LandCoverEDA

Anna Wolford

2024-03-22

1.

Summarize, in your own words, the main reasons for analyzing the data. Discuss the goals of the analysis in the context of the problem and why those goals are important (1 paragraph).

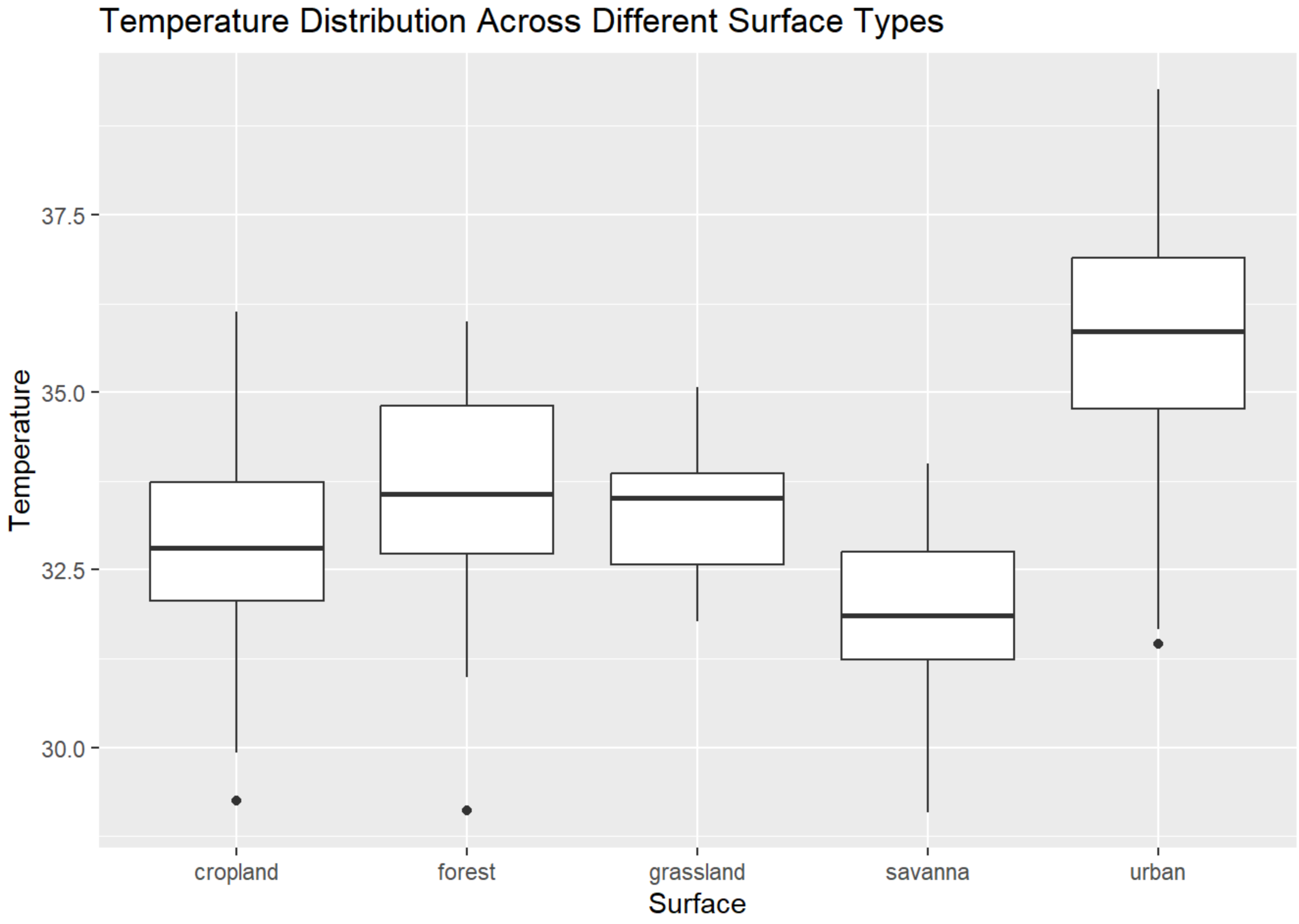
The primary goal of analyzing the SurfaceTemps.txt dataset is to understand the urban heat island effect in Houston, Texas, by investigating temperature variations across different ground surfaces. This analysis aims to determine whether there are temperature discrepancies among various surface types and identify which surface tends to be the hottest. Additionally, the analysis seeks to examine the impact of cloud cover interference on temperature measurements taken by satellites, providing insights into temperature readings in affected areas. Understanding these temperature patterns is crucial for addressing the numerous challenges posed by heat islands, including increased energy demand, air pollution, health risks, and water quality issues, thereby facilitating the development of effective mitigation strategies for urban heat islands. I lived in Texas for 17 months and I can vouch for the heat there - I swear eeach single degree hotter than 90F made a massive difference.

