dply

Rob Haywardr

Thursday, August 28, 2014

This is the video from [R-videos]{<http://rmarkdown.rstudio.com>}

Load the data.

library(dplyr)  
library(ggplot2)  
train = read.csv("../../Data/train.csv")  
head(train)

## PassengerId Survived Pclass  
## 1 1 0 3  
## 2 2 1 1  
## 3 3 1 3  
## 4 4 1 1  
## 5 5 0 3  
## 6 6 0 3  
## Name Sex Age SibSp  
## 1 Braund, Mr. Owen Harris male 22 1  
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female 38 1  
## 3 Heikkinen, Miss. Laina female 26 0  
## 4 Futrelle, Mrs. Jacques Heath (Lily May Peel) female 35 1  
## 5 Allen, Mr. William Henry male 35 0  
## 6 Moran, Mr. James male NA 0  
## Parch Ticket Fare Cabin Embarked  
## 1 0 A/5 21171 7.250 S  
## 2 0 PC 17599 71.283 C85 C  
## 3 0 STON/O2. 3101282 7.925 S  
## 4 0 113803 53.100 C123 S  
## 5 0 373450 8.050 S  
## 6 0 330877 8.458 Q

key functions in dplyr \* *select:* select columns from a dataframe \* *filter:* select rows from a data frame based upon criteria \* *group\_by:* group by a factor variable \* *summarize:* allows you to do summary stats based upon the grouped variable \* *arrange:* a better way to order the data set

### selecting columns and rows of a dataframe

selecting columns

VariablesThatICareAbout = select(train, Survived, Pclass, Sex, Age, SibSp, Parch, Fare)   
head(VariablesThatICareAbout)

## Survived Pclass Sex Age SibSp Parch Fare  
## 1 0 3 male 22 1 0 7.250  
## 2 1 1 female 38 1 0 71.283  
## 3 1 3 female 26 0 0 7.925  
## 4 1 1 female 35 1 0 53.100  
## 5 0 3 male 35 0 0 8.050  
## 6 0 3 male NA 0 0 8.458

Everything except for PassengerId:

VariablesThatICareAbout = select(train, -PassengerId)   
head(VariablesThatICareAbout)

## Survived Pclass Name  
## 1 0 3 Braund, Mr. Owen Harris  
## 2 1 1 Cumings, Mrs. John Bradley (Florence Briggs Thayer)  
## 3 1 3 Heikkinen, Miss. Laina  
## 4 1 1 Futrelle, Mrs. Jacques Heath (Lily May Peel)  
## 5 0 3 Allen, Mr. William Henry  
## 6 0 3 Moran, Mr. James  
## Sex Age SibSp Parch Ticket Fare Cabin Embarked  
## 1 male 22 1 0 A/5 21171 7.250 S  
## 2 female 38 1 0 PC 17599 71.283 C85 C  
## 3 female 26 0 0 STON/O2. 3101282 7.925 S  
## 4 female 35 1 0 113803 53.100 C123 S  
## 5 male 35 0 0 373450 8.050 S  
## 6 male NA 0 0 330877 8.458 Q

Columns Survived to Age

VariablesThatICareAbout = select(train, Survived:Age)   
head(VariablesThatICareAbout)

## Survived Pclass Name  
## 1 0 3 Braund, Mr. Owen Harris  
## 2 1 1 Cumings, Mrs. John Bradley (Florence Briggs Thayer)  
## 3 1 3 Heikkinen, Miss. Laina  
## 4 1 1 Futrelle, Mrs. Jacques Heath (Lily May Peel)  
## 5 0 3 Allen, Mr. William Henry  
## 6 0 3 Moran, Mr. James  
## Sex Age  
## 1 male 22  
## 2 female 38  
## 3 female 26  
## 4 female 35  
## 5 male 35  
## 6 male NA

Selecting rows: only first class passengers

FirstClass = filter(train, Pclass == 1)

