DATA SCIENCE WITH R

Class 4 – Data Manipulation in R

Topic 1

★ Manipulating Data Using Base R ★

INDEX

Manipulating data using base R

Using dplyr to manipulate data

Working with date objects

Merging tables

Missing value treatment

Using reshape2() to transpose data

Manipulating Character Strings

Using sqldf

Data Manipulation: Base R

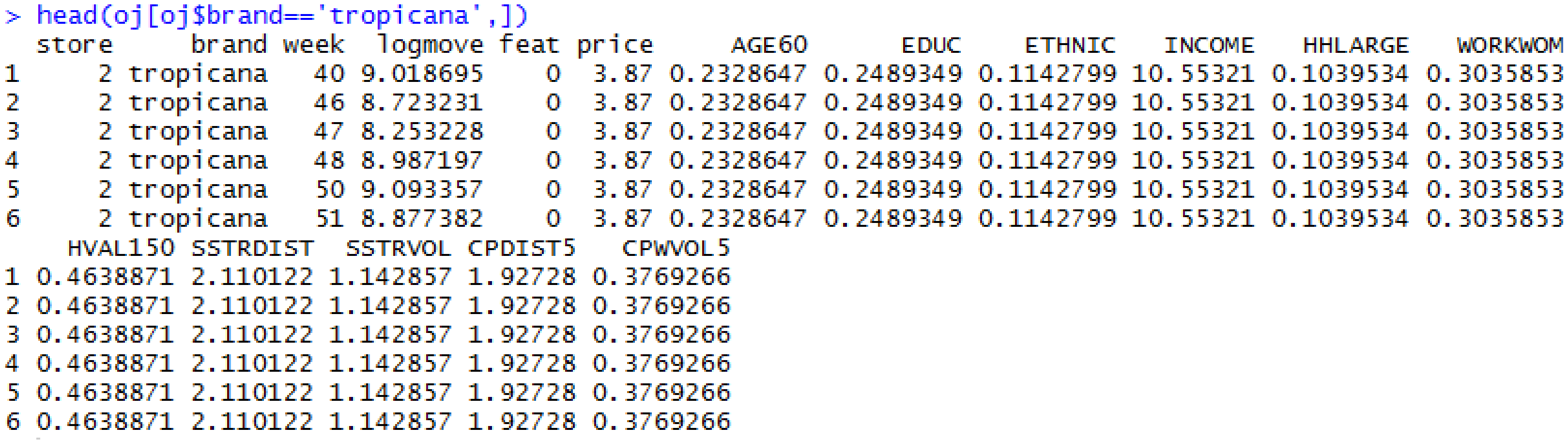
Data Manipulation: Base R

- Sub-setting data
- Selecting specified columns
- Adding new columns
- Reordering data (Ascending/Descending order)
- Group wise operations
- Producing contingency tables

Sub-setting data

Manipulating data: Base R (Sub setting)

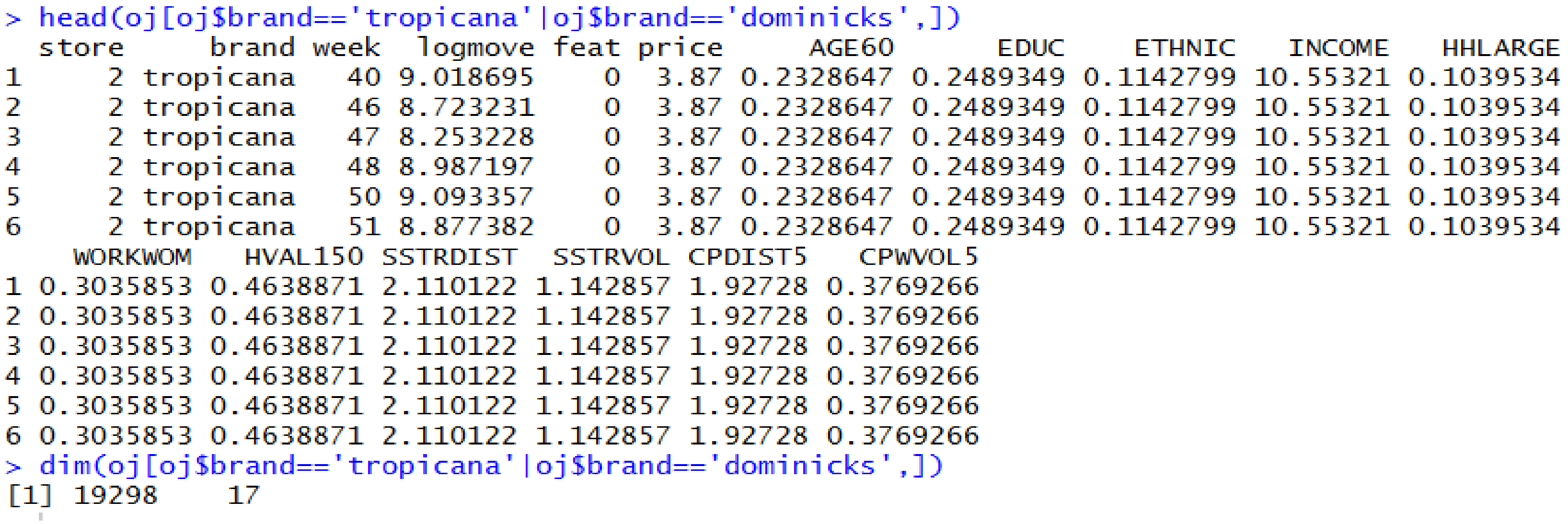
- Sub setting: Selecting a sub set of rows across all columns



```
> head(oj[oj$brand=='tropicana',])
  store     brand week  logmove feat price     AGE60      EDUC    ETHNIC   INCOME   HHLARGE   WORKWOM
1     2 tropicana   40 9.018695    0  3.87 0.2328647 0.2489349 0.1142799 10.55321 0.1039534 0.3035853
2     2 tropicana   46 8.723231    0  3.87 0.2328647 0.2489349 0.1142799 10.55321 0.1039534 0.3035853
3     2 tropicana   47 8.253228    0  3.87 0.2328647 0.2489349 0.1142799 10.55321 0.1039534 0.3035853
4     2 tropicana   48 8.987197    0  3.87 0.2328647 0.2489349 0.1142799 10.55321 0.1039534 0.3035853
5     2 tropicana   50 9.093357    0  3.87 0.2328647 0.2489349 0.1142799 10.55321 0.1039534 0.3035853
6     2 tropicana   51 8.877382    0  3.87 0.2328647 0.2489349 0.1142799 10.55321 0.1039534 0.3035853
    HVAL150 SSTRDIST  SSTRVOL CPDIST5   CPWVOL5
1 0.4638871 2.110122 1.142857 1.92728 0.3769266
2 0.4638871 2.110122 1.142857 1.92728 0.3769266
3 0.4638871 2.110122 1.142857 1.92728 0.3769266
4 0.4638871 2.110122 1.142857 1.92728 0.3769266
5 0.4638871 2.110122 1.142857 1.92728 0.3769266
6 0.4638871 2.110122 1.142857 1.92728 0.3769266
```

