Assignment4

Apichat Photi-A

2023-04-01

Data Visualization Assignment 4: Final Project

Requirements We will finish this class by giving you the chance to use what you have learned in a practical context, by creating data visualizations from raw data.

Choose a dataset of interest from the City of Toronto's Open Data Portal (https://www.toronto.ca/city-government/data-research-maps/open-data/) or Ontario's Open Data Catalogue (https://data.ontario.ca/).

Using R and one other data visualization software (Excel or free alternative, Tableau Public, Python, any other tool you prefer), create two distinct visualizations from your dataset of choice.

For each visualization, describe and justify with course content or scholarly sources: - What software did you use to create your data visualization?

Answer I use tableau public and post it at https://public.tableau.com/views/COVID-19inTorontobyP_Apichat/Dashboard1?:language=en-US&:display_count=n&:origin=viz_share_link

- Who is your intended audience?

Answer The audience for the COVID-19 dashboard in Tableau are individuals and organizations interested in monitoring the spread and impact of COVID-19. This could include public health officials, policymakers, healthcare professionals, researchers, and the general public.

- What information or message are you trying to convey with your visualization?

Answer The dashboard created in Tableau displays information on the number of fatal and active cases, as well as monthly case counts from January 2020 to March 2023. Additionally, the dashboard includes a waffle chart that depicts the distribution of cases across different areas in Toronto. Apart from this, a graph representing the monthly case counts for Toronto has been created in RStudio.

- What design principles (substantive, perceptual, aesthetic) did you consider when making your visualization? How did you apply these principles? With what elements of your plots?

Answer For **substantive design principles**, I used a line plot to represent the number of positive cases each month, and a bubble plot to represent each age group. Additionally, I used a treemap to show the reported cases from each neighborhood area in Toronto.

Regarding **perceptual design principles**, I chose a color palette that is suitable for various groups of audiences, making the visualization easy to understand.

For **aesthetic design principles**, I paid attention to the overall look and feel of the visualization, ensuring that it is visually appealing and cohesive. For example, I used colors that are easy to understand when creating the graphs. Additionally, I maintained consistent font sizes and colors throughout the dashboard to create a cohesive and organized layout.

- How did you ensure that your data visualizations are reproducible? If the tool you used to make your data visualization is not reproducible, how will this impact your data visualization?

Answer To ensure that my code is easy to understand and reproduce, I wrote it using descriptive variable names and comments where necessary. I believe that comments are the best way to communicate with other coders and help them understand the code better. Additionally, I shared my code and data with others using **GitHub**, a cloud-based collaboration tool that makes it easy for others to reproduce my visualizations and build upon them.

To ensure transparency and ease of replication, I also documented the sources of all data used in my visualizations. This includes any pre-processing steps performed on the data, so that others can understand how the data was cleaned and transformed. In my case, the data used was from **Toronto Open Data**, which is freely available to everyone. Overall, by following these best practices, I can ensure that my code is well-documented, easy to understand, and can be easily reproduced by myself or others.

- How did you ensure that your data visualization is accessible?

Answer The dashboard is available in tableau public.

- Who are the individuals and communities who might be impacted by your visualization?

Answer The COVID-19 data visualization dashboard might impact a wide range of individuals and communities. Here are some groups that may be affected:

General Public - The dashboard provides information about the number of positive cases, active cases, and fatalities related to COVID-19, which can help the general public understand the severity of the pandemic in their area. **Healthcare professionals** - The visualization can be used by healthcare professionals to track the spread of the virus, identify areas with high numbers of cases, and develop strategies to manage the pandemic each month. **Government officials** - The dashboard can be used by government officials to make informed decisions about policies and interventions to manage the pandemic. **Businesses** - The visualization can help businesses understand the impact of the pandemic on their operations, such as by showing areas with high numbers of cases that may be risky for business operations. **Vulnerable communities** - The dashboard can be used to identify areas with high numbers of cases that may be at a higher risk for COVID-19, such as communities with high rates of poverty or limited access to healthcare.

- How did you choose which features of your chosen dataset to include or exclude from your visualization?

Answer When choosing features of a dataset to include or exclude in a visualization, I considered factors such as relevance, clarity, completeness and accuracy. Relevant and useful features for the visualization were selected to achieve a complete and accurate representation of the data. By taking these factors into account, I could create a clear, understandable, and informative visualization that accurately represents the data.

- What ‘underwater labour’ contributed to your final data visualization product?

