MEMORANDUM

For the purpose of this project, we found the key metrics of success and plan a specific set of marketing strategies for Walmart, one of the retailers covered in the *comScore E-retailer Dataset.* The key metrics include understanding the hours with most traction, along with the persona archetype of the retail giant's loyal customer base—both factors measured against the *click-to-checkout rate.* After conducting multiple HiveQL queries on Ambari, coding in R Studio, reiterating Poisson and logistic regressions, clustering via RapidMiner, followed by visualizing the insights using Tableau, we found out that Walmart has to: 1) integrate targeted Advertising at a certain group per hour and 2) boost its loyal customer base using coupons and loyalty system. By doing so, the retail giant will achieve a competitive advantage superior to the competing players in its market. As such, we will discuss further both methods and action items.

Understanding Walmart, its business needs and industry helps keep the end in mind

Walmart (est. 1962) is a supermarket that promises "everyday low-cost" products. Therefore, the strength of its brand identity heavily relies in how the company can deliver a wide range of goods from many different areas with such low price points.[1] *Appendix A* highlights the company's strengths, weaknesses, and the situational landscape of competitors and the retail industry in which Walmart dominates. Overall, the value proposition of Walmart also depends on its ability to offer a wide range of product options based on each customer's unique needs besides the products' level of affordability.

Nowadays, customers tend to purchase online. Based on a study by comScore, buyers have a significantly higher tendency to check out their cart and make a purchase online—over their mobile phones than their desktop, specifically.[2] Consequently, Walmart is challenged to keep up with such behavioral change, which is widely adopted by customers internationally. The company then acquired *Jet.com* to compensate for the missing puzzle piece in its e-commerce initiatives, as compared to Amazon.[3] In terms of other incumbents in the low-priced retail market; Walmart has always been at the forefront and the first-mover. Target, Costco, and Sam's Club have been its indirect competitors. However, if Walmart plans to enter the e-commerce market, it will compete head-to-head with Amazon. Through this project, we will help Walmart do so by specifically targeting certain segment at a certain hour and growing its loyal user base.

Data Understanding leads our team to formulate profitable research questions

Our team decided to use Walmart as the retailer to consult, because it has the most data points within the dataset gathered by comScore. Extracting just the click count data of Walmart, we summarized its 4V's: 1) Volume – the full Walmart dataset has a total of 26,701 rows and 18 columns, together they display a sum of 480,618 instances*,* 2) Velocity – between the period of March X to Z of Y, each click was recorded per given session, URL domain, directory, and page—recording an enormous amount of data in a short span of 30-minute each session**,** (March 4-24, 2014) 3) Variety – there are two types of characteristic groups in this case, that is the clickstream dataset captured real-time and the demographics dataset pre-loaded from the CRM database system, and 4) Veracity – each dataset shows a reasonable amount of structure and consistency, both sets combined still maintain the same data integrity.

After knowing the basic knowledge of the datasets, our team then built connections between the eighteen building blocks, for example: session_id and age, clickthrough per session_id per time_period, etc. Hence, we generated research questions as follows: 1) in which hour does Walmart gain the most traction? And who populates most of such traction? I.e. when should Walmart target their Ads to such specific group of

people with similar characteristics? 2) which customer persona is the most likely to check out their purchase carts?

Analysis using R Studio and Hive taught us about the hour for highest traction

The first research question we had as mentioned earlier was the timing in which Walmart can target its customers most intensely due to its highest traffic measured by the click count-to-checkout rate. Instead of using the total click count where crazy or abnormal customer could skew the summary takeaway heavily—we grouped the original into buckets of "session_id." In addition to increasing the robustness of the data, such grouping by "session_id" technique also helped us to get more clarity and a sense of intuition on the demographic information of the customers who clicked per session, identified through their machine ID's. The step-by-step methodology that we followed as our procedure is as follows:

1. First, we group by "session_id" to eliminate counting the clicks from the same user twice or more. The biggest factors behind us choosing to group by "session_id" are:
    a. #1 Observation: we realized that the demographic variables "age", "gender," "income," "children," "ethnicity," and "education" did not change during one single session.
    b. #2 Observation: we also learned that "day" and "hour" also did not change substantially in a single session.
2. Following, we also take the sum of "duration" and "click" amounts for a single session so we know, in this session, what does the customer's total shopping duration and how many clicks does he or she made in this session.
3. For "checkout," we take the maximum of each single session because once this customer has ever proceeded to the checkout page; we consider he or she bought one or more item(s) to distinguish those who bought something and those who did not. (Note: However, we could not guarantee a statement of purchase due to the fact that we do not have the cart-abandonment-rate or the purchase data. Moreover, the customers could have just closed the tab without buying anything. Despite the limited amount of data, we think our approach is fair and will not produce large error.)
4. In addition, we also create two new variables called "ccrate" and "hourday".
    a. 'ccrate' means 'Click to Check out Rate', which equals to (check-out/clicks). It tells us how many clicks does a customer clicked before he or she checked out (if he or she did checked out, if not the ccrate = 0).
    b. 'hourday' equals to 100*day+hour, which gives us a intuitive and straightforward sense of the hour and day when the session happened. (e.g. 213 means 13:00 or 1 PM on Tuesday)

As a result, we generated a new dataset in addition to the one that we originally had, comprising the full Walmart data not grouped by "session_id." Such approach allows us to learn about the hour with most traction and the delicacies of each type of customer segment.

Customers tend to consistently buy more on Wednesday and weekends' peak hours

To further explore the dataset, we asked deeper questions in regard to timing of visitors finally landing on the checkout page. Such questions include: 1) which day is most popular for users to shop and check out? 2) Who is the specific target segment that shopped at such hour? Thus, we figured that Wednesday shows the highest traction, followed by Friday evening hours, Saturday and Sunday all-day. *Appendix B* shows a diagram of when is the most popular time for Walmart customers to shop. Based on our research combined with the analysis, we also found that on Wednesdays in particular, traffic picks up as shoppers

likely see an end to their work week approaching. Focus declines and shopping favorability increases.[5] As such, we would recommend that Walmart conduct a time-specific targeted advertising campaigns on a specific target market to those who are between 30 to 40 years old starting from 5PM until midnight. *Appendix C* further highlights the other series of hours most popular among users—that is from Friday evening through Saturday and Sunday all-day, with Sunday evening 3-8PM showing the most traction and highest 'ccrate' (click to checkout rate). Similarly, we also approached the data through Poisson and logistic regressions along with clustering to identify the persona of the most loyal customer base of Walmart.

Iterative process of logistic regressions and basic analysis paint the picture of personas

For the regression analysis, we decided to study how the factors decide if the customer is checking out. Since the result is either "0" (not checking out) or "1" (checking out), we applied logistic regression in this particular model. The tool for this task is *Python Jupyter Notebook.* The detailed process is listed on *Appendix D.* We first obtained the data with sessions combined from Hive query. Then we cleaned the data and created dummy variables for categorical data. After we put the relevant factors into the model, we got the result in the table. There isn't a very shocking finding from the result but it did provide a couple of insights below:

1. Age, clicks and duration have positive influence on checking out
2. More children might make people more hesitate to check out
3. Males are less likely to check out than females, despite both having negative coefficient
4. High-income class (household income more than 75,000) is actually less likely to check out than mid-income class and low-income class

Such insights provide a confirmation on the fact that Walmart needs to re-focus its branding on the already-loyal customers—those who are budget-conscious, price-sensitive, and are in their middle age as previously discussed. That said we further explore to build stronger personas through digging deeper in the R studio and clustering via Rapid Miner.

