FALL 2020

Big Data Project

MIS 6346.002 - Big Data - F20

**Group 5**

Anchal Jain

Harshit Gupta

Rasika Pawar

Rhea D'souza

Palash Jain

Pei-Wen Liu

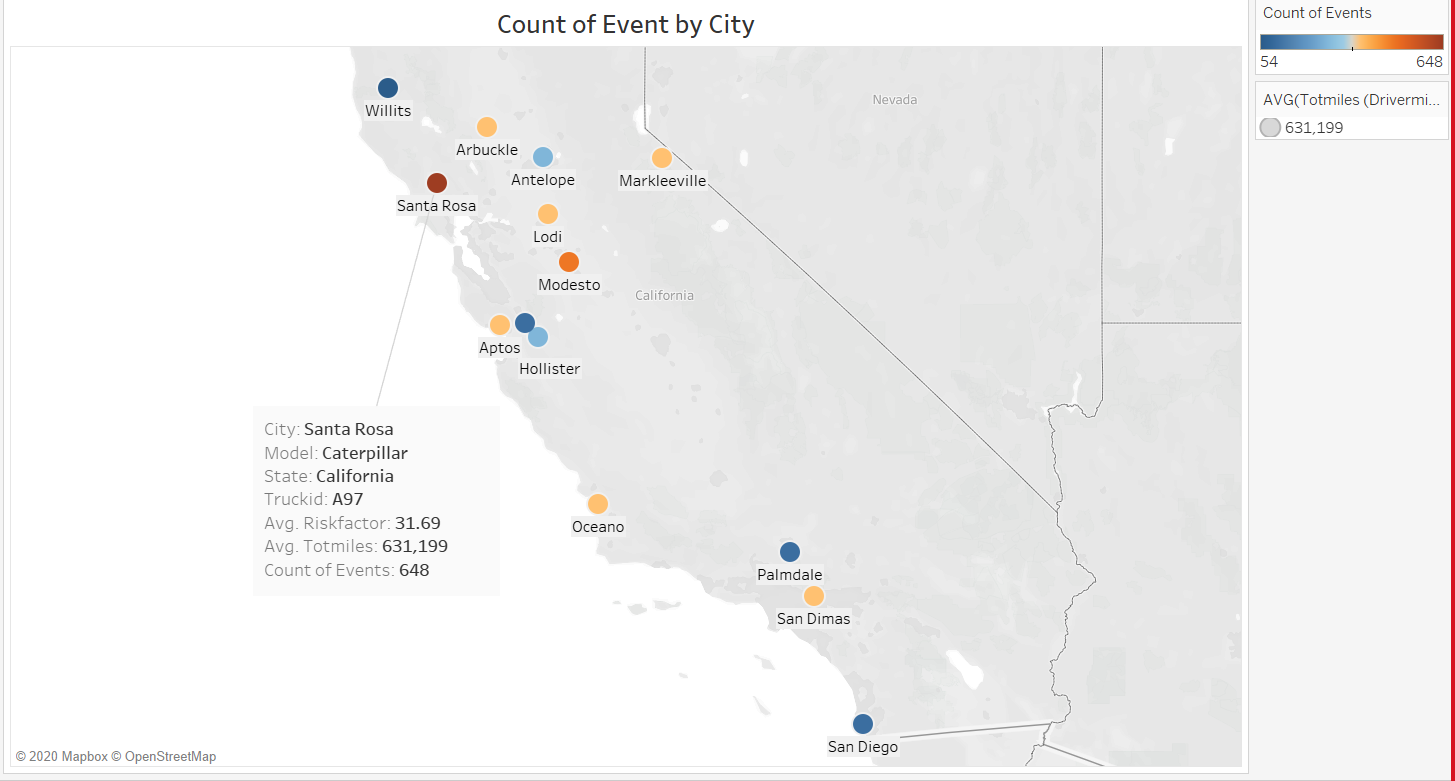
Aditya Bairagi

After importing the data from Cloudera Impala to Tableau by using IP address for the Impala - (Running on HUE server). Further, we identified relationships and associations between the different table columns and performed joins on those data tables and then loaded these data tables in Tableau for creating visualizations and dashboards.

