**Data Science Project Protocol**

***Predicting weekly dengue cases based on climate data in Brazil***

*Author: Gali Epstein Shochet*

*25.9.2021*

*BIDS class*

# Introduction

Dengue is a tropical and subtropical mosquito-borne infection that can cause severe illness and death. Dengue is endemic in over 100 countries and approximately two-fifths of the world’s population is currently at risk for dengue fever with an estimated 50 million infections annually [1]. During the past 30 years, dengue fever has dramatically expanded its geographical range and shortened its epidemic cycle in many places. According to the World Health Organization (WHO), the most important vector for dengue, *Aedes aegypti*, which is a predominantly urban mosquito species that favors particular environments such as locations where water storage is common and waste disposal services are inadequate [2].

Dengue is the far most reported disease, as it is also the oldest, and together with the others, it has a cycle that is attributed to the mosquito. In the summers usually, there is an increase in the reported cases, following the idea that the mosquito procreates intensively in this period because of the higher temperatures and more frequent rain. [3,4] Species are affected by both climatic and non-climatic factors. Climatic change can impose physiological constraints on species and therefore can affect species distributions to varying degrees. The relationship between climate and the distribution of a species throughout a landscape varies due to local adaptation and other factors, such as dispersion con­straints related to habitat availability [5].

There are four distinct but antigenically related types or serotypes of DENV (DENV-1 to DENV-4). The outcomes of infection by any of the DENV types range from asymptomatic, subclinical to symptomatic infections [6]. While there has been no definitive association of the distinct DENV types with clinical course of disease, there are reports suggesting that DENV-2 and DENV-3 cause severe disease more frequently than the other serotypes and that DENV-4 causes a milder illness [7].

Brazil’s map is divided according to regions. The area of Minas Gerais (MG), which is presented in the data of this project, is considered the most infected, with > 300,000 annual cases [8][Brazilian Ministry of Health website, accessed at http://portalsaude.saude.gov.br/]. Over the last 30 years, extensive dengue epidemics have occurred in Brazil, characterized by the emergence and re-emergence of different serotypes, a change in the epidemiological profile and an increase in the number of severe and fatal cases. The variations in reports, as well as variations in serotypes and the severity of disease make predictions more difficult, as the complete data is not always available. This is especially true to areas without available free clinics and routine random checks to detect asymptomatic or mild cases [9]. In a recent report, it was shown that the number of cases reported as of week 30 of 2019 (2,029,342) is higher than the annual totals reported during the most recent two years (2017-2018) and, to date, is lower than the annual totals reported during the 2015-2016 epidemic cycle. This current analysis consists of the years 2013-2018, which vary greatly in the number of new cases, a fact that might make predictions more difficult.

With no vaccine or specific treatment to mitigate the natural history of the disease, a tool that can help clinicians detect and provide timely treatment is of utmost importance. In this project, I aimed to predict the number of new weekly cases in the following 3 weeks of the climate data in a specific region in Brazil.

# Methodology (Project design)

## Data

The dataset is composed of data from two Brazilian governmental institutions, the INMET and DATASUS/SINAN, and about 5 cities of a region (Zona da Mata) of the Minas Gerais State. There were two types of data, climate data (5 files), and one large disease-patient data file. Data was obtained from kaggle:

<https://www.kaggle.com/joaopedromedeiros/climatechikungunya-zika-and-dengue>.

For context, climate data were collected using automatic stations that are spread throughout the country. They are unevenly spread, then only a few cities have this kind of data, in this specific context only 5 cities in the area have one. On the other hand, disease data is largely available through DATASUS/SINAN, even though data can be untrustworthy or with lost sequences of time. Currently, there is no available data for external validation. In total, there were 67,000 records of infections with Dengue, Zika, and Chikungunya in the patient file. In addition, there were records from 5 cities with daily climate data (e.g. atmospheric pressure, wind, temperatures and humidity).

The ‘SISTEMA DE INFORMAÇÃO DE AGRAVOS DE NOTIFICAÇÃO 3.0’ (2015) and google translate were used for understanding some of the parameters, as all variable names were in portuguese.

**Data processing and flat file creation in SQL**

The total number of rows in the ‘all patients file’ was 67002. Of them, 65000 were dengue. Therefore, I only focused on these cases. These were filtered out using the ‘WHERE’ statement. Of the date variables, the first, earlier, date is the ‘DT\_SIN\_PRI’, which is most probably the date of the beginning of signals (‘sinais de alarme’) and then there is the ‘DT\_INVEST’, which is the investigation date that has a few NAs and the DT\_NOTIFIC, which is the notification date. The DT\_SIN\_PRI variable was used to extract the year, the month and the week for each record. It was chosen since it had no NA or NULL values.

In the 5 climate data files, the date variable was similarly used. Daily data was aggregated into weekly averages. Max and min values were also recorded for each week (instead of daily max / min).

**Time frame**

Climate data was available from 5 towns in the same region (MG), in five separate files. Measurement date ranges for each town are described in the table below:

|  |  |  |
| --- | --- | --- |
| **Town name** | **Starting date** | **End date** |
| Manhuacu | 29/09/2010 | 1/1/2020 |
| Muriae | 31/08/2006 | 1/1/2020 |
| Vicosa | 16/09/2005 | 1/1/2020 |
| Coronel Pacheco | 20/10/2012 | 1/1/2020 |
| Juiz de fora | 27/05/2007 | 1/1/2020 |

However, the patient file only started on 03/01/2013 and ended at the end of 2018. Therefore, climate data files were only taken from the beginning of 2013 until the end of 24/12/2018. Since it was decided that the outcome will be weekly, and not daily, a cumulative week variable was created (cum\_week). This variable started at 1 and continued until 313. It was later used in order to merge climate data with patient numbers following 3 weeks, using another variable (outcome\_week, see below).

Inclusion criteria

Subjects with a recorded infection with Dengue, with available data on the climate in their region between 01.01.2013 and 24.12.2018.

Exclusion criteria

Cases that were earlier than 2013 or later than the end of 2018.

**Outcome (y)**

There will be one row per week, with the number of new weekly cases as the outcome. The data was aggregated, and patient number (num\_patients) was available for each week. Since there was a gap between the climate and the outcome (takes time for mosquitos to grow), it was decided as 3 weeks. Thus, another variable was created, outcome\_week, which is the cum\_week+3. Then patient data and climate data were joined. (i.e. ON climate\_data.cum\_week = patient\_data.outcome\_week).

**EDA strategy**

Exploratory data analysis (EDA) was performed on the weekly flat file that was generated above using R. Variable distributions, missing data and outliers were examined. Variable distribution was examined using histograms for numeric variables, as well as a check for normal distribution. Mean and median were compared to detect skewed averages using the describe() function. In addition, a relationship analysis was performed using the pairs () function to describe the distribution of each parameter, as well as its distribution vs. the other variables.

Additionally, to find correlations between variables, other than the outcome, a correlation matrix was created. Those that were found to have a strong correlation (above +/-0.5) were individually analyzed by scatterplots, with the addition of a linear trend line using ggplot2.

**Missing data**

Although many missing values were found in the original files (see appendix excel file), most of these variables were not included in the flat file, since they were mostly not relevant and repetitive. Moreover, missing data in the climate files was avoided since variables were aggregated into weekly averages.

In the flat file during the EDA process, missing values were tested by ‘ExploreData’ library - plot\_missing().

