**CP2403: Project**

Dengue fever is a mosquito-borne disease that occurs in tropical and sub-tropical parts of the world. In mild cases, symptoms are similar to the flu: fever, rash, and muscle and joint pain. In severe cases, dengue fever can cause severe bleeding, low blood pressure, and even death.

Because it is carried by mosquitoes, the transmission dynamics of dengue are [related to climate variables](http://ehp.niehs.nih.gov/wp-content/uploads/121/11-12/ehp.1306556.pdf) such as temperature and precipitation. In recent years dengue fever has been spreading. Historically, the disease has been most prevalent in Southeast Asia and the Pacific islands.

Your task is to explore the Dengue dataset provided 2 Latin American cities (San Juan and Iquitos) and predict the number of dengue cases based on environmental variables describing changes in temperature, precipitation, vegetation, and more.

The project has 2 parts and the due date for each part is as following: -

|  |  |  |  |
| --- | --- | --- | --- |
| **Parts** | **Marks** | **Relevant Material** | **Due Date** |
| Part 1 | 10% | Lecture 1 – 4 | Week 6,  Thursday, 29 March 2017  5pm |
| Part 2 | 30% | Lecture 5 – 12 | Week 13,  Friday, 25 May 2017  5pm |

**Part 1: 10%**

**Exploration & Visualization**

Apply appropriate data management techniques to the Dengue dataset provided. Eg. Change blank data to nan, recoding data, etc.

From the Dengue dataset provided,

1. Create a box plot to show the distribution of number of dengue cases in both San Juan (sj) or Iquitos (iq) city.
2. For either San Juan (sj) or Iquitos (iq) city, create a histogram to show the distribution of number of dengue cases.

y: number of week

x: number of case

1. For either San Juan (sj) or Iquitos (iq) city, create a line chart to show the number of dengue cases for each week.

x:number of case

1. For either San Juan (sj) or Iquitos (iq) city, create a bubble chart to show the relationship between average temperature (station\_avg\_temp\_c), total precipitation (station\_precip\_mm) and number of dengue cases. Use number of dengue cases as the bubble size.u

Hint: Refer to Lectures 2, 3 and 4 and Practicals 2, 3 and 4 for help on data management and data visualisation

**Part 2: 30%**

**ANOVA - Hypothesis 1**

We want to find out if the mean number of dengue cases per month is equal for the two cities, San Juan (sj) or Iquitos (iq).

Hint: Refer to Lecture 5 and Practical 5 for help on ANOVA analysis

**Multiple Regression - Hypothesis 2**

We can predict the number of dengue cases in each city based on temperature, precipitation, dew point, humidity and vegetation index. Using the following variables listed below (all or some variables), use multiple and/or polynomial regression to predict the number of dengue cases for either San Juan (sj) or Iquitos (iq). Evaluate your regression model using qq plot and standardized residual plots

* station\_max\_temp\_c – Maximum temperature
* station\_min\_temp\_c – Minimum temperature
* station\_avg\_temp\_c – Average temperature
* station\_precip\_mm – Total precipitation
* station\_diur\_temp\_rng\_c – Diurnal temperature range
* precipitation\_amt\_mm – Total precipitation
* reanalysis\_sat\_precip\_amt\_mm – Total precipitation
* reanalysis\_dew\_point\_temp\_k – Mean dew point temperature
* reanalysis\_air\_temp\_k – Mean air temperature
* reanalysis\_relative\_humidity\_percent – Mean relative humidity
* reanalysis\_specific\_humidity\_g\_per\_kg – Mean specific humidity
* reanalysis\_precip\_amt\_kg\_per\_m2 – Total precipitation
* reanalysis\_max\_air\_temp\_k – Maximum air temperature
* reanalysis\_min\_air\_temp\_k – Minimum air temperature
* reanalysis\_avg\_temp\_k – Average air temperature
* reanalysis\_tdtr\_k – Diurnal temperature range
* ndvi\_se – Pixel southeast of city centroid
* ndvi\_sw – Pixel southwest of city centroid
* ndvi\_ne – Pixel northeast of city centroid
* ndvi\_nw – Pixel northwest of city centroid

Hint: Refer to Lecture 7 and 8 and Practical 7 and 8 for help on linear, multiple and polynomial regression

**Time Series Analysis - Hypothesis 3:**

Assuming that there is a seasonality trend in the number of dengue cases, perform time series analysis for either San Juan (sj) or Iquitos (iq).

Hint: Refer to Lecture 10 and Practical 10 for help on time series analysis

**Additional Note:**

The Dengue dataset provided is part of the DengAI: Predicting Disease Spread competition hosted by DrivenData (<https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/page/80/>). As CP2403 students, you are welcome to extend your project and participate in the competition. However, participation in the competition is **NOT A REQUIREMENT** for CP2403.

**Project – Part 1 (10%) Rubric**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Criteria** | **Exemplary (9, 10)** | **Good (7, 8)** | **Satisfactory (5, 6)** | **Limited (2, 3, 4)** | **Very Limited (0, 1)** |
| **Data Management** | Applied excellent data management techniques to the dataset provided | Exhibits aspects of exemplary (left) and satisfactory (right) | Applied satisfactory data management techniques to the dataset provided | Exhibits aspects of satisfactory (left) and very limited (right) | Applied limited or no data management techniques to the dataset provided |
| **Data Visualisation -**  **Box Plot** | Created excellent box plot with appropriate title, axis and legend label | Created satisfactory box plot but chart labels such as title, axis and legend label are missing | Created limited or no box plot  No title, axis and legend label |
| **Data Visualisation - Histogram** | Created excellent histogram with appropriate title, axis and legend label | Created satisfactory histogram but chart labels such as title, axis and legend label are missing | Created limited or no histogram  No title, axis and legend label |
| **Data Visualisation – Line Chart** | Created excellent line chart with appropriate title, axis and legend label | Created satisfactory line chart but chart labels such as title, axis and legend label are missing | Created limited or no line chart  No title, axis and legend label |
| **Data Visualisation – Bubble Chart** | Created excellent bubble chart with appropriate title, axis and legend label | Created satisfactory bubble chart but chart labels such as title, axis and legend label are missing | Created limited or no bubble chart  No title, axis and legend label |

**Project – Part 2 (30%) Rubric**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Criteria** | **Exemplary (9, 10)** | **Good (7, 8)** | **Satisfactory (5, 6)** | **Limited (2, 3, 4)** | **Very Limited (0, 1)** |
| **ANOVA** | Applied excellent data management techniques to the dataset provided | Exhibits aspects of exemplary (left) and satisfactory (right) | Applied satisfactory data management techniques to the dataset provided | Exhibits aspects of satisfactory (left) and very limited (right) | Applied limited or no data management techniques to the dataset provided |
| **Regression**  **(Worth Double)** | Applied excellent regression techniques (linear, multiple, polynomial) | Applied satisfactory regression techniques (linear, multiple, polynomial) | Applied limited or no regression techniques (linear, multiple, polynomial) |
| **Regression Model Validation** | Created appropriate regression model validation graphs and excellent interpretation validation graphs | Created appropriate regression model validation graphs and satisfactory interpretation validation graphs | Created no regression model validation graphs and no interpretation validation graphs |
| **Time Series Analysis** | Excellent time series analysis, applying all the steps in time series analysis | Satisfactory time series analysis, applying some the steps in time series analysis | Limited or no time series analysis, applying non or limitee the steps in time series analysis |