Answer To obtain the required dataset for the COVID-19 dashboard, collaboration was necessary between various healthcare facilities such as clinics, hospitals, and primary care units in all areas of Toronto. The dataset included information on all patients who received treatment or underwent self-treatment at home or home isolation. The data also required updating of patient outcomes from secondary to tertiary care units or hospitals in Toronto. By collaborating with healthcare facilities and ensuring accurate data collection, a comprehensive dataset was obtained for the COVID-19 dashboard.

Your final submission document should include: - Two data visualizations - Written descriptions for each data visualization - Link to your dataset of choice - Complete and commented code as an appendix (for your visualization made with R, and for the other, if relevant)

This assignment is intentionally open-ended - you are free to create static or dynamic data visualizations, maps, or whatever form of data visualization you think best communicates your information to your audience of choice!

Total word count should not exceed (as a maximum) 1000 words

Why am I doing this assignment?

This assignment tests your ability to apply the skills and knowledge acquired throughout the class and assesses learning outcomes 1, 2, and 3: 1. Develop ability to create and customize data visualizations start to finish in R 2. Build an understanding of general design principles for creating accessible/equitable data visualizations in R and other software 3. Build an understanding of data visualization as purposeful/telling a story (and the ethical/professional implications thereof)

```r
#install.packages("opendatatoronto")
library(opendatatoronto)
library(dplyr)
```

```
##
## 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)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.4
## v ggplot2   3.4.1     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.1     v tidyr     1.3.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
# get package
package <- show_package("64b54586-6180-4485-83eb-81e8fae3b8fe")
package
```

```
## # A tibble: 1 x 11
##   title             id    topics civic_issues publisher excerpt dataset_category
##   <chr>             <chr> <chr>  <chr>        <chr>     <chr>   <chr>
## 1 COVID-19 Cases i~ 64b5~ Health <NA>         Toronto ~ Line-l~ Table
## # i 4 more variables: num_resources <int>, formats <chr>, refresh_rate <chr>,
## #   last_refreshed <date>
```

```r
# get all resources for this package
resources <- list_package_resources("64b54586-6180-4485-83eb-81e8fae3b8fe")

# identify datastore resources; by default, Toronto Open Data sets datastore resource format to CSV for
datastore_resources <- filter(resources, tolower(format) %in% c('csv', 'geojson'))

# load the first datastore resource as a sample
data <- filter(datastore_resources, row_number()==2) %>% get_resource()
data
```

```
## # A tibble: 395,602 x 15
##     X_id Assigned_ID Outbreak.Associated Age.Group      Neighbourhood.Name FSA
##    <int>       <int> <chr>               <chr>          <chr>              <chr>
##  1     1           1 Sporadic            50 to 59 Years Willowdale East    M2N
##  2     2           2 Sporadic            50 to 59 Years Willowdale East    M2N
##  3     3           3 Sporadic            20 to 29 Years Parkwoods-Donalda  M3A
##  4     4           4 Sporadic            60 to 69 Years Church-Yonge Corr~ M4W
##  5     5           5 Sporadic            60 to 69 Years Church-Yonge Corr~ M4W
##  6     6           6 Sporadic            50 to 59 Years Newtonbrook West   M2R
##  7     7           7 Sporadic            80 to 89 Years Milliken           M1V
##  8     8           8 Sporadic            60 to 69 Years Willowdale West    M2N
##  9     9           9 Sporadic            50 to 59 Years Willowdale East    M2N
## 10    10          10 Sporadic            60 to 69 Years Henry Farm         M2J
## # i 395,592 more rows
## # i 9 more variables: Source.of.Infection <chr>, Classification <chr>,
## #   Episode.Date <chr>, Reported.Date <chr>, Client.Gender <chr>,
## #   Outcome <chr>, Ever.Hospitalized <chr>, Ever.in.ICU <chr>,
## #   Ever.Intubated <chr>
```