Only first class male passengers

FirstClass = filter(train, Pclass == 1 & Sex == "male")

Ordering the rows

train = arrange(train, Fare, Pclass)  
head(train)

## PassengerId Survived Pclass Name Sex Age  
## 1 264 0 1 Harrison, Mr. William male 40  
## 2 634 0 1 Parr, Mr. William Henry Marsh male NA  
## 3 807 0 1 Andrews, Mr. Thomas Jr male 39  
## 4 816 0 1 Fry, Mr. Richard male NA  
## 5 823 0 1 Reuchlin, Jonkheer. John George male 38  
## 6 278 0 2 Parkes, Mr. Francis "Frank" male NA  
## SibSp Parch Ticket Fare Cabin Embarked  
## 1 0 0 112059 0 B94 S  
## 2 0 0 112052 0 S  
## 3 0 0 112050 0 A36 S  
## 4 0 0 112058 0 B102 S  
## 5 0 0 19972 0 S  
## 6 0 0 239853 0 S

Multiple commands in a single step The main value is the ability to pipe the output from one command into another use %>% to pipe one command into another

# average fare by passenger class

select(train, Pclass, Fare, Survived) %>%  
 group\_by(Pclass) %>% #don't have to specify the dataset a second time  
 summarize(AvgFare = mean(Fare), ProbSurvived = mean(Survived), N = length(Fare))

## Source: local data frame [3 x 4]  
##   
## Pclass AvgFare ProbSurvived N  
## 1 1 84.15 0.6296 216  
## 2 2 20.66 0.4728 184  
## 3 3 13.68 0.2424 491

# probability of survival by gender

select(train, Sex, Survived) %>%  
 group\_by(Sex) %>%  
 summarize(ProbSurvived = mean(Survived))

## Source: local data frame [2 x 2]  
##   
## Sex ProbSurvived  
## 1 female 0.7420  
## 2 male 0.1889

# both categories at the same time!

(group by two variables)

A = select(train, Pclass, Sex, Fare, Survived) %>%  
 group\_by(Pclass, Sex) %>% #don't have to specify the dataset a second time  
 summarize(AvgFare = mean(Fare), ProbSurvived = mean(Survived), N = length(Fare))  
A

## Source: local data frame [6 x 5]  
## Groups: Pclass  
##   
## Pclass Sex AvgFare ProbSurvived N  
## 1 1 female 106.13 0.9681 94  
## 2 1 male 67.23 0.3689 122  
## 3 2 female 21.97 0.9211 76  
## 4 2 male 19.74 0.1574 108  
## 5 3 female 16.12 0.5000 144  
## 6 3 male 12.66 0.1354 347

### Variable creation

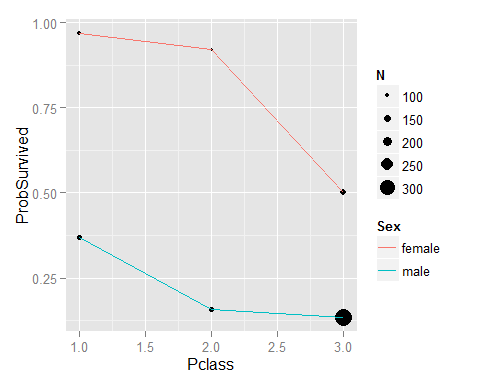
Use the mutate function

train = mutate(train  
 , AgeByFare = Age \* Fare  
 , Age2 = Age^2)  
#allows you to create variables in one step   
#without specifying the data frame a bunch of times  
train$Age2 = train$Age^2

### ggplot teaser

There will be another tutorial on ggplot Just wanted to show this to motivate using ggplot! dplyr makes it easier to produce grouped plots in ggplot

ggplot(A, aes(x = Pclass, y = ProbSurvived)) +   
 geom\_point(aes(size = N)) +   
 geom\_line(aes(by = Sex, color = Sex))

 survival probability unaffected for male 2 or 3 class, but big difference for females in 2nd and 3rd class