Discovering Significant Economic, Demographic and Environmental Predictors of Music's Happiness

Cole Spinale and Nicholas Maynard

5/25/2023

Project Overview:

In this project we would like to explore how a countries economic standing, demographics, and climate can affect the music that they listen. Specifically we would like to do this through the lens of a songs "happiness". Do wealthier countries listen to happier music than poorer countries? How does weather affect this listening data? Is increased rainfall associated with higher rates of listening to sad music? Does warmer weather correlate to listening to happier songs? To learn which factors can have the greatest impact on listening habits we will try to find a suitable regression model using backwards stepping to lower our models complexity and to try and capture the most significant predictors. Furthermore we will create a decision tree model that sifts through our list of predictors to find the most important features in predicting valence, a songs happiness.

Data Descriptions:

We acquired our data set containing environmental predictors from kaggle.com, where the creator of the dataset gathered their data from Google Earth Engine. This data set contains environmental variables for each country like max temperature in the warmest month (Celsius), number of cloudy days per year, and annual rainfall (mm) to name a few.

Variables: temp_max_warmestMonth - Measures the hottest recorded temperature in a country in degrees Celsius data type - numeric
range - 16.42866 to 37.03348

temp_mean_annual - Measures the mean temperature in a country in degrees Celsius
data type - numeric
range - -6.83169 24.99333

