Predoc Data Task – Short Answer Document

September 19, 2020

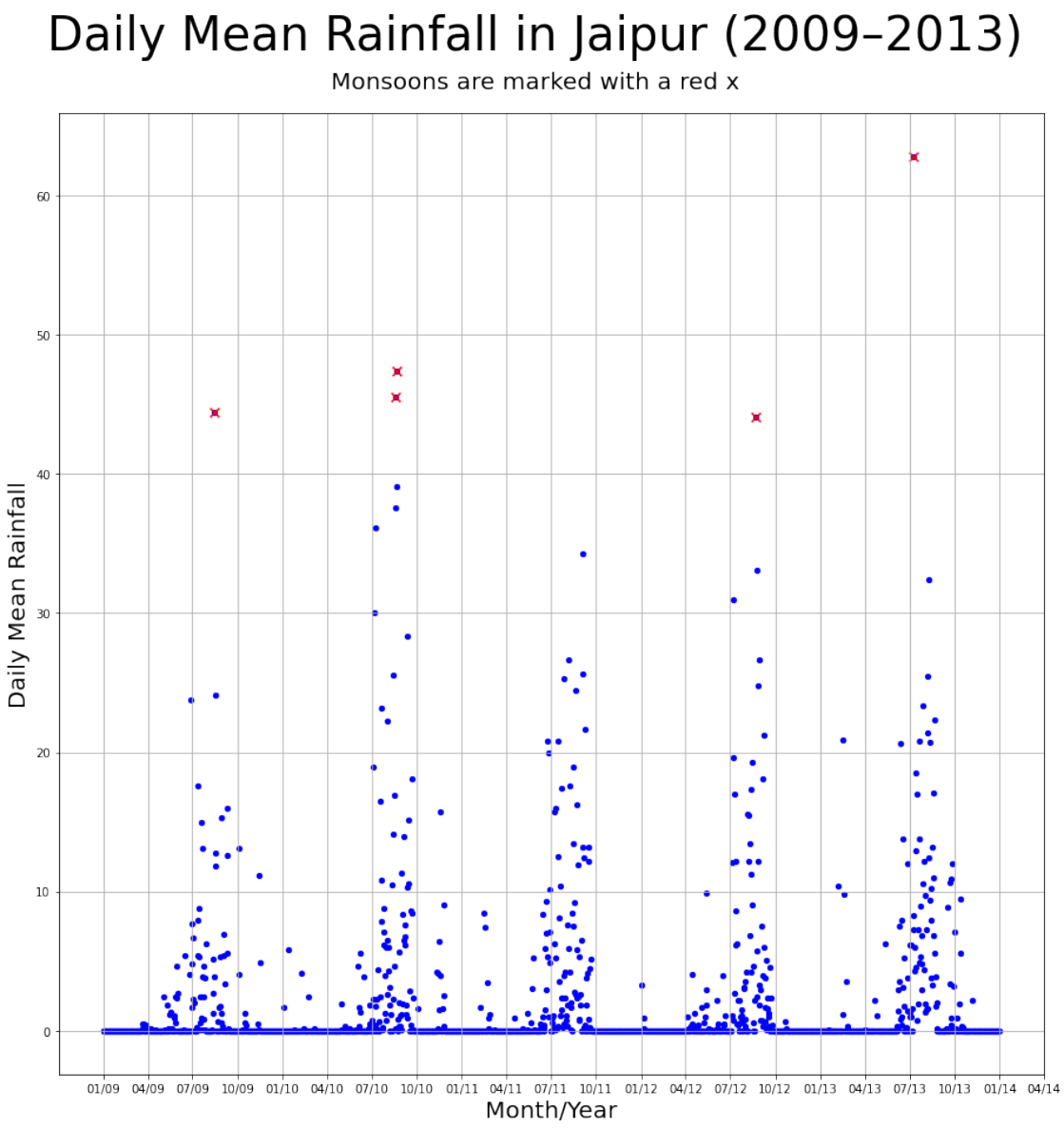
1 2: Data Cleaning

Computing the distance between every temperature/rain coordinate and every state centroid took quite a bit of time, over an hour for both datasets. I am using a 2.9 GHz Intel i5 6th Generation (Skylake) laptop processor, not a particularly fast machine but it is not outdated either. For a very large dataset, more computing power will be required. If computing power is not available, then it may be preferable to replace `vincenty` with `great_circle`, which is much less accurate but is around twice as fast to compute.

2 3: Data Exploration

2.1 2

[20]:



I have highlighted the points in the topmost sector of the graph as monsoons. All of these heavy rainfall events occur in between months July and October. Therefore I conclude that monsoon season generally starts in July and ends in October.

3 4: Estimation and Causal Inference

3.1 1

$mortality = \beta_0 + \beta_1 temperature$

3.2 2

A β_1 coefficient of 0.05 would indicate that a rise of 1 degree Celsius increase in temperature leads to a 0.05 increase in the mortality rate. A standard error of 0.02 indicates that the data closely hugs the line of best fit.

3.3 3

If temperature becomes too low, mortality will go up. If temperature becomes too high, mortality will go up. There is a middle range where mortality will go down. This indicates we should introduce a squared parameter to account for this. The new model is $mortality = \beta_0 + \beta_1 temperature + \beta_2 temperature^2$.

3.4 4

Adding in control variables to such as fixed effects (age, gender, urban/rural) or random effects (humidity) can help seek and out and mitigate effects from potential confounding variables. By adding control variables to our equation, their effect on *mortality* will be separated from *temperature*, and therefore the coefficient on the *temperature* will be more accurate.

3.5 5

$mortality = \beta_0 + \beta_1 temperature + \beta_2 temperature^2 + \beta_3 age + \beta_4 gender + \beta_5 urban + \beta_6 humidity$

In this model, *age*, *gender*, and *urban* are 0-1 dummy variables.

The standard error will be measured by taking the sum of squared residuals between the line of best fit and the data points.