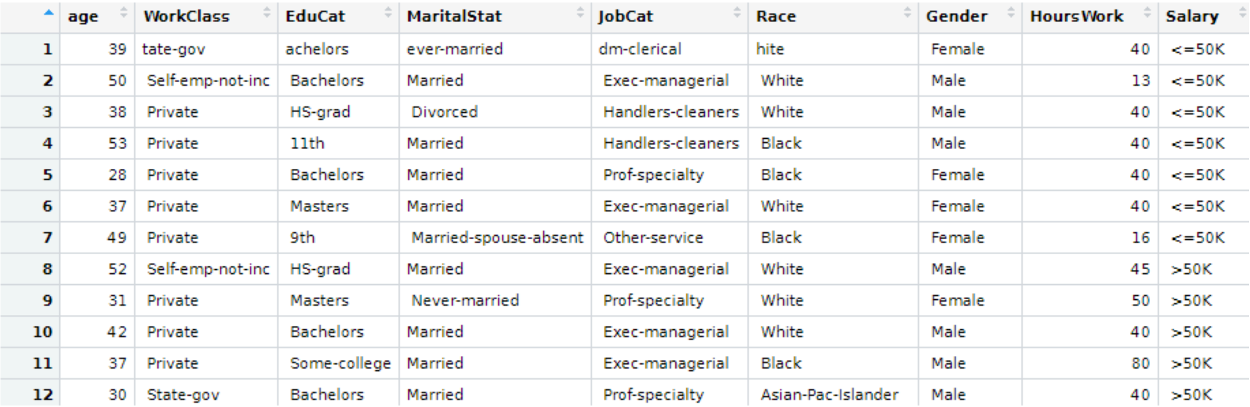
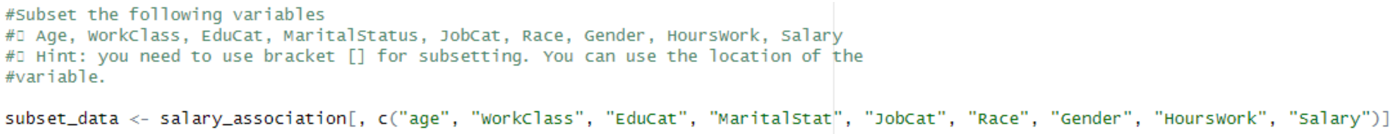
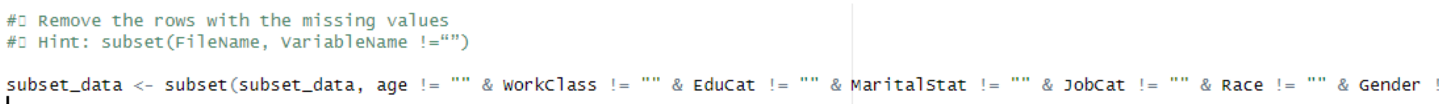
Advanced Association Rules



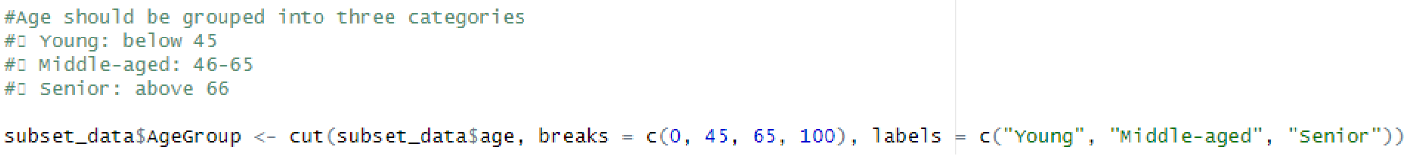
	age	WorkClass	EduCat	MaritalStat	JobCat	Race	Gender	HoursWork	Salary
1	39	tate-gov	achelors	ever-married	dm-clerical	hite	Female	40	<=50K
2	50	Self-emp-not-inc	Bachelors	Married	Exec-managerial	White	Male	13	<=50K
3	38	Private	HS-grad	Divorced	Handlers-cleaners	White	Male	40	<=50K
4	53	Private	11th	Married	Handlers-cleaners	Black	Male	40	<=50K
5	28	Private	Bachelors	Married	Prof-specialty	Black	Female	40	<=50K
6	37	Private	Masters	Married	Exec-managerial	White	Female	40	<=50K
7	49	Private	9th	Married-spouse-absent	Other-service	Black	Female	16	<=50K
8	52	Self-emp-not-inc	HS-grad	Married	Exec-managerial	White	Male	45	>50K
9	31	Private	Masters	Never-married	Prof-specialty	White	Female	50	>50K
10	42	Private	Bachelors	Married	Exec-managerial	White	Male	40	>50K
11	37	Private	Some-college	Married	Exec-managerial	Black	Male	80	>50K
12	30	State-gov	Bachelors	Married	Prof-specialty	Asian-Pac-Islander	Male	40	>50K



