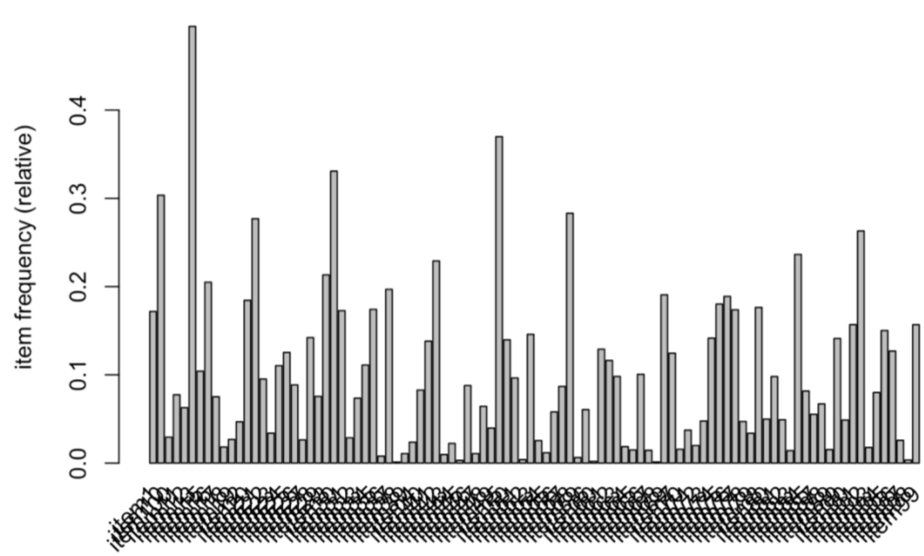
~~~~
### Create a frequent item plot, and a frequent item table.
```{r}
txn = read.transactions("Dataset/AssociationRules.csv")
#frequent item plot
itemFrequencyPlot(txn)
```
~~~~




~~~~
```{r}
#frequent item table
tab <- itemFrequency(txn)
head(tab)
```
~~~~

~~~~
  item1  item10 item100  item11  item12  item13 
 0.1718  0.3035  0.0294  0.0774  0.0628  0.4948 
~~~~

~~~~
#### a. Determine the most frequent item bought in the store.
```{r}
tail(sort(tab),1)
```
~~~~

~~~~
item13 
0.4948 
~~~~

~~~~
#### b. How many items were bought in the largest transaction?
```{r}
max(colSums(txn@data))
```
~~~~

~~~~
[1] 25
~~~~

Mine the Association rules with a minimum Support of 1% and a minimum Confidence of 0%.

```{r}
rules = apriori(txn, parameter=list(support=0.01, confidence=0.0))
```

R Console | data.frame 1 x 12 | data.frame 1 x 7

```
Apriori

Parameter specification:

Algorithmic control:

Absolute minimum support count: 100

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[98 item(s), 10000 transaction(s)] done [0.01s].
sorting and recoding items ... [89 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 done [0.02s].
writing ... [11524 rule(s)] done [0.01s].
creating S4 object  ... done [0.00s].
```

c. How many rules appear in the data?

Number of rules will appear in "Writing ...[... rule(s)]"

In this task, number of rules is 11524

d. How many rules are observed when the minimum confidence is 50%.

```{r}
rules = apriori(txn, parameter=list(support=0.01, confidence=0.5))
```

R Console | data.frame 1 x 12 | data.frame 1 x 7

```
Apriori

Parameter specification:

Algorithmic control:

Absolute minimum support count: 100

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[98 item(s), 10000 transaction(s)] done [0.01s].
sorting and recoding items ... [89 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 done [0.02s].
writing ... [1165 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

Number of rules is 1165

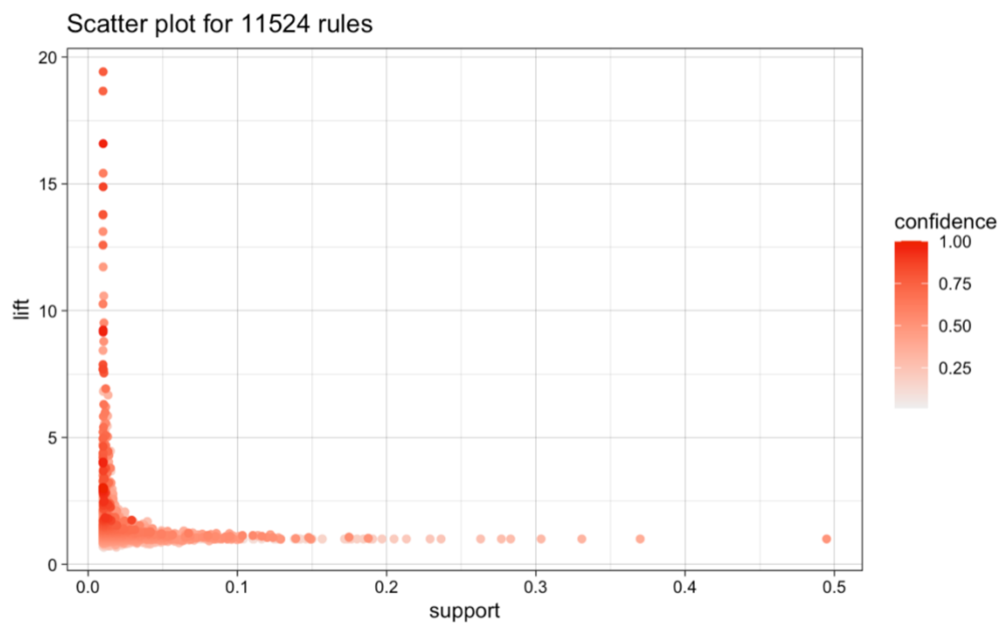
f. Identify the positioning of the interesting rules.

The interesting rules have high confidence and high lift, they would be located on the top left of the plot.

Compare support and lift.

g. Create a scatter plot measuring support vs. lift; record your observations.

