DATA SCIENCE WITH R

Class 4 – Data Manipulation in R

Topic 2

★ Using dplyr to Manipulate Data ★

© Jigsaw Academy Education Pvt Ltd

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

© Jigsaw Academy Education Pvt Ltd

Data Manipulation: dplyr

© Jigsaw Academy Education Pvt Ltd

Manipulating data: dplyr

- dplyr: Whats and Whys
- Sub-setting data using filter()
- Selecting columns using select()
- Adding new columns using mutate()
- Ordering data using arrange()
- Summarizing using summarize() and group_by()
- Using functional pipelines to do more than one manipulation task

© Jigsaw Academy Education Pvt Ltd

Manipulating data: dplyr

- Base R: Good for Medium sized data sets, Awkward Syntax
- dplyr: Faster and elegant syntax
- dplyr: Dataframes
- install.packages(“dplyr”)
- library(dplyr)

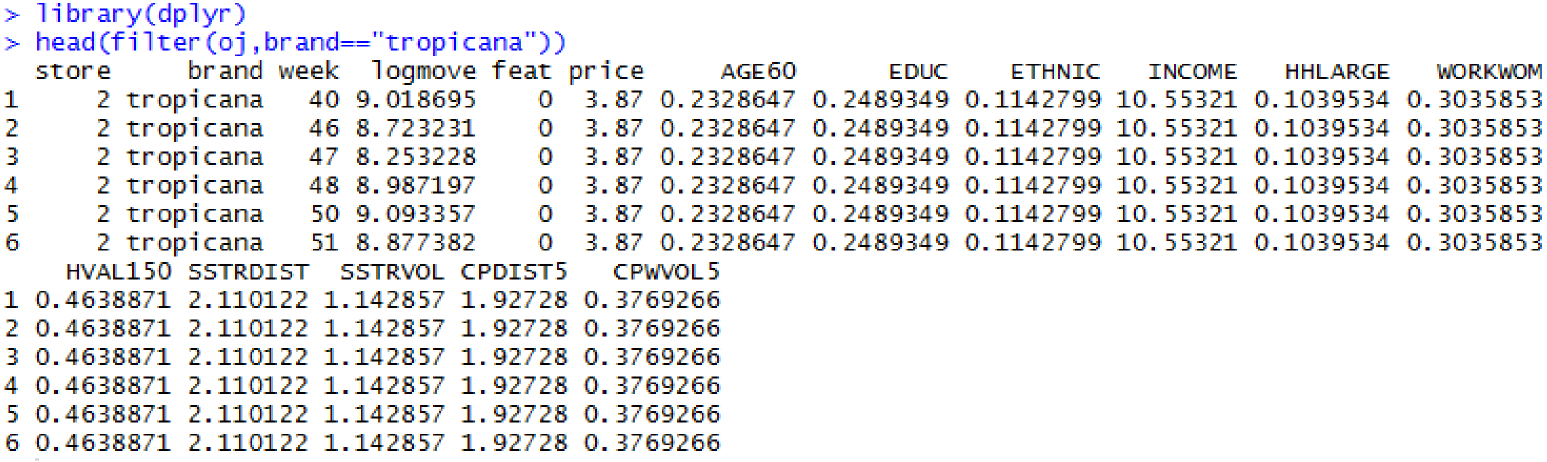
© Jigsaw Academy Education Pvt Ltd

Sub-setting: filter()

© Jigsaw Academy Education Pvt Ltd

Manipulating data: dplyr

- Sub-setting the data using filter(), base R equivalents: logical subsets and which()
- Only that portion of data such that brand bought is “tropicana”