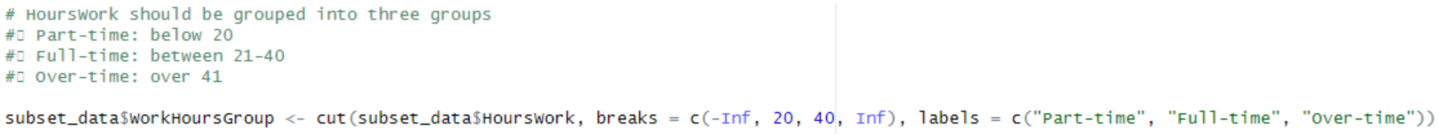
```
#Subset the following variables
#□ Age, WorkClass, EduCat, MaritalStatus, JobCat, Race, Gender, HoursWork, Salary
#□ Hint: you need to use bracket [] for subsetting. You can use the location of the
#variable.

subset_data <- salary_association[, c("age", "WorkClass", "EduCat", "MaritalStat", "JobCat", "Race", "Gender", "HoursWork", "Salary")]
```



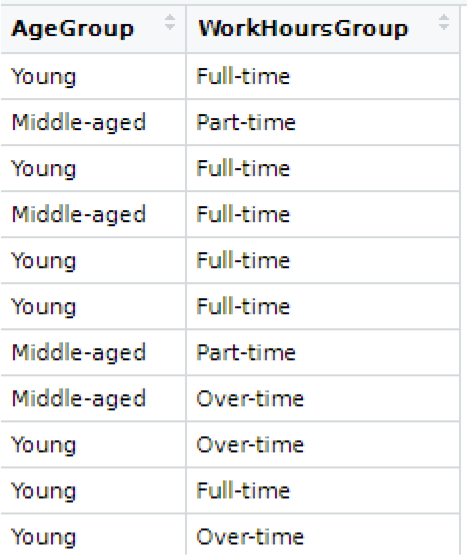
```
#□ Remove the rows with the missing values
#□ Hint: subset(FileName, VariableName !=“”)

subset_data <- subset(subset_data, age != "" & WorkClass != "" & EduCat != "" & MaritalStat != "" & JobCat != "" & Race != "" & Gender !
```



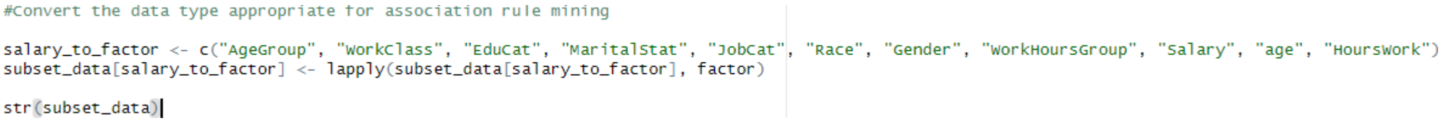
```
#Age should be grouped into three categories
#□ Young: below 45
#□ Middle-aged: 46-65
#□ Senior: above 66

subset_data$AgeGroup <- cut(subset_data$age, breaks = c(0, 45, 65, 100), labels = c("Young", "Middle-aged", "Senior"))
```



