LGBT Lives: Data Analysis for Different Cities

Janine Wu[1] (wux13@rpi.edu)

[1]Rensselaer Polytechnic Institute 110 8th St., Troy, NY, 12180 United States

Abstract

LGBT stands for lesbian, gay, bisexual and transgender. As a group that is not widely accepted by many culture and societies, LGBT's are having difficult lives in many cases. For instance, in some countries or states, it is illegal to be homosexual and have relationship with a partner. At the same time, for the states that allow homosexual partners to marry, there are still people who are hostile against the LGBT group.

This poster mainly focuses on uncovering the pattern that makes a city friendly to LGBT groups which allows them to have good lives there. Three hypotheses include 1)The city's openness towards LGBT group affects LGBT rights and safety. 2) The score for LGBT nightlife affects LGBT lives. 3)There exist some patterns of cities that are good for LGBTs to live in. All hypotheses have been proved by analysis results and workable models have been developed for each hypothesis. Openness in the city has been found as the critical element that determines the city's suitability for LGBT groups.

Problem area

The main focus of this poster is to uncover the pattern of what makes a city good for LGBT groups to live in. There are various different factors involved, which include chance of developing friendships and relationships, nightlife strongness, openness in the city, safety and LGBT rights. Three hypotheses in total have been made to test they relationship with each other and the city's suitability for LGBTs.

3 Hypotheses:

1. The city's openness towards LGBT group affects LGBT rights and safety.
2. The score for LGBT nightlife affects LGBT lives.
3. There exist some patterns of cities that are good for LGBTs to live in.

