Introduction to Data Wrangling

Setia

Data Wrangling

Data Wrangling (data preparation) adalah proses pembersihan (cleaning), penataan (structuring), dan pengayaan (enriching) data mentah ke dalam format yang siap dilakukan analisas berikutnya (downstream analysis). Data wrangling menjadi penting saat ini karena data yang ada lebih beragam dan tidak terstruktur. Biasanya terdapat enam langkah (berulang) dalam proses data wrangling:

1. Discovering: mengetahui berbagai sumber data (datasets) yang ada dan diperlukan .
2. Structuring: Pengaturan data, yang diperlukan karena data mentah datang dalam berbagai bentuk dan ukuran. Satu kolom dapat berubah menjadi beberapa baris untuk analisis yang lebih mudah.
3. Cleaning: Proses pembersihan data dari errors, ourtliers, salah ketik, missing data, non response, dll.
4. Enriching: Proses pengayaan data, bagai mana menggabungkan berbagai sumber data atau database sehingga didapat informasi tambahan dari data tersebut.
5. Validating: Merupakan proses untuk melakukan validasi dengan aturan yang ada (validation rules). Validasai termasuk verifikasi, konsistensi, qualitas (quality), dan keamanan (security)
6. Publishing: Menyediakan data yang telah "clean" untuk analisa lebih lanjut (down stream analysis)

Relational Data dengan dplyr

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.4

##
## 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
```

```r
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.4.4

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

## v tibble  1.4.1     v purrr   0.2.4
## v tidyr   0.8.1     v stringr 1.4.0
## v readr   1.1.1     v forcats 0.2.0

## Warning: package 'tibble' was built under R version 3.4.3
## Warning: package 'tidyr' was built under R version 3.4.4
```

```
## Warning: package 'readr' was built under R version 3.4.3
```

```
## Warning: package 'purrr' was built under R version 3.4.3
```

```
## Warning: package 'forcats' was built under R version 3.4.3
```

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

```
library(nycflights13)
```

```
## Warning: package 'nycflights13' was built under R version 3.4.4
```

```
data("airlines")
head(airlines)
```

```
## # A tibble: 6 x 2
##   carrier name
##   <chr>   <chr>
## 1 9E      Endeavor Air Inc.
## 2 AA      American Airlines Inc.
## 3 AS      Alaska Airlines Inc.
## 4 B6      JetBlue Airways
## 5 DL      Delta Air Lines Inc.
## 6 EV      ExpressJet Airlines Inc.
```

```
data("airports")
head(airports)
```

```
## # A tibble: 6 x 8
##   faa   name                             lat   lon   alt    tz dst   tzone
##   <chr> <chr>                          <dbl> <dbl> <int> <dbl> <chr> <chr>
## 1 04G   Lansdowne Airport               41.1 -80.6  1044 -5.00 A     Amer~
## 2 06A   Moton Field Municipal Airport   32.5 -85.7   264 -6.00 A     Amer~
## 3 06C   Schaumburg Regional             42.0 -88.1   801 -6.00 A     Amer~
## 4 06N   Randall Airport                 41.4 -74.4   523 -5.00 A     Amer~
## 5 09J   Jekyll Island Airport           31.1 -81.4    11 -5.00 A     Amer~
## 6 0A9   Elizabethton Municipal Airport  36.4 -82.2  1593 -5.00 A     Amer~
```

```
data("planes")
tail(planes)
```

```
## # A tibble: 6 x 9
##   tailnum  year type         manufacturer   model engi~ seats speed engine
##   <chr>   <int> <chr>        <chr>          <chr> <int> <int> <int> <chr>
## 1 N996DL   1991 Fixed wing ~ MCDONNELL DOU~ MD-88     2   142    NA Turbo~
## 2 N997AT   2002 Fixed wing ~ BOEING         717-~     2   100    NA Turbo~
## 3 N997DL   1992 Fixed wing ~ MCDONNELL DOU~ MD-88     2   142    NA Turbo~
## 4 N998AT   2002 Fixed wing ~ BOEING         717-~     2   100    NA Turbo~
## 5 N998DL   1992 Fixed wing ~ MCDONNELL DOU~ MD-88     2   142    NA Turbo~
## 6 N999DN   1992 Fixed wing ~ MCDONNELL DOU~ MD-88     2   142    NA Turbo~
```

