Task 5 for Machine Learning

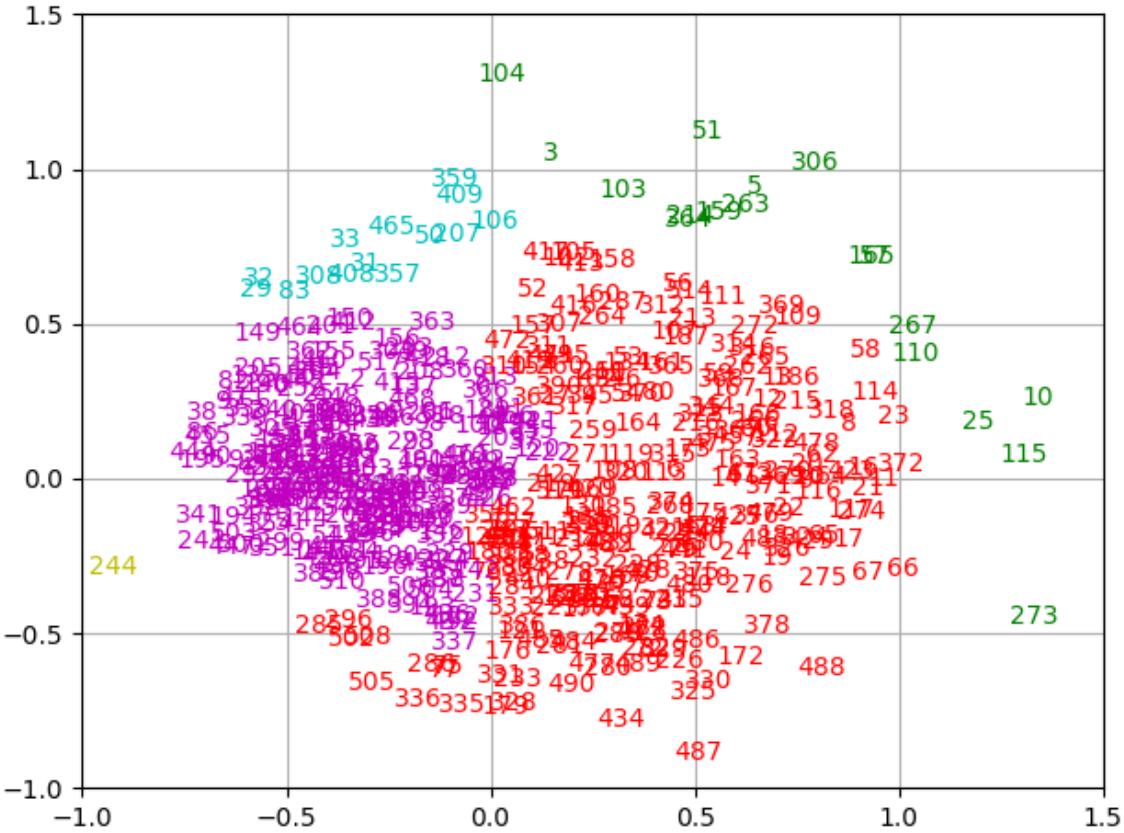
Eduardo Sánchez López
José Alejandro Libreros

Outliers management

For this task first we had to delete the possible outliers of each dataset so the feature selection is as clear as possible. For that we choosed to do hierarchical clustering and delete the elements that were more distant. We repeated this process as many times until we considered there weren't more outliers.