Data Analytics and Modeling

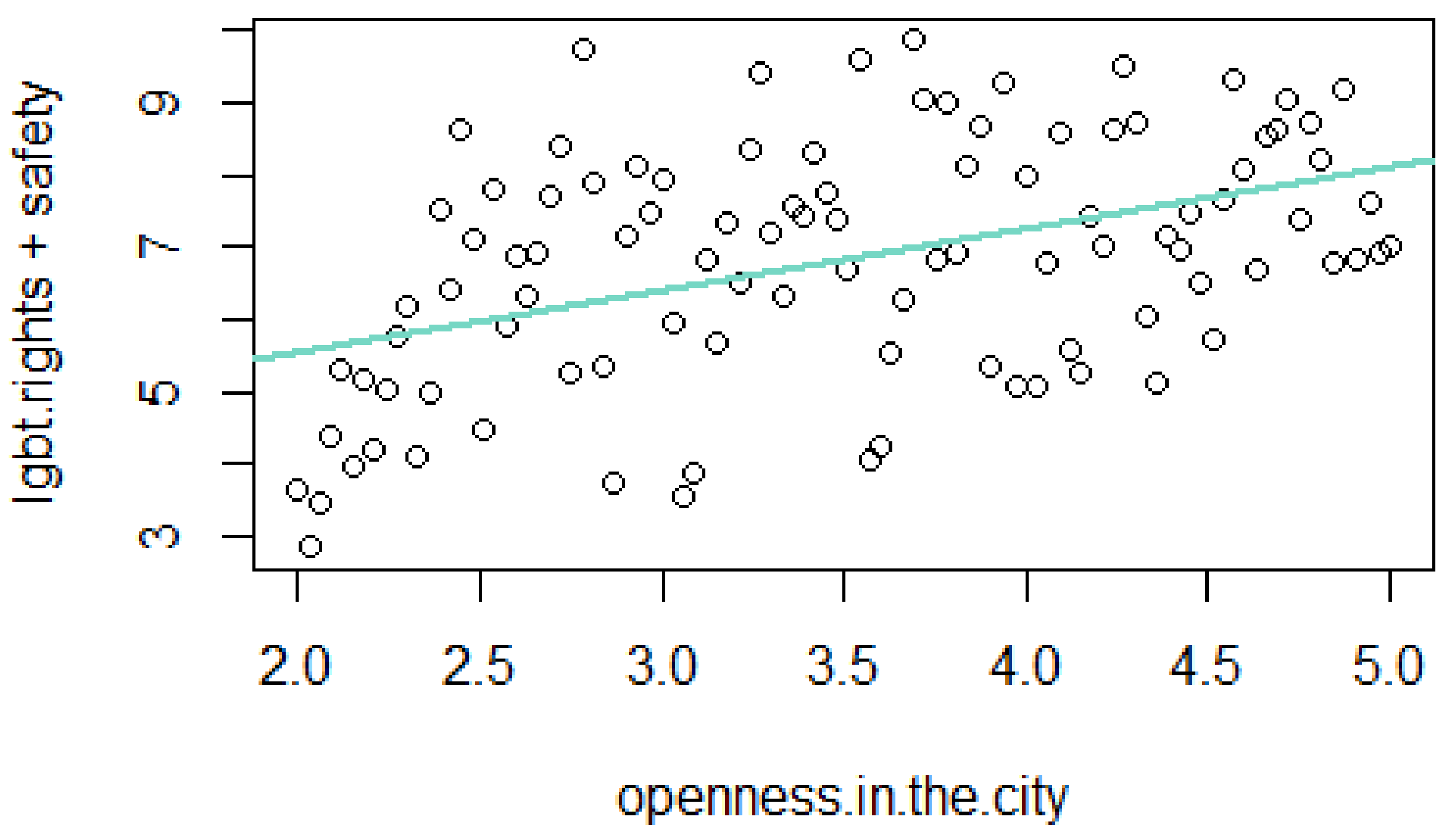
Distribution

According to the histogram, LGBT rights shows an exponential distribution. Total score shows a normal distribution. All other elements are equally distributed among score levels, but have less density on the two ends, which are the max and the min.

Openness vs. LGBT Rights and Safety

Linear regression model is used for discovering the relationship between openness in the city and LGBT rights and safety.

The model produced:
LGBT right + safety = 3.86 + 0.85 openness in the city ($p < 0.01$)
This model is highly significant but is an underfit model. A polynomial regression model might work better.
Plot of data points and linear regression model (green line):

