Ryerson University

Alberto Luis Rondon Rosario
Student Number: 500922389
CKME 136 - Data Analytics: Capstone Course - Fall 2018

Predicting Number of Dengue Infections Using Historical Climate Data Campinas, Brazil.

I. INTRODUCTION

The Dengue fever is one of the most recognized epidemics that currently affect hundreds of millions of individuals, mostly residing in tropical countries. As there's no immunization treatment developed to date by medicine, science has put its efforts on evaluating factors that could strongly influence the up rise of the level of infections in specific areas. And more importantly, identifying the likeability of epidemic outbreaks that could be potentially prevented by controlling agents in the environment. Trying to fight against this public health threat, governments of a variety of countries have supported this developments and organizations have enforced the proper measurements and distribution of data related to influential meteorological factors.

Hypothetically, Brazil's government is looking to support metropolitan areas like Campinas on identifying vulnerable seasons along with important factors that contribute to the increase of dengue infections in the area.

Two main research questions can be outlined:

- The program primarily aims to organize a calendar of educational / monitoring visits to the community to help prevent the increase on the levels of infection, set emergency alerts and allocate resources to prevent and treat the population affected by the infection

- Another important objective to tackle is to be able to proactively statistically infer the amount of cases that could be reported in the area.

II. LITERATURE REVIEW

In the light of the dramatic fluctuations observed among the variables used for this analysis. It's valid to question if there are any parameters or antecedents identified that must be taken into consideration when approaching the problem. A case worth to mention is the studies carried out in Noumea, France, where seasonal variability has been recorded and analyzed over 40 years[1]. Two variables seem to play an important role, temperature and relative humidity, more specifically, the following parameter can be cited:

[1] Collaboration conducted by the Aix-Marseille University, the Noumea Hospital Centre, the New Caledonia Pasteur Institute, Météo France in New Caledonia, the Department of Social and Sanitary Affairs (DASS) and the Secretariat

“If the temperature exceeds 32°C for more than 12 days in January, February and March – or during the southern summer – and humidity exceeds 95% for less than 12 days in January, a wave of dengue fever occurs. And vice versa: when cooler summer temperatures are combined with wetter weather, an epidemic occurs in almost all cases.”

At the same time, it’s important to mention that one of the main takeaways is the reassurance of the importance of local climate data to evaluate these epidemics. Factors like previous local dengue transmissions and lag times have been proved to have an impact on the spread of the disease. It’s also possible that transmissions could be delayed several months after and specific weather condition that can be classified as favorable. The Aedes Aegypti Mosquito eggs can resist desiccation for an average if 18.3 weeks (4.5 months) which exposes underlying correlation between the variables and the data points used for the analysis. In this case, the variable quarter of the year is validated as a relevant factor that aligns with the fluctuations on Dengue diagnoses.

Methods implemented variate taking in consideration different approaches. SVR machine learning algorithm alongside structural risk minimization to control the generalized error seem to be the most exhaustive methods to accurately track the phenomena, as presented in a robust case study performed to evaluate Dengue outbreak dynamic in China[2].

III. DATASET DICTIONARY

Dataset name
Dengue, Temperatura e Chuvas em Campinas-SP

A. Description
The document contains the number of confirmed monthly cases of dengue in the municipality of Campinas / SP from 1998 to 2014. In addition, data was collected on rainfall, average, minimum and maximum temperature in the city.

For almost every attribute the dataset presents a very low coefficient of variation, except for the main dependent variable, Cases_conf. Taking in consideration the evolution presented, it’s very noticeable that time has played an important role into the spread of this infection and there might be a wide spread of factors that have directly or indirectly favored not only the abrupt increase of cases reported, but also an important growth of the population of mosquitoes that carry this disease – such as the main specie, the Aedes aegipty.

---

of the South Pacific (SPC). Reference: https://en.ird.fr/the-media-centre/scientific-newssheets/410-predicting-outbreaks-of-dengue-fever-according-to-climate

[2] Developing a dengue forecast model using machine learning: A case study in China. https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0005973#sec005

B. Source

This material is publicly available in the Kaggle.com, released by Renan Gomez under CC0 1.0 Universal (CC0 1.0) Public Domain Dedication. The reports of dengue cases were collected from the SES (State Department of Health) and SINAM (Information System of Notification Diseases). The climatic data were obtained from tables found at ciiagro.sp.gov.br. Link: https://bit.ly/2OdeRsh

C. Dataset Dictionary

Date
The month in format YYYY-MM-DD
Attribute type: Date

Cases_conf
Total dengue cases confirmed in the specified month
Attribute type: quantitative – Discrete.
Summary: Min: 0 / Max: 20,428 / Mean:358 / Sd:1,742

Rain
Total amount of rainfall in the city during a specified month. Measurement recorded in mm.
Attribute type: Quantitative - Continuous
Summary: Min: 0 / Max:453 / Mean: 113 / Sd: 93