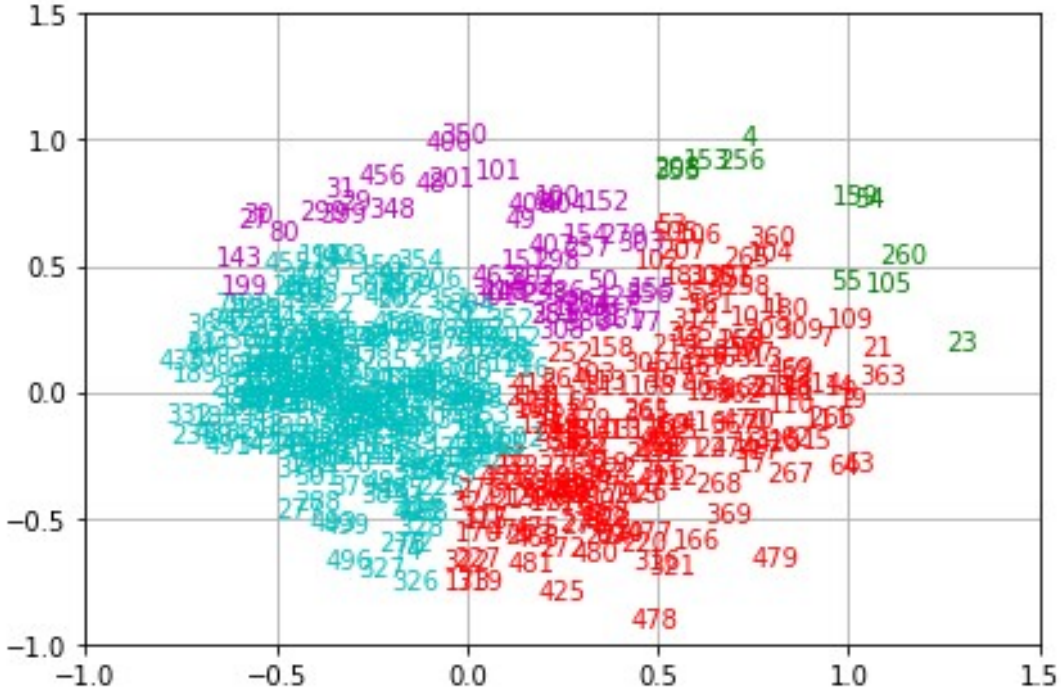
Iquitos:
Cut = 10

1st execution:



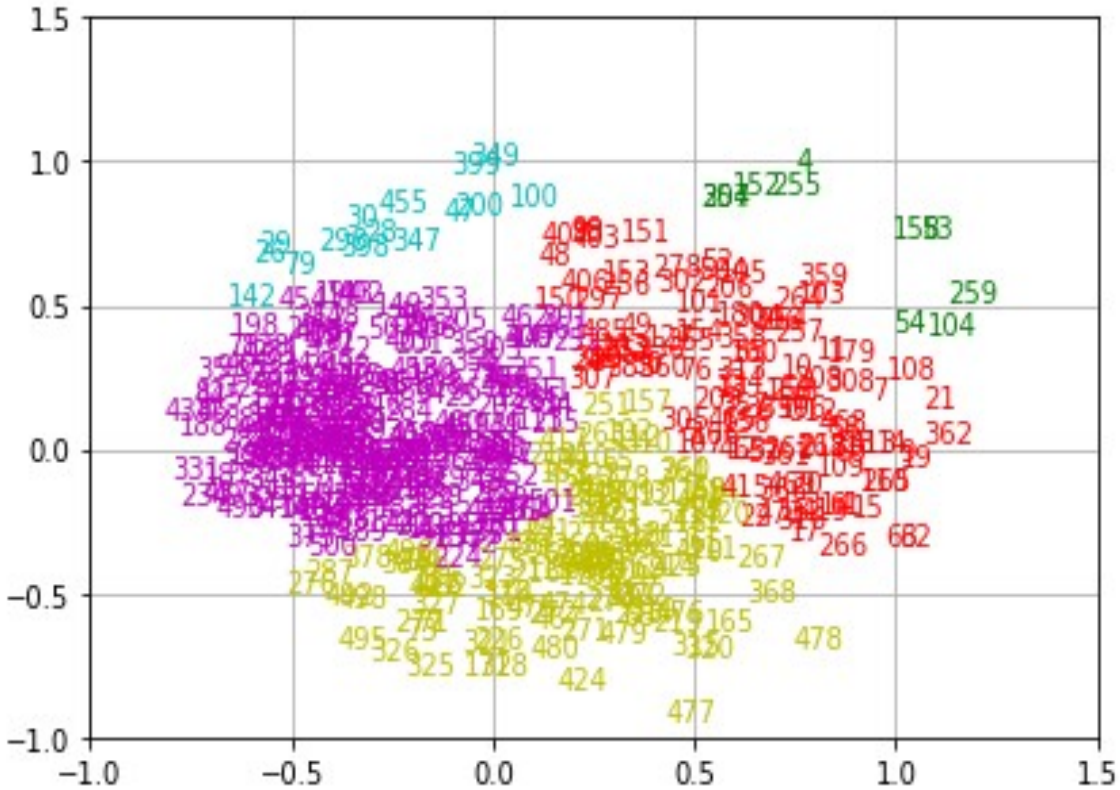
Outliers deleted: 244, 104, 3, 103, 51, 306, 10, 115, 273

2nd execution:



Outliers deleted: 23

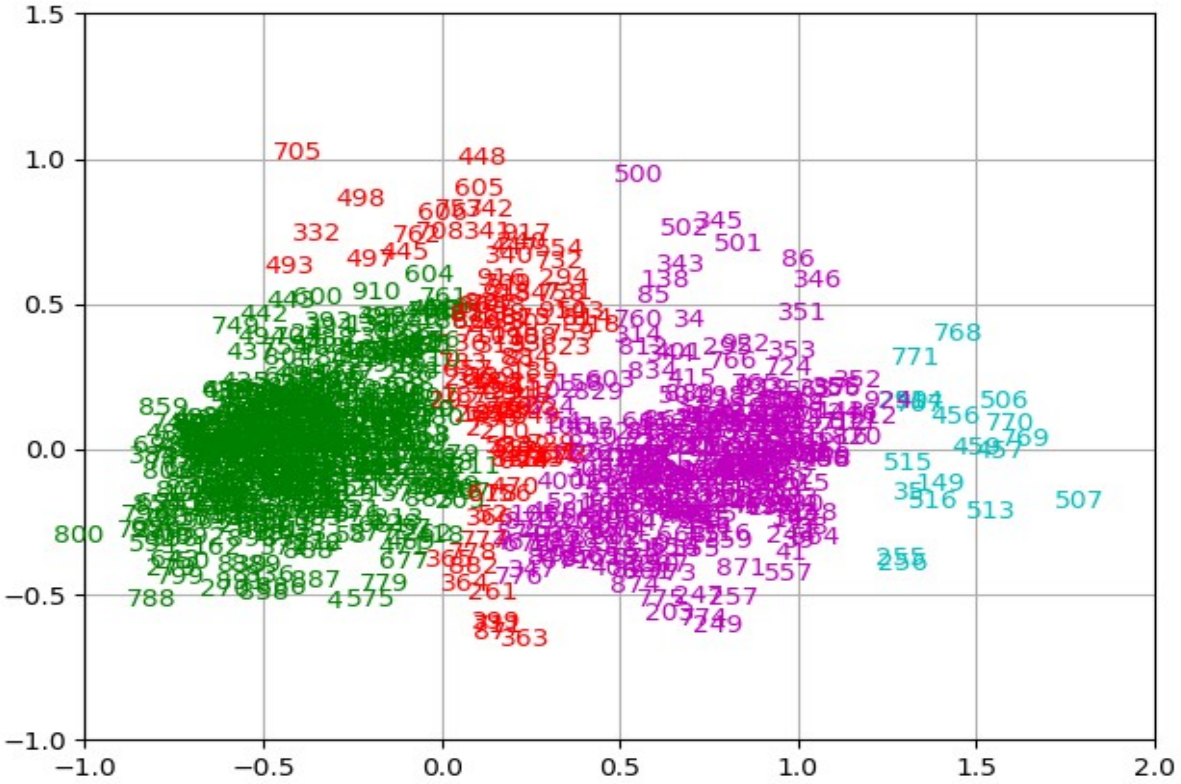
3rd iteration:



We considered to delete some outliers of the third group, but ultimately decided against it because the distances between the elements is not so big.