**Outlier detection and treatment**

As a first step, a table with variable description was created. The mean and the median were compared to see whether the distribution is not skewed. Then, an outlier data frame was created using a specified function (see EDA notebook) and summarized to see the number or outliers for each parameter. For those above count of 1, a boxplot was created. Using the ‘boxplot.stats(final\_data$wk\_avg\_hum)$out’ command, the actual values and their amount were found. In addition, the week (row) of these values were also retrieved using:

out <- boxplot.stats(final\_data$var)$out

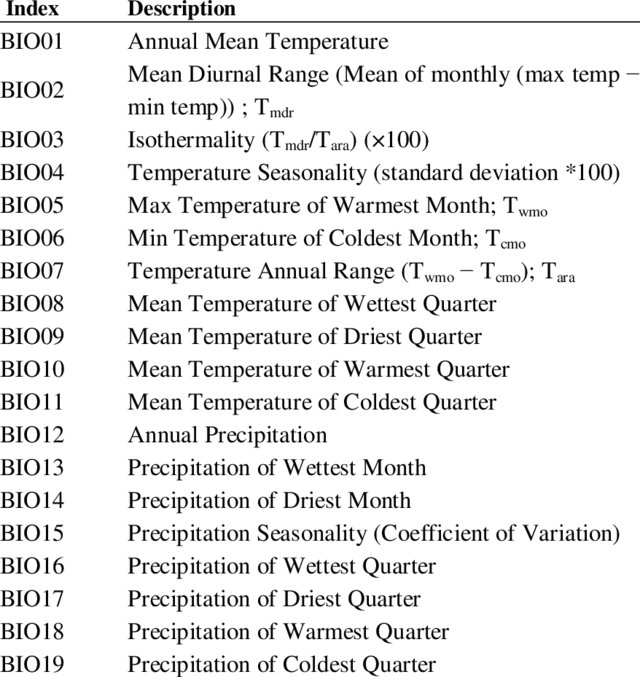
out\_ind <- which(final\_data$var %in% c(out))

Finally, the distribution of the outcome variable was plotted against each variable. Then, a correlation test was done to find the R and the P-value of the connection using spearman analysis. Outliers were treated extremely carefully, since the size of the data was limited. The ‘rules’ for outlier removal were considered, in regard to the effect on both the distribution and the correlation with the outcome variable.

**Data enrichment**

Variables that were found to have a skewed, not normal distribution, were transformed into log variables and to categorical Y/N variables.

Additional parameters were added based on recent publications, as described by Michael S. O’Donnell and Drew A. Ignizio (2012) [5] and others. Calculations can be found in the creating new parameters notebook. The following parameters were calculated and added to the data following the EDA process:



**Feature Selection**

Following data enrichment, a feature selection was done using Python several with selected models, to count whether they significantly affect the outcome. These included LASSO, Random forest, SVM and gradient boosting. Each of the models generated a T/F parameter for each variable, which was additively summed up in a table. All models were summed to test the amount of times each variable was significant in all models.

## Models

**Dataset subsetting**

Following feature selection, data was divided into a test, dev and train subsets.

The ‘mechkar’ package was used, as well as a ‘train-test’ function (see Dataset preprocessing notebook). The Y/N variable was used to estimate balancing of the data, since the tab1 function only accepts categorical values. The dimensions were estimated using the dim function. Dataset distribution was plotted using ggplot.

Data was stratified using oversampling. Although the outcome is numeric (linear), for simplicity the analysis of data balance was done on the Y/N variable, which was found to be equally important.

**Model selection**

Model selection was done using R. A function ‘regressionMetrics’ was used to create a comparative table between models. Then, six models were trained using the ‘train’ data, and compared to the outcome of the ‘dev’ data, as shown in the ‘Models’ notebook. Predictions were plotted against the outcome to visualize the degree of correlation.

The best model was employed on the ‘test’ data to create the final prediction.

## Deployment of the model

This is an experimental model, which is expected to be validated by external, larger, data sources. In my view, this model is only a small part of a larger predictive platform that will include many more parameters, including demographic and medical data, that were not available in this analysis.

QA will be done on the basis of other available published models. The assessment will be done based on past events, that will be compared to the model predictions.

The final user of these predictions are public health officials or doctors that have available climate data. Once validated, such a tool could be used to predict the need for excessive amounts of tests and to make hospital beds more available in the case of a high number prections. This is similar to the COVID-19 situation, in which multiple parameters and multiple teams are combined to predict future events / waves.

Ideally, at the end of the process, it is expected that entering certain climate variables, the region etc. will result in the prediction of the amount of patients in the following 3 weeks.

The end user will be introduced to the app with the available fields to fill out, which will then give an outcome result estimate. This model will be constantly updated with new available data, and validated using medical records from government official sources.

Such a model needs to be updated at least once a month in order to maintain accuracy. Therefore, a designated team is needed for data entry and validation of extreme values. A limit of min-max values should be set for each parameter to avoid entry errors.

In the case of unknown values, the model will gather data from last week as a reference.

# Results

**EDA and data cleansing**

The flat file included 310 rows (weeks), with 21 variables. Of them, two were repetitive following the join process (cum\_week and outcome\_week) and thus were removed early on.

The timeframe was set from the beginning of 2013 through the end of 2018, with a common cumulative week parameter that was used to join the data.

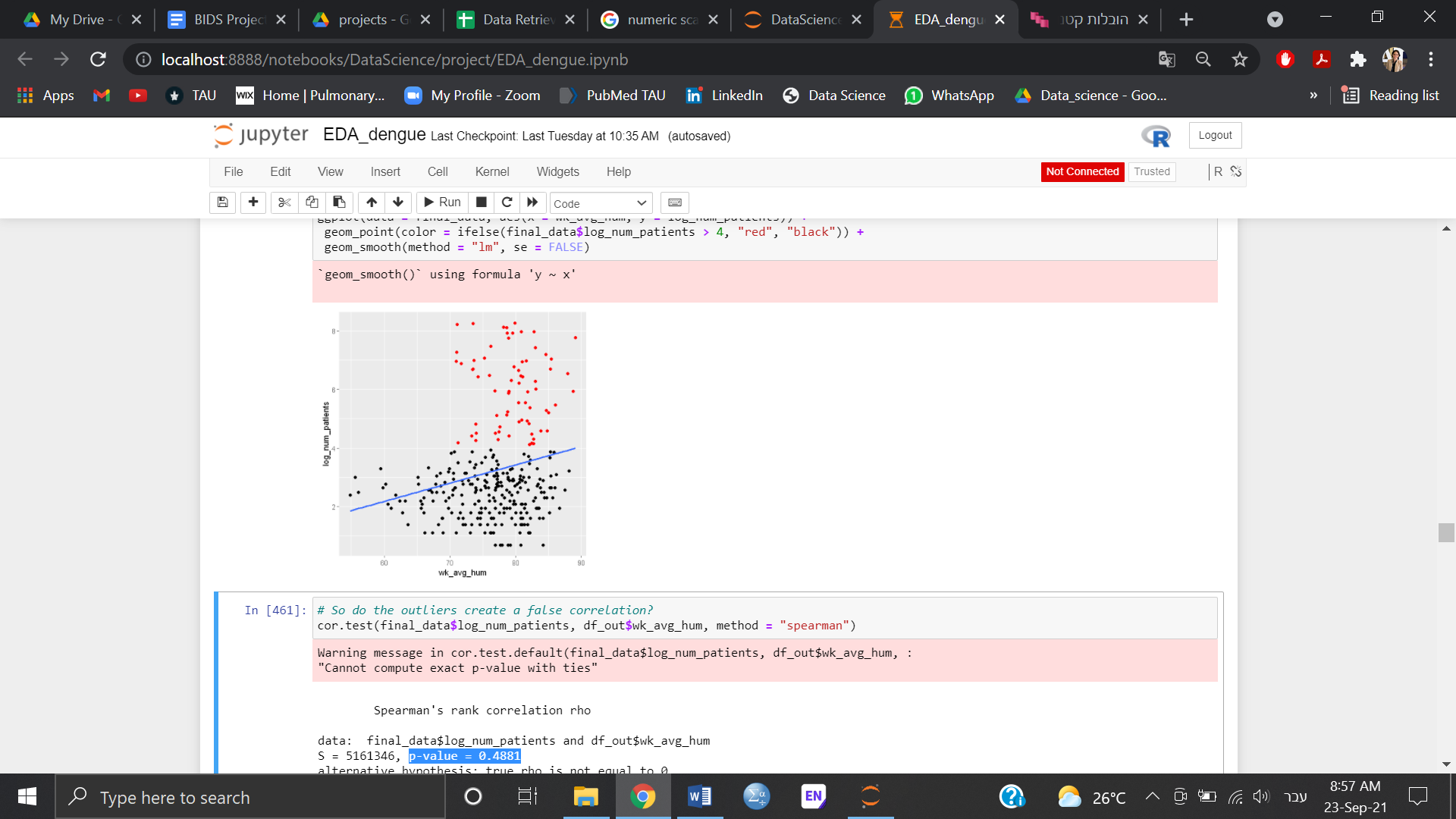
It was found that there were no missing values in the file, as it is a weekly average. No further actions were needed. Outliers were detected as described in the methods section. Since the distribution of the outcome wasn’t normal and more ‘poussin’ like, the variables were plotted against the logarithmic (log\_num\_patients). Similarly, it was found that another parameter was not normally distributed (wk\_avg\_precipitation), which was also transformed. Additionally, these variables were used to create a Y/N parameter, since the log did not result in a normal distribution.

