INFX 573: Problem Set 2 - Data Wrangling

Pierre Augustamar

Due: Monday, October 18, 2016

Collaborators: Derrick Phoebe

Instructions:

Before beginning this assignment, please ensure you have access to R and RStudio.

1. Download the `problemset2.Rmd` file from Canvas. Open `problemset2.Rmd` in RStudio and supply your solutions to the assignment by editing `problemset2.Rmd`.
2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.
3. Be sure to include well-documented (e.g. commented) code chucks, figures and clearly written text chunk explanations as necessary. Any figures should be clearly labeled and appropriately referenced within the text.
4. Collaboration on problem sets is acceptable, and even encouraged, but each student must turn in an individual write-up in his or her own words and his or her own work. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.
5. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit PDF`, rename the R Markdown file to `YourLastName_YourFirstName_ps2.Rmd`, knit a PDF and submit the PDF file on Canvas.

Setup:

In this problem set you will need, at minimum, the following R packages.

```r
# Load standard libraries
library(tidyverse)
library(nycflights13)
library(jsonlite)
```

Problem 1: Open Government Data

Use the following code to obtain data on the Seattle Police Department Police Report Incidents.

```r
police_incidents <- fromJSON("https://data.seattle.gov/resource/7ais-f98f.json")
dim(police_incidents)  # Determine the number of rows and columns
```

```
## [1] 1000   19
```

```r
head(police_incidents) # Check out the first five rows of the data
```

```
##   offense_code_extension   offense_type general_offense_number
## 1                      0         EQUALS             2016239258
## 2                      0   ASSLT-NONAGG             2016340018
## 3                      1 VEH-THEFT-AUTO             2016340045
## 4                      0 THEFT-SHOPLIFT             2016339816
## 5                      0   ASSLT-NONAGG             2016339898
## 6                      1 VEH-THEFT-AUTO             2016339682
##   offense_code rms_cdw_id year zone_beat     latitude
## 1         2903     949463 <NA>      <NA>         <NA>
## 2         1313    1038931 2016        E2 47.615837097
## 3         2404    1038930 2016        U1 47.667503357
## 4         2303    1038854 2016        L3 47.721984863
## 5         1313    1038866 2016        U2 47.659805298
## 6         2404    1038799 2016        N1 47.700145721
##   summarized_offense_description       date_reported
## 1                           <NA>                <NA>
## 2                        ASSAULT 2016-09-19T14:25:00
## 3                  VEHICLE THEFT 2016-09-19T13:21:00
## 4                    SHOPLIFTING 2016-09-19T12:14:00
## 5                        ASSAULT 2016-09-19T11:33:00
## 6                  VEHICLE THEFT 2016-09-19T10:19:00
##   occurred_date_or_date_range_start summary_offense_code month
## 1                              <NA>                 <NA>  <NA>
## 2               2016-09-19T13:00:00                 1300     9
## 3               2016-09-18T15:00:00                 2400     9
## 4               2016-09-19T10:12:00                 2300     9
## 5               2016-09-19T11:33:00                 1300     9
## 6               2016-09-17T17:00:00                 2400     9
##   census_tract_2000 location.latitude location.needs_recoding
## 1              <NA>              <NA>                      NA
## 2         7500.5009      47.615837097                   FALSE
## 3         4400.4003      47.667503357                   FALSE
## 4          100.5005      47.721984863                   FALSE
## 5         5301.3002      47.659805298                   FALSE
## 6         1400.3013      47.700145721                   FALSE
##   location.longitude         hundred_block_location district_sector
## 1               <NA>                           <NA>            <NA>
## 2      -122.31816864            16XX BLOCK OF 11 AV               E
## 3     -122.315200806         52XX BLOCK OF 12 AV NE               U
## 4     -122.293640137 127XX BLOCK OF LAKE CITY WY NE               L
## 5     -122.314323425      NE 43 ST / BROOKLYN AV NE               U
## 6     -122.366722107          8XX BLOCK OF NW 97 ST               N
##        longitude occurred_date_range_end
## 1           <NA>                    <NA>
## 2 -122.318168640                    <NA>
## 3 -122.315200806     2016-09-19T13:00:00
## 4 -122.293640137                    <NA>
## 5 -122.314323425                    <NA>
## 6 -122.366722107     2016-09-19T07:00:00
```

```r
str(police_incidents)  # Check out the structure of each objects
```

```
## 'data.frame':    1000 obs. of  19 variables:
##  $ offense_code_extension           : chr  "0" "0" "1" "0" ...
##  $ offense_type                     : chr  "EQUALS" "ASSLT-NONAGG" "VEH-THEFT-AUTO" "THEFT-SHOPLIFT"
##  $ general_offense_number           : chr  "2016239258" "2016340018" "2016340045" "2016339816" ...
##  $ offense_code                     : chr  "2903" "1313" "2404" "2303" ...
##  $ rms_cdw_id                       : chr  "949463" "1038931" "1038930" "1038854" ...
##  $ year                             : chr  NA "2016" "2016" "2016" ...
##  $ zone_beat                        : chr  NA "E2" "U1" "L3" ...
##  $ latitude                         : chr  NA "47.615837097" "47.667503357" "47.721984863" ...
##  $ summarized_offense_description   : chr  NA "ASSAULT" "VEHICLE THEFT" "SHOPLIFTING" ...
##  $ date_reported                    : chr  NA "2016-09-19T14:25:00" "2016-09-19T13:21:00" "2016-09-1
##  $ occurred_date_or_date_range_start: chr  NA "2016-09-19T13:00:00" "2016-09-18T15:00:00" "2016-09-1
##  $ summary_offense_code             : chr  NA "1300" "2400" "2300" ...
##  $ month                            : chr  NA "9" "9" "9" ...
##  $ census_tract_2000                : chr  NA "7500.5009" "4400.4003" "100.5005" ...
##  $ location                         :'data.frame':   1000 obs. of  3 variables:
##   ..$ latitude      : chr  NA "47.615837097" "47.667503357" "47.721984863" ...
##   ..$ needs_recoding: logi  NA FALSE FALSE FALSE FALSE FALSE ...
##   ..$ longitude     : chr  NA "-122.31816864" "-122.315200806" "-122.293640137" ...
##  $ hundred_block_location           : chr  NA "16XX BLOCK OF 11 AV" "52XX BLOCK OF 12 AV NE" "127XX
##  $ district_sector                  : chr  NA "E" "U" "L" ...
##  $ longitude                        : chr  NA "-122.318168640" "-122.315200806" "-122.293640137" ...
##  $ occurred_date_range_end          : chr  NA NA "2016-09-19T13:00:00" NA ...
```