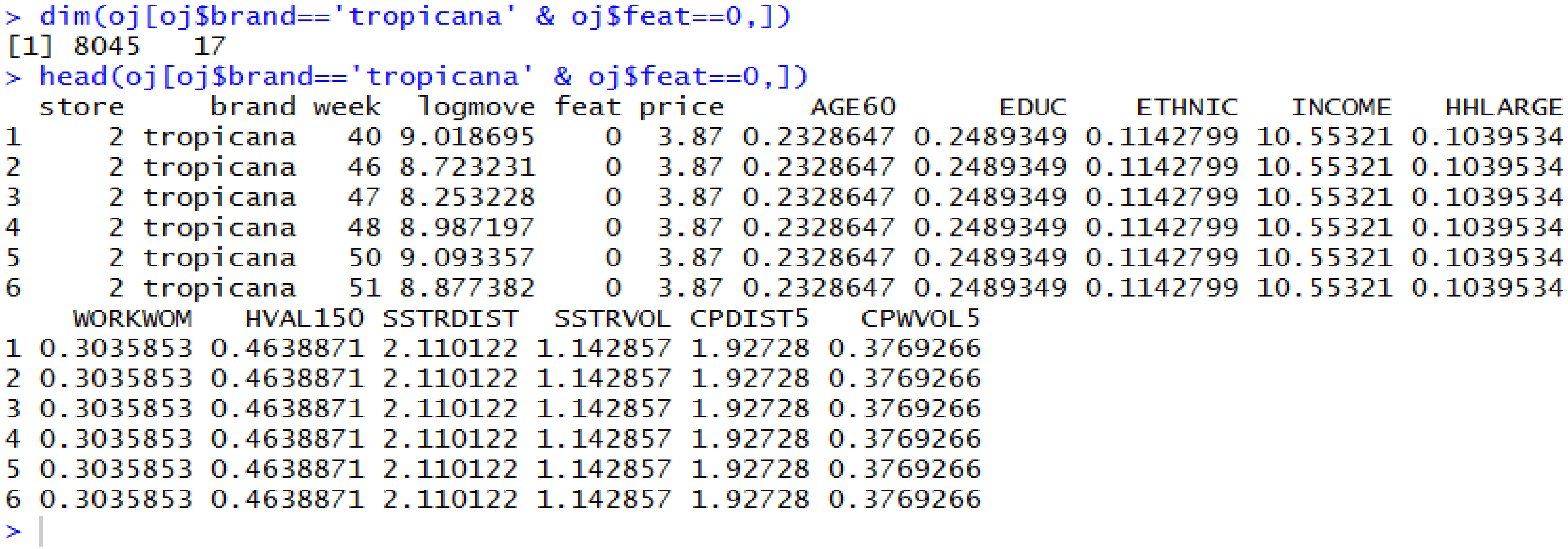
Manipulating data: Base R (Sub setting)

- Can use multiple conditions, | (or), & (and) operator



```
> head(oj[oj$brand=='tropicana'|oj$brand=='dominicks',])
  store     brand week  logmove feat price     AGE60      EDUC    ETHNIC   INCOME    HHLARGE
1     2 tropicana   40 9.018695    0  3.87 0.2328647 0.2489349 0.1142799 10.55321 0.1039534
2     2 tropicana   46 8.723231    0  3.87 0.2328647 0.2489349 0.1142799 10.55321 0.1039534
3     2 tropicana   47 8.253228    0  3.87 0.2328647 0.2489349 0.1142799 10.55321 0.1039534
4     2 tropicana   48 8.987197    0  3.87 0.2328647 0.2489349 0.1142799 10.55321 0.1039534
5     2 tropicana   50 9.093357    0  3.87 0.2328647 0.2489349 0.1142799 10.55321 0.1039534
6     2 tropicana   51 8.877382    0  3.87 0.2328647 0.2489349 0.1142799 10.55321 0.1039534
    WORKWOM   HVAL150 SSTRDIST  SSTRVOL CPDIST5   CPWVOL5
1 0.3035853 0.4638871 2.110122 1.142857 1.92728 0.3769266
2 0.3035853 0.4638871 2.110122 1.142857 1.92728 0.3769266
3 0.3035853 0.4638871 2.110122 1.142857 1.92728 0.3769266
4 0.3035853 0.4638871 2.110122 1.142857 1.92728 0.3769266
5 0.3035853 0.4638871 2.110122 1.142857 1.92728 0.3769266
6 0.3035853 0.4638871 2.110122 1.142857 1.92728 0.3769266
> dim(oj[oj$brand=='tropicana'|oj$brand=='dominicks',])
[1] 19298    17
```

