**Level 6000**

**1. Abstract and Introduction**

The motivation for this project is to call people’s attention for the lives of the minorities. In this project, the focus is LGBT lives in different cities and countries. 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 project focus on developing the pattern that makes a city friendly to LGBT groups which allows them to have good lives there. The hypothesis for this study includes:

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.

**2. Data Description and Exploratory Data Analytics**

The data used for this project 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 in json format on the website. The data is pulled from the website and converted it into csv format to make the analysis process easier. The datasets include the following columns:

rank: The rank of LGBT life of the city. Ranked by the column Total from highest to lowest.  
city: the city that is being described   
country: the country the city is in  
dating: dating score(1-5) – chance of developing friendships and relationships with other members of the LGBT community   
lgbt.nightlife: nightlife score(1-5) – LGBT night scene score, strong night scene helps some cities to establish themselves as leading gay destinations  
openness.in.the.city: openness in city score (1-5) – can feel proud and confident in who you are in the community  
safety: safety score(1-5) – being protected against hate crime and other forms of prejudice  
lgbt.rights – rights score(1-5) – the rights LGBTs have in the community, can be affected by laws  
total: the total score of the city, calculated by summing up the score from above five elements  
filter.order: used for ranking

**3. Analysis**

Before going into specific analysis, completeness of the dataset is checked first. The result of the function any(is.na(data)) returned FALSE, indicating that there is no missing value in the dataset. After that, the original dataset is subset. Since ranking, is somehow the same with total score and total score provides more information, ranking and filter.score will not be used in further analysis. At the same time, City and Country will not be used in the analysis, so Rank, City, and Country is subset out from the dataset. The new formed data frame contains dating, lgbt.nightlife, openness.in.the.city, safety, lgbt.rights, and total and will be used for later analysis.

Boxplot is used to check for any outlier and data points that is not valid. No outliers in this case has been found. Since for five elements, the max value can only be 5, any data points over 5 should be excluded and for total, data points over 25 should be excluded. No invalid data points found in this case.

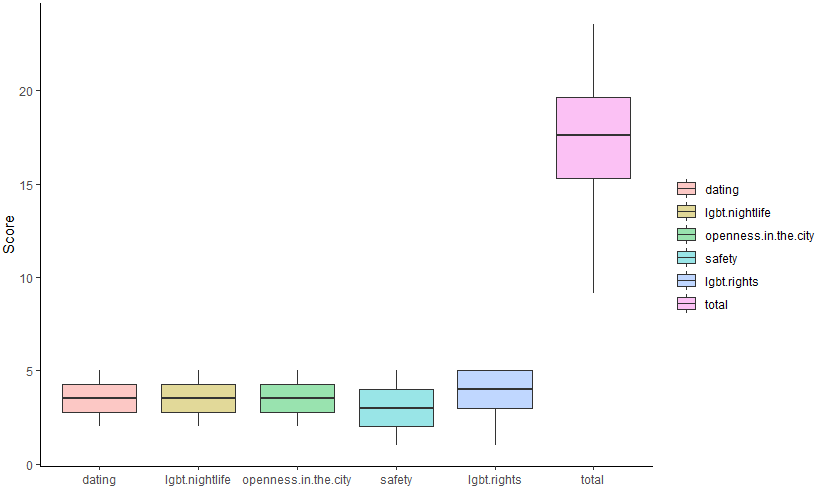


Figure 1: Boxplot of columns

The validity of data points is double checked with the summary() functions and no column of the five element have a maximum higher than 5, nor a minimum smaller than 1. The max for total score is 23.56, which is smaller than 25 and a minimum of 9.14. Considering the minimum score of the five elements, the min of total is reasonable and valid. Thus, all data points in the data set have been confirmed valid.

Histograms of the five elements has been generated to check for distributions. LGBT rights score shows trend of an exponential distribution. All other elements are equally distributed among most score levels, but are less distributed on the two ends, which is near maximum and minimum. The total score shows a trend of normal distribution.