(a) Describe, in detail, what the data represents.

The data represent offenses reported by officers from the Seattle Police Department when dispatched to investigate a crime reported in the city of Seattle.

(b) Describe each variable and what it measures. Be sure to note when data is missing. Confirm that each variable is appropriately cast - it has the correct data type. If any are incorrect, recast them to be in the appropriate format.

- offense_code_extension - A unique extension code that identifies a particular offense.
- offense_type - A value that identifies the related offense. tar
- general_offense_number - A unique number identifies general offense.
- offense_code - A single code for an offense.
- rms_cdw_id - a number identifies a residential location
- year - The year that the offense was recorded
- zone_beat - Areas that individual patrol officers are assigned responsibility for.
- latitude - Angles that uniquely define a location
- summarized_offense_description - A summary of the offense which may include loitering, disorderly conduct, harassment among others.
- date_reported - The date that a crime was reported to the police.
- occurred_date_or_date_range_start - Approximate date and or date range that a crime started.
- summary_offense_code - a classification of offenses that are defined by a related code.
- month - The month that the offense was recorded
- census_tract_2000 - A permanent geographical entities within a county that identify areas of crime being reported.
- location - A combination of longitude and latitudes coordinates to identify a specific place
- hundred_block_location - A location of a crime about a grid's map from the city center
- district_sector - A mapping on how police precincts are organized.
- longitude - Angles that uniquely define a location
- occurred_date_range_end - Approximate date and or date range that a crime started.

Response to incorrect data type: * offense_code_extension - No change * offense_type - No change * general_offense_number - No change * offense_code - No change * rms_cdw_id - Change to an integer * year - I would change character to numerice value * zone_beat - No change * latitude - I would change this to a floating number * summarized_offense_description - No change * date_reported _ I would change this to a date time * occurred_date_or_date_range_start - I would change this to a date time * summary_offense_code - type is fine * month - I would change character to numerice value * census_tract_2000 - No change * location - No change * hundred_block_location - No change * district_sector - No change * longitude - I would change this to a floating value * occurred_date_range_end - I would change this to a date type