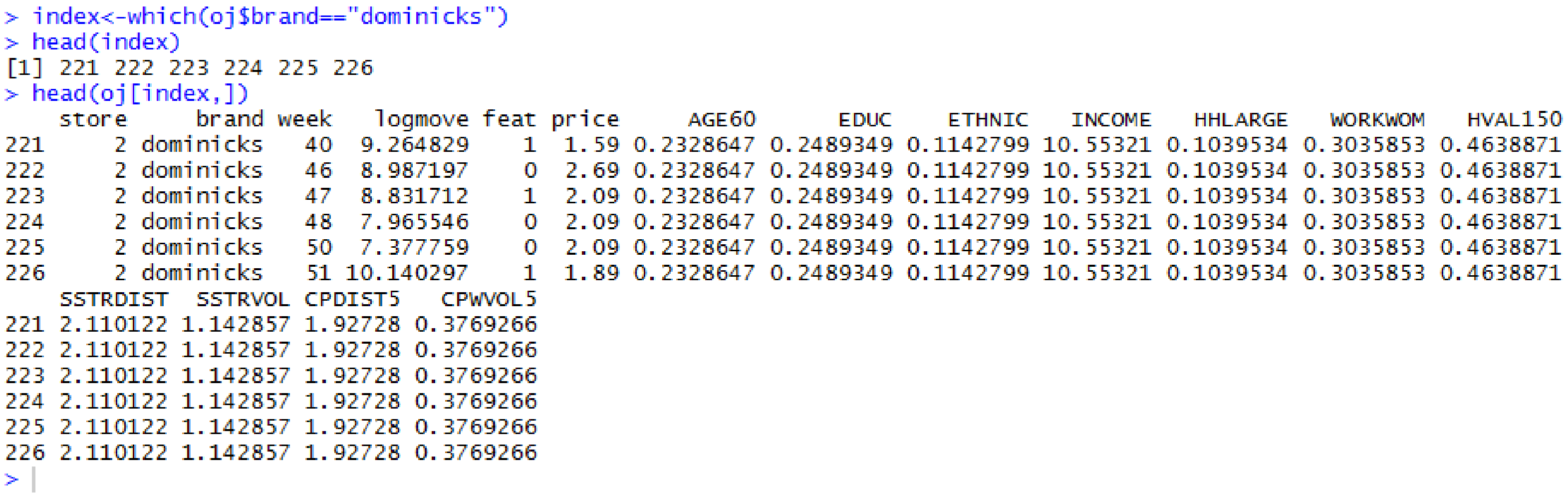
Manipulating data: Base R (Sub setting)



```
> dim(oj[oj$brand=='tropicana' & oj$feat==0,])
[1] 8045   17
> head(oj[oj$brand=='tropicana' & oj$feat==0,])
  store     brand week  logmove feat price     AGE60      EDUC    ETHNIC   INCOME    HHLARGE
1     2 tropicana   40 9.018695    0  3.87 0.2328647 0.2489349 0.1142799 10.55321 0.1039534
2     2 tropicana   46 8.723231    0  3.87 0.2328647 0.2489349 0.1142799 10.55321 0.1039534
3     2 tropicana   47 8.253228    0  3.87 0.2328647 0.2489349 0.1142799 10.55321 0.1039534
4     2 tropicana   48 8.987197    0  3.87 0.2328647 0.2489349 0.1142799 10.55321 0.1039534
5     2 tropicana   50 9.093357    0  3.87 0.2328647 0.2489349 0.1142799 10.55321 0.1039534
6     2 tropicana   51 8.877382    0  3.87 0.2328647 0.2489349 0.1142799 10.55321 0.1039534
    WORKWOM   HVAL150 SSTRDIST  SSTRVOL CPDIST5    CPWVOL5
1 0.3035853 0.4638871 2.110122 1.142857 1.92728 0.3769266
2 0.3035853 0.4638871 2.110122 1.142857 1.92728 0.3769266
3 0.3035853 0.4638871 2.110122 1.142857 1.92728 0.3769266
4 0.3035853 0.4638871 2.110122 1.142857 1.92728 0.3769266
5 0.3035853 0.4638871 2.110122 1.142857 1.92728 0.3769266
6 0.3035853 0.4638871 2.110122 1.142857 1.92728 0.3769266
> 
```

Manipulating data: Base R (Sub setting)