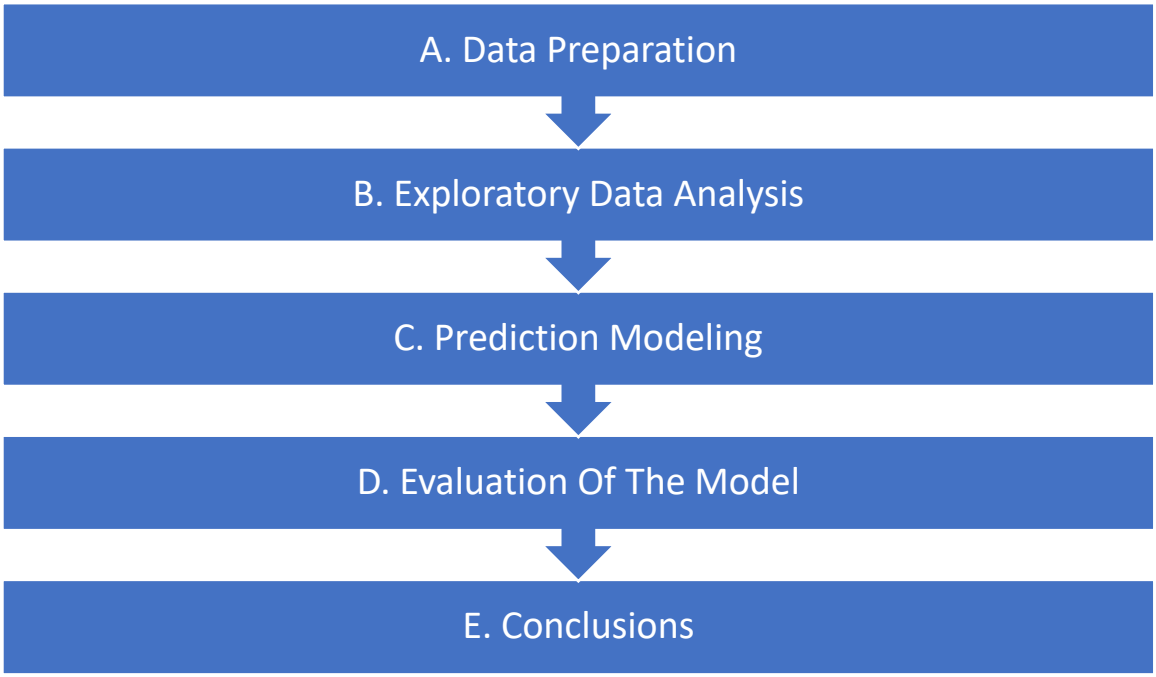
Avg_temp
Average temperature in the city during a specified month (°C)
Attribute type: Quantitative - Continuous
Summary: Min: 16 / Max: 27 / Mean: 22 / Sd: 2

Min_temp
Daily average temperature (°C)
Average temperature in the city during a specified month
Attribute type: Quantitative - Continuous
Summary: Min: 8 / Max: 24 / Mean: 18 / Sd: 3

Max_temp
Maximum temperature in the city during a specified month (°C)
Attribute type: Quantitative - Continuous
Summary: Min:19 / Max: 30 / Mean: 26 / Sd: 2

Quarter
Quarter of the year taking in consideration the month of the register.
Attribute type: Qualitative – Ordinal
Labels: 1 – 2 – 3 – 4

IV. Approach



A. DATA PREPARATION

Integrate Quarter variable to the dataset

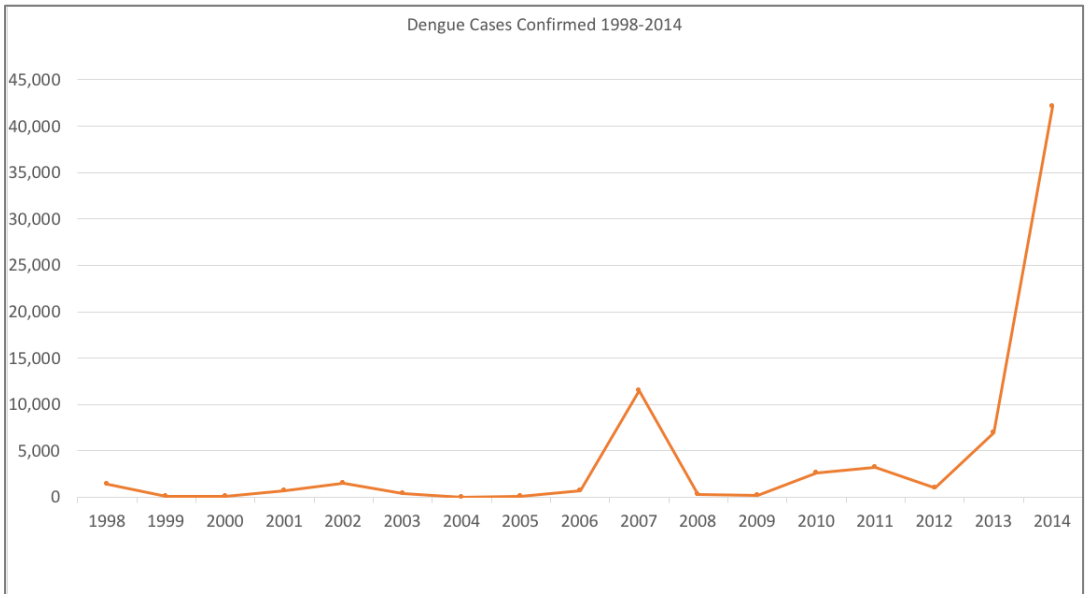
The Quarter variable was populated with corresponding labels for the Date variable.

