COMPSCIX 415.2 Homework 1

Rajat Jain

6/5/2018

Contents

My Repository

My Github repository for my assignments can be found at this URL: https://github.com/rajatmnnit/compscix-415-2-assignments

Libraries

```
library(mdsr)
library(tidyverse)
```

Exploring Data

Load data from Package

```
data("WorldCities")
```

Data Summary - Question# 1

WorldCities data set contains 23018 observations and 10 variables. Some of the variable names are: code, name, latitude, longitude, country, countryRegion, population, regionCode, region, date

A quick glimpse at the data

```
# Using glimpse function from dplyr
glimpse(WorldCities)
```

```
## Observations: 23,018
## Variables: 10
## $ code          <int> 3040051, 3041563, 290594, 291074, 291696, 292223...
## $ name          <chr> "les Escaldes", "Andorra la Vella", "Umm al Qayw...
## $ latitude      <dbl> 42.50729, 42.50779, 25.56473, 25.78953, 25.33132...
## $ longitude     <dbl> 1.53414, 1.52109, 55.55517, 55.94320, 56.34199, ...
## $ country       <chr> "AD", "AD", "AE", "AE", "AE", "AE", "AE", "AE", ...
## $ countryRegion <chr> "8", "7", "7", "5", "6", "3", "4", "6", "1", "4"...
## $ population    <dbl> 15853, 20430, 44411, 115949, 33575, 1137347, 263...
## $ regionCode    <int> 1033, 1037, 2, 2, 20, 11, 4, 6, 16, 15, 275, 4, ...
## $ region        <chr> "Europe/Andorra", "Europe/Andorra", "Asia/Dubai"...
## $ date          <chr> "10/15/08", "5/30/10", "11/3/12", "11/30/12", "1...
```

Extraction

Top 200 Rows

```
WorldCities <- head(WorldCities, 200) # 200 rows
```

Countries

```
country_col <- WorldCities$country
unique(country_col)
```

```
## [1] "AD" "AE" "AF" "AG" "AI" "AL" "AM" "AO" "AR"
```

Regions - Question# 2

```
unique(WorldCities$region)
```

```
##  [1] "Europe/Andorra"                 "Asia/Dubai"
##  [3] "Asia/Kabul"                     "America/Antigua"
##  [5] "America/Anguilla"               "Europe/Tirane"
##  [7] "Asia/Yerevan"                   "Africa/Luanda"
##  [9] "America/Argentina/Buenos_Aires" "America/Argentina/Cordoba"
## [11] "America/Argentina/Salta"        "America/Argentina/Tucuman"
## [13] "America/Argentina/San_Juan"
```

The tidy way - Question# 3

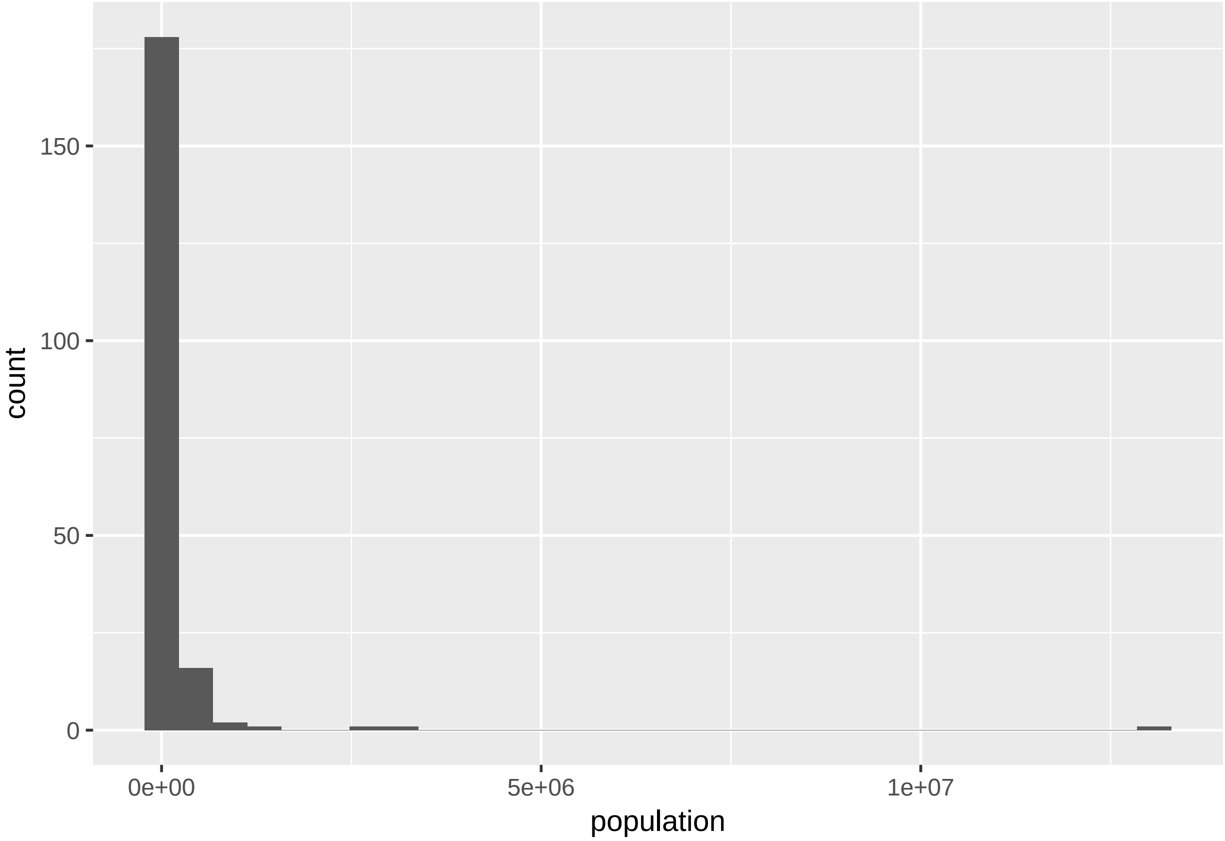
```
WorldCities %>% select(country) %>% unique()
```

```
##     country
## 1        AD
## 3        AE
## 15       AF
## 65       AG
## 66       AI
## 67       AL
## 87       AM
## 104      AO
## 131      AR
```

