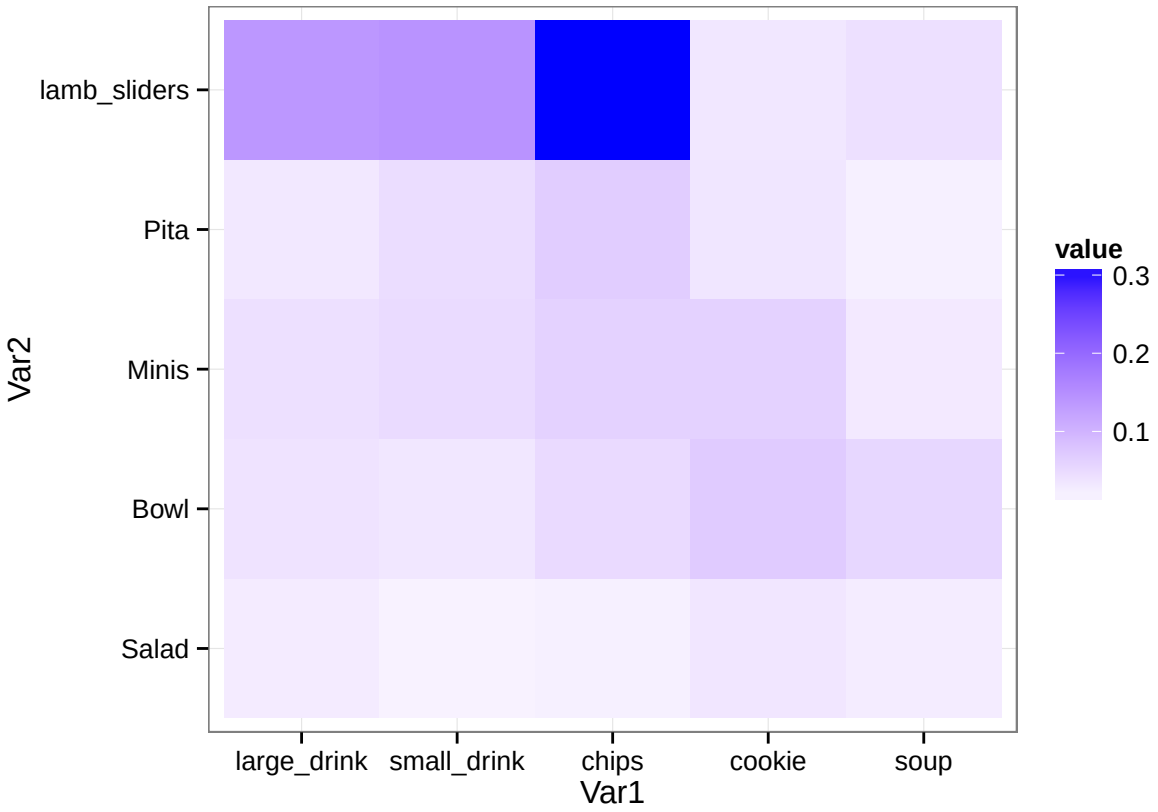

```
mealprotein <- reshape2::melt(cor(df2_orders[,protein],df2_orders[,meal]))
ggplot(mealprotein, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() +
  scale_fill_gradient2(low = "red", high = "blue", mid = "white") +
  theme_bw()
```

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



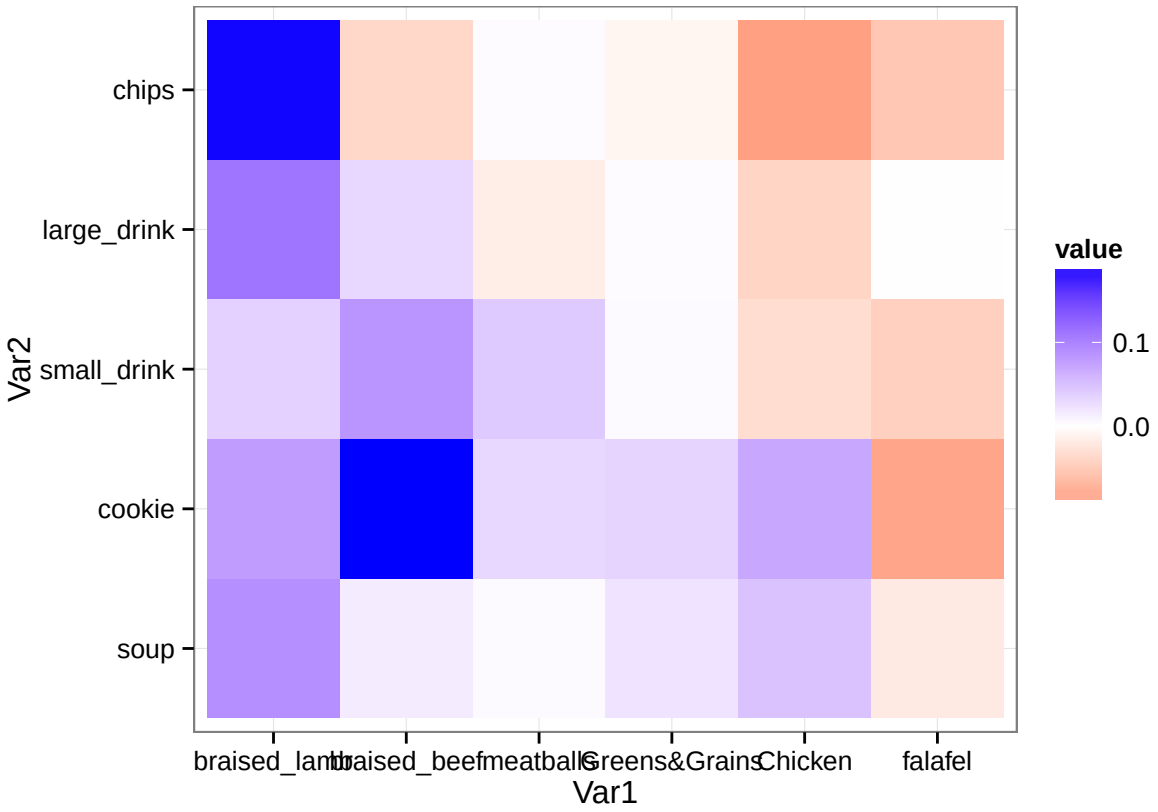

```
mealside <- reshape2::melt(cor(df2_orders[,side],df2_orders[,meal]))
ggplot(mealside, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() +
  scale_fill_gradient2(low = "red", high = "blue", mid = "white") +
  theme_bw()
```

```
## 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 data set. 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 largest 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 tendency 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.

Inital Aims

In our efforts to increase our customer base and volume of sales, it is necessary to find the most expedient avenues to accomplish this goal. A common business practice incentives new customers through a promotion, either by offering a discount or complimentary gift. Here at Cava Grill, we are also considering such a promotion, but in order to fully commit to such a program, we must evaluate the impact that such a campaign will have and whether it is worth the revenue forfeit by the promotion. This proposal will set for a framework for implementing such a promotion and evaluating it's success or failure. The evaluation will seek to prove two separate, but connected outcomes:

1. The promotion will decrease the time until a new customer's first purchase, and
2. The accelerated timeline will be maintained in the subsequent purchase habits of customers.

General Test Structure - A|B Testing

At its base, this type of test requires some type of comparison in order to be thoroughly evaluated. In this case, we will need a sample of potential customers that was exposed to the new promotion (treatment) and a sample of potential customers who was not exposed (control). Additionally, we have to determine the best way to identify new customers and not apply the promotion to existing customers. Leaving the challenge of removing repeat customers aside for the moment, we can consider different ways in which we might segment our customer into our treatment and control groups:

1. Offering the new customer promotion at certain restaurant locations

For this first method, we would create in store advertisements and promotions at different store locations. Here our treatment group would consist of new customers attained at locations with the promotion and our control group would consist of new customers attained at locations not offering the promotion. While this method would most likely be the easiest to implement, there are several drawbacks.

- Inherently, different store locations will have access to different subsets of the customer population (e.g. some locations will have much higher proportions of working professionals whereas another may be comprised mostly of college students)

- This model does not let us place individual customers in the treatment or control group as they may have access to multiple locations.

- Restaurant choice might very well present itself as a confounding factor. Let's say that we were able to pair each of our restaurants such that each pair shares access to population subsets with nearly identical characteristics; customers with access to a treatment and control location might very well opt to make their first purchase at the treatment location in order to take advantage of the promotion.

