Weather Project Analysis Report

By: Jasleen Kaur (301557247),

Arshdeep Kaur (301540172)

Table of Content:

Overview:

This project examines the correlation between weather conditions and tourist visitation in various cities. It involves a multi-step process, including data extraction, preprocessing, analysis, visualization, hypothesis testing, and machine learning techniques. The analysis focuses on weather data from the Global Historical Climatology Network (GHCN) for 2016 and 2017 and tourist destination information from Wikipedia for 2018.

Workflow:

1. Data Extraction:

Tourist Data Identification:

To begin the project, we manually identified valid station IDs for cities from Wikipedia's list of the top 100 tourist destinations. We extracted these station IDs from GHCN map by searching individual cities and getting the valid station IDs. We tried to get the airport station IDs, if available and made sure that the station end data is not before 2020. The data with station IDs, city and country names and the number of visitors is available in the 'station_cities.csv' file.

We were unable to obtain the city names from the latitude and longitude data provided in the ghcnd-stations.txt cluster. We attempted to use the geopy library, but it only returned addresses in various formats, which made it difficult to extract the city and country names. Therefore, we decided to extract the station ID manually for the top 100 tourist cities, as it seemed like a more sensible option.

Weather Data Collection:

We accessed the GHCN data for weather information from the cluster. Initially, we filtered the dataset to extract the required data for the years 2016, 2017, and 2018. We removed the unwanted columns, such as qflag, sflag, mflag, and obstime. After that, we transformed the data from a long format to a wide format, and created separate columns for tmin, tmax, and prep to make calculations more straightforward. Next, we merged the station_cities data with the GHCN data to obtain the required data for the analysis. We utilized broadcast with station, as it is a smaller dataset. The output folder named 'ghcn-output' was generated by the 'weather_data_extraction.py' file, which contains the 2016 and 2017 data.

Please note that we tried to obtain data for 2018, as the tourist data available is for that year. However, none of the station IDs we are using have data for that year.

2. Data Preprocessing:

Data Processing:

The data we extracted was still quite large, containing individual TMAX, TMIN and PRCP values for every month in 2016 and 2017 for all the cities. We noticed that some of the rows had null values when we looked at the data in ghcn-output. To address this, we decided to use mean imputation to fill in the null values for columns TMAX, TMIN, and PRCP with specific averages. We also filtered the yearly data to get separate data frames for 2016 and 2017.

For each year, we calculated the average TMAX, TMIN, and PRCP for all the cities and joined the datasets. We then created a dataset with station ID, city, country, Number of visitors, avg_tmax_2016, avg_tmin_2016, avg_prcp_2016, avg_tmax_2017, avg_tmin_2017, avg_prcp_2017. This data was saved in a json.gz file located in the 'calcout' folder.

3. Extraction and Calculations:

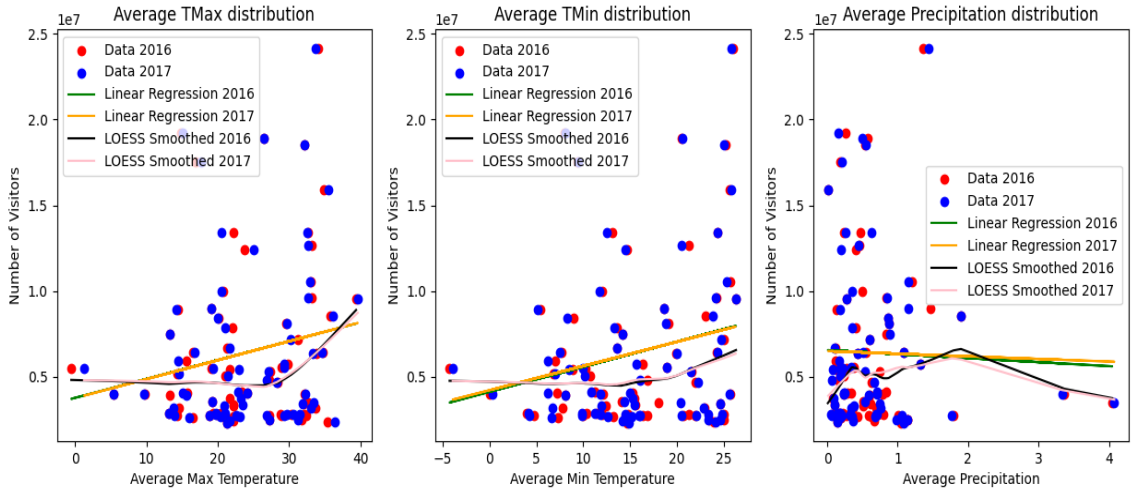
Plots:

We used the extracted data to create scatter plots for the three averages: avgTMAX, avgTMIN, and avgPRCP. In each scatter plot, the x-axis represents the respective average values, while the y-axis represents the number of visitors. The red dots in the plot represent the 2016 data, and the blue dots represent the 2017 data. We added a green best-fit line to the 2016 data and an orange best-fit line to the 2017 data. In addition, we included a black line representing the best-fit line using loess smoothing on the 2016 data and a pink line representing the best-fit line using loess smoothing on the 2017 data. The plots yielded the following R-values:

	T-Max	T-Min	PRCP
2016	0.18322747385676053	0.21478718168726027	-0.03236144148889365
2017	0.18164507314557002	0.21028234995891226	-0.02120327600922051

Correlation coefficient: The values provided indicate the strength and direction of the linear relationship between the number of tourists visiting a city and the corresponding weather attributes for the specified years. An R-value closer to 1 or -1 signifies a more substantial positive or negative correlation, respectively. Values closer to 0 indicate a weaker correlation.

In this context, the T-Min attribute for both years shows a slightly stronger positive linear correlation compared to T-Max. However, the Precipitation attribute exhibits a weak negative linear correlation. This suggests a slight inverse relationship between precipitation and tourist visitation. Which means that cities with higher average Tmax and Tmin and lower average precipitation tend to have more tourists visiting them.



(From the above graphs we can see that there is positive relationship between Tmax and Tmin with number of visitors and negative relationship between precipitation and number of visitors)

Hypothesis Testing:

We used the 'testing.py' script to conduct hypothesis tests and employed t-tests to predict the relationship between weather attributes and tourist numbers. As the number of datasets is greater than 40 so we assume that the data is normally distributed and proceeded with the ttest. The results are as follows:

1. T_value:

	T-Max	T-Min	PRCP
2016	-11.39129091768319	-11.391305683779908	-11.39133232356949
2017	-11.391290857186446	-11.391305913998469	-11.39133237880197

2. p-value:

	T-Max	T-Min	PRCP
2016	7.801251895642995e-22	7.800556526908867e-22	7.799302155942365e-22
2017	7.801254744698219e-22	7.80054568588881e-22	7.799299555454896e-22

P-value analysis: Our study aimed to investigate the relationship between weather conditions and the number of tourists in a city. To do this, we formulated a null hypothesis that there is no relationship between weather and tourism. We conducted t-test statistics on different weather conditions, including average maximum temperature, average minimum temperature, and average precipitation, for the years 2016 and 2017. The results show a remarkably low p-value of approximately 7.8e-22 which is unlikely due to random chance, indicating strong evidence against the null hypothesis. Hence, we reject the null hypothesis and accept the alternative hypothesis that there is a statistically significant relationship between weather conditions and

the number of tourists visiting a city. This analysis underscores a consistent and robust correlation between weather and tourism, signifying the impact of weather on tourism.

Machine Learning:

After our initial analysis, which indicated a correlation between weather conditions and tourist numbers, we decided to explore deeper by using machine learning techniques to predict tourist visitation based on weather attributes. We utilized the 'machineLearning.py' script on 2016 weather data to train a random forest regressor model that could forecast tourist numbers. (Note: we do not use the averaged weather data rather all the dataset available with us for year 2016 for our 75 stations as it is better to train the data on larger dataset). To achieve this, we used techniques such as train_test_split for dataset division and standardScaler for weather value scaling. Surprisingly, we achieved a test data score of 0.614. Through our exploration, we found that using all three factors—maximum temperature, minimum temperature, and precipitation—yielded more accurate predictions. We applied this model to predict 2017 tourist figures, while the predictions did not perfectly mirror the actual numbers, they were somewhat close. The model's accuracy, while not exceptionally high, still indicated a reasonable ability to predict tourist visitation based on weather data using machine learning models.

4. Results:

After conducting a thorough analysis, it was found that there is a consistent and significant correlation between weather conditions such as maximum temperature (TMAX), minimum temperature (TMIN) and precipitation (PRCP), and the number of tourists visiting various cities from 2016 to 2017. By rejecting the null hypothesis and successfully using machine learning in predicting the outcomes, the statistical significance and practical impact of weather on tourist visitation patterns in different cities were strongly confirmed. These findings reinforce the idea that weather plays a critical role in influencing tourist behavior, which in turn can shape tourism patterns across different cities.