```r
#looka at the data
head(data)
```

```
## # A tibble: 6 x 15
##    X_id Assigned_ID Outbreak.Associated Age.Group      Neighbourhood.Name  FSA
##   <int>       <int> <chr>               <chr>          <chr>               <chr>
## 1     1           1 Sporadic            50 to 59 Years Willowdale East     M2N
## 2     2           2 Sporadic            50 to 59 Years Willowdale East     M2N
## 3     3           3 Sporadic            20 to 29 Years Parkwoods-Donalda   M3A
## 4     4           4 Sporadic            60 to 69 Years Church-Yonge Corri~ M4W
## 5     5           5 Sporadic            60 to 69 Years Church-Yonge Corri~ M4W
## 6     6           6 Sporadic            50 to 59 Years Newtonbrook West    M2R
## # i 9 more variables: Source.of.Infection <chr>, Classification <chr>,
## #   Episode.Date <chr>, Reported.Date <chr>, Client.Gender <chr>,
## #   Outcome <chr>, Ever.Hospitalized <chr>, Ever.in.ICU <chr>,
## #   Ever.Intubated <chr>
```

```r
#select potential columns
data |> select("Age.Group", "Neighbourhood.Name", "Source.of.Infection", "Client.Gender") |>
  sample_n(10)
```

```
## # A tibble: 10 x 4
##    Age.Group      Neighbourhood.Name           Source.of.Infection Client.Gender
##    <chr>          <chr>                        <chr>              <chr>
##  1 40 to 49 Years Danforth-East York           No Information     MALE
##  2 30 to 39 Years Pelmo Park-Humberlea         No Information     MALE
##  3 20 to 29 Years Westminster-Branson          No Information     FEMALE
##  4 80 to 89 Years Bayview Woods-Steeles        Community          FEMALE
##  5 30 to 39 Years Church-Yonge Corridor        No Information     MALE
##  6 60 to 69 Years Glenfield-Jane Heights       Household Contact  MALE
##  7 70 to 79 Years Annex                        Community          FEMALE
##  8 40 to 49 Years Birchcliffe-Cliffside        Household Contact  FEMALE
##  9 60 to 69 Years Agincourt South-Malvern West No Information     MALE
## 10 30 to 39 Years Long Branch                  Community          MALE
```

```
#install libraries
#install.packages("maps")
#install.packages("dplyr")
library(maps)
```

```
##
## Attaching package: 'maps'

## The following object is masked from 'package:purrr':
##
##     map
```

```
library(dplyr)
library(ggplot2)
glimpse(data)
```

```
## Rows: 395,602
## Columns: 15
## $ X_id                <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,~
## $ Assigned_ID         <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,~
## $ Outbreak.Associated <chr> "Sporadic", "Sporadic", "Sporadic", "Sporadic", "S~
## $ Age.Group           <chr> "50 to 59 Years", "50 to 59 Years", "20 to 29 Year~
## $ Neighbourhood.Name  <chr> "Willowdale East", "Willowdale East", "Parkwoods-D~
## $ FSA                 <chr> "M2N", "M2N", "M3A", "M4W", "M4W", "M2R", "M1V", "~
## $ Source.of.Infection <chr> "Travel", "Travel", "Travel", "Travel", "Travel", ~
## $ Classification      <chr> "CONFIRMED", "CONFIRMED", "CONFIRMED", "CONFIRMED"~
## $ Episode.Date        <chr> "2020-01-22", "2020-01-21", "2020-02-05", "2020-02~
## $ Reported.Date       <chr> "2020-01-23", "2020-01-23", "2020-02-21", "2020-02~
## $ Client.Gender       <chr> "FEMALE", "MALE", "FEMALE", "FEMALE", "MALE", "MAL~
## $ Outcome             <chr> "RESOLVED", "RESOLVED", "RESOLVED", "RESOLVED", "R~
## $ Ever.Hospitalized   <chr> "No", "Yes", "No", "No", "No", "No", "No", "Yes", ~
## $ Ever.in.ICU         <chr> "No", "No", "No", "No", "No", "No", "No", "No", "N~
## $ Ever.Intubated      <chr> "No", "No", "No", "No", "No", "No", "No", "No", "N~
```