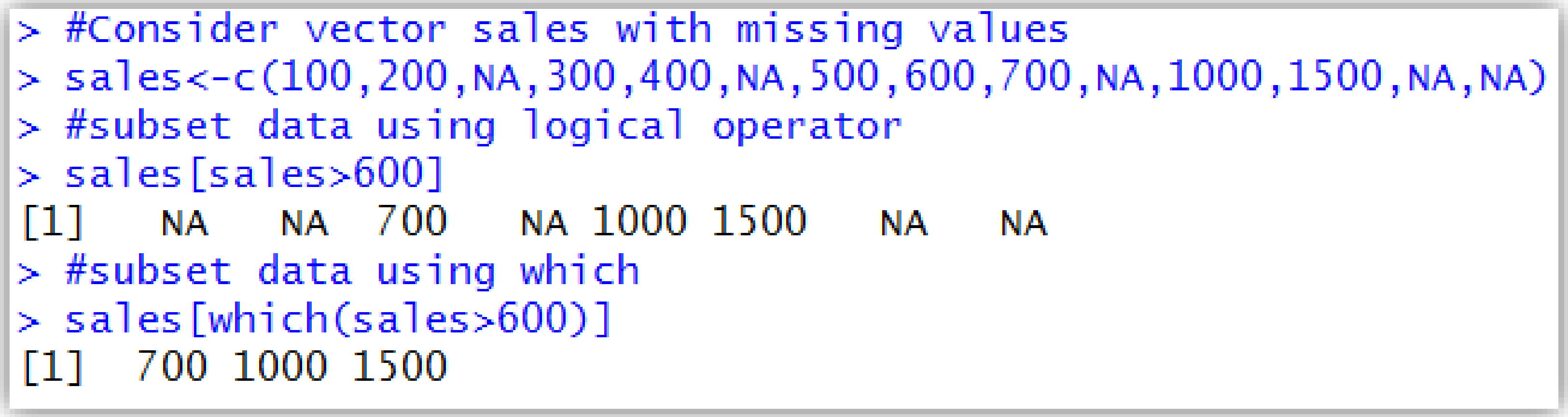
- So, far logical sub-setting is discussed.
- Use which() operator to get the index for specific rows



```
> index<-which(oj$brand=="dominicks")
> head(index)
[1] 221 222 223 224 225 226
> head(oj[index,])
    store     brand week   logmove feat price     AGE60      EDUC    ETHNIC   INCOME   HHLARGE   WORKWOM   HVAL150
221     2 dominicks   40  9.264829    1  1.59 0.2328647 0.2489349 0.1142799 10.55321 0.1039534 0.3035853 0.4638871
222     2 dominicks   46  8.987197    0  2.69 0.2328647 0.2489349 0.1142799 10.55321 0.1039534 0.3035853 0.4638871
223     2 dominicks   47  8.831712    1  2.09 0.2328647 0.2489349 0.1142799 10.55321 0.1039534 0.3035853 0.4638871
224     2 dominicks   48  7.965546    0  2.09 0.2328647 0.2489349 0.1142799 10.55321 0.1039534 0.3035853 0.4638871
225     2 dominicks   50  7.377759    0  2.09 0.2328647 0.2489349 0.1142799 10.55321 0.1039534 0.3035853 0.4638871
226     2 dominicks   51 10.140297    1  1.89 0.2328647 0.2489349 0.1142799 10.55321 0.1039534 0.3035853 0.4638871
    SSTRDIST  SSTRVOL CPDIST5   CPWVOL5
221 2.110122 1.142857 1.92728 0.3769266
222 2.110122 1.142857 1.92728 0.3769266
223 2.110122 1.142857 1.92728 0.3769266
224 2.110122 1.142857 1.92728 0.3769266
225 2.110122 1.142857 1.92728 0.3769266
226 2.110122 1.142857 1.92728 0.3769266
> 
```

Logical vectors Vs. which

- which() removes NA values in the logical vector
- It only returns the indices where the logical vector is TRUE

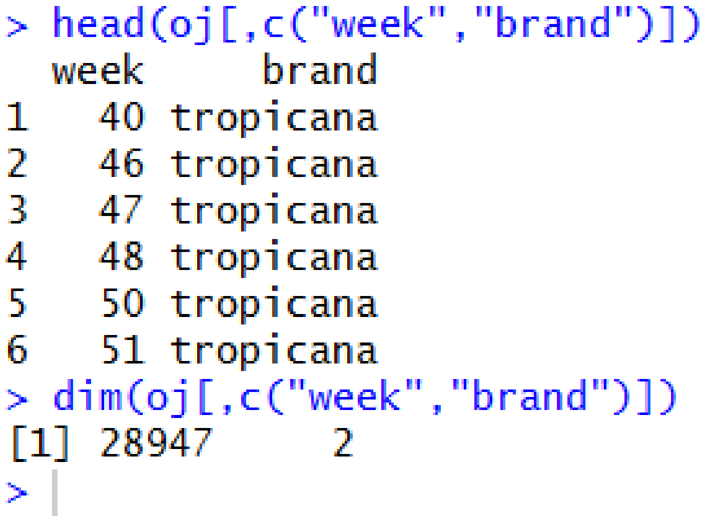


```
> #Consider vector sales with missing values
> sales<-c(100,200,NA,300,400,NA,500,600,700,NA,1000,1500,NA,NA)
> #subset data using logical operator
> sales[sales>600]
[1]   NA   NA  700   NA 1000 1500   NA   NA
> #subset data using which
> sales[which(sales>600)]
[1]  700 1000 1500
```

Selecting Columns

Manipulating data: Base R (Selecting)

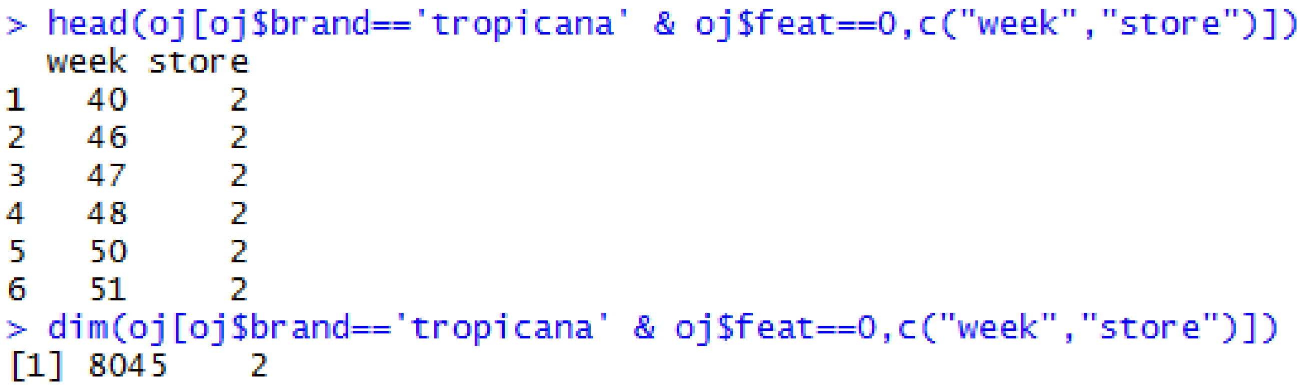
- Selecting a specified set of columns



```
> head(oj[,c("week","brand")])
  week     brand
1   40 tropicana
2   46 tropicana
3   47 tropicana
4   48 tropicana
5   50 tropicana
6   51 tropicana
> dim(oj[,c("week","brand")])
[1] 28947     2
> 
```