Resolve approach to several missing values on the Rain variable.

The original sources for the dataset were consulted to evaluate the nature of the aforementioned. After confirmation, these elements actually represent the absences of rain in the area during the registered period. The empty slots where replaced by the value 0.

Evaluate the nature of possible outliers.

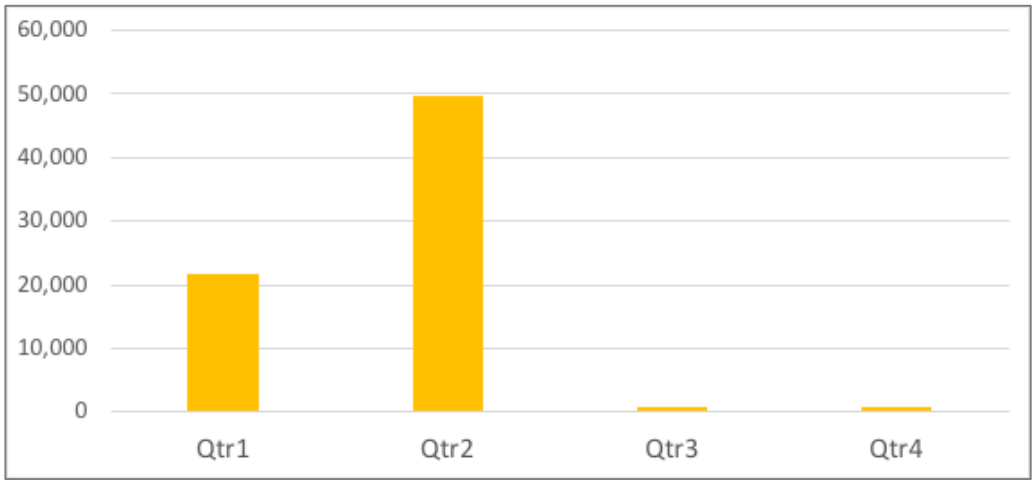
Outliers are registered as legitimate part of the dataset between Q1 and Q2 on 2014. Further steps have been included in the Exploratory Data Analysis to better understand the impact of this set of registers and how to integrate them in the experiment.



