Exploratory data analysis with R

PLAN 372 Homework 2

In this assignment, you will work with a dataset to perform exploratory data analysis and answer a research or policy question using R. There are two datasets you can choose, both from Wake County. One is a dataset of calls to the fire department between 2007 and the present. The other is a dataset of restaurant inspection scores (those plaques you see with a rating for how sanitary a restaurant is, from 0-100). Both datasets are available on Sakai.

There are several questions listed below for each dataset. Choose **one** dataset, and please answer all of the questions for that dataset. For many, you could answer them either using a table or a plot. You can choose to use either plots or tables, but please include at least one table and at least one plot in your final writeup.

To submit the assignment, submit a writeup with the answers to the questions on Sakai. If you use a plot (from ggplot) or a table (e.g. from summarize) to answer a question, please include that table in your writeup. Keep track of your code in a Github repository, and include a link to the repository in your writeup. You can either make the repository public, or add me (Github username: mattwigway) as a collaborator, so that I can access your code. Please use comments liberally to describe your thought process behind your code, and to note which part of your code is answering which question. In addition to documenting your process, this will allow me to award partial credit in the event that you did not get the right answer for a question, but approach the problem correctly.

The datasets may contain missing data. Make sure to note in your report if there were missing data that would affect the answer to a particular question, and how much data were missing.

The point value for each question is noted below. In addition to per-question points, you will get:
1 point for including at least one plot in your writeup
1 point for including at least one table in your writeup
1 point for code committed to Github, with comments describing the functionality of each section of code
1 point for writing code that runs from top to bottom without errors (note: if you don't get this point but the code does run without errors on your computer, let me know and we can look at it in class. I won't penalize you if the code ran on your computer but doesn't on mine due to different system configurations, software versions, etc.)

You need to submit both the writeup and the code on Github to receive credit for this assignment.

Partial credit is possible throughout the assignment.

Restaurant dataset

The restaurant dataset contains records of the most recent health inspection for food-service establishments in Wake County. County health officials are curious to gain a better understanding of the overall picture of food safety in the county, in order to better target enforcement efforts, and have asked you use this dataset to answer the following questions:

1. Visualize the overall distribution of inspection scores using a histogram. [1 point]
2. Some restaurants have been in business much longer than others. Is there any trend in terms of how highly older vs. newer restaurants score on their inspections? [0.5 points]
3. Wake County is the most populous county in North Carolina, and there are many cities in it. Do the inspection scores vary by city? Note that the city column contains some differently spelled city names; make sure to clean those up so that there is only one estimated value per city. The recode function that we used for creating a weekend/weekday variable in the SFpark exercise will be useful here, and you may also be interested in the str_to_upper function. [1 point]
4. Wake County employs a whole team of inspectors. It is possible that some inspectors may be more thorough than others. Do inspection scores vary by inspector? [0.5 points]
5. It is possible that some extreme results from the previous questions are due to small sample sizes in a particular city, for a particular inspector, or in a particular time period. Look at the sample sizes in each of your groups. Do you think this is an explanation for the results you came to above? [0.5 point]
6. The data file contains records for many types of food-service facility (e.g. restaurants, food trucks, etc.). Are the scores for restaurants higher than other types of facility? [0.5 point]
7. Since restaurants are where the general public is most likely to interact with the food-service system, Wake County Public Health is particularly interested in sanitation in restaurants. Repeat the analyses above (1-5) for restaurants specifically. [2 points]

The dataset contains the following columns:
OBJECTID – numeric identifier of a particular inspection
HSISID – numeric identifier of a particular restaurant
SCORE – the sanitation score of the restaurant (0-100)
DATE_ - the date of the inspection
DESCRIPTION – a description of any issues found at the inspection
TYPE – Whether this was a normal inspection, or a re-inspection after correction issues from a previous inspection
INSPECTOR – who conducted this inspection
PERMITID – the Food Service Permit identifier for the facility
NAME – the name of the facility
RESTAURANTOPENDATE – when the restaurant first opened
FACILITYTYPE – whether this is a restaurant, food truck, etc.

Each row represents the most recent inspection for a particular food-service establishment.

Fire department dataset

This dataset represents all calls to the Wake County Fire Department since 2007. The fire department wishes to optimize their services and deploy their personnel to maximize the population they can serve, and minimize response time, and have asked you to use this dataset to answer several questions.

1. How long does it take Wake County Fire to respond to incidents, on average (i.e. the time between when an incident is dispatched and when firefighters arrive on the scene)? (hint: you can subtract lubridate date columns from each other). [1 point]
2. Does this response time vary by station? What stations have the highest and lowest average response times? [0.5 points]
3. Have Wake County Fire's response times been going up or down over time? What might be the reason for these changes? [0.5 points]
4. At what times of day are fire calls most likely to occur? [1 point]
5. The dataset contains all types of fire department calls, other than emergency medical services (which are removed to protect privacy). The codes for the different incident types can be found on page 3-22 of the National Fire Incident Reporting System Complete Reference Guide. How many calls to Wake County Fire are recorded in this dataset, and how many of them are actual fires? [0.5 points]
6. It is reasonable that firefighters might respond more quickly to some types of incidents than others (e.g., a building fire, code 111 might be higher priority than a cat stuck in a tree, code 542). Using the reference guide linked above to determine appropriate incident codes, evaluate the average response time to actual fires. Is this response time faster than the average response time for all incidents? [0.5 points]
7. Repeat the analysis for questions 2-4 for actual fires, rather than all incidents. [2 points]

The dataset contains the following columns:
X, Y – longitude and latitude of incident
OBJECTID, incident_number – numeric identifiers for each incident
incident_type – numeric code for the type of the incident (e.g. house fire, smell of gas, etc)
incident_type_description – plain-English description of the incident type code
arrive_date_time – when firefighters arrived at the scene of the incident
dispatch_date_time – when the call came in and firefighters were dispatched to the incident
cleared_date_time – when firefighters cleared the incident and left the scene
exposure – if the fire spread from another fire (e.g. a house fire started by a car fire)
platoon – which shift responded to the incident
station – which fire station responded to the incident
address, address2 – address and address line 2 for the incident
apt_room – apartment or suite number, if applicable
GlobalID – numeric identifier
CreationDate, EditDate – when the incident was entered into the system, and when the record was last edited

Creator, Editor – who entered the incident, and who edited it most recently