Spotify Data

Yanzun Jiang,

October 10, 2024

Table of contents

1 Introduction

To conduct the analysis of songs on Spotify, a dataset downloaded from Spotify (2024) was utilized, as described in Section 2. Based on the initial findings, it was observed that higher air temperatures are correlated with higher water temperatures, while higher water temperatures coincided with an increase in waterfowl activity in some extent(Section 3). Also, supplementary insights are provided in Section A.

2 Data

2.1 Overview

The dataset used in this analysis is the songs of David Tao sourced from Spotify (Spotify (2024)). It records various aspects of the songs by David from 1997 to present.

The variables analyzed in this study are listed below:

`album_release_date`: the date that the album was released

`energy`: the energy of the song

`key`: the main key of the song

`valence`: the degree of happiness of the song

`tempo`: the pace of the song

The dataset was accessed using the `spotifyr` package (Thompson et al. (2022)). For the analysis, the R programming language was employed (R Core Team (2023)), utilizing the `tidyverse` (Wickham et al. (2019)) package for data cleaning, transformation, visualization, and the `here` package (Müller (2020)) accessing the data in this analysis. Afterward, the cleaned dataset was processed and tested using additional functions from the `tidyverse` package (Wickham et al. (2019)).

2.2 Preview of the Dataset

3 Results

3.1 Gain from Distribution of the Variables

The histogram of air temperatures, as shown in **?@fig-air**, reveals that most recorded air temperatures at Toronto beaches fall between 12°C and 30°C, with a noticeable peak around 20°C. The distribution is approximately symmetric, indicating no significant skew, meaning

that air temperatures are evenly spread around the central value. This suggests that beach monitoring predominantly occurs during periods of moderate and comfortable weather, which is conducive to outdoor activities. Temperatures below 8°C and above 33°C are much less frequent, indicating that extreme weather conditions are rare, emphasizing the temperate climate in Toronto during beach usage periods.

In contrast, the water temperature histogram, seen in **?@fig-water**, is left-skewed, meaning that while most temperatures cluster between 8°C and 25°C, there is a tail extending toward lower temperatures. A peak is observed around 18°C, representing the most common water temperatures during beach season. The left skew indicates that colder water temperatures are infrequent but present, typically occurring outside peak summer months. This distribution reflects how water retains heat more slowly than air, remaining relatively cool even as air temperatures rise, which is characteristic of large bodies of water.

Lastly, the waterfowl observations histogram, as seen in **?@fig-fowl**, displays a highly right-skewed distribution. Most days had a low count of waterfowl, with the majority of observations falling between 0 and 50 observations. However, a long tail extends to the right, indicating that while large counts of waterfowl are rare, they do occur on certain days. This right skew suggests occasional spikes in bird activity, possibly linked to seasonal migrations or environmental conditions that attract more birds to the beaches. These sporadic peaks could have implications for water quality, as larger groups of waterfowl might increase the risk of contamination.

These histograms (**?@fig-air**, **?@fig-water**, and **?@fig-fowl**) offer valuable insights into the key environmental variables affecting Toronto beaches. The symmetric air temperature distribution, left-skewed water temperature, and highly right-skewed waterfowl observations each reveal distinct patterns that inform public health monitoring and beach management decisions, ensuring that the environment remains safe and enjoyable for visitors.

A Appendix

A.1 Data Cleaning

The data cleaning process involved tidying the dates, filtering out useless columns from the raw dataset, and filtering out observations with NAs.

A.2 Summary Statistics of the Data

The summary statistics provides key statistical insights into the environmental variables analyzed at Toronto beaches. For each variable, the summarize tables (**?@tbl-summarise-air**, **?@tbl-summarise-water**, and **?@tbl-summarise-fowl**) reports the mean, median, variance, minimum value, maximum value, and interquartile range of those vairables, offering a concise overview of range, central tendencies, and variability in the data.

References

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Spotify. 2024. "Spotify Music Data." Data retrieved from https://open.spotify.com.

Thompson, Charlie, Daniel Antal, Josiah Parry, Donal Phipps, and Tom Wolff. 2022. *Spotifyr: R Wrapper for the 'Spotify' Web API*. https://github.com/charlie86/spotifyr.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.