Data Extraction and Manipulation

Divya Krishnan

Monday, October 19, 2015

Data Extraction and Manipulation

```r
# Stardard libraries
library(jsonlite)
library(dplyr)
library(ggplot2)
library(tidyr)
library(RSocrata)
library(acs)
# Explicitly added the package of reshape2 for melt function
library(reshape2)
```

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

```r
# Importing the JSON file from URL
policeIncidents <- fromJSON("https://data.seattle.gov/resource/7ais-f98f.json")
# Exploring the dataset
head(policeIncidents)
```

```
##   offense_code             offense_type census_tract_2000
## 1            X          DISTURBANCE-OTH          100.3004
## 2         1313             ASSLT-NONAGG        10702.3009
## 3         2404           VEH-THEFT-AUTO         3100.5011
## 4         5707                 TRESPASS         7200.1079
## 5         1206 ROBBERY-STREET-BODYFORCE         8100.3011
## 6            X          DISTURBANCE-OTH         8400.1004
##         date_reported location.needs_recoding location.longitude
## 1 2016-02-21T08:58:00                   FALSE     -122.290794373
## 2 2016-02-21T08:36:00                   FALSE     -122.370819092
## 3 2016-02-21T08:09:00                   FALSE     -122.398323059
## 4 2016-02-21T07:15:00                   FALSE     -122.349639893
## 5 2016-02-21T07:11:00                   FALSE     -122.342880249
## 6 2016-02-21T06:14:00                   FALSE     -122.325515747
##   location.latitude occurred_date_range_end zone_beat
## 1      47.720947266     2016-02-21T08:58:00        L3
## 2      47.543205261                    <NA>        W3
## 3      47.688655853     2016-02-21T08:00:00        J2
## 4      47.618579865                    <NA>        Q3
## 5       47.60981369                    <NA>        M1
## 6      47.614086151                    <NA>        E3
##   offense_code_extension district_sector     hundred_block_location
## 1                     21               L    125XX BLOCK OF 35 AV NE
## 2                      0               W 65XX BLOCK OF SYLVAN WY SW
## 3                      1               J     83XX BLOCK OF 32 AV NW
## 4                      0               Q      3XX BLOCK OF DENNY WY
## 5                      0               M   19XX BLOCK OF WESTERN AV
## 6                     21               E      SUMMIT AV / E PIKE ST
##   summarized_offense_description month general_offense_number year
## 1                    DISTURBANCE     2              201662362 2016
## 2                        ASSAULT     2              201662321 2016
## 3                  VEHICLE THEFT     2              201662337 2016
## 4                       TRESPASS     2              201662300 2016
## 5                        ROBBERY     2              201662295 2016
## 6                    DISTURBANCE     2              201662265 2016
##        longitude summary_offense_code     latitude rms_cdw_id
## 1 -122.290794373                    X 47.720947266     644277
## 2 -122.370819092                 1300 47.543205261     644273
## 3 -122.398323059                 2400 47.688655853     644307
## 4 -122.349639893                 5700 47.618579865     644306
## 5 -122.342880249                 1200 47.609813690     644251
## 6 -122.325515747                    X 47.614086151     644210
##   occurred_date_or_date_range_start
## 1               2016-02-21T08:00:00
## 2               2016-02-21T08:36:00
## 3               2016-02-20T20:40:00
## 4               2016-02-21T07:15:00
## 5               2016-02-21T07:00:00
## 6               2016-02-21T04:24:00
```

```r
# Getting the dmensions of the dataset
dim(policeIncidents)
```

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

```r
# Exploring the variables
colnames(policeIncidents)
```

```
##  [1] "offense_code"
##  [2] "offense_type"
##  [3] "census_tract_2000"
##  [4] "date_reported"
##  [5] "location"
##  [6] "occurred_date_range_end"
##  [7] "zone_beat"
##  [8] "offense_code_extension"
##  [9] "district_sector"
## [10] "hundred_block_location"
## [11] "summarized_offense_description"
## [12] "month"
## [13] "general_offense_number"
## [14] "year"
## [15] "longitude"
## [16] "summary_offense_code"
## [17] "latitude"
## [18] "rms_cdw_id"
## [19] "occurred_date_or_date_range_start"
```