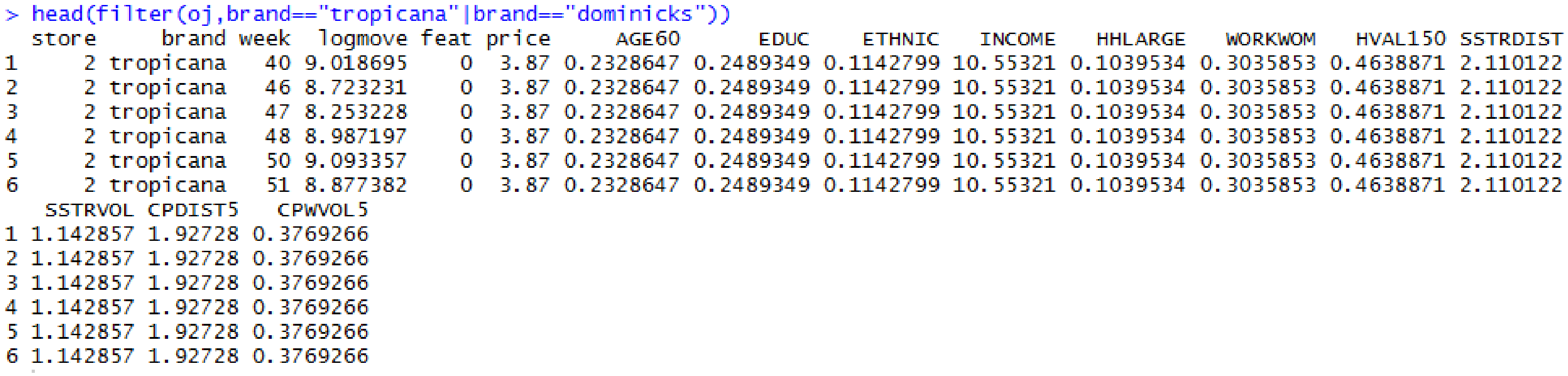


```
> library(dplyr)
> head(filter(oj,brand=="tropicana"))
  store     brand week  logmove feat price     AGE60      EDUC    ETHNIC   INCOME   HHLARGE   WORKWOM
1     2 tropicana   40 9.018695    0  3.87 0.2328647 0.2489349 0.1142799 10.55321 0.1039534 0.3035853
2     2 tropicana   46 8.723231    0  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
```

© Jigsaw Academy Education Pvt Ltd

Manipulating data: dplyr

- Sub-setting the data using filter(), base R equivalents: logical subsets and which()
- Only that portion of data such that brand bought is “tropicana” or “dominicks”



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

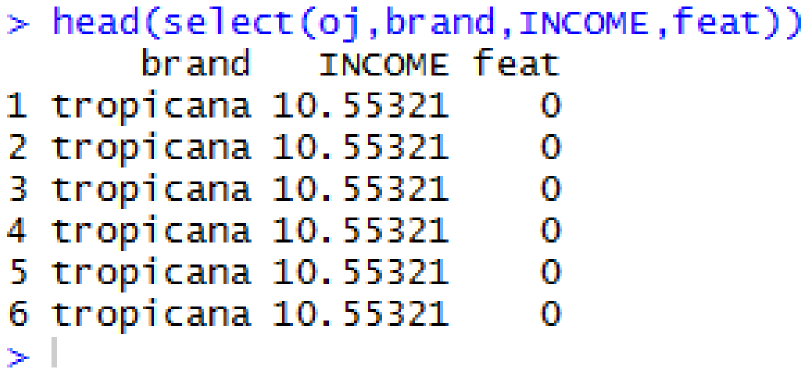
© Jigsaw Academy Education Pvt Ltd

Selecting Columns: select()

© Jigsaw Academy Education Pvt Ltd

Manipulating data: dplyr

- Selecting columns from data using select(), base R equivalents: index subsets
- Selecting columns brand and income

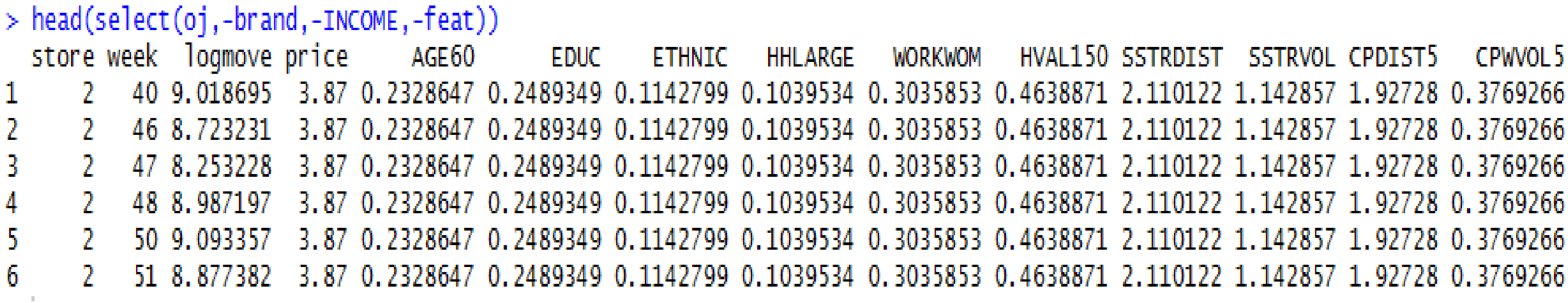


```
> head(select(oj,brand,INCOME,feat))
      brand   INCOME feat
1 tropicana 10.55321    0
2 tropicana 10.55321    0
3 tropicana 10.55321    0
4 tropicana 10.55321    0
5 tropicana 10.55321    0
6 tropicana 10.55321    0
> 
```