```{r warning=FALSE}
plot(rules, measure = c("support", "lift"), shading = "confidence", jitter = 0)
```

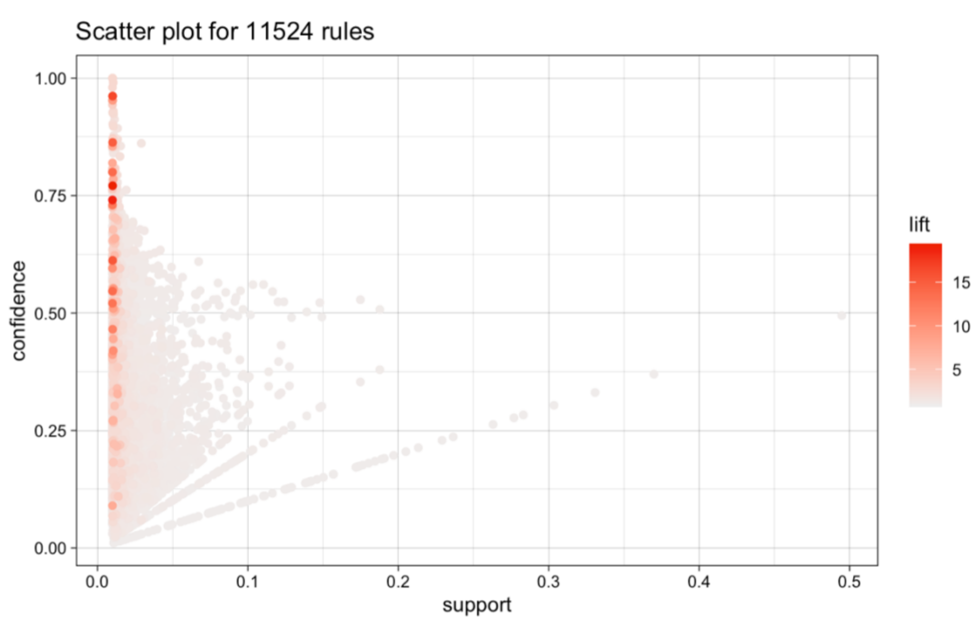


e. Explain how the specified confidence impacts the number of rules.

The specified confidence, say 50%, reduces the number of rules by only considering the transactions that have at least a pair of items at least 50% of the time.

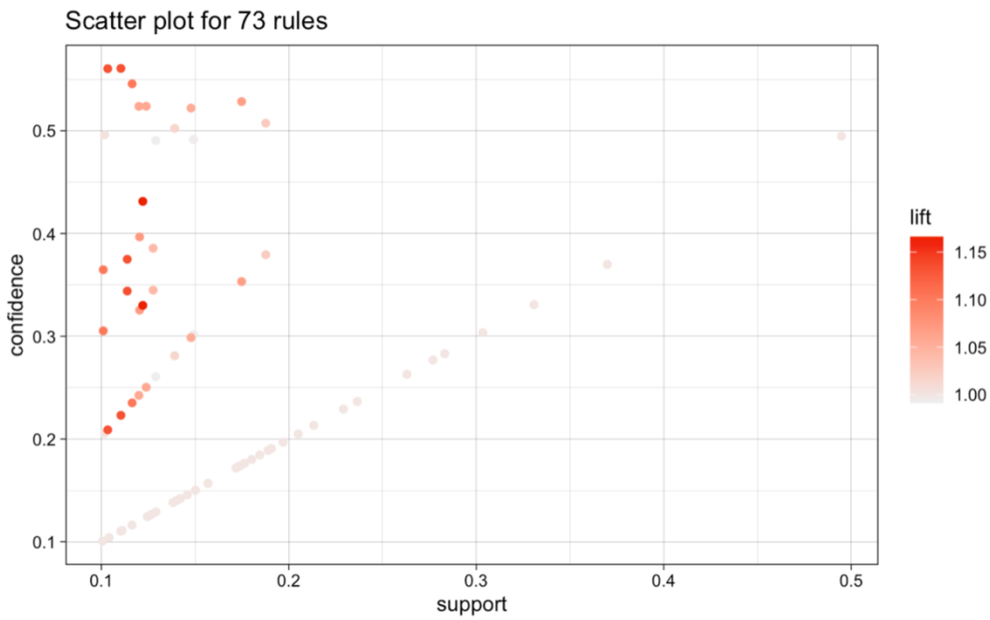
Create a scatter plot comparing the parameters support and confidence on the axis, and lift with shading.

