Water Temperature at Beaches in Toronto

Yu, Lexun

Sep 12, 2024

This is a Quarto file that downloads a dataset using opendatatoronto, cleans it, and makes a graph.

Plan

The dataset I am interested in would need to have the date, and the water temperature. A quick sketch of a dataset that would work is Figure 1a, I am interested in the water temperature each month, the table would be like Figure 1b:

Date	Water Temp
2021/7/1	21
2021/7/2	22
2021/7/3	23
2021/7/4	24
2021/7/5	21
2021/7/6	23
2021/7/7	27
2021/7/8	28
2021/7/9	25
2021/7/10	30
2021/7/11	31
2021/7/12	32
2021/7/13	33

(a) Quick sketch of a dataset

Year	Water Temp
2008	21
2009	22
2010	23
2011	24
2012	21
2013	23
2014	27
2015	28
2016	25
2017	30
2018	31
2019	32

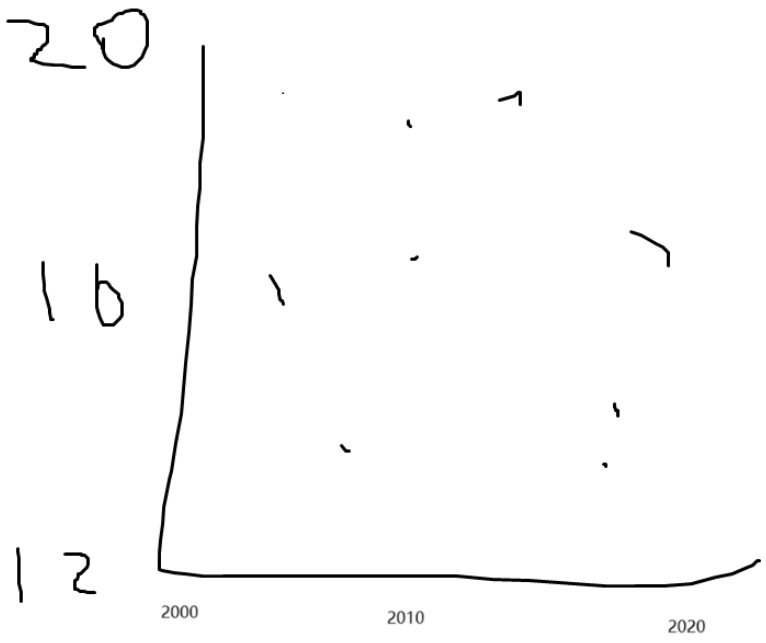
(b) Quick sketch of a table of the average water temperature each month

Figure 1: Sketches of a potential dataset and table related to water temperature.

Them I will draw a geom_point graph like Figure 2a:

Simulate

This document uses R Core Team (2024) and Wickham (2016)



(a) Quick sketch of a graph

Figure 2: Sketches of a potential dataset and graph related to water temperature.

After examining the raw data, I found that there is only data between May and September. So, I am only generating simulated data between May and September

```
set.seed(853)

simulated_temp_data <-
    tibble(
        date = rep(x = as.Date("2016-05-01") + c(0:152), times = 1),
        water_temp = rpois(
            n = 153 * 1,
            lambda = 16
        )
    )

head(simulated_temp_data)
```

```
# A tibble: 6 x 2
  date       water_temp
  <date>          <int>
1 2016-05-01         14
2 2016-05-02         15
3 2016-05-03         29
4 2016-05-04         10
5 2016-05-05          8
6 2016-05-06         12
```

Acquire and display the raw data

The following terminal output displays the raw data obtained from opendatatoronto.

```
# A tibble: 6 x 13
  `_id` dataCollectionDate beachName       windSpeed windDirection airTemp rain
  <int> <chr>              <chr>               <int> <chr>            <int> <chr>
1     1 2010-08-03         Marie Curtis P~         5 SW                  31 Yes
2     2 2010-08-03         Sunnyside Beach         5 SW                  31 Yes
3     3 2010-08-03         Hanlan's Point~         5 SW                  31 Yes
4     4 2010-08-03         Gibraltar Poin~         5 SW                  31 Yes
5     5 2010-08-03         Centre Island ~         5 SW                  31 Yes
6     6 2010-08-03         Ward's Island ~         5 SW                  31 Yes
# i 6 more variables: rainAmount <int>, waterTemp <dbl>, waterFowl <int>,
#   waveAction <chr>, waterClarity <chr>, turbidity <dbl>
```

Clean the data

Read the csv

```
raw_toronto_beaches_data <-
  read_csv(
    file = "input/toronto_beaches.csv",
    show_col_types = FALSE
  )
head(raw_toronto_beaches_data)
```

```
# A tibble: 6 x 13
  `_id` dataCollectionDate beachName       windSpeed windDirection airTemp rain
  <dbl> <date>             <chr>               <dbl> <chr>           <dbl> <chr>
1     1 2010-08-03         Marie Curtis P~         5 SW                 31 Yes
2     2 2010-08-03         Sunnyside Beach         5 SW                 31 Yes
3     3 2010-08-03         Hanlan's Point~         5 SW                 31 Yes
4     4 2010-08-03         Gibraltar Poin~         5 SW                 31 Yes
5     5 2010-08-03         Centre Island ~         5 SW                 31 Yes
6     6 2010-08-03         Ward's Island ~         5 SW                 31 Yes
# i 6 more variables: rainAmount <dbl>, waterTemp <dbl>, waterFowl <dbl>,
#   waveAction <chr>, waterClarity <chr>, turbidity <dbl>
```