© Jigsaw Academy Education Pvt Ltd

Manipulating data: dplyr

- Selecting columns from data using select(), base R equivalents: index subsets
- Dropping columns brand and income



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

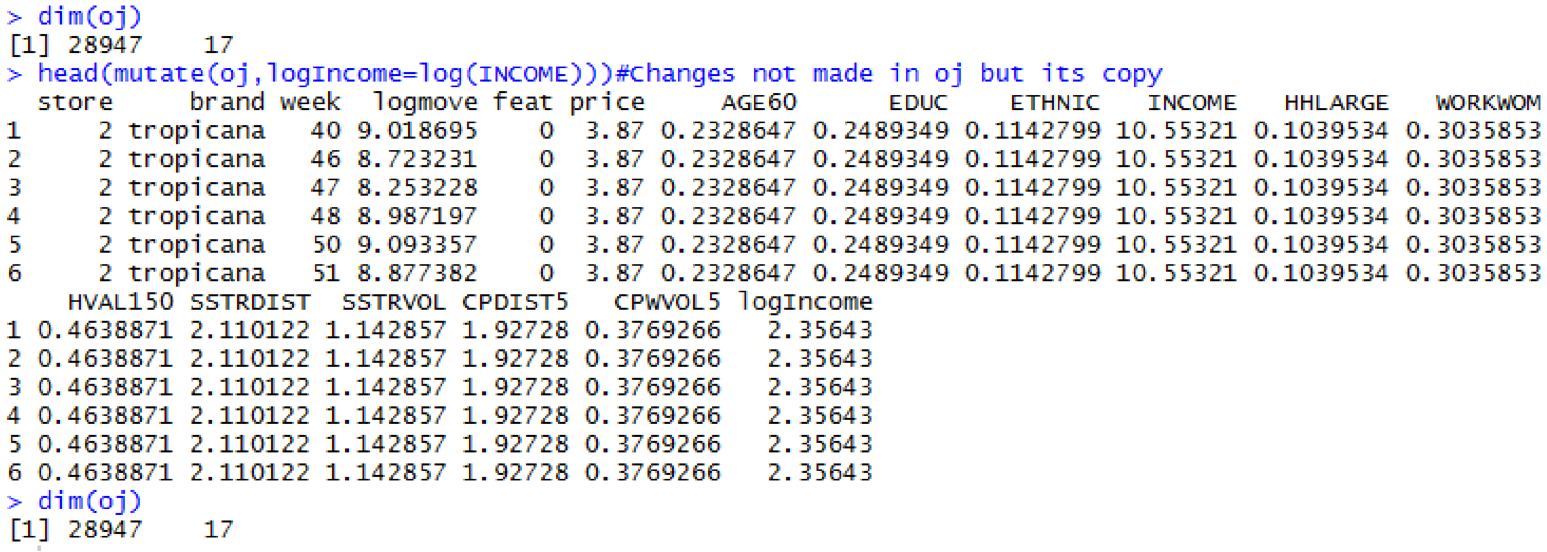
© Jigsaw Academy Education Pvt Ltd

Creating New Columns: mutate()

© Jigsaw Academy Education Pvt Ltd

Manipulating data: dplyr

- Adding columns to data using mutate(),
- Adding a new column, log(income)



```
> dim(oj)
[1] 28947    17
> head(mutate(oj,logIncome=log(INCOME)))#Changes not made in oj but its copy
  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 logIncome
1 0.4638871 2.110122 1.142857 1.92728 0.3769266   2.35643
2 0.4638871 2.110122 1.142857 1.92728 0.3769266   2.35643
3 0.4638871 2.110122 1.142857 1.92728 0.3769266   2.35643
4 0.4638871 2.110122 1.142857 1.92728 0.3769266   2.35643
5 0.4638871 2.110122 1.142857 1.92728 0.3769266   2.35643
6 0.4638871 2.110122 1.142857 1.92728 0.3769266   2.35643
> dim(oj)
[1] 28947    17
```