Overall, the following parameters included outliers: wk\_atm\_p, wk\_avg\_hum, avg\_max\_blast\_w, max\_blast\_w and avg\_wind\_speed. For each one, a scatter plot was done, as shown in the figure below, in addition to a boxplot.

Then, a correlation test was done to find the R and the P-value of the connection. Notably, there was a ‘high patient group’ that was discovered, that included patient log counts above 4 (red).

Therefore, even though there was not always a linear correlation, the relation to the patient group (high/ low) was also considered. To keep the connection with maximal values and outcome, outliers were carefully replaced with values close to the max/ min values to maintain the correlation, as described in:

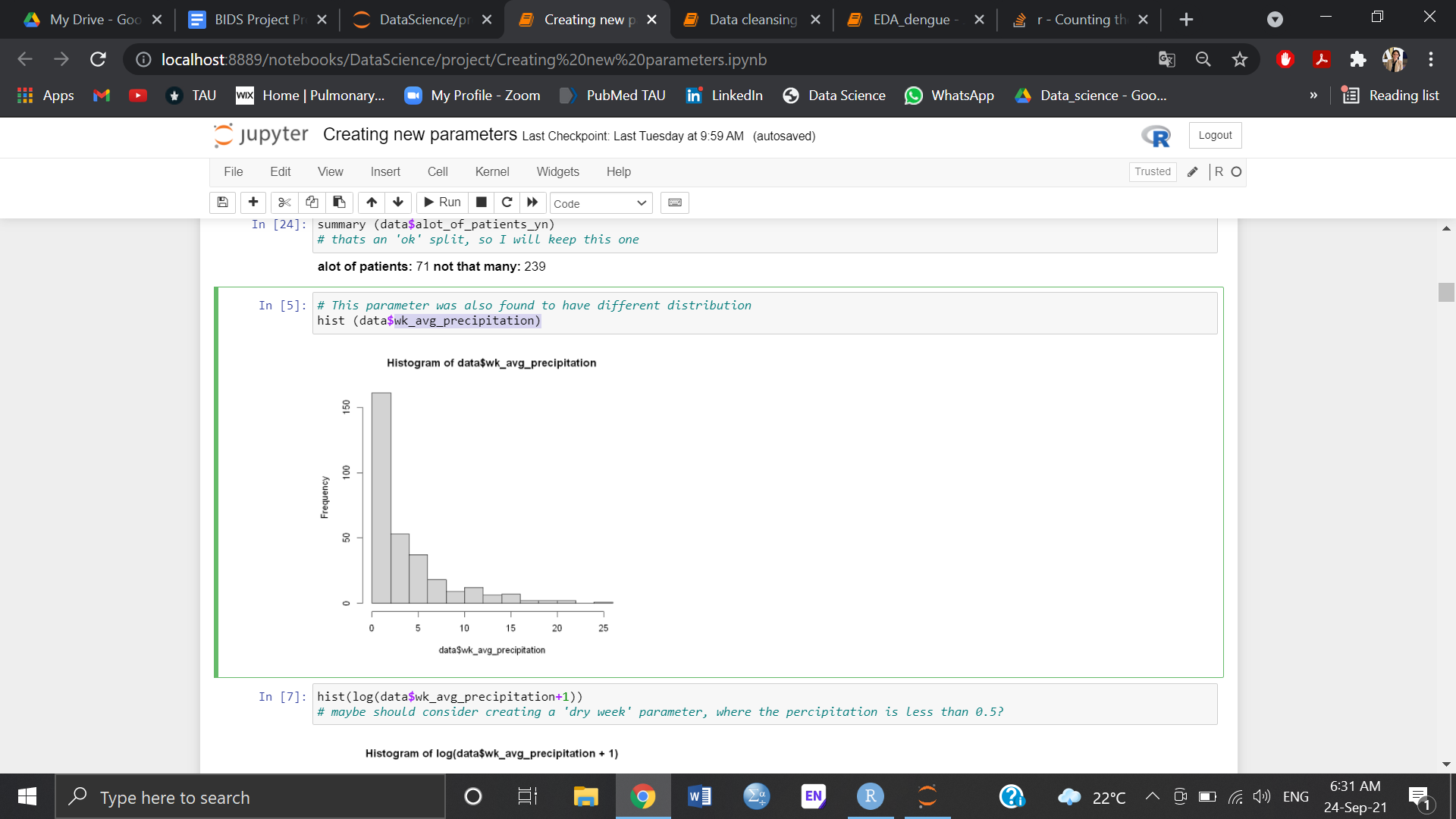
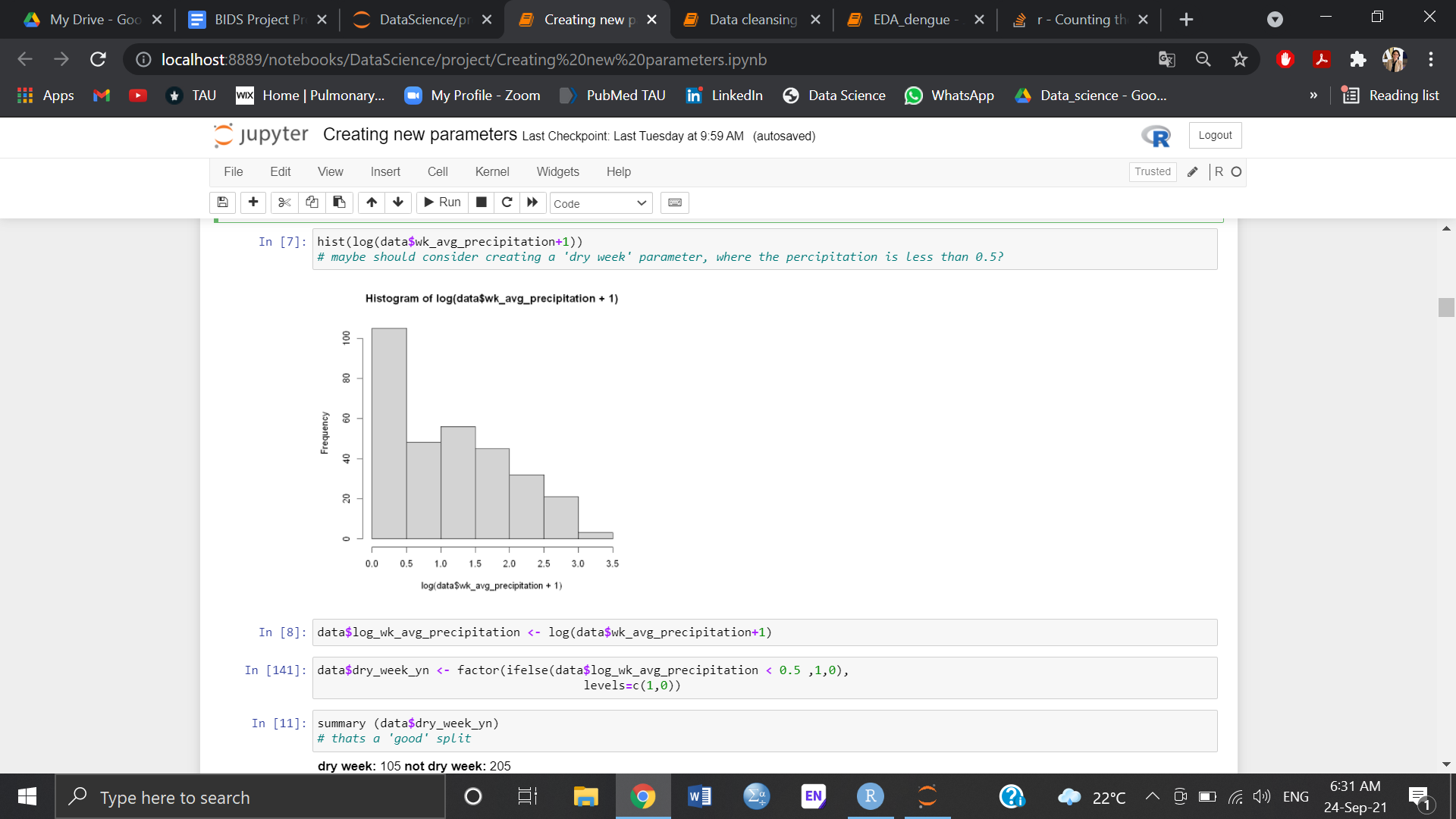
http://qsel.columbia.edu/formhub.R/demo/RemoveOutliers.html. Notably, all outliers were in the ‘low patient group’ (i.e. the black). This could be of importance as extreme conditions may affect mosquito survival.



