Dengue Prediction

Introduction

A. Problem Statement

Predict the number of dengue cases each week (in each location i.e. San Juan and Iquitos) based on environmental variables describing changes in temperature, precipitation, vegetation, and more.

B. Data Files

File	Description
Submission Format	The format that the submission to the competition must be in.
Test Data Features	The features for the testing dataset
Training Data Features	The features for the training dataset.
Training Data Labels	The number of dengue cases for each row in the training dataset.

C. Submission Format

The submission format will have 4 columns – city (sj = San Juan, iq = Iquitos), year, weekofyear and the predicted number of dengue cases for that week.

	A	B	C	D	E
1	city	year	weekofyear	total_cases	
2	sj	2008	18	0	
3	iq	2008	18	0	
4	sj	2008	20	0	
5	iq	2008	20	0	

Exploratory Data Analysis

A. Summary statistics

	San Juan	Iquitos
Number of columns	24	
Number of useable features	20 (exclude city, year, weekofyear, week_start_date,	
Number of observations (train)	936	520
Number of years (train)	19 (1999 to 2008)	11 (2000 to 2010)

Number of dengue cases (train)	31993 (34 per week on average)	3934 (7 per week on average)
Number of observations (test)	260	156
Number of years (test)	6 (2008 to 2013)	4 (2010 to 2013)

B. Features

[1]

	CDR Normalized Difference Vegetation Index	NCEP Climate Forecast System Reanalysis	CDR PERSIANN Precipitation Product	GHCN daily climate data
0	ndvi_ne	reanalysis_air_temp_k	precipitation_amt_mm	station_avg_temp_c
1	ndvi_nw	reanalysis_avg_temp_k		station_diur_temp_rng_c
2	ndvi_se	reanalysis_dew_point_temp_k		station_max_temp_c
3	ndvi_sw	reanalysis_max_air_temp_k		station_min_temp_c
4		reanalysis_min_air_temp_k		station_precip_mm
5		reanalysis_precip_amt_kg_per_m2		
6		reanalysis_relative_humidity_percent		
7		reanalysis_sat_precip_amt_mm		
8		reanalysis_specific_humidity_g_per_kg		
9		reanalysis_tdtr_k		

C. Feature Distributions

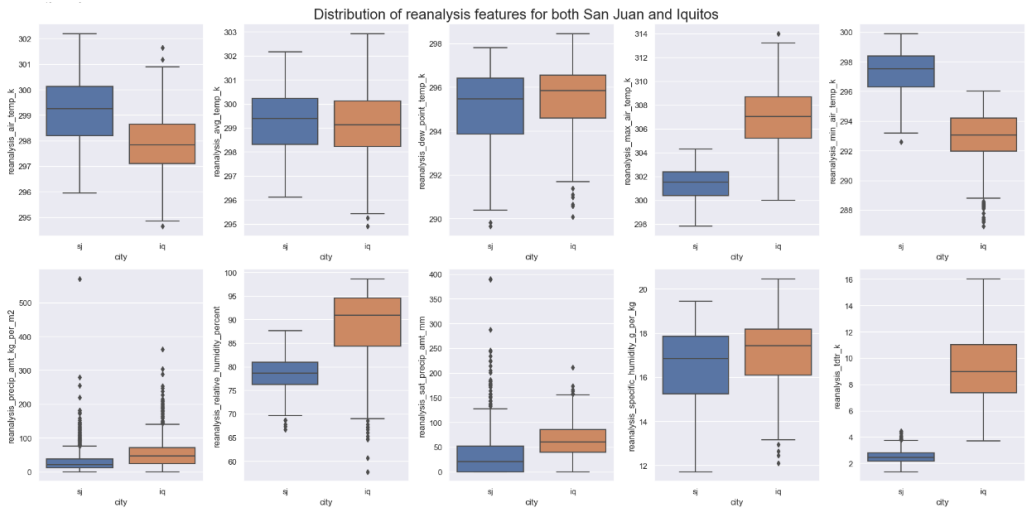
1. Reanalysis

"Reanalysis data provide the most complete picture currently possible of past weather and climate.