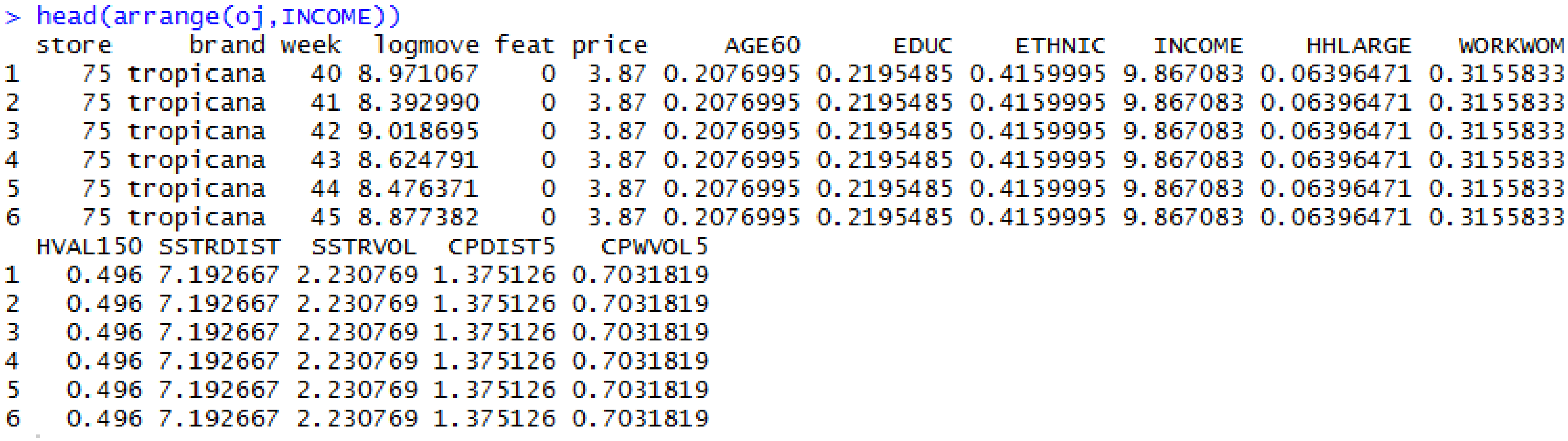
© Jigsaw Academy Education Pvt Ltd

Ordering data: arrange()

© Jigsaw Academy Education Pvt Ltd

Manipulating data: dplyr

- Ordering data using order_by(),
- Order whole data by income in ascending order

