Chong.Rick_RC-2

Rick Chong

January 21, 2018

```
library(knitr)
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 2.2.1     v purrr   0.2.4
## v tibble  1.4.1     v dplyr   0.7.4
## v tidyr   0.7.2     v stringr 1.2.0
## v readr   1.1.1     v forcats 0.2.0
```

```
## -- Conflicts -------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readr)
library(nycflights13)
glimpse(flights)
```

```
## Observations: 336,776
## Variables: 19
## $ year           <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013,...
## $ month          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ day            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ dep_time       <int> 517, 533, 542, 544, 554, 554, 555, 557, 557, 55...
## $ sched_dep_time <int> 515, 529, 540, 545, 600, 558, 600, 600, 600, 60...
## $ dep_delay      <dbl> 2, 4, 2, -1, -6, -4, -5, -3, -3, -2, -2, -2, -2...
## $ arr_time       <int> 830, 850, 923, 1004, 812, 740, 913, 709, 838, 7...
## $ sched_arr_time <int> 819, 830, 850, 1022, 837, 728, 854, 723, 846, 7...
## $ arr_delay      <dbl> 11, 20, 33, -18, -25, 12, 19, -14, -8, 8, -2, -...
## $ carrier        <chr> "UA", "UA", "AA", "B6", "DL", "UA", "B6", "EV",...
## $ flight         <int> 1545, 1714, 1141, 725, 461, 1696, 507, 5708, 79...
## $ tailnum        <chr> "N14228", "N24211", "N619AA", "N804JB", "N668DN...
## $ origin         <chr> "EWR", "LGA", "JFK", "JFK", "LGA", "EWR", "EWR"...
## $ dest           <chr> "IAH", "IAH", "MIA", "BQN", "ATL", "ORD", "FLL"...
## $ air_time       <dbl> 227, 227, 160, 183, 116, 150, 158, 53, 140, 138...
## $ distance       <dbl> 1400, 1416, 1089, 1576, 762, 719, 1065, 229, 94...
## $ hour           <dbl> 5, 5, 5, 5, 6, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 5,...
## $ minute         <dbl> 15, 29, 40, 45, 0, 58, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ time_hour      <dttm> 2013-01-01 05:00:00, 2013-01-01 05:00:00, 2013...
```

```
glimpse(weather)
```

```
## Observations: 26,130
## Variables: 15
## $ origin     <chr> "EWR", "EWR", "EWR", "EWR", "EWR", "EWR", "EWR", "E...
## $ year       <dbl> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 201...
## $ month      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ day        <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ hour       <int> 0, 1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, ...
## $ temp       <dbl> 37.04, 37.04, 37.94, 37.94, 37.94, 39.02, 39.02, 39...
```

```
## $ dewp       <dbl> 21.92, 21.92, 21.92, 23.00, 24.08, 26.06, 26.96, 28...
## $ humid      <dbl> 53.97, 53.97, 52.09, 54.51, 57.04, 59.37, 61.63, 64...
## $ wind_dir   <dbl> 230, 230, 230, 230, 240, 270, 250, 240, 250, 260, 2...
## $ wind_speed <dbl> 10.35702, 13.80936, 12.65858, 13.80936, 14.96014, 1...
## $ wind_gust  <dbl> 11.918651, 15.891535, 14.567241, 15.891535, 17.2158...
## $ precip     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ pressure   <dbl> 1013.9, 1013.0, 1012.6, 1012.7, 1012.8, 1012.0, 101...
## $ visib      <dbl> 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10,...
## $ time_hour  <dttm> 2012-12-31 18:00:00, 2012-12-31 19:00:00, 2012-12-...
```

Question 1

```
# Q1.1 Suppose you restrict the weather and flights tables to observations from January 1, 2013. Call t
#Answer: 19366 rows

flightjan1 <- flights %>%
  filter(year==2013, month==1, day==1)
glimpse(flightjan1)
```

```
## Observations: 842
## Variables: 19
## $ year           <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013,...
## $ month          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ day            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ dep_time       <int> 517, 533, 542, 544, 554, 554, 555, 557, 557, 55...
