Data Manipulation with dplyr

Ni

```
library("dplyr")
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
setwd("~/Dropbox/Coursera/RStudio/Data")
nobel<-read.csv("NobelPrize.csv")
dim(nobel)
```

```
## [1] 969  18
```

```
#glimpse(nobel)
```

Transforming data

select(),arrange(),filter(), mutate()

```
# Select the columns
data_selected1<-nobel %>%
  select(Year,Category,Prize.Share,Full.Name,Sex,
         Birth.City,Birth.Country,Organization.Country)
head(data_selected1)
```

```
##   Year   Category Prize.Share                   Full.Name  Sex
## 1 1901  Chemistry         1/1 Jacobus Henricus van 't Hoff Male
## 2 1901 Literature         1/1              Sully Prudhomme Male
## 3 1901   Medicine         1/1       Emil Adolf von Behring Male
## 4 1901      Peace         1/2            Jean Henry Dunant Male
## 5 1901      Peace         1/2               Frédéric Passy Male
## 6 1901    Physics         1/1       Wilhelm Conrad Röntgen Male
##           Birth.City     Birth.Country Organization.Country
```

```
## 1          Rotterdam       Netherlands             Germany
## 2              Paris            France
## 3  Hansdorf (Lawice)  Prussia (Poland)             Germany
## 4             Geneva       Switzerland
## 5              Paris            France
## 6 Lennep (Remscheid) Prussia (Germany)             Germany
```

```
# sort the data
data_sorted1<- data_selected1 %>%
  arrange(desc(Year)) %>% # from high to low
  filter(Category=="Medicine")
head(data_sorted1)
```

```
##   Year Category Prize.Share           Full.Name    Sex           Birth.City
## 1 2016 Medicine         1/1    Yoshinori Ohsumi   Male              Fukuoka
## 2 2015 Medicine         1/4 William C. Campbell   Male             Ramelton
## 3 2015 Medicine         1/4       Satoshi Ōmura   Male Yamanashi Prefecture
## 4 2015 Medicine         1/2           Youyou Tu Female      Zhejiang Ningbo
## 5 2014 Medicine         1/2        John O'Keefe   Male         New York, NY
## 6 2014 Medicine         1/4     May-Britt Moser Female             Fosnavåg
##              Birth.Country     Organization.Country
## 1                    Japan                    Japan
## 2                  Ireland United States of America
## 3                    Japan                    Japan
## 4                    China                    China
## 5 United States of America           United Kingdom
## 6                   Norway                   Norway
```

```
# Or combine multiple conditions together
data_sorted2<- data_selected1 %>%
  arrange(desc(Year)) %>% # from high to low
  filter(Category=="Medicine",Sex=="Female")
head(data_sorted2)
```

```
##   Year Category Prize.Share               Full.Name    Sex       Birth.City
## 1 2015 Medicine         1/2               Youyou Tu Female  Zhejiang Ningbo
## 2 2014 Medicine         1/4         May-Britt Moser Female         Fosnavåg
## 3 2009 Medicine         1/3   Elizabeth H. Blackburn Female Hobart, Tasmania
## 4 2009 Medicine         1/3         Carol W. Greider Female    San Diego, CA
## 5 2008 Medicine         1/4 Françoise Barré-Sinoussi Female            Paris
## 6 2004 Medicine         1/2            Linda B. Buck Female      Seattle, WA
##              Birth.Country     Organization.Country
## 1                    China                    China
## 2                   Norway                   Norway
## 3                Australia United States of America
## 4 United States of America United States of America
## 5                   France                   France
## 6 United States of America United States of America
```

Aggregating Data

- count()

