Data Tidying Project

Alex Horde

6/13/2018

Get Data

Data Set 1: Sales from the Retail Trade and Food Services Report from the US Census. This dataset only covers Department Stores, though the report covers a wide range of retail types. [1992-2016]

Data Set 2 US Retail Sales by Store Type with Growth Rate [2009-2014]

```
#1992-2016
#https://data.world/retail/department-store-sales

GET("https://query.data.world/s/gdk7iwtlisq6vkktmybqqr7hjjty5s", write_disk(tf <- tempfile(fileext = ".x
```

```
## Response [https://download.data.world/file_download/retail/department-store-sales/retail-trade-repor
##   Date: 2018-10-23 10:29
##   Status: 200
##   Content-Type: application/vnd.ms-excel
##   Size: 62.5 kB
## <ON DISK>  /tmp/Rtmpx70DvR/file523a33c52590.xls
```

```
df1 <- read_excel(tf)

#2009-2014
# https://data.world/garyhoov/retail-sales-growth
GET("https://query.data.world/s/py7kinxvyuxjpzwdjs2ti4wdmui6bi", write_disk(tf <- tempfile(fileext = ".x
```

```
## Response [https://download.data.world/file_download/garyhoov/retail-sales-growth/US%20Retail%20Sales%
##   Date: 2018-10-23 10:29
##   Status: 200
##   Content-Type: application/vnd.ms-excel
##   Size: 169 kB
## <ON DISK>  /tmp/Rtmpx70DvR/file523a599e073c.xls
```

```
df2 <- read_excel(tf)
## the the first row and make that the column names of the data frame
colnames(df2) <- df2[1,]
```

Save Raw Data

```
## use saveRDS() to save each object as a .rds file
setwd("/cloud/project/data_tidying_project/data/raw_data/")
saveRDS(df1, "df_department.rds")
saveRDS(df2, "df_retail.rds")
```

Wrangle Data

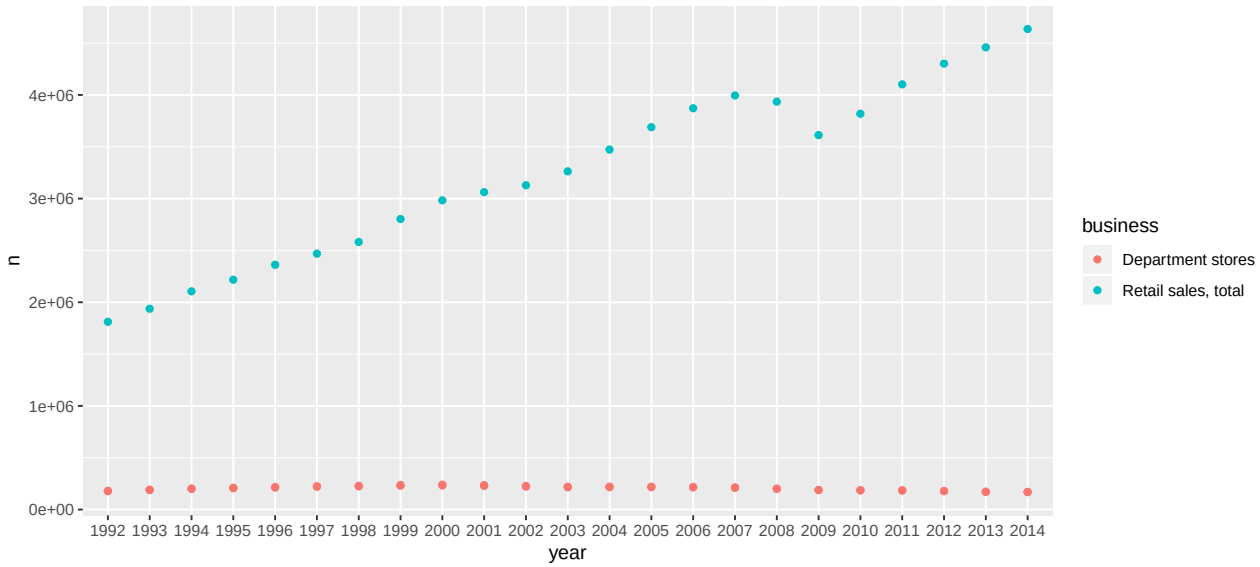
```
## an example working with df2
## let's wrangle!
```

```r
df_retail <- df2 %>%
  ## remove the r from the column names of df2
  magrittr::set_colnames(gsub("r","",df2[1,])) %>%
  ## add a new column called "business"
  mutate(business = gsub("[...]|[.]","",`Kind of business`)) %>%
  ## filter to include Retail sales or Department stores sales
  filter(grepl('Retail sales, total |Department stores', business)) %>%
  ## only look at columns with year information in them
  select(.,c(matches('19|20'),business)) %>%
  ## take year column and collapse them into a single column
  gather(., "year", "n", 1:(ncol(.)-1)) %>%
  ## make sure the count column `n` is numeric
  mutate(n=as.numeric(n)) %>%
  ## filter to only include the businesses we're interested in
  filter(business == "Retail sales, total "| business=="Department stores ")
```