'Click-to-checkout' rate leads to refocusing the marketing dollars in mid-age, mid-income

We asked more research questions such as: 1) Is there a difference of the 'click-to-checkout' rate between genders? 2) Is there a difference of shopping duration time between genders? 3) Is there a statistical relationship between 'click-to-checkout' rate and ages? 4) What is the difference of the age distribution between those who checkout and those who does not? 5) From a marketing perspective, which group of customers is Walmart's 'Loyal customers' and when should Walmart send them advertisement emails or messages to maximize revenue? As such, we attained results that:

1. The mean 'click-to-check' rate is almost the same for males and females. (0.25876% vs 0.25881%)
2. Females tend to do more window-shopping than males, as reflected by the mean duration time. (293 vs 359)
3. According to the two-sample t test, the age differs significantly between those who buy something and those who do not. (p-value = 0.01318)

We find the age in the 'buy' group is larger than age in the 'nobuy' group. People around 22-30 years old consist of the largest amount of 'buy' group, while people who age 24-42 consist of the largest amount of 'nobuy' group. The t-test and the density plot further testify that middle-age people are most likely to

shop on Walmart. Thereby, we conclude the middle-age group (age >= 24 and <= 40) as Walmart's 'Loyal Customers'. Now, we know who the target customers that we should focus on. We want to know when (Hour and Day) do our 'Loyal Customers' are most likely to buy something (as reflected by checkout) so we could send them promotional email or discount opportunity at that time to let them buy more, increasing our revenue.

Clustering further leads us to understanding the 'when' and 'who' Walmart should target

Initially, we looked at the characteristics of all web surfers for Walmart, in order to determine eventually their propensity to make a purchase. However, this led to heavily unfavorable data, reflecting that any cluster had almost no willingness to purchase. This was later found to be because the number of sessions involving purchases was only about 5% of total sessions. In order to reflect the customers who actually made purchases, we needed to further clean the data:

- Condensed the above dataset, grouped by session_id, further to only incorporate those sessions involving a purchase. This was because it allows us to see what type of person actually completes a purchase, not just all those who visit Walmart's site.
- Categorized the timing, in order to make it easier to convert nominal variables into fewer dummy variables for cluster analysis.

    - Hours of the day were separated into 4 groups, 6 hours each: "Early Morning" before 6am, "Morning" before 12pm, "Afternoon" before 6pm, "Night" before 12am.
    - Days of the week were separated into either "Weekday" (Mon-Thurs) or "Weekend" (Fri-Sun)

Through conducting cluster analysis, we found that the amount of clusters did not provide too many stark differences between clusters. From 3 clusters to 10+ clusters, the data still reflected largely the same kinds of purchasers. However, we selected 5 clusters because of the proportions of purchasers within each cluster - the proportions within some of the clusters for processes with more clusters (6+ clusters) were very small, sometimes representing 1% of the sample. We wanted to focus on clusters that were little more even in population, thus we selected 5.

Clustering also helps us build a stronger understanding of customers' behaviors

There was little variability across clusters for most attributes, although magnitudes differed. This low variability may have been due to the smaller sample size of 340 purchases, or perhaps due to many purchases performed by the same person, albeit on different machines, since this form of cluster analysis is not the most robust. Walmart has many customers, and perhaps the same type of person makes purchases online, while most of the underrepresented customer types in this analysis visit the store.
Members of the first cluster have completed some higher education, and have a lower-middle income. This was also the only group that included Spanish-dominant ethnicities.

Mostly lower income purchasers who mostly purchased at night, but also purchased a lot in the afternoon represented the second cluster. Members of the third cluster were the youngest, which corresponded with the lowest amount of clicks per session and the lowest session duration. These customers were also the only group to purchase more during on weekdays. Conversely, the fourth cluster was the oldest, corresponding with the largest amount of clicks per session, and duration 10 times as long as the quickest cluster. This cluster also had the highest proportion of customers with incomes over $100,000, and was the most likely to purchase at night. The last cluster had the most even gender balance, where most clusters were ⅔ women. This cluster had the most customers with unknown levels of education.

Based on cluster population, we were able to learn which cluster completed the most purchases, which tells us an estimate of which kind of customer is the most likely to purchase. Our analysis indicated that Cluster 3 had the highest proportion of purchase sessions, representing about 34% of purchases sessions. This means that Walmart should cater towards these customers more, in order to retain key customer types.

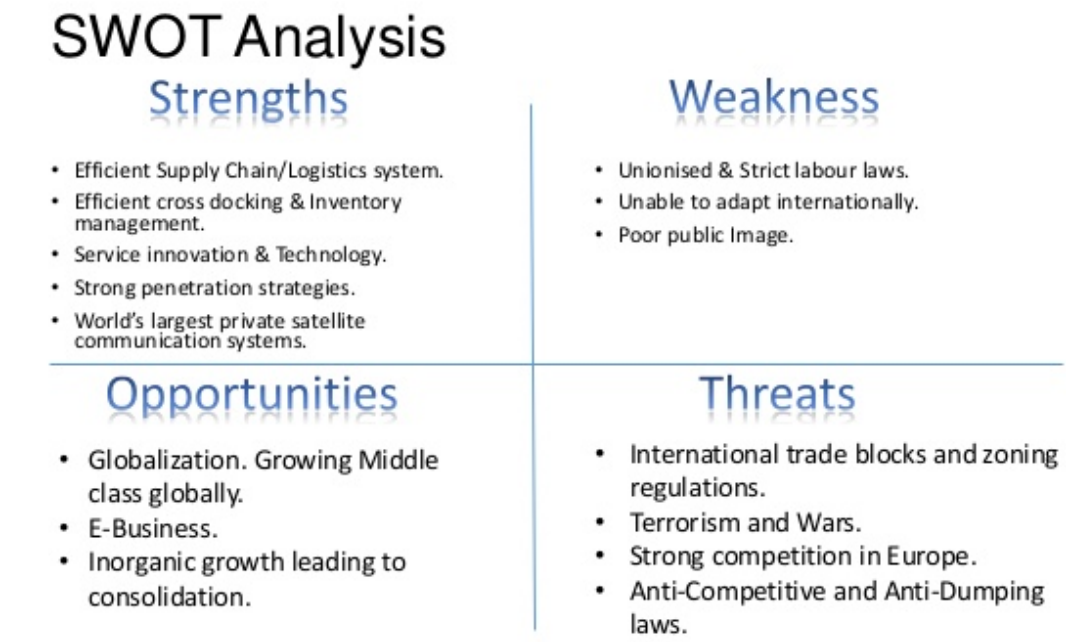
After conducting the Cluster Analysis, we decided that Walmart should target these customers. Cluster 3 is made up of the youngest customers, who most likely value time saving the most, hence their involvement with online purchases. Walmart should advertise with a user-friendly interface that lets customers cruise through the checkout process with ease. By catering to their largest purchasing group, they will gain loyalty, which will ultimately result in higher revenues for Walmart.