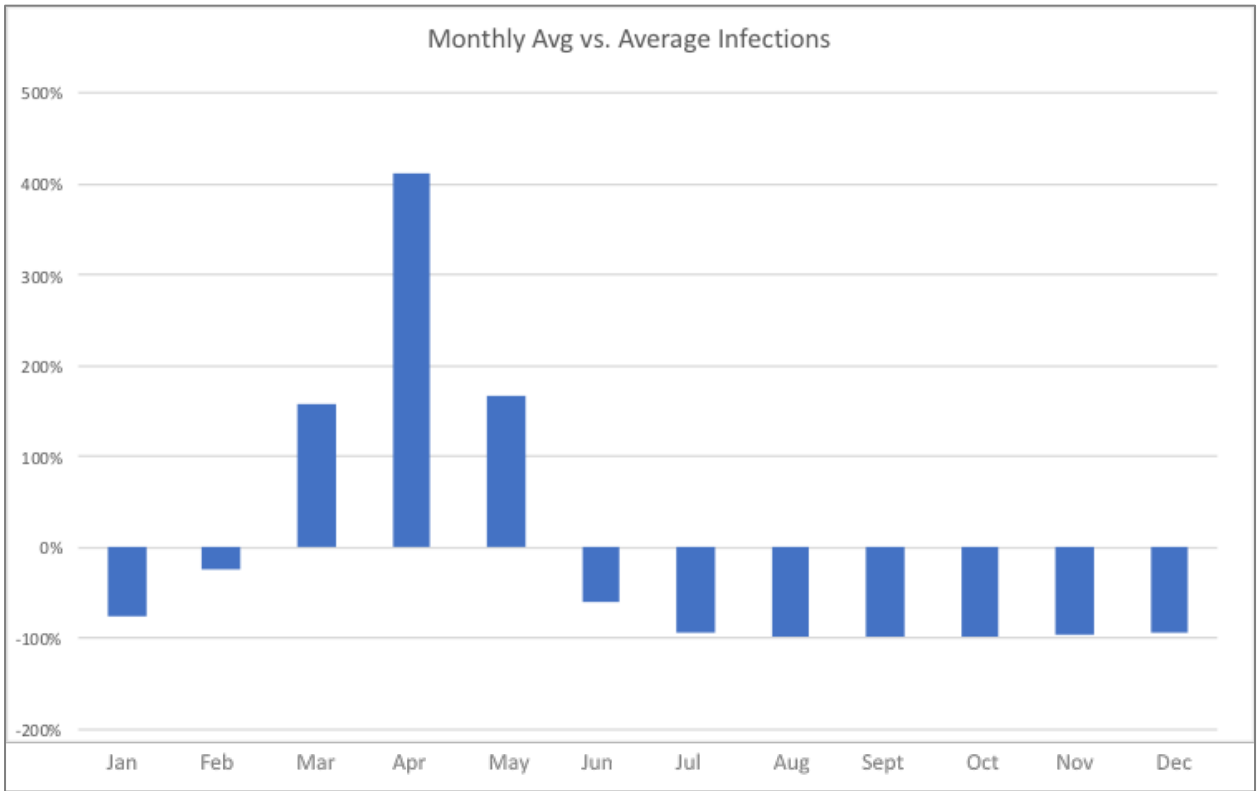
B. EXPLORATORY DATA ANALYSIS

General Exploration

The quarter of the year is likely to be an important attribute for the model, taking in consideration that most infections are likely to occur on Q2 every year.

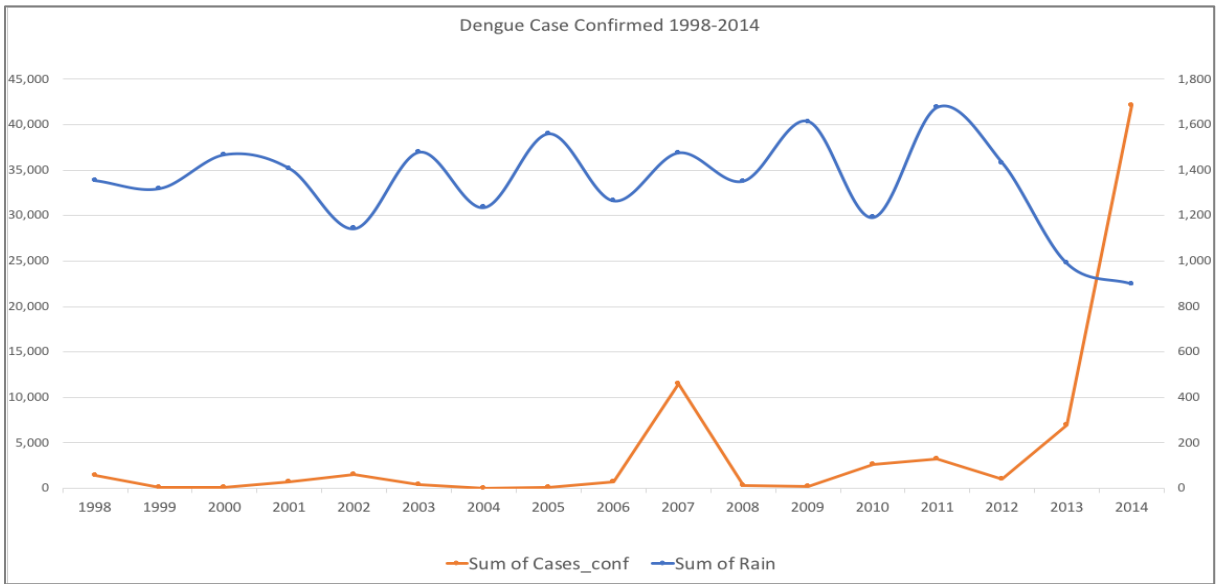


March, April and May are the months that consistently record the highest amount of cases registered. The likeability of being infected is importantly higher on this time of the year.



For 2014 we can observe a dramatic increase of infections for the specified period, which somewhat discards the possibility of classifying these registers as illegitimate outliers and rather invites to further evaluate their impact before modeling.

At a first glance, a prolonged decrease on the levels of rain seem to have an impact on the evolution of the dependent variable. It's important to note that other climatic factors not contemplated on the dataset could have affected these registrations, i.e relative humidity.



Multivariate Analysis
Correlation Matrix:

Evaluation of correlation among the independent variables and between the independent and dependent variables. Given the nonparametric distribution of the attributes in the dataset the Spearman method was implemented. **Tool:** R Spearman correlation analysis.