They are a blend of observations with past short-range weather forecasts rerun with modern weather forecasting models. They are globally complete and consistent in time and are sometimes referred to as 'maps without gaps'." [2]

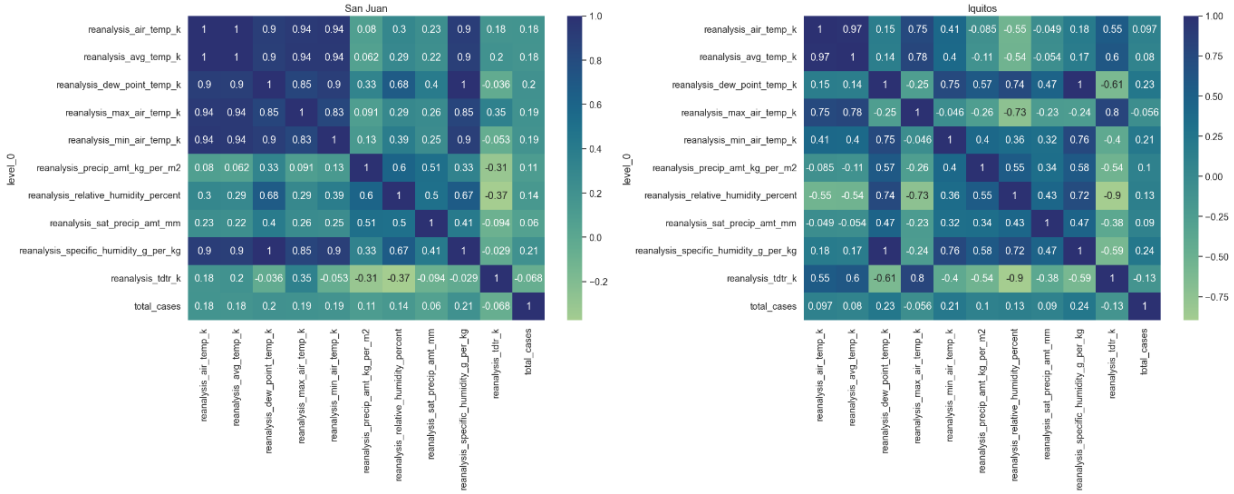
Almost similar trends for both cities. Iquitos seems to be more humid than San Juan.

[1] https://dengueforecasting.noaa.gov/docs/Metadata.pdf
[2] https://www.ecmwf.int/en/about/media-centre/focus/2020/fact-sheet-reanalysis



Checking correlations:

No strong correlations

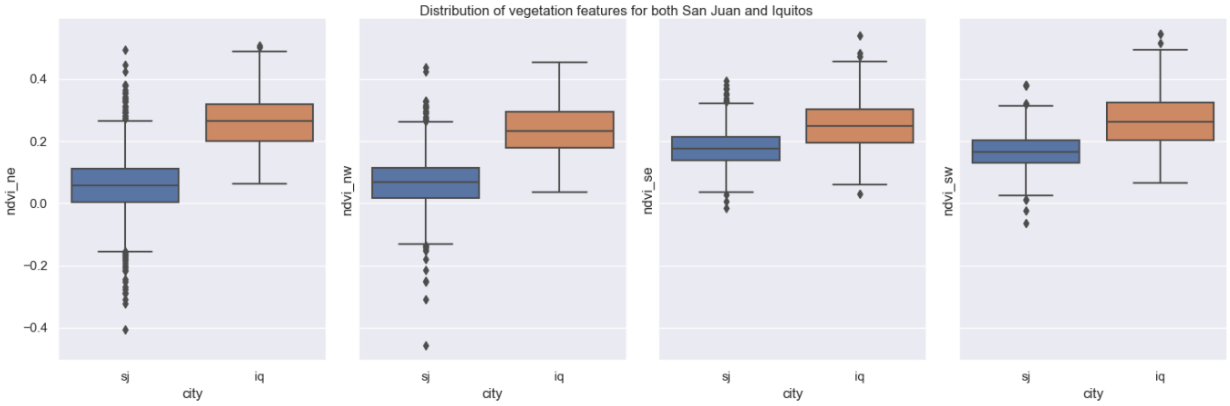


2. Vegetation

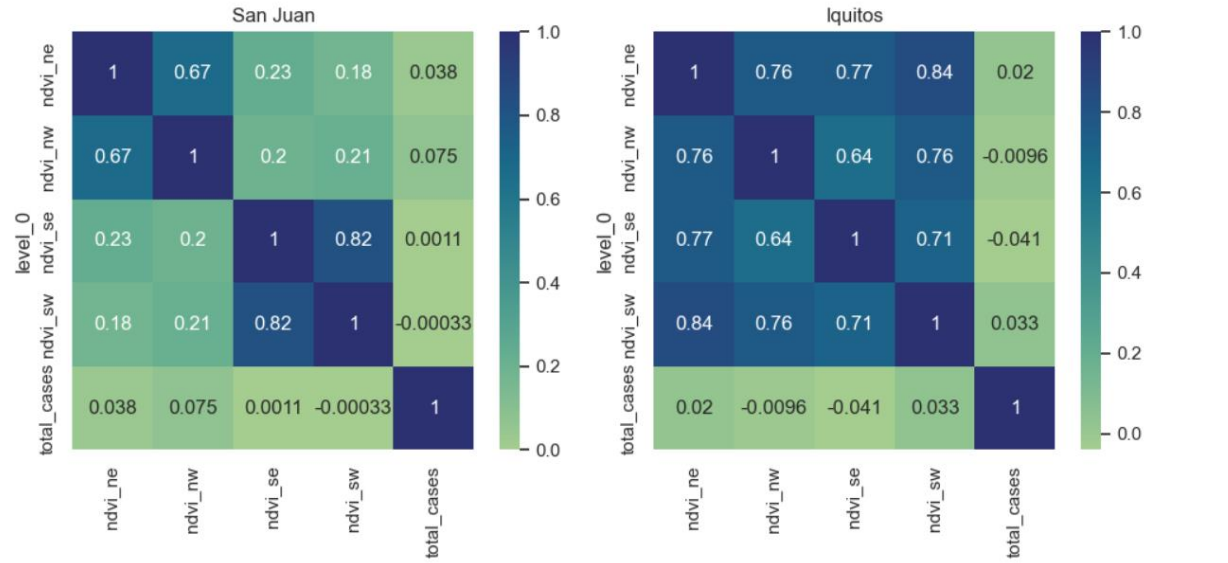
"NDVI always ranges from -1 to +1…For example, when you have negative values, it's **highly likely that it's water**. On the other hand, if you have an NDVI value close to +1, there's a high possibility that it's **dense green leaves**. But when NDVI is close to zero, there are likely no green leaves and it could even be an **urbanized area**." [3]

[3] https://gisgeography.com/ndvi-normalized-difference-vegetation-index/

There seems to be more vegetation in Iquitos compared to San Juan. Iquitos also seems to have fewer water bodies (cases where ndvi < 0). There are outlier areas in ndvi where there is more vegetation.



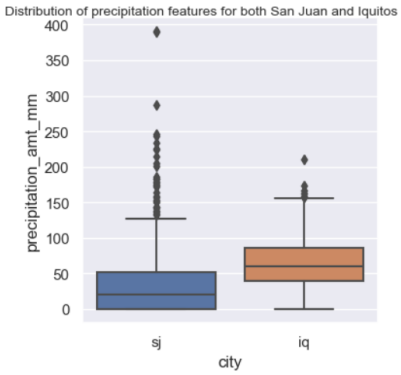
Checking correlations:



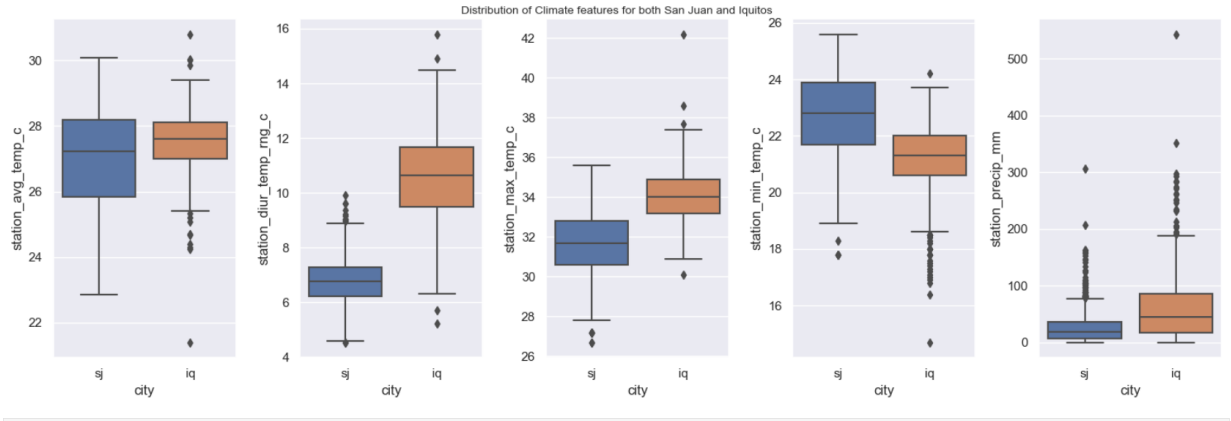
For San Juan, NW and NE have more water bodies as compared to SW and SE so their ndvi's correlation is slightly higher with total cases.

For Iquitos, there doesn't seem to be any apparent trend.