(a) Describe, in detail, what the data represents. The data represents initial police reports taken down by police officers when responding to incidents around Seattle (Seattle Police Department Police Report Incident, 2010). The datset has 1000 observations and 19 variables.
References -
[1] Seattle Police Department Police Report Incident | Data.Seattle.Gov | Seattle's Data Site. (2010, July 28). Retrieved October 18, 2015, from https://data.seattle.gov/Public-Safety/Seattle-Police-Department-Police-Report-Incident/7ais-f98f

(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. The data set has the following variables -
[1] “offense_code” “offense_type” “census_tract_2000” “date_reported”
[5] “location” “zone_beat” “offense_code_extension” “district_sector”
[9] “hundred_block_location” “summarized_offense_description” “month” “general_offense_number”
[13] “year” “longitude” “summary_offense_code” “latitude”
[17] “rms_cdw_id” “occurred_date_or_date_range_start” “occurred_date_range_end”

Description for the variables - The data set has the following variables -
[1] “offense_code” - The offense code for the police incident.
[2] “offense_type” - The type of offense such as identity theft, disturbance, burgulary etc.
[3] “census_tract_2000” - This is the census tract 2000 data.
[4] “date_reported” - The date the incident was reported on.
[5] “location” - The data frame having the information such as needs_recoding, longitude and latitude.
[6] “zone_beat” - The different zones that Seattle city is categorized into.
[7] “offense_code_extension” - Different code extensions to the offense.
[8] “district_sector” - The district sector of the incident.

[9] “hundred_block_location” - The hundred block region of the incident.
[10] “summarized_offense_description” - The summarized description of the offense. For example the offense type of “ASSLT-NONAGG” & “ASSLT-AGG-WEAPON” are categorized as “ASSAULT” in this variable. This variable gives a more broader category for the offense.
[11] “month” - The month that the incident occured.
[12] “general_offense_number” - The general offense number given to an incident.

[13] “year” - The year that the incident occured.
[14] “longitude” - The longitude measurement of the location of the incident.
[15] “summary_offense_code” - The summarized offense code of the incident.
[16] “latitude” - The latitude measurement of the location of the incident.
[17] “rms_cdw_id” - Unique row identifier (Seattle Police Department Police Report Incident, 2010).
[18] “occurred_date_or_date_range_start” - The date the incident occured or started. [19] “occurred_date_range_end” - The date the incident ended or the report was closed.
The rms_cdw_id is the unique code identifying each record. The following variables identify different types details about the offense, such as different offense codes and its description -
[1] “offense_code” “offense_type” “offense_code_extension” “summarized_offense_description” [5] “general_offense_number” “summary_offense_code”

The dataset also has variables describing the date and time of the offenses [1] “date_reported” “occurred_date_or_date_range_start” “occurred_date_range_end”

All the variables are in character data type. Though character datatype makes sense for most variables, but some variables need to be recast as numeric due to the nature of value recorded.

Recasting of the variables are as follows -
All the variables except the ones below have the appropriate datatype. The following variables need to be recasted in the format:Variable - Existing datatype - To recasted datatype
policeIncidents$location$longitude - character - numeric
policeIncidents$location$latitude - character - numeric
policeIncidents$longitude - character - numeric
policeIncidents$latitude - character - numeric
policeIncidents$month - character - numeric
policeIncidents$year - character - numeric

