Weather of Popular Destinations Analysis

A CMPT 353 Final Project

<Names Removed>

Course: CMPT353 - Summer 2023

Instructor: Greg Baker

Introduction

In this report, we present an analysis of the weather between top tourist destinations and cities that did not make it on the coveted list. Our curiosity led us to explore whether there exists a trend in weather conditions among popular tourist destinations and how significantly they differ from their less traveled counterparts.

Problem

The project idea was originally to find anything that we could say about the weather in popular cities to travel. Although people may not choose a travel destination only based on the weather of the destination, we are interested in finding any relationships between weather and popular destinations. We can find cities that can potentially be a popular travel desitination based on the weather trend we find by analyzing weather data. By investigating weather data for both popular and unpopular tourist destinations, we aim to compare weather conditions between these two groups and discern whether there are any distinctive weather patterns that make certain destinations more appealing to travelers. Through data analysis, modeling, and visualization, we will delve deeper into the impact of weather on destination popularity and provide valuable insights for travelers, tourism stakeholders, and decision-makers.

Data

For our analysis, we utilized two primary sources of data to gain a perspective on weather conditions in popular and unpopular tourist destinations. Firstly, we obtained a list of top tourist destinations based on a Wikipedia curated list, List of cities by international visitors, based on Euromonitor's ranking. After scraping the data using BeautifulSoup, we extracted the top 65 cities that attract international travelers. These destinations served as crucial reference points to our analysis, allowing us to make inferences on weather with significant tourist appeal.

Secondly, historical weather data was collected from the Global Historical Climatology Network (GHCN) through the SFU's Hadoop cluster generously provided by Greg Baker. We specifically collected daily weather records from the years 2015 to 2019 to analyze pre-pandemic weather patterns. The data collected was daily records of an extensive range of weather parameters. We were interested in maximum temperature (tmax), minimum temperature (tmin) and precipitation (prcp).

Data Cleaning

To prepare the weather data for analysis, we utilized three separate Python scripts: get_popular_cities_data.py, get_unpopular_cities_data.py, and get_random_cities_data.py. We created get_popular cities_data.py to obtain weather data of 40 popular cities in the ranking. Initially, we only collected 25 popular cities in the Euromonitor ranking because it was quite time-consuming to find stations in these cities from our GHCN datasets. However, during data analysis, we discovered that some features, such as tmin and precipitation, failed the normality test and were unsuitable for modeling. As a result, we expanded the data collection to include weather data for 40 popular cities to ensure a more comprehensive analysis, as suggested in Greg's lecture. This method was mentioned in Greg's lecture. Additionally, we have top 65 destinations in our csv file because some cities were not in the datasets. We used 40 cities in the ranking that we could find from the datasets. In get_unpopular cities_data.py, we collected 40 unpopular tourist destinations and in get_random_cities_data.py, we collected 20 random stations as our test data. In these three files, we selected the name of the station, station code, tmax, tmin, and prcp, and dropped all the other features such as latitude and longitude. Also, the 'avg_tmax' and ‘avg_tmin’ column in the weather data was °C×10 so we divided by 10. Then we joined all the tmax, tmin, and prcp dataframe into a single weather dataframe using the name of the station and station code. Finally, we filtered the weather data to include only the cities we planned to use in our analysis. After cleaning the data, we stored it in JSON format to make it easily usable for further analysis.

Analysis Techniques

Basic Statistics

We calculated the mean values of the average temperature, minimum temperature, and precipitation by applying python mean function and rounded it two decimal points. [see Basic Statistics]

Inferential Statistics

We used t-tests to see if there is any difference on the mean value of each weather variable (tmax, timn, prcp) between popular cities and unpopular cities. In order to use t-tests, we also performed normality tests and equal variances tests which proves our data sets satisfy the assumptions of t-tests. [see Inferential Statistics]

ML Classification

We first take training data and validation data out of combined data of popular cities and unpopular cities. Then, we used a voting classifier to model our dataset. The five voted classifiers we used are bayesian classifier, k-nearest neighbors classifier, support-vector machine classifier, and decision tree classifier. Using that model, we predicted the potential popular travel destinations out of randomly chosen cities.
[see Machine Learning Classification]

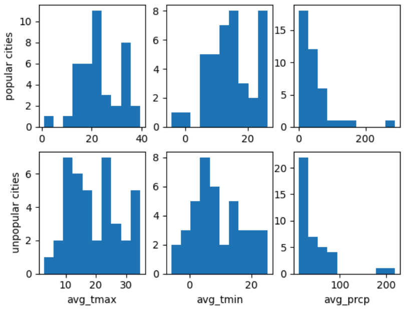
Results and Findings

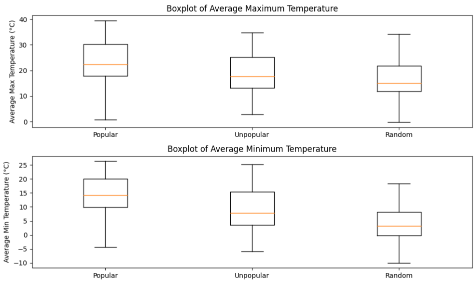
Our results revealed several important findings on our datasets of popular and unpopular tourist destinations. Our analysis comprises basic statistics, inferential statistics and machine learning techniques yielding insights on top tourist destinations and non-listed cities.