Clean names

```
cleaned_beaches_data <-
    clean_names(raw_toronto_beaches_data)
head(cleaned_beaches_data)
```

```
# A tibble: 6 x 13
     id data_collection_date beach_name wind_speed wind_direction air_temp rain
  <dbl> <date>               <chr>           <dbl> <chr>              <dbl> <chr>
1     1 2010-08-03           Marie Cur~          5 SW                    31 Yes
2     2 2010-08-03           Sunnyside~          5 SW                    31 Yes
3     3 2010-08-03           Hanlan's ~          5 SW                    31 Yes
4     4 2010-08-03           Gibraltar~          5 SW                    31 Yes
5     5 2010-08-03           Centre Is~          5 SW                    31 Yes
6     6 2010-08-03           Ward's Is~          5 SW                    31 Yes
# i 6 more variables: rain_amount <dbl>, water_temp <dbl>, water_fowl <dbl>,
#   wave_action <chr>, water_clarity <chr>, turbidity <dbl>
```

keep only the necessary date and water temperature information

```
cleaned_beaches_data <-
  cleaned_beaches_data |>
  select(
    data_collection_date,
    water_temp
  )

cleaned_beaches_data <-
  cleaned_beaches_data |>
  rename(
    date = data_collection_date,
    temp = water_temp
  )

head(cleaned_beaches_data)
```

```
# A tibble: 6 x 2
  date        temp
  <date>     <dbl>
1 2010-08-03  22.6
2 2010-08-03  21.9
3 2010-08-03  24.3
4 2010-08-03  21.3
5 2010-08-03  21.3
6 2010-08-03  21.4
```

```
names(cleaned_beaches_data)
```

```
[1] "date" "temp"
```

Write the new csv

```
write_csv(
  x = cleaned_beaches_data,
  file = "output/cleaned_beaches_data.csv"
)
```

```
beaches_clean <-
    read_csv("output/cleaned_beaches_data.csv", show_col_types = FALSE)
```

Group the cleaned data by year.

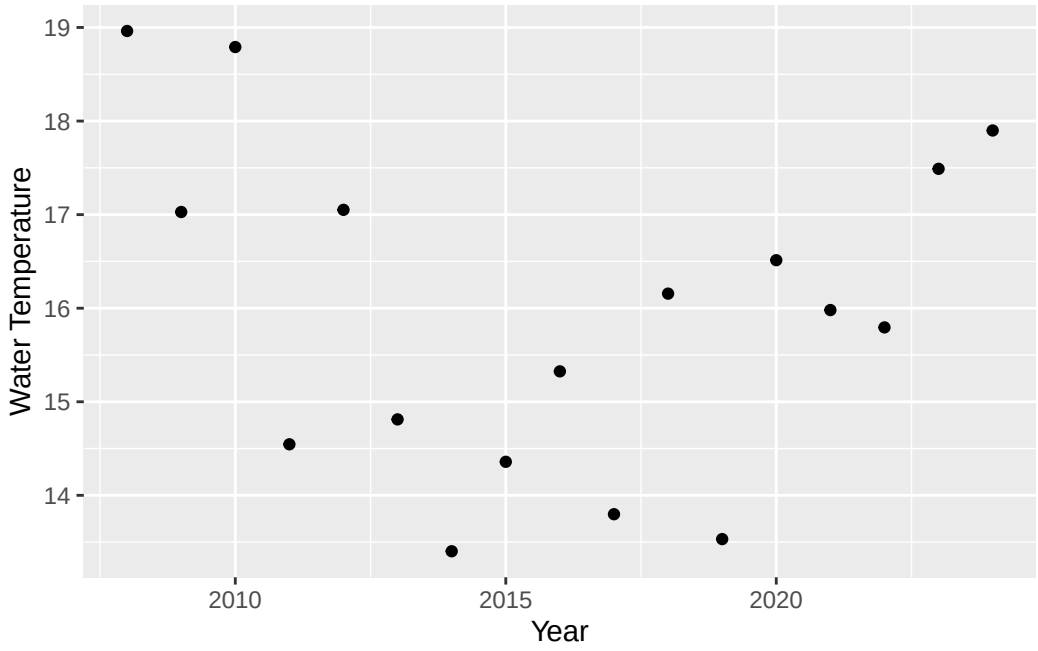
```
summary_data <- beaches_clean |>
  mutate(temp_year = year(date)
  ) |>
  arrange(year(date)) |>
  drop_na(temp) |>
  summarise(number_temp = mean(temp),
            .by = temp_year)

write_csv(summary_data, file = "output/cleaned_beaches_data_year.csv")
summary_data |> kable()
```

temp_year	number_temp
2008	18.96261
2009	17.02802
2010	18.79092
2011	14.54550
2012	17.05103
2013	14.81171
2014	13.40234
2015	14.35840
2016	15.32493
2017	13.79816
2018	16.15567
2019	13.53229
2020	16.51292
2021	15.98010
2022	15.79472
2023	17.48933
2024	17.89932

Explore

I can now make a graph of how water temperature change over time.

```r
summary_data |>
  ggplot(aes(x = temp_year, y = number_temp)) +
  geom_point() +
  labs(x = "Year", y = "Water Temperature") +
  scale_color_brewer(palette = "Set1") +
  theme(legend.position = "bottom")
```



Bibliography

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.