Missing Data -
1. offense_code - This variable has values of 'X', which are probably refer to missing data.
2. summary_offense_code - This variable has values of 'X', which are probably refer to missing data.
3. occured_date_range_end - This variable has NA values signifying missing data.
References -
[1] Seattle Police Department Police Report Incident | Data.Seattle.Gov | Seattle's Data Site. (2010, July 28). Retrieved October 18, 2015, from https://data.seattle.gov/Public-Safety/Seattle-Police-Department-Police-Report-Incident/7ais-f98f

```r
# Exploring the variables
colnames(policeIncidents)
```

```
##  [1] "offense_code"
##  [2] "offense_type"
##  [3] "census_tract_2000"
##  [4] "date_reported"
##  [5] "location"
##  [6] "occurred_date_range_end"
##  [7] "zone_beat"
##  [8] "offense_code_extension"
##  [9] "district_sector"
## [10] "hundred_block_location"
## [11] "summarized_offense_description"
## [12] "month"
## [13] "general_offense_number"
## [14] "year"
## [15] "longitude"
## [16] "summary_offense_code"
## [17] "latitude"
## [18] "rms_cdw_id"
## [19] "occurred_date_or_date_range_start"
```

```r
# Summary of each variable
summary(policeIncidents)
```

```
##  offense_code       offense_type       census_tract_2000 
##  Length:1000        Length:1000        Length:1000       
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##  date_reported
##  Length:1000
##  Class :character
##  Mode  :character
##  location.needs_recoding  location.longitude  location.latitude
##  Mode :logical       Length:1000         Length:1000
##  FALSE:1000          Class :character    Class :character
##  NA's :0             Mode  :character    Mode  :character
##  occurred_date_range_end  zone_beat         offense_code_extension
##  Length:1000             Length:1000        Length:1000
##  Class :character        Class :character   Class :character
##  Mode  :character        Mode  :character   Mode  :character
##  district_sector    hundred_block_location summarized_offense_description
##  Length:1000        Length:1000            Length:1000
##  Class :character   Class :character       Class :character
##  Mode  :character   Mode  :character       Mode  :character
##     month           general_offense_number     year
##  Length:1000        Length:1000            Length:1000
##  Class :character   Class :character       Class :character
##  Mode  :character   Mode  :character       Mode  :character
##   longitude         summary_offense_code   latitude
##  Length:1000        Length:1000          Length:1000
##  Class :character   Class :character     Class :character
##  Mode  :character   Mode  :character     Mode  :character
##   rms_cdw_id        occurred_date_or_date_range_start
##  Length:1000        Length:1000
##  Class :character   Class :character
##  Mode  :character   Mode  :character
```

```
# Exploring different types of variable category
names(select(policeIncidents,contains("offense")))
```

```
## [1] "offense_code"                   "offense_type"
## [3] "offense_code_extension"         "summarized_offense_description"
## [5] "general_offense_number"         "summary_offense_code"
```

```
names(select(policeIncidents,contains("date")))
```

```
## [1] "date_reported"                     "occurred_date_range_end"
## [3] "occurred_date_or_date_range_start"
```

```
# Exploring the class and mode of the variables
sapply(policeIncidents,class)
```

```
##                       offense_code                      offense_type
##                        "character"                       "character"
##                  census_tract_2000                     date_reported
##                        "character"                       "character"
##                           location           occurred_date_range_end
##                       "data.frame"                       "character"
##                          zone_beat            offense_code_extension
##                        "character"                       "character"
##                    district_sector            hundred_block_location
##                       "character"                       "character"
##    summarized_offense_description                             month
##                       "character"                       "character"
##            general_offense_number                              year
##                       "character"                       "character"
##                         longitude              summary_offense_code
##                       "character"                       "character"
##                          latitude                        rms_cdw_id
##                       "character"                       "character"
## occurred_date_or_date_range_start
##                       "character"
```

```
sapply(policeIncidents,mode)
```

```
##                      offense_code                      offense_type
##                       "character"                       "character"
##                 census_tract_2000                     date_reported
##                       "character"                       "character"
##                          location           occurred_date_range_end
##                            "list"                       "character"
##                         zone_beat            offense_code_extension
##                       "character"                       "character"
##                   district_sector            hundred_block_location
##                       "character"                       "character"
##    summarized_offense_description                             month
##                       "character"                       "character"
##            general_offense_number                              year
##                       "character"                       "character"
##                         longitude              summary_offense_code
##                       "character"                       "character"
##                          latitude                        rms_cdw_id
##                       "character"                       "character"
## occurred_date_or_date_range_start
##                       "character"
```

```
# Exploring the location data frame
names(policeIncidents$location)
```

```
## [1] "needs_recoding" "longitude"      "latitude"
```

```
# Recasting some of the variables from character to numeric data type
policeIncidents$location$longitude<-as.numeric(policeIncidents$location$longitude)
policeIncidents$location$latitude<-as.numeric(policeIncidents$location$latitude)
policeIncidents$longitude<-as.numeric(policeIncidents$longitude)
policeIncidents$latitude<-as.numeric(policeIncidents$latitude)
policeIncidents$month<-as.numeric(policeIncidents$month)
policeIncidents$year<-as.numeric(policeIncidents$year)
# Recasting date variables to Date datatype
policeIncidents$date_reported<-as.Date(policeIncidents$date_reported)
# Recasting date variables to Date datatype after formatting them using sub function
policeIncidents$occurred_date_or_date_range_start<-as.Date(sub("T"," ",policeIncidents$occurred_date_or
policeIncidents$occurred_date_range_end<-as.Date(sub("T"," ",policeIncidents$occurred_date_range_end))
```

```r
#View(policeIncidents)
```

```r
# Copying the recasted policeIncidents dataset to tidy it
policeIncidents.tidy<-policeIncidents

#Checking the needs_recording column inside location data frame
summary(policeIncidents.tidy[,"location"]["needs_recoding"])
```