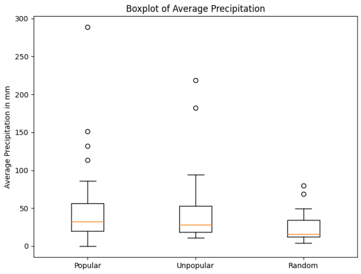
Basic Statistics

(from stats.py)

Through basic statistical analysis, we were able to plot a bar graph and two boxplots with temperatures and precipitation. We looked into popular cities, unpopular cities and random cities. [*Please note that generating random cities will always be different, this data is based on what is in our repo upon submitting.*]







Upon seeing the bar graph of average precipitation, it becomes evident that the data exhibited a right-skewed distribution. To ensure this skewness, we applied data transformation techniques on the average precipitation. After the transformation, we observed more bell-curved shaped [see After Transforming for more details]. As shown in the boxplot for average precipitation, we noticed some outliers but decided to keep them. A further detailed of the boxplot can be found in the table of means and standard deviations below:

	Popular Cities	Unpopular Cities
Maximum Temperature Mean (°C)	23.17	19.49
Minimum Temperature Mean (°C)	14.5	9.07
Precipitation Mean (mm)	48.04	43.91
Maximum Temperature Standard Deviations (°C)	8.23	8.32
Minimum Temperature Standard Deviations (°C)	7.42	8.01
Precipitation Standard Deviations (mm)	51.6	43.3

Here, we noticed that the standard deviation for precipitation is quite large for both popular cities and unpopular cities, yielding ±51.6 and ±43.3 respectively. This suggests that precipitation varies greatly and it is very spread out.

We wondered if the means differ from one another before we come to any conclusions about the weather in top tourist destinations. There is more on this later in the report.

Inferential Statistics

(from infer_stats.py)

Before we can perform t-tests, we have to ensure that we satisfy the t-tests assumptions of the two populations: normal distribution and with equal variances. Normality tests and equal variance tests are conducted to show if our datasets satisfy the assumptions. Also these tests were used to see if we could even do a Bayesian Classifier as it also has the assumptions of having data to be normally distributed. We conducted normality tests on popular cities, unpopular cities and combined cities to check if our datasets satisfies the assumption.

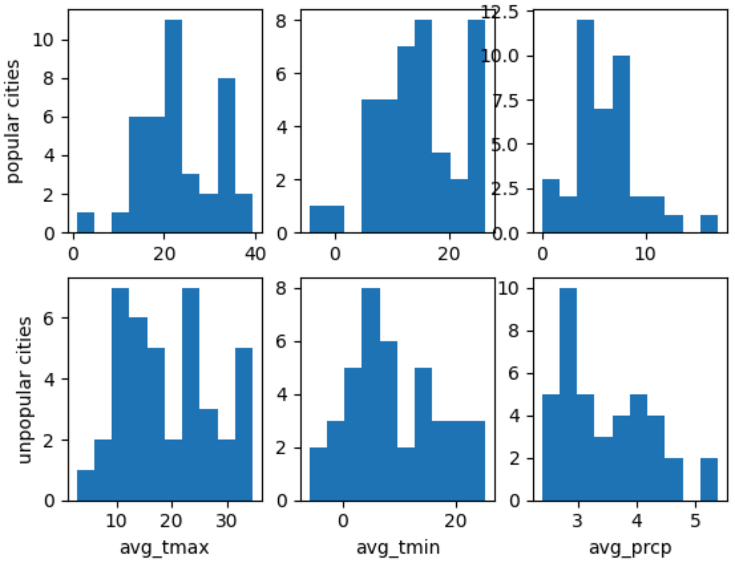
Normal Distribution Tests

pvalue	avg_tmax	avg_tmin	avg_prcp
Popular cities	0.8362 -> Normally Distributed	0.9360 -> Normally Distributed	**2.7808e-11 -> NOT normally distributed**
Unpopular cities	0.1618 -> Normally Distributed	0.2928 -> Normally Distributed	**1.7913e-09 -> NOT normally distributed**
Combined cities (popular & unpopular)	0.3673 -> Normally Distributed	0.2533 -> Normally Distributed	**4.7574e-16 -> NOT normally distributed**

For avg_tmax and avg_tmin, we see that the normality tests on average precipitation of popular cities, unpopular cities, and combined cities have failed tests with $p<0.05$. For Normal Distribution, we are looking for $p > 0.05$ to reject the null hypothesis. This is apparent from the very right skewed histograms of average precipitation above [see Basic Statistics]. We then conducted transformation techniques to normalize our precipitation data.