## $ sched_dep_time <int> 515, 529, 540, 545, 600, 558, 600, 600, 600, 60...
## $ dep_delay      <dbl> 2, 4, 2, -1, -6, -4, -5, -3, -3, -2, -2, -2, -2...
## $ arr_time       <int> 830, 850, 923, 1004, 812, 740, 913, 709, 838, 7...
## $ sched_arr_time <int> 819, 830, 850, 1022, 837, 728, 854, 723, 846, 7...
## $ arr_delay      <dbl> 11, 20, 33, -18, -25, 12, 19, -14, -8, 8, -2, -...
## $ carrier        <chr> "UA", "UA", "AA", "B6", "DL", "UA", "B6", "EV",...
## $ flight         <int> 1545, 1714, 1141, 725, 461, 1696, 507, 5708, 79...
## $ tailnum        <chr> "N14228", "N24211", "N619AA", "N804JB", "N668DN...
## $ origin         <chr> "EWR", "LGA", "JFK", "JFK", "LGA", "EWR", "EWR"...
## $ dest           <chr> "IAH", "IAH", "MIA", "BQN", "ATL", "ORD", "FLL"...
## $ air_time       <dbl> 227, 227, 160, 183, 116, 150, 158, 53, 140, 138...
## $ distance       <dbl> 1400, 1416, 1089, 1576, 762, 719, 1065, 229, 94...
## $ hour           <dbl> 5, 5, 5, 5, 6, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 5,...
## $ minute         <dbl> 15, 29, 40, 45, 0, 58, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ time_hour      <dttm> 2013-01-01 05:00:00, 2013-01-01 05:00:00, 2013...
```

```
weatherjan1 <- weather %>%
  filter(year==2013, month==1, day==1)
glimpse(weatherjan1)
```

```
## Observations: 69
## Variables: 15
## $ origin     <chr> "EWR", "EWR", "EWR", "EWR", "EWR", "EWR", "EWR", "E...
## $ year       <dbl> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 201...
## $ month      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ day        <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ hour       <int> 0, 1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, ...
## $ temp       <dbl> 37.04, 37.04, 37.94, 37.94, 37.94, 39.02, 39.02, 39...
```

```
## $ dewp       <dbl> 21.92, 21.92, 21.92, 23.00, 24.08, 26.06, 26.96, 28...
## $ humid      <dbl> 53.97, 53.97, 52.09, 54.51, 57.04, 59.37, 61.63, 64...
## $ wind_dir   <dbl> 230, 230, 230, 230, 240, 270, 250, 240, 250, 260, 2...
## $ wind_speed <dbl> 10.35702, 13.80936, 12.65858, 13.80936, 14.96014, 1...
## $ wind_gust  <dbl> 11.918651, 15.891535, 14.567241, 15.891535, 17.2158...
## $ precip     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ pressure   <dbl> 1013.9, 1013.0, 1012.6, 1012.7, 1012.8, 1012.0, 101...