In addition, for each parameter, the rows that included the outliers were compared in order to detect an extreme week. However, no such week was found, as outliers were randomly distributed. Moreover, the size of the data is relatively small. A fact that encouraged maintaining as much information as possible. No correlation was found between the outliers and their week. Thus, as mentioned above, extreme values were reduced to fit the max or the min values to maintain trends in the data.

**Data enrichment**

Nineteen additional parameters were calculated based on previous publications as described in the methods section (see creating new parameters notebook). Moreover, since the EDA showed that the outcome variable wasn’t normally distributed, an additional log patient variable was added, as well as a Y/N variable for the high patient group, which included the log\_num\_patients above 4 for table 1 creation (alot\_of\_patients\_yn). The wk\_avg\_precipitation parameter was transformed into a log variable. However, this process did not result in a normal distribution. Therefore, a categorical parameter of ‘dry\_week\_yn’ was created, for log\_wk\_avg\_precipitation < 0.5. This resulted in a relatively even distribution of dry week= 105 and not dry week = 205. A t-test showed that this variable is significantly (p= 1.171e-07) affecting the log outcome. It was also found to significantly affect the Y/N variable of the outcome (p = 0.00034, chi-square). In total, the final data included 46 parameters.

 → 

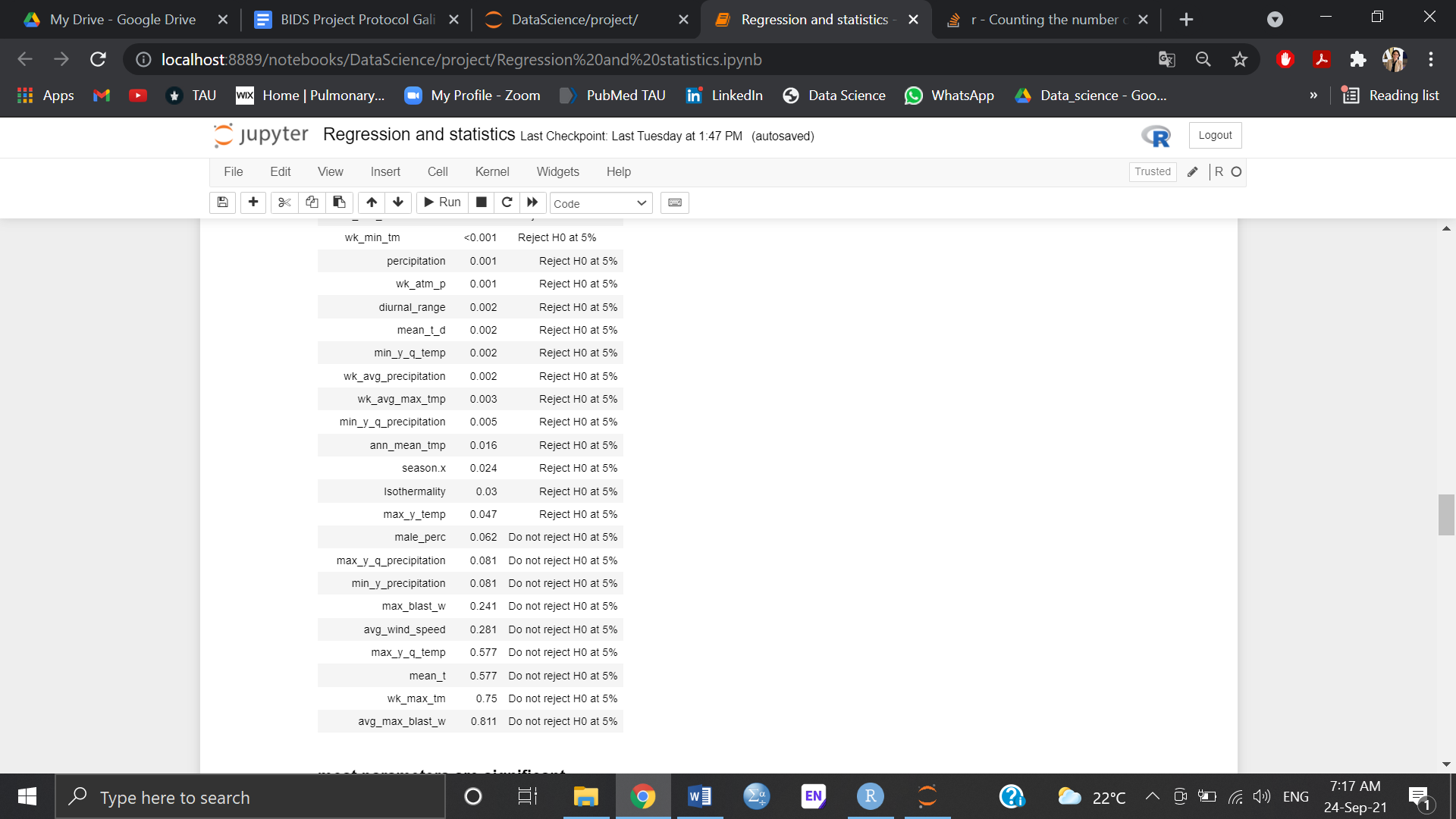
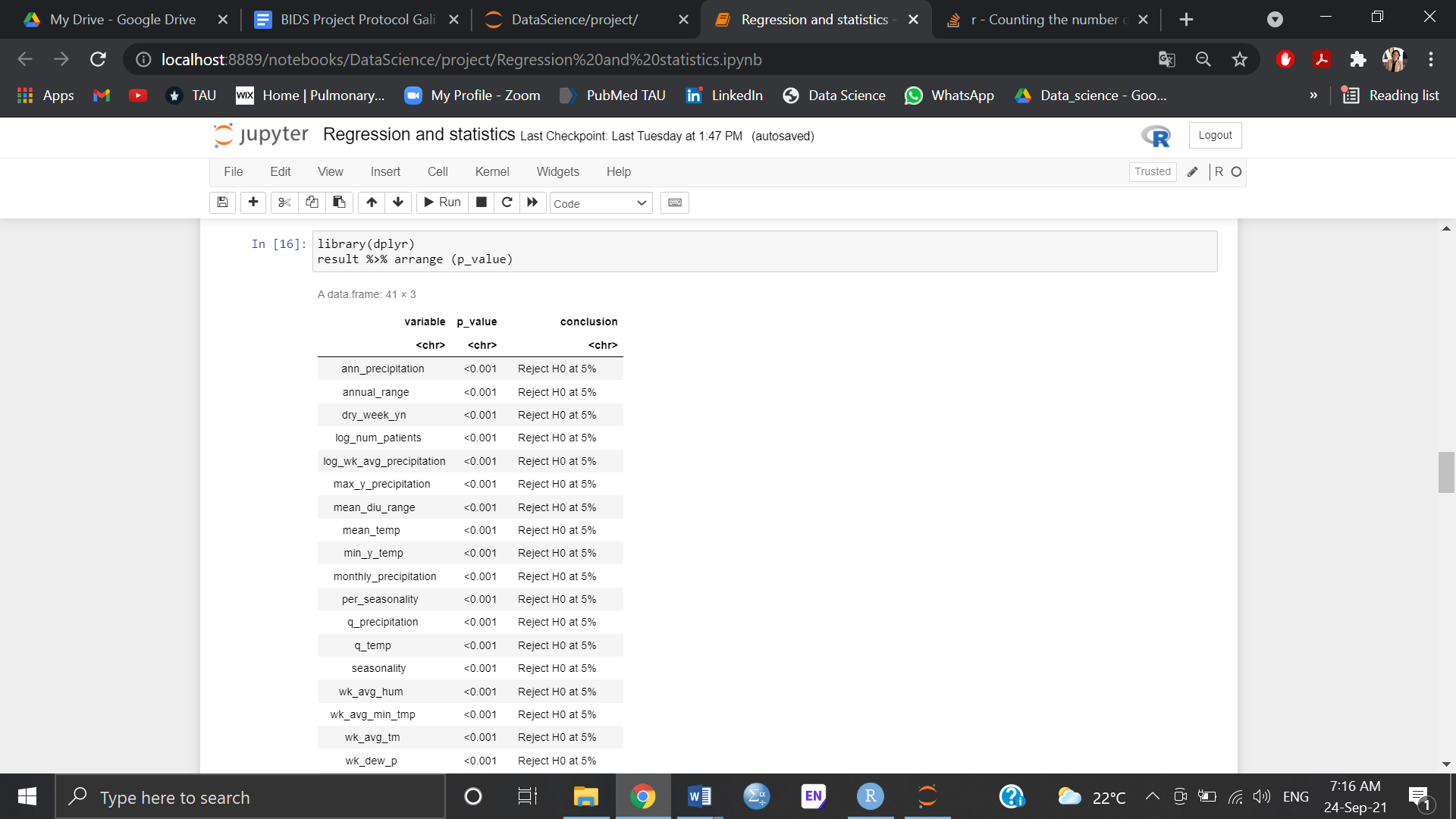
**Feature selection, Regressions and Statistics**

A feature selection test was performed and presented in the table below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Lasso** | **RandomForest** | **GradientBoost** | **SVM** | **Sum** |
| wk\_atm\_p | 1 | 0 | 0 | 1 | 2 |
| wk\_dew\_p | 0 | 1 | 0 | 0 | 1 |
| wk\_avg\_max\_tmp | 0 | 0 | 0 | 0 | 0 |
| wk\_max\_tm | 0 | 0 | 0 | 0 | 0 |
| wk\_avg\_tm | 0 | 0 | 0 | 0 | 0 |
| wk\_min\_tm | 0 | 0 | 0 | 0 | 0 |
| wk\_avg\_min\_tmp | 0 | 0 | 0 | 0 | 0 |