Manipulating data: Base R

- Selecting + Sub-setting

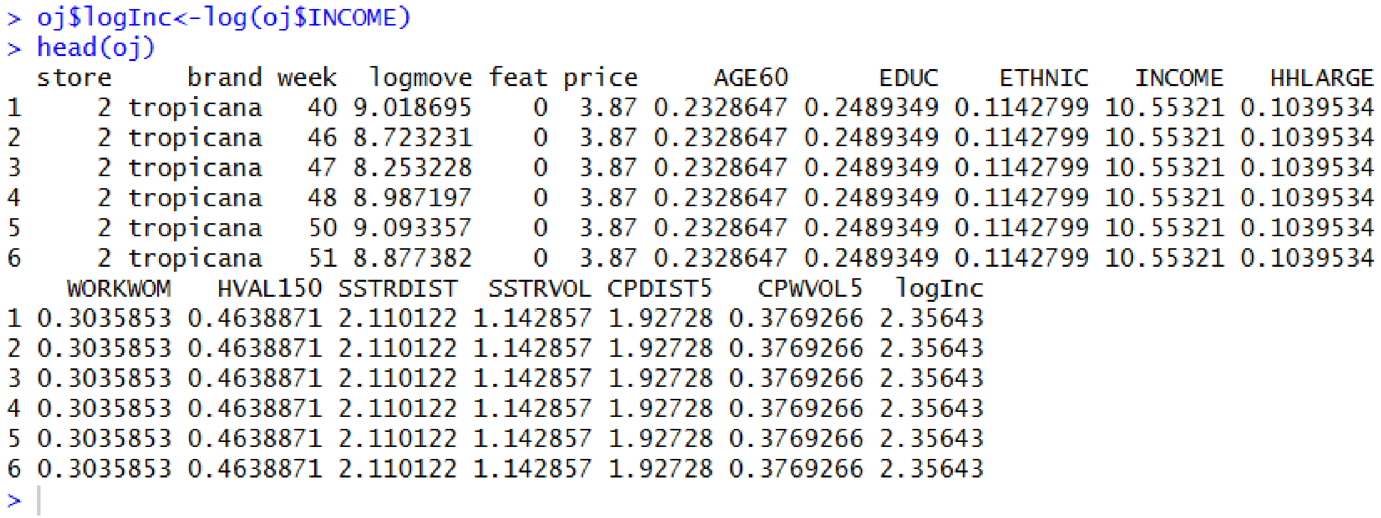


```
> head(oj[oj$brand=='tropicana' & oj$feat==0,c("week","store")])
  week store
1   40     2
2   46     2
3   47     2
4   48     2
5   50     2
6   51     2
> dim(oj[oj$brand=='tropicana' & oj$feat==0,c("week","store")])
[1] 8045    2
```

Adding new columns

Manipulating data: Base R

- Adding new columns

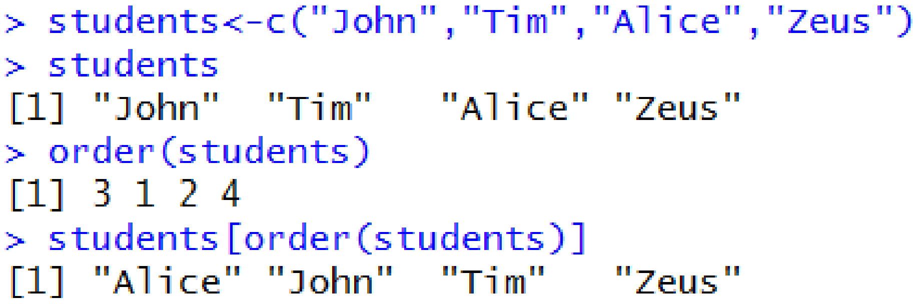


```
> oj$logInc<-log(oj$INCOME)
> head(oj)
  store     brand week  logmove feat price     AGE60      EDUC    ETHNIC   INCOME    HHLARGE
1     2 tropicana   40 9.018695    0  3.87 0.2328647 0.2489349 0.1142799 10.55321 0.1039534
2     2 tropicana   46 8.723231    0  3.87 0.2328647 0.2489349 0.1142799 10.55321 0.1039534
3     2 tropicana   47 8.253228    0  3.87 0.2328647 0.2489349 0.1142799 10.55321 0.1039534
4     2 tropicana   48 8.987197    0  3.87 0.2328647 0.2489349 0.1142799 10.55321 0.1039534
5     2 tropicana   50 9.093357    0  3.87 0.2328647 0.2489349 0.1142799 10.55321 0.1039534
6     2 tropicana   51 8.877382    0  3.87 0.2328647 0.2489349 0.1142799 10.55321 0.1039534
    WORKWOM   HVAL150 SSTRDIST  SSTRVOL CPDIST5   CPWVOL5  logInc
1 0.3035853 0.4638871 2.110122 1.142857 1.92728 0.3769266 2.35643
2 0.3035853 0.4638871 2.110122 1.142857 1.92728 0.3769266 2.35643
3 0.3035853 0.4638871 2.110122 1.142857 1.92728 0.3769266 2.35643
4 0.3035853 0.4638871 2.110122 1.142857 1.92728 0.3769266 2.35643
5 0.3035853 0.4638871 2.110122 1.142857 1.92728 0.3769266 2.35643
6 0.3035853 0.4638871 2.110122 1.142857 1.92728 0.3769266 2.35643
> 
```