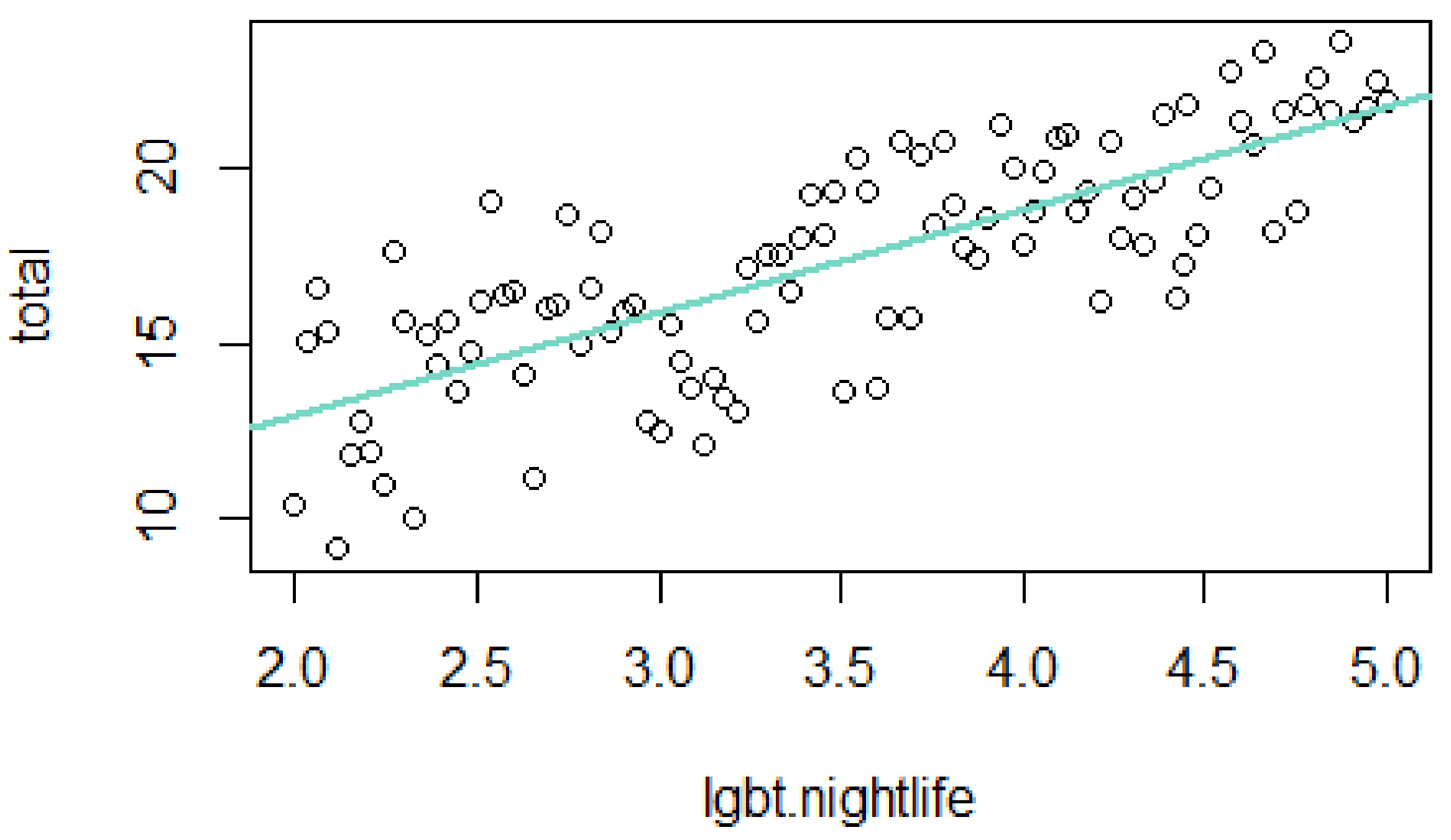


LGBT Nightlife vs. City's Suitability

Linear regression model is used for discovering the relationship between LGBT nightlife and city's suitability for LGBT groups.

The model produced:
Total score of life in city = 7.09 + 2.92 LGBT nightlife score.