2. Randomly extend the promotion to our website and app visitors

This second method would allow us to directly place all of our website visitors into either the treatment or control group. By controlling for IP address, we will track the unique visitors to our page (or app users without a linked purchase) and then compare the total number of new customers who either used the promotion or did not across all of our locations. A small drawback to this approach is the inability to exclude repeat customers from the

3. Run a universal promotion for a short period of time, using past customer information as our control

The third approach tends to be a fairly common practice. This method would allow us to use all of our existing customer metrics and data collection as we would simply compare the trend before the promotion to the trend after the promotion. This type of quasi-experiment is beneficial for we would no longer have to deal with the randomization problem on new customers, however, it might create a confounding effect in that customers already introduced to Cava will be included once the treatment goes live, resulting in a potentially longer trial time for the experiment.

4. Use third party marketing with a treatment and control advertisement

Finally, the last approach is the use of third party marketing to manage the release of the promotion to our consumer population. By using a platform such as a Facebook ad campaign, despite the additional cost, we would be able to target customers both by location and multiple consumer demographics. Additionally, we can obtain an estimate for the number of customers being introduced through our advertisement. A drawback to this method is that we are unable to directly link potential new customers who where introduced to the control ad to those who weren't. However, this method does provide some distinct advantages:

- Our split of our customers into our testing and control group will be done automatically and done with fairly good precision by controlling for unique user

- With the addition of more consumer information, we can control for differences in the make up of the sample size when applying our test for effect

- Through the use of media tracking and page tagging, we can subset our population further to follow just those reached by the test

The Repeat Customer Problem

Turning back to the consideration that we set aside, we want to make sure that we are not including repeat customers within our promotion experiment. The best way to accomplish this task would be to encourage new customers to sign up and use our app to enroll in the promotion and pay for their meals. By adding in this additional component, while we would be excluding those unwilling to use our app, we could directly track those who downloaded the app via one of our advertisements and whether they were exposed to the promotion. Additionally, we can directly track how long it takes from exposure to download to purchase. Additionally, we can exclude from the analysis consumers whose credit cards are already in our purchase database.

Sample Size Calculation

Determining the proper sample size for this test depends on our current ad to app download conversion rate, the increase at which we find the promotion monetarily feasible, and the confidence with which we are comfortable making that conclusion. Given that information, we could calculate the difference in proportion based on the following equation:

$$n = \frac{z^2 pq}{d^2}$$

For example, if our initial conversion rate was 2%, we wanted to use a 95% confidence interval, and we needed an increased conversion of 0.5% to justify the promotion, our sample size would be as follows:

$$n = \frac{(1.96)^2(0.02)(0.98)}{0.05^2} = 3011.81$$

Evaluation of Statistical Test

The test statistic for the difference of proportion test would be as follows:

$$z^* = \frac{(\hat{p} - p_0) - d_0}{\sqrt{\hat{p}_1(1-\hat{p}_1) + \hat{p}_2(1-\hat{p}_1)}}\sqrt{n}$$

We can consider the test to be significant if our test statistic is greater than 1.96; where $d_0 :=$ the desired difference in conversion rate.

Further Extrapolation

An additional application of this test would be to apply the demographic information we obtained through the 3rd party ad placement to further extrapolate the effect that the promotion will have on our consumer population. Either through the use of a survey weighting scheme or the application of a hierarchical linear model, we may be able to obtain closer estimates for the projected increase in business.

Evaluation of the Initial Aims

The first objective expressed in this proposal was determining that this promotion will decrease the time until a customer's first purchase. Given the collection method for our promotion experiment, we can consider the time until purchase as the time from a consumer's initial exposure to a marketing campaign to the time of their first

purchase.

The second objective expressed in this proposal was determining that the time between subsequent purchases also decreased. For this aim, we can actually use a much larger selection of our internal data, but we can consider time between purchases to be the average amount of time between purchases for any particular customer.

Both of these objectives can be evaluated though the use of maximum likelihood estimation. Thus for each set, we will construct probability distributions for the waiting times for customers between exposure and first purchase and between subsequent purchases respectively. In the former case, we can create this distribution from the customers within the control group that made a purchase. In the latter case, we can use large collection of data that corresponds to the delay between purchases given that these customers did not participate in the promotion. These probability distributions will be constructed through solving the following optimization problem:

$$\underset{\Theta}{\operatorname{argmax}}\ P(\Theta|x_1, ..., x_n)$$

Where $\Theta :=$ the parameterization of a given probability distribution

Evaluating our objectives can thus be accomplished by calculating the probability of the data, given the result of the optimization problem above [$P(x_1^*, ..., x_n^*|\Theta_{ML})$]. If that probability is significantly small, we can conclude that the times for the experimental groups follows a different distribution, for which we can create a new ML estimate.