Ordering data

Ordering

- order() returns the element order that results in a sorted vector

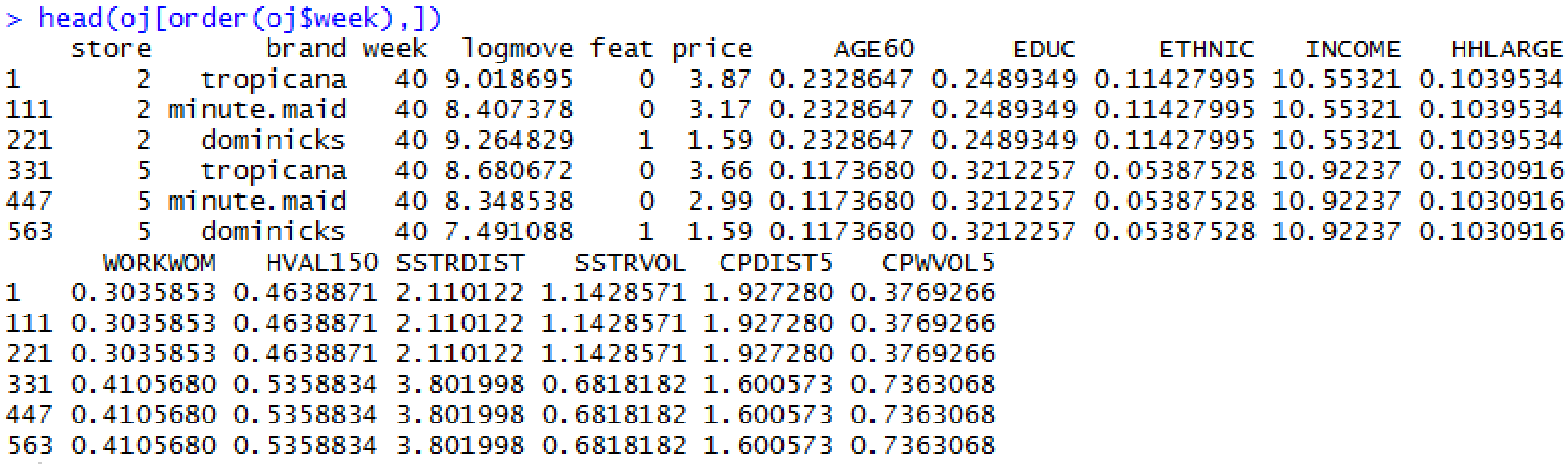


```
> students<-c("John","Tim","Alice","Zeus")
> students
[1] "John"  "Tim"   "Alice" "Zeus"
> order(students)
[1] 3 1 2 4
> students[order(students)]
[1] "Alice" "John"  "Tim"   "Zeus"
```

- Application: Very useful for sorting dataframes

Manipulating data: Base R

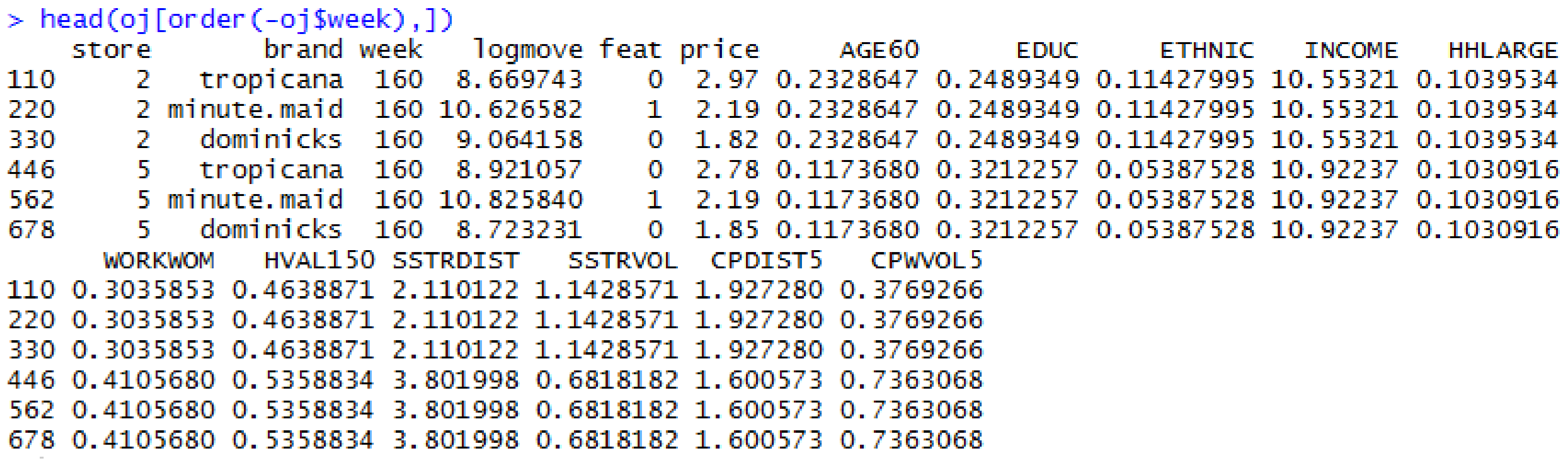
- Ordering data



```
> head(oj[order(oj$week),])
    store       brand week  logmove feat price     AGE60      EDUC     ETHNIC   INCOME   HHLARGE
1       2   tropicana   40 9.018695    0  3.87 0.2328647 0.2489349 0.11427995 10.55321 0.1039534
111     2 minute.maid   40 8.407378    0  3.17 0.2328647 0.2489349 0.11427995 10.55321 0.1039534
221     2   dominicks   40 9.264829    1  1.59 0.2328647 0.2489349 0.11427995 10.55321 0.1039534
331     5   tropicana   40 8.680672    0  3.66 0.1173680 0.3212257 0.05387528 10.92237 0.1030916
447     5 minute.maid   40 8.348538    0  2.99 0.1173680 0.3212257 0.05387528 10.92237 0.1030916
563     5   dominicks   40 7.491088    1  1.59 0.1173680 0.3212257 0.05387528 10.92237 0.1030916
      WORKWOM   HVAL150 SSTRDIST   SSTRVOL  CPDIST5   CPWVOL5
1   0.3035853 0.4638871 2.110122 1.1428571 1.927280 0.3769266
111 0.3035853 0.4638871 2.110122 1.1428571 1.927280 0.3769266
221 0.3035853 0.4638871 2.110122 1.1428571 1.927280 0.3769266
331 0.4105680 0.5358834 3.801998 0.6818182 1.600573 0.7363068
447 0.4105680 0.5358834 3.801998 0.6818182 1.600573 0.7363068
563 0.4105680 0.5358834 3.801998 0.6818182 1.600573 0.7363068
```