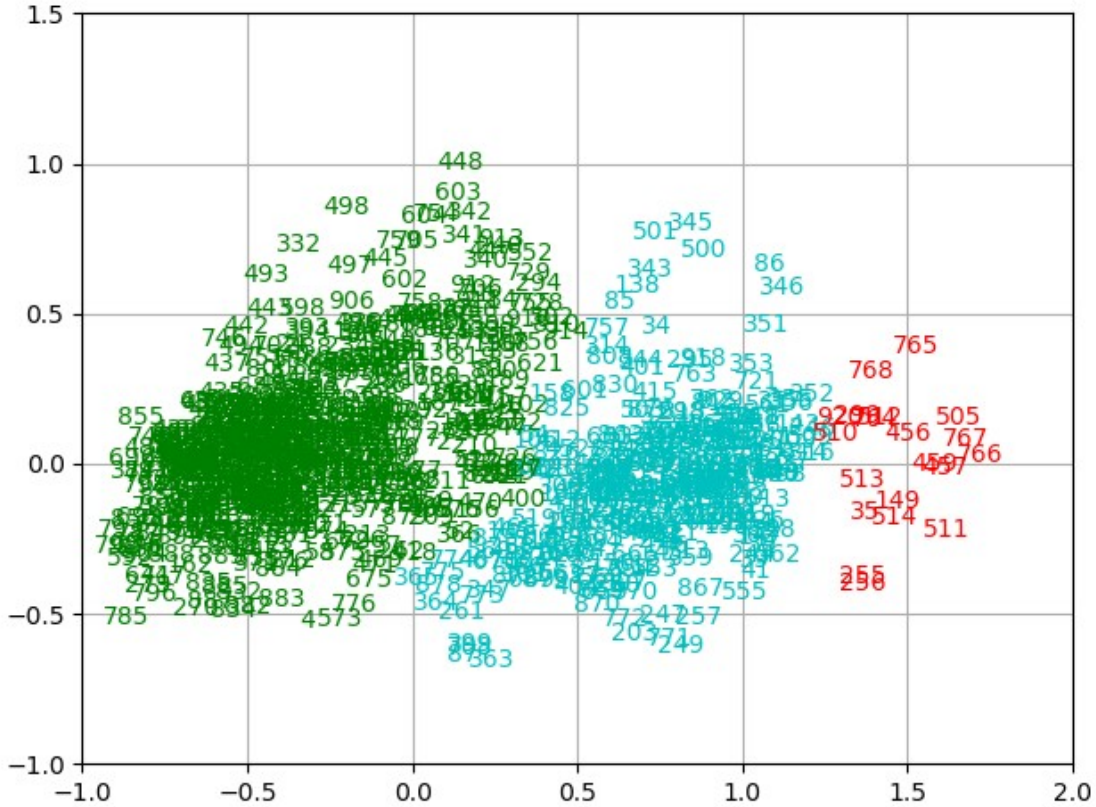
San Juan:
Cut = 14

1st iteration



Outliers deleted: 507, 500, 705 and 800

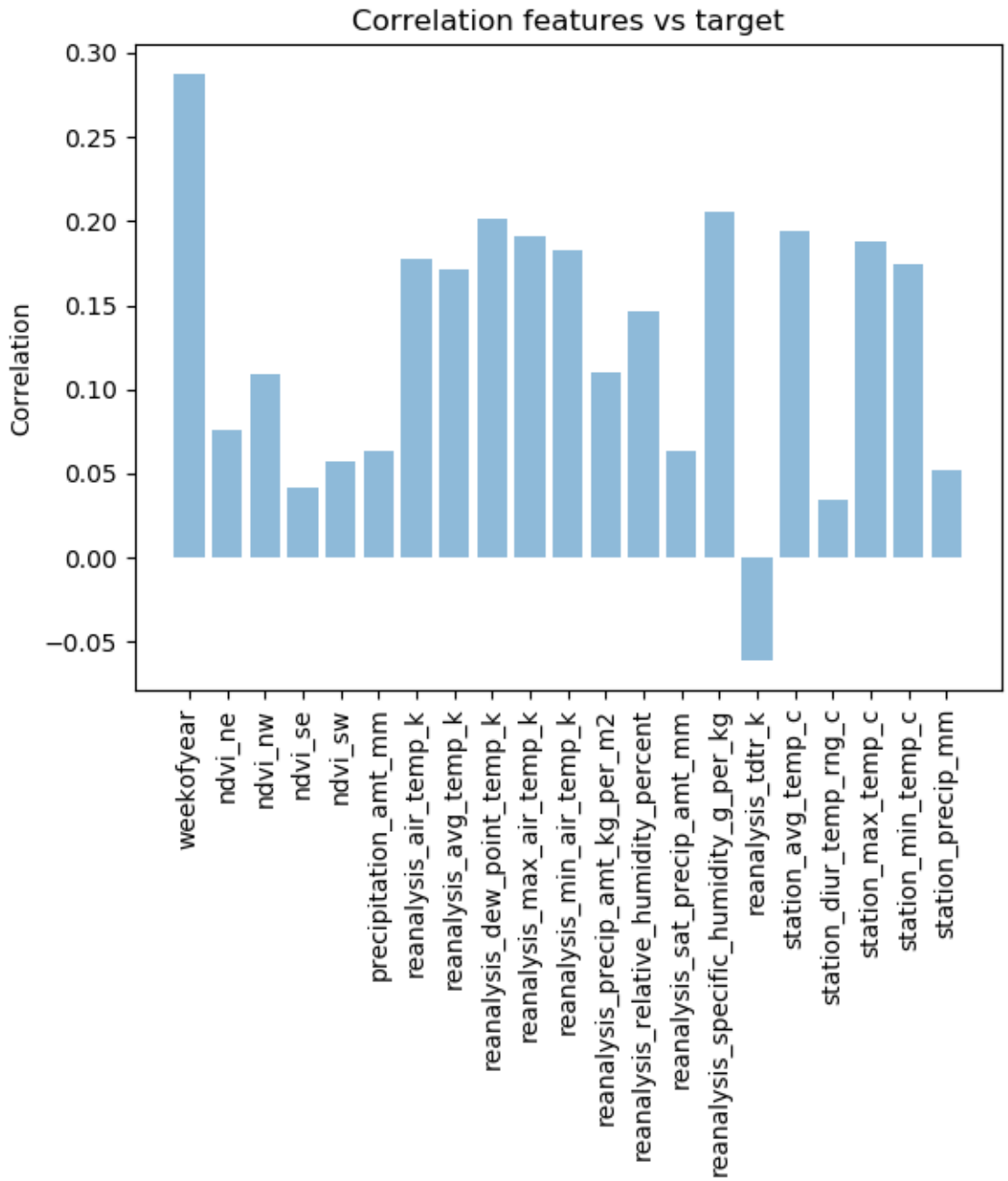
2nd iteration