Next steps: we recommend that Walmart tailors its advertising and grow a loyal fan base

Through our insights, we have concluded that Walmart has a large opportunity to take advantage of by using big data analysis. Through Cluster Analysis, we determined that Walmart should mainly target younger customers through marketing campaigns. This finding was supported by our regression and statistical analysis, reflecting that there is a significant difference in age as a factor to making a purchase. In both Cluster and Statistical Analysis, gender was not found to have much a difference in checkouts, although it was a factor in session duration. Furthermore, we explored the timing of purchases, determining that Wednesdays and peak hours during the weekends were the times involving most purchases. In order to capitalize on these insights, Walmart should focus on advertising to customers with these characteristics, and by catering to its most frequent customers, the company will grow its customer base into more loyal customers, which will contribute to its long-term success.

Appendices

Appendix A: Business Understanding - Competitor's Landscape (SWOT + PESCE)



Appendix B: Walmart's user persona, visitors and the Hive queries behind the analysis

dayofweek	hour	duration	gender_id	age	ıh_income_ic	children_id	ethnicity_id	Checkout
1	0	0	2	22	84003	0	10	1
1	0	0	2	38	84004	1	10	1
1	0	0	2	44	84005	1	10	1
1	0	0	2	45	84003	0	10	1
1	0	1	1	25	84005	0	10	1
1	0	1	1	43	84002	1	10	1
1	0	1	1	45	84001	1	12	1
1	0	1	1	55	84006	0	10	1
1	0	1	2	38	84004	1	10	1
1	0	1	2	44	84005	1	10	1
1	0	2	1	43	84002	1	10	1
1	0	2	2	23	84001	0	10	1
1	0	2	2	57	84002	1	10	1
1	0	3	1	25	84005	0	10	1
1	0	3	1	37	84004	1	10	1

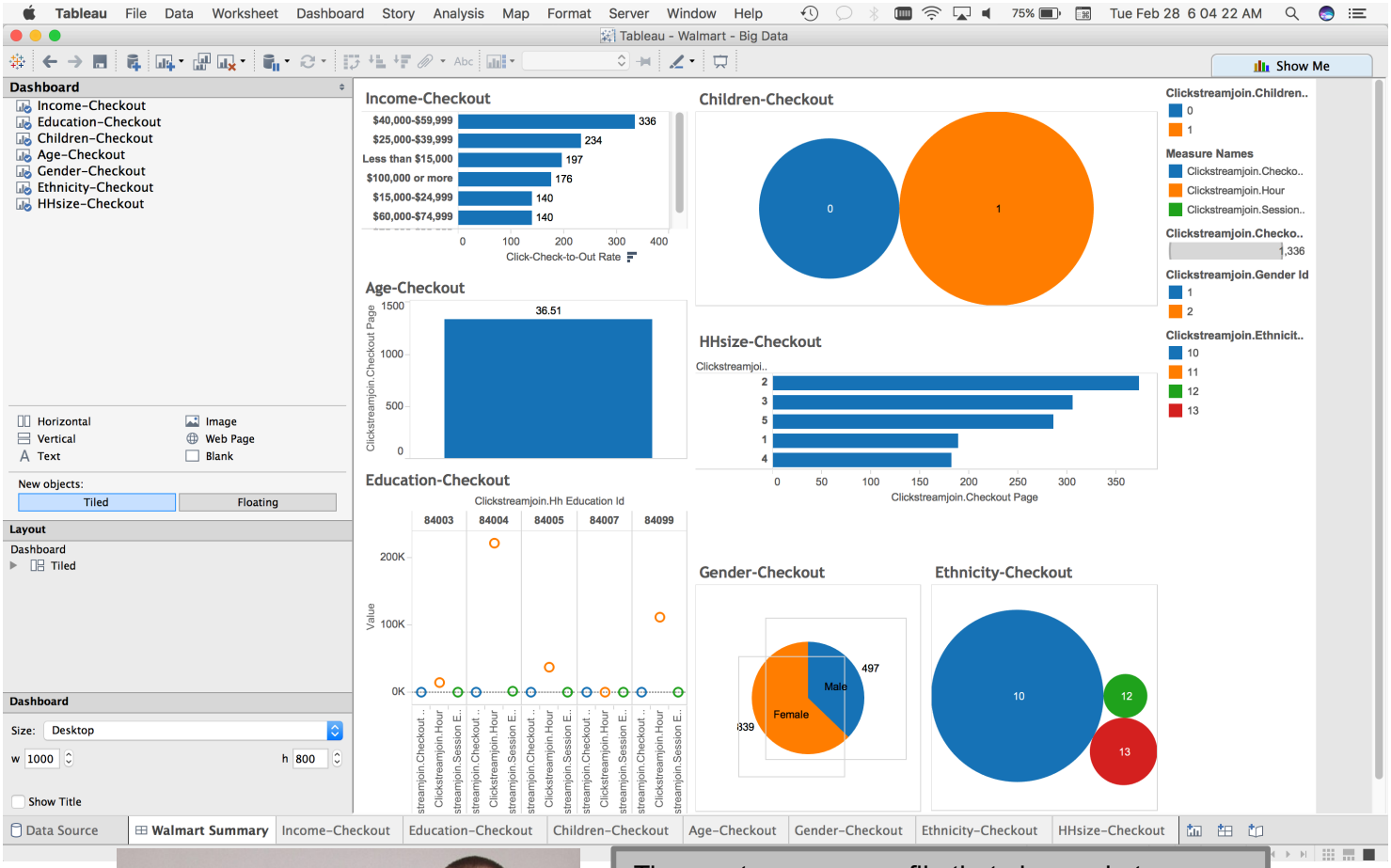
Persona Insights of Those Who Checked Out

TIME	GENDER	AGE	HOUSEHOLD	ETHNICITY
Dayofweek:	**Gender:**	**Age:**	**Income:**	**Ethnicity:**
Most common	Most common	Most common	Most common	Most common
7	2	23	84004	10
SUNDAY	**FEMALE**	**23 YEARS OLD**	**$40,000-$59,999**	**NON-HISPANIC**
Average	Average	Average	Average	Average
4.066843555	1.612613532	36.17434432	84003.77073	10.33191142
THURSDAY	**FEMALE**	**36 YEARS OLD**	**$25,000-$39,999**	**NON-HISPANIC**

Hour:	Duration:	Children:
Most common	Most common	Most common
19	120	1
7PM	**120 MINUTES**	**YES**
Average	Average	Average
14.54632105	43.57301765	0.719767323
2-3PM	**44 MINUTES**	**YES**

Conclusion: A Common Scenario

On Sunday evening or night between 2 to 8 PM, many mid-age non-Hispanic women around the age of 36 years old with one or more children who have a median household income in between $40,000-$59,999 visited Walmart's website, clicked around for about 45 minutes, and then proceeded to the checkout page.



