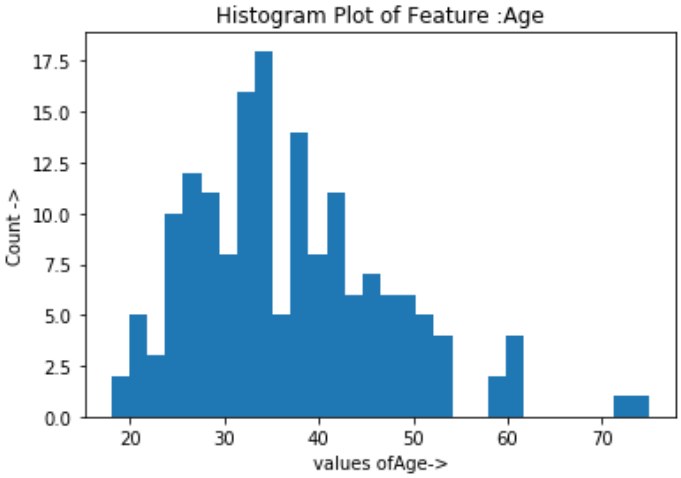
Report:Task B

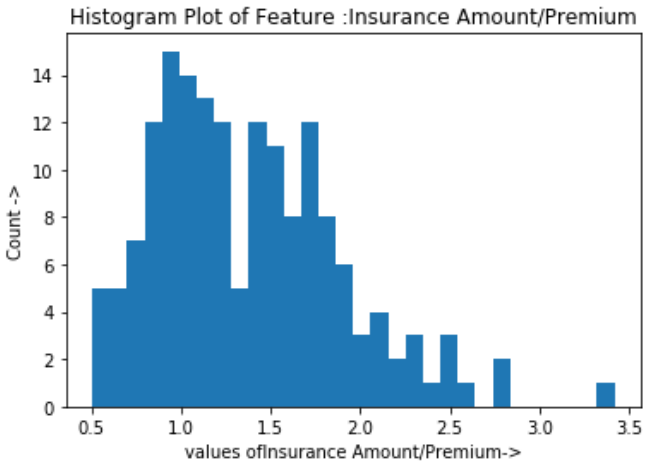
Given Objective: Predict the Loss Amount for insurance polices using historical trends and features.

Sub Task1: Create a sample data set
A car insurance based sample data set has been created with following features:

- Allstate claims data set on Kaggle was used as reference for creating a sample data set.(ref: https://www.kaggle.com/c/allstate-claims-severity/data)
- Features used for the sample data set creation included:
  1. Age of the owner/Driver: As the title suggests, it is the age of person.
  2. Car Model: This indicates if is small/medium/high range car encoded as (1/2/3).
  3. Car Age: Duration since the car was purchased (0 to 10 years)
  4. Car Cost: Amount paid to get the ownership
  5. Insurance amount: Insurance amount paid
  6. Accident Severity : 1 to 5 depending of the survey assessment with 1 being less severe and 5 being heavy damage.
  7. Loss Amount : Loss amount was computed on case by case basis to incorporate human error and noise.
- All the features from 1 to 6 were generated using random distributions like Rayleigh distribution or uniform distribution.
- Around 200 data samples were created and around 166 of them were annotated. The remaining samples were discarded.

The histogram of few features is as shown below:





The rest of the histograms and code for generation can be found at:

Prediction Model:

- Linear Regression has been used, as it is the simplest possible prediction model which can be tested easily
- The data is normalized with respect each features to avoid unwanted outputs.
- Regression Fitting is done at the beginning to obtain a linear equation.
- 90 sample points are used for training and 75 are used for testing.
- Sklearn library has been used to incorporate Linear Regression Model

The code and the details of execution is available at :

Results:

- It can be seen from below sample table that the linear regression model developed can predict results accurately up to first digit.
- The above observation has been quantified by using L1 norm which shows that the error in the first digit is zero for most of the cases while the second decimal digit accuracy varies form 0.2 to 0.9
- The accuracy measure provided by scikit learn indicates a score result between 0.69 to 0.75 which tells that the accuracy of each prediction is high as 1 is regarded as perfect score.

```
   Actual Loss Amount  Predicted Loss Amount   L1 Norm
0            0.060441               0.086051  0.025610
1            0.613878               0.587146  0.026732
2            0.257143               0.293654  0.036511
3            0.259331               0.352031  0.092700
4            0.522449               0.433201  0.089248
Mean L1 Norm for 75 test samples
0.08879388490608407
Standard Deviation of L1 Norm for 75 test samples
0.06362811957056849
```

Future Work:

- The fit score from scikit learn indicates a good score for a simplest possible linear regression model which provides from for further exploration in terms of features to get better results.
- Use of complex models like Neural Network or Reinforcement Learning based models may yield even better results.