(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.

```
##  needs_recoding
##  Mode :logical
##  FALSE:1000
##  NA's :0
```

```r
#Extracting latitude and longitude from the data frame location embedded in policeIncidents
policeIncidents.tidy$location_longitude<-policeIncidents.tidy[,"location"]["longitude"]
policeIncidents.tidy$location_latitude<-policeIncidents.tidy[,"location"]["latitude"]
# Removing location dataframe
policeIncidents.tidy<-subset(policeIncidents.tidy,select=-location)

summary(policeIncidents.tidy)
```

```
##  offense_code       offense_type       census_tract_2000
##  Length:1000        Length:1000        Length:1000
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  date_reported        occurred_date_range_end  zone_beat
##  Min.   :2016-02-13   Min.   :2015-11-11      Length:1000
##  1st Qu.:2016-02-14   1st Qu.:2016-02-13      Class :character
##  Median :2016-02-16   Median :2016-02-15      Mode  :character
##  Mean   :2016-02-16   Mean   :2016-02-14
##  3rd Qu.:2016-02-18   3rd Qu.:2016-02-17
##  Max.   :2016-02-21   Max.   :2016-02-21
##  offense_code_extension district_sector    hundred_block_location
##  Length:1000            Length:1000        Length:1000
##  Class :character       Class :character   Class :character
##  Mode  :character       Mode  :character   Mode  :character
##
##
##
##  summarized_offense_description    month         general_offense_number
##  Length:1000                    Min.   : 1.000   Length:1000
##  Class :character               1st Qu.: 2.000   Class :character
##  Mode  :character               Median : 2.000   Mode  :character
##                                  Mean   : 2.041
##                                  3rd Qu.: 2.000
##                                  Max.   :12.000
##       year        longitude      summary_offense_code    latitude
##  Min.   :2014   Min.   :-122.4   Length:1000          Min.   :47.50
##  1st Qu.:2016   1st Qu.:-122.3   Class :character     1st Qu.:47.60
##  Median :2016   Median :-122.3   Mode  :character     Median :47.62
##  Mean   :2016   Mean   :-122.3                        Mean   :47.63
##  3rd Qu.:2016   3rd Qu.:-122.3                        3rd Qu.:47.67
##  Max.   :2016   Max.   :-122.3                        Max.   :47.73
##   rms_cdw_id        occurred_date_or_date_range_start
##  Length:1000        Min.   :2014-01-01
##  Class :character   1st Qu.:2016-02-13
##  Mode  :character   Median :2016-02-15
##                     Mean   :2016-02-12
##                     3rd Qu.:2016-02-17
##                     Max.   :2016-02-21
##  location_longitude.longitude location_latitude.latitude
##  Min.   :-122.41139           Min.   :47.49896
##  1st Qu.:-122.34565           1st Qu.:47.60088
##  Median :-122.33043           Median :47.61645
##  Mean   :-122.33073           Mean   :47.62676
##  3rd Qu.:-122.31547           3rd Qu.:47.66865
##  Max.   :-122.25001           Max.   :47.73394
```

```r
# Checking if multiple columns of longitude have the same data
diff<-policeIncidents.tidy$longitude-policeIncidents.tidy$location_longitude
summary(diff)
```

```
##    longitude
##  Min.   :0
##  1st Qu.:0
##  Median :0
##  Mean   :0
##  3rd Qu.:0
##  Max.   :0
```

```r
# Checking if multiple columns of latitude have the same data
diff<-policeIncidents.tidy$latitude-policeIncidents.tidy$location_latitude
summary(diff)
```

```
##     latitude
##  Min.   :0
##  1st Qu.:0
##  Median :0
##  Mean   :0
##  3rd Qu.:0
##  Max.   :0
```

```r
#Removing the extra columns
policeIncidents.tidy<-subset(policeIncidents.tidy,select=-c(location_longitude,location_latitude))
# Dataset details after tidying
summary(policeIncidents.tidy)
```

```
##  offense_code       offense_type       census_tract_2000 
##  Length:1000        Length:1000        Length:1000       
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##                                                          
##                                                          
##                                                          
##  date_reported        occurred_date_range_end  zone_beat        
##  Min.   :2016-02-13   Min.   :2015-11-11      Length:1000       
##  1st Qu.:2016-02-14   1st Qu.:2016-02-13      Class :character  
##  Median :2016-02-16   Median :2016-02-15      Mode  :character  
##  Mean   :2016-02-16   Mean   :2016-02-14                        
##  3rd Qu.:2016-02-18   3rd Qu.:2016-02-17                        
##  Max.   :2016-02-21   Max.   :2016-02-21                        
##  offense_code_extension district_sector    hundred_block_location
##  Length:1000            Length:1000        Length:1000           
##  Class :character       Class :character   Class :character      
##  Mode  :character       Mode  :character   Mode  :character      
##                                                                  
##                                                                  
##                                                                  
##  summarized_offense_description     month        general_offense_number
##  Length:1000                    Min.   : 1.000   Length:1000           
##  Class :character               1st Qu.: 2.000   Class :character      
##  Mode  :character               Median : 2.000   Mode  :character      
##                                 Mean   : 2.041                         
##                                 3rd Qu.: 2.000                         
##                                 Max.   :12.000                         
##       year        longitude      summary_offense_code    latitude    
##  Min.   :2014   Min.   :-122.4   Length:1000          Min.   :47.50  
##  1st Qu.:2016   1st Qu.:-122.3   Class :character     1st Qu.:47.60  
##  Median :2016   Median :-122.3   Mode  :character     Median :47.62  
##  Mean   :2016   Mean   :-122.3                        Mean   :47.63  
##  3rd Qu.:2016   3rd Qu.:-122.3                        3rd Qu.:47.67  
##  Max.   :2016   Max.   :-122.3                        Max.   :47.73  
##   rms_cdw_id        occurred_date_or_date_range_start
##  Length:1000        Min.   :2014-01-01               
##  Class :character   1st Qu.:2016-02-13               
##  Mode  :character   Median :2016-02-15               
##                     Mean   :2016-02-12               
##                     3rd Qu.:2016-02-17               
##                     Max.   :2016-02-21
```

```
colnames(policeIncidents.tidy)
```

```
##  [1] "offense_code"                     
##  [2] "offense_type"                     
##  [3] "census_tract_2000"                
##  [4] "date_reported"                    
##  [5] "occurred_date_range_end"          
##  [6] "zone_beat"                        
##  [7] "offense_code_extension"           
##  [8] "district_sector"                  
##  [9] "hundred_block_location"
## [10] "summarized_offense_description"
## [11] "month"
## [12] "general_offense_number"
## [13] "year"
## [14] "longitude"
## [15] "summary_offense_code"
## [16] "latitude"
## [17] "rms_cdw_id"
## [18] "occurred_date_or_date_range_start"
```

```
# Melting the date variables and removing NA values
policeIncidents.tidy<-melt(
  data=policeIncidents.tidy,
  id=c("offense_code","offense_type","census_tract_2000","zone_beat","offense_code_extension","district
  variable.name = "date_type",
  value.name = "date_value",
  na.rm=TRUE
)


# Melting the location variables and removing NA values
policeIncidents.tidy<-melt(
  data=policeIncidents.tidy,
  id=c("offense_code","offense_type","census_tract_2000","zone_beat","offense_code_extension","district
  variable.name = "location_measure",
  value.name = "location_value",
  na.rm=TRUE
)

# Exploring variables of the offense category
head(select(policeIncidents.tidy,contains("offense")))
```

```
##   offense_code             offense_type offense_code_extension
## 1            X          DISTURBANCE-OTH                     21
## 2         1313             ASSLT-NONAGG                      0
## 3         2404            VEH-THEFT-AUTO                     1
## 4         5707                 TRESPASS                      0
## 5         1206 ROBBERY-STREET-BODYFORCE                      0
## 6            X          DISTURBANCE-OTH                     21
##   summarized_offense_description general_offense_number
## 1                    DISTURBANCE              201662362
## 2                        ASSAULT              201662321
## 3                  VEHICLE THEFT              201662337
## 4                       TRESPASS              201662300
## 5                        ROBBERY              201662295
## 6                    DISTURBANCE              201662265
##   summary_offense_code
## 1                    X
## 2                 1300
## 3                 2400
## 4                 5700
## 5                 1200
## 6                    X
```

```
# Melting the different types of offense codes and removing NA values
policeIncidents.tidy<-melt(
  data=policeIncidents.tidy,
  id=c("offense_type","census_tract_2000","zone_beat","district_sector","hundred_block_location","summa
  variable.name = "offense_code_type",
  value.name = "offense_code_value",
  na.rm=TRUE
)

# Viewing the tidied dataset
#View(policeIncidents.tidy)
# Exploring the tidied dataset
head(policeIncidents.tidy)
```

```