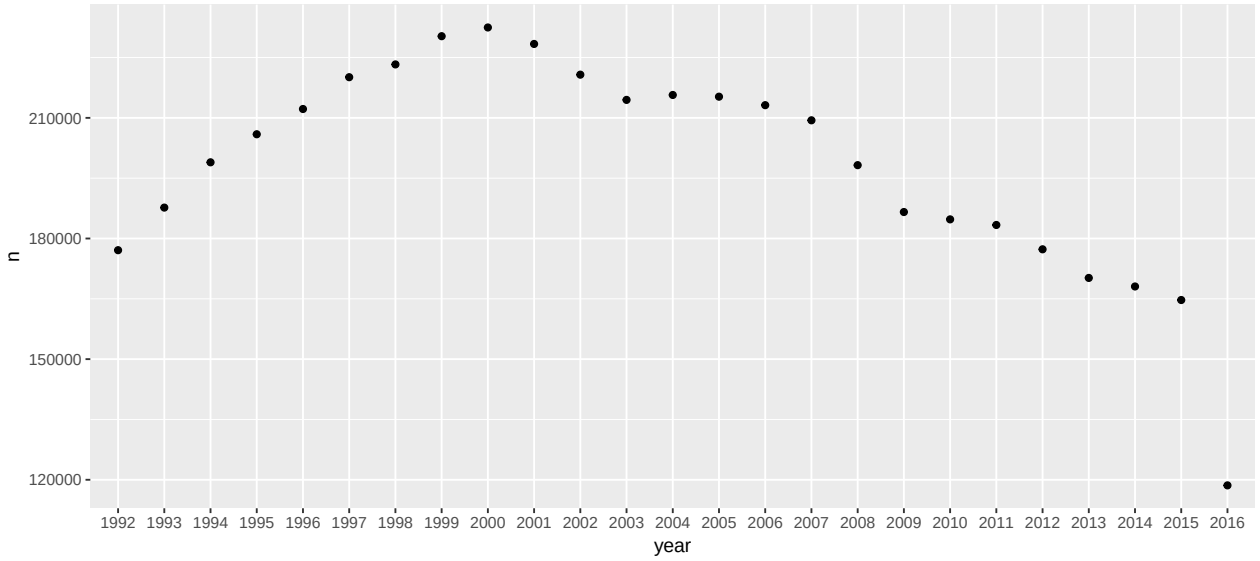
```r
## now, your turn!
## work with df1
df_department <- df1 %>%
  ## split Period column into one column called "month" and one called "year"
  separate(Period, into=c("month", "year"), extra = "drop", remove=FALSE) %>%
  ## add a column `value` which contains the
  ## information from the `Value (in millions)`
  mutate(value=`Value (in millions)`) %>%
  ## group the data frame by the `year` column
  group_by(year) %>%
  ## Summarize the data by creating a new column
  ## call this column `n`
  ## have it contain the sum of the `value` column
  summarize(n=sum(value)) %>%
  ### create a new column called `business`
  ## set the value of this column to  be "department stores"
  ## for the entire data set
  mutate(business="department stores") %>%
  ## reorder column names to be : business, year, n
  select(business, year, n)
```

```r
## Now, combine the two data frames
df_total <- left_join(df_department, df_retail, by = c('business','year','n'))
```

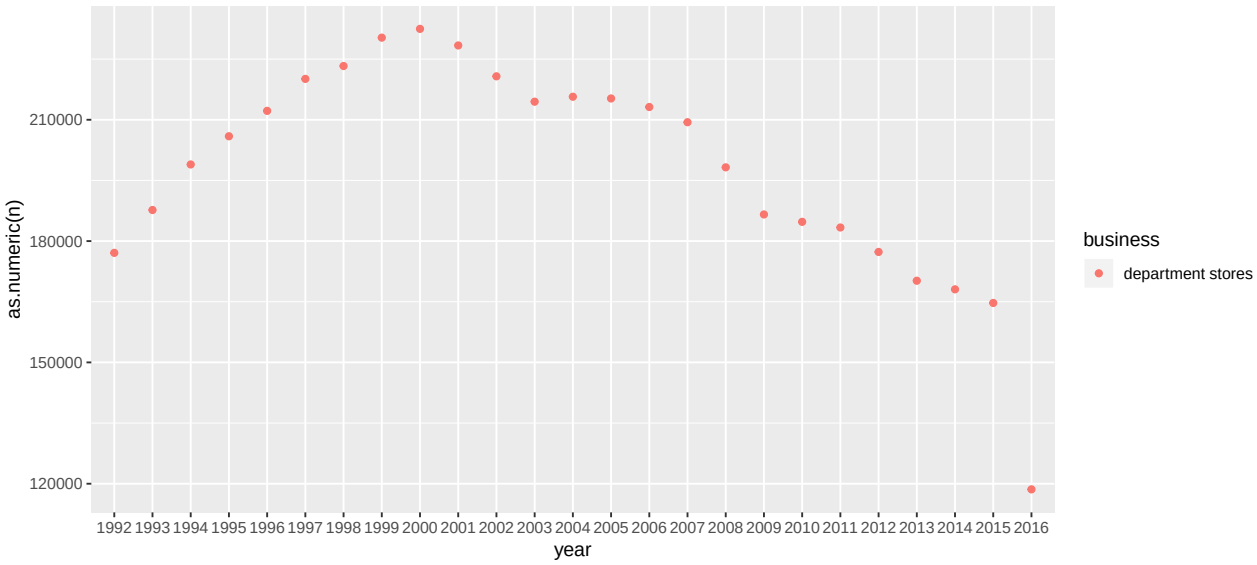
```r
## Plot Retail Sales data
ggplot(df_retail, aes(x=year,y=n,colour=business)) +
  geom_point()
```




```
## Plot Department Sales data
ggplot(df_department, aes(x=year,y=n)) +
  geom_point()
```




```
## Plot Combined Data
ggplot(df_total, aes(x=year,y=as.numeric(n), colour=business)) +
  geom_point()
```


as.numeric(n)
210000
180000
150000
120000
1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016
year
business
department stores