```{r warning=FALSE}
rules <- apriori(txn,parameter =list(supp=0.01,conf =0.0),control = list(verbose = FALSE))
plot(rules, jitter = 0)
```



j. Using the interaction tool for a scatter plot, identify 3 rules that appear in at least 10% of the transactions by coincidence.

```{r warning=FALSE}
rules1 <- apriori(txn,parameter =list(supp=0.1,conf =0.0),control = list(verbose = FALSE))
plot(rules1,  jitter = 0)
```



Identify 3 rules that appear in at least 10% of the transactions by coincidence:

item37 -> item13

item20 -> item13

item3 -> item13

Identify the most interesting rules by extracting the rules in which the Confidence is >0.8. Observe the output of the data table for the most interesting rules.

```{r}
subrules <- subset(rules,rules@quality$confidence>0.8)
inspect(subrules)
```

Description: df [38 × 8]

	lhs		rhs	support	confidence	coverage	lift	count
	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<int>
[1]	{item55}	=>	{item34}	0.0100	0.8547009	0.0117	7.693077	100
[2]	{item83}	=>	{item13}	0.0119	0.8439716	0.0141	1.705682	119
[3]	{item23}	=>	{item13}	0.0292	0.8613569	0.0339	1.740818	292
[4]	{item10, item44}	=>	{item13}	0.0101	0.8487395	0.0119	1.715318	101
[5]	{item20, item23}	=>	{item13}	0.0114	0.9120000	0.0125	1.843169	114
[6]	{item23, item5}	=>	{item13}	0.0105	0.8400000	0.0125	1.697656	105
[7]	{item49, item56}	=>	{item15}	0.0101	0.9528302	0.0106	9.153028	101
[8]	{item15, item49}	=>	{item56}	0.0101	0.8632479	0.0117	14.883584	101
[9]	{item49, item56}	=>	{item84}	0.0100	0.9433962	0.0106	3.988990	100
[10]	{item49, item56}	=>	{item30}	0.0105	0.9905660	0.0106	2.994456	105

1-10 of 38 rows

Previous 1 2 3 4 Next

k. Sort the rules stating the highest lift first. Provide the 10 rules with the lowest lift. Do they appear to be coincidental (Use lift = 2 as baseline for coincidence)? Why or why not?

```{r}
a <- sort(subrules, by = "lift")
inspect(a)
#Provide the 10 rules with the lowest lift
inspect(tail(a,10))
```

data.frame 38 x 8 | data.frame 10 x 8

Description: df [38 × 8]

	lhs <chr>	<chr>	rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>	count <int>
[1]	{item15, item30, item49}	=>	{item56}	0.0101	0.9619048	0.0105	16.584565	101
[2]	{item15, item49}	=>	{item56}	0.0101	0.8632479	0.0117	14.883584	101
[3]	{item30, item49, item56}	=>	{item15}	0.0101	0.9619048	0.0105	9.240199	101
[4]	{item49, item56}	=>	{item15}	0.0101	0.9528302	0.0106	9.153028	101
[5]	{item30, item56, item77}	=>	{item15}	0.0100	0.8196721	0.0122	7.873892	100
[6]	{item55}	=>	{item34}	0.0100	0.8547009	0.0117	7.693077	100
[7]	{item30, item49, item84}	=>	{item77}	0.0101	0.8080000	0.0125	4.651698	101
[8]	{item30, item49, item56}	=>	{item84}	0.0100	0.9523810	0.0105	4.026981	100
[9]	{item15, item30, item49}	=>	{item84}	0.0100	0.9523810	0.0105	4.026981	100
[10]	{item49, item56}	=>	{item84}	0.0100	0.9433962	0.0106	3.988990	100

1-10 of 38 rows   Previous 1 2 3 4 Next

