12-709 Data Analytics for Engineered Systems (CEE) - Fall 2020
Homework 3
Part A "soft deadline" October 16 (no submission required!)
Parts A and B Due on Canvas: October 28 before class - 15% penalty per day late

Basic R Model Building

Learning Objectives: This individual assignment is intended to:

- Practice basic R operations and experience programming-based data analytics;
- Revisit familiar data analysis approaches and explore how done in R;
- Manage imperfect data and perform analysis on it in R;
- Demonstrate your ability to do data cleaning and data manipulation in R;
- Practice exploratory data analysis, visualization and analytics tools in R.

[I am making two parts, to encourage you to start early. Homework 3B will be posted soon.]

General Directions:
As mentioned in lecture and syllabus, writing skills are a key component of your homework grades (and your future careers). Please take care in writing, editing, formatting, and printing your work before submitting it. Your deliverable should look professional. While I have asked you to 'document' your processes, the final result of this assignment can be submitted as a series of homework questions and answers and does NOT need to be organized as a project report. Use charts and significant figures appropriately.

You may use any tools, e.g., spreadsheet or SQL in parallel (e.g., to ensure you're getting the right answers and check your work). However, all of the *analysis* tasks should be completed with code, ideally R, but feel free to use another like Python. No penalty for inefficient code.

Deliverables: You must submit **SHORT** three-part answers, including code used, and resulting tables/charts. You can submit (a) a single PDF file of your three-part answers from Microsoft Word (with charts) as well as R code, **OR** (b) a single R markdown or notebook **knitted PDF** with questions, answers, and code chunks all in one place. Regardless of format, you need to generate results from running code – you cannot just take screenshots of data frames as answers. Please upload separate files, not ZIP files to make grading easier.

Please submit all work for parts A and B together to Canvas. Markdown format is required for Part B, but **we will award 2 bonus points for doing both parts in markdown or notebook format.**

Please follow the Canvas **Discussion** thread "Questions and Answers about Homework 3".

Overall hint: if we ask for something multiple times (such as "do this for each city"), it often means that coding it as a loop or iteration will save time!

Introduction

It is often said that 80% of data analysis is spent on the cleaning and preparing data. It's not just a first step, but must be repeated many times over the course of analysis as new problems about the data come to light. Real-world data tend to have errors, be incomplete, noisy, and

inconsistent. Data cleaning methods attempt to normalize the data, fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

In this assignment, you'll work on a "dirty" dataset and perform analysis on it. As mentioned above, some of your time will be spent cleaning and manipulating the data. By completing the assignment, you will become familiar with the most practical Base R (and later, tidyverse) operations used in data analytics. Note this is our first homework in R, and **you do not need to use any statistical tools beyond summary statistics** in R (e.g., mean, variance, counts, etc.). That said, you are welcome to use higher-level statistical tools - if you feel appropriate.

Data Description
You are provided with "raw" data from OpenWeatherMap, a service that archives weather data from around the world. The data is hourly weather for several European cities over a 5-year period. It is a large CSV file. There is also a Data Dictionary posted on Canvas.

Part A [25 pts]: Exploring Base R Features

Use only Base R functions and operators (including your own written functions using Base R) to answer the questions in Part A. Please do not use other packages, and write code such that the requested answer is explicitly provided in the console results (i.e., don't just write code to have it create a result in the workspace).

Question 1 [10 pts] – **Initial data processing and exploratory data analysis**

Note: each of these can be less than 5 lines of code, often only 1 line!

~~a) [2 pts]: Read in the raw CSV file, and create a data frame called *raw_data* that contains only the date (dt or dt_iso), city_id, and five other variables (temp, humidity, wind_speed, rain_1h, and weather_id). As needed, change the variable classes, and specify factors, to make sure the rest of your analysis will work correctly!~~

~~b) [2 pts]: How many observations are there overall in the raw data? Also, make a 'table' to show how many raw observations are there for each of the city_id's specified.~~

~~c) [2 pts]: Perform a high-level EDA of the data frame created in 1a. As always, describe the dataset, and identify potential data problems from the EDA results (but do not fix/clean the data yet). Discuss which of the weather variables have the least or most NA (missing) values. **Note: this three-part answer should be relatively longer to discuss the EDA results!**~~

~~d) [1 pts] Find the mean and standard deviation of temperature and wind speed using all of the raw data.~~

~~e) [1 pts] Calculate again the means of the two variables but after 'trimming' 2% of data from each end. Describe the results and what they tell you about data cleaning needed. Hint: use help if needed.~~

f) [2 pts] Consider the following table that matches the ids and names of the cities:

Id	city_name
3143244	Oslo
2950159	Berlin
2988507	Paris
3169070	Rome
2643743	London
6458923	Lisbon
1111111	Pittsburgh

i) Create a dataframe named *city_df* that contains the information provided in the table above. Name the columns "Id" and "city_name" as shown.

ii) **Join** *raw_data* and *city_df* data frames to create a data frame called *joined_data* that has an additional column with the city name associated with each observation (column must be called "city_name").

Question 2 [8 pts] – Data Manipulation and Turning data into information

a) [2 pts] Modify the *joined_data* data frame by removing: (1) all duplicate records from the data frame, (2) any records where the temperature is equal to zero and (3) any records where the humidity is greater than 100. Name the resulting data frame *removed_joined_data*.

b) [2 pts]: What proportion of the observations in *removed_joined_data* (overall, not at city level) is the temperature below 0 degrees Celsius AND the wind speed is greater than 5?

c) [4 pts]: Make PivotTable-like summaries (for each city, and showing city names) for the following using *removed_joined_data*:
i) Average temperature and wind speed
ii) Standard deviation of the temperature and wind speed
ii) Minimum and maximum temperature and wind speed
iv) Frequency of 'clear' (weather ID=800) conditions

Question 3 [7 pts] Data visualization in Base R

a) [3 pts]: Assume you want to understand whether there are data gaps in the hourly records for each city (e.g., days/weeks/months with no records due to sensor failure). Make a new variable that tracks 'time between consecutive records' for *removed_joined_data* AND for each city create a table of the 'time between' to try to find whether any or all of the cities have gaps. Discuss whether they are generally day/week/month long type gaps.