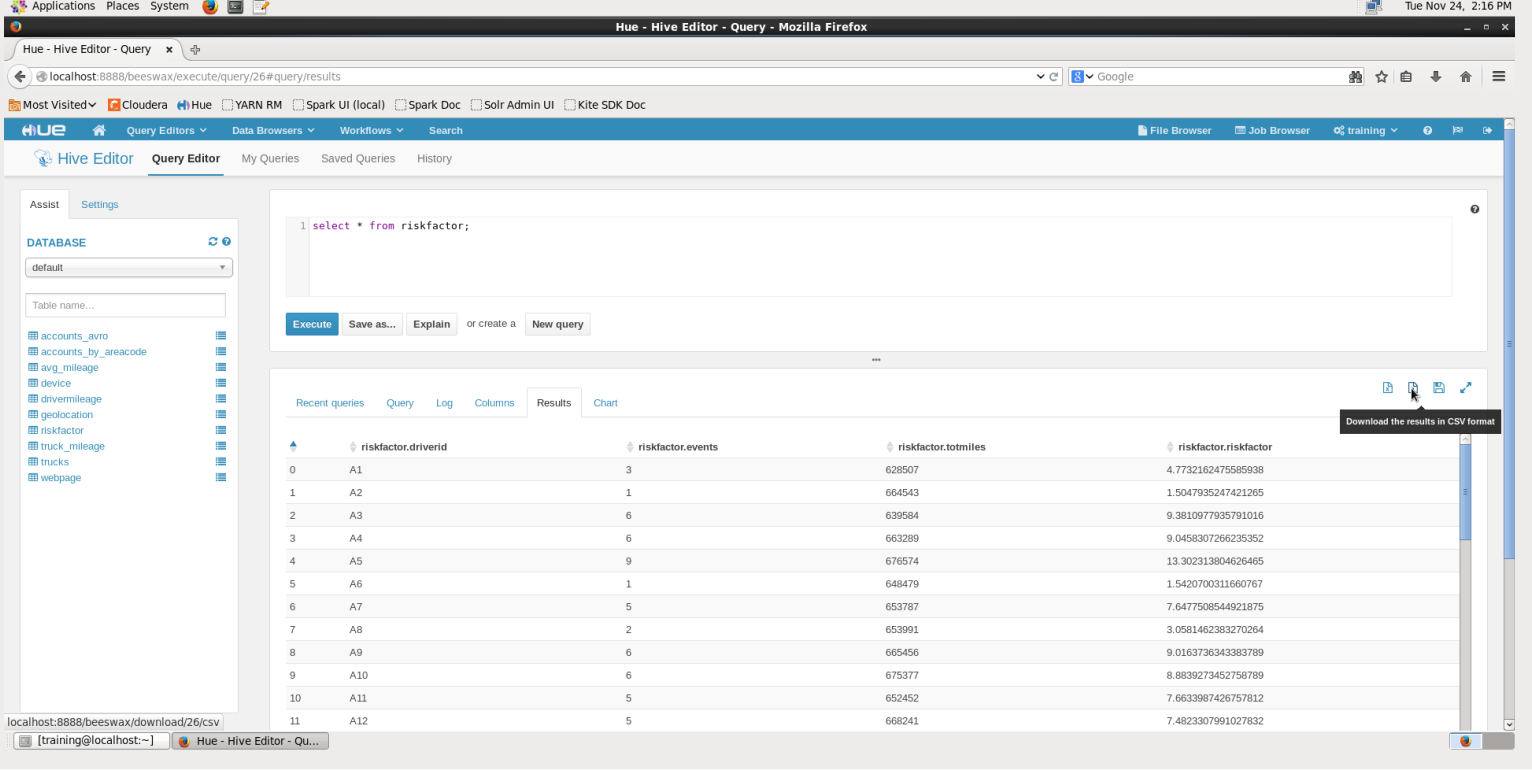
Some of the noticeable points in these dashboards are:

1). It provides information about the location of 15 drivers who are under the highest risk on scaled risk factor of 0-10 with 9.902 for A97 being the highest scaled risk factor and 2.932 being the lowest.

2). Driver Id A97 undergoes the highest risk as it undergoes a lot of risky events like Over Speeding, Lane Departure, Normal, unsafe tail distance etc. in the California region.