Source: http://i.dailymail.co.uk/i/pix/2013/07/24/article-2377056-1AFA7173000005DC-978_634x893.jpg

The most common profile that shopped at Walmart is mid-age, non-Hispanic women with one or more children, mid-level income. As such, they tend to shop on between Thursdays through Sunday (2-8PM). Thus, Walmart should focus targeted advertising at this segment at such

Source: http://cdn.moneycrashers.com/wp-content/uploads/2015/12/asian-american-family-918x516.jpg

Hive Queries:

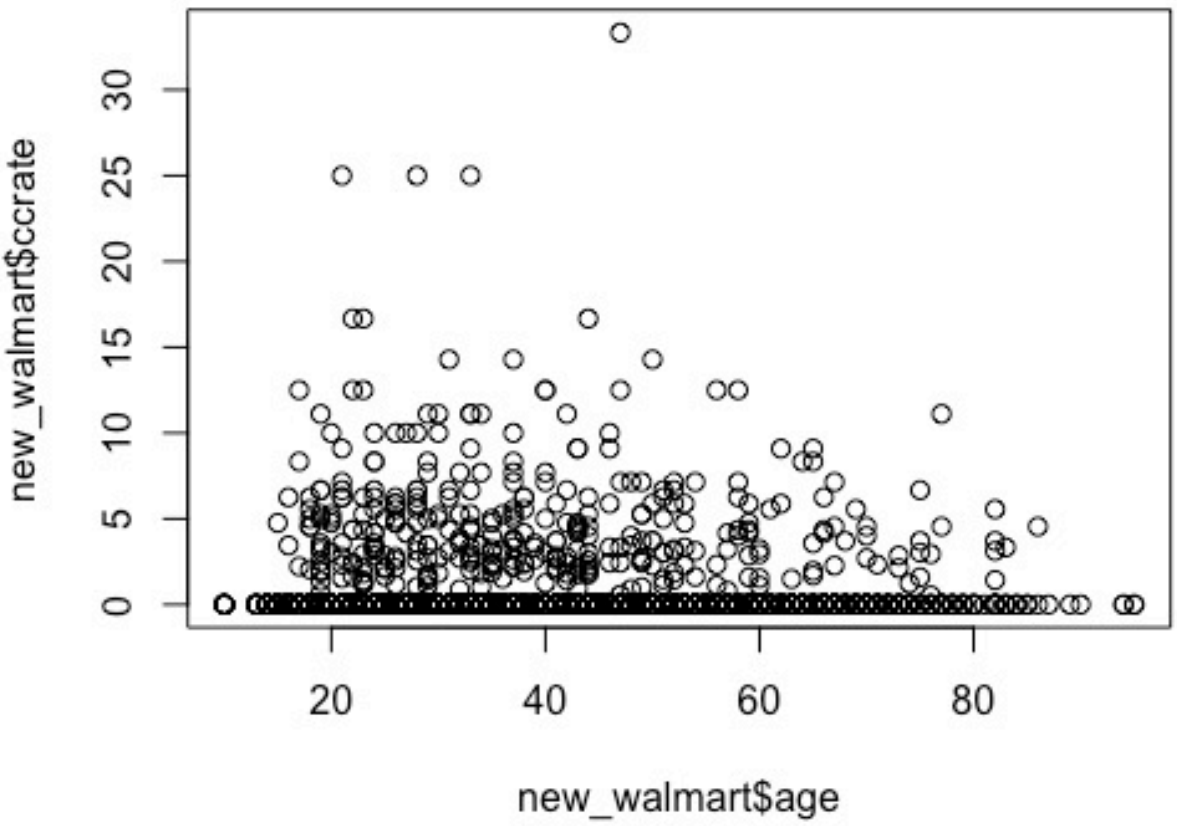
```
create database clickstream;
use clickstream;
create external table clickstream_log
(
url_host string,
url_dir string,
url_page string,
machine_id string,
duration int,
checkout_page int,
time string,
session_id string,
session_exit string
)
row format delimited
fields terminated by '\t'
stored as textfile
location '/user/admin1/clickstream_analytics/clickstream';

create external table demographics
(
time_period_id int,
machine_id2 string,
age int,
gender_id string,
hh_income_id string,
hh_size_id string,
hh_education_id string,
children_id string,
ethnicity_id string
)
row format delimited
fields terminated by '\t'
stored as textfile
```

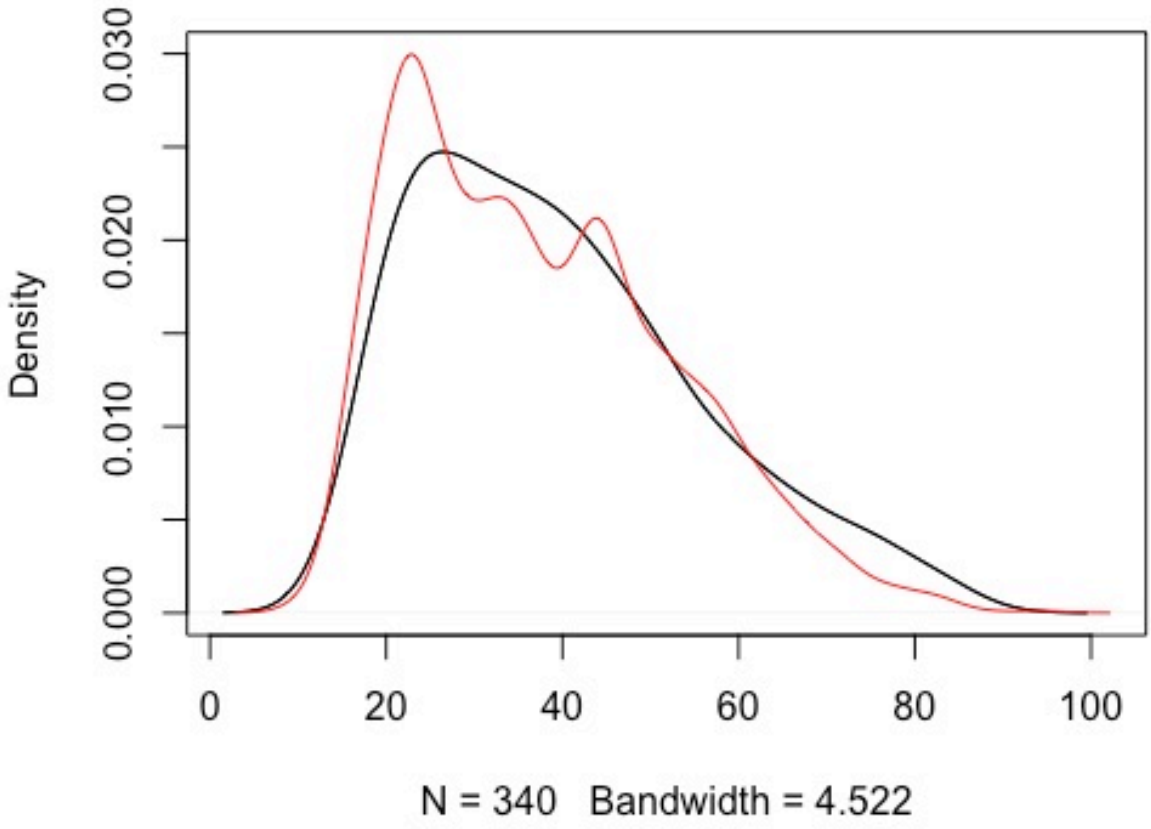
```
location '/user/admin1/clickstream_analytics/demographics';

create table clickstream.clickstream_hour as
select hour(time) as hour, *
from clickstream.clickstream_log;

create table clickstream.clickstream_hour_day as
select from_unixtime(unix_timestamp(time),'u') as dayofweek, *
from clickstream.clickstream_hour;

select dayofweek, count(*) as clickcount
from clickstream.clickstream_hour_day
where url_host like '%walmart%'
group by dayofweek;

SELECT dayofweek, hour, count(*) as clickcount
from clickstream.clickstream_hour_day
where url_host like '%walmart%'
group by dayofweek, hour;

// Line 50-53 is part of Exercise 1. Visualization was conducted on
both platforms, Hive view and Excel.

create table clickstream.clickstreamjoin as
select * from clickstream_hour_day left outer join
demographics
on clickstream_hour_day.machine_id = demographics.machine_id2;

select dayofweek, gender_id, count(*) as clickcount
from clickstreamjoin
where url_host like '%walmart%'
group by dayofweek, gender_id;

select dayofweek, hour, gender_id, count(*) as clickcount
from clickstreamjoin
```