## $ visib      <dbl> 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10,...
## $ time_hour  <dttm> 2012-12-31 18:00:00, 2012-12-31 19:00:00, 2012-12-...
```

```
flt.tbl <- table(flightjan1$origin)
flt.wea <- table(weatherjan1$origin)
sum(flt.tbl*as.numeric(flt.wea))
```

```
## [1] 19366
```

```
#Q1.2 How many rows and columns would you obtain from executing this command? (Suggestion: Figure this
#Answer: 2,933,293,231 rows

flt.fulltbl <- table(flights$origin)
flt.fullwea <- table(weather$origin)
sum(flt.fulltbl*as.numeric(flt.fullwea))
```

```
## [1] 2933293231
```

Question 2

```
x <- read_csv('C:/Users/rick_/Desktop/winter class/Data exploration/assignment 2/dummypossmall.csv')
```

```
## Parsed with column specification:
## cols(
##   CustomerID = col_integer(),
##   Outlet = col_character(),
##   Item = col_character(),
##   price = col_double(),
##   Date = col_date(format = ""),
##   quantity = col_integer()
## )
```

```
xc1 <- x %>% complete(Date, Outlet, Item, fill=list(quantity=0))
xc2 <- x %>% complete(Date, nesting(Outlet, Item, price),
fill=list(quantity=0))

#Q2.1 You have 5 days of data. Evanston sells 5 items, Chicago sells 3. How many observations do you ha
#Answer: observations of original table = 12. Observations if include zeros
#for missing combinations will be 50 for xc1 and 40 for xc1 (using xc2
#because it is more accurate)

glimpse(x)
```

```
## Observations: 12
## Variables: 6
## $ CustomerID <int> 3, 14, 33, 5, 2, 22, 29, 27, 34, 10, 84, 42
## $ Outlet     <chr> "EV", "EV", "CH", "EV", "EV", "EV", "CH", "EV", "CH...
## $ Item       <chr> "hotdog", "hotdog", "lacroix", "sandwich", "hotdog"...
```

```
## $ price      <dbl> 5.0, 5.0, 2.0, 8.0, 5.0, 2.0, 5.5, 5.0, 2.0, 8.0, 3...
## $ Date       <date> 2017-11-06, 2017-11-08, 2017-11-07, 2017-11-06, 20...
## $ quantity   <int> 8, 2, 5, 1, 9, 3, 8, 7, 12, 8, 3, 4
```

```
glimpse(xc1)
```

```
## Observations: 50
## Variables: 6
## $ Date       <date> 2017-11-06, 2017-11-06, 2017-11-06, 2017-11-06, 20...
## $ Outlet     <chr> "CH", "CH", "CH", "CH", "CH", "EV", "EV", "EV", "EV...
## $ Item       <chr> "coffee", "hotdog", "lacroix", "muffin", "sandwich"...
## $ CustomerID <int> 34, NA, NA, NA, NA, NA, 3, NA, NA, 5, NA, NA, 33, N...
## $ price      <dbl> 2, NA, NA, NA, NA, NA, 5, NA, NA, 8, NA, NA, 2, NA,...
## $ quantity   <dbl> 12, 0, 0, 0, 0, 0, 8, 0, 0, 1, 0, 0, 5, 0, 0, 0, 9,...
```

```
glimpse(xc2)
```

```
## Observations: 40
## Variables: 6
## $ Date       <date> 2017-11-06, 2017-11-06, 2017-11-06, 2017-11-06, 20...
## $ Outlet     <chr> "CH", "CH", "CH", "EV", "EV", "EV", "EV", "EV", "CH...
## $ Item       <chr> "coffee", "hotdog", "lacroix", "coffee", "hotdog", ...
## $ price      <dbl> 2.0, 5.5, 2.0, 2.0, 5.0, 2.0, 3.0, 8.0, 2.0, 5.5, 2...
## $ CustomerID <int> 34, NA, NA, NA, 3, NA, NA, 5, NA, NA, 33, NA, 2, 22...
## $ quantity   <dbl> 12, 0, 0, 0, 8, 0, 0, 1, 0, 0, 5, 0, 9, 3, 0, 8, 0,...
```

```
#Q2.2 Describe in words the differences between xc1 and xc2. Which is correct?
#Answer: xc1 has 50 observations while xc2 has 40 observations. xc2 is
#different from xc1 due to the command nesting(Outlet, Item, price), which
#was used to find all unique combinations of Outlet, Item, price,then
#including those not found in the data, supply each variable as a separate
#argument. As a result, xc2 will have more complete data from xc1, for
#example xc2 will have hotdog price for CH on 2017-11-06, which was
#essentially obtained from 2017-11-10 hotdog sales. In xc1, as there is no
#nesting, the price is 0 which is not true.

#Q2.3 What is average daily revenue for each location?