| wk\_avg\_hum | 1 | 0 | 0 | 0 | 1 |
| wk\_min\_hum | 1 | 0 | 0 | 0 | 1 |
| avg\_max\_blast\_w | 0 | 0 | 0 | 0 | 0 |
| max\_blast\_w | 0 | 0 | 0 | 0 | 0 |
| avg\_wind\_speed | 0 | 0 | 0 | 0 | 0 |
| wk\_avg\_precipitation | 0 | 0 | 0 | 0 | 0 |
| num\_patients | 0 | 1 | 1 | 1 | 3 |
| male\_perc | 0 | 0 | 0 | 0 | 0 |
| log\_num\_patients | 1 | 1 | 1 | 0 | 3 |
| log\_wk\_avg\_precipitation | 0 | 0 | 0 | 0 | 0 |
| dry\_week\_yn | 0 | 0 | 0 | 0 | 0 |
| alot\_of\_patients\_yn | 1 | 1 | 1 | 0 | 3 |
| season.x | 0 | 0 | 0 | 0 | 0 |
| ann\_mean\_tmp | 0 | 0 | 0 | 0 | 0 |
| diurnal\_range | 0 | 0 | 0 | 0 | 0 |
| mean\_diu\_range | 0 | 0 | 0 | 0 | 0 |
| max\_y\_temp | 0 | 0 | 0 | 0 | 0 |
| min\_y\_temp | 0 | 0 | 0 | 0 | 0 |
| annual\_range | 0 | 0 | 0 | 0 | 0 |
| Isothermality | 0 | 0 | 0 | 0 | 0 |
| mean\_temp | 0 | 0 | 0 | 0 | 0 |
| ann\_precipitation | 0 | 0 | 0 | 0 | 0 |
| monthly\_precipitation | 1 | 0 | 0 | 0 | 1 |
| min\_y\_precipitation | 0 | 0 | 0 | 0 | 0 |
| q\_precipitation | 0 | 1 | 0 | 0 | 1 |
| min\_y\_q\_precipitation | 0 | 0 | 0 | 0 | 0 |
| max\_y\_q\_precipitation | 0 | 0 | 0 | 0 | 0 |
| q\_temp | 0 | 1 | 0 | 0 | 1 |
| mean\_t | 0 | 0 | 0 | 0 | 0 |
| mean\_t\_d | 0 | 0 | 0 | 0 | 0 |
| percipitation | 1 | 0 | 0 | 1 | 2 |
| max\_y\_q\_temp | 0 | 0 | 0 | 0 | 0 |
| min\_y\_q\_temp | 0 | 0 | 0 | 0 | 0 |
| max\_y\_precipitation | 0 | 0 | 0 | 0 | 0 |
| seasonality | 0 | 0 | 0 | 0 | 0 |
| per\_seasonality | 0 | 0 | 0 | 0 | 0 |

Unfortunately, as shown in the above table, all multivariate analyses resulted in a low number of significant parameters. In an attempt to proceed with the right variables, a ‘table 1’ was created, using a Y/N variable (alot\_of\_patients\_yn). Since it was found that the ‘alot\_of\_patients\_yn’ is a relatively large group, 71 / 239, a univariate table was created in order to understand general trends in the data. Additionally, a linear model was done for the log parameter of the outcome. In that analysis, only the ann\_mean\_tmp, wk\_min\_hum, ann\_precipitation, min\_y\_precipitation and q\_precipitation were significant.