2.

Summarize the main features and patterns in the data in words and graphics. For each graphical summary (figure/plot) drawn, a written description of the main features of the graphic should also be provided. Written descriptions should be in the context of the problem (e.g. do not refer to an explanatory variable as x) and generally avoid statistical jargon/notation.

```
ggplot(temps, aes(x = Surface, y = Temp)) +
  geom_boxplot() +
  labs(x = "Surface", y = "Temperature") +
  ggtitle("Temperature Distribution Across Different Surface Types")
```

```
## Warning: Removed 126 rows containing non-finite values (`stat_boxplot()`).
```



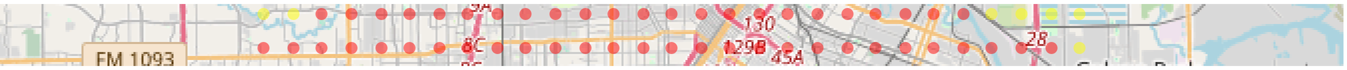
In attempts to address research questions 1 and 2, I created side-by-side boxplots. It appears that the urban surface has a higher median temperature in celsius than all other surfaces. We will need further analysis to see if this is a significant difference.

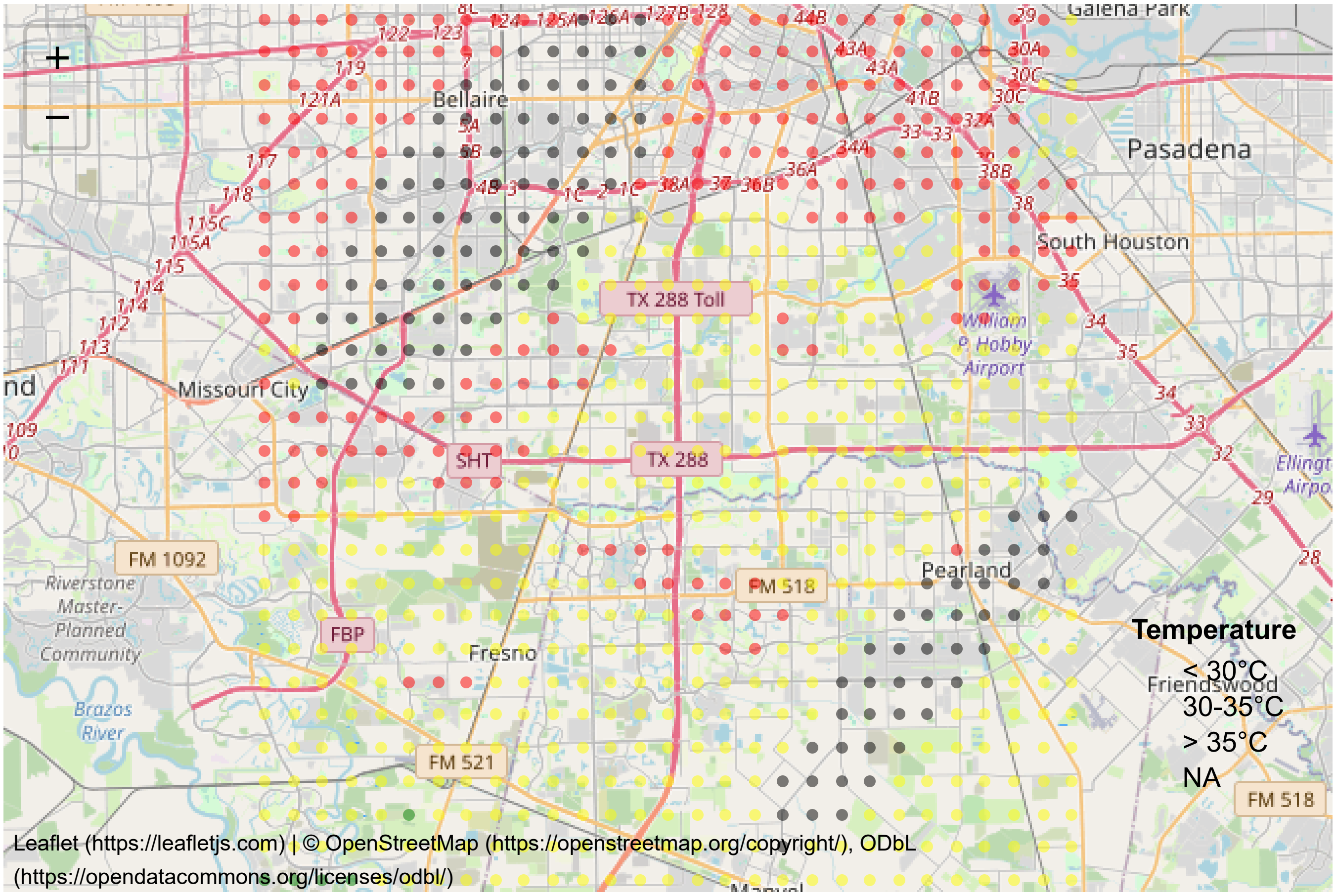
```
filtered_temps <- temps %>%
  filter(!is.na(Lat) & !is.na(Lon))

temp_colors <- cut(filtered_temps$Temp, breaks = c(-Inf, 30, 35, Inf), labels = c("green", "yellow", "red"))
temp_colors[is.na(filtered_temps$Temp)] <- "grey"
```

```
## Warning in `[<-.factor`(`*tmp*`, is.na(filtered_temps$Temp), value = "grey"):
## invalid factor level, NA generated
```

```
leaflet(filtered_temps) %>%
  addTiles() %>%
  addCircleMarkers(~Lon, ~Lat, radius = 3, color = ~temp_colors, stroke = FALSE, fillOpacity = 0.5,
                   popup = paste("Temperature:", round(filtered_temps$Temp, 2), "°C<br>",
                                 "Surface Type:", filtered_temps$Surface)) %>%
  addLegend(position = "bottomright", colors = c("green", "yellow", "red", "grey"),
            labels = c("< 30°C", "30-35°C", "> 35°C", "NA"), title = "Temperature")