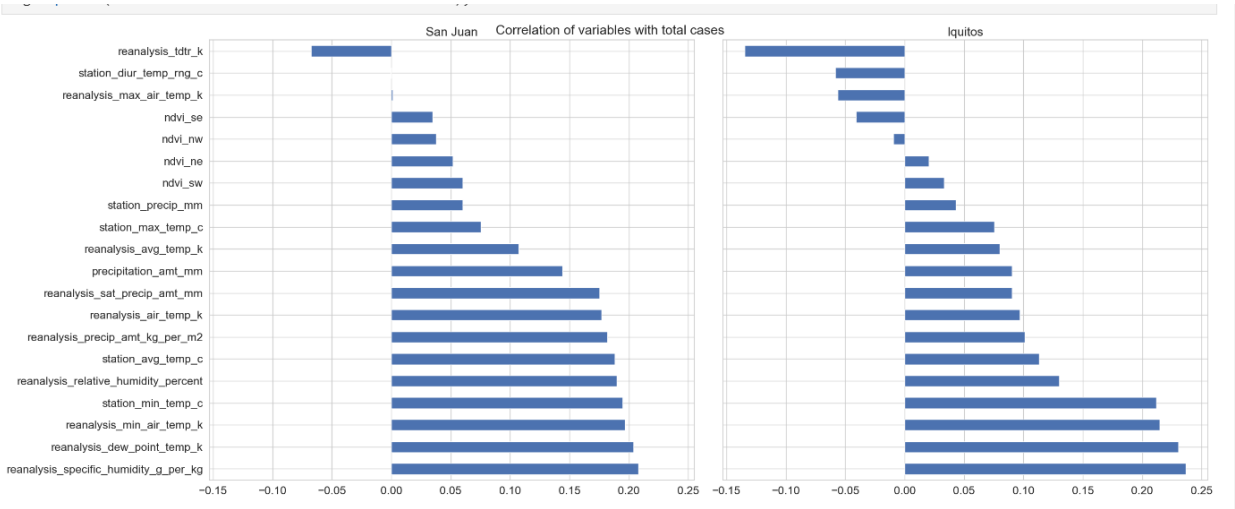
3. Precipitation



4. Climate

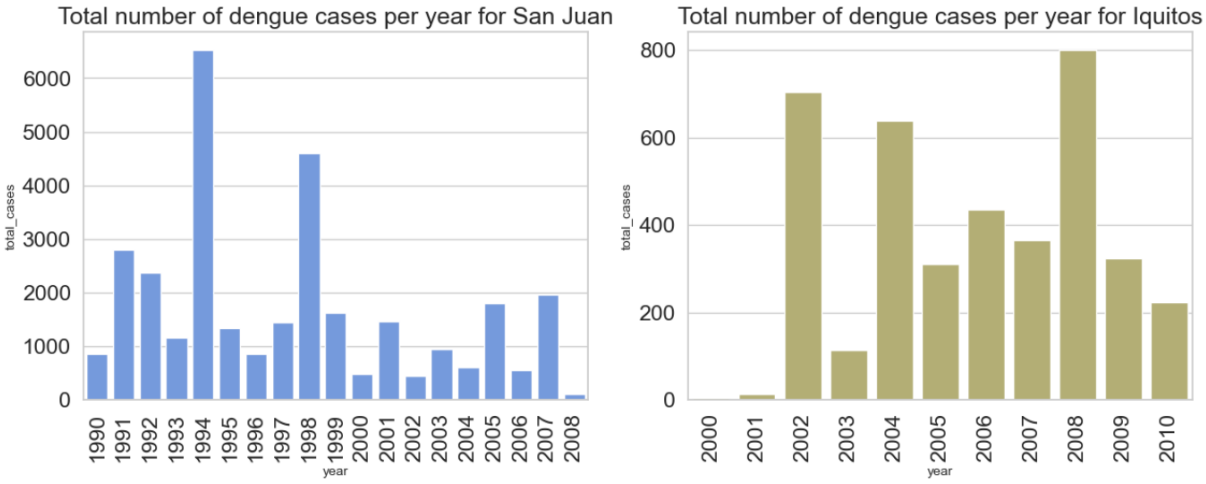


Overall correlations. For San Juan, vegetation variables are more correlated with total number of cases than Iquitos. For Iquitos, temperature related variables are more correlated with total number of cases



D. Sanity Checks

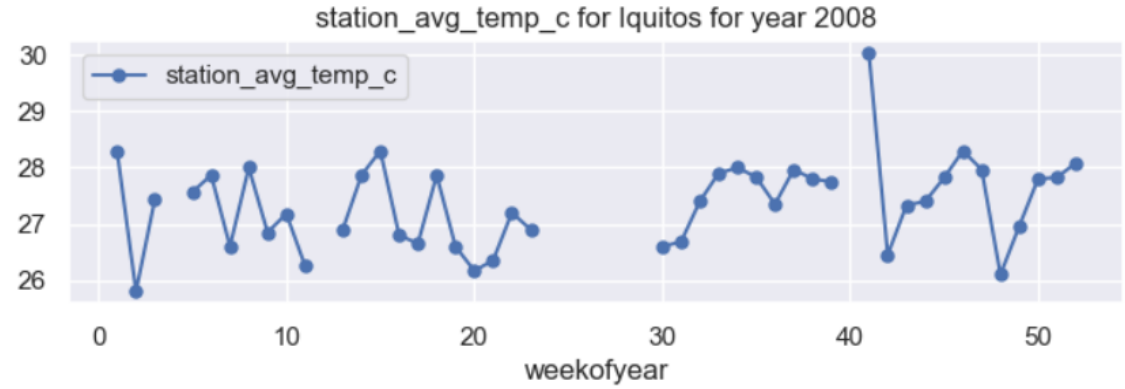
- There is no value equal to -9999 in climate and precipitation variables? Or any other unique value that is assigned to missing values? No
- There are no gaps in weeks? No
- Any year, where there were no cases of dengue/unusually low? In 2000 for Iquitos there were only 4 cases and San Juan has only __ cases for 2008. We have data after 26th week for Iquitos for year 2000, and data till 17th week for San Juan for 2008
- Were there any apparent outbreaks? San Juan in 1994 and 1998. Iquitos in 2002 and 2008



E. Imputing Missing values

Ndvi_ne is 20% missing for San Juan and station average temperature is 7% missing for Iquitos. The rest are below 5% missing