```
#Provide the 10 rules with the lowest lift
inspect(tail(a,10))
```

data.frame 38 x 8 | data.frame 10 x 8

Description: df [10 × 8]

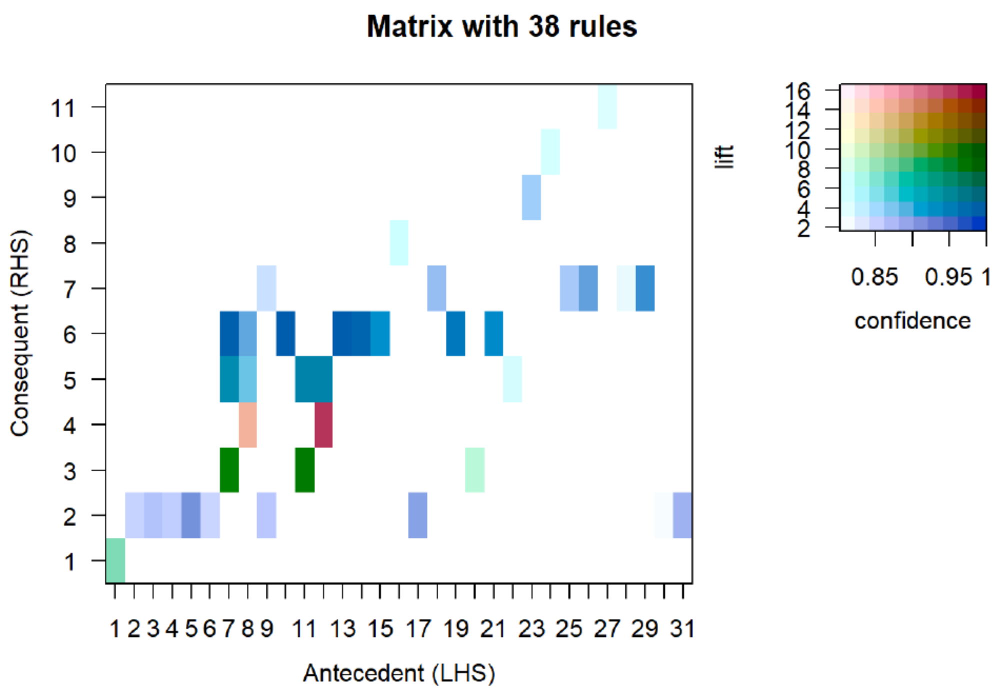
	lhs <chr>	<chr>	rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>	count <int>
[1]	{item16, item25, item77}	=>	{item5}	0.0104	0.8062016	0.0129	2.179512	104
[2]	{item20, item23}	=>	{item13}	0.0114	0.9120000	0.0125	1.843169	114
[3]	{item5, item82, item99}	=>	{item13}	0.0134	0.8933333	0.0150	1.805443	134
[4]	{item3, item84, item95}	=>	{item13}	0.0108	0.8780488	0.0123	1.774553	108
[5]	{item23}	=>	{item13}	0.0292	0.8613569	0.0339	1.740818	292
[6]	{item82, item99}	=>	{item13}	0.0154	0.8555556	0.0180	1.729094	154
[7]	{item10, item44}	=>	{item13}	0.0101	0.8487395	0.0119	1.715318	101
[8]	{item83}	=>	{item13}	0.0119	0.8439716	0.0141	1.705682	119
[9]	{item23, item5}	=>	{item13}	0.0105	0.8400000	0.0125	1.697656	105
[10]	{item30, item95, item96}	=>	{item13}	0.0118	0.8027211	0.0147	1.622314	118

1-10 of 10 rows

Create a Matrix-based visualization of two measures with colored squares. The two measures should compare confidence and lift (have recorded = FALSE). Note that 4 interesting rules stand out on the graph.

```{r}
plot(subrules, method="matrix",shading = c("lift","confidence"), control = list(reorder = FALSE))
```

```
## Itemsets in Antecedent (LHS)
##  [1] "{item55}"                "{item83}"                "{item23}"
##  [4] "{item10,item44}"         "{item20,item23}"         "{item23,item5}"
##  [7] "{item49,item56}"         "{item15,item49}"         "{item82,item99}"
## [10] "{item15,item49,item56}" "{item30,item49,item56}" "{item15,item30,item49}"
## [13] "{item49,item56,item84}" "{item15,item49,item84}" "{item49,item77,item84}"
## [16] "{item30,item49,item84}" "{item5,item82,item99}"  "{item13,item82,item99}"
## [19] "{item15,item56,item77}" "{item30,item56,item77}" "{item15,item56,item84}"
## [22] "{item15,item30,item56}" "{item22,item3,item41}"  "{item10,item22,item41}"
## [25] "{item25,item34,item77}" "{item16,item34,item77}" "{item20,item25,item41}"
## [28] "{item16,item25,item77}" "{item16,item61,item77}" "{item30,item95,item96}"
## [31] "{item3,item84,item95}"
## Itemsets in Consequent (RHS)
##  [1] "{item34}" "{item13}" "{item15}" "{item56}" "{item84}" "{item30}"
##  [7] "{item5}"  "{item77}" "{item10}" "{item3}"  "{item92}"
```



m. What can you infer about rules represented by a dark blue color?

Rules in a dark (deep) blue color suggest that we are likely to see these itemsets paired together by coincidence making them interesting but not important rules

Extract the three rules with the highest lift.

n. Record the Rules. Explain why these rules vary from the rules in Step 3.