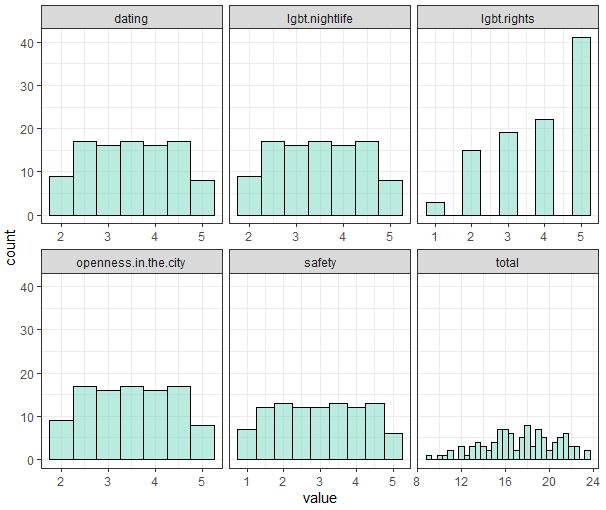


Figure 2: Histograms of the columns in the data frame

**4. Model Development and Application of model(s)**

**Hypothesis 1**

Trying to prove hypothesis 1, the city’s openness towards LGBT group affects LGBT rights and safety, a plot graph is generated to see the relationship between them. The graph indicates a somehow linear model. Thus, linear regression model will be used for hypothesis 1 and predicting the LGBT rights level and safety level by the openness in the city.

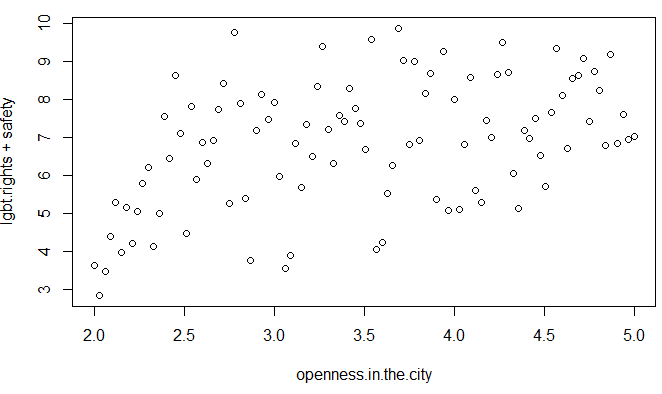


Figure 3: Plot of openness in the city vs. LGBT rights+safety

The linear regression model produced is the following:

LGBT right + safety = 3.86 + 0.85 openness in the city

Both the intercept and the coefficient of openness in the city have significant code of 3 \*s, indicating that they are highly significant. The model has as p value of less than 0.001, which indicates the result is highly significant. The analysis proves hypothesis 1, the city’s openness towards LGBT group affects LGBT rights and safety.

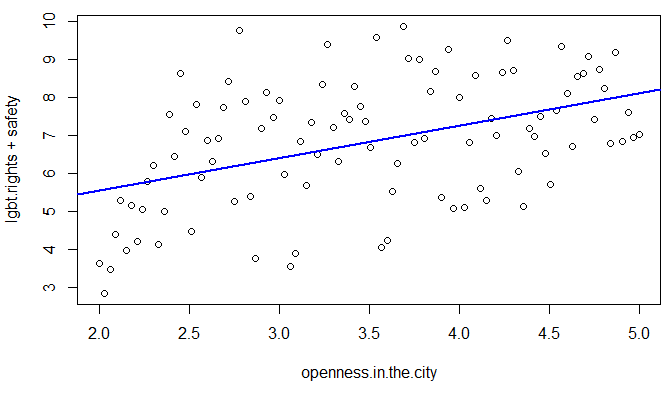


Figure 4: Plot of openness in the city vs. LGBT rights+safety with linear regression model. Blue line = linear regression model, points = data points

The model is a working model, but not a good fit to the data points. When the model is plotted along with data points, it can be seen that although the overall trend of the model fits the trend of data points, this model is an underfit model. The data is scattered over the plot, while the model fits only the trend of increment and missed some curving of data. The model fitting fails because data points are too scattered around, thus linear regression might not be a good choice.

**Hypothesis 2**

For hypothesis 2, the score for LGBT nightlife affects LGBT lives, a plot graph is generated to see the relationship between them. The graph shows a clear linear relationship between LGBT nightlife and the total score of lives. Linear regression model will be used to predict the total score using LGBT nightlife score.