##                offense_type census_tract_2000 zone_beat district_sector
## 1          DISTURBANCE-OTH          100.3004        L3               L
## 2             ASSLT-NONAGG        10702.3009        W3               W
## 3           VEH-THEFT-AUTO         3100.5011        J2               J
## 4                 TRESPASS         7200.1079        Q3               Q
## 5 ROBBERY-STREET-BODYFORCE         8100.3011        M1               M
## 6          DISTURBANCE-OTH         8400.1004        E3               E
##       hundred_block_location summarized_offense_description month year
## 1    125XX BLOCK OF 35 AV NE                    DISTURBANCE     2 2016
## 2 65XX BLOCK OF SYLVAN WY SW                        ASSAULT     2 2016
## 3     83XX BLOCK OF 32 AV NW                  VEHICLE THEFT     2 2016
## 4      3XX BLOCK OF DENNY WY                       TRESPASS     2 2016
## 5   19XX BLOCK OF WESTERN AV                        ROBBERY     2 2016
## 6      SUMMIT AV / E PIKE ST                    DISTURBANCE     2 2016
##   rms_cdw_id     date_type date_value location_measure location_value
## 1     644277 date_reported 2016-02-21        longitude      -122.2908
## 2     644273 date_reported 2016-02-21        longitude      -122.3708
## 3     644307 date_reported 2016-02-21        longitude      -122.3983
## 4     644306 date_reported 2016-02-21        longitude      -122.3496
## 5     644251 date_reported 2016-02-21        longitude      -122.3429
## 6     644210 date_reported 2016-02-21        longitude      -122.3255
##   offense_code_type offense_code_value
## 1      offense_code                  X
## 2      offense_code               1313
## 3      offense_code               2404
## 4      offense_code               5707
## 5      offense_code               1206
## 6      offense_code                  X
```

```
colnames(policeIncidents.tidy)
```

```
##  [1] "offense_type"                   "census_tract_2000"
##  [3] "zone_beat"                      "district_sector"
##  [5] "hundred_block_location"         "summarized_offense_description"
##  [7] "month"                          "year"
##  [9] "rms_cdw_id"                     "date_type"
## [11] "date_value"                     "location_measure"
## [13] "location_value"                 "offense_code_type"
## [15] "offense_code_value"
```

```r
summary(policeIncidents.tidy)
```

```
##  offense_type       census_tract_2000   zone_beat
##  Length:19296       Length:19296       Length:19296
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  district_sector    hundred_block_location summarized_offense_description
##  Length:19296       Length:19296           Length:19296
##  Class :character   Class :character       Class :character
##  Mode  :character   Mode  :character       Mode  :character
##
##
##
##      month             year       rms_cdw_id
##  Min.   : 1.000   Min.   :2014   Length:19296
##  1st Qu.: 2.000   1st Qu.:2016   Class :character
##  Median : 2.000   Median :2016   Mode  :character
##  Mean   : 2.047   Mean   :2016
##  3rd Qu.: 2.000   3rd Qu.:2016
##  Max.   :12.000   Max.   :2016
##                              date_type      date_value
##  date_reported                    :8000   Min.   :2014-01-01
##  occurred_date_range_end          :3296   1st Qu.:2016-02-14
##  occurred_date_or_date_range_start:8000   Median :2016-02-16
##                                           Mean   :2016-02-14
##                                           3rd Qu.:2016-02-17
##                                           Max.   :2016-02-21
##   location_measure location_value                 offense_code_type
##  longitude:9648    Min.   :-122.41   offense_code          :4824
##  latitude :9648    1st Qu.:-122.33   offense_code_extension:4824
##                    Median : -37.38   general_offense_number:4824
##                    Mean   : -37.35   summary_offense_code  :4824
##                    3rd Qu.:  47.62
##                    Max.   :  47.73
##  offense_code_value
##  Length:19296
##  Class :character
##  Mode  :character
##
##
##
```

```r
apply(policeIncidents.tidy,2,class)
```

```
##                   offense_type              census_tract_2000
##                    "character"                    "character"
##                      zone_beat                district_sector
##                    "character"                    "character"
##         hundred_block_location summarized_offense_description
##                    "character"                       "character" 
##                          month                              year 
##                    "character"                       "character" 
##                     rms_cdw_id                         date_type 
##                    "character"                       "character" 
##                     date_value                  location_measure 
##                    "character"                       "character" 
##                 location_value                 offense_code_type 
##                    "character"                       "character" 
##             offense_code_value 
##                    "character"
```

```
# Exporting the cleaned dataset
save(policeIncidents.tidy,file="policeIncidentsTidy.Rdata")
```