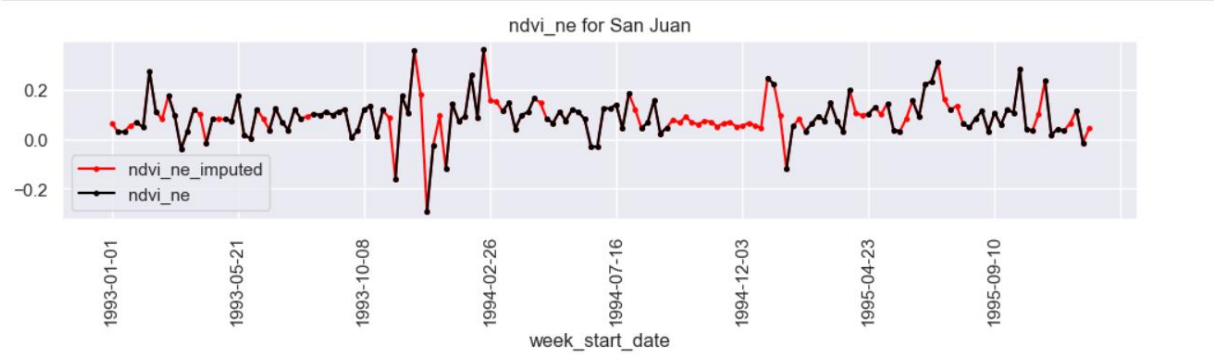
	precent_missing_sj	precent_missing_iq
index	0.000000	0.000000
city	0.000000	0.000000
year	0.000000	0.000000
weekofyear	0.000000	0.000000
week_start_date	0.000000	0.000000
ndvi_ne	20.405983	0.576923
ndvi_nw	5.235043	0.576923
ndvi_se	2.029915	0.576923
ndvi_sw	2.029915	0.576923
precipitation_amt_mm	0.961538	0.769231
reanalysis_air_temp_k	0.641026	0.769231
reanalysis_avg_temp_k	0.641026	0.769231
reanalysis_dew_point_temp_k	0.641026	0.769231
reanalysis_max_air_temp_k	0.641026	0.769231
reanalysis_min_air_temp_k	0.641026	0.769231
reanalysis_precip_amt_kg_per_m2	0.641026	0.769231
reanalysis_relative_humidity_percent	0.641026	0.769231
reanalysis_sat_precip_amt_mm	0.961538	0.769231
reanalysis_specific_humidity_g_per_kg	0.641026	0.769231
reanalysis_tdtr_k	0.641026	0.769231
station_avg_temp_c	0.641026	7.115385
station_diur_temp_rng_c	0.641026	7.115385
station_max_temp_c	0.641026	2.692308
station_min_temp_c	0.641026	1.538462
station_precip_mm	0.641026	3.076923



1. ARIMA:

   All values that are missing for more than 1% of the time are imputed using ARIMA with the following parameters:
   - p (order of AR): number of lags of signal to be used as predictors = 4 (using almost 1 month (4 weeks) of past data to predict next missing value)
- q (order of MA): number of lagged forecast errors that should go into the ARIMA Model = 4 (errors for last 4 weeks of data)
- d (order of integration): minimum number of differencing needed to make the series stationary.
  Used Augmented Dickey Fuller test (ADF) to determine this.
  The time series seems stationary so used d = 0

An example:



2. Forward Fill:

All values that were missing for less than 1% of the time were just forward filled

A. Feature Importance using Random Forest

a. Static data:

One approach can be to consider that data as static i.e. we have a bunch of columns (this week's data) and we predict the number of cases happening this week. Example:

station_max_temp_c	station_min_temp_c	station_precip_mm	total_cases
29.4	20.0	16.0	4
31.7	22.2	8.6	5
32.2	22.8	41.4	4
33.3	23.3	4.0	3
35.0	23.9	5.8	6

