**CP2403: Project – Part 2 – UPDATES Task Sheet**

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  (Completed) | 10% | Module 1 - 4 | Week 6,  Thursday, 29 March 2017  5pm |
| Part 2 | 30% | Module 5 - 10 | Week 13,  Friday, 25 May 2017  5pm |

**Part 2: 30%**

**A) ANOVA - Hypothesis 1**

We want to find out if the mean number of dengue cases (total\_cases) in a week is equal for the cities San Juan and Iquitos.

Ensure you complete, zip and submit both the ‘**CP2403 - Project – Part 2 – ANOVA.docx**’ and ‘**Project-Part2-ANOVA-template.ipynb**’ files.

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

**B) Multiple Regression - Hypothesis 2**

We want to predict the number of dengue cases per week (total\_cases) in each city based on temperature, precipitation, dew point, humidity and vegetation index. Using the variables listed below, use multiple and/or polynomial regression to predict the number of dengue cases for San Juan (sj) and Iquitos (iq). You can use **ALL** the variables listed or select **SOME** variables for the regression analysis. Justify your selection.

Evaluate your regression model using qq plot and standardized residual plots.

**(There are multiple correct answers for this section)**

Ensure you complete, zip and submit both the ‘**CP2403 - Project – Part 2 – Regression.docx**’ and ‘**Project-Part2-Regression-template.ipynb**’ files.

* 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 Module 7, Module 8, Practical 7 and Practical 8 for help on linear, multiple and polynomial regression

**C) 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).

Ensure you complete, zip and submit both the ‘**CP2403 - Project – Part 2 – TS.docx**’ and ‘**Project-Part2-TS-template.ipynb**’ files.

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 2 (30%) Rubric**

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