# CP2403 - Project – Part 2 - ANOVA

First Name: Anh Duc

Last Name: Dang

We want to predict the weekly dengue cases (total\_cases) based on the following explanatory variables provided.

* 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

|  |
| --- |
| **1: Scatter plots between each explanatory variable and response variable** |
|  |
| **2: List all the explanatory variables selected for regression analysis. Justify your selection** |
| List of all the explanatory variables:   * station\_avg\_temp\_c – Average temperature * station\_precip\_mm – Total precipitation * reanalysis\_specific\_humidity\_g\_per\_kg – Mean specific humidity * reanalysis\_precip\_amt\_kg\_per\_m2 – Total precipitation * reanalysis\_avg\_temp\_k – Average air temperature   The chosen explanatory variables combined 3 main feature which affect the number of dengue total\_cases include: temperature, precipitation and humidity. The variables consist of the data from the station and the real analysis so that we can find out better regression between total\_cases and explanatory variables. |
| **3: Regression analysis results** |
|  |
| **4: Regression equation/line** |
| total\_cases= 0.0004 + 6.1157(station\_avg\_temp\_c\_c) + 0.0370(station\_precip\_mm\_c) + 0.0116( reanalysis\_avg\_temp\_k\_c) + 0.1158(reanalysis\_precip\_amt\_kg\_per\_m2\_c) + 1.2960( reanalysis\_specific\_humidity\_g\_per\_kg\_c \*\* 2) |
| **5: qqplot** |
|  |
| **6: Conclusion from qqplot** |
| The qqplot show that the plot have the left skew and a heavy tailed. The skewness greater  than the theoretical distribution. But overrall, the values lie along a line distributrion has the simmilar shape as theoretical distribution that we have supposeds  The qqplot is bad fit |
| **7: percentage of observations over 2 standardized deviation** |
| 4.676753782668501 |
| **8: percentage of observations over 2.5 standardized** |
| 3.43878954607978 |
| **9: Conclusion from observations over 2 std and 2.5 std** |
| * 1% of observation have > 2.5 standardized residual   + (3.4387) * And 5% of observation have > 2.0 standardized residual   + (4.6767) * Poor fit |

**Scenario 5**

Experiment with multiple and polynomial regression between oilperperson(x1), co2emissions

(x2), relectricperperson (x3) and employrate (y)

|  |
| --- |
| **1: Regression Analysis results** |
|  |
| **2: Regression line** |
|  |
| **3: qqplot** |
|  |
| **4: Conclusion from qqplot** |
|  |
| **5: percentage of observations over 2 standardized deviation** |
|  |
| **6: percentage of observations over 2.5 standardized** |
|  |
| **7: Conclusion from observations over 2 std and 2.5 std** |
|  |