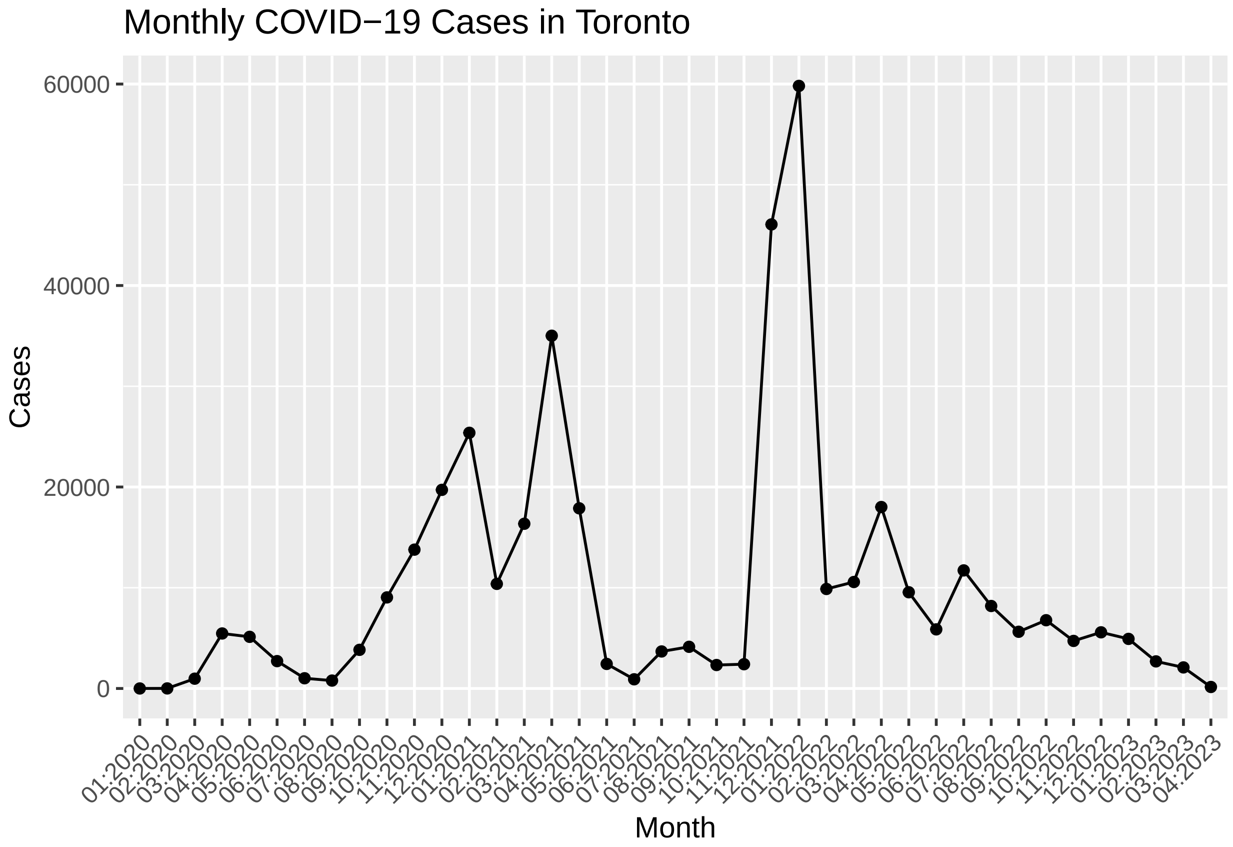
```
# Load required packages
library(tidyverse)
library(lubridate)
library(sf)
```

```
## Linking to GEOS 3.10.2, GDAL 3.4.2, PROJ 8.2.1; sf_use_s2() is TRUE
```

```
#install.packages("leaflet")
library(leaflet)
library(dplyr)

# Set working directory and download data from Toronto Open Data website
#setwd("~/Desktop")
#download.file("https://ckan0.cf.opendata.inter.prod-toronto.ca/download_resource/15c71d2a-1d9c-41e7-b7
#              destfile = "covid19_data.csv")

# Read data and filter for Toronto cases only
covid_data <- data
view(covid_data)
# Convert date to proper format
covid_data$Reported.Date <- ymd(covid_data$`Reported.Date`)

# Create new columns for day, week, and month
covid_data <- covid_data %>%
  mutate(Day = day(Reported.Date),
         Year = year(Reported.Date),
         Month = month(Reported.Date, label = TRUE))
```

```
#manage the data


#create new column for month and year
df <- data %>%
  mutate(Month = format(as.Date(data$`Reported.Date`), "%m"),
         Year = format(as.Date(data$`Reported.Date`), "%Y"))

#group by month
count_by_month <- df %>%
  group_by(Year, Month) %>%
  count()

count_by_month$month_year <- paste(count_by_month$Month, count_by_month$Year, sep=":")

#arrange the data
count_by_month$month_year <- factor(count_by_month$month_year, levels = count_by_month$month_year)
```

```
#create ggplot object
p <- ggplot(data = count_by_month, aes(x = month_year, y = n)) +
  geom_point() + geom_line(aes(group = "1")) +
  labs(title = "Monthly COVID-19 Cases in Toronto",
       x = "Month",
       y = "Cases") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1))

# Display plot
p
```


Monthly COVID−19 Cases in Toronto
Cases
60000
40000
20000
0
Month
01:2020
02:2020
03:2020
04:2020
05:2020
06:2020
07:2020
08:2020
09:2020
10:2020
11:2020
12:2020
01:2021
02:2021
03:2021
04:2021
05:2021
06:2021
07:2021
08:2021
09:2021
10:2021
11:2021
12:2021
01:2022
02:2022
03:2022
04:2022
05:2022
06:2022
07:2022
08:2022
09:2022
10:2022
11:2022
12:2022
01:2023
02:2023
03:2023
04:2023