	Cases_conf	Rain	Avg_temp	Min_temp	Max_temp
Cases_conf	1.00	0.08	0.20	0.22	0.06
Rain	0.08	1.00	0.57	0.55	0.47
Avg_temp	0.20	0.57	1.00	0.89	0.86
Min_temp	0.22	0.55	0.89	1.00	0.68
Max_temp	0.06	0.47	0.86	0.68	1.00

Source Code -
Markdown Document provided or Git Hub"
https://github.com/AlbertoRondon/DataAnalytics_CapstoneCourse_Fall2018

Correlation among the variables:

Rain and temperature hold the most significant correlations present in the dataset. Rain <> Average Temperature (0.57), Rain <> Minimum Temperature (0.55) and Rain <> Maximum Temperature (0.47).

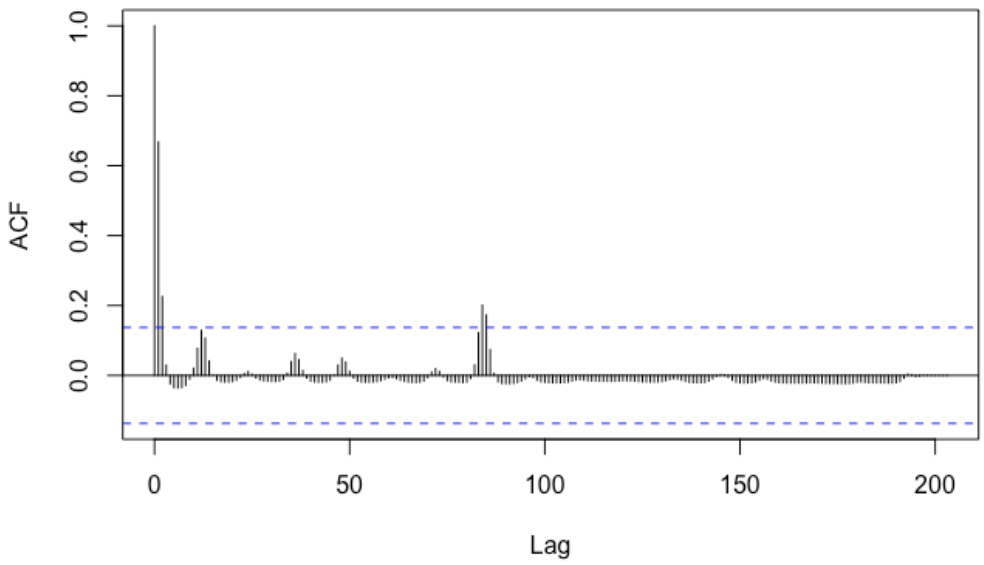
Correlation among independent and dependent variables:

None of the variables evaluated seems to be strongly correlated to the dependent variable Cases_conf. The most relevant values are present for Avg_temp and Min_Temp showing 0.20 and 0.22 results respectively.

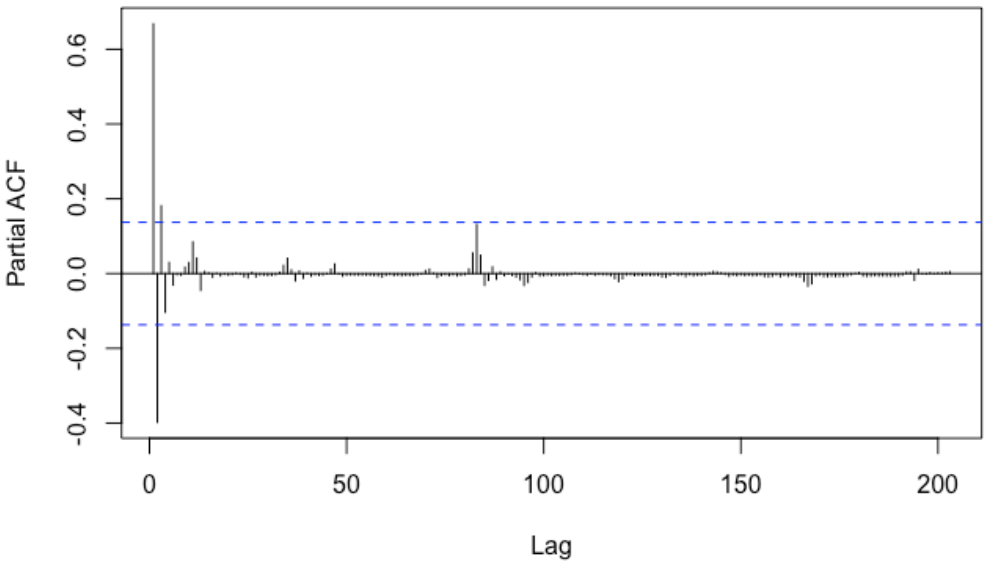
Auto-correlation of the Dependent Variable.