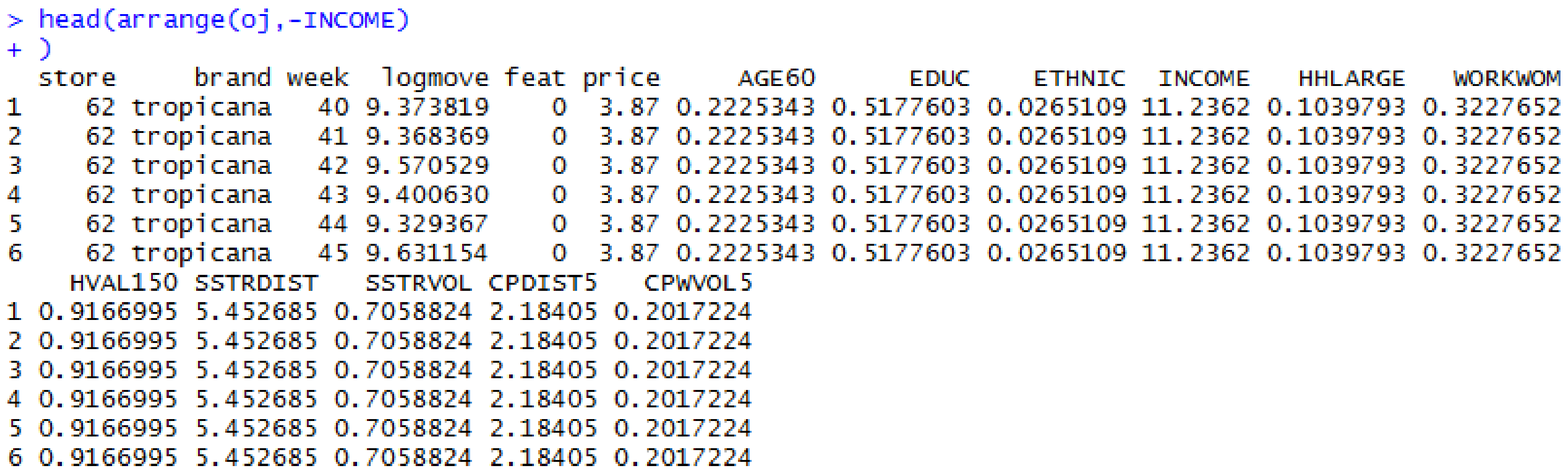


```
> head(arrange(oj,INCOME))
  store     brand week  logmove feat price     AGE60      EDUC    ETHNIC   INCOME     HHLARGE   WORKWOM
1    75 tropicana   40 8.971067    0  3.87 0.2076995 0.2195485 0.4159995 9.867083 0.06396471 0.3155833
2    75 tropicana   41 8.392990    0  3.87 0.2076995 0.2195485 0.4159995 9.867083 0.06396471 0.3155833
3    75 tropicana   42 9.018695    0  3.87 0.2076995 0.2195485 0.4159995 9.867083 0.06396471 0.3155833
4    75 tropicana   43 8.624791    0  3.87 0.2076995 0.2195485 0.4159995 9.867083 0.06396471 0.3155833
5    75 tropicana   44 8.476371    0  3.87 0.2076995 0.2195485 0.4159995 9.867083 0.06396471 0.3155833
6    75 tropicana   45 8.877382    0  3.87 0.2076995 0.2195485 0.4159995 9.867083 0.06396471 0.3155833
  HVAL150 SSTRDIST  SSTRVOL  CPDIST5   CPWVOL5
1   0.496 7.192667 2.230769 1.375126 0.7031819
2   0.496 7.192667 2.230769 1.375126 0.7031819
3   0.496 7.192667 2.230769 1.375126 0.7031819
4   0.496 7.192667 2.230769 1.375126 0.7031819
5   0.496 7.192667 2.230769 1.375126 0.7031819
6   0.496 7.192667 2.230769 1.375126 0.7031819
```

© Jigsaw Academy Education Pvt Ltd

Manipulating data: dplyr

- Ordering data using order_by(),
- Order whole data by income in descending order



```
> head(arrange(oj,-INCOME)
+ )
  store     brand week  logmove feat price     AGE60      EDUC    ETHNIC  INCOME   HHLARGE   WORKWOM
1    62 tropicana   40 9.373819    0  3.87 0.2225343 0.5177603 0.0265109 11.2362 0.1039793 0.3227652
2    62 tropicana   41 9.368369    0  3.87 0.2225343 0.5177603 0.0265109 11.2362 0.1039793 0.3227652
3    62 tropicana   42 9.570529    0  3.87 0.2225343 0.5177603 0.0265109 11.2362 0.1039793 0.3227652
4    62 tropicana   43 9.400630    0  3.87 0.2225343 0.5177603 0.0265109 11.2362 0.1039793 0.3227652
5    62 tropicana   44 9.329367    0  3.87 0.2225343 0.5177603 0.0265109 11.2362 0.1039793 0.3227652
6    62 tropicana   45 9.631154    0  3.87 0.2225343 0.5177603 0.0265109 11.2362 0.1039793 0.3227652
    HVAL150 SSTRDIST   SSTRVOL CPDIST5   CPWVOL5
1 0.9166995 5.452685 0.7058824 2.18405 0.2017224
2 0.9166995 5.452685 0.7058824 2.18405 0.2017224
3 0.9166995 5.452685 0.7058824 2.18405 0.2017224
4 0.9166995 5.452685 0.7058824 2.18405 0.2017224
5 0.9166995 5.452685 0.7058824 2.18405 0.2017224
6 0.9166995 5.452685 0.7058824 2.18405 0.2017224
```