```
data("weather")
tail(weather)
```

```
## # A tibble: 6 x 15
##   origin  year month   day  hour  temp  dewp humid wind~ wind~ wind~ prec~
##   <chr>  <dbl> <dbl> <int> <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 LGA     2013  12.0    30    13  37.0  21.9  54.0   340  17.3  20.7     0
## 2 LGA     2013  12.0    30    14  36.0  19.9  51.8   340  13.8  21.9     0
## 3 LGA     2013  12.0    30    15  34.0  17.1  49.5   330  17.3  21.9     0
## 4 LGA     2013  12.0    30    16  32.0  15.1  49.2   340  15.0  23.0     0
## 5 LGA     2013  12.0    30    17  30.9  12.9  46.7   320  17.3  NA       0
## 6 LGA     2013  12.0    30    18  28.9  10.9  46.4   330  18.4  NA       0
## # ... with 3 more variables: pressure <dbl>, visib <dbl>, time_hour <dttm>
```

```
data("flights")
head(flights)
```

```
## # A tibble: 6 x 19
##    year month   day dep_t~ sche~ dep_~ arr_~ sche~ arr_~ carr~ flig~ tail~
##   <int> <int> <int>  <int> <int> <dbl> <int> <int> <dbl> <chr> <int> <chr>
## 1  2013     1     1    517   515  2.00   830   819  11.0 UA     1545 N142~
## 2  2013     1     1    533   529  4.00   850   830  20.0 UA     1714 N242~
## 3  2013     1     1    542   540  2.00   923   850  33.0 AA     1141 N619~
## 4  2013     1     1    544   545 -1.00  1004  1022 -18.0 B6      725 N804~
## 5  2013     1     1    554   600 -6.00   812   837 -25.0 DL      461 N668~
## 6  2013     1     1    554   558 -4.00   740   728  12.0 UA     1696 N394~
## # ... with 7 more variables: origin <chr>, dest <chr>, air_time <dbl>,
## #   distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
```

```
setwd("C:/Users/stis/Documents/Training R Data Science Pusdiklat")
country <- read.csv("CountryData.csv")
dim(country)
```

```
## [1] 256  77
```

```
## check if unique key

planes %>%
  count(tailnum) %>%
  filter(n>1)
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
## # A tibble: 0 x 2
## # ... with 2 variables: tailnum <chr>, n <int>
```

```
#join ##

flights2 <- flights %>%
  select(year:day, hour, origin, dest, tailnum, carrier)

head(flights2)
```

```
## # A tibble: 6 x 8
##    year month   day  hour origin dest  tailnum carrier
##   <int> <int> <int> <dbl> <chr>  <chr> <chr>   <chr>
## 1  2013     1     1  5.00 EWR    IAH   N14228  UA
## 2  2013     1     1  5.00 LGA    IAH   N24211  UA
## 3  2013     1     1  5.00 JFK    MIA   N619AA  AA
## 4  2013     1     1  5.00 JFK    BQN   N804JB  B6
## 5  2013     1     1  6.00 LGA    ATL   N668DN  DL
## 6  2013     1     1  5.00 EWR    ORD   N39463  UA
```

```r
head(airlines)
```

```
## # A tibble: 6 x 2
##   carrier name
##   <chr>   <chr>
## 1 9E      Endeavor Air Inc.
## 2 AA      American Airlines Inc.
## 3 AS      Alaska Airlines Inc.
## 4 B6      JetBlue Airways
## 5 DL      Delta Air Lines Inc.
## 6 EV      ExpressJet Airlines Inc.
```

```r
## add full airline by carrier ##
 flights2 %>%
   select(-origin, -dest) %>%
   left_join(airlines, by="carrier")
```