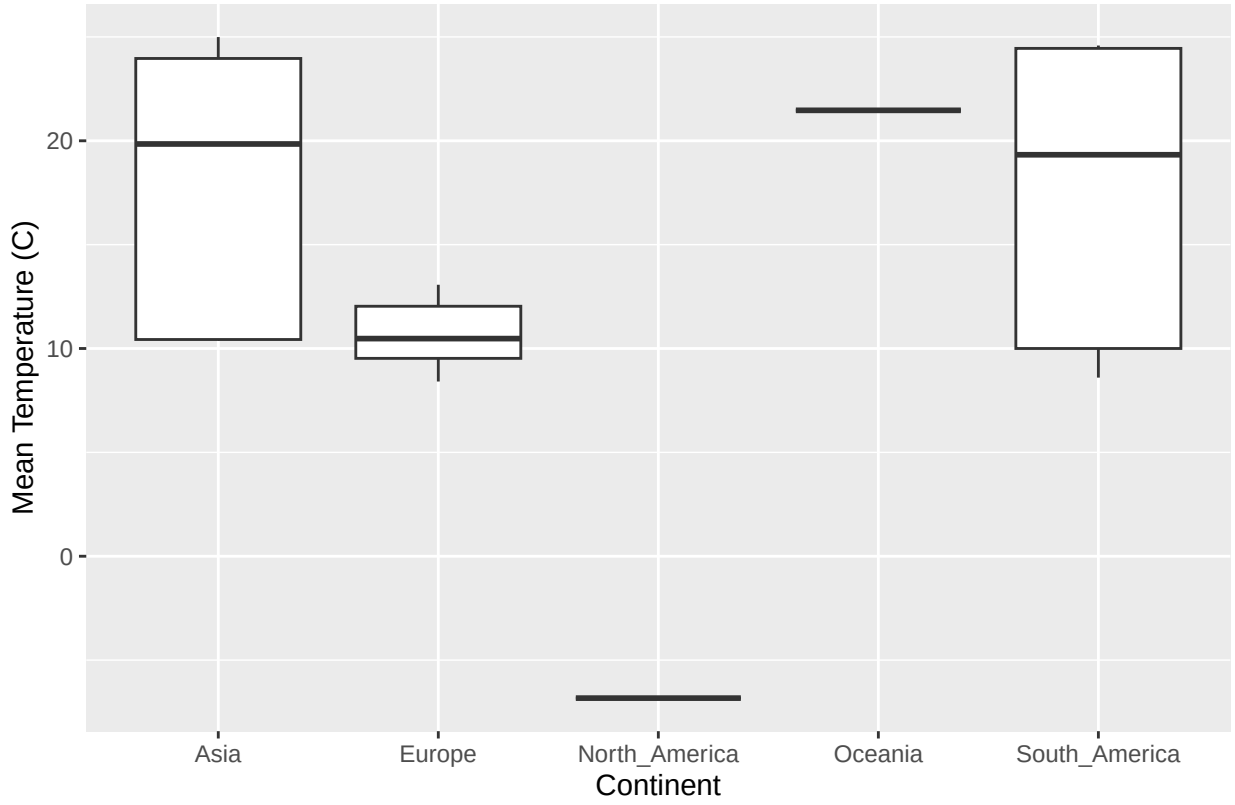


Figure 1: Annual Mean Temperature by Continent

Figure 1 shows the mean annual temperature for each continent. The boxplots show the distribution of mean annual temperature for all represented countries in each continent. There is only one country in both North America and Oceania that had complete data across all three data sets we combined. We see that Asia and Oceania have the highest mean annual temperature, with the continent's medians both lying above 30 degrees Celsius. We hypothesize that the songs that the people in these continents listen to will generally be happier than those from North America and Europe.

rain_mean_annual - Measures the mean annual rain in millimeters in a country
data type - numeric
range - 295.9783 to 2728.3552

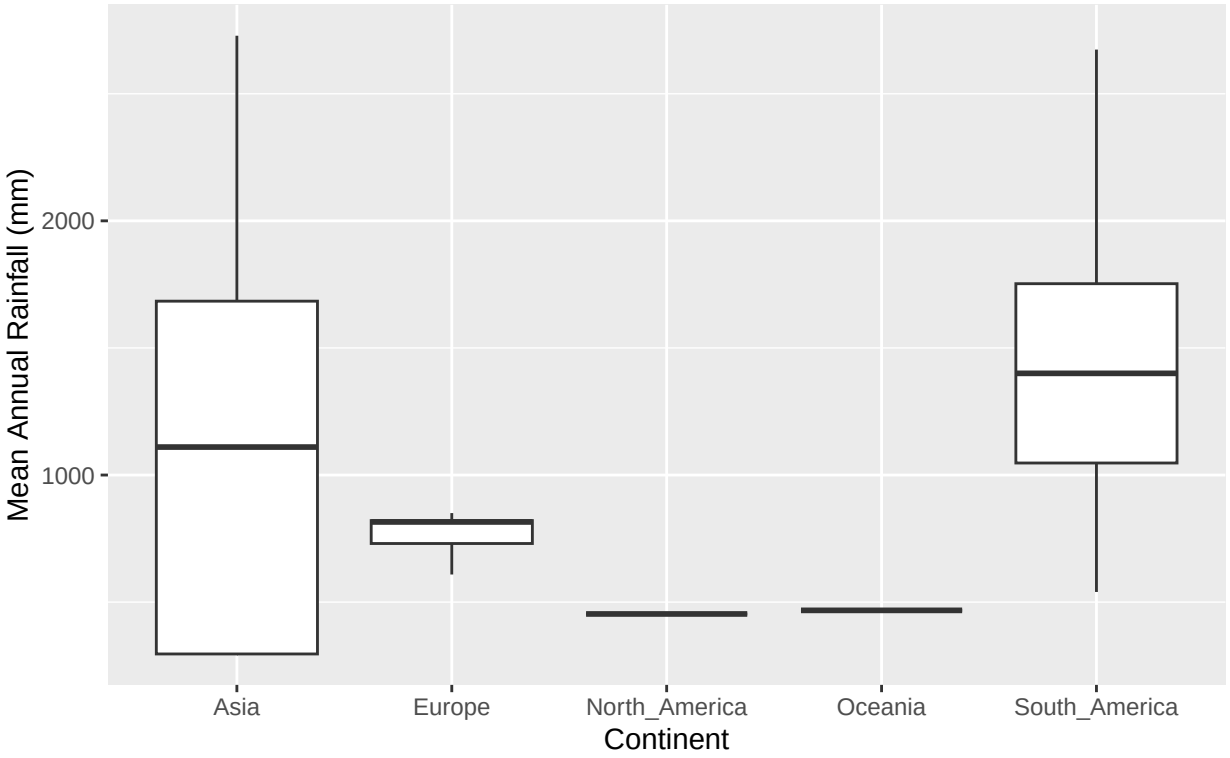


Figure 2: Average Rainfall by Continent

Figure 2 shows the distribution of mean annual rainfall for all represented countries in each continent. We see that Asia and South America have higher median mean annual rainfall than other continents. North America and Oceania's medians are almost 3 times less than South America's. We believe that these continents that experience more rainfall will generally be listening to more sad music and therefore have a lower associated valence.