The model generated is following:

Total score of life = 7.09 + 2.92 LGBT nightlife score

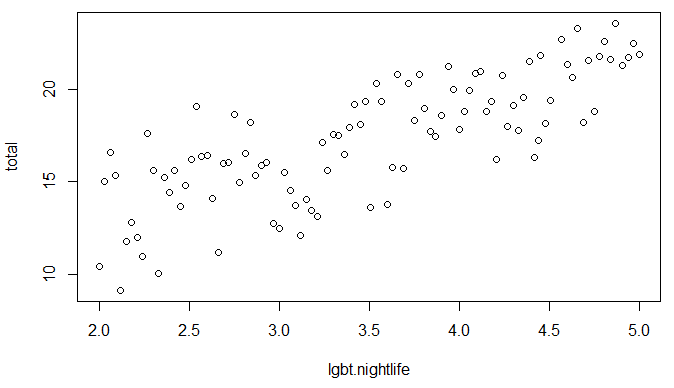


Figure 5: Plot of LGBT nightlife vs. total score

Both the intercept and coefficient of nightlife have 3 significant stars, indicating they are highly significant. The model has a p value of less than 0.001, indicating the result is highly significant. This model uncovers the relationship between LGBT nightlife and total score, proving hypothesis 2: the score for LGBT nightlife affects LGBT lives.

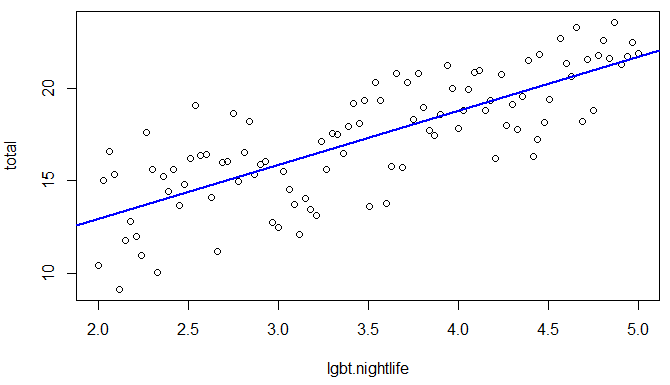


Figure 6: Plot of LGBT nightlife vs. total score with linear regression model. Blue line = linear regression model, points = data points

The model can be considered a fit for the dataset. Although there are points where it does not fit the points exactly, it passes where most points are and fits the overall trend of the dataset. There are some kind of curves seen in the plot; thus, using a polynomial regression model might help to get a better fit.

**Hypothesis 3**

Hypothesis 3 states that there exist some patterns of cities that are good for LGBTs to live in. In order to prove hypothesis 3, the goal is to find a relationship between elements and total score. As linear regression worked for the previous two hypotheses, it has been tired for hypothesis 3 as well. The linear regression model built from data is the following:

Total score = -7.1\*10-15 + dating + LGBT nightlife + openness in the city + safety + LGBT rights

All coefficients have 3\* of significant code, indicating all of them are significant. The p value is less than 0.001, indicating the result is significant. All the coefficients are 1 and the intercept is very close to 0. This result does not provide any additional information. It is known that the total score is the sum of five element score, while this model is saying the same thing. Although it is valid, but it does not help with proving or disproving hypothesis 3. Thus, clustering the tried next.

Weighted k-nearest neighbor classification is used to generate clustering model. However, KNN models is not a good fit for this dataset. It does not correctly predict the total score. For all levels of total score, the KNN model does not show a good performance. Thus, the KNN model is discarded.

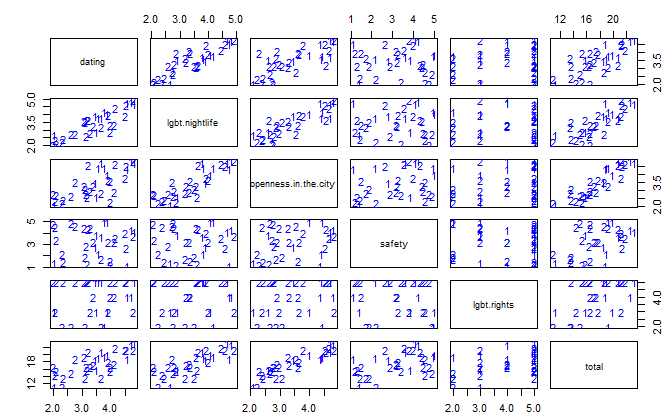


Figure 7: KNN model

After KNN model, conditional inference tree model is used, trying to find the pattern of the city that is good for LGBTs to live in. Ctree model worked well with finding characteristics of a good city for LGBTs. Among the decision tree, all nodes have a p value less than 0.05 indicating a significant result.