```
## # A tibble: 336,776 x 7
##     year month   day  hour tailnum carrier name
##    <int> <int> <int> <dbl> <chr>   <chr>   <chr>
##  1  2013     1     1  5.00 N14228  UA      United Air Lines Inc.
##  2  2013     1     1  5.00 N24211  UA      United Air Lines Inc.
##  3  2013     1     1  5.00 N619AA  AA      American Airlines Inc.
##  4  2013     1     1  5.00 N804JB  B6      JetBlue Airways
##  5  2013     1     1  6.00 N668DN  DL      Delta Air Lines Inc.
##  6  2013     1     1  5.00 N39463  UA      United Air Lines Inc.
##  7  2013     1     1  6.00 N516JB  B6      JetBlue Airways
##  8  2013     1     1  6.00 N829AS  EV      ExpressJet Airlines Inc.
##  9  2013     1     1  6.00 N593JB  B6      JetBlue Airways
## 10  2013     1     1  6.00 N3ALAA  AA      American Airlines Inc.
## # ... with 336,766 more rows
```

```r
 ## join using all matched variables

 flights2 %>%
   left_join(weather)
```

```
## Joining, by = c("year", "month", "day", "hour", "origin")
```

```
## # A tibble: 336,776 x 18
##     year month   day  hour orig~ dest  tail~ carr~  temp  dewp humid wind~
##    <dbl> <dbl> <int> <dbl> <chr> <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl>
##  1  2013  1.00     1  5.00 EWR   IAH   N142~ UA     39.0  28.0  64.4   260
##  2  2013  1.00     1  5.00 LGA   IAH   N242~ UA     39.9  25.0  54.8   250
##  3  2013  1.00     1  5.00 JFK   MIA   N619~ AA     39.0  27.0  61.6   260
##  4  2013  1.00     1  5.00 JFK   BQN   N804~ B6     39.0  27.0  61.6   260
##  5  2013  1.00     1  6.00 LGA   ATL   N668~ DL     39.9  25.0  54.8   260
##  6  2013  1.00     1  5.00 EWR   ORD   N394~ UA     39.0  28.0  64.4   260
##  7  2013  1.00     1  6.00 EWR   FLL   N516~ B6     37.9  28.0  67.2   240
##  8  2013  1.00     1  6.00 LGA   IAD   N829~ EV     39.9  25.0  54.8   260
##  9  2013  1.00     1  6.00 JFK   MCO   N593~ B6     37.9  27.0  64.3   260
## 10  2013  1.00     1  6.00 LGA   ORD   N3AL~ AA     39.9  25.0  54.8   260
## # ... with 336,766 more rows, and 6 more variables: wind_speed <dbl>,
## #   wind_gust <dbl>, precip <dbl>, pressure <dbl>, visib <dbl>, time_hour
## #   <dttm>
```

Handling Time Variable

```
 ## Dates and Time Handling

 library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 3.4.4
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
today()
```

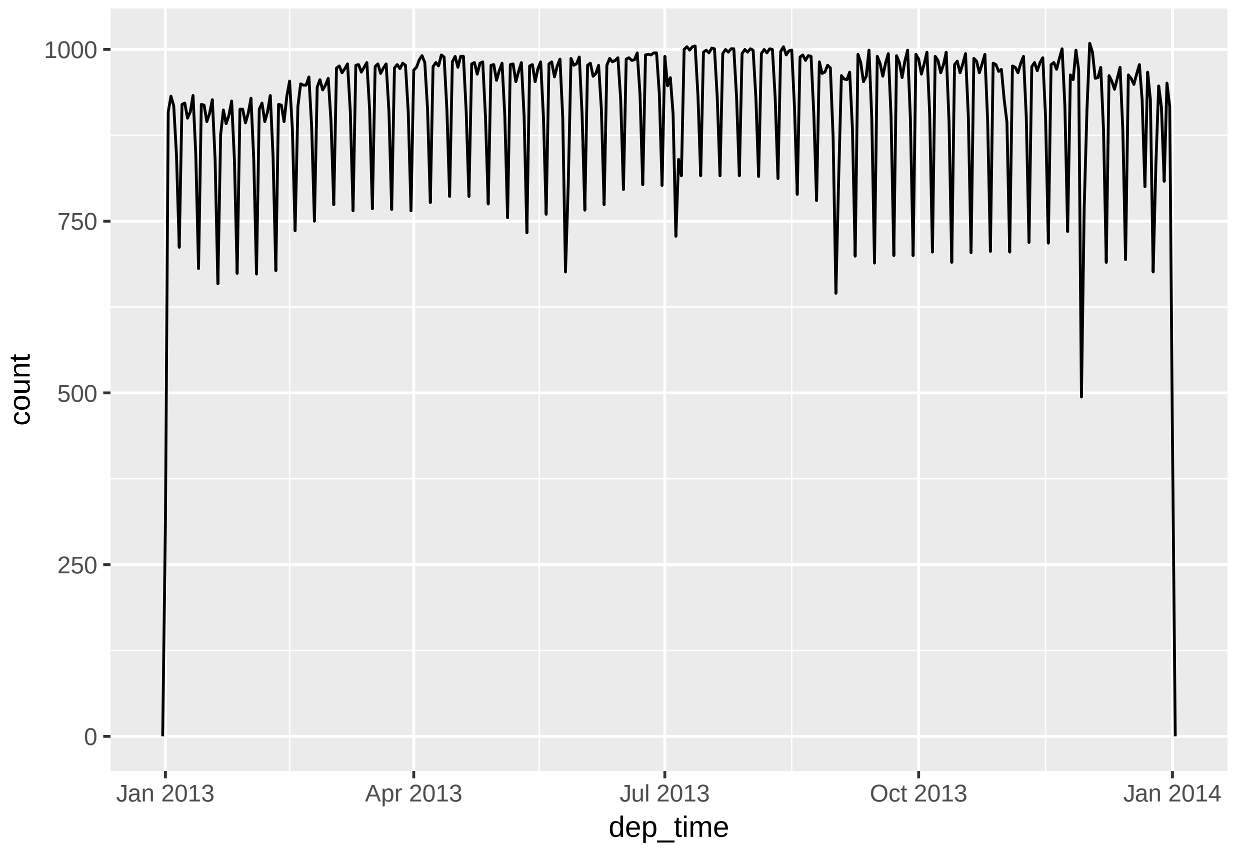
```
## [1] "2019-02-24"
```