cloudiness - Measures the number of cloudy days in a country in a year
data type - numeric
range - 88.59925 to 293.09511

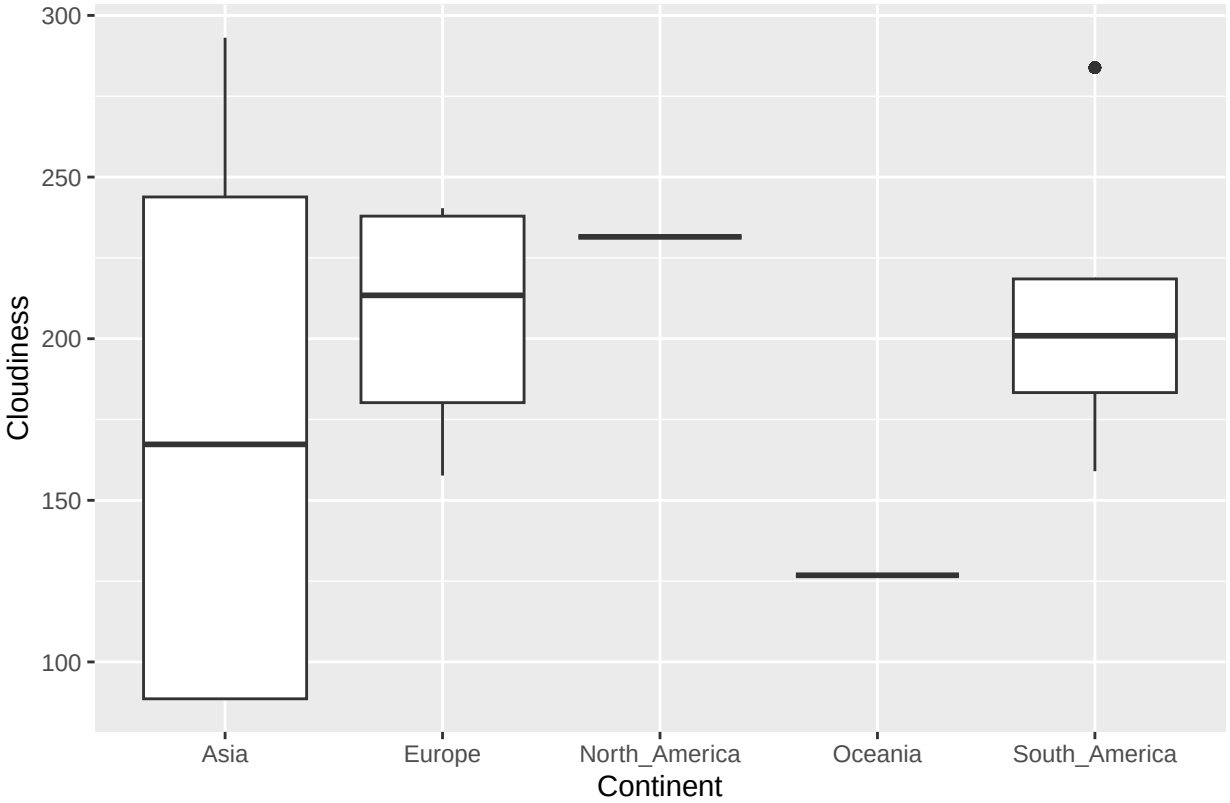


Figure 3: Cloudiness by Continent

Note: this data was taken from different years but we can assume that climates in each of these countries has stayed relatively constant over time so it should not cause any problems with our analyses.

We also acquired our dataset on economic and demographic factors by country from kaggle.com where it was initially pulled from The United Nations Statistics Divisions system UNData. This dataset contains variables like population density per square kilometer, percent of the population living in urban communities, and GDP per capita.