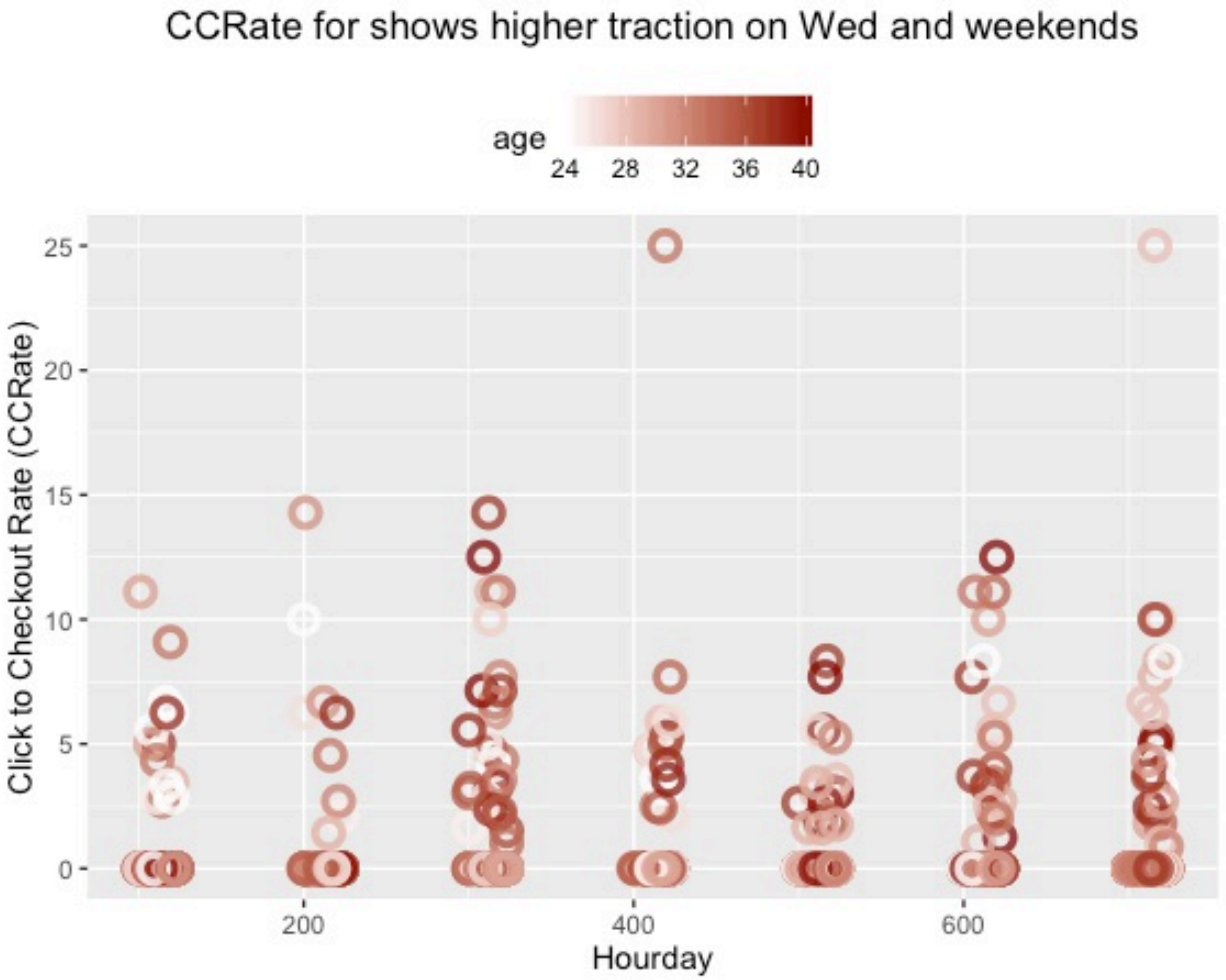
```
where url_host like '%walmart%'
group by dayofweek, hour, gender_id;

select dayofweek, hour, gender_id, count(distinct session_id) as clickcount
from clickstreamjoin
where url_host like '%walmart%'
group by dayofweek, hour, gender_id;

select dayofweek, hour, gender_id, count(distinct checkout_page) as checkout
from clickstreamjoin
where url_host like '%walmart%'
group by dayofweek, hour, gender_id;

select dayofweek, hour, duration, gender_id, age, hh_income_id, children_id,
ethnicity_id, count(distinct checkout_page) as checkout
from clickstreamjoin
where url_host like '%walmart%'
group by dayofweek, hour, duration, gender_id, age, hh_income_id, children_id,
ethnicity_id;
```