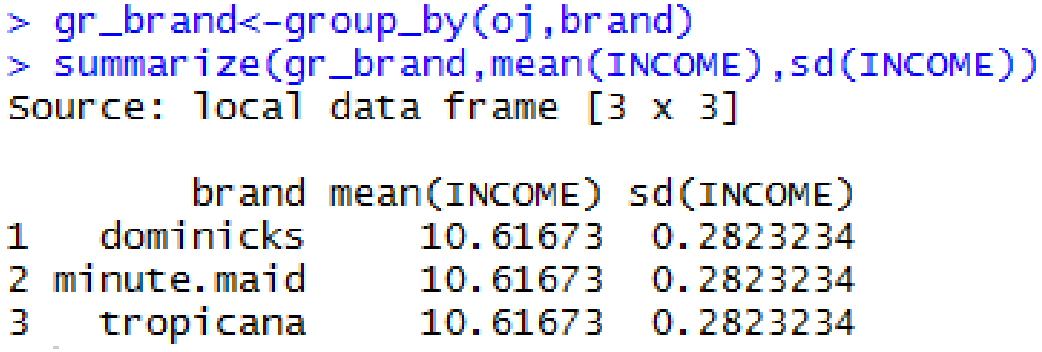
© Jigsaw Academy Education Pvt Ltd

Summarizing data:
summarize() and group_by()

© Jigsaw Academy Education Pvt Ltd

Manipulating data: dplyr

- Summarizing data using summarize() and group_by()
- group_by() makes grouped table, summarize() can take this grouped table and produce summaries for different columns
- Mean level of income and standard deviation of income for each brand of orange juice



```
> gr_brand<-group_by(oj,brand)
> summarize(gr_brand,mean(INCOME),sd(INCOME))
Source: local data frame [3 x 3]

        brand mean(INCOME) sd(INCOME)
1   dominicks     10.61673  0.2823234
2 minute.maid     10.61673  0.2823234
3   tropicana     10.61673  0.2823234
```

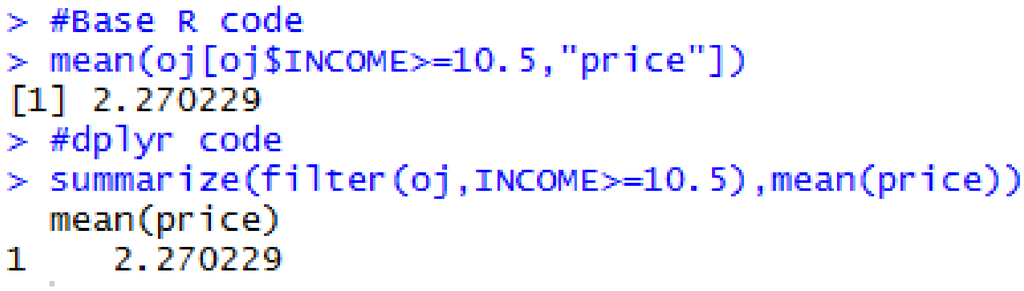
© Jigsaw Academy Education Pvt Ltd

Functional Pipelines: %>%

© Jigsaw Academy Education Pvt Ltd

Manipulating data: dplyr

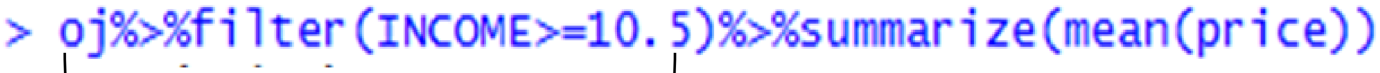
- dplyr becomes a powerful tool when combined with %>% (pipe) operator
- Several data manipulation tasks can be accomplished in just one line of code
- Traditionally functional composition is achieved by using nested function calls
- For example, Find the mean price for all people whose income is >=10.5



```
> #Base R code
> mean(oj[oj$INCOME>=10.5,"price"])
[1] 2.270229
> #dplyr code
> summarize(filter(oj,INCOME>=10.5),mean(price))
  mean(price)
1    2.270229
```