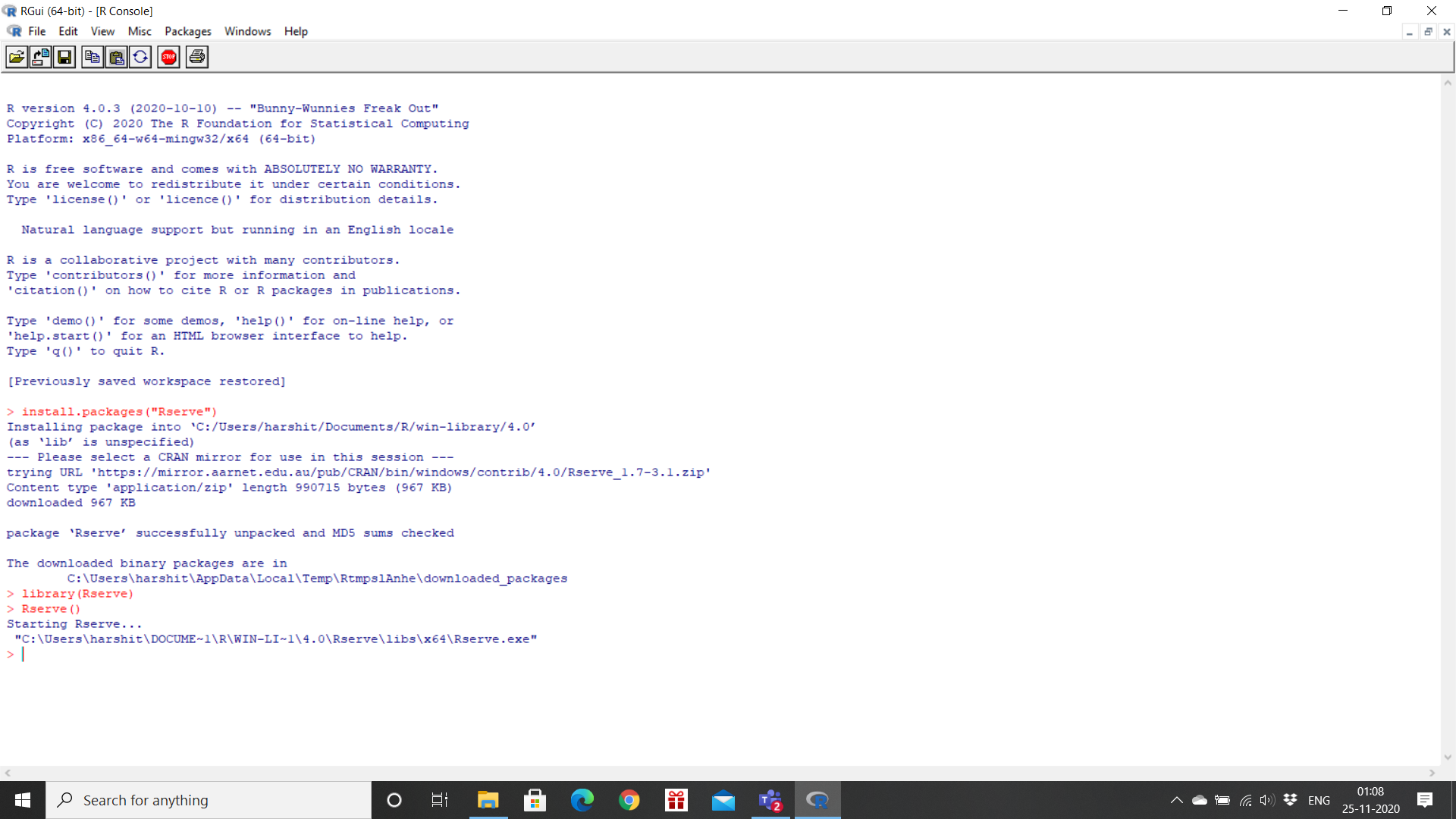


We have downloaded the riskfactor table in .csv format from virtual machine in our local system :