```





Leaflet (https://leafletjs.com) | © OpenStreetMap (https://openstreetmap.org/copyright/), ODbL (https://opendatacommons.org/licenses/odbl/)

In attempt to answer the research question “What is the temperature at locations where the satellite was interfered by cloud cover?”, I created a leaflet interactive map that shots the temperature and the surface type on the map. It appears that the NAs are in areas with higher temperatures (denoted by green and red colors). This would suggest that the areas that are affected by cloud cover are probably above 30 degrees celsius. Again, more analysis is required to be able to make a more accurate prediction.

3.

Discuss aspects of the data that would create correlation between observations. If applicable, quantify how much cross-observation correlation is present in the data. This is spatially correlated data. This means that for any given locational point, the point next to it is likely similar to it. Supposedly you can quantify how much corss-observation correlation is present in the data by performing moran’s test? I haven’t done that before so I’m not quite sure how you’d do it.

4.

Posit an appropriate statistical model that could be applied to analyze the data. Discuss why the proposed statistical method would be useful in achieving the goals mentioned in point #1. The only model I’ve heard of that we could do to answer research question #3 and account for the spatial correlation would be a Spatial Regression model. I believe that would account for the spatial dependency in order to make accurate more predictions for question #3.

5.

Identify one aspect of the analysis that you don’t know how to do. I do not know how to handle spatially correlated data. At all! But I am so excited to learn!!