© Jigsaw Academy Education Pvt Ltd

Manipulating data: dplyr

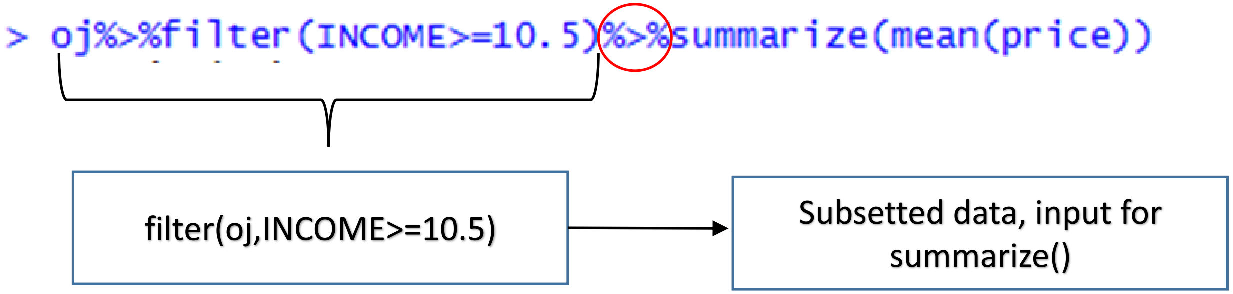


```
> oj%>%filter(INCOME>=10.5)%>%summarize(mean(price))
```

filter(oj,INCOME>=10.5)

© Jigsaw Academy Education Pvt Ltd

Manipulating data: dplyr



```
> oj%>%filter(INCOME>=10.5)%>%summarize(mean(price))
```

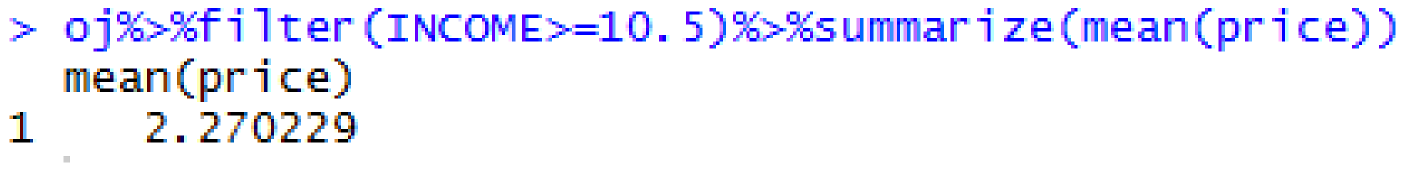
filter(oj,INCOME>=10.5)

Subsetted data, input for summarize()

© Jigsaw Academy Education Pvt Ltd

Manipulating data: dplyr

- Clearly the code looks very messy, using a %>% operator, we can make it more readable



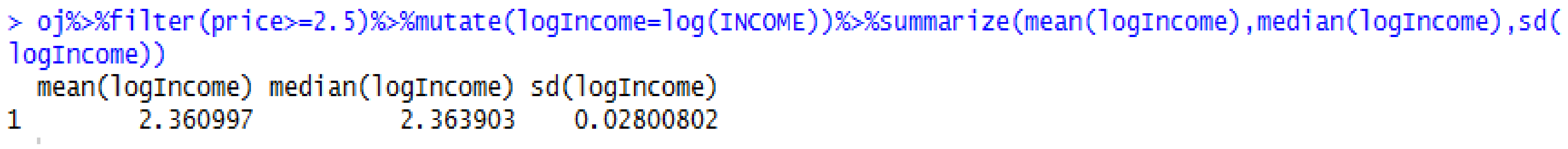
```
> oj%>%filter(INCOME>=10.5)%>%summarize(mean(price))
  mean(price)
1    2.270229
```

- This can be easily read as:
- Take data oj, filter it based on income
- Take this filtered data frame and compute the mean of price

© Jigsaw Academy Education Pvt Ltd

Manipulating data: dplyr

- Subset the data based on price>=2.5, create a column logIncome, compute the mean, standard deviation and median of column logIncome



```
> oj%>%filter(price>=2.5)%>%mutate(logIncome=log(INCOME))%>%summarize(mean(logIncome),median(logIncome),sd(
logIncome))
  mean(logIncome) median(logIncome) sd(logIncome)
1        2.360997          2.363903    0.02800802
```

© Jigsaw Academy Education Pvt Ltd

RECAP

- dplyr: better manipulation functionality
- Sub-setting data using filter()
- Selecting columns using select()
- Adding new columns using mutate()
- Ordering data using arrange()
- Summarizing using summarize() and group_by()
- Using functional pipelines to do more than one manipulation task

© Jigsaw Academy Education Pvt Ltd