After Transforming

pvalue	avg_tmax	avg_tmin	avg_prcp
Popular cities	0.8362-> Normally Distributed	0.9360-> Normally Distributed	**0.0015 -> NOT normally distributed**
Unpopular cities	0.1618 -> Normally Distributed	0.2928 -> Normally Distributed	**0.1622 -> Normally Distributed**
Combined cities (popular & unpopular)	0.3673 -> Normally Distributed	0.2533 -> Normally Distributed	**0.0033 -> NOT normally distributed**



We transformed the precipitation data to make it normally distributed by taking square root of the precipitation value of popular cities and combined cities data and by taking log of unpopular cities data. Although we still failed the normality tests for the precipitation data of popular cities and combined cities, we can assume that our data is now normally distributed based on the central limit theorem. The precipitation graph of our data looks close to normal distribution and the size of our dataset of popular cities and unpopular cities is above 40 and the size of our dataset of popular cities is above 80. Therefore, we conclude that our datasets are normally distributed.

Equal Variances Tests

We conducted Levene's test of equal variances on the relevant weather parameters. We wanted to check the appropriateness of using the t-test.

	avg_tmax	avg_tmix	avg_prcp
Combined_cities (popular & unpopular)	0.5530 -> Equal Variance	0.3933 -> Equal Variance	0.3933 -> Equal Variance

We see that the datasets passed the tests with $p > 0.05$, therefore our dataset has equal variances.

T-Tests

We were able to do t-tests as it satisfied its assumptions of normally-distributed data and having equal variances [See Normal Distribution Tests and Equal Variances Tests]. This statistical test allowed us to compare each weather component that we were interested in. This allowed us to understand the significant differences in weather conditions by two groups namely popular and unpopular cities. Our hypothesis are as follows:

For maximum temperature:
H_0: The average maximum temperature between popular and unpopular cities are the same.
H_a: There is a difference in average maximum temperature between popular and unpopular cities.

For minimum temperature:
H_0: The average minimum temperature between popular and unpopular cities are the same.
H_a: There is a difference in average minimum temperature between popular and unpopular cities.

For precipitation:
H_0: The average precipitation between popular and unpopular cities are the same.
H_a: There is a difference in average precipitation between popular and unpopular cities.

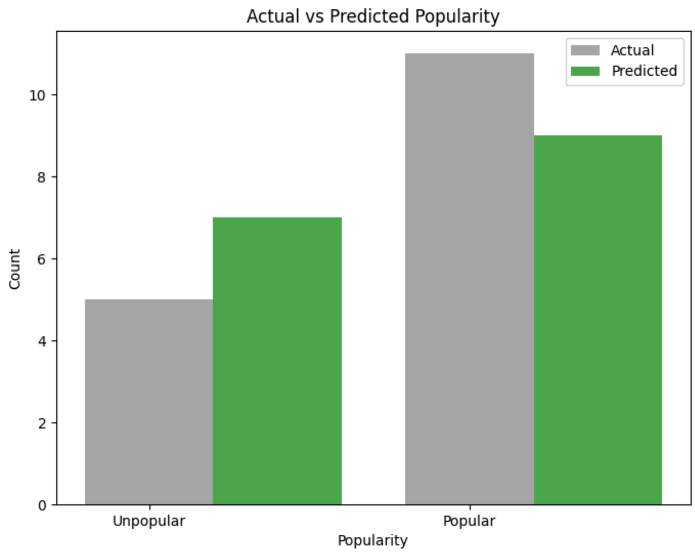
	p-value	Conclusion
Maximum Temperature	0.044	Reject the NULL Hypothesis
Minimum Temperature	0.002	Reject the NULL Hypothesis
Precipitation	8.031	Reject the NULL Hypothesis

According to our t-tests, we can accept that there is a difference in means for maximum temperatures in popular and unpopular cities. The case is the same for minimum temperatures and precipitation.

With this, we were curious to see the cities that have potential to be in the list of top tourist destinations based on weather, more on this in Machine Learning Classification part of our report.

Machine Learning Classification

(from modeling.py)



Our graph shows that predicted popularity in unpopular destinations were a bit overestimated and slightly underestimated for popular destinations. Our model did a decent job at recognizing popular and unpopular cities based on our weather conditions with the scores below.