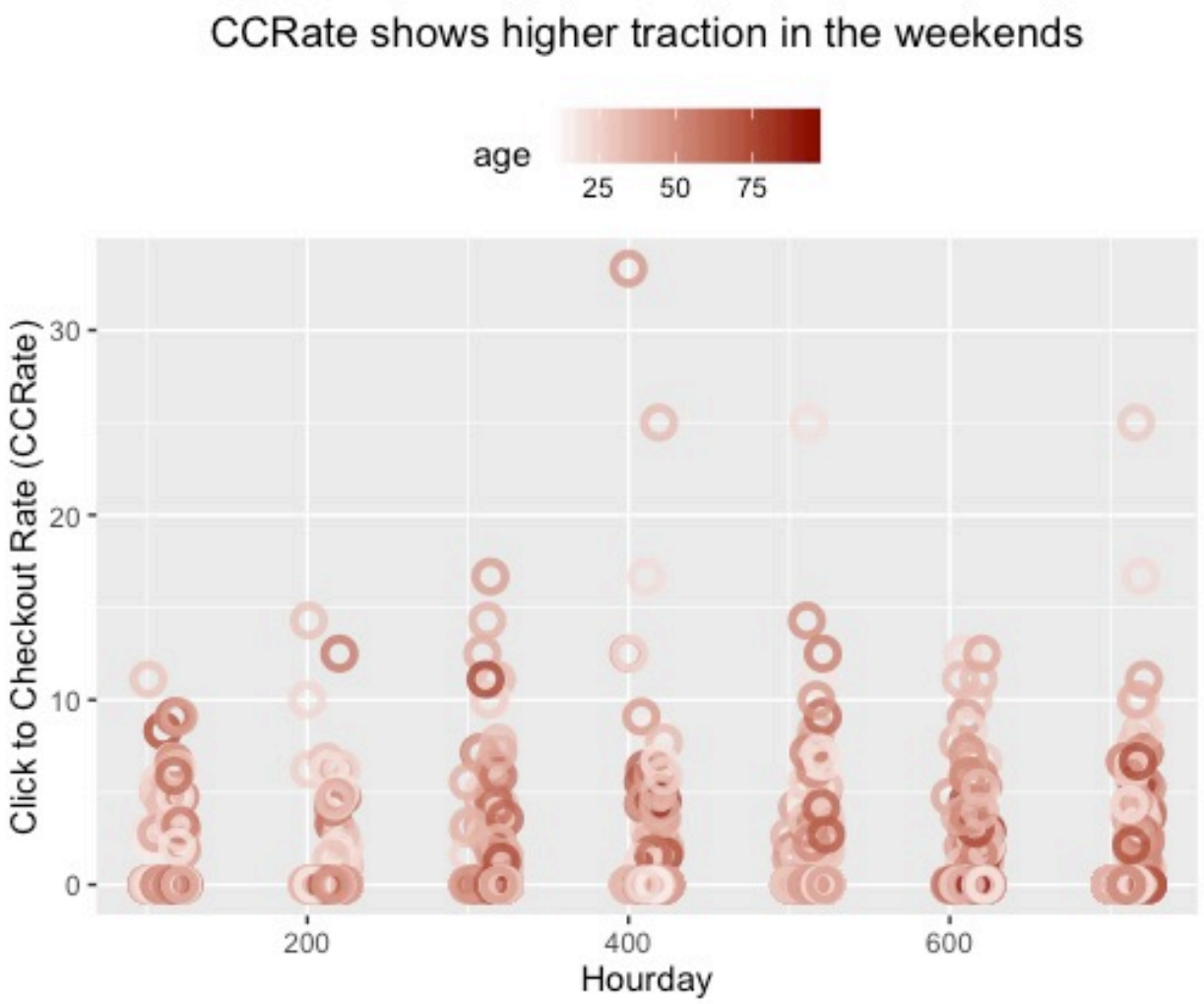
Appendix C: R and RStudio

Data Visualization:

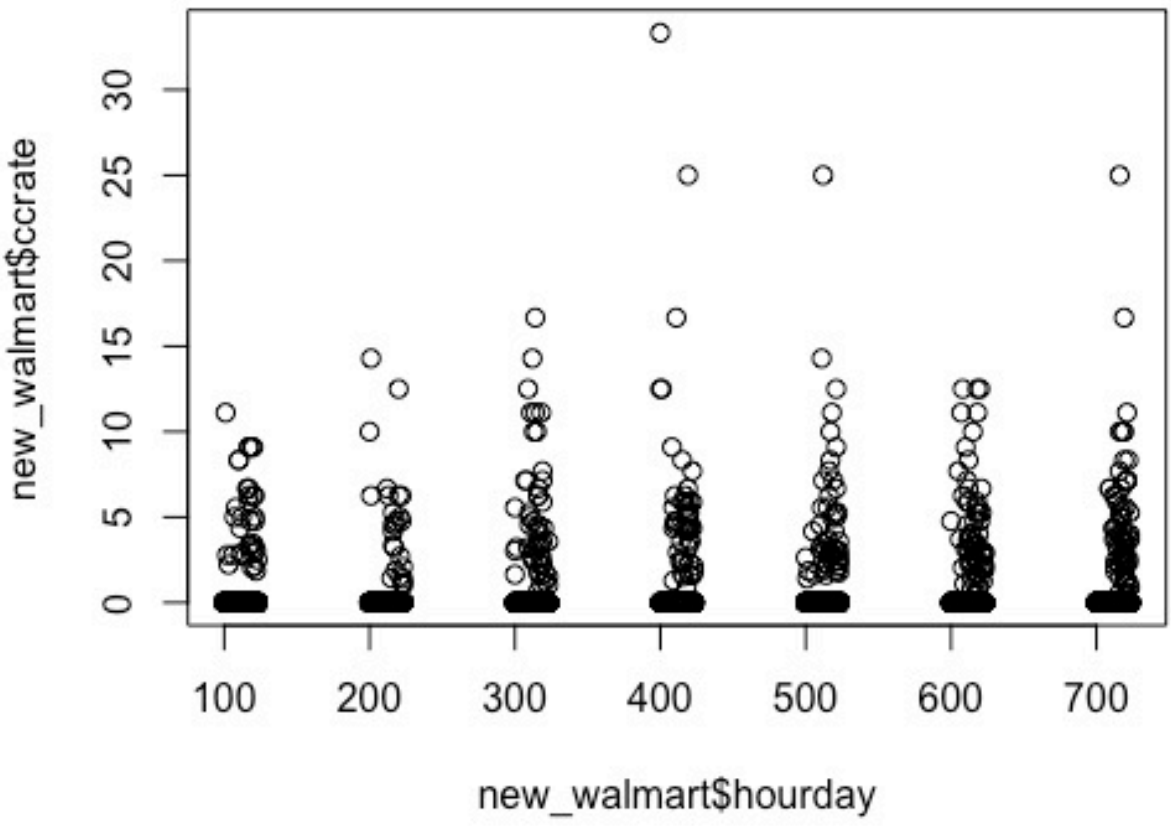



Density
0.000
0.010
0.020
0.030
0
20
40
60
80
100
N = 340 Bandwidth = 4.522


CCRate for shows higher traction on Wed and weekends
age
24
28
32
36
40
Click to Checkout Rate (CCRate)
25
20
15
10
5
0
200
400
600
Hourday


CCRate shows higher traction in the weekends
age
25
50
75
Click to Checkout Rate (CCRate)
30
20
10
0
200
400
600
Hourday


Click-to-Check Rate
Click-to-Check Rate
40
30
20
10
0
200
400
600
Hourday
Gender
1
2



Further Insights:

```
> mean(new_walmart$ccrate[new_walmart$gender == 1])
[1] 0.2587591
> mean(new_walmart$ccrate[new_walmart$gender == 2])
[1] 0.2588153

> mean(new_walmart$duration[new_walmart$gender == 1])
[1] 293.0181
> mean(new_walmart$duration[new_walmart$gender == 2])
[1] 359.3059
```

The coding lines used within the R Studio:

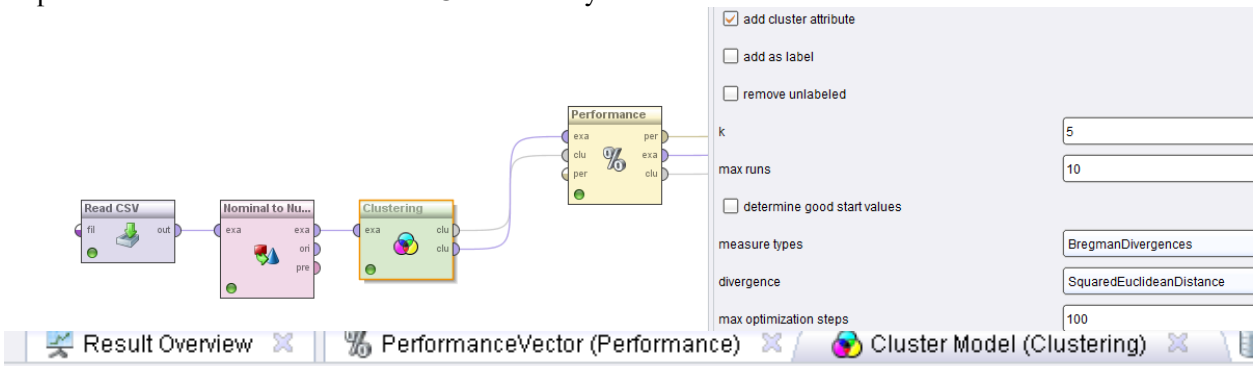
```
f <- file.choose()
walmart <- read.csv(f)


#########################################################
# Get groupby_session_id
age <- tapply(walmart$age, walmart$session, "max")
day <- tapply(walmart$dayofweek, walmart$session, "max")
hour <- tapply(walmart$hour, walmart$session, "max")
duration <- tapply(walmart$duration, walmart$session, "sum")
gender <- tapply(walmart$gender_id, walmart$session, "max")
income <- tapply(walmart$income, walmart$session, "max")
children <- tapply(walmart$children, walmart$session, "max")
ethnicity <- tapply(walmart$ethnicity, walmart$session, "max")
checkout <- tapply(walmart$checkout, walmart$session, "max")
click <- tapply(walmart$click, walmart$session, "sum")
education <- tapply(walmart$education, walmart$session, "max")
new_walmart <- data.frame(age, day, hour, duration, gender, income,
children, ethnicity, education, checkout, click)
walmart <- data.frame(check_out, hour, day, hour_day, duration, age,
gender, income, edu, children, ethnicity)
View(new_walmart)
write.csv(new_walmart, file = 'walmart_groupby_session')

#########################################################
new_walmart['ccrate'] <- 100*new_walmart$checkout/new_walmart$click
new_walmart['hourday'] <- 100*new_walmart$day+new_walmart$hour
attach(new_walmart)
## MALE
mean(new_walmart$ccrate[new_walmart$gender == 1])
mean(new_walmart$ccrate[new_walmart$gender == 2])

mean(new_walmart$duration[new_walmart$gender == 1])
mean(new_walmart$duration[new_walmart$gender == 2])

mean(new_walmart$click[new_walmart$gender == 1])
mean(new_walmart$click[new_walmart$gender == 2])

mean(new_walmart$duration[new_walmart$gender == 1])
mean(new_walmart$duration[new_walmart$gender == 2])

table(new_walmart$day, new_walmart$checkout)
check_day <- (table(new_walmart$day, new_walmart$checkout))
barplot(check_day[, 2])

library(ggplot2)
ggplot(data = new_walmart, aes(new_walmart$day,
fill=new_walmart$gender)) + geom_bar()
```

```
#########################################################################
ggplot(data = new_walmart, aes(x= new_walmart$hourday, y =
new_walmart$ccrate , color = (age))) + geom_point(shape = 1, size = 3,
alpha = 0.8, stroke = 2) + scale_colour_gradient(low = "white", high =
"dark red")+
  xlab ('Hourday') + ylab ('Click to Checkout Rate
(CCRate)')+ggtitle('CCRate shows higher traction in the
weekends')+labs(colour = 'age')+theme(legend.position="top")

## Middle age
middle_walmart <- new_walmart[new_walmart$age >= 24 & new_walmart$age
<= 40, ]
ggplot(data = middle_walmart, aes(x= middle_walmart$hourday, y =
middle_walmart$ccrate , color = (age))) + geom_point(shape = 1, size =
3, alpha = 0.8, stroke = 2) + scale_colour_gradient(low = "white",
high = "dark red")+
  xlab ('Hourday') + ylab ('Click to Checkout Rate
(CCRate)')+ggtitle('CCRate for shows higher traction on Wed and
weekends')+labs(colour = 'age')+theme(legend.position="top")

## Wednesday
wed_walmart <- new_walmart[new_walmart$day == 3, ]
wed_ccrate <- wed_walmart$ccrate
other_ccrate <- new_walmart$ccrate[new_walmart$day != 3]
t.test(wed_ccrate, other_ccrate)
#########################################################################
## CCR and age
CCR <- new_walmart$ccrate
age2 <- new_walmart$age^2
age <- new_walmart$age
quadratic.model <-lm(CCR ~ age+age2)
fitdistr(CCR, densfun = 'Poisson')
summary(quadratic.model)

plot(new_walmart$ccrate~new_walmart$income)
#########################################################################
buy_walmart <- new_walmart[new_walmart$ccrate != 0,]
nobuy_walmart <- new_walmart[new_walmart$ccrate == 0,]
plot(buy_walmart$ccrate~buy_walmart$income)
par(mfrow = c(1,2))
hist(buy_walmart$age, breaks = 15)
hist(nobuy_walmart$age, breaks = 15)
#########################################################################
plot.multi.dens <- function(s)
{
  junk.x = NULL
  junk.y = NULL
  for(i in 1:length(s)) {
    junk.x = c(junk.x, density(s[[i]])$x)
```

```
    junk.y = c(junk.y, density(s[[i]])$y)
  }
  xr <- range(junk.x)
  yr <- range(junk.y)
  plot(density(s[[1]]), xlim = xr, ylim = yr, main = "")
  for(i in 1:length(s)) {
    lines(density(s[[i]]), xlim = xr, ylim = yr, col = i)
  }
}

buy_age <- buy_walmart$age
nobuy_age <- nobuy_walmart$age

buy_rate <- buy_walmart$hourday
nobuy_rate <- nobuy_walmart$hourday
t.test(buy_age, nobuy_age)
par(mfrow = c(1,1))
plot.multi.dens( list(buy_age, nobuy_age))
plot.multi.dens( list(buy_rate, nobuy_rate))
```

Appendix D: Clustering – RapidMiner

RapidMiner Process and Results for Cluster Analysis



Result Overview | PerformanceVector (Performance) | Cluster Model (Clustering)

Text View | Folder View | Graph View | Centroid Table | Centroid Plot View | Annotations