Variables: GDP_per_cap - Measures a countries GDP per capita
data type - numeric
range - 1614.2 to 51352.2

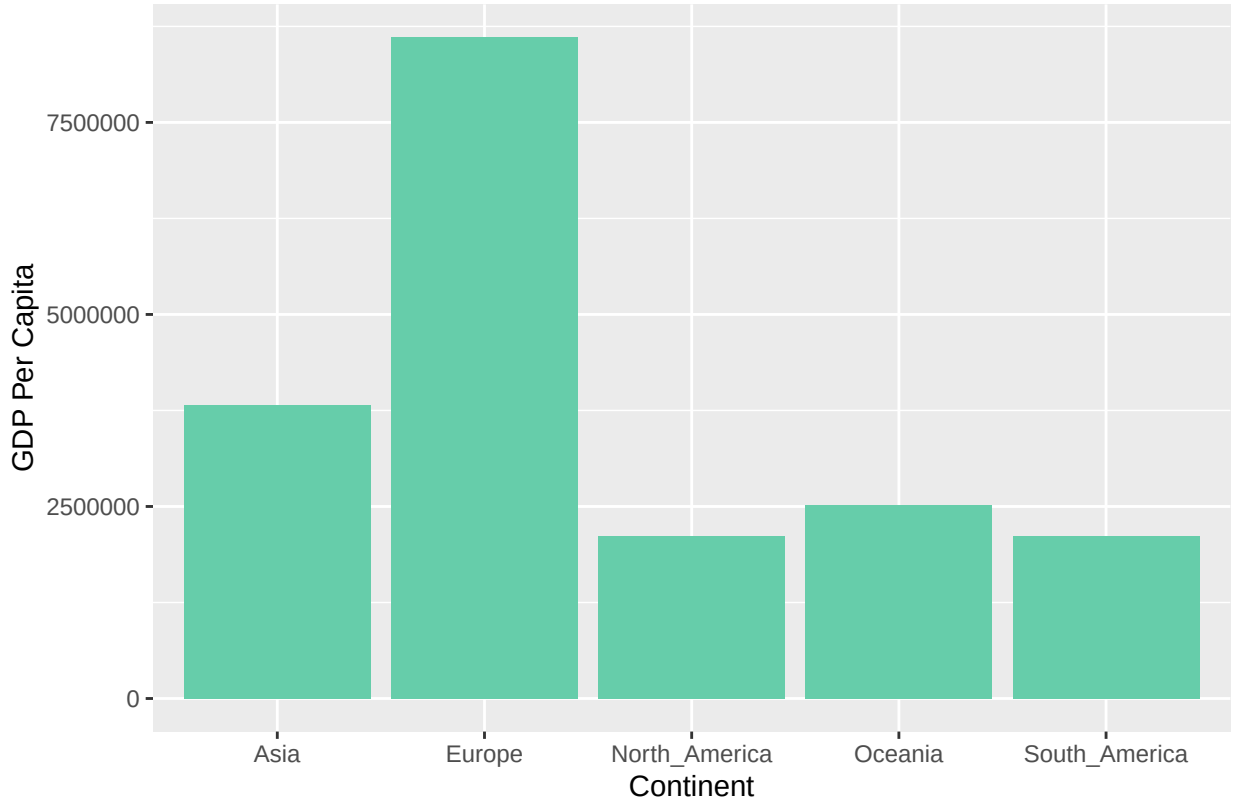


Figure 4: GDP Per Capita by Continent

pop_dens - Measures a countries population density per kilometer squared
data type - numeric
range - 3.178732 to 407.384502