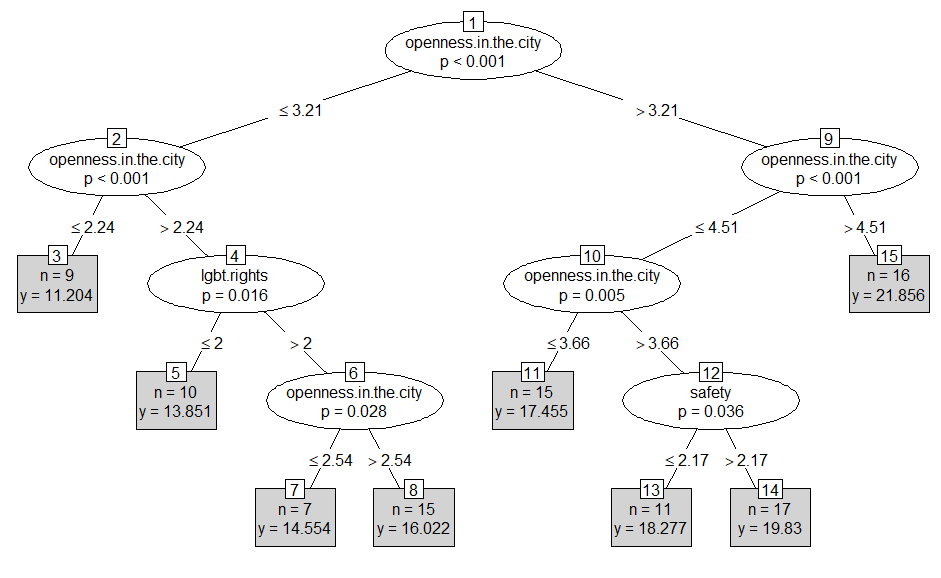


Figure 8: Ctree model

According to the ctree model, for cities with openness score less than or equal to 3.21, they have a lower total score. Among the lower score cities, LGBT rights contribute a bit to the total score. For cities openness score less than or equal to 3.21, but larger than 2.24, if LGBT right score is less than or equal to 2, cities have an average score of 13.85; if the right score is larger than 2, if openness is over 2.54, the average score is 16.02, else have an average of 14.56.

For cities with openness score more than 3.21, they tend to have a higher total score (total score of over 17). The highest total score group of cities have an openness score larger than 4.51. For cities have openness score larger than 3.21, but smaller than or equal to 3.66, they have an average score of 17.46. For cities with openness score larger than 3.66 but smaller than or equal to 4.51, if the safety score is smaller than or equal to 2.17, they have average score of 18.28; if the safety score is larger than 2.17, the average score is 19.83.

This model shows that the best cities all have a high openness score, and safety is just a bonus point for that. The result of ctree model proves hypothesis 3 by finding a pattern for the cities that are good for LGBTs to live in. The critical point is openness of the city, with LGBT rights and safety involved in some cases.

**5. Conclusions and Discussion**

During the data analysis process, the following results has been found:

1. Openness in the city is positively related with LGBT rights and safety,  
   p < 0.001.
2. LGBT nightlife is positively related with total score, p < 0.001.
3. 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 three hypotheses made has all been proved:

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.

The three models have been developed for each hypothesis:

1. Linear regression for openness and LGBT right and safety. This model is a working model, but it is an underfit model.
2. Linear regression for LGBT nightlife and LGBT lives. This model is a good model. It fits the data well, but a polynomial regression model might do a better job.
3. Conditional inference tree for patterns of good cities for LGBT to live in. This tree model has a high significance.

During the analysis, linear regression models did not take many efforts to find. For the tree model, linear regression and KNN models have been tried, but they did not work. Thus, tree model is tried and used. Polynomial regression models can be used for better regression models. Later directions 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.

**Reference:**

Dataset: <https://www.nestpick.com/best-lgbt-cities/>

Packages: kknn, rpart, party, ggplot2, tidyr, resample.