Data Cleaning in R

Aseem Mehrotra

07/09/2021

Cleaning the COVID Data for Analysis

Data source: **Bing**

Load Packages

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.4     ✓ dplyr   1.0.7
## ✓ tidyr   1.1.3     ✓ stringr 1.4.0
## ✓ readr   2.0.1     ✓ forcats 0.5.1
```

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

```
library(ggplot2)
library(sqldf)
```

```
## Loading required package: gsubfn
```

```
## Loading required package: proto
```

```
## Warning in doTryCatch(return(expr), name, parentenv, handler): unable to load sh
ared object '/Library/Frameworks/R.framework/Resources/modules//R_X11.so':
##   dlopen(/Library/Frameworks/R.framework/Resources/modules//R_X11.so, 6): Librar
y not loaded: /opt/X11/lib/libSM.6.dylib
##   Referenced from: /Library/Frameworks/R.framework/Versions/4.1/Resources/module
s/R_X11.so
##   Reason: image not found
```

```
## Could not load tcltk.  Will use slower R code instead.
```

```
## Loading required package: RSQLite
```

```
library(lubridate)
```

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

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

Read Data

```
## Reading the data

CovidData <- read.csv("BingCovidDataUpdated6sep2021.csv")
```

Data Cleaning Steps

```
## Change name of Updated column to UpdateDate with Date data type

CovidData <- CovidData %>%
  mutate(UpdatedDate = mdy(Updated))

## Filter data

CovidDataA <- CovidData %>%
  select(UpdatedDate, Country_Region, AdminRegion1, AdminRegion2, Latitude, Longitu
de, ConfirmedChange, Confirmed, RecoveredChange, Recovered, DeathsChange, Deaths) %
>%
  filter(Country_Region == "India") %>%
  filter(AdminRegion1 != "") %>%
  filter(AdminRegion2 != "")

## Arrange Data
CovidDataA <- CovidDataA %>%
  arrange(AdminRegion1, AdminRegion2)
CovidDataB <- sqldf("SELECT UpdatedDate AS Date__Date, Country_Region AS Country, A
dminRegion1 AS State, AdminRegion2 AS District, Latitude, Longitude, ConfirmedChang
e AS NewCases, Confirmed AS TotalCases, RecoveredChange AS NewRecovered, Recovered
AS TotalRecovered, DeathsChange AS NewDeaths, Deaths AS TotalDeaths FROM CovidDataA
ORDER BY State, District, Date__Date DESC", method = "name__class")
```

Exporting Data

```
## Export Data in CSV
write.csv(CovidDataB, "IndiaData.csv")
```