Technocolabs Data Science Internship

Project Report

Traffic Mortality Reduction in USA states

By Ishan Sarkar

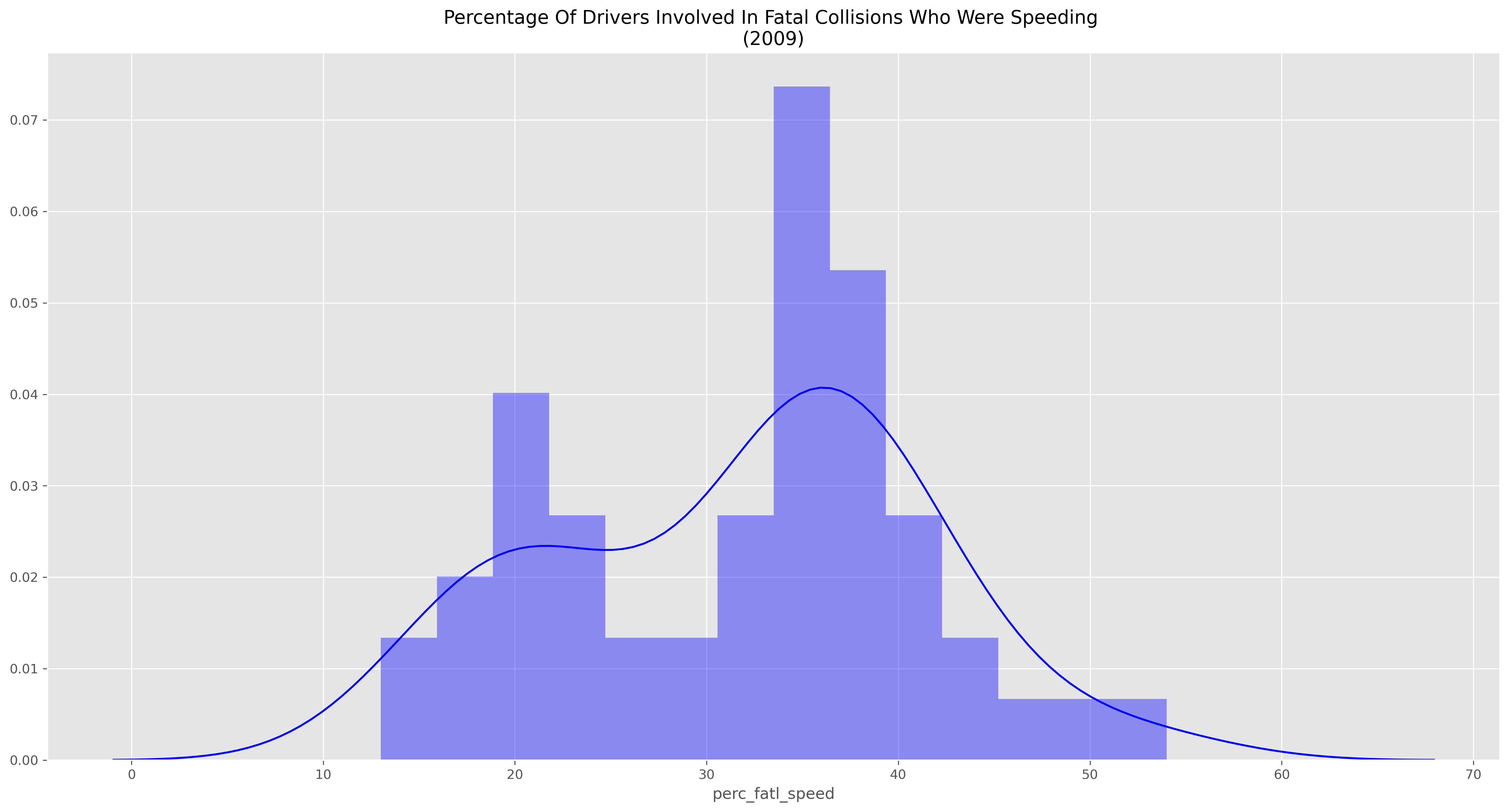