Manipulating data: Base R

- Ordering data



```
> head(oj[order(-oj$week),])
    store       brand week   logmove feat price     AGE60      EDUC     ETHNIC   INCOME   HHLARGE
110     2   tropicana  160  8.669743    0  2.97 0.2328647 0.2489349 0.11427995 10.55321 0.1039534
220     2 minute.maid  160 10.626582    1  2.19 0.2328647 0.2489349 0.11427995 10.55321 0.1039534
330     2   dominicks  160  9.064158    0  1.82 0.2328647 0.2489349 0.11427995 10.55321 0.1039534
446     5   tropicana  160  8.921057    0  2.78 0.1173680 0.3212257 0.05387528 10.92237 0.1030916
562     5 minute.maid  160 10.825840    1  2.19 0.1173680 0.3212257 0.05387528 10.92237 0.1030916
678     5   dominicks  160  8.723231    0  1.85 0.1173680 0.3212257 0.05387528 10.92237 0.1030916
      WORKWOM   HVAL150 SSTRDIST   SSTRVOL  CPDIST5   CPWVOL5
110 0.3035853 0.4638871 2.110122 1.1428571 1.927280 0.3769266
220 0.3035853 0.4638871 2.110122 1.1428571 1.927280 0.3769266
330 0.3035853 0.4638871 2.110122 1.1428571 1.927280 0.3769266
446 0.4105680 0.5358834 3.801998 0.6818182 1.600573 0.7363068
562 0.4105680 0.5358834 3.801998 0.6818182 1.600573 0.7363068
678 0.4105680 0.5358834 3.801998 0.6818182 1.600573 0.7363068
```

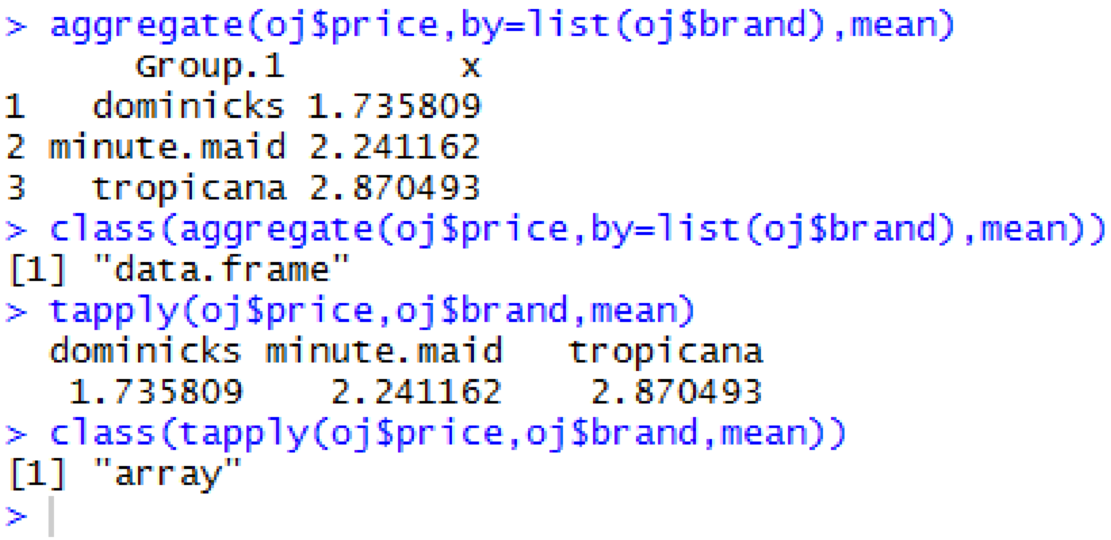
Manipulating data: Base R

- Subsetting data: Using logical subsets and which() statement
- Selecting columns: Using column names at column index
- Adding new columns: Use of $ operator
- Re-ordering data: order()
- Group Wise Summaries
- Producing Contingency tables