#Answer: see table
q2.3 <- xc2 %>%
  group_by(Outlet, Date) %>%
  mutate(revenue=quantity*price) %>%
  summarize(total_rev=sum(revenue)) %>%
  summarize('average daily revenue'=mean(total_rev))

kable(q2.3)
```

Outlet	average daily revenue
CH	15.6
EV	45.0

```
#Q2.4 Create a table showing average daily revenue, by item, for each location.
#Answer: see table
q2.4 <- xc2 %>%
  group_by(Outlet, Item, Date) %>%
  mutate(revenue=quantity*price) %>%
```

```
  summarize(total_rev=sum(revenue)) %>%
  summarize('average daily revenue'=mean(total_rev)) %>%
  spread(key=Item, value='average daily revenue', fill=0)
kable(q2.4)
```

Outlet	coffee	hotdog	lacroix	muffin	sandwich
CH	4.8	8.8	2.0	0.0	0.0
EV	1.6	26.0	1.2	1.8	14.4
#Question	3				

```
library(babynames)
library(stringr)
glimpse(babynames)
```

```
## Observations: 1,858,689
## Variables: 5
## $ year <dbl> 1880, 1880, 1880, 1880, 1880, 1880, 1880, 1880, 1880, 188...
## $ sex  <chr> "F", "F", "F", "F", "F", "F", "F", "F", "F", "F", "F", "F...
## $ name <chr> "Mary", "Anna", "Emma", "Elizabeth", "Minnie", "Margaret"...
## $ n    <int> 7065, 2604, 2003, 1939, 1746, 1578, 1472, 1414, 1320, 128...
## $ prop <dbl> 0.072384329, 0.026679234, 0.020521700, 0.019865989, 0.017...
```

```
#Q3.1 How many unique babynames are there in the babynames data?
#Answer: 95,025
distinct(babynames, name)
```

```
## # A tibble: 95,025 x 1
##    name
##    <chr>
##  1 Mary
##  2 Anna
##  3 Emma
##  4 Elizabeth
##  5 Minnie
##  6 Margaret
##  7 Ida
##  8 Alice
##  9 Bertha
## 10 Sarah
## # ... with 95,015 more rows
```

```
#Q3.2 How many people in the data have the name "James"?
#Answer: 5,144,205
tmpJames <- babynames %>%
  filter(grepl(pattern="^James$", name)) %>%
  summarize(total=sum(n))

glimpse(tmpJames)
```

```
## Observations: 1
## Variables: 1
## $ total <int> 5144205
```