Taking in consideration the stationary nature of the variable Case_conf auto-correlation was evaluated to identify possible lag times on the shifts and the influence of past infections on future ones.

ACF Analysis showed no important pattern on the long term, but there is a presence of a gradual decreasing pattern on the two subsequent months of a particular record, more specifically a correlation of 0.668 and 0.226 respectively.



On the other hand, PACF showed a negative correlation for the first subsequent month and a weak positive one for the second.



This ratifies that the dependent variable could be used as a predictor for a predictive model of higher complexity. Further evaluation - i.e estimation of ideal lag order for a potential ARIMA model - yields a ideal space of 1-2 lags, information that also complies with more exhaustive experimentations cited in the literary review.

Source Code -
Markdown Document provided or Git Hub"
https://github.com/AlbertoRondon/DataAnalytics_CapstoneCourse_Fall2018

Pre-Modeling Conclusions

Outlier treatment for the dependent variable

Taking in consideration that all records in the dataset might hold relevant information and the high variability of the dependent variable, instead of eliminating or replacing the outliers using techniques such as Windsorizing and Trimming, the dependent variable was transformed using the natural log to reduce the influence of the outliers in the regression.

Dimensionality Reduction:

Several iterations were generated to determine the highest level of accuracy using different amounts of variables. The final set of variables included in the model would be disclosed in the Conclusions section.

C. PREDICTIVE MODELING

Model: Multiple Linear Regression in R.

Model was fit for both, the original dataset and the dataset that included a log-transformed set of infections confirmed.

Multiple Linear Regression – Original Dataset

```
Call:
lm(formula = Cases_conf ~ Rain + Min_temp, data = Dengue)

Residuals:
   Min     1Q Median     3Q    Max
  -661   -434   -297   -138  19933

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -284.21     689.14   -0.41     0.68
Rain           -2.12       1.44   -1.48     0.14
Min_temp       49.77      42.01    1.18     0.24

Residual standard error: 1740 on 201 degrees of freedom
Multiple R-squared:  0.0121,    Adjusted R-squared:  0.00225
F-statistic: 1.23 on 2 and 201 DF,  p-value: 0.295
```

Multiple Linear Regression – Log-Transformed Dependent Variable Dataset

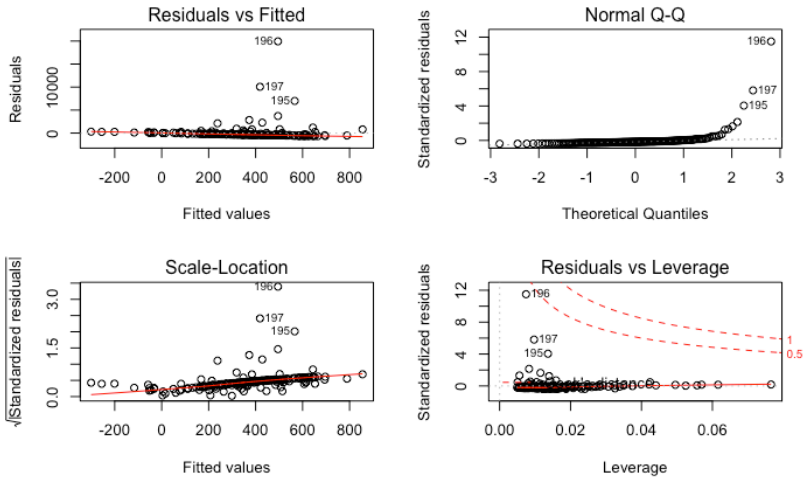
```
Call:
lm(formula = Cases_conf ~ Rain + Min_temp, data = Dengue_log)

Residuals:
   Min     1Q Median     3Q    Max
-3.734 -1.420 -0.368  1.315  6.543

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.72551    0.82091    0.88   0.3779
Rain        -0.00207    0.00171   -1.21   0.2287
Min_temp     0.15214    0.05005    3.04   0.0027 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.07 on 201 degrees of freedom
Multiple R-squared:  0.0444,    Adjusted R-squared:  0.0349
F-statistic: 4.67 on 2 and 201 DF,  p-value: 0.0104
```

Source Code -
Markdown Document provided or Git Hub"
https://github.com/AlbertoRondon/DataAnalytics_CapstoneCourse_Fall2018

D. EVALUATION

Evaluation of Residuals:

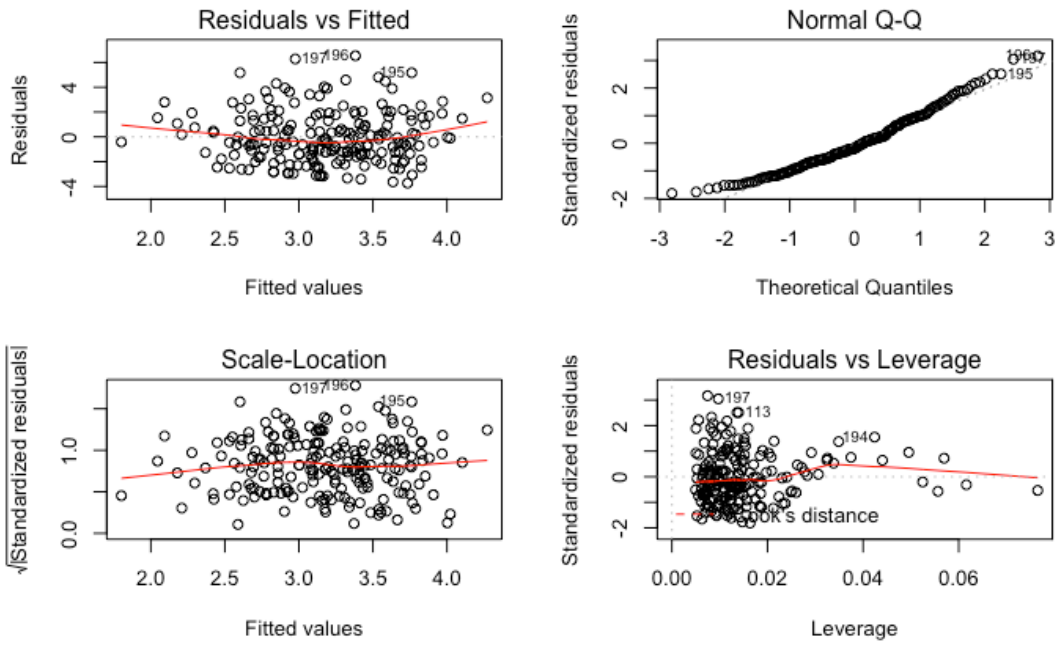
Multiple Linear Regression – Original Dataset



In the evaluation we can observe that there's no distinctive pattern among the residuals, the data points follow the line drawn by the fit, hence the relationship is more likely not be explained by the model. Also, residuals are not equally spread along the predictions, what supposes heteroskedasticity in the variance. We can also observe that among the outliers there are cases that importantly influence the behaviour de model. Basic assumptions have not been satisfied by the model what suggests its rejection.

Source Code -
Markdown Document provided or Git Hub"
https://github.com/AlbertoRondon/DataAnalytics_CapstoneCourse_Fall2018

Multiple Linear Regression – Log-Transformed Dependent Variable Dataset



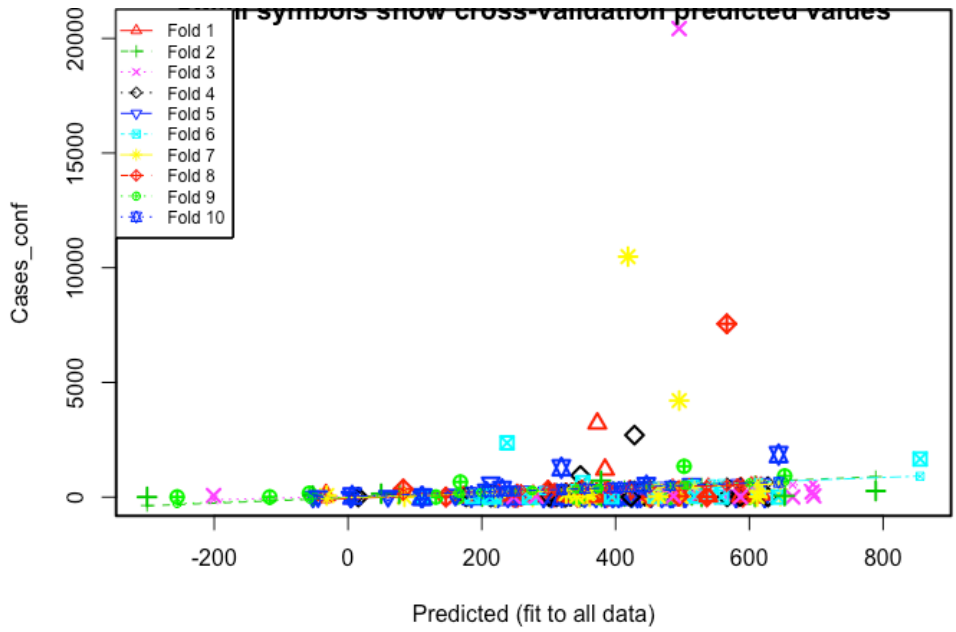
For the Log-transformed variable the residuals are randomly spread, and the fit appears to be able to explain the relationship. Residuals present good homoscedasticity and we can't visualize the Cook's Distance lines, what indicates that no extreme influential values could dramatically affect the results of the presented model. Basic assumptions have been satisfied by the model what and by residual inspections is the favorable model to predict the infections.