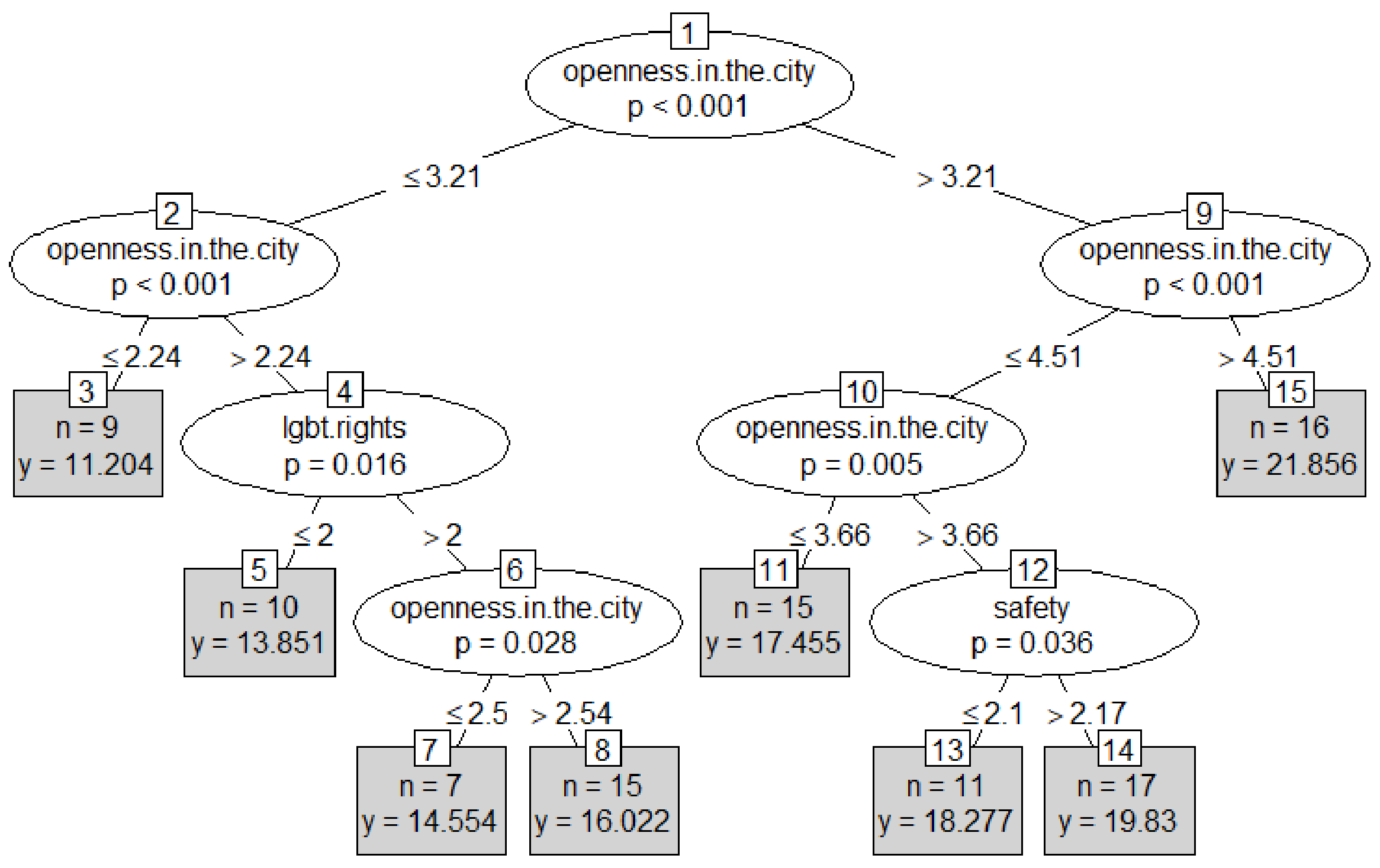
This model is highly significant and is a fit model.
Plot of data points and linear regression model (green line):



Pattern for Cities

Linear regression models has been tried for discovering the pattern and get a result of Total score = -7.1*10-15 + dating + LGBT nightlife + openness in the city + safety + LGBT rights, which does not indicate anything. Weighted k-nearest neighbor classification is then tired, but the prediction does not have good fit rate.

Conditional inference tree model is then tried and found working, with p values less than 0.01 for all nodes.