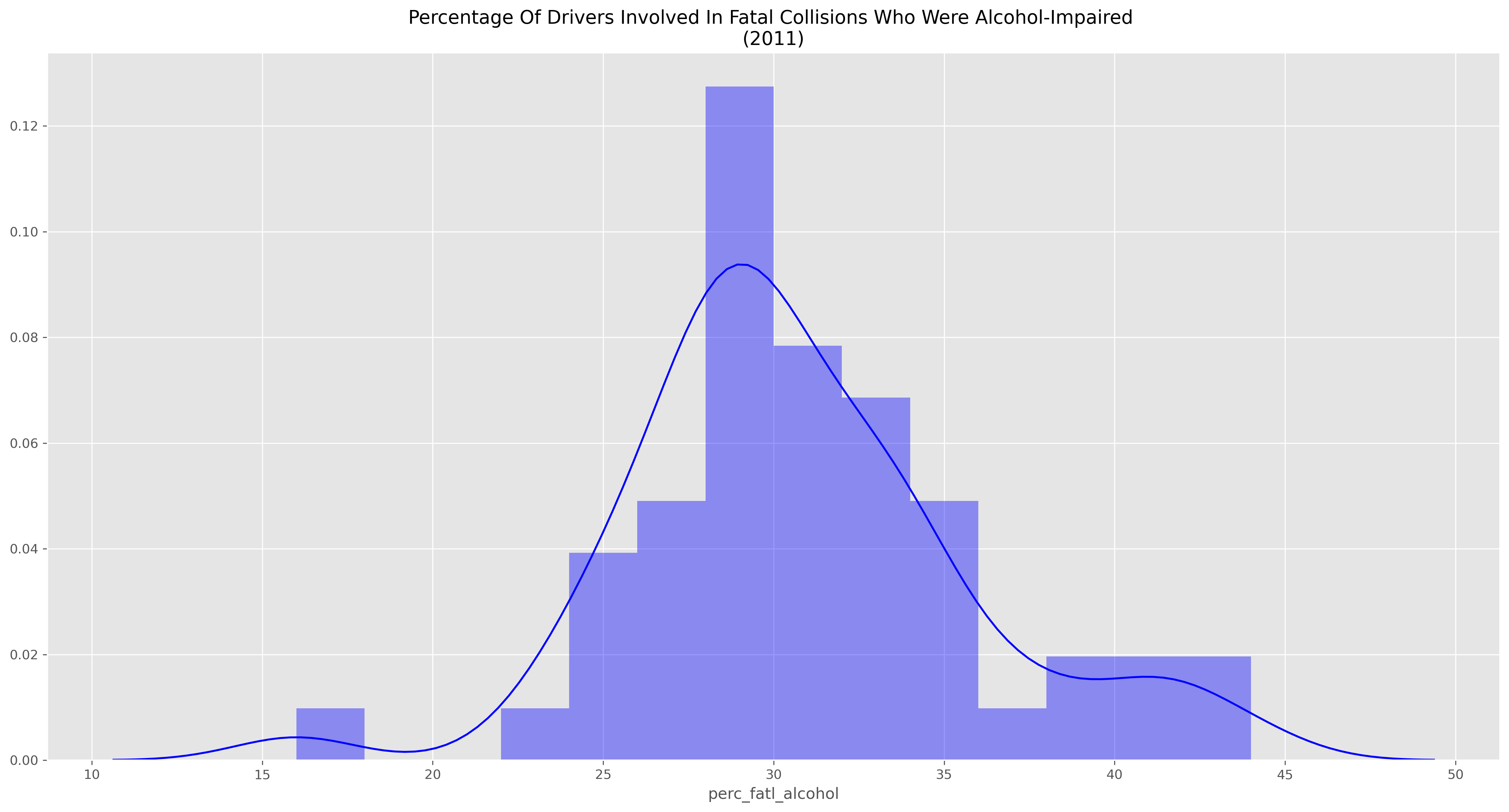
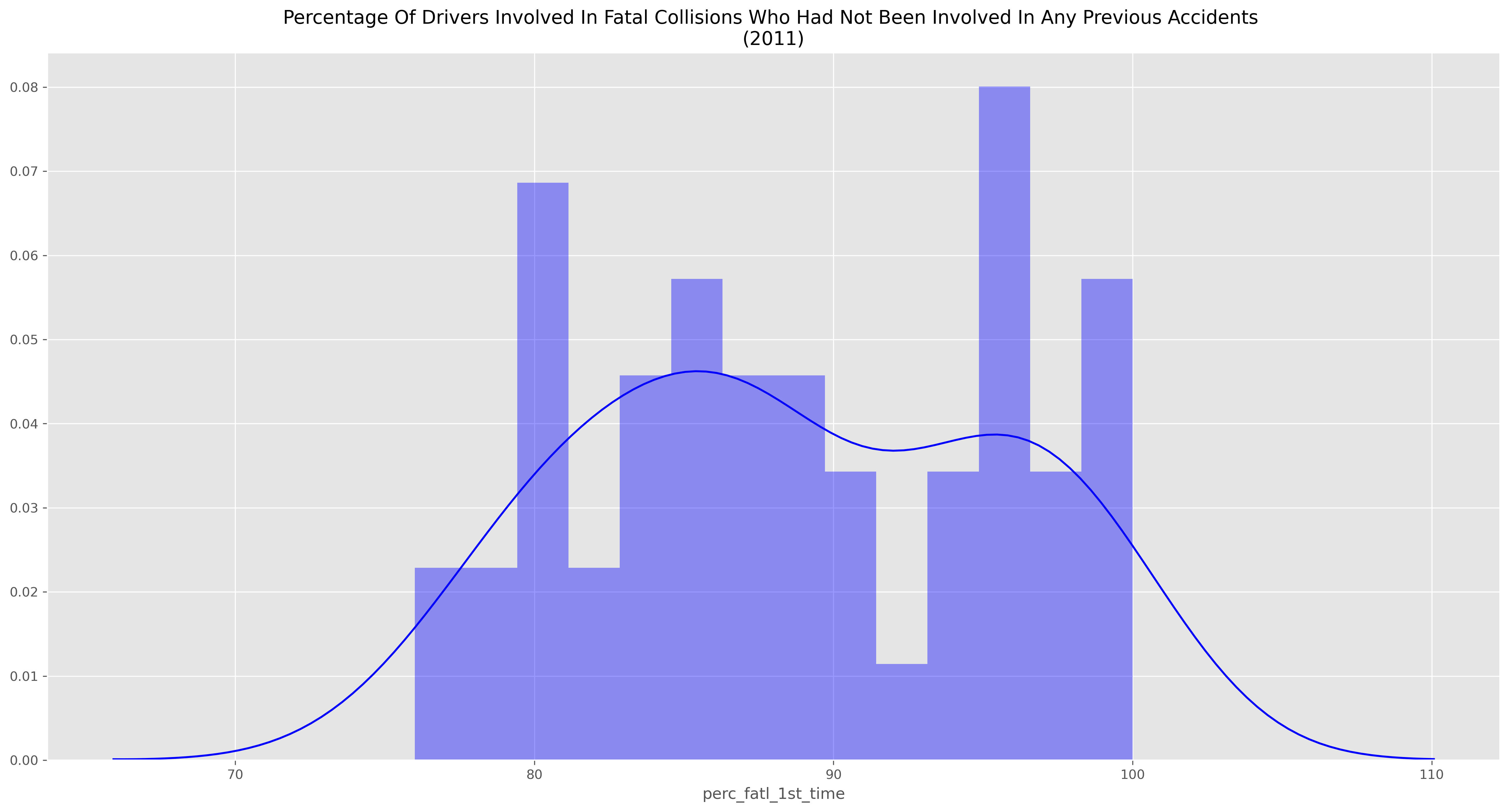
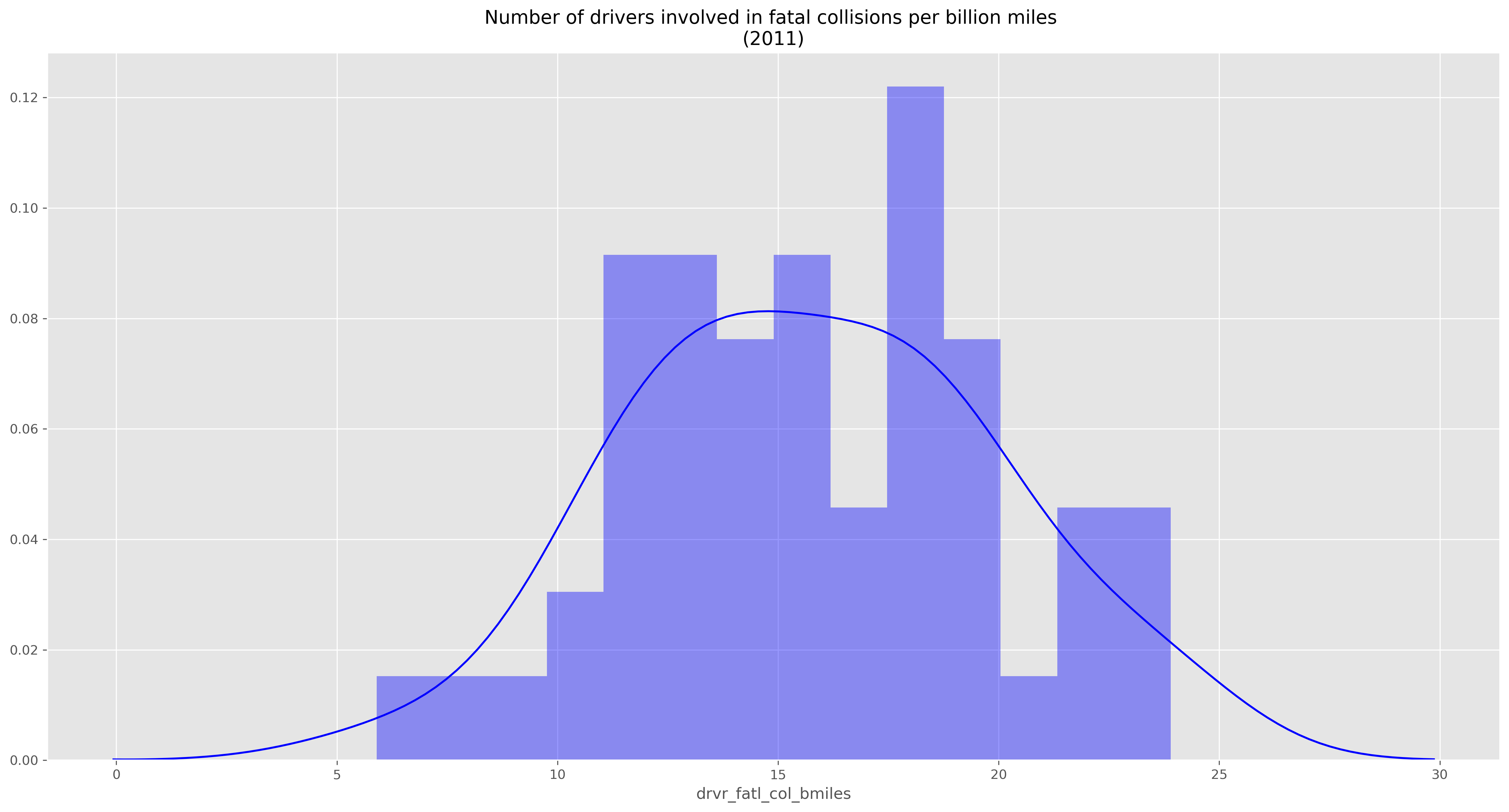
UG 3rd Year, Computer Science