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
DayType = Weekend	0.551	0.542	0.452	0.632	0.594
DayType = Weekday	0.449	0.458	0.548	0.368	0.406
TimeOfDay = Night	0.500	0.458	0.539	0.632	0.688
TimeOfDay = Afternoon	0.372	0.385	0.296	0.211	0.188
TimeOfDay = Morning	0.064	0.115	0.104	0.158	0.094
TimeOfDay = EarlyMorning	0.064	0.042	0.061	0	0.031
gender = 2	0.679	0.729	0.600	0.789	0.562
gender = 1	0.321	0.271	0.400	0.211	0.438
income = 84003	0.179	0.208	0.183	0.105	0.188
income = 84004	0.244	0.208	0.209	0.368	0.188
income = 84007	0.154	0.156	0.113	0.263	0.125
income = 84001	0.077	0.115	0.148	0.053	0.062
income = 84002	0.141	0.094	0.122	0	0.188
income = 84005	0.103	0.125	0.122	0.211	0.094
income = 84006	0.103	0.094	0.104	0	0.156
children = 1	0.756	0.760	0.783	0.737	0.562
children = 0	0.244	0.240	0.217	0.263	0.438
ethnicity = 10	0.795	0.854	0.878	0.737	0.781
ethnicity = 13	0.167	0.115	0.104	0.263	0.156
ethnicity = 11	0.013	0	0	0	0
ethnicity = 12	0.026	0.031	0.017	0	0.062
education = 84004	0.423	0.573	0.452	0.316	0.375
education = 84005	0.128	0.104	0.130	0.158	0.062
education = 84099	0.423	0.292	0.383	0.526	0.469
education = 84003	0.026	0.031	0.035	0	0.094
age	40.628	41.458	35.887	44.368	42.438
duration	1297.013	713.521	304.452	3323.895	2063.812
click	40.064	28.010	15.191	82.211	55.219

Appendix E: Results and queries from Python

Out[212]:

Logit Regression Results

Dep. Variable:	checkout	**No. Observations:**	4487
Model:	Logit	**Df Residuals:**	4472
Method:	MLE	**Df Model:**	14
Date:	Wed, 01 Mar 2017	**Pseudo R-squ.:**	0.1927
Time:	22:13:07	**Log-Likelihood:**	-684.57
converged:	True	**LL-Null:**	-847.99
		LLR p-value:	2.928e-61

	coef	**std err**	**z**	**P>|z|**	**[95.0% Conf. Int.]**
duration	4.182e-05	0.000	0.265	0.791	-0.000 0.000
age	0.0111	0.005	2.168	0.030	0.001 0.021
children	-0.0778	0.185	-0.421	0.674	-0.440 0.284
click	0.0628	0.006	10.856	0.000	0.051 0.074
weekend_weekdays	-0.8594	5.51e+06	-1.56e-07	1.000	-1.08e+07 1.08e+07
weekend_weekend	-1.0433	5.51e+06	-1.89e-07	1.000	-1.08e+07 1.08e+07
hours_afterwork	-0.9183	nan	nan	nan	nan nan
hours_workhours	-1.0300	nan	nan	nan	nan nan
gender_Female	-0.8748	nan	nan	nan	nan nan
gender_Male	-1.0377	nan	nan	nan	nan nan
income_high income	-0.7000	nan	nan	nan	nan nan
income_low income	-0.5306	nan	nan	nan	nan nan
income_mid income	-0.5809	nan	nan	nan	nan nan
edu_high income	-0.8061	nan	nan	nan	nan nan
edu_low edu	-0.9605	nan	nan	nan	nan nan

```
# coding: utf-8

# In[162]:

import pandas as pd
import bs4
from bs4 import BeautifulSoup
from scipy import stats
from sklearn import datasets, linear_model
from sklearn.linear_model import LogisticRegression
import statsmodels.api as sm


# In[188]:

data=pd.read_csv('walmart_groupby_session.csv')


# In[189]:

data.head()


# In[190]:

#selecting the relevent columns
df=data[["day","hour","duration","checkout","age","gender","income","e
ducation","children","click"]]


# In[191]:

#excluding the rows with unknown information
df= df[df["gender"]!=99]
df= df[df["income"]!=84099]
df= df[df["education"]!=84099]
df= df[df["children"]!=99]

#drop the rows with blank
df.dropna()


# In[192]:

#check the dataframe
df.head()
```

```
# In[195]:

def day(dayofweek):
    if dayofweek==6:
        return "weekend"
    elif dayofweek==7:
        return "weekend"
    else:
        return "weekdays"

df["weekend"] = df["day"].apply(day)


# In[196]:

def hour(hour):
    if hour>17:
        return "afterwork"
    elif hour<3:
        return "afterwork"
    else:
        return "workhours"

df["after5"] = df["hour"].apply(hour)


# In[197]:

def gender(gender):
    if gender==1:
        return "Male"
    if gender==2:
        return "Female"

df["Male"]=df["gender"].apply(gender)


# In[198]:

def income(income):
    if income==84001:
        return "low income"
    elif income==84002:
        return "low income"
    elif income==84003:
        return "low income"
    elif income==84004:
        return "mid income"
    elif income==84005:
```

```
        return "mid income"
    elif income==84006:
        return "high income"
    elif income==84007:
        return "high income"

df["income_level"] = df["income"].apply(income)


# In[199]:

def education(education):
    if education==84002:
        return "low edu"
    elif education==84003:
        return "low edu"
    elif education==84004:
        return "low edu"
    elif education==84005:
        return "high income"
    elif education==84007:
        return "high income"

df["edu_level"] = df["education"].apply(education)


# In[200]:

df.head()


# In[207]:

dummy_weekend = pd.get_dummies(df['weekend'], prefix='weekend')
dummy_after5 = pd.get_dummies(df['after5'], prefix='hours')
dummy_gender = pd.get_dummies(df['Male'], prefix='gender')
dummy_income = pd.get_dummies(df['income_level'], prefix='income')
dummy_edu = pd.get_dummies(df['edu_level'], prefix='edu')


# In[208]:

cols_to_keep = ['checkout','duration','age',"children","click"]
merge = df[cols_to_keep].join(dummy_weekend.ix[:,])
merge1=merge.join(dummy_after5.ix[:, ])
merge2=merge1.join(dummy_gender.ix[:, ])
merge3=merge2.join(dummy_income.ix[:, ])
merged=merge3.join(dummy_edu.ix[:, ])
```

```
# In[209]:

merged.head()


# In[211]:

train_cols = merged.columns[1:]
logit = sm.Logit(merged['checkout'], merged[train_cols])
result = logit.fit_regularized()


# In[212]:

result.summary()


# In[20]:
```

References

- http://www.businessinsider.com/meet-the-average-wal-mart-shopper-2014-9
- http://www.retaildive.com/news/who-is-the-new-wal-mart-customer/412078/
- http://fortune.com/2015/10/14/walmart-affluent-consumers/
- http://www.kantarretail.com/wp-content/uploads/2016/06/Kantar-Retail-Breakthrough-Insights-First-Half-2016.pdf
- https://www.ama.org/publications/eNewsletters/Marketing-News-Weekly/Pages/kantar-retail-shopper-data-walmart-kohls-target-kmart.aspx
- https://www.remarkety.com/online-shopping-trends-reveal-best-days-and-times-to-email-customers)
- www.walmart.com/about-us
- https://www.weblinc.com/blog/trends-when-do-people-shop-online/