Detailed Summary of R work:

❖ USA (2014,2015)

1) The libraries that we needed – MASS, ISLR, xlsx, car, cluster - are installed and called.
2) We read the csv file for 2014, 2015 and 2016.
3) We removed the NA values from both the files.
4) The respective data is converted into data-frame.
5) Then a new column named aggravated_assault_2016 is added in data for 2014 for the prediction of mean of violent crime per 10.000 of population for 2016.
6) Then another column named m2014 is added in data for 2014 in which mean value of violent crime of 2014 is calculate per 10,000 of population.
7) Then multiple variable linear regression is done for m2014 where m2014 is dependent variable and all types of violent crime are independent variables.
   - We noticed that p value for 'Murder and slaughter' is very high as compared to other variables which means it does not fit the data to a large extent so we removed it from the fit and again did linear regression 'lm.fit14'.
8) We find the confidence interval for estimate of coefficients.
9) Then we predicted the value of m2014 for all values of aggravated assault of 2014.
10) Then we predicted the value of m2014 for those value of aggravated assault which are in 2016 which gives the approximated mean violent crime per 10,000 of population for 2016.
    - We observed the trends in 2014 and 2015 and concluded that:
      - m2014 and m2015 are almost same as analyzed by null-hypothesis so m2016 will also be approximately same.
      - By multiple linear regression, we found that the trends in p-value for both 2014 and 2015 are almost same so same can be predicted for 2016 and in both the years, aggravated assault value has the least p-value so we used the value of aggravated assault of 2016 and predicted m2016 by predicting m2014 for aggravated assault of 2016.
      - We even observed that they are approximately same as m2014.
11) We plotted m2014 vs aggravated assault of 2014. We also did abline plot and we observed that plot is done for 3 variables as we can see that most data is concentrated in starting only so rest points can be considered as outliers.
12) We plotted hatvalues for lm.fit14 which gives leverage statistics.
13) Then index of maximum variable is found.
14) Same is done for 2015 but prediction for m2016 is done only by 2014 data.
15) This is followed by performing null hypothesis which was that m2014 and m2015 are equal because if they are equal, then our base of predictions for 2016 is correct. We found that our null hypothesis is accepted and hence out predictions.

❖ LA

1) We read the csv files for LA and separately for 2014 and 2015.
2) Then the data is converted to data-frame.
3) Generalized logistic regression is done because we have some non-numeric values.

- Crime.Code.Descent is the dependent variable and Victim sex, victim age, victim descent, time occurred, Latitude and longitude are the independent variables.
- We observed that p value for Victim's age is very low which shows that it is a major contributor in glm

4) We found coefficients for the glm fit. It is done to see the effect of negative and positive parameters. Positive parameters are directly proportional to dependent variables while negative is inversely proportional.
5) Then we calculated probabilities that a particular crime will happen for given values of independent variables.
6) The same is done for the data of LA for 2015.