```
now()
```

```
## [1] "2019-02-24 20:51:19 +07"
```

```
flights2 <-  flights %>%
   select(year, month, day, hour, minute) %>%
   mutate(
     dep_time = make_datetime(year, month, day, hour,minute)
   )


 flights2 %>%
   ggplot(aes(dep_time)) +
   geom_freqpoly(binwidth=86400)
```



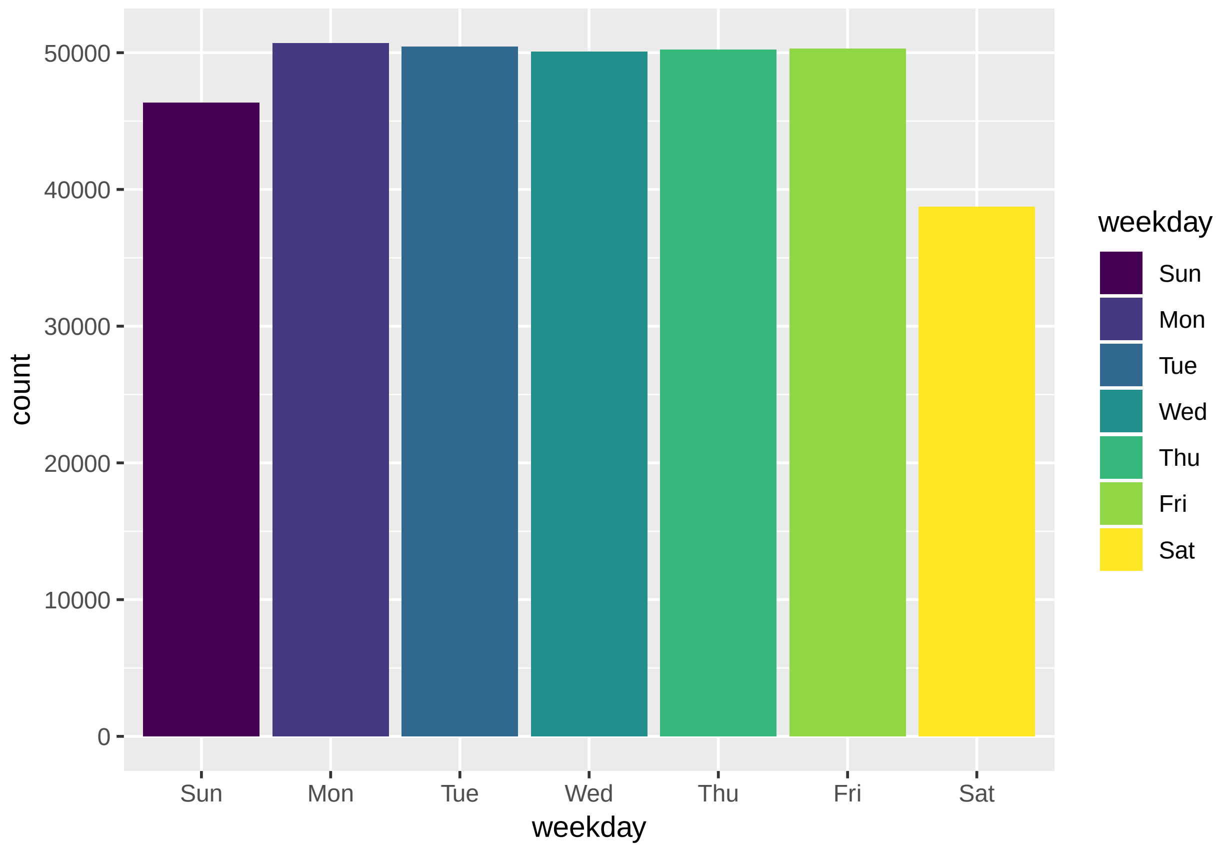
```
## Get  name of the day ##