```
#Q3.3 How many people in the data have names beginning "Jam"?
#Answer: 5,804,740
tmpJam <- babynames %>%
  filter(grepl("^Jam", name)) %>%
  summarize(total=sum(n))

glimpse(tmpJam)
```

```
## Observations: 1
## Variables: 1
## $ total <int> 5804740
```

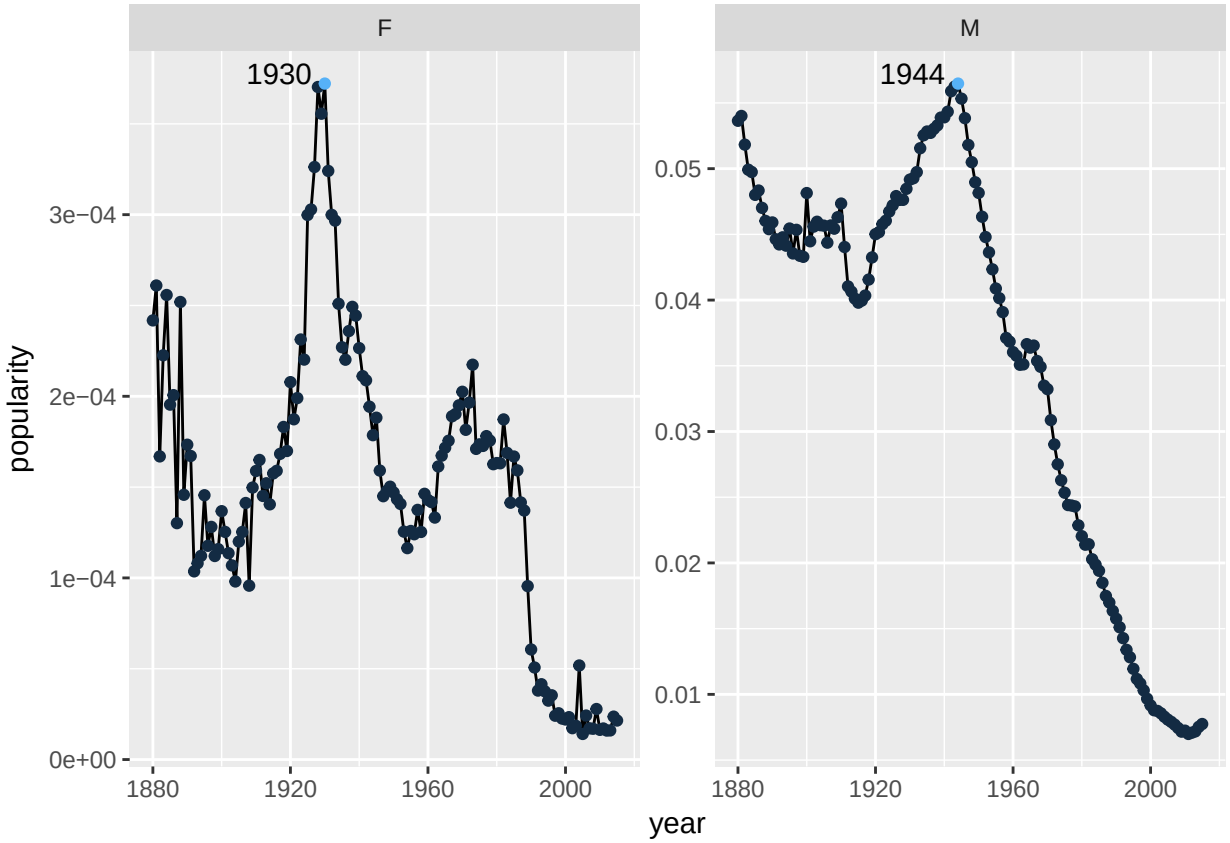
```
#Q3.4 How many people in the data have names containing "jam", ignoring case?
#Answer: 6,524,762
tmpjam <- babynames %>%
  filter(grepl("jam", name, ignore.case=TRUE)) %>%
  summarize(total=sum(n))

glimpse(tmpjam)
```

```
## Observations: 1
## Variables: 1
## $ total <int> 6524762
```

```
#Q3.5 What is the sex breakdown for people in the data whose name contains "jam" and is not "James", ig
#Answer:See table. Year peak in 1930 for female, 1944 for male

tmpsex <- babynames %>%
  filter(!grepl("^James$", name)) %>%
  filter(grepl("jam", name, ignore.case=TRUE))%>%
  group_by(sex) %>%
  summarize(total_count=sum(n))

kable(tmpsex)
```

sex	total_count
F	378056
M	1002501

```
#Q3.6 Construct a plot showing the popularity of "James" between 1880 and 2015, by sex. Define populari
#Answer: See chart

tmpsex2 <- babynames %>%
  mutate(isJames=ifelse(name=="James", n, 0)) %>%
  group_by(sex, year) %>%
  summarize(total_population=sum(n),total_James=sum(isJames),popularity=total_James/total_population)


tmpsex2 %>% mutate(color = ifelse(popularity==max(popularity), 1,0)) %>% mutate(maxyear=ifelse(popularit
```




```
#Q3.7 How many total occurrences of "James" are there considering only decadal years, i.e. 1880, 1890,
#Answer: I use an alternative method  through regex. The answer is 522642

tmpJamesDecal <- babynames %>%
  filter(grepl("0$", year)) %>%
  mutate(isJames=ifelse(name=="James", n, 0)) %>%
  summarize(total_population=sum(n),total_James=sum(isJames))
```