Univariate analysis using the y/n patient variable:



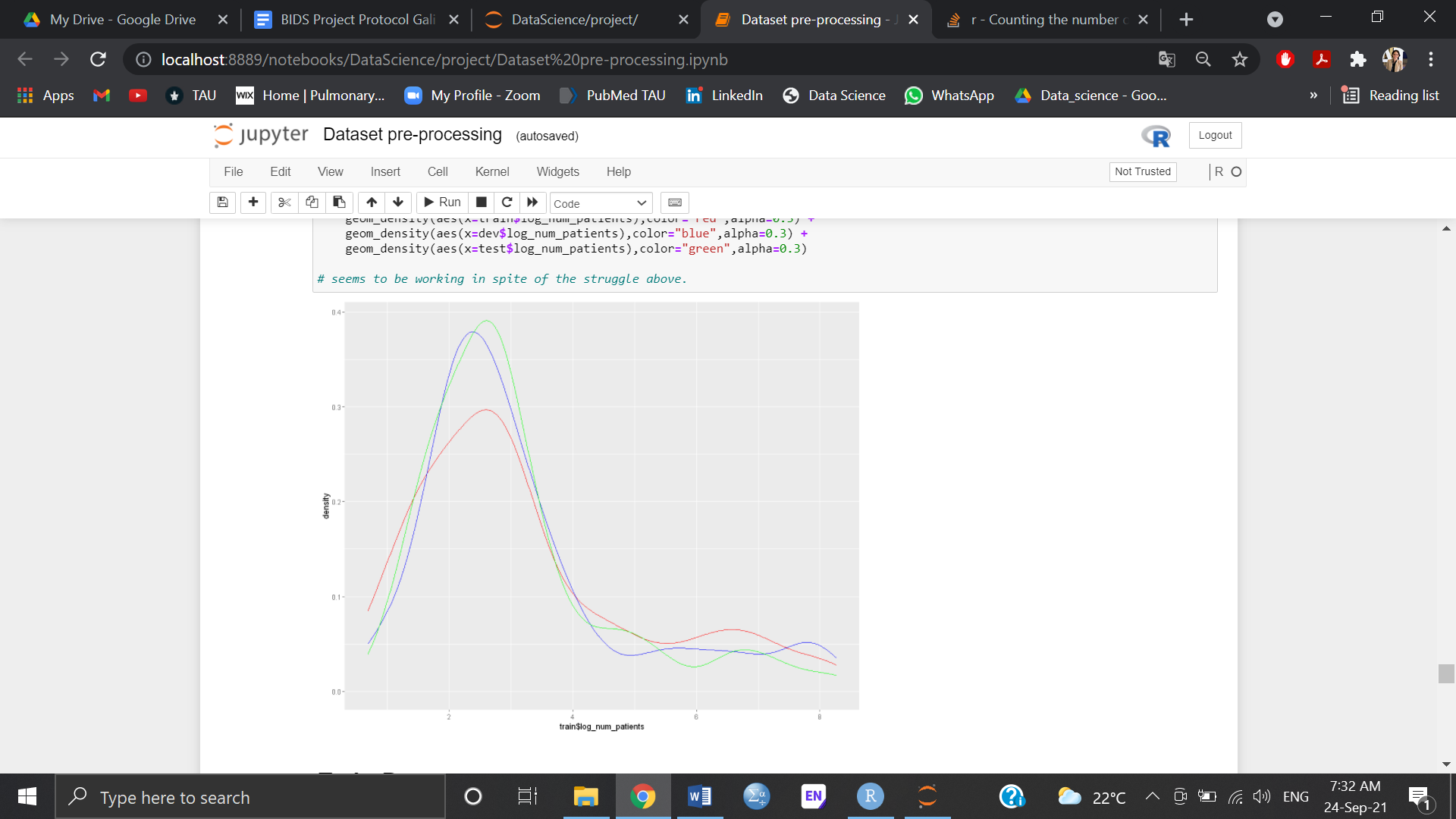
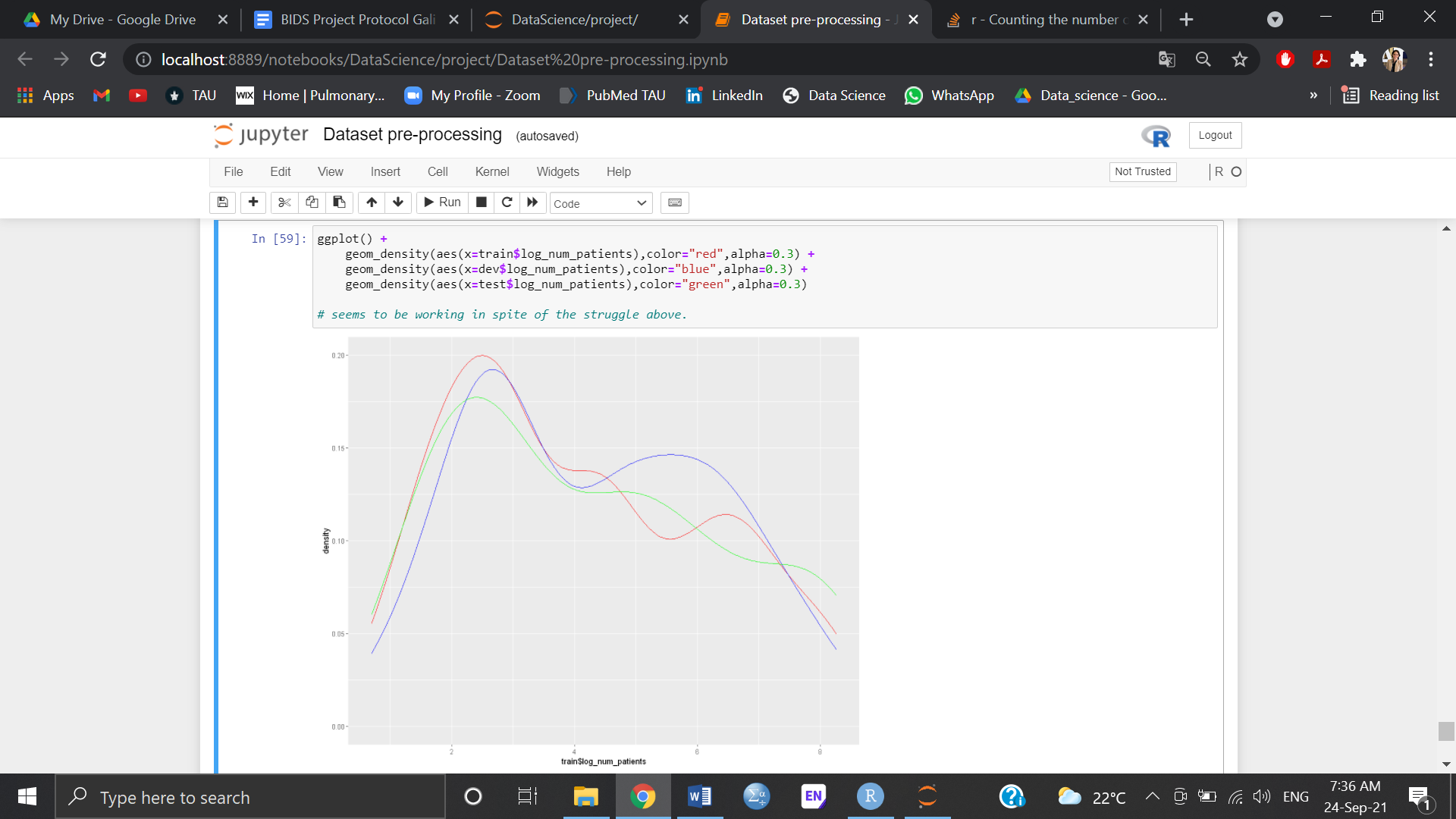
Following all analyses, it was decided to drop:

Male\_perc, max\_y\_q\_precipitation, min\_y\_precipitation, max\_blast\_w, avg\_wind\_speed, max\_y\_q\_temp, mean\_t, wk\_max\_tm, avg\_max\_blast\_w and diurnal\_range since they were repetitive and did not appear in any analysis and were not significant in the univariate analysis.

The year, month and season were also removed, since they correlated too strongly with other parameters.

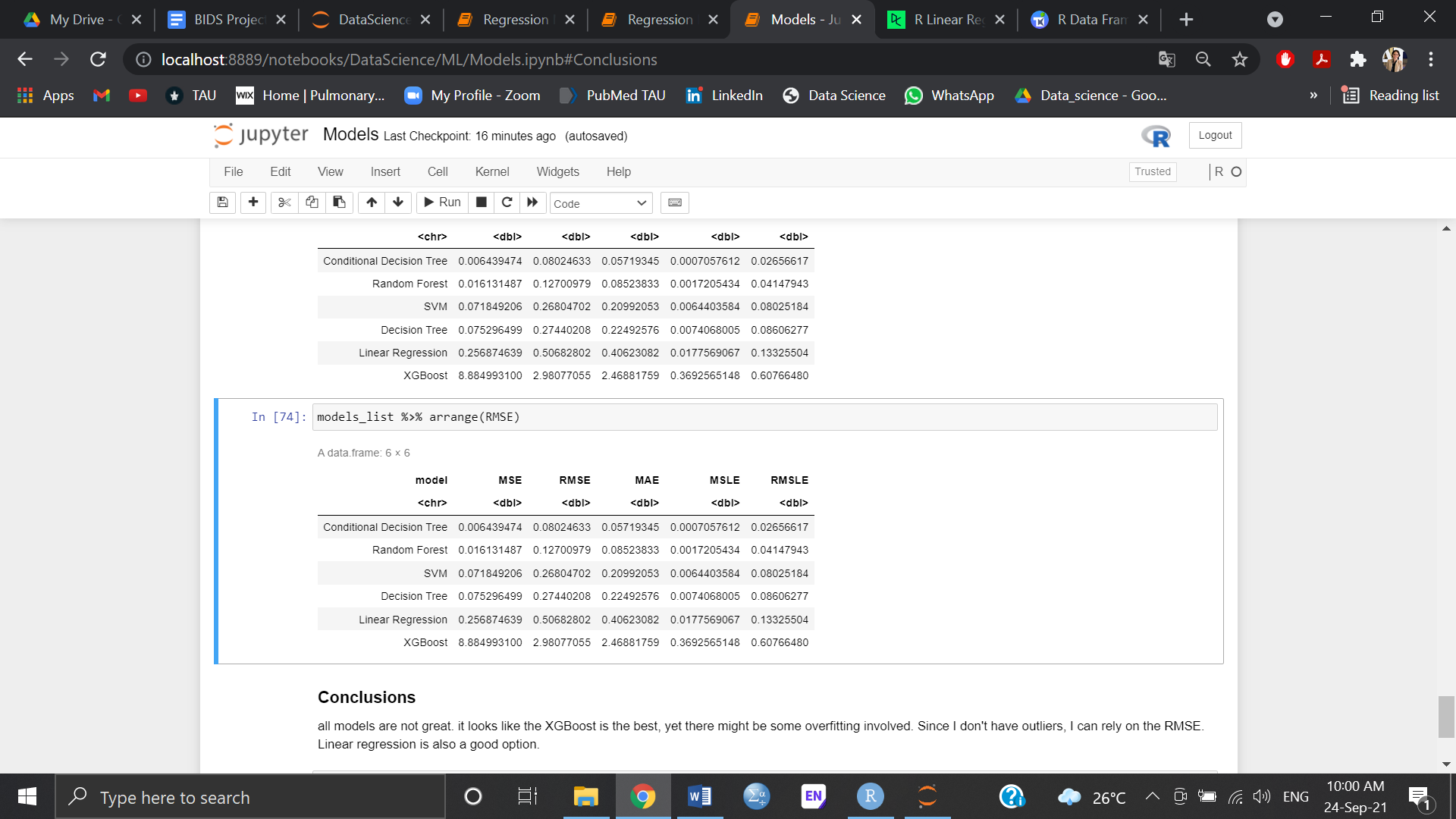
**Data subsetting**

Since the data is relatively small, 310 rows and 33 columns, caution was needed for dividing the data. The Y/N variable was used to estimate balancing of the data. Following the first division for test and temp, it was found that the test included 32 vs. 32 (50% Y), while the temp was 62 vs. 248 (less than 30% Y). Then, another division was similarly performed, that included 32 vs. 32 (50% Y) for dev and 50 vs. 198 in the train. Although the numbers are relatively low, data distribution was similar in all datasets (see left panel below). However, the ‘high patient group’, which is the main (interesting) outcome, was under-represented (see right side of the left panel below). Thus a balance was done using over-sampling, in order to maintain an adequate number of ‘alot\_of\_patients\_yn’ in the data. As shown in the right panel below, the data remained similar between the subsets, yet the outcome variable was properly distributed, as the high proportion of patients is expected to account for about a third of the patients.