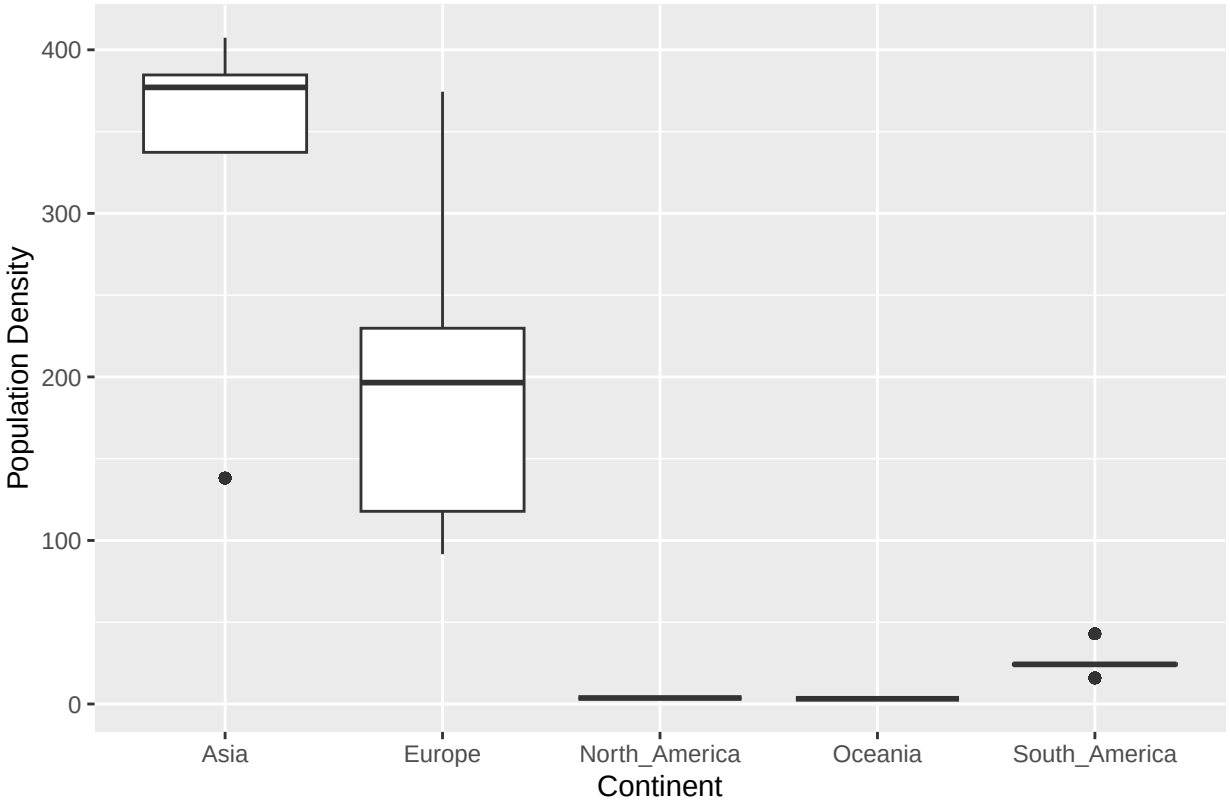


Figure 5: Population Density by Continent

Urban_pop - Measures the percent of a countries population living in urban communities
data type - numeric
range - 32.7 to 97.9

Note: this data is mostly taken from 2017 and if no data from 2017 was available, the next closest year is used instead.

Our Spotify dataset was also acquired from kaggle.com but the data was originally extracted from Spotify's API, later in the year 2019. The dataset includes different variables that describe aspects of songs like liveliness, acousticness, and also valence, the main response variable in our analysis. Valence is a measure of how happy a song is from 0 to 1, 1 being Happy by Pharrell Williams and 0 being Tears in Heaven by Eric Clapton.

Variables: Valence - Measures of a songs happiness on a scale from 0 to 1
data type - numeric
range - 5 to 98