```{r}
subrules2 <- head(sort(rules, by="lift"), 3)
inspect(subrules2)
```

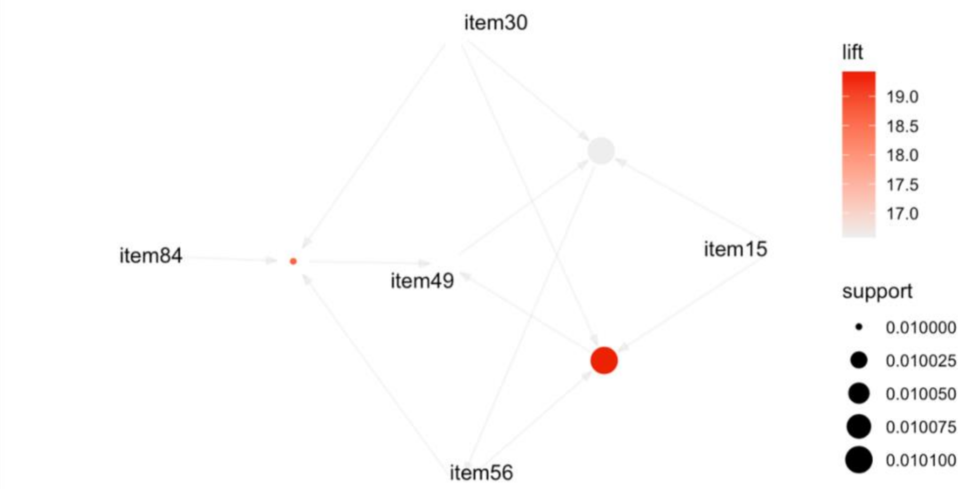
Description: df [3 × 8]

	lhs <chr>	<chr>	rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>	count <int>
[1]	{item15, item30, item56}	=>	{item49}	0.0101	0.7709924	0.0131	19.42046	101
[2]	{item30, item56, item84}	=>	{item49}	0.0100	0.7407407	0.0135	18.65846	100
[3]	{item15, item30, item49}	=>	{item56}	0.0101	0.9619048	0.0105	16.58456	101

3 rows

These rules vary from earlier because the associations between these items happen more than expected (high lift), but they do not occur more than 80% of the time.

o. Create a Graph-based visualization with items and rules as vertices
```{r}
plot(subrules2, method="graph")
```



p. Based on your observations, explain how you would expect association rules to relate to order (i.e. the number of items contained in the rule).

- Support and order have a strong inverse relationship

- These rules vary from earlier because the associations between these items happen more than expected, but they do not occur more than 80% of the time.

Create a training set from the first 8,000 transactions. Create a testing set from the last 2,000 transactions. Run the algorithm on each dataset. Compare the results.
```{r}
train_set = head(txn, 8000)
test_set = tail(txn, 2000)
rules_train = apriori(train_set, parameter = list(supp =0.01, conf = 0.8),control = list(verbose = FALSE))
inspect(rules_train)
rules_test = apriori(test_set, parameter = list(supp =0.01, conf = 0.8),control = list(verbose = FALSE))
inspect(rules_test)
```

data.frame
50 x 8

data.frame
80 x 8

Description: df [50 × 8]

	lhs <chr>	<chr>	rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>	count <int>
[1]	{item55}	=>	{item34}	0.010125	0.8526316	0.011875	7.698705	81
[2]	{item83}	=>	{item13}	0.012125	0.8508772	0.014250	1.727669	97
[3]	{item23}	=>	{item13}	0.028250	0.8432836	0.033500	1.712251	226
[4]	{item10, item44}	=>	{item13}	0.010000	0.8602151	0.011625	1.746630	80
[5]	{item20, item23}	=>	{item13}	0.010375	0.8924731	0.011625	1.812128	83
[6]	{item23, item5}	=>	{item13}	0.010125	0.8181818	0.012375	1.661283	81
[7]	{item49, item56}	=>	{item15}	0.010375	0.9540230	0.010875	9.251132	83
[8]	{item15, item49}	=>	{item56}	0.010375	0.8829787	0.011750	15.456958	83
[9]	{item49, item56}	=>	{item84}	0.010250	0.9425287	0.010875	4.034366	82
[10]	{item49, item56}	=>	{item30}	0.010750	0.9885057	0.010875	2.986422	86

1-10 of 50 rows Previous 1 2 3 4 5 Next