Response to recast data type:

```
 # change the data type for year, month, latitude, longitude, date_reported, occurred_date_or_date_rang

recastData = transform(police_incidents, year=as.integer(year), month=as.integer(month),
                      latitude=as.double(latitude), longitude=as.double(longitude),
                      date_reported=as.data.frame.Date(date_reported),
                      occurred_date_or_date_range_start=as.data.frame.Date(occurred_date_or_date_range_s
                      occurred_date_range_end=as.data.frame.Date(occurred_date_range_end))

str(recastData)
```

```
## 'data.frame':    1000 obs. of  19 variables:
##  $ offense_code_extension           : chr  "0" "0" "1" "0" ...
##  $ offense_type                     : chr  "EQUALS" "ASSLT-NONAGG" "VEH-THEFT-AUTO" "THEFT-SHOPLIFT"
##  $ general_offense_number           : chr  "2016239258" "2016340018" "2016340045" "2016339816" ...
##  $ offense_code                     : chr  "2903" "1313" "2404" "2303" ...
##  $ rms_cdw_id                       : chr  "949463" "1038931" "1038930" "1038854" ...
##  $ year                             : int  NA 2016 2016 2016 2016 2016 2016 2016 2016 2016 ...
##  $ zone_beat                        : chr  NA "E2" "U1" "L3" ...
##  $ latitude                         : num  NA 47.6 47.7 47.7 47.7 ...
##  $ summarized_offense_description   : chr  NA "ASSAULT" "VEHICLE THEFT" "SHOPLIFTING" ...
##  $ date_reported                    :'data.frame':   1000 obs. of  1 variable:
##   ..$ date_reported: chr  NA "2016-09-19T14:25:00" "2016-09-19T13:21:00" "2016-09-19T12:14:00" ...
##  $ occurred_date_or_date_range_start:'data.frame':   1000 obs. of  1 variable:
##   ..$ occurred_date_or_date_range_start: chr  NA "2016-09-19T13:00:00" "2016-09-18T15:00:00" "2016-0
##  $ summary_offense_code             : chr  NA "1300" "2400" "2300" ...
##  $ month                            : int  NA 9 9 9 9 9 9 9 9 9 ...
##  $ census_tract_2000                : chr  NA "7500.5009" "4400.4003" "100.5005" ...
##  $ location                         :'data.frame':   1000 obs. of  3 variables:
##   ..$ latitude      : chr  NA "47.615837097" "47.667503357" "47.721984863" ...
##   ..$ needs_recoding: logi  NA FALSE FALSE FALSE FALSE FALSE ...
##   ..$ longitude     : chr  NA "-122.31816864" "-122.315200806" "-122.293640137" ...
##  $ hundred_block_location           : chr  NA "16XX BLOCK OF 11 AV" "52XX BLOCK OF 12 AV NE" "127XX 
##  $ district_sector                  : chr  NA "E" "U" "L" ...
##  $ longitude                        : num  NA -122 -122 -122 -122 ...
##  $ occurred_date_range_end          :'data.frame':   1000 obs. of  1 variable:
##   ..$ occurred_date_range_end: chr  NA NA "2016-09-19T13:00:00" NA ...
```

```
head(recastData)
```

```
##   offense_code_extension   offense_type general_offense_number
## 1                      0         EQUALS             2016239258
## 2                      0   ASSLT-NONAGG             2016340018
## 3                      1 VEH-THEFT-AUTO             2016340045
## 4                      0 THEFT-SHOPLIFT             2016339816
## 5                      0   ASSLT-NONAGG             2016339898
## 6                      1 VEH-THEFT-AUTO             2016339682
##   offense_code rms_cdw_id year zone_beat latitude
## 1         2903     949463   NA      <NA>       NA
## 2         1313    1038931 2016        E2 47.61584
## 3         2404    1038930 2016        U1 47.66750
## 4         2303    1038854 2016        L3 47.72198
## 5         1313    1038866 2016        U2 47.65981
## 6         2404    1038799 2016        N1 47.70015
##   summarized_offense_description       date_reported
## 1                           <NA>                <NA>
## 2                        ASSAULT 2016-09-19T14:25:00
## 3                  VEHICLE THEFT 2016-09-19T13:21:00
## 4                    SHOPLIFTING 2016-09-19T12:14:00
## 5                        ASSAULT 2016-09-19T11:33:00
## 6                  VEHICLE THEFT 2016-09-19T10:19:00
##   occurred_date_or_date_range_start summary_offense_code month
## 1                              <NA>                 <NA>    NA
## 2               2016-09-19T13:00:00                 1300     9
## 3               2016-09-18T15:00:00                 2400     9
## 4               2016-09-19T10:12:00                 2300     9
## 5               2016-09-19T11:33:00                 1300     9
## 6               2016-09-17T17:00:00                 2400     9
##   census_tract_2000 location.latitude location.needs_recoding
## 1              <NA>              <NA>                      NA
## 2         7500.5009      47.615837097                   FALSE
## 3         4400.4003      47.667503357                   FALSE
## 4          100.5005      47.721984863                   FALSE
## 5         5301.3002      47.659805298                   FALSE
## 6         1400.3013      47.700145721                   FALSE
##   location.longitude         hundred_block_location district_sector
## 1               <NA>                           <NA>            <NA>
## 2      -122.31816864            16XX BLOCK OF 11 AV               E
## 3     -122.315200806         52XX BLOCK OF 12 AV NE               U
## 4     -122.293640137 127XX BLOCK OF LAKE CITY WY NE               L
## 5     -122.314323425      NE 43 ST / BROOKLYN AV NE               U
## 6     -122.366722107          8XX BLOCK OF NW 97 ST               N
##   longitude occurred_date_range_end
## 1        NA                    <NA>
## 2 -122.3182                    <NA>
## 3 -122.3152     2016-09-19T13:00:00
## 4 -122.2936                    <NA>
## 5 -122.3143                    <NA>
## 6 -122.3667     2016-09-19T07:00:00
```