flights2 <- flights2 %>%
  mutate(weekday= wday(dep_time, label=T))

head(flights2)
```

```
## # A tibble: 6 x 7
##    year month   day  hour minute dep_time            weekday
##   <int> <int> <int> <dbl>  <dbl> <dttm>              <ord>
## 1  2013     1     1  5.00   15.0 2013-01-01 05:15:00 Tue
## 2  2013     1     1  5.00   29.0 2013-01-01 05:29:00 Tue
## 3  2013     1     1  5.00   40.0 2013-01-01 05:40:00 Tue
## 4  2013     1     1  5.00   45.0 2013-01-01 05:45:00 Tue
## 5  2013     1     1  6.00    0   2013-01-01 06:00:00 Tue
## 6  2013     1     1  5.00   58.0 2013-01-01 05:58:00 Tue
```

```
flights2 %>% ggplot(aes(x=weekday)) +
  geom_bar(aes(fill=weekday))
```



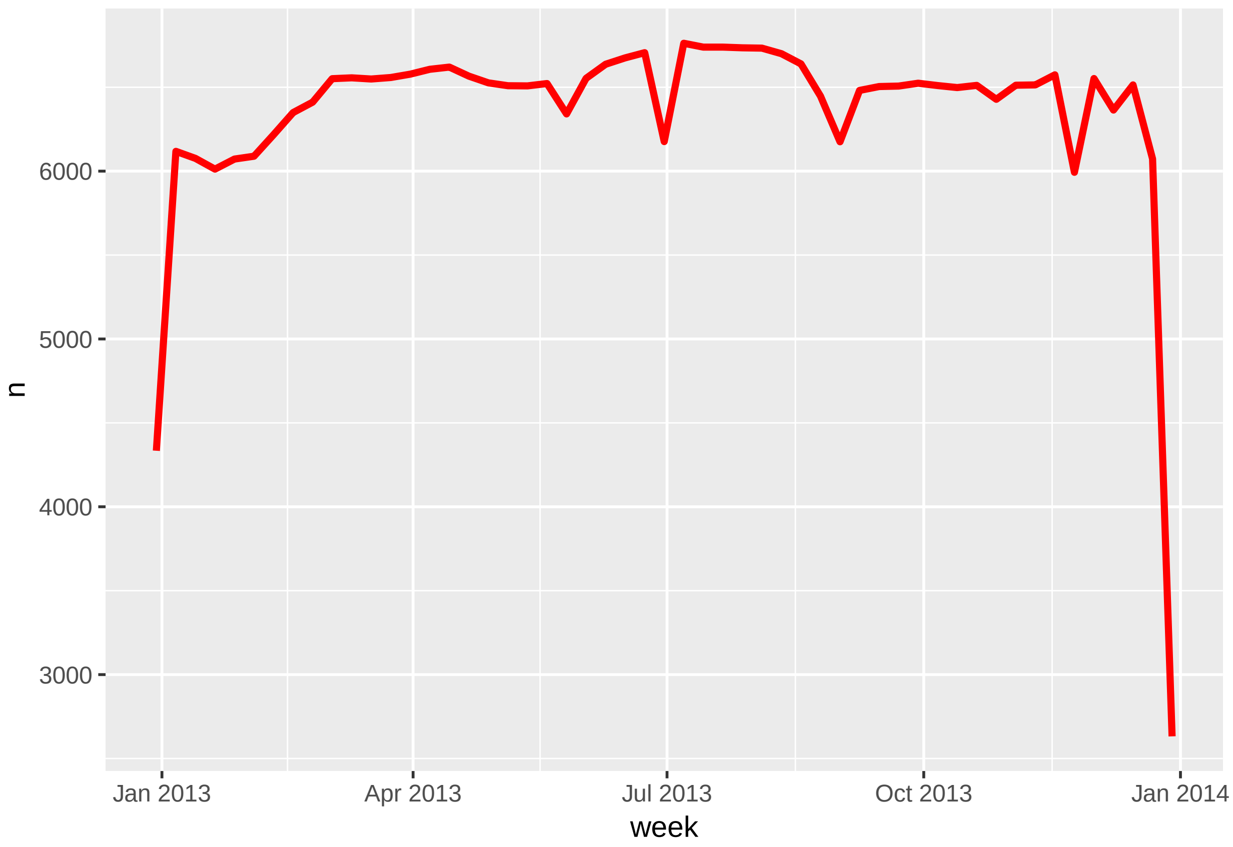
```
## rounding Time ##