```
rules_test = apriori(test_set, parameter = list(supp =0.01, conf = 0.8),control = list(verbose = FALSE))
inspect(rules_test)
```

data.frame 50 x 8

data.frame 80 x 8

Description: df [80 × 8]

	lhs <chr>	<chr>	rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>	count <int>
[1]	{item83}	=>	{item13}	0.0110	0.8148148	0.0135	1.616696	22
[2]	{item23}	=>	{item13}	0.0330	0.9295775	0.0355	1.844400	66
[3]	{item10, item72}	=>	{item30}	0.0105	0.9545455	0.0110	2.892562	21
[4]	{item10, item44}	=>	{item5}	0.0105	0.8076923	0.0130	2.145265	21
[5]	{item10, item44}	=>	{item13}	0.0105	0.8076923	0.0130	1.602564	21
[6]	{item23, item95}	=>	{item77}	0.0110	0.9166667	0.0120	5.253104	22
[7]	{item23, item77}	=>	{item95}	0.0110	0.8148148	0.0135	5.758409	22
[8]	{item23, item95}	=>	{item20}	0.0105	0.8750000	0.0120	4.807692	21
[9]	{item23, item95}	=>	{item13}	0.0115	0.9583333	0.0120	1.901455	23
[10]	{item23, item77}	=>	{item20}	0.0110	0.8148148	0.0135	4.477004	22

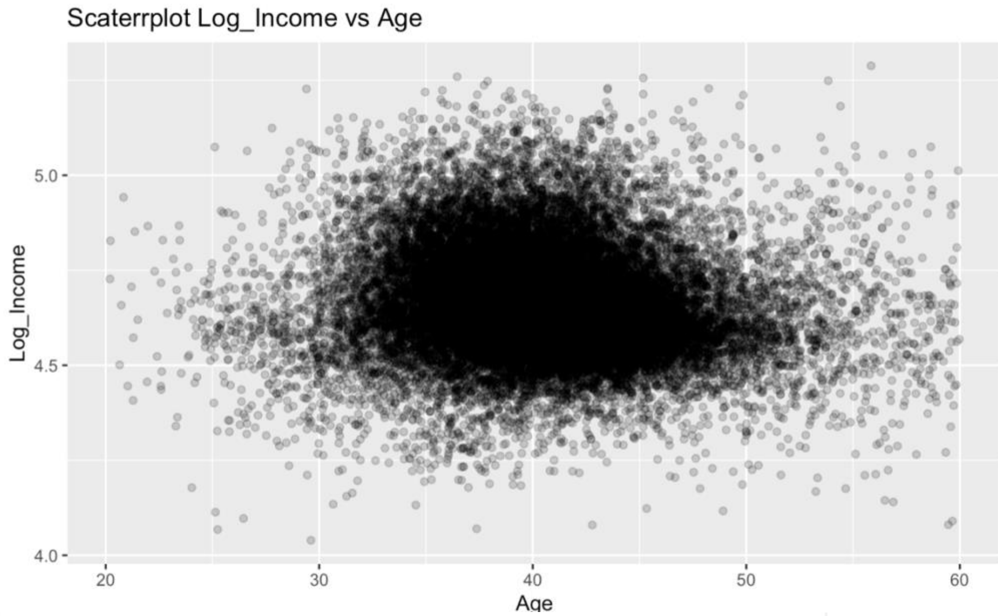
1-10 of 80 rows Previous 1 2 3 4 5 6 ... 8 Next

We see that majority of the rules that are present in the training set are also present in the hold out set with similar support and confidences.

```
############## Exercise 3.2 ##############
## Gather and Prepare Data
```{r}
data = read.csv("Dataset/zeta.csv")
#Remove all  meanhouseholdincome duplicates (only females records should be in the dataset)
data = subset(data, data$sex == 'F')
#Remove the columns zcta and sex
data = subset(data, select = -c(zcta, sex))
#Remove outliers
##8 < meaneducation < 18
data = subset(data, meaneducation <18 & meaneducation >8)
##10,000 < meanhouseholdincome < 200,000
data <- subset(data, meanhouseholdincome <200000 & meanhouseholdincome >10000)
##0 < meanemployment < 3
data <- subset(data, meanemployment <3 & meanemployment >0)
##20 < meanage < 60
data <- subset(data, meanage <60 & meanage >20)
#Create a variable called log_income = log10(meanhouseholdincome)
data$log_income <- log10(data$meanhouseholdincome)
#Rename the columns
names(data)[names(data)=="meanage"] <- "age"
names(data)[names(data)=="meaneducation"] <- "education"
names(data)[names(data)=="meanemployment"] <- "employment"
```
```

~~~
## Linear Regression Analysis

### a. Create a scatter plot showing the effect age has on log_income and paste it here.  Do you see any linear relationship between the two variables?
```{r message=FALSE}
library(ggplot2)
ggplot(data,aes(x= age, y=log_income)) +geom_point(alpha=0.2) +labs(x="Age",y="Log_Income",title="Scaterrplot Log_Income vs Age")
```
~~~




~~~
```{r}
#correlation
cor(data$age, data$log_income)
```
~~~

~~~
[1] -0.108803
~~~

From the scatter plot We can see, there seems to appear to be a very weak inverse linear relationship between the two variables.

In addition, the correlation cor= `r cor(data$age, data$log_income)` between the two variables is low, indicating that there is only a weak relationship between them.

b. Create a linear regression model between log_income and age. What is the interpretation of the t-value? What kind of t-value would indicate a significant coefficient?