- wt() -weigh your count by particular variables rather than finding the number of counties.
- group_by() -if you group_by(X, Y) then summarize, the result will still be grouped by X.
- summarize() -summarize(newColName=mean()/max()/sum()...)
- ungroup()
- top_n()

```
data_top10<-nobel %>%
  select(Year,Category,Sex,Organization.Country) %>%
  count(Organization.Country,sort=TRUE)%>%
  top_n(10,Organization.Country)
head(data_top10)
```

```
##                  Organization.Country   n
## 1            United States of America 363
## 2                      United Kingdom  91
## 3                         Switzerland  22
## 4                              Sweden  17
## 5                         Netherlands  11
## 6 Union of Soviet Socialist Republics  11
```

```
data_selected2<-nobel %>%
  select(Year,Category,Sex,Organization.Country) %>%
  group_by(Category) %>%
  filter(Sex=="Female")
head(data_selected2)
```

```
## # A tibble: 6 x 4
## # Groups:   Category [4]
##    Year Category   Sex    Organization.Country
##   <int> <chr>      <chr>  <chr>
## 1  1903 Physics    Female ""
## 2  1905 Peace      Female ""
## 3  1909 Literature Female ""
## 4  1911 Chemistry  Female "France"
## 5  1926 Literature Female ""
## 6  1928 Literature Female ""
```

Selecting and transforming data

- select -select a range using *colname1:colname10*
- select helpers -contain("keyword") -starts_with(): select only cols that starts with a particular prefix -ends_with() -last_col() -select(-colname): remove column
- rename(newcolname=existingcolname) or select(col1,col2, newcolname=col3)
- transmute() -a combination of select and mutate -returns a subset of a columns that are transformed and changed -keeps only specified variables For example, dataset %>% transmute(col1,col2,colAA=col3/col4)

Visualizing with ggplot2()

- %in% example: filter(name %in% c("Steven", "Thomas", "Matthew"))
- Visualize those three names as a line plot over time, with each name represented by a different color. ggplot(selected_names, aes(x = year, y = number, color = name)) + geom_line()

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

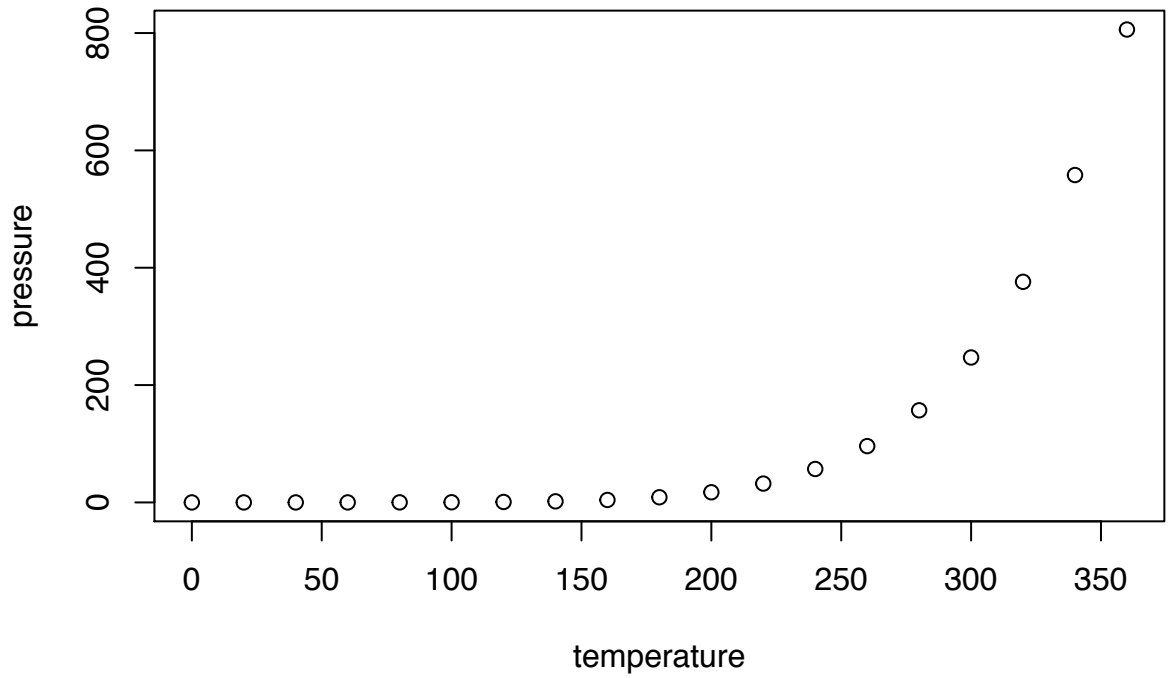
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.