Source Code -
Markdown Document provided or Git Hub"
https://github.com/AlbertoRondon/DataAnalytics_CapstoneCourse_Fall2018

Cross Validation

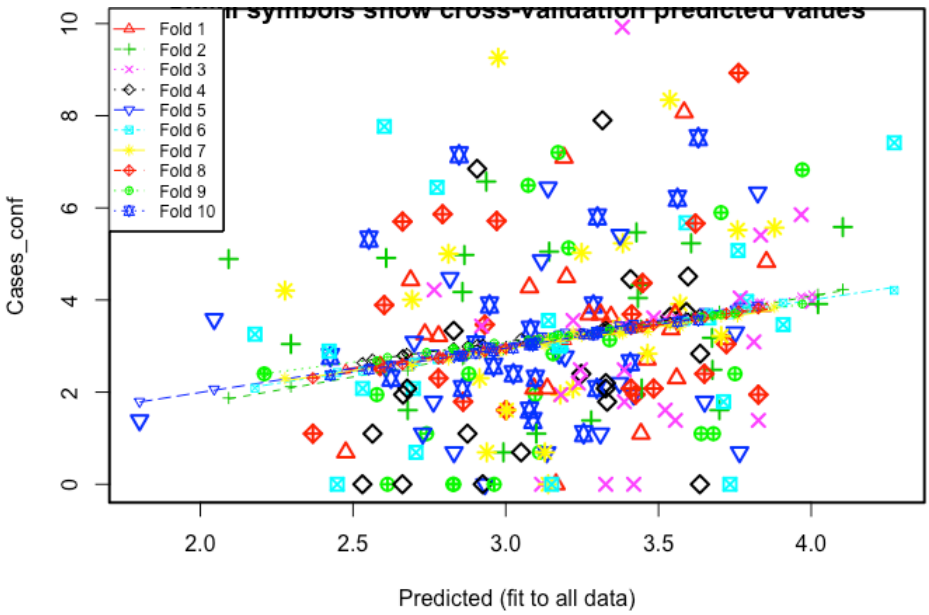
Given the high variability of the predicted variable, 10 folds Cross-Validation have been chosen as the ideal method to assess the results using a randomized original and logged dataset respectively.

Multiple Linear Regression – Original Dataset



Residuals present a similar behaviour as the described in the previous exploration. Cross-validated Standard Error yielded equals 381.9.

Multiple Linear Regression – Logged Dependent Variable Dataset



Residuals present a similar behaviour as the described in the previous exploration. Cross-validated Standard Error yielded equals 0.45.

We can easily compare both standard errors in the same scale calculating the Log of the SE for the original dataset in R: Log (381.91) = 5.95.

Source Code -
Markdown Document provided or Git Hub"
https://github.com/AlbertoRondon/DataAnalytics_CapstoneCourse_Fall2018

Dimensionality Reduction

The model will be fit with the log-transformed dataset taking in consideration that the error is importantly minimized in this experiment and the distribution of the residual follows the favorable assumptions for linear models. To further potentialize the performance observed, few complementary iterations where executed subtracting the variables Maximum Temperature and Average Temperature gradually.

Iteration No. 1: Subtracting Average Temperature: Standard Error: 0.46
Iteration No. 2: Subtracting Maximum Temperature: Standard Error: 0.46

As noted, the elimination of any variable from the model has little to none effect in the calculation of the residuals and the prediction of dependent variable. Parting from this experiment and the Ockham's Razor principle, only Rain and Minimum Temperature would be fit into the model. All the experiments in the document would also reflect this resolution.

Source Code -
Markdown Document provided or Git Hub"
https://github.com/AlbertoRondon/DataAnalytics_CapstoneCourse_Fall2018

E. CONCLUSIONS

- March, April and May are the months that consistently record the highest amount of cases registered. The likeability of being infected is importantly higher on this time of the year.
- Iterations eliminating less correlated variables where executed and yielded little to none improvement in the standard error reported. The final model includes only the Rain and Minimum Temperature variables.
- The best performing model includes a log-transformed variable of the Dengue cases registered, presenting a cross-validated standard error of 0.45.
- The final model must contemplate the SE registered to improve accuracy. Final equation could implement the following format: y= x(a) + z(b) + SE.
- When implementing the model for further prediction is important to treat the outputs to return them to their original scale.
- Relative humidity has proved to be an important factor in experiments held in other geolocations[3]. The local measurement or computation of this variable could highly impact the accuracy of this model.
- Auto-correlation for each variable contemplated is an important aspect to take into account to develop more advanced models that could involve time series analysis.

[3] Developing a dengue forecast model using machine learning: A case study in China.
https://doi.org/10.1371/journal.pntd.0005973