STA234 HW-5

DUE DATE: 3/6/2022 by 10PM in Moodle

In this problem we will use the *midwest* data in ggplot2 package. First read about the data to understand the variables. In steps stated below, we will make a scatterplot between two numerical variables, then we will add more variables to the plot using aesthetics like color and size, and enhance plot in various ways. Do the following parts:

(a) [5 points] Make a scatterplot of area (x-axis) versus total population (y-axis). Label axes and add title. What do you notice? Use options(scipen=999) to turn off scientific notation like 1e+06 and redo the plot. Add a chart number (Fig.1 etc.) with title and subtitle to your plot.

(b) [5 points] Identify all places (state and county) with total population above 1000000 and report them.

(c) [7.5 points] Delete the above outlier and redo the scatterplot. You can use any way to remove this, here are some options:

- change x and y axes limits for the plot using two additional settings:
  xlim(c(0, 0.1))
  ylim(c(0, 1000000))
- use function coord_cartesian() with same limits as above.
- create a subset of the Midwest data with population total less than or equal to 1000000 and then redo the scatter plot.

(d) [5 points] Using geom_smooth() add linear regression model. What can you say about the relation between population and area? Using geom_smooth() with R's default method "loess" add a fitted non-linear curve instead of lm. Read help file for the geom_smooth() to explain this "loess" and tell what do you observe is different in lm and R's default.

(e) [2.5 points] Change color of the dots in the above plot.

(f) [7.5 points] Change color of the dots based on the state (categorical) variable. Show how to change the color of the state dots from the R's default choice in above plot.

(g) [5 points] Use theme_bw() and theme_classic() to change the theme of the above plot.

(h) [7.5 points] Have the dot size vary by popdensity (continuous) variable in above plot. What do you observe?

(i) [5 points] Modify legend for state and popdensity to States and Density respectively.

(j) [7.5 points] Change state names to actual names of the state, replace IL with Illinois, IN with Indiana and so on.

(k) [7.5 points] Switch the order of legend for state and popdensity using guides() function.

(l) [5 points] Remove legend from the plot.

(m) [5 points] Make legend move to the left side.

(n) [5 points] Make legend move to the bottom and horizontal.

(o) [7.5 points] Filter subset of data with poptotal values > 300,000 and call this dataframe as midwest_sub. Report how many counties satisfy this criterion.

(p) [5 points] Create a variable in midwest_sub data for large county which satisfies poptotal>300,000 as follows:

```
midwest_sub$large_county <- ifelse(midwest_sub$poptotal > 300000, midwest_sub$county, "")
```

(q) [7.5 points] In the scatter plot of area versus population, with dots colored for states and dots varying in size for popdensity, add text to highlight the identifies counties using function geom_text().

(r) [5 points] Change the background color of your plot.

Problem 2 [50 points]: Write a summary of your project, including information about

- Introduction: Tell us what problem of interest. Why is this problem interesting and important? Introduce recent research done in area related to your problem. You can pack all this together to motivate us. Do keep it short, to the point, and interesting.
- Data: Tell us about the data resource and explain dimensions of the data, variables in the data, and how does this data relate to your research questions.
- Initial findings: Perform exploratory data analysis (EDA) including summaries and data visualizations for one research goal. Show main plot(s) and findings. Make sure to add labels, titles etc. to make your tables and graphs informative.
- share one advanced data visualization with any required information on how this plot is useful for the study and what it tells about your project (interpretation). Show your R skills, creativity, and advanced work in R.

Note: This is an initial report.