(c) Produce a clean dataset, according to the rules of tidy data discussed in class. Export the data for future analysis using the Rdata format.

```
#Since the data can not be pivoted by using gather or spread, I will retrieve key columns and leave out

# Select a subset or relevant information as well as rename some of the variables
incidents <- police_incidents %>%
            select(date = date_reported,
                  location=hundred_block_location,
                  offense = offense_code,
                  category=offense_type,
                  long= longitude,
                  lat=latitude)

# Filter incidens with missing date
incidents <- incidents %>% filter(!is.na(date))

# Filter incidents with missing latitude
incidents <- incidents %>% filter(!is.na(lat))

# Filter incidents with missing locations
incidents <- incidents %>% filter(!is.na(location))


# save as an .rds file
saveRDS(incidents, file="incidents_data.rds")
```

(d) Describe any concerns you might have about this data. This may include biases, missing data, or ethical concerns.

Response concerns:

Seattle police department has some useful information. However, I have noted that some columns have NA values or an x value. These are questionable values that made it difficult to understand the reasoning for not having accurate information. For instance, there were numerous records with an offense code/Summary code = "x" that does not make any sense. Does this mean that there are offenses that have no related offense code? Also, there are a number of entries with no occurred date range. Again, this seems odd that there will be incidents with no date range. It will be a nigtmare to investiage those incidents with proper information. when the offenses.

Also, the report contains block location as well as longituge, and latitude. These are prime information that advertisers use to target a specific demographic.Thus, it will not be impossible that this open to door to discrimination because of high rate of crimes and on the other hand, marketers coulde use this information to sell houses for areas that are not listed as high crimes.

Problem 2: Wrangling the NYC Flights Data

In this problem set we will use the data on all flights that departed NYC (i.e. JFK, LGA or EWR) in 2013. You can find this data in the `nycflights13` R package.

(a) Importing Data:

Load the data.

```
# Edit me.
str(flights)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    336776 obs. of  19 variables:
##  $ year          : int  2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 ...
##  $ month         : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ day           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ dep_time      : int  517 533 542 544 554 554 555 557 557 558 ...
##  $ sched_dep_time: int  515 529 540 545 600 558 600 600 600 600 ...
##  $ dep_delay     : num  2 4 2 -1 -6 -4 -5 -3 -3 -2 ...
##  $ arr_time      : int  830 850 923 1004 812 740 913 709 838 753 ...
##  $ sched_arr_time: int  819 830 850 1022 837 728 854 723 846 745 ...
##  $ arr_delay     : num  11 20 33 -18 -25 12 19 -14 -8 8 ...
##  $ carrier       : chr  "UA" "UA" "AA" "B6" ...
##  $ flight        : int  1545 1714 1141 725 461 1696 507 5708 79 301 ...
##  $ tailnum       : chr  "N14228" "N24211" "N619AA" "N804JB" ...
##  $ origin        : chr  "EWR" "LGA" "JFK" "JFK" ...
##  $ dest          : chr  "IAH" "IAH" "MIA" "BQN" ...
##  $ air_time      : num  227 227 160 183 116 150 158 53 140 138 ...
##  $ distance      : num  1400 1416 1089 1576 762 ...
##  $ hour          : num  5 5 5 5 6 5 6 6 6 6 ...
##  $ minute        : num  15 29 40 45 0 58 0 0 0 0 ...
##  $ time_hour     : POSIXct, format: "2013-01-01 05:00:00" "2013-01-01 05:00:00" ...
```

```r
str(airports)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1396 obs. of  7 variables:
##  $ faa : chr  "04G" "06A" "06C" "06N" ...
##  $ name: chr  "Lansdowne Airport" "Moton Field Municipal Airport" "Schaumburg Regional" "Randall Air
##  $ lat : num  41.1 32.5 42 41.4 31.1 ...
##  $ lon : num  -80.6 -85.7 -88.1 -74.4 -81.4 ...
##  $ alt : int  1044 264 801 523 11 1593 730 492 1000 108 ...
##  $ tz  : num  -5 -5 -6 -5 -4 -4 -5 -5 -5 -8 ...
##  $ dst : chr  "A" "A" "A" "A" ...
```

```r
str(airlines)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    16 obs. of  2 variables:
##  $ carrier: chr  "9E" "AA" "AS" "B6" ...
##  $ name   : chr  "Endeavor Air Inc." "American Airlines Inc." "Alaska Airlines Inc." "JetBlue Airway
```

```r
str(planes)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    3322 obs. of  9 variables:
##  $ tailnum     : chr  "N10156" "N102UW" "N103US" "N104UW" ...
##  $ year        : int  2004 1998 1999 1999 2002 1999 1999 1999 1999 1999 ...
##  $ type        : chr  "Fixed wing multi engine" "Fixed wing multi engine" "Fixed wing multi engine"
##  $ manufacturer: chr  "EMBRAER" "AIRBUS INDUSTRIE" "AIRBUS INDUSTRIE" "AIRBUS INDUSTRIE" ...
##  $ model       : chr  "EMB-145XR" "A320-214" "A320-214" "A320-214" ...
##  $ engines     : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ seats       : int  55 182 182 182 55 182 182 182 182 182 ...
##  $ speed       : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ engine      : chr  "Turbo-fan" "Turbo-fan" "Turbo-fan" "Turbo-fan" ...
```

(b) Data Manipulation:

Use the flights data to answer each of the following questions. Be sure to answer each question with a written response and supporting analysis.

- How many flights were there from NYC airports to Seattle in 2013?

```
# Count each of the destination flights to Seattl
flights %>%
  select(dest) %>%
  filter(dest == "SEA") %>%
  summarise(NY_To_Sea_Total_Flights= n())
```

```
## # A tibble: 1 × 1
##   NY_To_Sea_Total_Flights
##                     <int>
## 1                    3923
```

- How many airlines fly from NYC to Seattle?

```
# Edit me.
   mergeAirlinesAndFlights =  merge(flights, airlines, by="carrier")
    mergeAirlinesAndFlights %>%
        select(dest) %>%
        filter (dest == "SEA") %>%
        summarise(total_Airlines=n())
```

```
##   total_Airlines
## 1           3923
```

- How many unique air planes fly from NYC to Seattle?

```
# Edit me.
mergeAirlinesAndFlights =  merge(flights, airlines, by="carrier")
mergeAirlinesAndFlights %>%
    select(dest, carrier) %>%
    filter (dest == "SEA") %>%
    summarise(unique_Carrier_Count =n_distinct(carrier))
```

```
##   unique_Carrier_Count
## 1                    5
```

- What is the average arrival delay for flights from NYC to Seattle?

```
# Edit me.
flights %>%
  select(origin, dest) %>%
  filter(dest == "SEA") %>%
  mutate(average_Flights_Delay_To_Seattle = mean(flights$arr_delay, na.rm=TRUE))
```

```
## # A tibble: 3,923 × 3
##    origin  dest average_Flights_Delay_To_Seattle
##     <chr> <chr>                            <dbl>
## 1     EWR   SEA                         6.895377
## 2     JFK   SEA                         6.895377
## 3      EWR    SEA                           6.895377
## 4      EWR    SEA                           6.895377
## 5      JFK    SEA                           6.895377
## 6      EWR    SEA                           6.895377
## 7      EWR    SEA                           6.895377
## 8      JFK    SEA                           6.895377
## 9      JFK    SEA                           6.895377
## 10     EWR    SEA                           6.895377
## # ... with 3,913 more rows
```

- What proportion of flights to Seattle come from each NYC airport?

```
# Edit me.
m =  flights %>%
       select(origin, dest) %>%
       filter(dest == "SEA") %>%
       group_by(origin) %>%
       tally()

m %>%
  mutate((proportion = m$n/sum(m$n)*100))
```

```
## # A tibble: 2 × 3
##   origin     n `(proportion = m$n/sum(m$n) * 100)`
##    <chr> <int>                               <dbl>
## 1    EWR  1831                            46.67346
## 2    JFK  2092                            53.32654
```