Hint: think of how you would do this in Excel – its similar in Base R. And of course, don't forget that if you get stuck you can always search the Internet for help! Just cite sources used.

b) [4 pts]: For each of the following, create two boxplots for the hourly data for London, one using *joined_data* and the other using *removed_joined_data* (you will make four total boxplots):
(i) temperature (variable *temp*)
(ii) hourly rainfall (*rain_1h*).

Make sure the plots are properly formatted and labeled (don't worry about color!). Discuss outliers identified by plots in terms of number of data points and how much the outliers differ.

Part B [30 pts]: Data analysis with the R tidyverse package

The following questions *should* be done by leveraging functions in the tidyverse R package (including pipe). Of course, you can still use some Base R code, but any answers using no tidyverse functions will receive a small deduction. *And remember that markdown is required*.

Question 4 [10 pts]: Tidying the dataset

a) [1 pt] Begin again with the "raw" data from the CSV file, and again create a dataframe called *raw_data_tidy* including only the date (dt or dt_iso), city_id, and seven other variables (temp, humidity, wind_speed, rain_1h, snow_1h, weather_main and weather_description). Note the last two are slightly different than before.

b) [1 pt] Remove duplicate values from *raw_data_tidy*.

c) [2 pts] Show all unique text weather_descriptions associated with the weather_main values of Cloudy and Rain in *raw_data_tidy*. Use assumptions to reduce the number of unique weather_descriptions (e.g., observe similarities in the weather_description text strings, and make assumptions to combine them to slightly reduce the number of weather_descriptions by using regular expressions!)

d) [6 pts] Using your findings from Part A on other problems with the raw data (outliers, missing values, data gaps, etc.), use tidyverse to perform at least **three** types of data cleaning to fix **major** data errors in *raw_data_tidy*. Two of these should be filling missing values and fixing outliers. Of course, be sure to document your cleaning steps and discuss and show how the clean data compares to the raw data. Name the cleaned dataframe *clean_data_tidy*.

Question 5 [6 pts]: Custom functions, iterations, conditions, vectorized methods

a) [2 points] Define a custom R function to compute apparent temperature. We use wind chill temperature for apparent temperature when temperatures are at or below 50 °F and wind speeds above 3 mph. The wind chill temperature (units of degrees F) can be calculated as follows[1]:

$$T_{wc} = 35.74 + 0.6215T - 35.75*V^{0.16} + 0.4275*T(V^{0.16})$$

where T is air temperature in degrees Fahrenheit and V is wind speed in miles per hour

b) [4 pts] Using your apparent temperature function above, add a column to *clean_data_tidy* for apparent temperatures of each hourly temperature value. Do this with three different methods: (1) by using a for loop, (2) with map or apply, and (3) with a dplyr function. For each way, record the elapsed run time (by using the R Sys.time function). Compare the elapsed times needed to run each of the three ways and discuss which seems to be the most efficient.

Question 6 [12 pts]: Assessing seasonal differences in weather

Answer the following questions with clean_data_tidy assuming northern meteorological temperate seasons (winter is December-February, spring is March-May, summer is June-August, and autumn/fall is September-November).

[1] See example calculator and link to formula for wind chill on https://www.weather.gov/epz/wxcalc_windchill

a) [2 pts] Make a data structure of the total raining hours for each city during each season in every year. Show the values in the data structure for 2013 for each city and each season.

b) [2 pts] Make a data structure of the total hours where the temperature is below 0 degrees Celsius for each city during each season in every year. Show the values in the data structure for 2013 for each city and each season.

The remaining questions use the *ggplot2* package in tidyverse. We have given minimum design requirements for each chart, but we encourage you to explore using additional features available to make the charts look even nicer (e.g., using color even when not asked to do it)! **Note: since you are able to use ggplot, the graphics should look great!**

c) [2 pts] Use ggplot2 to create a bar graph of the number of raining hours for each season in the dataset for London (over all of the years). Be sure to label the x and y-axes and to create a title/caption for the graph. You do not need to use color for this plot.

d) [2 pts] Use ggplot2 to make a single box plot of the temperatures for each of the six cities. Label each axis and provide a title. You do not need to use color. [To be clear, this is a single chart that contains 6 box plots.]

e) [2 pts] Use ggplot2 to create a line graph of the number of raining hours for each season in the dataset for all of the cities. Be sure to label the x and y-axes and to create a title/caption for the graph. Please create a legend showing a correspondence between the line color used and the names of the cities. [To be clear, this is a single line graph!]

f) [2 pts] Use ggplot2 and facets, or another package like gridExtra[2] to create a panel of two line graphs side by side, of raining hours and hours where the temperature is below 0 degrees Celsius for each season in the dataset for all of the cities. [To be clear, the same 'raining hours' chart from part e should be on the left, and the second plot is of hours with temperature below 0 (question 6b) following a similar method as in part e should be on the right.]

Question 7 [2 pts]: Write a summary (about one page in length) that discusses what aspects of this assignment were easy and hard to do in R (and think about how long it would have taken to do similar things in Excel or Tableau). From your experience, in which situations would you want to keep using R and in which would you seek different tools?

[2] We intentionally did not show or discuss this package in class. The whole point of this question is to demonstrate how quickly you can search for information needed about an unknown package and do it.