# Case Study on TN Plantation 2015-2016

## Problem Statement:

Based on the facts obtained from the dataset **“**[**Area & Production of Plantation Crops by District-wise in Tamil Nadu for the Year 2015-16**](https://tn.data.gov.in/catalog/area-production-productivity-plantation-crops-district-wise-tamil-nadu-year-2015-16)**”,** Tamil Nadu government needs to publish a statistics on the below questions:

* Analyse which of the district have maximum and least production of
  + Tea
  + Bamboo
  + Rubber
* Total area of each plantation under all districts
* Identify the whether there are any relation between bamboo, tea and rubber plantation of each of the district

Data Set:

  

## Glimpse of Data Set:

All the attached data set contains below schema:

* **Serial\_Number:** Integer Type
* **District:** String Type
* **Area:** Integer Type
* **Production:** Integer Type

## Instructions:

**Inference of each data:**

* **Bamboo-plantation-15-16.txt:**
* The file does not contain header, therefore define the schema using Struct Type
* **Column District:**
  + Should contain String data only, but the field has special characters inside the data
  + **Clue:** Use regex and replace function to remove the special character under the District column
* **Column Area:**
  + Column Area should contain only integer data only, but the data has alphabets between the data
  + **Clue:** Use regex and replace function to remove the alphabets under the Area column
* **Tea-plantation-15-16.txt:**
* The file does not contain header, therefore define the schema using Struct Type
* The file provided is a **“Sequence File”,** meaning it does not have newline character between each row
* **Clue:** Replace every 5th delimiter using “\n” character to format the file.
* **Rubber-plantation-15-16.txt:**
* Clean file with header
* Cast each column to the above data type mentioned under the schema section

**General Rules that applies to all Data Set:**

* All the data set attached has “|” delimiter
* Add Column for each of the above data set having column name as “**Plantation Type**” with schema “**String Type**” and value substring from the file name (ex: **Bamboo**-plantation-15-16.txt)
* Add a column for each of the above data set having column names as “**Productivity**” with schema “**Float Type**” and value as result of (Value of production) divided by (Value of Area)

## Transformation:

* Once the file is cleansed and populated into the Dataframe
* Perform **union** function with the above 3 Dataframe
* Write the solution to derive results for the above business problem using Spark Sql

**Evaluation Parameter:**

* Check whether data is ingested and converted to data frame
* Check whether the file cleansing is applied with proper regex and replace logic inside a function for bamboo plantation
* Check whether the sequence file is converted with the 5th occurrence of pipe delimited to newline character
* Manual correction directly on the file is not encouraged
* Evaluate the Spark SQL transformation logic written actually derives the solution for the business problem
* Evaluate the outcome and achieved what is expected