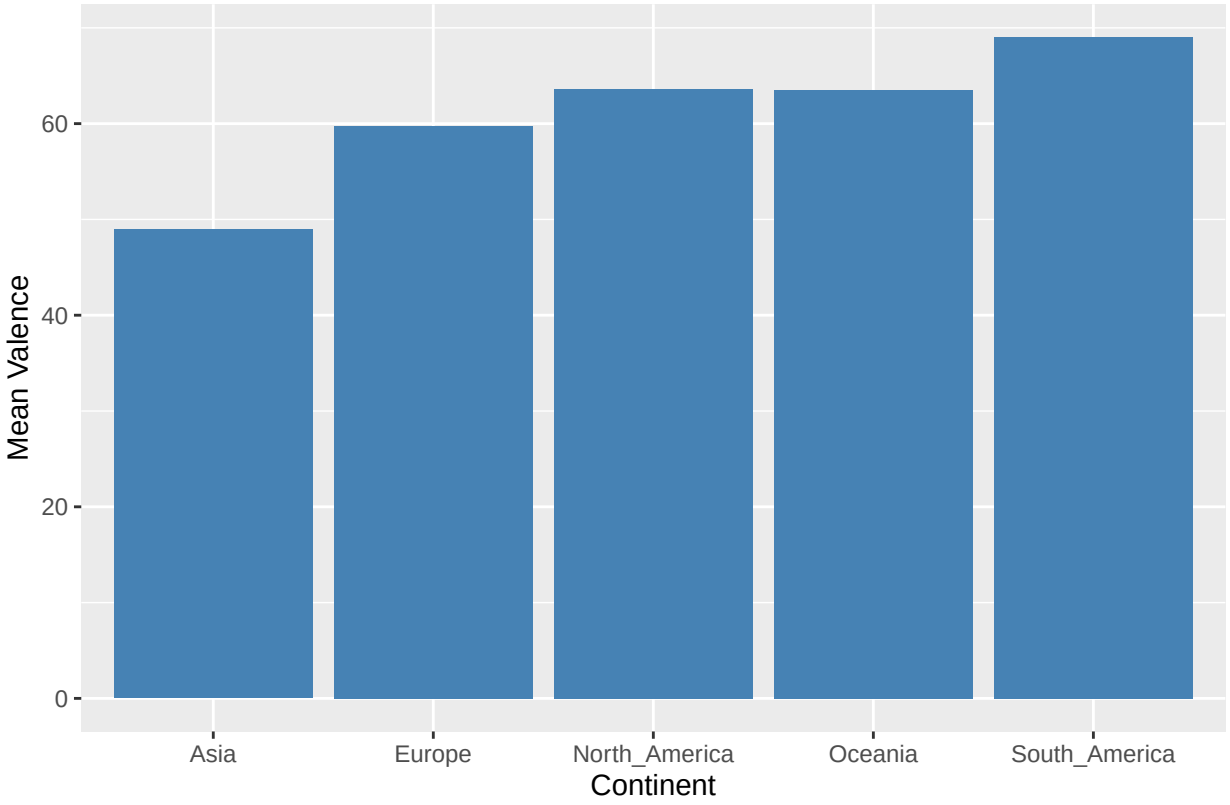


Figure 6: Mean Valence by Continent

Table 1: Number of Unique Countries Per Continent

Continent	Countries
Asia	4
Europe	5
North_America	1
Oceania	1
South_America	4

From our data sets we were able to successfully aggregate data for 15 countries, as seen in Table 1. We see that we have under representation in continents like North America and Oceania which may effect our models.

Analysis

Table 2: Backward Step Regression Summary

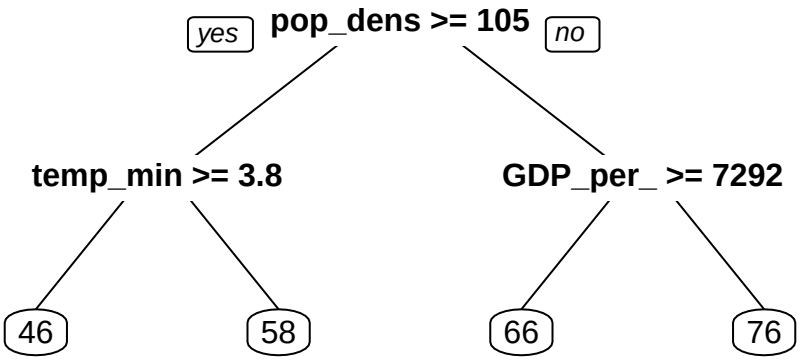
term	estimate	std.error	statistic	p.value
(Intercept)	-2.4828052	14.3058044	-0.1735523	0.8622653
cloudiness	0.1885623	0.0432006	4.3648085	0.0000146
rain_mean_annual	-0.0150106	0.0035219	-4.2620922	0.0000229
pop_dens	-0.0284661	0.0056453	-5.0424292	0.0000006
GDP_per_cap	-0.0003637	0.0000885	-4.1109985	0.0000439
Urban_pop	0.3619307	0.0665690	5.4369232	0.0000001
temp_max_warmestMonth	0.9553619	0.2672552	3.5747179	0.0003736

We chose to explore the variables we did earlier because they happened to be the best predictors of a songs valence. We first created a simple multiple linear regression model that predicts valence off of a countries cloudiness, annual temperature, annual rainfall, population density, GDP per capita, percenct urban population, and maximum and minimum temperature. This model had way to many predictor variables however, and we thought it would be best to use backwards selection to lessen the number of predictors and improve the models AIC. Doing this backwards stepping resulted in the AIC for the model being decreased by 2 but also gave a slightly lower multiple R-sqaured moving from 0.1021 to 0.09904. These R-squared values are extremely low and indicate that a vary low proportion of the variance seen in valence can actually be described by these predictor variables.

Table 2 shows that the maximum temperature in the warmest month has a large coefficient, almost 3 times larger than the next largest coefficient for the percent of the population living in urban communities. The associated p-value for maximum temperature and urban population are both small as well and seem to be our best predictors of valence based off the model.