[ishanss2018@gmail.com](mailto:ishanss2018@gmail.com)

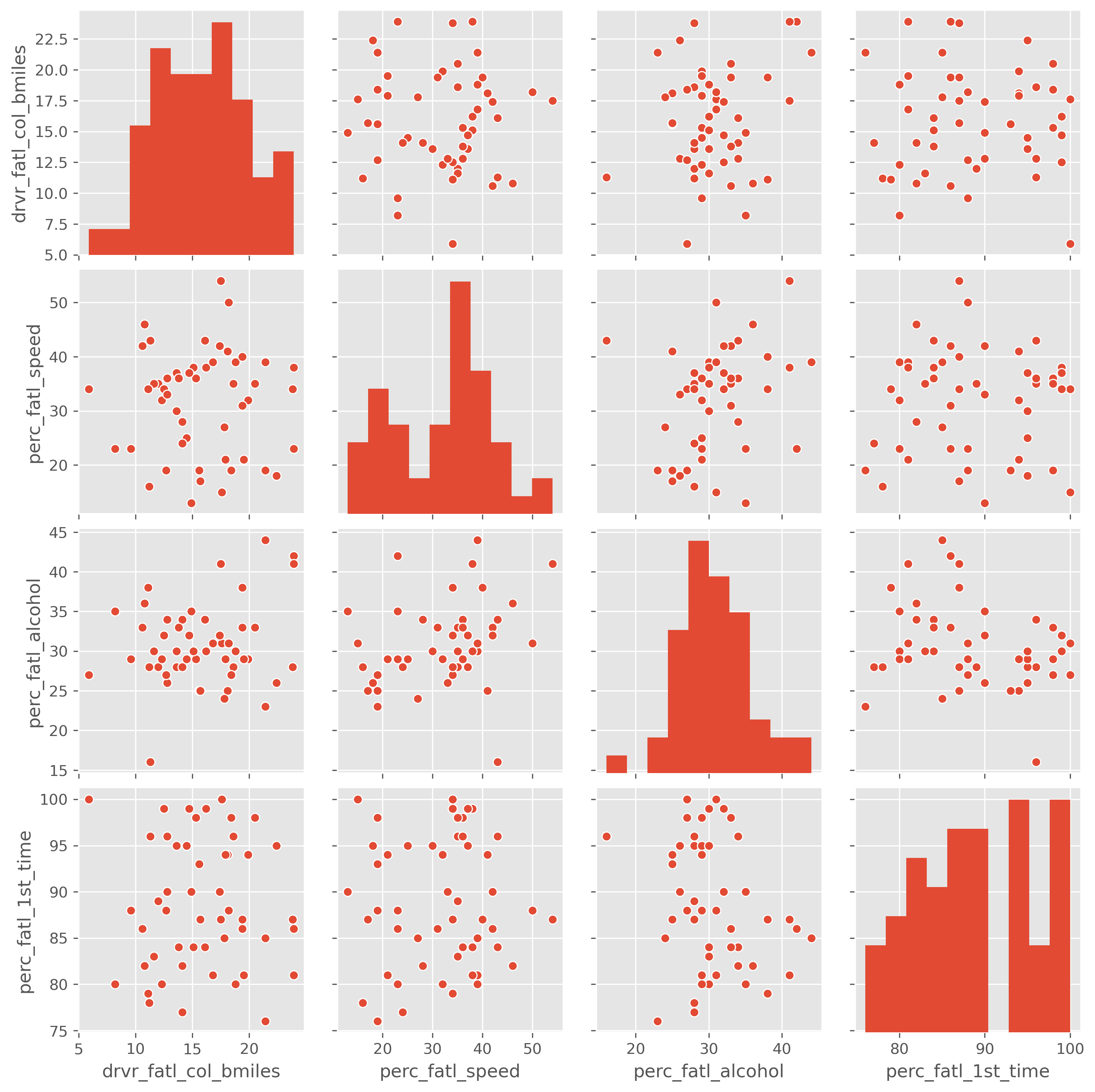
Delhi, India

The aim of the project was to analyze the given dataset containing percentage of fatalities collisions on road in each state of USA. Also data pertaining to percentage of cases due to alcohol consumption, speeding and first timers in accident were provided. Eventually it was required to determine which states required more attention in terms of policies.