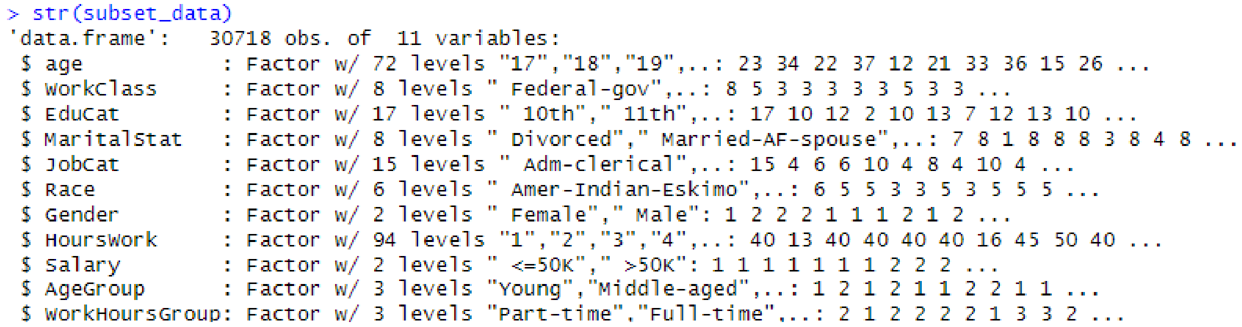
```
# HoursWork should be grouped into three groups
#□ Part-time: below 20
#□ Full-time: between 21-40
#□ Over-time: over 41

subset_data$WorkHoursGroup <- cut(subset_data$HoursWork, breaks = c(-Inf, 20, 40, Inf), labels = c("Part-time", "Full-time", "Over-time"))
```



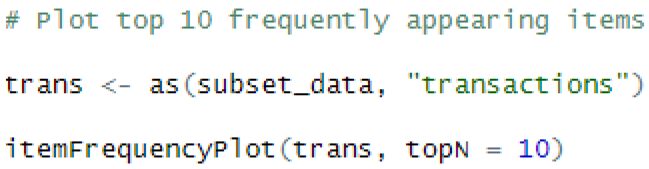
AgeGroup	WorkHoursGroup
Young	Full-time
Middle-aged	Part-time
Young	Full-time
Middle-aged	Full-time
Young	Full-time
Young	Full-time
Middle-aged	Part-time
Middle-aged	Over-time
Young	Over-time
Young	Full-time
Young	Over-time



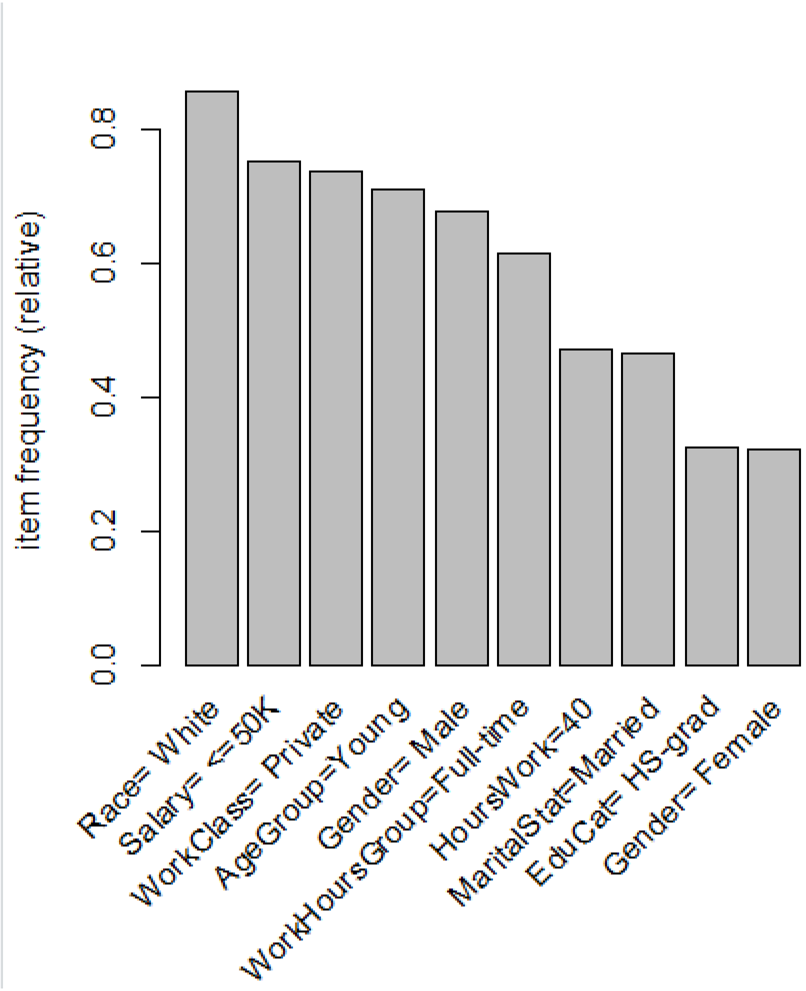
```
#Convert the data type appropriate for association rule mining

salary_to_factor <- c("AgeGroup", "WorkClass", "EduCat", "MaritalStat", "JobCat", "Race", "Gender", "WorkHoursGroup", "Salary", "age", "HoursWork")
subset_data[salary_to_factor] <- lapply(subset_data[salary_to_factor], factor)

str(subset_data)
```

