640 Project Proposal

Milestone 2

Isabella Sturm

2021-09-09

**Introduction**

**Background**

Seattle has a reputation as being the rainiest city, even nicknamed the rainy city. On average it rains 150 days per year, accumulating an average of 37 inches. While Seattle might not get the most rain in terms of inches of precipitations, it is still one of the largest complaints about Seattle. Perhaps, it may be the most well-known “fact” about Seattle. Despite the rainy weather, Seattle is a city with so much to offer for outdoor activities, including the famous Pike’s Place Market and many walking or hiking trails and parks. One can plan to experience both mountains and beaches in the area, but there is just one problem: How can a visitor or a Seattle native plan for the outdoor offerings of Seattle while also avoiding rain?

**Problem Statement**

Tourists coming to Seattle want to be able to enjoy the best of Seattle and not be stuck indoors because of their disdain for rain. Residents of Seattle want to be able to take their dogs to the park or walk along their pier and ideally avoid the rain.

Given a date of travel, what is the likelihood it will not rain?

How likely is the chance of rain given the high and low temperature for the day?

**Scope**

The scope of this project focuses on Seattle precipitation data from 1948 through 2017. The only factors at this point will be Date (month, week, day), minimum daily temperature and maximum daily temperature determining whether there was (and for this project will be) rain.

**Document Overview**

This document will outline the following:

* Preliminary Requirements
  + Technical Approach
  + Data Sources, Data Plan
  + Analysis
  + Requirement Development
  + Model Deployment
  + Testing and Evaluation
* Expected Results
* Project Execution
  + Project Plan
  + Project Risk

**Preliminary Requirements**

**Technical Approach**

To accomplish this project, data will be downloaded from Kaggle and if necessary, I can find more Seattle weather data in Kaggle or through a weather API and import the data into R and Python. From there data will be explored, analyzed and modelled so predictions about the data can be made.

**Data Sources & Data Plan**

The initial data source I found is from Kaggle: <https://www.kaggle.com/rtatman/did-it-rain-in-seattle-19482017> .

This dataset includes the date of the record (DATE), the min temperature (TMIN) and max temperature (TMAX) of that day, the amount of precipitation (PRCP), and whether or not it rained (RAIN).

|  |  |  |
| --- | --- | --- |
| Column Name | Type | Description |
| DATE | Date, yyyy-mm-dd | Date of record |
| PRCP | Float/Decimal number to 2 places | Amount of precipitation in inches |
| TMAX | Integer | Max temperature in Fahrenheit |
| TMIN | Integer | Min temperature in Fahrenheit |
| RAIN | Boolean | Indicates whether or not it rained on the observed day |

If more data/variables are needed, more data can be collected using a weather API such as OpenWeatherMap: <https://openweathermap.org/history> .

**Analysis**

Using Python via Jupyter Notebook and R via RStudio, initial exploratory analysis will be performed first to gain insights about the variables and determine any correlation and useful statistics that can help in predicting whether it will rain in Seattle. From here, I can test categorical models and find a good fit model for predicting precipitation in Seattle.

**Requirement Development**

To develop for this project, first preparing the data will be necessary. This includes cleaning any data or filling in missing values. I also plan on calculating additional variables from the date field including month, week of year, and day of week. After performing a first round of exploratory analysis, it might be necessary to collect more data or variables using weather APIs or finding another dataset that can be merged with the current dataset. Once initial insights are gained, model testing and evaluation is next. Additional research may be necessary in helping determine the proper model for the data.

**Model Deployment**

For this project, since the data is not live it cannot really be deployed, but we can split the data and make predictions retroactively and use the data to see how accurately the data outcome was predicted.

**Testing and Evaluation**

The data will be split up into training and testing data to test the model. Depending on the model selected, there are different ways to evaluate the model. For instance, it would be good to check R-squared value and graph residuals. A confusion matrix could also help in evaluating the model.

**Expected Results**

Weather is variant though predictable to some degree. Based on my knowledge of Seattle weather, I wouldn’t expect it to rain much in the summer months through September, but I would expect more rainy days between October and March. These are the months that are considered the rainiest in Seattle.

Generally, however, there are two possible outcomes in the data for predicting if it will rain in Seattle: TRUE and FALSE. TRUE indicating it will rain, and FALSE indicating it will not rain in Seattle.

**Execution**

**Project Plan**

While plans can change, below is an outline of the plan I will attempt to follow throughout the term to complete this project:

* Week 1: Come up with project plan.
* Week 2: Research data sets and come up with project proposal.
* Week 3: Begin exploration of the data. Prepare questions to ask about the data and select variables to use for analysis.
* Week 4: Prepare data for further analysis and modelling. Revisit data exploration stage if necessary.
* Week 5: Finish up preliminary analysis.
* Week 6-8: Test and evaluate models.
* Week 9: Begin presentation preparation of intermediate results so far.
* Week 10-11: Finalize models and visualizations showing results of the analysis.
* Week 11-12: Finalize presentation.

**Project Risk**

There are several risks that can be associated with any data science project. For this project, one potential risk is not having enough variables in the data to make an accurate prediction. One solution to this is to combine the current dataset with other data that can be pulled from a weather API. As with any project, another risk for this project is improper analytics, as a learning data scientist, there are many errors I can make during data analysis that can affect the progress or accuracy of the entire project. The best I can do to mitigate this risk is to do reduce impact of biases and be diligent in researching and performing techniques learned in this course and previous courses while analyzing the data.