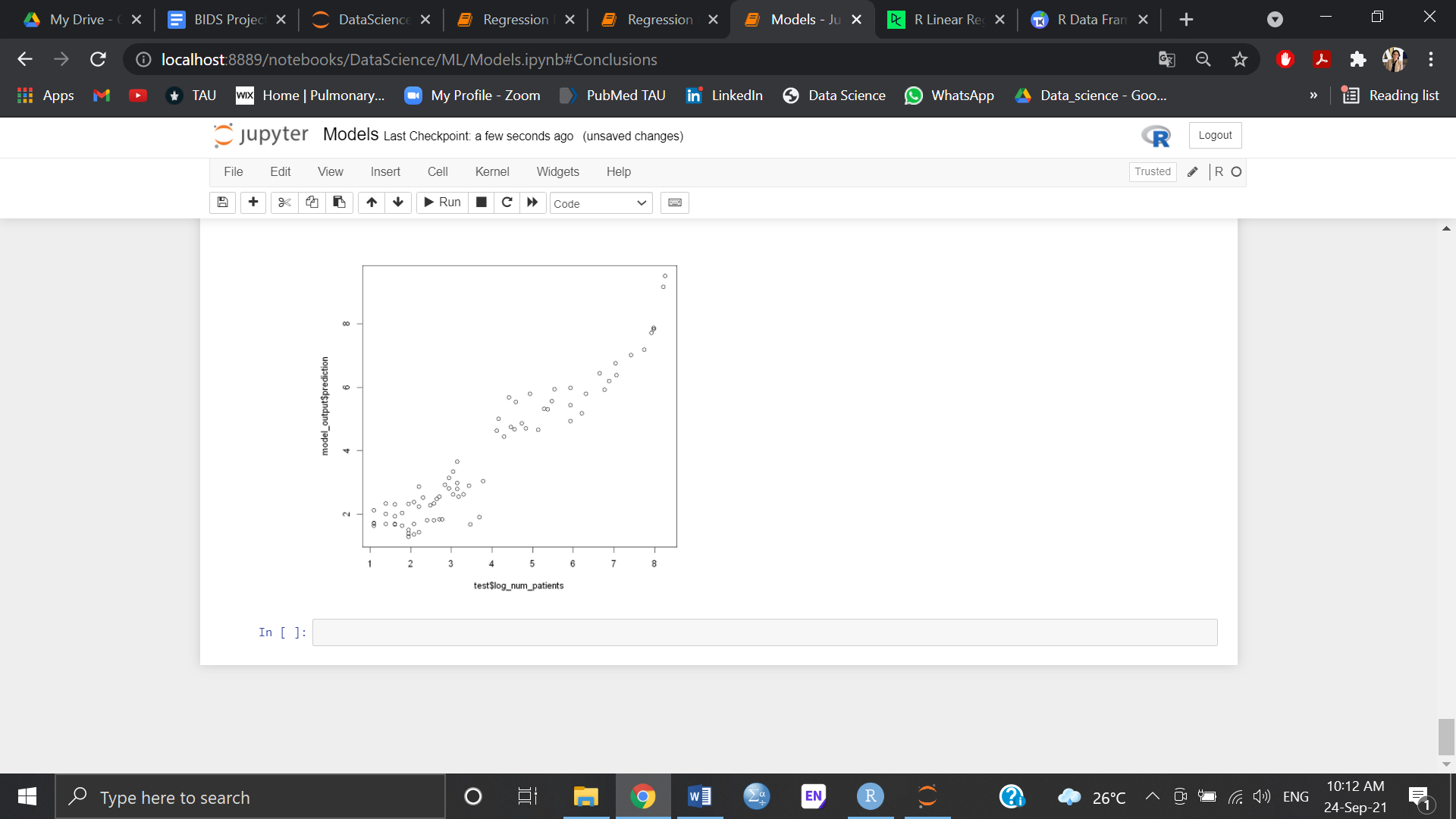
 

**Model selection**

A total of six supervised models were employed on the train data. Then the dev data was used to estimate model accuracy. Metrics were sorted by RMSE, as well as MAE, which showed a similar order. The metrics of these models are shown below:



The linear model was chosen based on the metrics above. XGBoost was excluded due to a high possibility of overfitting, as was seen in the plot. Finally, the Test data was modeled using the lm, which resulted in the following prediction:



# 

# Discussion and Conclusion

In this project, I aimed to create a model to predict the future weekly number of patients, based on climate data that was available for a 5 year period in a specific region in Brazil. First, all data was aggregated to weekly stats. The number of patients for each week was added, with a lag of 3 weeks as the outcome. This part, which is not shown in the notebooks, was actually a large portion of the work.

During the EDA process, it was found that there is an ‘alot\_of\_patients’ group, consisting of numbers of patients above log 4. This group was also used during the analysis.

Additionally, some outliers were discovered, and were handled consciously, without additions of NA variables, in order to maintain trends in the data, which correlated significantly with the alot\_of\_patients. This was also done due to the relatively small data available.

Since data was so limited, additional parameters were calculated based on literature review, which resulted in doubling the amount of variables, including some seasonality variables.

Throughout the analysis, there was the dilemma whether to maintain the linear (numeric) analysis, using the log\_patient variable, or to go further with the cluster/ classification models approach with the Y/N parameter. I chose the first option. The models did not result in great predictions. This was already evident in the feature selection process that showed that most variables, although significant in the univariate analysis, were not significant in the multivariate analysis. In the following step, data was divided into 3 parts, using established techniques, and balanced by addition of measurements to make sure the outcome is balanced. Finally, the test data was used for the prediction of the outcome. A prediction was generated, which actually reflected the two groups of patient numbers. I believe that given the limited data in this assay, the prediction is relatively accurate.

Study limitations included a very small dataset. Although the original files consisted of 65000 rows in the patient file and multiple years in the climate data, aggregations and data cleansing in sql during the flat file generation resulted in a small dataset. The upside of this process is that it resulted in no missing variables, as they were all excluded in the process of aggregation and time frame selection. Another limitation was that this was not a time series model, which did not take into account past events of the previous week (i.e. internal correlation). Adding this feature could improve predictions, as there are ‘waves’ of infection.

In conclusion, this model could be carefully used for the specific area of the ‘Minas Gerais’. More parameters, other than climate (e.g. waste disposal areas, sanitary conditions), are needed in order to improve this model.

**References**

1. San Martin, J.L.; Brathwaite-Dick, O. [Integrated strategy for dengue prevention and control in the Region of the Americas]. *Rev Panam Salud Publica* **2007**, *21*, 55-63, doi:10.1590/s1020-49892007000100011.

2. In *Dengue: Guidelines for Diagnosis, Treatment, Prevention and Control: New Edition*; WHO Guidelines Approved by the Guidelines Review Committee; Geneva, 2009.

3. Nunes, P.C.G.; Daumas, R.P.; Sanchez-Arcila, J.C.; Nogueira, R.M.R.; Horta, M.A.P.; Dos Santos, F.B. 30 years of fatal dengue cases in Brazil: a review. *BMC Public Health* **2019**, *19*, 329, doi:10.1186/s12889-019-6641-4.

4. Cardoso-Leite, R.; Vilarinho, A.C.; Novaes, M.C.; Tonetto, A.F.; Vilardi, G.C.; Guillermo-Ferreira, R. Recent and future environmental suitability to dengue fever in Brazil using species distribution model. *Trans R Soc Trop Med Hyg* **2014**, *108*, 99-104, doi:10.1093/trstmh/trt115.

5. O’Donnell, M.S.; Ignizio, D.A. Bioclimatic Predictors for Supporting Ecological Applications in the Conterminous United States. *USGS* **2012**.

6. Halstead, S.B. Pathogenesis of dengue: challenges to molecular biology. *Science* **1988**, *239*, 476-481, doi:10.1126/science.3277268.

7. Balmaseda, A.; Hammond, S.N.; Perez, L.; Tellez, Y.; Saborio, S.I.; Mercado, J.C.; Cuadra, R.; Rocha, J.; Perez, M.A.; Silva, S.; et al. Serotype-specific differences in clinical manifestations of dengue. *Am J Trop Med Hyg* **2006**, *74*, 449-456.

8. Fares, R.C.; Souza, K.P.; Anez, G.; Rios, M. Epidemiological Scenario of Dengue in Brazil. *Biomed Res Int* **2015**, *2015*, 321873, doi:10.1155/2015/321873.

9. Brathwaite Dick, O.; San Martin, J.L.; Montoya, R.H.; del Diego, J.; Zambrano, B.; Dayan, G.H. The history of dengue outbreaks in the Americas. *Am J Trop Med Hyg* **2012**, *87*, 584-593, doi:10.4269/ajtmh.2012.11-0770.