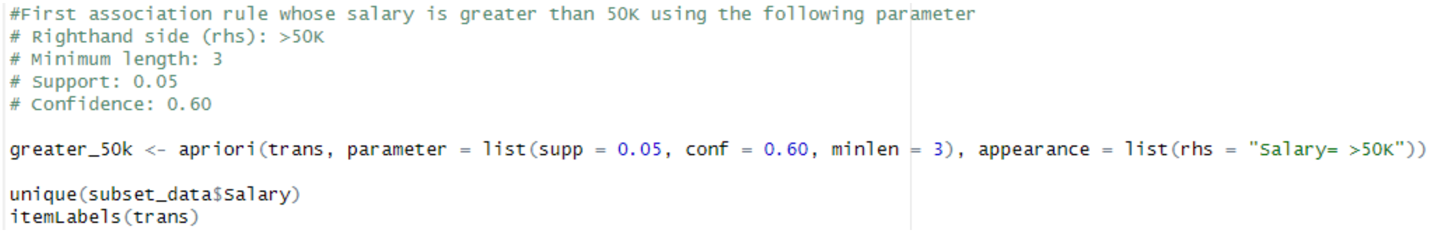


```
> str(subset_data)
'data.frame':	30718 obs. of  11 variables:
 $ age           : Factor w/ 72 levels "17","18","19",..: 23 34 22 37 12 21 33 36 15 26 ...
 $ WorkClass     : Factor w/ 8 levels " Federal-gov",..: 8 5 3 3 3 3 3 5 3 3 ...
 $ EduCat        : Factor w/ 17 levels " 10th"," 11th",..: 17 10 12 2 10 13 7 12 13 10 ...
 $ MaritalStat   : Factor w/ 8 levels " Divorced"," Married-AF-spouse",..: 7 8 1 8 8 8 3 8 4 8 ...
 $ JobCat        : Factor w/ 15 levels " Adm-clerical",..: 15 4 6 6 10 4 8 4 10 4 ...
 $ Race          : Factor w/ 6 levels " Amer-Indian-Eskimo",..: 6 5 5 3 3 5 3 5 5 5 ...
 $ Gender        : Factor w/ 2 levels " Female"," Male": 1 2 2 2 1 1 1 2 1 2 ...
 $ HoursWork     : Factor w/ 94 levels "1","2","3","4",..: 40 13 40 40 40 40 16 45 50 40 ...
 $ Salary        : Factor w/ 2 levels " <=50K"," >50K": 1 1 1 1 1 1 1 2 2 2 ...
 $ AgeGroup      : Factor w/ 3 levels "Young","Middle-aged",..: 1 2 1 2 1 1 2 2 1 1 ...
 $ WorkHoursGroup: Factor w/ 3 levels "Part-time","Full-time",..: 2 1 2 2 2 2 1 3 3 2 ...
```

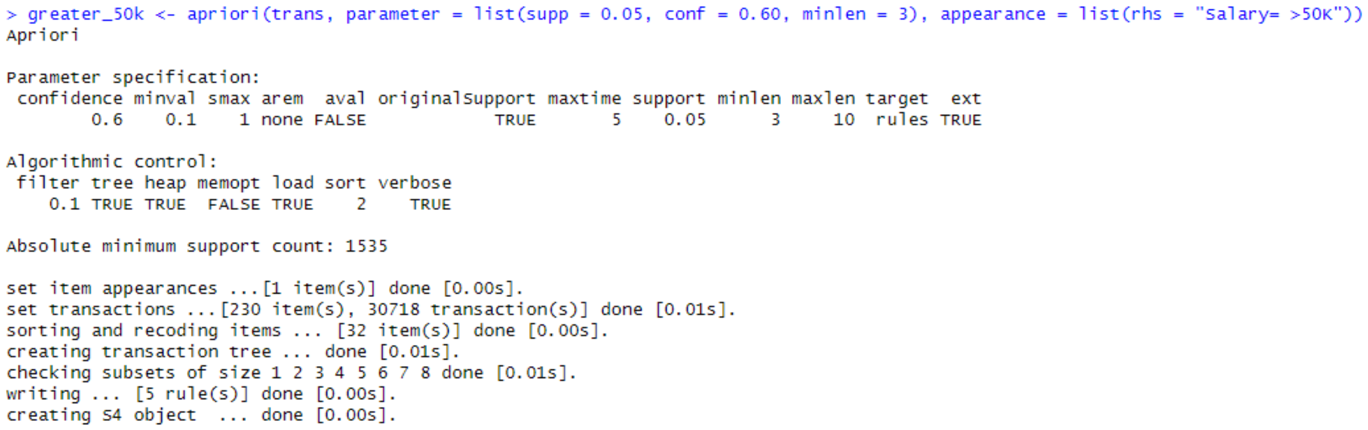


```
# Plot top 10 frequently appearing items

trans <- as(subset_data, "transactions")

itemFrequencyPlot(trans, topN = 10)
```