(d) Describe any concerns you might have about this data. This may include biases, missing data, or ethical concerns. The greatest ethical concern is the privacy of the victims who are involved in the police incident. The open dataset makes the infromation available to anyone without any verification or procedures. The longitude and latitude variables reveal the exact location of the incident. Also the police incident report can be combined with other sources of data and it can be used to cause greater discomfort for the victims. For example, a newspaper report about the items stolen in a house from a particular neighbourhood combined with the police incident report can give valuable information to the the wrong people. Hence, making the victims more vulnerable to future police incidents. The missing data in offense_code, summary_offense_code and occured_date_range_end variables represent incomplete information. The greater cause of concern is that we don't know how the missing values are represented in the different variables. For example what if '0' in offense_code_extension represents missing data. It is not easy to figure out all the meta data regarding the dataset and hence it is possible that people doing data analysis on the dataset can come to worng conclusions because of insufficient background on the dataset.

Exploring 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 and describe in a short paragraph how the data was collected and what each variable represents.

The nycflights data was collected from the Bureau of transportation statistics about the flights that departed from NYC airports in 2013. The dataset flights has 336776 observation of flights departing from NYC airports and 16 variables that capture the flight departure data. The dataset has the followiong variables -
[1] "year" - Year of departure which is 2013 for all the records
[2] "month" - Month of flight departure
[3] "day" - Day of departure
[4] "dep_time" - departure time of flight
[5] "dep_delay" - departure delay
[6] "arr_time" - arrival time
[7] "arr_delay" - arrival delay
[8] "carrier" - airline carrier (abbreviated in 2 letters)
[8] "tailnum" - Tail number of the plane
[9] "flight" - Flight number
[10] "origin" - Origin of the flight which is one of the NYC airports
[11] "dest" - Destination airport for the flight
[13] "air_time" - Amount of time spent in air bny the flight
[14] "distance" - The distance flown by the flight