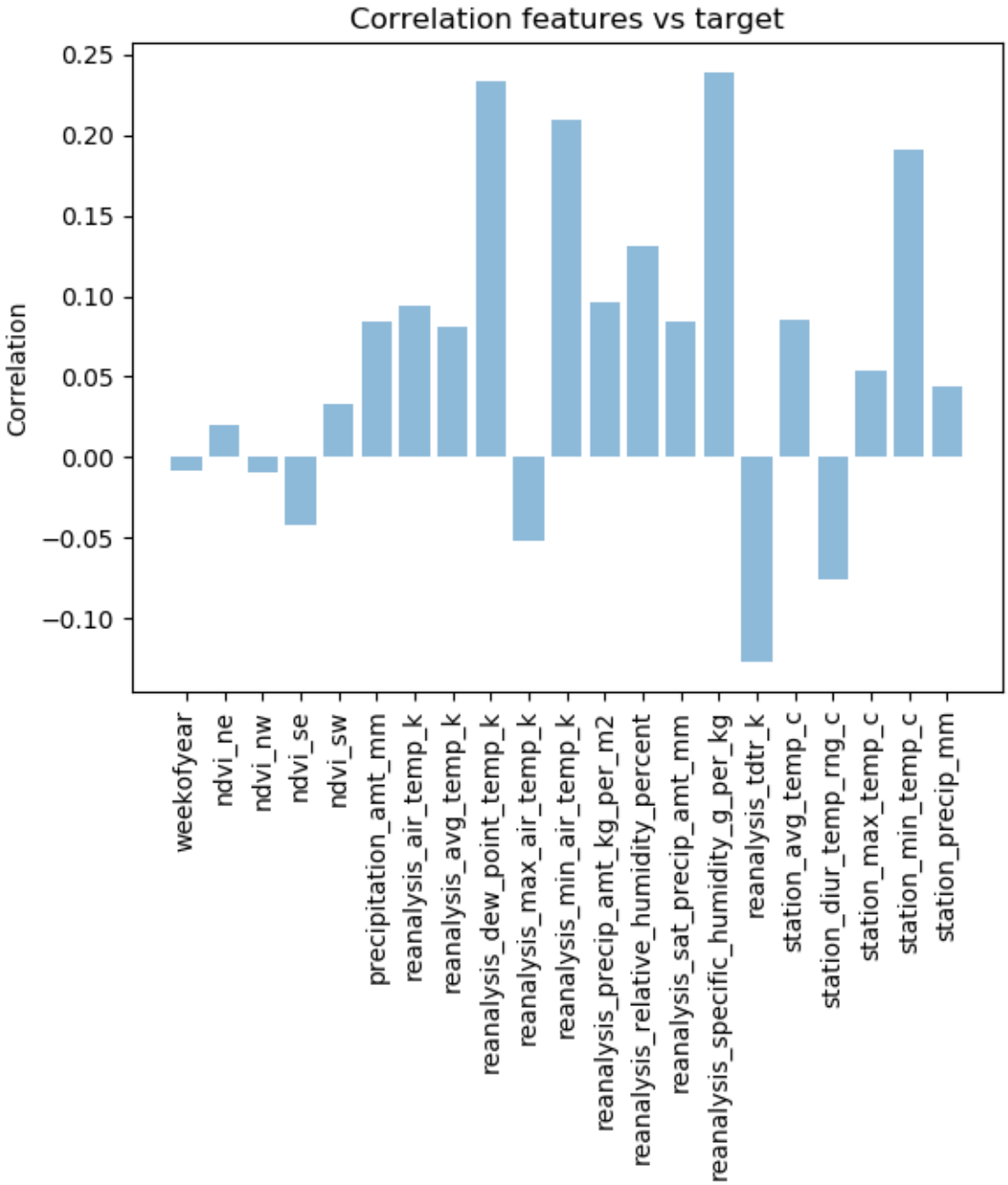
Feature selection analysis

Now for the feature selection we did the correlation between the target feature and the features.