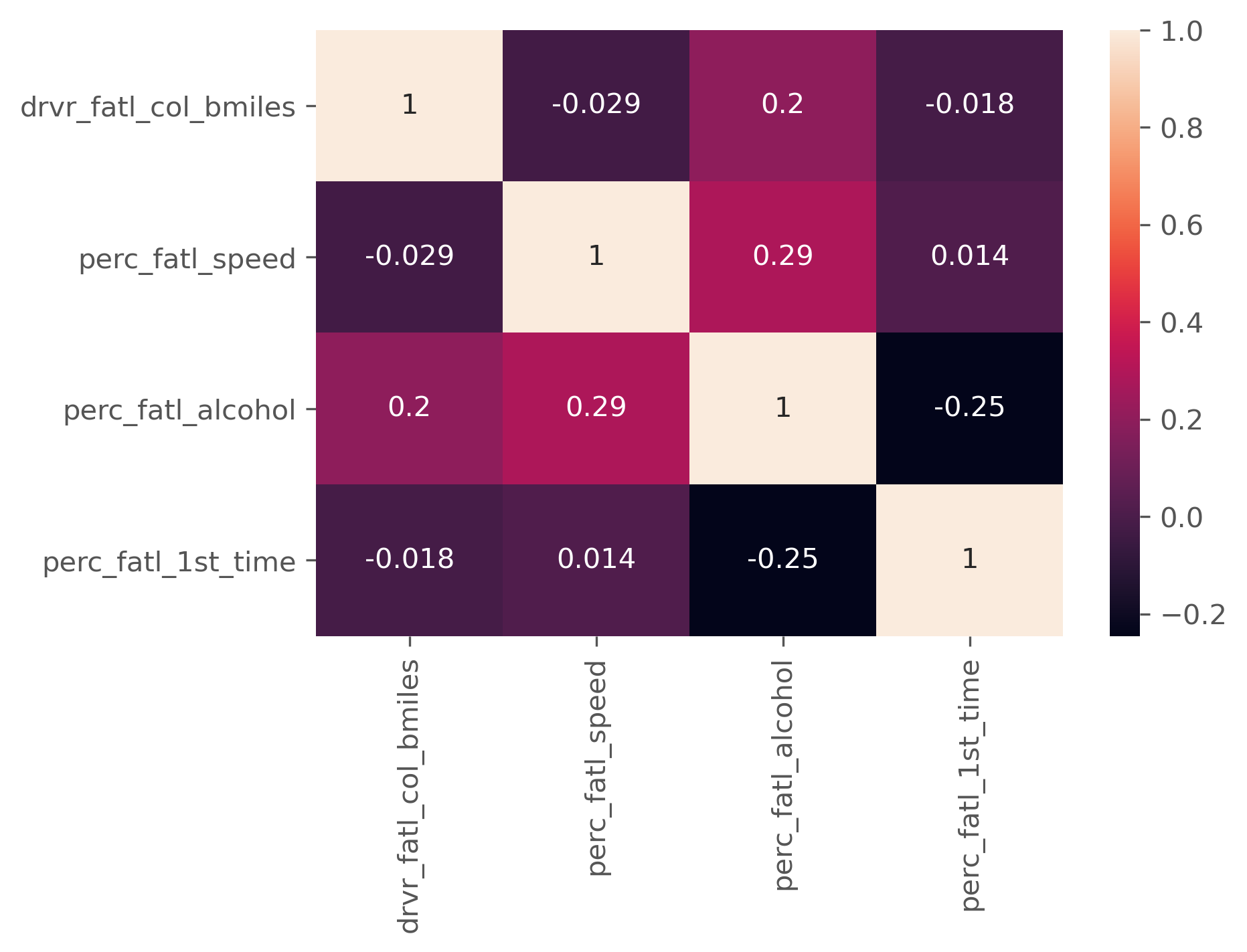
First of all the data was cleaned by removing the description text and organizing it into a dataframe. Basic statistical information for each feature was fetched. The values didn’t give much observable idea about the data. Therefore visualizations using graphs were made. Most of the features showed close to Normal Distribution behavior. The skewness was handled by Scaling to normalize the values to get better convergence of algorithms.



Pairplots were generated to check the correlation between the features visually.



All feature pairs showed up to weak correlation in the pairplot. Therefore numerical correlation values using Pearson Correlation coefficient were calculated.