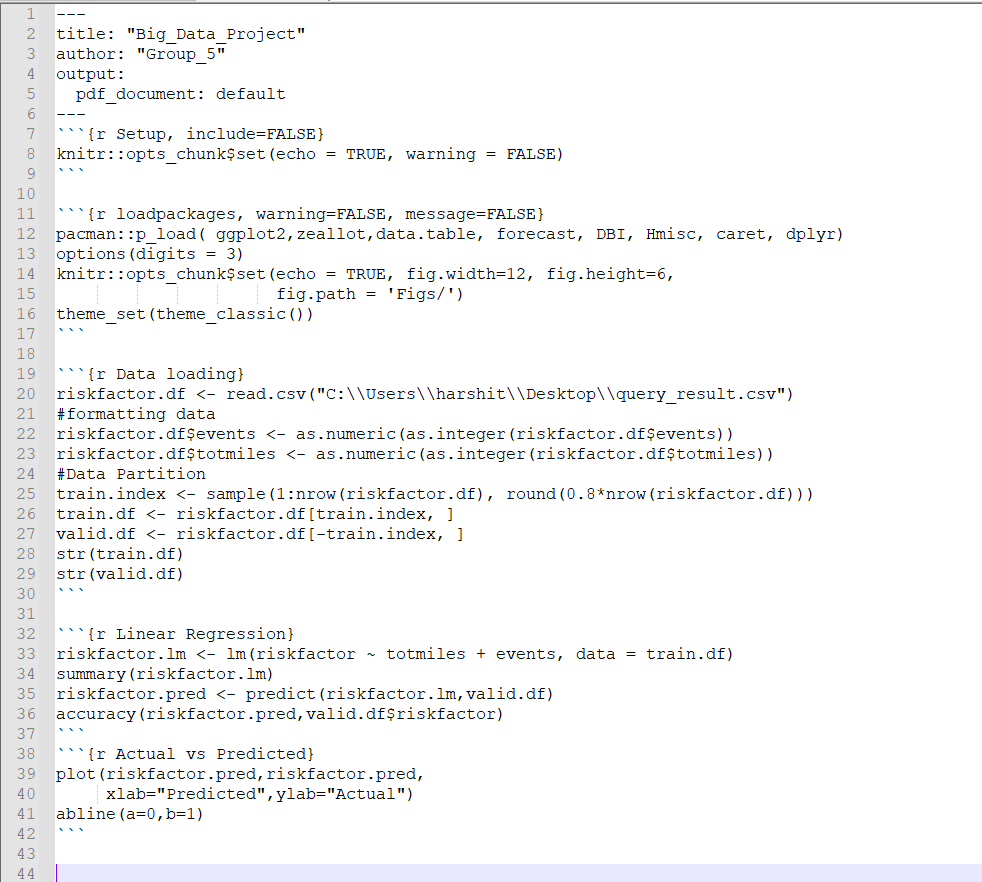
**R Implementation using Linear Regression:**

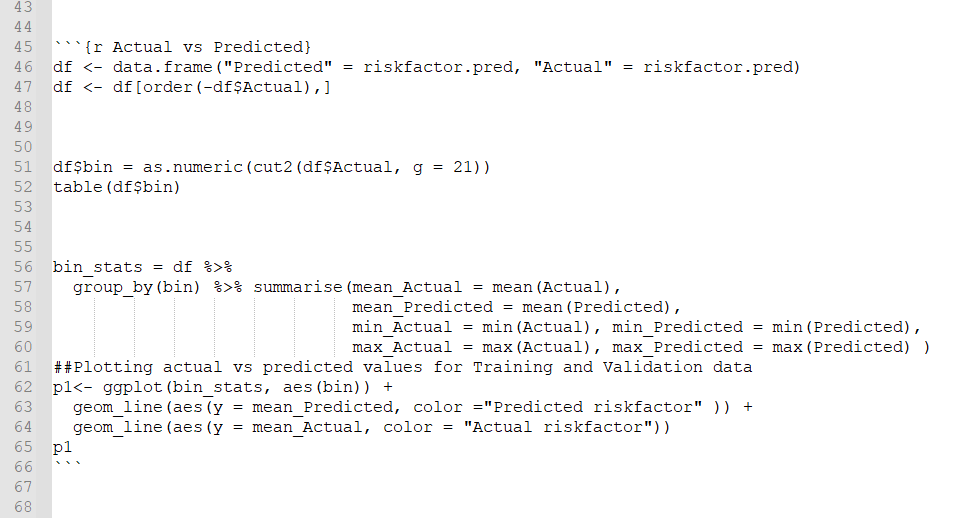
Here we are trying to implement Linear Regression using R in Tableau. In order to connect R with Tableau we have to use Rserve() package to establish connectivity between the two. We need to import package Rserve using import.package(“Rserve”) as shown in the screenshot :



**R Code:**

After establishing a connection between R and tableau, we can download data from local machine i.e. from hive and import it in .csv format. This .csv file is an input to the Linear regression model. We used this .csv to load data into R. We have partitioned this dataset into training dataset(80%) and validation dataset(20%).