To explore these variables further we will create a decision tree model that shows the most influential of these predictors. Using the decision tree may also allow us to potentially see any non-linear relationships between the predictors and valence.

Figure 7: Tree Model of Important Predictors



The decision tree model we made shows that population density is the most important feature in predicting valence. The minimum temperature in the coldest month as well as GDP per capita seemed to be the next most important features. This conflicts with the outcome of our regression model that seemed to show that the maximum temperature in the warmest month and the percent of the population living in urban communities were the best predictors.

This led us to believe that these variables may be associated to each other in some way. To look into this further we decided to check the correlations of these variables by making a correlation matrix using the corrplot package.

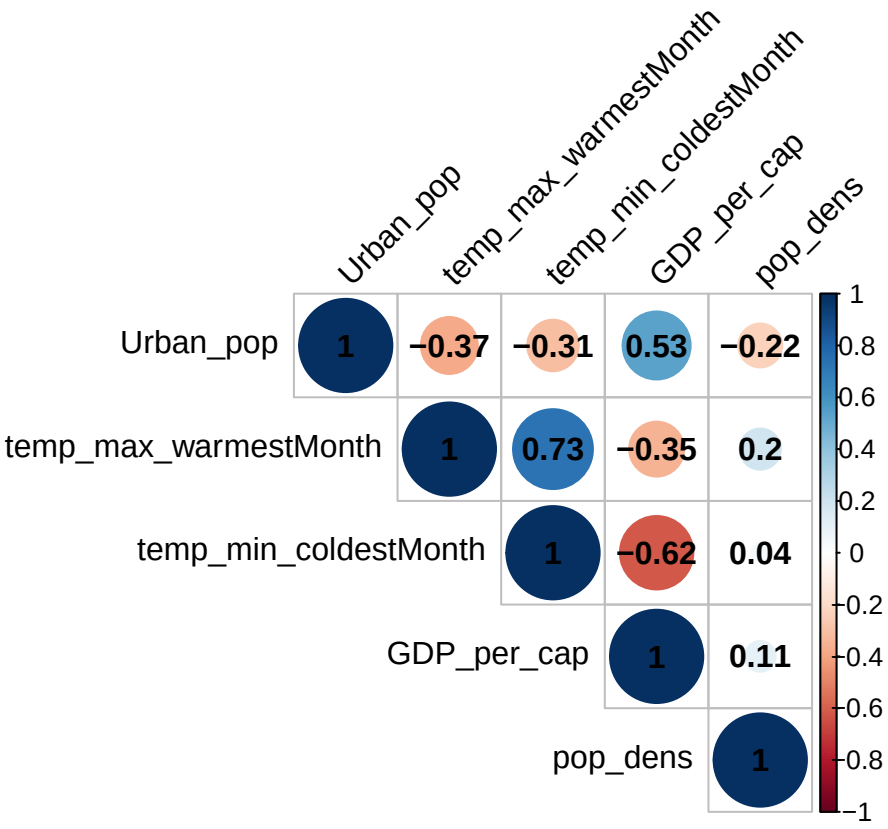


Figure 8

We see in Figure 8 that some of these variables have moderate to high correlations between them which may have been imapcting our models. Most notably the temperatures are highly correlated to one another with a correlation coefficient of 0.73. Minimum temperature with GDP per capita and urban population with GDP per capita also have a relatively high correlation coefficients. The correlations between our predictor variables may help to explain the low R-squared value we found in our backwards step model, and may explain why the significant features are different between the regression model and the decision tree.

Conclusion

In conclusion we found that some significant predictors of the valence of a song are temperature, population density, and urban population. We found through our regression model that the hotter your countries hottest day is, the more likely the songs they listen to throughout the year are happier. This matches with our initial hypothesis. We also found that GDP per capita was not very significant in predicting valence as noted in Table X, as it has an extremely low coefficient. Some unexpected predictors like population density and urban population proved to be better predictors of valence than rainfall and cloudiness, which we initially thought would be significant.

References:

https://www.kaggle.com/datasets/zanderventer/environmental-variables-for-world-countries
https://www.kaggle.com/datasets/sudalairajkumar/undata-country-profiles

https://www.kaggle.com/datasets/leonardopena/top-50-spotify-songs-by-each-country