Overview

Created a webscraper program to get data from Wikipedia table

Used Pandas to parse text into csv files

Used Pyspark to process data into usable format

Analyzed data using a 1 sided t-test.

Analyzed significance using Anova tests.

Introduction

This project analyzes whether weather in the 100 most visited cities around the world is milder than the weather in other cities with populations above 15000. Specifically, the Pandas is used to read reformat and read data alongside Pyspark, then, 1 sided t-tests are used to compare the temperature and precipitation of the cities and Anova tests are used to determine the findings' significance.

Data Gathering and Refinement

The *weather-3* dataset was acquired from the SFU cluster and is based off GHCN data. It is a collection of the weather throughout 2016 reported by a variety of different stations throughout the world. This data is read into a Pyspark dataframe using the *observation_schema* also found in the exercise 8 files.

The initial dataset was supplemented through a few files. *ghcn-stations.txt* provides latitude/longitude for the various reporting stations, and *ghcn-countries.txt* provides keys for the abbreviations GHCN uses to classify its stations. These files can be found on the GHCN website at https://www.ncei.noaa.gov/pub/data/ghcn/daily/ghcnd-stations.txt and https://www.ncei.noaa.gov/pub/data/ghcn/daily/ghcnd-countries.txt.

Another dataset was needed to match the stations and their latitude/longitude coordinates to physical locations – *cities15000.txt*. This file contains cities around the world with more than 15000 residents living in them alongside their latitude/longitude coordinates. This can be found at: https://download.geonames.org/export/dump/.

The approximately 100 most visited cities in the world were found on Wikipedia, at this link found on Coursys: https://en.wikipedia.org/wiki/List_of_cities_by_international_visitors.

Data Refinement

weather-3 is read into a spark dataframe and provides the initial frame which the other files add to or alter. It is also filtered to make sure that there are no "null" rows within it.

The txt files (*cities15000.txt, ghcn-countries.txt, ghcn-stations.txt*) are converted into csv files, which are then read into Spark dataframes. Using inner joins, the data is added to the *weather-3* set and is then used to find which city the reporting stations are within ~55km of. This is achieved by taking the absolute difference of their longitude and latitude coordinates. Lastly, the data is split into two sets; one for the popular cities, and one for all cities.

Considerations

The data used is dated to 2016, meaning that changes in climate and travel trends are not reflected in its results. Furthermore, latitude/longitude values are accurate only up to 4 decimal digits and are sourced from two different data providers. The method which cities are found from their latitude/longitude coordinates is also unrefined, relying upon the absolute difference in latitude and longitude not exceeding 0.35, or about 55km.

Analysis

The results point at there not being much significant results, other than the precipitation perhaps having a correlation.

Limitations

Much time was spent attempting to work with the geopy library during the development period, however, many errors such as http type, and the dataset's size caused the usage of geopy to become more undesirable. Working with APIs in general was difficult due to the chosen scope, however, this could have been mitigated in various ways, not limited to partitioning the dataset,

or otherwise refining the data to fit a narrower question. Were this project to be repeated, much more time would be spent on the analysis portion, as it is rather incomplete.