GroupWise operations

Manipulating data: Base R

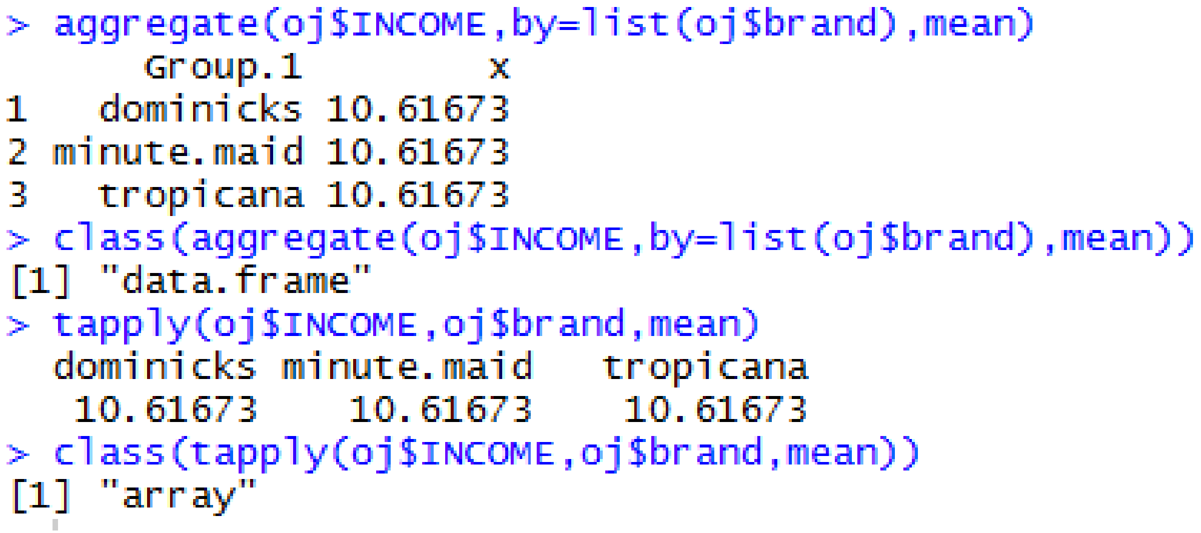
- GroupWise operations
- tapply(), aggregate()
- What is the mean price of each brand of juice across all stores?



```
> aggregate(oj$price,by=list(oj$brand),mean)
     Group.1        x
1   dominicks 1.735809
2 minute.maid 2.241162
3   tropicana 2.870493
> class(aggregate(oj$price,by=list(oj$brand),mean))
[1] "data.frame"
> tapply(oj$price,oj$brand,mean)
  dominicks minute.maid   tropicana 
   1.735809    2.241162    2.870493 
> class(tapply(oj$price,oj$brand,mean))
[1] "array"
> 
```

Manipulating data: Base R

- GroupWise operations
- tapply(), aggregate()
- What is the mean income level corresponding to brand of juice across all stores?



```
> aggregate(oj$INCOME,by=list(oj$brand),mean)
      Group.1        x
1   dominicks 10.61673
2 minute.maid 10.61673
3   tropicana 10.61673
> class(aggregate(oj$INCOME,by=list(oj$brand),mean))
[1] "data.frame"
> tapply(oj$INCOME,oj$brand,mean)
  dominicks minute.maid   tropicana
   10.61673    10.61673    10.61673
> class(tapply(oj$INCOME,oj$brand,mean))
[1] "array"
```

Contingency tables

Manipulating data: Base R

- Category wise counts: Contingency tables

Income	Age	Gender	Location
10,000,000	24	M	Arizona
20,000,000	32	F	California
15,000,000	28	M	Arizona
18,000,000	26	F	California

Manipulating data: Base R

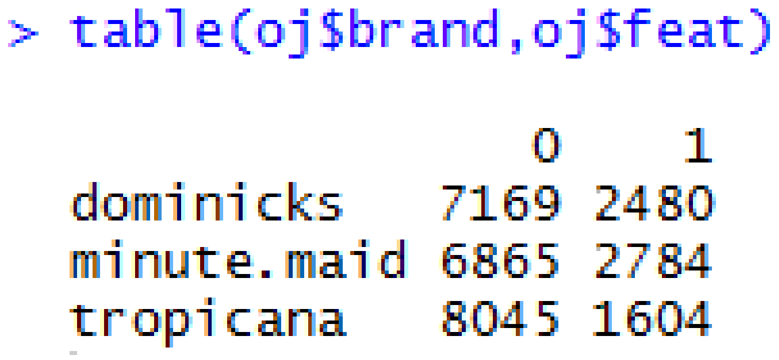
- Category wise counts: Contingency tables

Counts	California	Arizona
Male	0	2
Female	2	0

Income	California	Arizona
Male	0	10,000,000+15,000,000
Female	20,000,000+18,000,000	

Manipulating data: Base R

- Category wise counts: Contingency tables
- table(), xtab()
- Number of people who bought different brands categorized by presence of advertising campaigns

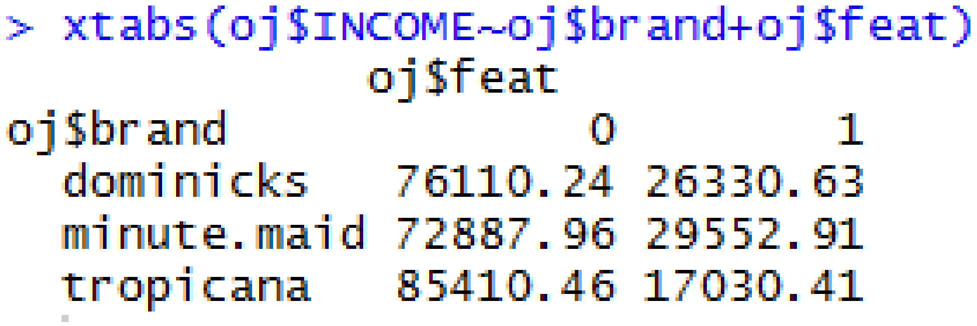


```
> table(oj$brand,oj$feat)

                 0    1
  dominicks   7169 2480
  minute.maid 6865 2784
  tropicana   8045 1604
```

Manipulating data: Base R

- Category wise counts: Contingency tables
- table(), xtab()
- Total income categorized by brand and presence of advertisements



```
> xtabs(oj$INCOME~oj$brand+oj$feat)
             oj$feat
oj$brand             0        1
  dominicks   76110.24 26330.63
  minute.maid 72887.96 29552.91
  tropicana   85410.46 17030.41
```

RECAP

- Sub-setting data
- Selecting specified columns
- Adding new columns
- Reordering data (Ascending/Descending order)
- Group wise operations
- Producing contingency tables