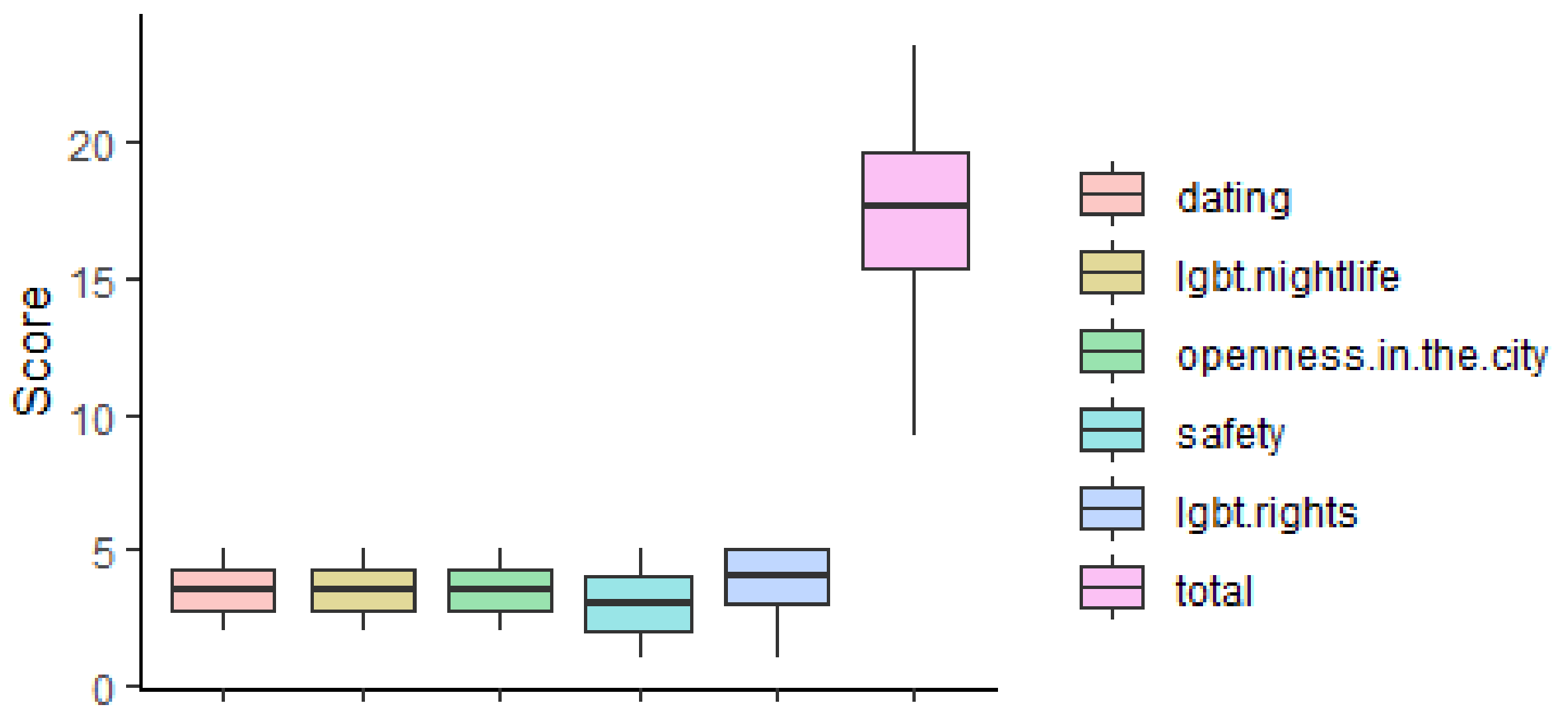
The Data

The dataset comes from Neskpick's Best LGBT Cities 2017. (https://www.nestpick.com/best-lgbt-cities/)