5. Limitations:

Our study has some limitations that need to be acknowledged. One of the most significant issues was the limited availability of complete weather datasets for all the desired cities. Although we had tourist data for the top 100 cities in 2018, we did not have the weather data for all of them, as only some cities had a station ID that we could use to obtain the necessary information. Furthermore, machine learning models require further fine-tuning to improve predictive accuracy. It is also worth noting that tourism is influenced by a variety of other factors besides weather, such as famous sites, food, and the world's wonders.

Additionally, we had access to yearly tourist numbers for the cities we studied, but it would have been better if we had access to monthly tourist data. With this data, we could have seen the effect of temperature on the tourist numbers in a particular city.

Project Experience Summary

Jasleen Kaur

During my time as a member of a team project, my focus was on analyzing the impact of weather on tourism in top international tourist destinations. I was actively involved in various aspects of the project, showcasing my skills in data extraction, analysis, and machine learning. My primary responsibilities and achievements were as follows:

1. Validating Station IDs: Conducted a thorough check of valid station IDs for the cities in our dataset. This was to ensure the accuracy and reliability of weather data by validating station IDs, which is a crucial step in maintaining data integrity.

2. Weather Data Extraction: Developed the `weather_data_extraction.py` script to extract weather data from the cluster. Implemented efficient data extraction procedures to gather relevant information for further analysis.

3. Data Processing and Calculations: Developed the `calculations.py` script, focusing on data processing and calculations. I calculated yearly averages of Tmin, Tmax and Prcp and applied data filtering techniques to isolate the relevant weather data for 2016 and 2017.

4. Machine Learning Implementation: I collaborated with my team member in developing the `machineLearning.py` script for machine learning analysis and implementing the Random Forest Regressor model to train and predict the number of visitors given tmax, tmin and prcp values of a city. We trained the model on 2016 data and successfully predicted tourist numbers in 2017.

5. Documentation and Collaboration: I contributed significantly to creating the project's final report and README file. I ensured clear code, methodologies, and results documentation, enhancing the project's transparency and replicability. I also collaborated closely with my team member by engaging in regular discussions to address challenges, share insights, and refine project strategies.

6. Git Collaboration: I supported my partner in navigating Git commands and version control. I facilitated smooth collaboration using Git, ensuring a well-organized and versioned codebase.

Throughout the project, I actively engaged in discussions with my partner, ensuring a comprehensive understanding of each aspect of the analysis. By combining technical expertise with effective collaboration, we successfully navigated the complexities of the project, resulting in a well-executed analysis of the relationship between weather conditions and tourism in top international destinations. This experience has further honed my data analysis, machine learning, and collaborative problem-solving skills, positioning me as a valuable contributor to future projects in similar domains.

Project Experience Summary

Arshdeep Kaur

Along with my team partner, I brainstormed the project, how we would go about it, and what tests and analysis tools we would use to answer our question of whether the weather impacts tourism worldwide. We looked at three primary weather data for this: maximum temperature, minimum temperature, and precipitation, including rainfall and snowfall. My primary responsibilities and achievements are as follows:

1. Manual extraction the station ID: Successfully executed the manual extraction of station IDs for the top 100 cities, demonstrating a meticulous approach and attention to detail. The acquired station IDs serve as a crucial foundation for subsequent data analysis, contributing to the efficiency and precision of our project.

2. Researching spark command: Conducted research for spark command for successful retrieval of files required for working of the project. Leveraged acquired knowledge to proficiently execute file copying and uploading tasks from local computer to the cluster and vice versa.

3. Inferential statistics: Designed and implemented the 'testing.py' program, showcasing analysis of weather parameters for 75 cities across the years 2016 and 2017 by generating three insightful graphs that provided a comprehensive overview of weather trends. Additionally, it conducts the t-test, which is essential in concluding whether there is a correlation between the weather parameters and the number of tourists.

4. Machine Learning Implementation: Collaborated with my project partner to implement ‘machineLearning.py’ which is a machine learning model to predict the score of the predictions. During the project, we conducted thorough experiments with various machine learning models, using a systematic approach to determine and implement the most efficient model that suits our dataset. This iterative process enabled us to enhance the predictive accuracy of our model and obtain valuable insights for predicting visitor numbers.

5. Documentation and collaboration: Coordinated with my project partner to create a comprehensive README file that details the data extraction and preprocessing steps, as well as the execution commands. Additionally, I co-authored the project report, which highlighted the project's workflow, methodologies, and critical findings.

Throughout the project, I made sure to communicate actively with my team members to ensure that we all understood the requirements and expectations. By staying involved and attentive to each team member's insights, we were able to effectively address challenges and deepen our understanding of the project's. Our continuous and transparent communication played a significant role in the project's success, highlighting my commitment to proficient and collaborative project management.