weekflight <- flights2 %>%
  count(week=floor_date(dep_time,"week"))

 head(weekflight)
```

```
## # A tibble: 6 x 2
##   week                    n
##   <dttm>              <int>
## 1 2012-12-30 00:00:00  4334
## 2 2013-01-06 00:00:00  6118
## 3 2013-01-13 00:00:00  6076
## 4 2013-01-20 00:00:00  6012
## 5 2013-01-27 00:00:00  6072
## 6 2013-02-03 00:00:00  6089
```

```
 ggplot(weekflight, aes(week,n)) +
   geom_line(col=2,lwd=1.25)
```



Perapian (Tidy) Data

Data yang berasal dari berbagai sumber dan biasanya tidak siap untuk digunakan dalam analisis harus dirapikan. Perapian data merupakan salah satu proses yang penting dalam data wrangling.

```
 ## Data From Barcelona

pop <- read.csv("barcelona-data-sets/population.csv")
head(pop)
```

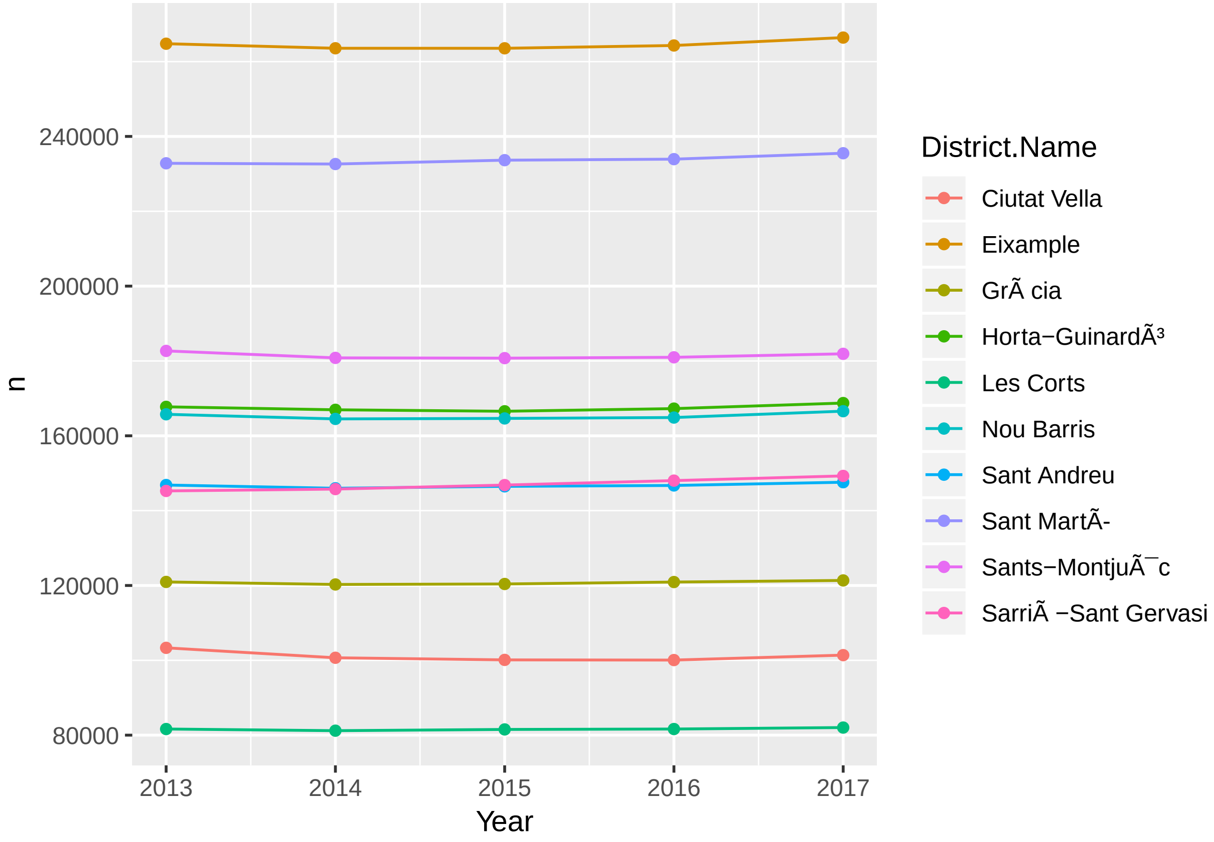
```
##   Year District.Code District.Name Neighborhood.Code
## 1 2017             1  Ciutat Vella                 1
## 2 2017             1  Ciutat Vella                 2
## 3 2017             1  Ciutat Vella                 3
## 4 2017             1  Ciutat Vella                 4
## 5 2017             2      Eixample                 5
## 6 2017             2      Eixample                 6
##                        Neighborhood.Name Gender Age Number
## 1                               el Raval   Male 0-4    224
## 2                        el Barri GÃ²tic   Male 0-4     50
## 3                         la Barceloneta   Male 0-4     43
## 4 Sant Pere, Santa Caterina i la Ribera   Male 0-4     95
## 5                          el Fort Pienc   Male 0-4    124
## 6                     la Sagrada FamÃlia   Male 0-4    191
```