Visualize

```
WorldCities %>% ggplot(aes(x = population)) +
  geom_histogram()
```

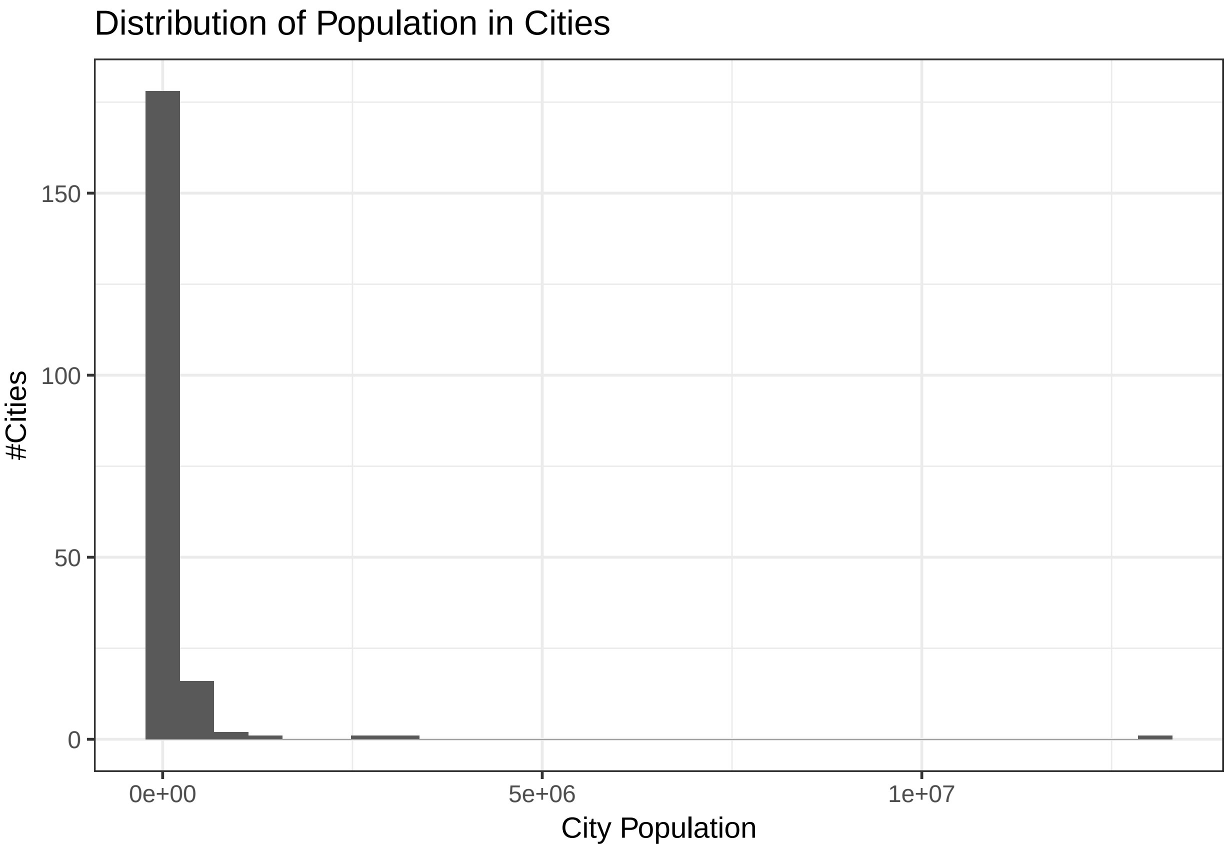
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Population Distribution - Question# 4

```
WorldCities %>% ggplot(aes(x = population)) +
  geom_histogram() +
  xlab("City Population") +
  ylab("#Cities") +
  ggtitle("Distribution of Population in Cities") +
  theme_bw()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


Distribution of Population in Cities
#Cities
150
100
50
0
0e+00
5e+06
1e+07
City Population