```{r}
linearMod <- lm(log_income ~ age, data)
print(linearMod)
summary(linearMod)
```

```
Call:
lm(formula = log_income ~ age, data = data)

Coefficients:
(Intercept)          age
   4.787748    -0.003074


Call:
lm(formula = log_income ~ age, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.65733 -0.08296 -0.01620  0.07178  0.67202

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.7877484  0.0064657   740.5   <2e-16 ***
age         -0.0030739  0.0001584   -19.4   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1366 on 31427 degrees of freedom
Multiple R-squared:  0.01184,	Adjusted R-squared:  0.01181
F-statistic: 376.5 on 1 and 31427 DF,  p-value: < 2.2e-16
```

The t-value tests whether or not there is a statistically significant relationship between the dependent variable and the independent variable, that is whether or not the beta coefficient of the independent variable is significantly different from zero.

Mathematically, for a given beta coefficient (b), the t-test is computed as t = (b - 0)/SE(b), where SE(b) is the standard error of the coefficient b. The t-value measures the number of standard deviations that b is away from 0. The higher the t-value, the more significant independent variable.

In our exercise, both the t-values for the intercept and age are highly significant, which means that there is a significant association between age and income.

c. What is the interpretation of the R-squared value? What kind of R-squared value would indicate a good fit?

The R-squared value is a goodness of fit measure. The R-squared ranges from 0 to 1 (i.e.: a number near 0 represents a regression that does not explain the variance in the dependent variable well and a number close to 1 does explain the observed variance in the dependent variable).
`![Formula of R-squared](image1.png)`

$$R^2 = 1 - \frac{SSE}{SST}$$

A high value of R-squared is a good indication.
In our exercise, the R-squared we get is 0.01184. Or roughly 1.2% of the variance found in the dependent variable (income) can be explained by the independent variable (age).

d. What is the interpretation of the F-statistic? What kind of F-statistic indicates a strong linear regression model?

F-statistic is a good indicator of whether there is a relationship between our independent and the dependent variables. The further the F-statistic is from 1 the better it is. However, how much larger the F-statistic needs to be depends on both the number of data samples and the number of model parameters.

`![Formula of F-statistic](image.png)`

$$F = \frac{\frac{\sum (y_{pred} - y_{mean})^2}{p-1}}{\frac{\sum (y - y_{pred})^2}{n-p}}$$

Formula of F-statistic

The F-statistic is used to determine if the model is actually doing better than just guessing the mean value of y as the prediction (the "null model").

If the linear model is really just estimating the same as the null model, then the F-statistic should be about 1.

A F-statistic that is much larger than 1 indicates a strong linear regression model.

e. View a detailed summary of the previous model. What is the R-squared value? Does this suggest that the model is a good fit? Why?

```{r}
summary(linearMod)
```

```
Call:
lm(formula = log_income ~ age, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.65733 -0.08296 -0.01620  0.07178  0.67202 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.7877484  0.0064657   740.5   <2e-16 ***
age         -0.0030739  0.0001584   -19.4   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

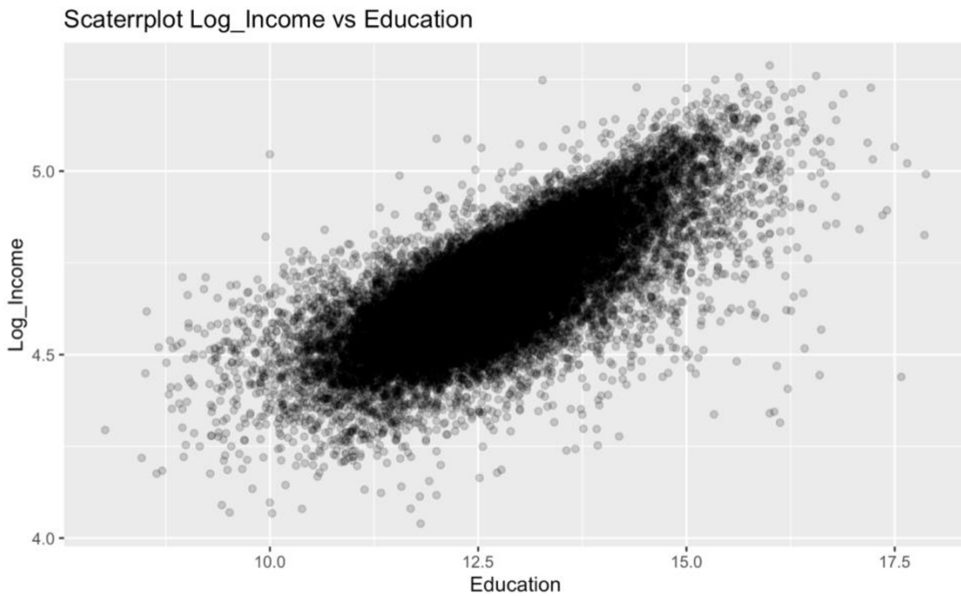
Residual standard error: 0.1366 on 31427 degrees of freedom
Multiple R-squared:  0.01184,	Adjusted R-squared:  0.01181 
F-statistic: 376.5 on 1 and 31427 DF,  p-value: < 2.2e-16
```

Multiple R-squared:0.01184

Adjusted R-squared: 0.01181

This R-squared value is very far from 1 and near to 0 suggests that the model is not a good fit.

