Lab-07

hello, my name is

March 11, 2022

Preface

The goal of this assignment is to help you gain more familiarity with using **ggplot** to visualize distributions. In this lab we provide less scaffolding and more open-ended questions. As always, please come to office hours and reach out to your teaching staff if you have any questions.

Data

We will work with data on Airbnb reviews from Inside Airbnb.[1] For this lab, we will just use the `listing_id` and variables that summarize the date of each review, both of which are contained in `reviews.csv`. Start by importing these data and assigning them to a name.

[1] Inside Airbnb is a mission driven activist project with the objective to: *Provide data that quantifies the impact of short-term rentals on housing and residential communities; and also provides a platform to support advocacy for policies to protect our cities from the impacts of short-term rentals.*

1. We'll start by making some simple visualizations of the distribution of the number of reviews for each `listing_id`. First, make a histogram. Make the first bar start right at zero, customize the bins or binwidth to suit your tastes, and use a named color to delineate between bars. Only include listings with 250 reviews or fewer.

2. Make a density plot of the distribution of the number of reviews for each `listing_id`. Include listings with any number of reviews and use a log scale for the x axis.

3. Compute the total number of reviews for all listings each month and assign it to a name. Make three visualizations of the the proportion of reviews each month: pie, stacked bars, and side-by-side bars. When we read in these data, the tidyverse's `read_csv` parsed the data and stored all the variables as numbers, which are continuous. In reality, `month` is a discrete variable that only takes on 12 levels. You may want to use `factor()` to tell R to treat `month` accordingly. For the side-by-side bars, make sure the x axis is labeled with the integers 1 through 12. Which do you think is most effective for highlighting the months with the least and most reviews? Explain your logic.

4. Compute the total number of reviews in each month *for each listing*. Plot the density of this count variable, faceting by month. Only include listings with 50 reviews or fewer. What, if any, conclusions can you draw from the resulting plot?

5. Aggregate the reviews data to the listing level. For each listing, create two variables: the first year it received a review, and the total number of reviews it has received to date. Use the resulting data frame to make an overlapping density plot of the distribution of the number of reviews for properties which entered the market in each year. Fill by year, telling R you want to treat year as a discrete variable by encoding it as a factor (see `?factor()`). Adjust the opacity to make it more readable. What, if any, conclusions can you draw from the resulting plot?

6. Reproduce the plot above as a ridgeline plot instead of an overlapping density plot. Fill all the densities with a single named color of your choice and color them white.

7. The ridgeline plot above in 6 is problematic insofar as it suggests a non-zero probability that listings have negative reviews. Fix that. What can you learn from this visualization? Is anything about it surprising?

8. Combine the reviews data with the data in `listings.csv` that we also used in example-07 and use it to reveal and/or visualize something that you couldn't learn with just `reviews.csv` or just `listings.csv` on their own.