Problem Set 2

Lauren Nichols

2025-09-14

Problem Set Instructions

Please read *Problem_Set_Instructions.pdf* before beginning this problem set. These instructions will be relevant to this problem set and the following (Problem Set 3).

Part I: Life Expectancy

For the first half of this assignment we will be exploring changes and differences in Average Life Expectancy using the Gapminder data set. These data provide summary statistics related to a variety of global trends for various countries across years. We will be using a subset of the full data set.

Load the gapminder data

- **If you are using R**, you can access the data directly through the `gapminder` package using the code below.

```r
library(gapminder)
library(tidyverse)

gap <- gapminder
```

- **If you are using python**, you will need to use the provided csv file (GapminderData.csv), and load it into your environment using pandas.

```python
import pandas as pd

df = pd.read_csv('GapminderData.csv')

print(df.head())
```

1 - Exploratory figure. (5 pt)

A) Make an **exploratory** figure that looks at life expectancy for all countries over time.

B) List two questions that emerge based on the visualization.

2A - Plot. (18 pts)

Filter the data to just the **United States** and **United Kingdom**.

Create a line plot that shows life expectancy over time (using the full time range in the data set).

2B - Questions (5 pts)

- Did you use the default y-axis scale limits, or did you manually set them? *(e.g., Did you change the minimum or maximum values on the y-axis?)*
- **If you changed the y-axis limits**, how did you decide what scale to use? **If you kept the defaults**, how does your plotting library determine those default axis limits?
- What is one potential *advantage* and one potential *disadvantage* of the y-axis scale you used (whether default or custom)?

3 - Plot. (18 pts)

Recreate the line plot showing **life expectancy over time** for the **United States** and the **United Kingdom**.

This time, also include data for **all other European and North American countries** to provide context for the trends in the U.S. and U.K. Design your plot so that the U.S. and U.K. remain the clear focus, and the added context supports rather than distracts from that focus.

4 - Plot. (18 pts)

Create a figure that effectively and efficiently represents the following title and subtitle (note that the data only goes through 2007, so we will pretend like 2007 is the current year):

Title: “The current Life Expectancy Gap for the U.S. compared to U.K. is the largest ever recorded”

Subtitle: “With the exception of a period in the 1970-80s, life expectancy has been consistently lower for the U.S. than the U.K.”

PART II: Traffic on NC 147

Scenario: The North Carolina Department of Transportation (NCDOT) is planning a project to repave a section of North Carolina Highway 147 (NC 147). This highway is a main artery for traffic traveling through Durham. NCDOT wants to minimize disruptions to traffic during the course of the repaving project.

Task: NCDOT has tasked us with identifying the best days and times for deploying the work crews. For this problem set, we will be exploring the data and sharing a few initial insights with the stakeholder. (Next problem set, we will develop and share our recommendations.)

Data Provided: NCDOT has provided traffic data from the previous year on this section of highway as a csv file (also available here: https://ncdot.ms2soft.com/TDMS.UI_Core/portal). It includes hourly traffic volume data for both directions on the highway.

The columns in the data are as follows:

- `LOCAL ID`: Unique identifier for the section of highway.
- `DIRECTION` : Direction of traffic; `EB`: East Bound; `WB`: West Bound.
- `MONTH`: Month number, `1`: January through `5`: May.
- `DOW`: Day of week, `1`: Sunday through `7`: Saturday.
- `DATE`: Calendar date in `"%m/%d/%y"` format.
- `TIME`: Time in hours.
- `VOLUME`: Total traffic volume.

Note: There are many approaches to processing and analyzing time-series data. We will be taking a very simplistic approach.

Load the data.

First, we will load the data and make sure all date and time values are formatted correctly:

In R:

```
library(tidyverse)
library(lubridate)

dat <- read_csv("HourlyDataByDirection_raw.csv",
    col_types = cols(DATE = col_date(format = "%m/%d/%y")))

#combine DATE and TIME and convert into a datetime object using lubridate::as_datetime

dat <- dat %>%
  mutate(datetime = as_datetime(paste(DATE, TIME)))
```

In python:

```
import pandas as pd
from datetime import datetime

#Read CSV with specific date parsing
dat = pd.read_csv("HourlyDataByDirection_raw.csv",
                  parse_dates=['DATE'],
                  date_format='%m/%d/%y')

#Combine DATE and TIME into datetime object
dat['datetime'] = pd.to_datetime(
    dat['DATE'].astype(str) + ' ' + dat['TIME'].astype(str)
    )

print(dat.head())
```

Exploratory plots:

1A. Hourly Data (10 pts)

Plot the raw hourly data as a line plot, with time on the x axis and `VOLUME` on the y axis. Facet the plot based on `DIRECTION`. Note that we have data for both sides of the highway (East Bound and West Bound) - so these need to be plotted as separate lines within each graph (2 pts)

1B. Answer the following (5 pts)

What insights can you get from these initial plots? In what ways does this plot mask patterns in the data?

2A. Monthly variation in traffic. (12 pts)

Make four ***exploratory*** plots, each with **month on the x axis**, **volume on the y axis**.

Try making these plots without using AI.

Create the following plot types:

a) A boxplot,

b) A violin plot (or half-violin plot),

c) A strip plot (also known as a jitter plot)

c) A bar plot showing total volume for each month

2B. Answer the following: (9 pts)

What insight or story does each plot reveal that the others do not? Briefly explain how each plot helps you understand the data in a unique way.