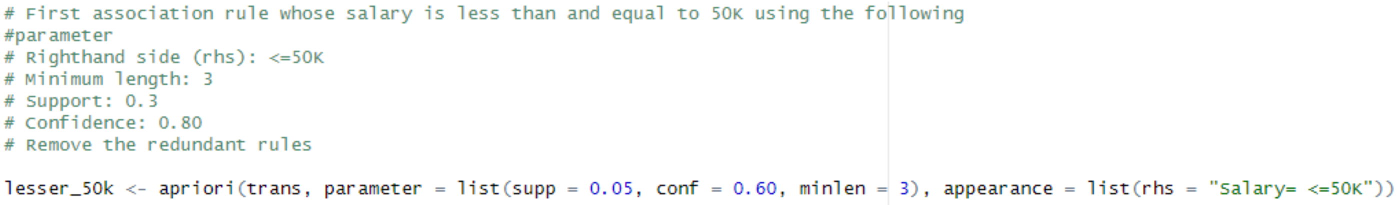




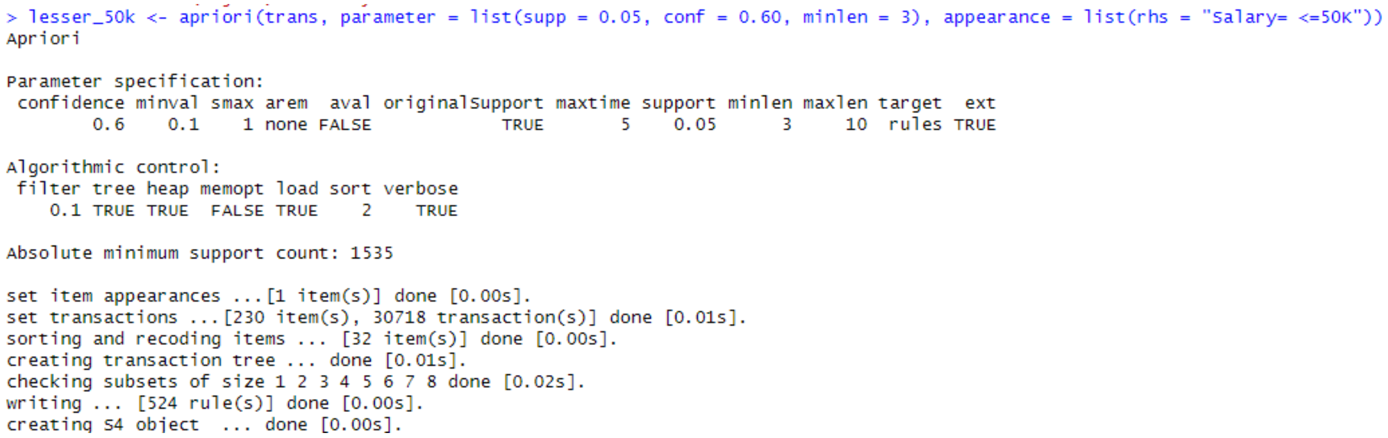
```
#First association rule whose salary is greater than 50K using the following parameter
# Righthand side (rhs): >50K
# Minimum length: 3
# Support: 0.05
# Confidence: 0.60

greater_50k <- apriori(trans, parameter = list(supp = 0.05, conf = 0.60, minlen = 3), appearance = list(rhs = "Salary= >50K"))

unique(subset_data$Salary)
itemLabels(trans)
```