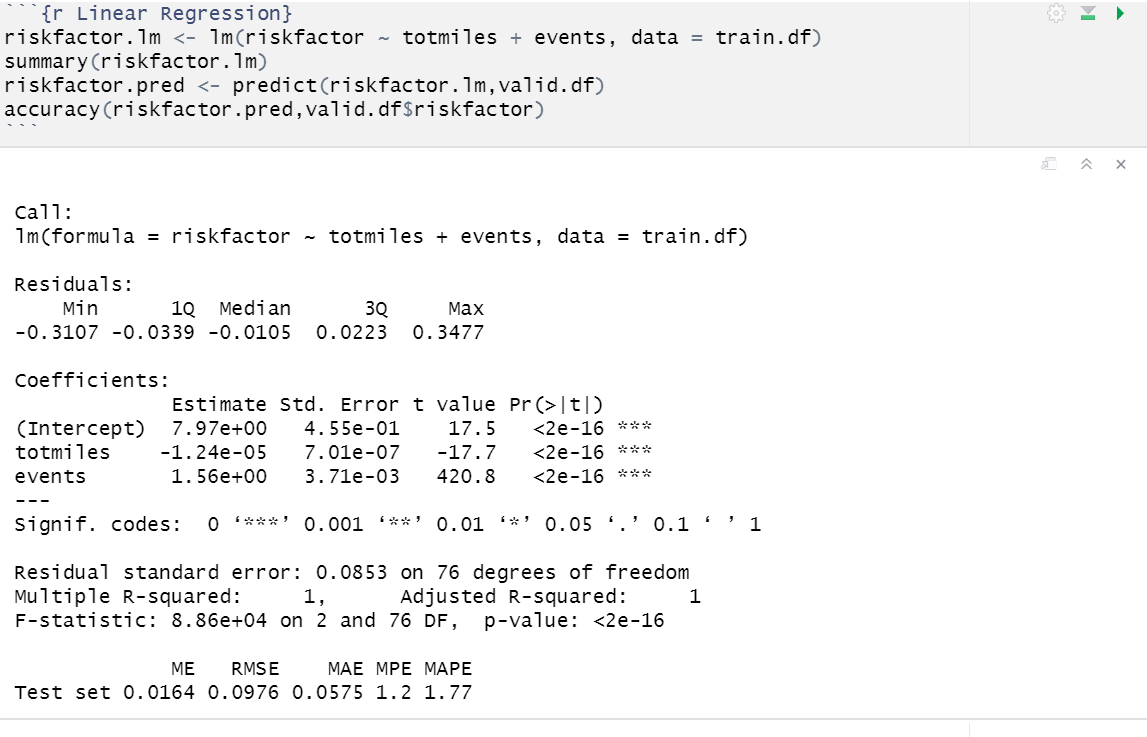


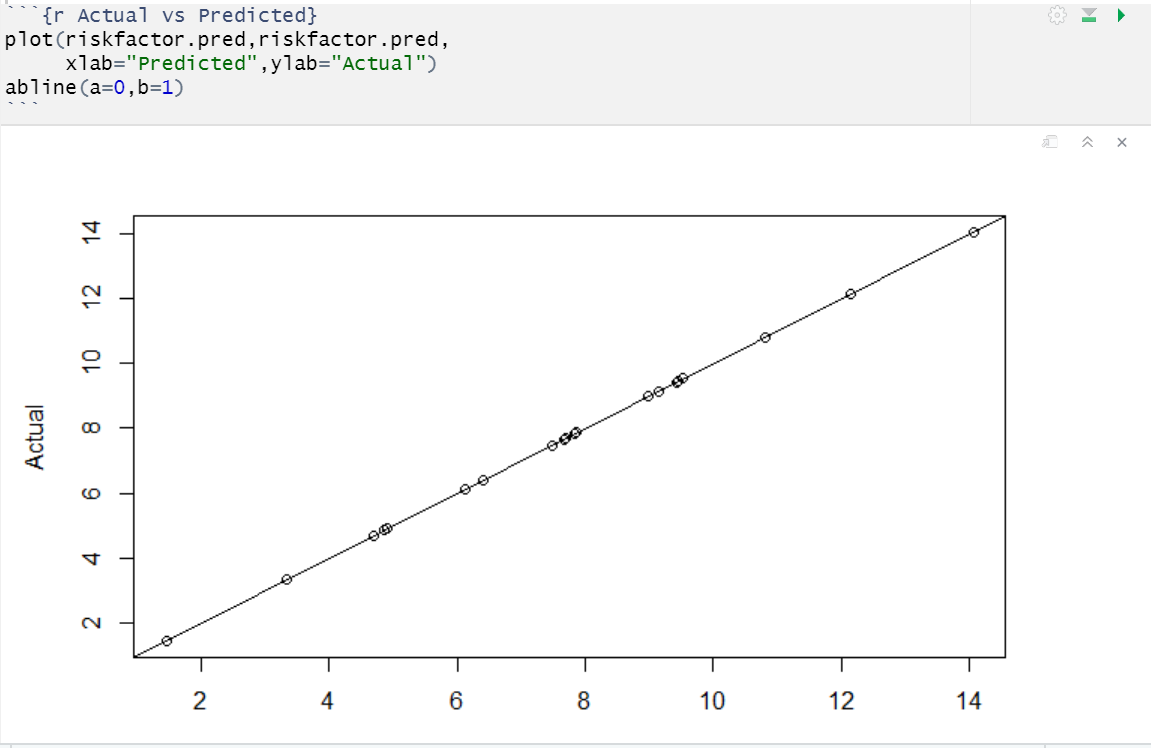


This multiple linear regression model is trained using the train dataset i.e. by using 80% of the total data. lm() function is used in R to create multiple linear regression model. The target variable i.e. the dependent variable is risk factor and the independent variables are total miles and events. Total miles and events columns are converted into numeric values before training the model. The trained model is then used to predict the values of risk factor using the validation dataset. We have then calculated the accuracy of our multiple linear regression model using the accuracy() function in R.Also, we can test the accuracy by using graphs, if the model fits save the model in .rds format in local (This needs to be referenced by the Tableau).

**Checking accuracy in R**

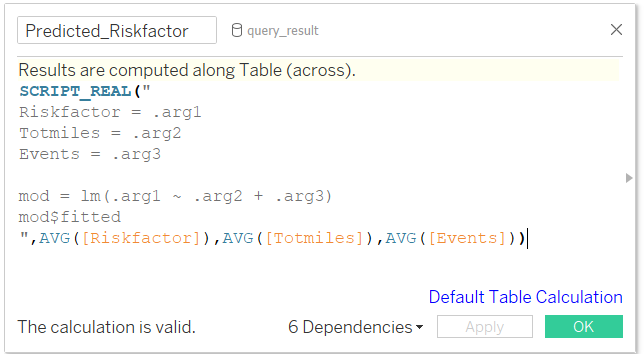
Summary() function is used to check the accuracy and all other realted values of the model in R





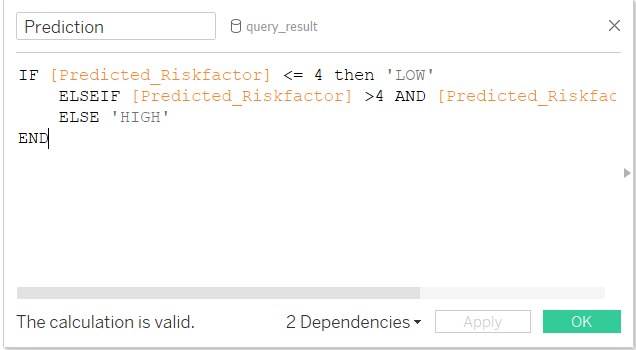
**Creating calculated fields in Tableau:**

R is integrated into the calculated fields in Tableau. In order to call R we can use SCRIPT\_REAL calculation. SCRIPT\_REAL means the output is a real number.



Here values from column Risk factor is passed into arg1, similarly from Totmiles in arg2 and from Events in arg3.

**Predicting Riskfactor :**



Here we are predicting riskfactor on the basis of following assumptions:

* if predicted riskfactor is less than or equal to 4 then risk factor is low.
* if predicted riskfactor is greater than 4 and less than 7 then risk factor is medium.
* if predicted riskfactor is greater than or equal to 7 then risk factor is high.



**Creation of Parameter RISKFACTOR\_STATUS and Dynamic chart:**

We have made a dynamic chart using the above calculated field i.e. Predicted\_Riskfactor and Prediction, and presented the insert and drop down filter through which we can select or enter driver’s id and can retrieve whether that driver id come under which category(Low, Medium,high).

