Homework - Week 11

your name here

2025-04-07

Preface

The goal of this assignment is to help you gain more familiarity with using **ggplot** to visualize geospatial data. As always, please come to office hours and reach out to your teaching staff if you have any questions.

Rendering

Since the maps in this homework contain a lot of data, rendering may be slower than usual. We recommend running your code interactively (e.g., in chunks) so you can work on each problem separately without needing to render the entire document repeatedly. Then you can render the final product for submission when ready.

Data

We will work with data on Airbnb listings from Inside Airbnb.[1] For this lab, we will start by using listing location data contained in `listings.csv`. Each row corresponds to a listing `id`, and variables summarize the key details for that listing. Import these data and assign them to a name.

```
listings <- read_csv("listings.csv")
```

[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 plotting locations without formally treating them as spatial data. Use `geom_point()` to make a scatter plot of listing locations. Adjust the transparency by setting `alpha = 0.1` to make it easier to see which locations have higher and lower densities of listings. Color by borough (e.g., Bronx, Brooklyn, etc.).

2. Use `read_sf()` to read in the shapefile `nybb_22a/nybb.shp` and assign it to `boroughs`. Then use `geom_sf()` to plot the spatial data contained in `boroughs`, and `fill` by `BoroName`.

3. Now we'll combine the listings data with these new spatial data. Use `count()` or `summarize()` to count the number of listings in each borough. Join the resulting data frame with `boroughs`. Use `geom_sf()` to plot the boroughs as in 2, but now `fill` by the number of Airbnb listings in each borough to make a choropleth map. Add `scale_fill_viridis_c()` to alter the default fill colors.[2]

[2]Several `viridis` scales are loaded with `tidyverse`, since they are included in the package `ggplot2`.

4. Now use the joined data to compute the number of listings per square mile, then use that to recreate your choropleth map from 3. Do the patterns look similar to or different from the map you created in 3? Do you think one is better than the other? Why or why not?

5. Let's do some data work. Use `st_as_sf()` to convert the Airbnb listings in Manhattan to an `sf` data frame. Set the coordinate reference system by including the argument `crs = st_crs("WGS84")` in your call to `st_as_sf()`. Assign this new object to `listing_locations`.

Next, create a sf data frame with one row that contains the location of Times Square. Use the same coordinate reference system ("WGS84").

Use `mutate()` and `st_distance()` to create a new variable in `listing_locations` that contains the distance between each listing and Times Square in kilometers.

Finally, start making a visualization using the `boroughs` map of Manhattan as a base map. Then plot the listings in Manhattan and `color` by the distance to Times Square. Adjust the transparency by setting `alpha = 0.1`, and add `scale_color_viridis_c()` to customize the color scheme.

6. Read in `review_summary.csv`.[3] Isolate the properties with at least 10 reviews and `review_scores_location` of at least 4. Join the `review_scores_location` variable to `listing_locations`. Make a plot similar to the one from 5 using the review data: Start making a visualization using the `boroughs` map of Manhattan as a base map. Then add the listings using `geom_sf()` and `color` by `review_scores_location`. Adjust the transparency by setting `alpha = 0.1`, and add `scale_color_viridis_c()` to customize the color scheme. Based on your knowledge of the location of Times Square from before, does proximity to Times Square seem to matter for location ratings?

[3] As a reminder from a previous assignment: each row corresponds to a listing `id`, and variables summarize all the reviews for that listing. For context, here is the reviews page for the first listing in the data. In the top left corner you can see that Airbnb reviews include an overall rating (`review_scores_rating`) and several sub-ratings for specific things (e.g., cleanliness, stored in the column `review_scores_cleanliness`).

7. Use `lm()` to estimate a linear regression model to confirm your visual analysis of whether proximity to Times Square matters for location ratings. Use `review_scores_location` as the dependent variable and distance to Times Square as the independent variable. Print a `summary()` of the results and comment on your findings. Do you think this is a good model of location ratings?