Name:- A B Eswara Krishna

eswar krishna

[Company name]  [Company address]

Mile STONE REPORT

** An introduction to the problem (based on your earlier Capstone submissions).**

Problem description is documented in the Github in the below link.

<https://github.com/eswarkrishna/datascience/blob/master/CapstoneProposal1.Rmd>

** A deeper dive into the data set:**

* What important fields and information does the data set have?

Actually there are lot of Information in the data set I chose. There are nearly 600+ Variables to be analysed and combined. After analysing all the dataset. I felt there are around 50 Variables which are important. I am targeting to get to around 50 variables to be analysed.

* What are its limitations i.e. what are some questions that you cannot answer with this data set?

It’s a really huge dataset in terms of the number of columns.

* What kind of cleaning and wrangling did you need to do?

Need to use DPlyr package to clean and merge multiple worksheets,

** Any preliminary exploration you’ve performed and your initial findings.**

I have already worked on this dataset previously, the data is pretty exhaustive and real.

You just have to clean and merge the data to get the actual data.

** Based on these findings, what approach are you going to take? How has your approach changed from what you initially proposed, if applicable?**

The approach hasn’t been changed much. Below are the tasks needed to be done based on my findings.

1) Filter the data from the Excel sheet 3 yrs 5 states

2) Tidy it and clean up unnecessary data

3) Plotting and visualisations using ggplot/tableau

4) Correlation analysis. We will be able able to figure out the relations between features / parameters

5) Statistical model building on the data like regression analysis

6) you will need additional test data to build predictions and its accuracy. Prediction score

7) Rmd file or word doc to be submitted for report .