[15] “hour” - departure time in hours
[16] “minute” - departure time in minutes

```
# Importing the nycflights13 dataset
library(nycflights13)
flights<-nycflights13::flights
# Information about the dataset
#?nycflights13::flights
# Exploring the dataset
head(flights)
```

```
## Source: local data frame [6 x 16]
##
##    year month   day dep_time dep_delay arr_time arr_delay carrier tailnum
##   (int) (int) (int)    (int)     (dbl)    (int)     (dbl)   (chr)   (chr)
## 1  2013     1     1      517         2      830        11      UA  N14228
## 2  2013     1     1      533         4      850        20      UA  N24211
## 3  2013     1     1      542         2      923        33      AA  N619AA
## 4  2013     1     1      544        -1     1004       -18      B6  N804JB
## 5  2013     1     1      554        -6      812       -25      DL  N668DN
## 6  2013     1     1      554        -4      740        12      UA  N39463
## Variables not shown: flight (int), origin (chr), dest (chr), air_time
##   (dbl), distance (dbl), hour (dbl), minute (dbl)
```

```
dim(flights)
```

```
## [1] 336776     16
```

```
colnames(flights)
```

```
##  [1] "year"      "month"     "day"       "dep_time"  "dep_delay"
##  [6] "arr_time"  "arr_delay" "carrier"   "tailnum"   "flight"
## [11] "origin"    "dest"      "air_time"  "distance"  "hour"
## [16] "minute"
```

```
summary(flights)
```

```
##       year          month             day           dep_time
##  Min.   :2013   Min.   : 1.000   Min.   : 1.00   Min.   :   1
##  1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 907
##  Median :2013   Median : 7.000   Median :16.00   Median :1401
##  Mean   :2013   Mean   : 6.549   Mean   :15.71   Mean   :1349
##  3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1744
##  Max.   :2013   Max.   :12.000   Max.   :31.00   Max.   :2400
##                                                  NA's   :8255
##    dep_delay          arr_time      arr_delay          carrier
##  Min.   : -43.00   Min.   :   1   Min.   : -86.000   Length:336776
##  1st Qu.:  -5.00   1st Qu.:1104   1st Qu.: -17.000   Class :character
##  Median :  -2.00   Median :1535   Median :  -5.000   Mode  :character
##  Mean   :  12.64   Mean   :1502   Mean   :   6.895
##  3rd Qu.:  11.00   3rd Qu.:1940   3rd Qu.:  14.000  
##  Max.   :1301.00   Max.   :2400   Max.   :1272.000  
##  NA's   :8255      NA's   :8713   NA's   :9430      
##    tailnum              flight        origin              dest          
##  Length:336776      Min.   :   1   Length:336776      Length:336776     
##  Class :character   1st Qu.: 553   Class :character   Class :character  
##  Mode  :character   Median :1496   Mode  :character   Mode  :character  
##                     Mean   :1972                                        
##                     3rd Qu.:3465                                        
##                     Max.   :8500                                        
##                                                                         
##     air_time        distance          hour           minute     
##  Min.   : 20.0   Min.   :  17   Min.   : 0.00   Min.   : 0.00  
##  1st Qu.: 82.0   1st Qu.: 502   1st Qu.: 9.00   1st Qu.:16.00  
##  Median :129.0   Median : 872   Median :14.00   Median :31.00  
##  Mean   :150.7   Mean   :1040   Mean   :13.17   Mean   :31.76  
##  3rd Qu.:192.0   3rd Qu.:1389   3rd Qu.:17.00   3rd Qu.:49.00  
##  Max.   :695.0   Max.   :4983   Max.   :24.00   Max.   :59.00  
##  NA's   :9430                   NA's   :8255    NA's   :8255
```

```r
sapply(flights,class)
```

```
##        year       month         day    dep_time   dep_delay    arr_time 
##   "integer"   "integer"   "integer"   "integer"   "numeric"   "integer" 
##   arr_delay     carrier     tailnum      flight      origin        dest 
##   "numeric" "character" "character"   "integer" "character" "character" 
##    air_time    distance        hour      minute 
##   "numeric"   "numeric"   "numeric"   "numeric"
```

(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?

There are 3923 flights from NYC airports to Seattle in 2013.

```r
# Number of flights from NYC airports to Seattle airports

flights %>%
  filter(dest=="SEA") %>%
    summarise(flight_count_NYC_to_SEA = n())
```

```
## Source: local data frame [1 x 1]
## 
##   flight_count_NYC_to_SEA
##                     (int)
## 1                    3923
```

- How many airlines fly from NYC to Seattle?

There are 5 airlines from NYC to Seattle.

```r
# Number of distinct airlines that fly from NYC to Seattle
flights %>%
  filter(dest=="SEA") %>%
    distinct(carrier) %>%
      summarise(airlines_count_NYC_to_SEA = n())
```

```
## Source: local data frame [1 x 1]
##
##   airlines_count_NYC_to_SEA
##                       (int)
## 1                         5
```

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

936 unique airplanes fly from NYC to Seattle

```r
# Number of unique air planes(given by tailnum column) that fly from NYC to Seattle
flights %>%
  filter(dest=="SEA") %>%
    distinct(tailnum) %>%
      summarise(planes_count_NYC_to_SEA = n())
```

```
## Source: local data frame [1 x 1]
##
##   planes_count_NYC_to_SEA
##                     (int)
## 1                     936
```

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

The average arrival delay for flights from NYC to Seattle is -1.099099

```r
# Average arrival delay for flights from NYC to Seattle
flights %>%
  filter(dest=="SEA") %>%
    summarise(avg_delay = mean(arr_delay, na.rm=TRUE))
```

```
## Source: local data frame [1 x 1]
##
##   avg_delay
##       (dbl)
## 1 -1.099099
```

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

About 46.7% of flights to Seattle come from EWR airport in New York and 53.3% of flights to Seattle come from JFK airport.

```
# Proportion of flights to Seattle from respective NYC airports
flights %>%
  filter(dest=="SEA") %>%
    group_by(origin) %>%
      summarise(num_flights = n()) %>%
        mutate(proportion_flights = num_flights*100/sum(num_flights))
```

```
## Source: local data frame [2 x 3]
##
##   origin num_flights proportion_flights
##    (chr)       (int)              (dbl)
## 1    EWR        1831           46.67346
## 2    JFK        2092           53.32654
```