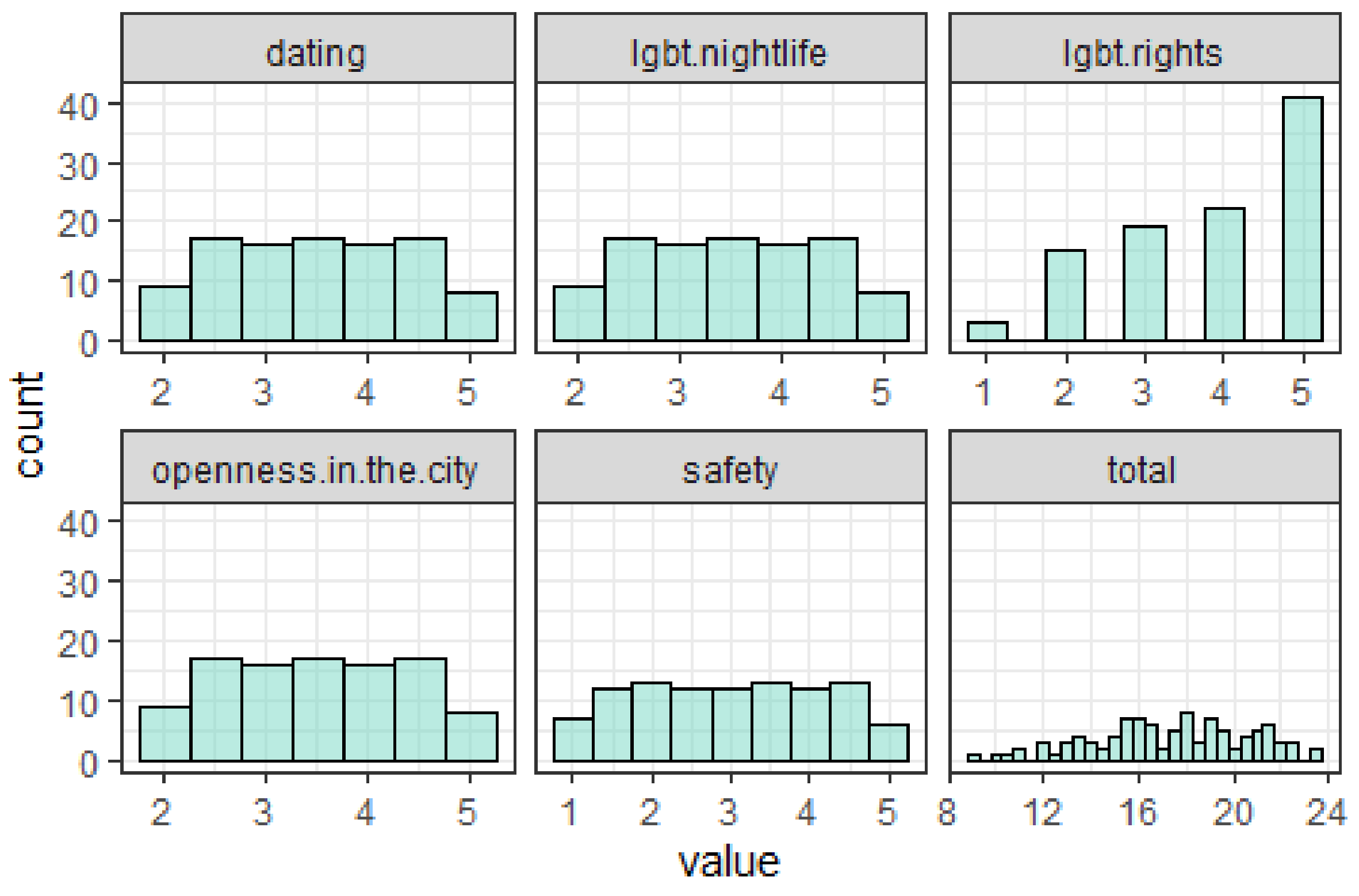
In the dataset, they recognized 100 cities with active LGBT communities. The dataset is originally in json format and is then pulled from the website and converted it into csv format for data analysis.

The dataset includes the following columns:
rank, city, country, dating, lgbt.nightlife, openness.in.the.city, safety, lgbt.rights, total, filter.order

Only dating, lgbt.nightlife, openness.in.the.city, safety, lgbt.rights, and total column are used in the data analysis process. All columns has been checked with any(is.na(df)) function and no missing values found. Summary(df) function and boxplots has been used to check for any outliers (single score not in the range 1-5 and total score not in range 5-25) and none of them has been found. Histograms of each column has been plotted with ggplot2 to check for distributions.



Boxplot for columns



Histograms for columns

Conclusion

All three hypothesis has been proved by the analysis result. Openness in the city is positively related with LGBT rights and safety, $p < 0.01$. LGBT nightlife is positively related with total score, $p < 0.01$.Openness is a critical factor that determines the score of the city: the higher the openness of the city is, the more suitable the city is for LGBT groups to live in.

The linear regression model for openness vs. LGBT right and safety is an underfit model. The linear regression model for LGBT nightlife and total score is a good fit. For both linear regression models, polynomial regression model might be a better approach. For the pattern of a good city, linear regression and kknn model both did not work, a conditional inference tree model is thus developed.

The next step of the process should focus on predicting the city ranking for cities all around the world with their characteristics. Furthermore, models developed can be correct to specific communities. Hopefully, this ranking and scoring technique can help LGBT groups to find the best suitable place for them to live in.

Glossary:
RPI – Rensselaer Polytechnic Institute
R – A program to process data and perform statistical analysis
Package (P) or Library (R) – software package to be loaded to perform extra tasks
Df, dataframe – Data manipulation structure in R

Poster: MT15A-08
Resources
Dataset: https://www.nestpick.com/best-lgbt-cities/
Ggplot2 package in R for visualization
Kknn, party, resample, rpart, tydyr packages in R