b. Time-series forecasting data

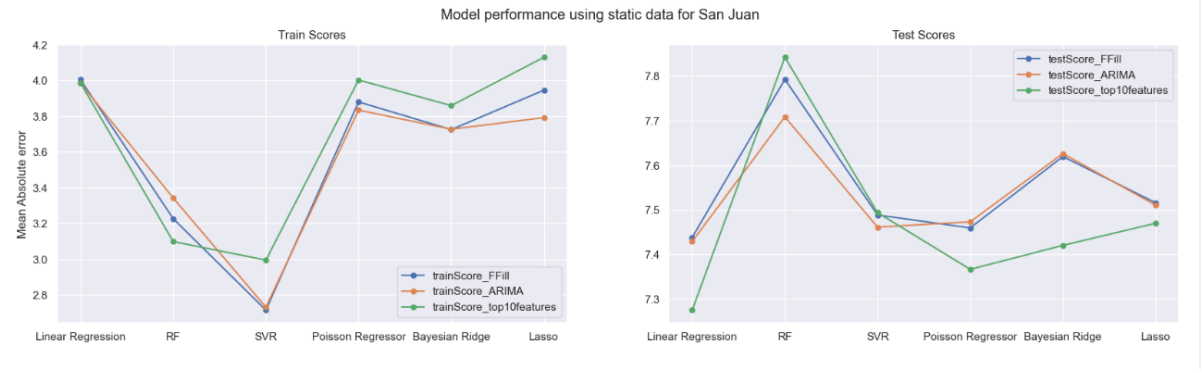
Another approach can be to use a moving window and reshape the data such that it becomes a time-series forecasting problem.

	ndvi_se(t-1)	station_avg_temp_c(t-1)	no_of_cases(t)
1	0.198483	25.442857	5
2	0.162357	26.714286	4

F. Cross validation to select models:

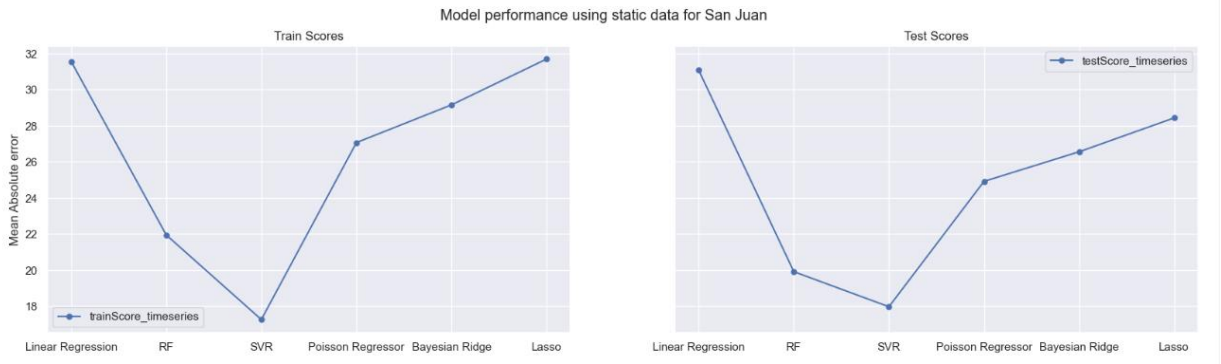
1. Static data:

	trainScore_FFill	testScore_FFill	trainScore_ARIMA	testScore_ARIMA	trainScore_top10features	testScore_top10features
Linear Regression	4.004173	7.438405	3.984435	7.429026	3.985931	7.275345
RF	3.226750	7.792724	3.342333	7.708590	3.098667	7.842244
SVR	2.715252	7.488533	2.731422	7.461290	2.995190	7.493739
Poisson Regressor	3.880183	7.459662	3.834298	7.473319	4.003011	7.366798
Bayesian Ridge	3.725062	7.619573	3.727294	7.626264	3.859532	7.420478
Lasso	3.945884	7.516061	3.792386	7.510525	4.130050	7.470039



2. Time Series Data

	trainScore_timeseries	testScore_timeseries
Linear Regression	31.559900	31.069020
RF	21.948073	19.913209
SVR	17.261421	17.971386
Poisson Regressor	27.065980	24.916965
Bayesian Ridge	29.154843	26.556164
Lasso	31.695609	28.432935



G. Final submission

SUBMISSIONS

Score	Submitted by	Timestamp
26.8678	hadiahameed	2022-11-25 20:11:41 UTC
27.9231	hadiahameed	2022-11-25 20:20:01 UTC