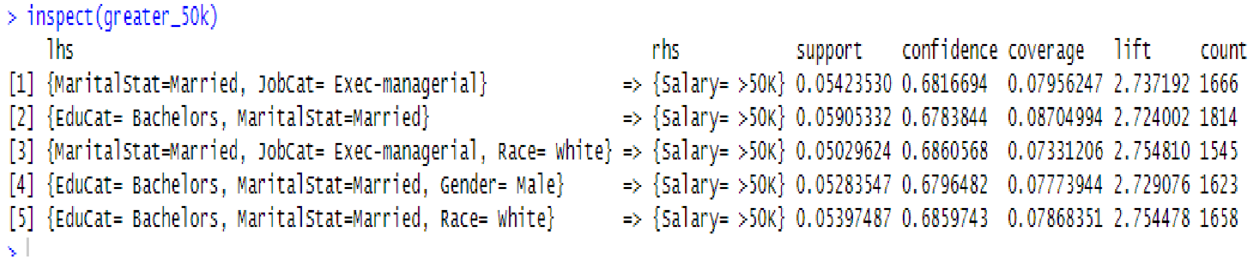
```
> greater_50k <- apriori(trans, parameter = list(supp = 0.05, conf = 0.60, minlen = 3), appearance = list(rhs = "Salary= >50K"))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target  ext
        0.6    0.1    1 none FALSE            TRUE       5    0.05      3     10  rules TRUE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 1535

set item appearances ...[1 item(s)] done [0.00s].
set transactions ...[230 item(s), 30718 transaction(s)] done [0.01s].
sorting and recoding items ... [32 item(s)] done [0.00s].
creating transaction tree ... done [0.01s].
checking subsets of size 1 2 3 4 5 6 7 8 done [0.01s].
writing ... [5 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```



```
# First association rule whose salary is less than and equal to 50K using the following
#parameter
# Righthand side (rhs): <=50K
# Minimum length: 3
# Support: 0.3
# Confidence: 0.80
# Remove the redundant rules

lesser_50k <- apriori(trans, parameter = list(supp = 0.05, conf = 0.60, minlen = 3), appearance = list(rhs = "Salary= <=50K"))
```



```
> lesser_50k <- apriori(trans, parameter = list(supp = 0.05, conf = 0.60, minlen = 3), appearance = list(rhs = "Salary= <=50K"))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target  ext
        0.6    0.1    1 none FALSE            TRUE       5    0.05      3     10  rules TRUE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 1535

set item appearances ...[1 item(s)] done [0.00s].
set transactions ...[230 item(s), 30718 transaction(s)] done [0.01s].
sorting and recoding items ... [32 item(s)] done [0.00s].
creating transaction tree ... done [0.01s].
checking subsets of size 1 2 3 4 5 6 7 8 done [0.02s].
writing ... [524 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```



```
> inspect(greater_50k)
    lhs                                                        rhs             support    confidence coverage   lift     count
[1] {MaritalStat=Married, JobCat= Exec-managerial}            => {Salary= >50K} 0.05423530 0.6816694  0.07956247 2.737192 1666
[2] {EduCat= Bachelors, MaritalStat=Married}                  => {Salary= >50K} 0.05905332 0.6783844  0.08704994 2.724002 1814
[3] {MaritalStat=Married, JobCat= Exec-managerial, Race= White} => {Salary= >50K} 0.05029624 0.6860568  0.07331206 2.754810 1545
[4] {EduCat= Bachelors, MaritalStat=Married, Gender= Male}     => {Salary= >50K} 0.05283547 0.6796482  0.07773944 2.729076 1623
[5] {EduCat= Bachelors, MaritalStat=Married, Race= White}      => {Salary= >50K} 0.05397487 0.6859743  0.07868351 2.754478 1658
>
```

Based on the results from the Greater than 50k Salary individuals. There were a total of 5 different metrics that on average earned more than 50k. With the highest degree of confidence being the individuals who are married and are executive/managerial staff. With a confidence of 68% and a Lift of 2.74 meaning they are 2.74 times more likely to earn more than 50k. Another interesting metric that can be seen is that most Educated White males earn more than 50k. With 3 out of the 5 metrics being educated with at least a Bachelor's Degree. The other metrics with there being a little over 524 different cases all say salaries less than 50k.