Portfolio Problem 2: Data Wrangling

Data Overview

The data for this portfolio problem is from the National Weather Service. The data includes sixteen months of forecasts and observations from 167 cities, as well as a separate data set with information about those cities and some other American cities. There is some data processing already completed here, so to cite your data source, you can just include the name of the National Weather Service (you don't need a link). **Be sure to read the assessment guidelines at the end of this file!**

You may use additional data beyond those that are given here, but the data source must be reliable and it must be cited.

You can import the data from the following URLs:

- https://www.math.carleton.edu/ckelling/data_science/weather_forecasts.csv
- https://www.math.carleton.edu/ckelling/data_science/forecast_cities.csv
- https://www.math.carleton.edu/ckelling/data_science/outlook_meanings.csv

weather_forecasts.csv

variable	class	description
date	date	date described by the forecast and observation
city	factor	city
state	factor	state or territory
high_or_low	factor	whether the forecast is for the high temperature or the low temperature
forecast_hours_before	integer	the number of hours before the observation (one of 12, 24, 36, or 48)
observed_temp	integer	the actual observed temperature on that date (high or low)
forecast_temp	integer	the predicted temperature on that date (high or low)
observed_precip	double	the observed precipitation on that date, in inches; note that some observations lack an indication of precipitation, while others explicitly report 0
forecast_outlook	factor	an abbreviation for the general outlook, such as precipitation type
possible_error	factor	either (1) "none" if the row contains no potential errors or (2) the name of the variable that is the cause of the potential error

forecast_cities.csv

variable	class	description
city	character	city
state	character	state or territory
lon	double	longitude
lat	double	latitude
koppen	character	Köppen climate classification
elevation	double	elevation in meters

variable	class	description
distance_to_coast	double	distance to coast in miles
wind	double	mean wind speed
elevation_change_four	double	greatest elevation change in meters out of the four closest points to this city in a collection of elevations used by the team at Saint Louis University
elevation_change_eight	double	greatest elevation change in meters out of the eight closest points to this city in a collection of elevations used by the team at Saint Louis University
avg_annual_precip	double	average annual precipitation in inches

outlook_meanings.csv

variable	class	description
forecast_outlook	character	an abbreviation for the general outlook, such as precipitation type
meaning	character	the meaning of that abbreviation

Assignment

Your goal is to learn which areas of the U.S. struggle with weather prediction and explore possible reasons why. Specifically, you will focus on the error in high and low temperature forecasting, and may wish to also consider precipitation and outlook.

You should write a **short** blog post describing your findings. I envision an introductory paragraph that provides some context to your data, and a couple paragraphs outlining your findings. I'm looking for something that is insightful and well-crafted. Include 1-2 (no more than 2) visualizations to help communicate your findings. That's it.

You should write your blog post in R Markdown, create any graphics using `ggplot2`, and use tools from this class for data wrangling. To submit your work, push both your R Markdown (.Rmd) file and knitted PDF document to GitHub and Gradescope. **Only one person needs to submit per group.** Do not forget to give your post an informative title!

Assessment

Below, I've included some items that I will consider for assessment of your second portfolio problem.

Reproducibility

- All necessary components of the problem were pushed to GitHub and linked to Gradescope
- The code and blog post are contained in a single .Rmd file
- The pdf does not contain visible code
- The .Rmd file contains all necessary code to reproduce the summaries/graphic(s)
- The .Rmd file does not contain any unnecessary code
- The .Rmd file does not produce any unnecessary content (package loading messages, warnings, etc.)
- Code is **well commented**, so a randomly selected classmate could easily navigate it

Scope

- Both weather_forecasts.csv and cities.csv were used in the analysis.
- An investigation of the accuracy of the forecasts is included
- An investigation of possible reasons why some areas struggle with accuracy is included
- **Depth of analysis:** pursue something insightful!

Data Wrangling

- Data sets were correctly imported
- Data sets were correctly joined
- Any necessary data wrangling/reshaping was correctly implemented
- Summary statistics were correctly calculated for the task at hand
- Factor levels, dates, and times were manipulated correctly
- Factor levels, dates, and times were readable/interpretable in any summaries/plots

Data Visualization

- The graphic(s) are correctly constructed and appropriate for the task at hand
- The graphic(s) are interpretable without reading the text of the blog post
- Aesthetics (color, shape, etc.) faithfully reflect the representation of the underlying data
- The graphic(s) have readable axis labels, units, and legends (where appropriate)
- The graphic(s) lists the source(s) of the data
- The graphic(s) contains no color, symbolism, or text that is irrelevant to the question it seeks to answer or argument that it seeks to make.

Communication

- An informative title is included for the blog post
- The blog post contains very few grammatical mistakes
- The blog post contains very few spelling mistakes
- The blog post clearly introduces the data set in your own words (it cannot be copied from the prompt)
- The data is faithfully represented in your analysis/communication
- The blog post clearly summarizes the major takeaways from your analysis
- An insightful result is communicated about weather forecasts