Training Score	Validation Score
0.921875	0.75

In our modeling part, we modeled the combined dataset of popular cities and unpopular cities to predict if a city can potentially be a popular travel destination. The validation score shows our modeling is able to sufficiently predict the popularity of cities although we still have room for improvement.

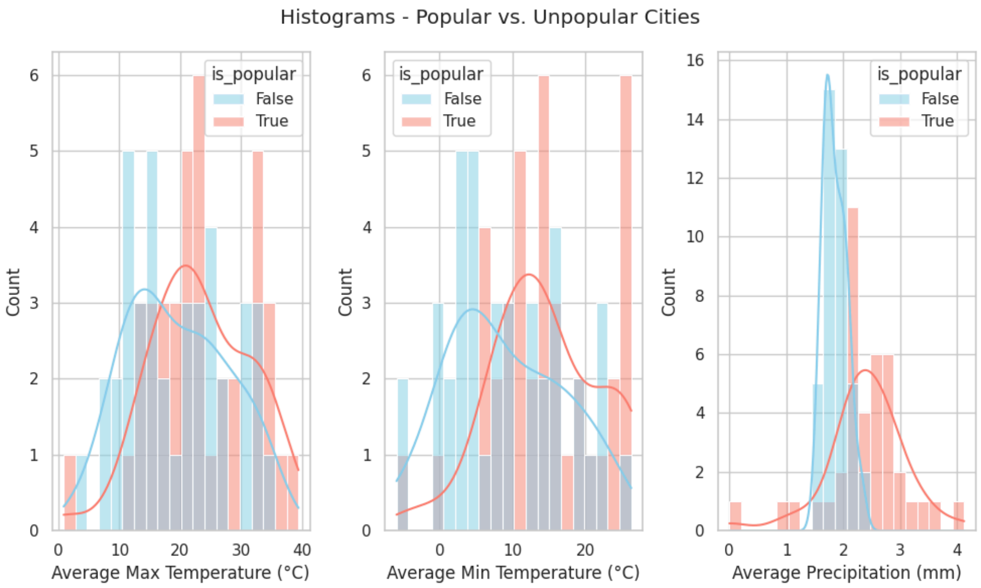
We also made our model predict popularity on random cities datasets and outputted a json file(test_data_predictions.json) which displayed our predictions. We can see some cities that have potential to be a top tourist attraction based on weather.

Limitations

In our project, we were not able to find weather data of cities. Instead, we used stations' weather data. Since we assume that the weather in a station is the same as an entire city, which would likely not be true, there is a probability that our analysis does not apply in real life. Also, we limited the number of weather variables to three (tmax, tmin, prcp) and did not consider any other variables. Therefore, our modeling can be very different from real life weather modeling which has much more weather variables such as humidity, snow fall, and so forth. In addition to the limitation of our dataset, we encountered some technical problems. To create the datasets of popular cities and unpopular cities, we needed to manually check if the station was located in a popular city or in an unpopular city since our original data was based on stations, not on cities. If we had more time, we could write code that could automatically check if a station in the original dataset was in a popular city or in an unpopular city. Then we were able to create datasets of a larger size and be able to produce more accurate analysis on our hypothesis.

Conclusion and Future Work

In conclusion, we can answer our question that top tourist destinations have a difference in maximum, minimum temperatures, and precipitation than unpopular tourist cities. Since there are a difference in means as concluded by the t-tests we conducted, we can conclude that the top tourist destinations have an average of maximum temperature of 23.17 °C ± 8.23, a minimum temperature of 14.5 °C ± 7.42 and a precipitation on average of 48.4mm ± 51.6. Since there is a large standard deviation for precipitation, we can conclude that on average, top tourist destinations have a tendency of having temperatures between 14.5°C and 23.17°C. This range should be what is considered comfortable for most tourists.



Our machine learning model prediction on our random cities suggests that there are some cities that have the potential to be a top tourist location based on weather alone. From the model, places that have weather similar to top tourist destinations are Camooweal Township (station":"ASN00037010") and Nevshir ("station":"TUM00017193"). These cities can work on how they can improve their tourism as they have the potential for making it to Wikipedia’s curated list, List of cities by international visitors.

This analysis we conducted on weather patterns in top tourist destinations and non-listed cities can serve as a valuable starting point for cities to explore their tourism potential based on weather conditions. By identifying the distinct weather characteristics that sets popular cities apart, smaller or less known cities can gain insights into factors that attract international travelers. Of course, there are other factors that contribute to tourism but this is a good start for cities with high-touristic cities like weather but by understanding the influences of weather on tourism preferences can inspire these cities to leverage their climate features to attract visitors and boost their tourism.