```r
dim(pop)
```

```
## [1] 70080     8
```

```r
pop2 <- pop %>%
  count(Year,District.Name, wt=Number)

ggplot(pop2, aes(Year, n)) +
  geom_line(aes(group=District.Name,color=District.Name))+
  geom_point(aes(color=District.Name))
```



```r
## Wide form ##

pop3 <- spread(pop2, key=Year, value=n)
head(pop3)
```

```
## # A tibble: 6 x 6
##   District.Name                `2013` `2014` `2015` `2016` `2017`
##   <fctr>                        <int>  <int>  <int>  <int>  <int>
## 1 Ciutat Vella                 103339 100685 100115 100070 101387
## 2 Eixample                     264780 263565 263558 264305 266416
## 3 "Gr\u00c3\u00a0cia"          120949 120273 120401 120918 121347
## 4 "Horta-Guinard\u00c3\u00b3" 167743 166950 166559 167268 168751
## 5 Les Corts                     81640  81200  81530  81642  82033
## 6 Nou Barris                   165748 164516 164648 164881 166579
 ## Long form ##

 pop4 <- pop3 %>%
   gather("2013","2014", "2015" ,"2016", "2017", key=Year, value="population")
 head(pop4)
```

```
## # A tibble: 6 x 3
##   District.Name                Year  population
##   <fctr>                       <chr>      <int>
## 1 Ciutat Vella                 2013      103339
## 2 Eixample                     2013      264780
## 3 "Gr\u00c3\u00a0cia"         2013      120949
## 4 "Horta-Guinard\u00c3\u00b3" 2013      167743
## 5 Les Corts                    2013       81640
## 6 Nou Barris                   2013      165748
```