````
### f. Create a scatter plot showing the effect education has on log_income.  Do you see any linear relationship between the two variables?
```{r message=FALSE}
ggplot(data,aes(x= education, y=log_income)) +geom_point(alpha=0.2) +labs(x="Education",y="Log_Income",title="Scaterrplot Log_Income vs Education")
```
````



This scatter plot seems to suggest that there is some sort of linear relationship between the two variables.The intercept seems to be positive.

````
### g. Analyze a detailed summary of a linear regression model between log_income and education.  What is the R-squared value?  Is the model a good fit? Is it better than the previous model?

```{r}
linearMod2 <- lm(log_income ~ education, data)
summary(linearMod2)
```
````

```
Call:
lm(formula = log_income ~ education, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.72721 -0.05349  0.00029  0.05796  0.64512 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.3896705  0.0067123   505.0   <2e-16 ***
education   0.1010797  0.0005311   190.3   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09369 on 31427 degrees of freedom
Multiple R-squared:  0.5354,	Adjusted R-squared:  0.5354 
F-statistic: 3.622e+04 on 1 and 31427 DF,  p-value: < 2.2e-16
```

Multiple R-squared: 0.5354

Adjusted R-squared: 0.5354

This R-squared value is much closer to 1 than our first model and suggests that the model is a decent fit. It is a better fit than the first model.

h. Analyze a detailed summary of a linear regression model between the dependent variable log_income, and the independent variables age, education, and employment. Is this model a good fit? Why? What conclusions can be made about the different independent variables?

```{r}
linearMod3 <- lm(log_income ~ education + age + employment, data)
summary(linearMod3)
```

```
Call:
lm(formula = log_income ~ education + age + employment, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.70315 -0.05023  0.00066  0.05213  0.64021 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.5123331  0.0076320  460.21   <2e-16 ***
education    0.0912653  0.0005980  152.61   <2e-16 ***
age         -0.0026030  0.0001109  -23.48   <2e-16 ***
employment   0.0663722  0.0019559   33.94   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09017 on 31425 degrees of freedom
Multiple R-squared:  0.5697,	Adjusted R-squared:  0.5697 
F-statistic: 1.387e+04 on 3 and 31425 DF,  p-value: < 2.2e-16
```

This model appears to be a good, but not perfect, fit because the R-squared value is somewhat close to 1.

The F-statistic is much larger than 1, and the p-value is extremely small, which indicates a strong model.

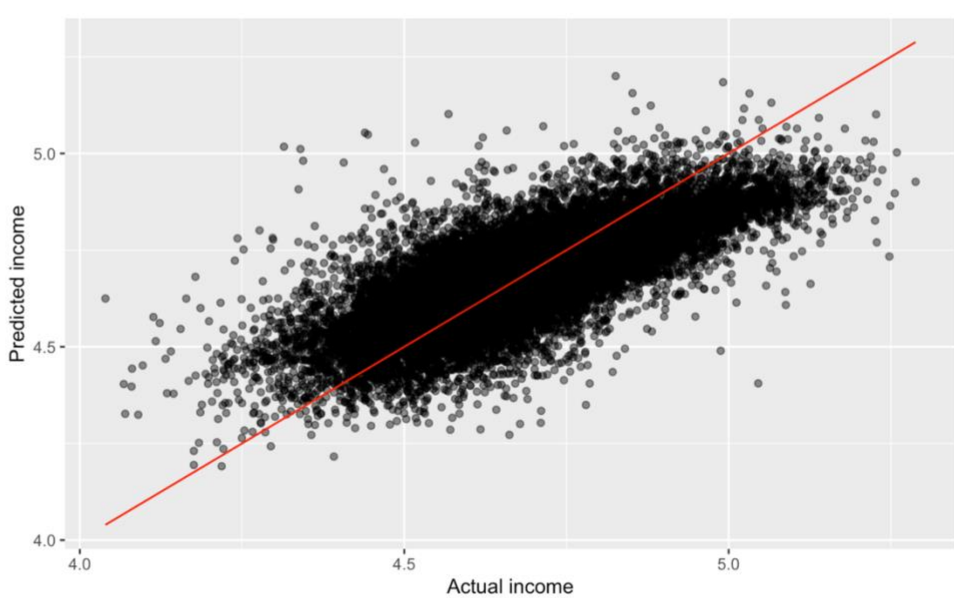
The independent variable age seems to have the weakest linear relationship because its coefficient and t-value are small.

i. Based on the coefficients of the multiple regression model, by what percentage would income increase/decrease for every unit of education completed, while all other independent variables remained constant?

For every unit of education completed, income increase 9.13%.

j. Create a graph that contains a y = x line and uses the multiple regression model to plot the predicted data points against the actual data points of the training set.

```{r}
ggplot() + geom_point(aes(x= data$log_income, y=fitted(linearMod3)), alpha=0.5) + geom_line(aes(x=data$log_income, y=data$log_income), col = 'red') +labs(x="Actual income", y="Predicted income")
```



k. How well does the model predict across the various income ranges?

In the graph, for lower incomes our model seems to over predict the income.

For higher incomes, our model seems to slightly under predict the income.

This graph indicates that our model provides reliable predictions around the median income range.