Iquitos:



San Juan:


Correlation features vs target
Correlation
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
weekofyear
ndvi_ne
ndvi_nw
ndvi_se
ndvi_sw
precipitation_amt_mm
reanalysis_air_temp_k
reanalysis_avg_temp_k
reanalysis_dew_point_temp_k
reanalysis_max_air_temp_k
reanalysis_min_air_temp_k
reanalysis_precip_amt_kg_per_m2
reanalysis_relative_humidity_percent
reanalysis_sat_precip_amt_mm
reanalysis_specific_humidity_g_per_kg
reanalysis_tdtr_k
station_avg_temp_c
station_diur_temp_rng_c
station_max_temp_c
station_min_temp_c
station_precip_mm

And the decision tree relevancy between those features and the target one. The depth choosen for this task Features not represented in this table had 0 relevancy or were cut because other considerations, we talk more on this matter later.

Iquitos:

Feature	Relevancy
weekofyear	0.196364
reanalysis_specific_humidity_g_per_kg	0.690909
station_avg_temp_c	0.112727

San Juan:

Feature	Relevancy
weekofyear	0.690413
ndvi_nw	0.0831107
reanalysis_dew_point_temp_k	0.156723
reanalysis_specific_humidity_g_per_kg	0.0569902
station_min_temp_c	0.0127634

For San Juan we have 5 good relevant features and also 3 of them have good correlation with our target feature (total_cases) so we roll out with them.

But for Iquitos we only have 3 features, so we choosed to also get one more feature, **reanalysis_dew_point_temp_k**, because is the one that's more correlated next to those 3.

Deleted Features.

The features that we deleted before doing this study were:

city: We are already splitting the dataset between the only 2 cities, so this field is useless.

week_start_date: Harder to process since it's a string and we already have *weekofyear,* which gives us the same information.

year: Two factors were taken into account for this decision:

- Test data vs Training data. We don't care about the year when we are predicting the number of dengue cases, since maybe the prediction is for years that don't appear in our training data.
As an example, how can we predict the cases of a certain week of 2015 if we are lacking years 2011, 2012, 2013 and 2014? There's a very big time gap.

- The decision tree gave too much relevancy to this feature. This is related to the previous point, we want to establish relations between the features and having a decision tree be almost solely decided by a feature was a very bad symptom, specially by a non-cyclic time feature.
Nonetheless this also can gives us information, for Iquitos the correlation between *total_cases* and *year* was very positive, and in San Juan it was very negative, maybe indicating that in Iquitos the number of Dengue cases goes up each year, and in San Juan it goes down.

Other considerations and comments.

All the nulls were replaced by the mean of that column. We considered to do the mean of the 3 upper and lower rows as we did in the previous tasks, but we observed better results with this approximation.

As for why we didn't use correlation between features for the selection, we considered that it was way too risky to start deleting features based on how correlated they were.
What we mean by this is that we didn't have a good criteria to delete one feature over the other one that's very correlated to it. They may be correlated with each other, but that doesn't mean that they are equally important for the prediction of the total cases of dengue, which is out ultimately goal.
For future tasks we can always go back if needed, but for now we consider that deleting features without a good criteria as for why we are doing it is not a clever idea.

As a final note, we started keeping our code more tidy and working more with the tools in our disposal. Now we fixed the issue presented in the last task and we started doing all the changes to our database internally in memory with pandas while executing the script, there was no need to manually open the .csv file in an external program like LibreOffice Calc to make changes to the dataset.
We also created a small file called RedPandas.py with all the different functions that we are using, like creating a decision tree or doing a Cross-Validation test to check what should be the desirable depth to create a decision tree, so that way we can reuse them in other tasks without having to rewrite/copy and paste code. Because of time constraint we didn't include the hierarchical clustering functions in that little framework, so we have to use the same script used in the second Task.