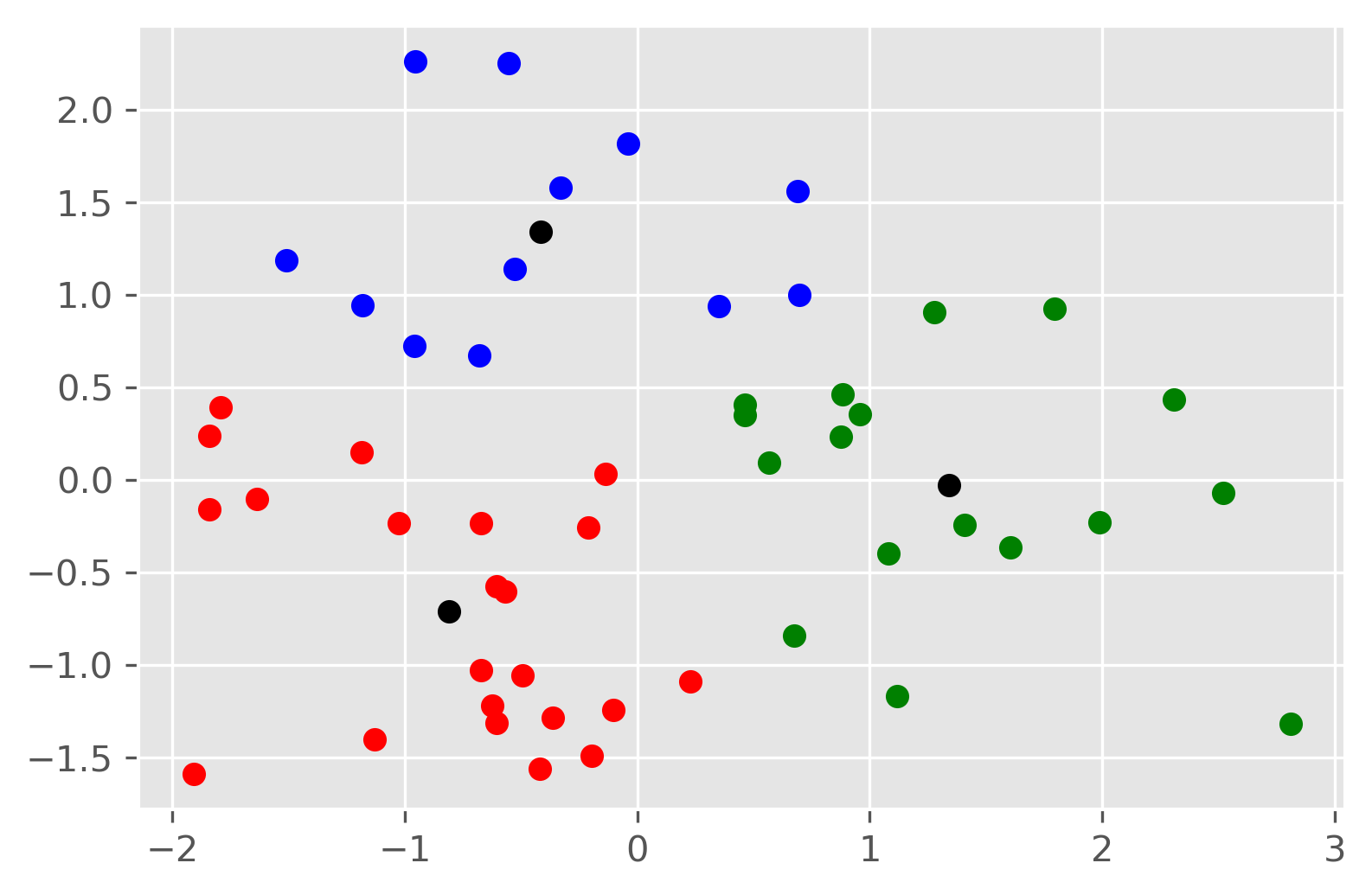
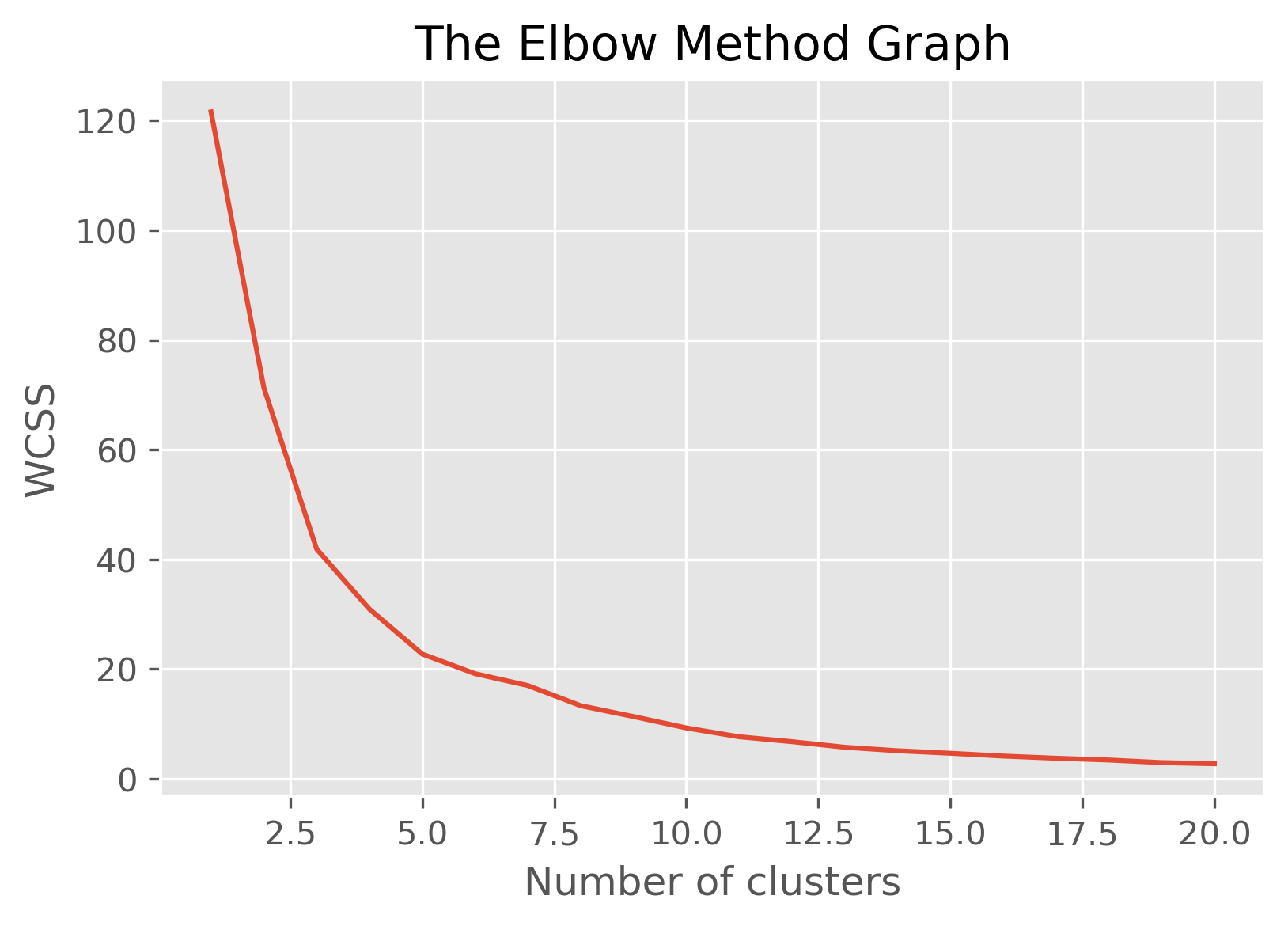
**Most feature pairs show very weak correlation. But the following show weak correlation-**

1. **Fatality due to alcohol and Target**
2. **Fatality due to alcohol and fatality due to speeding**
3. **Fatality due to alcohol and fatality who were not involved in any accident before**

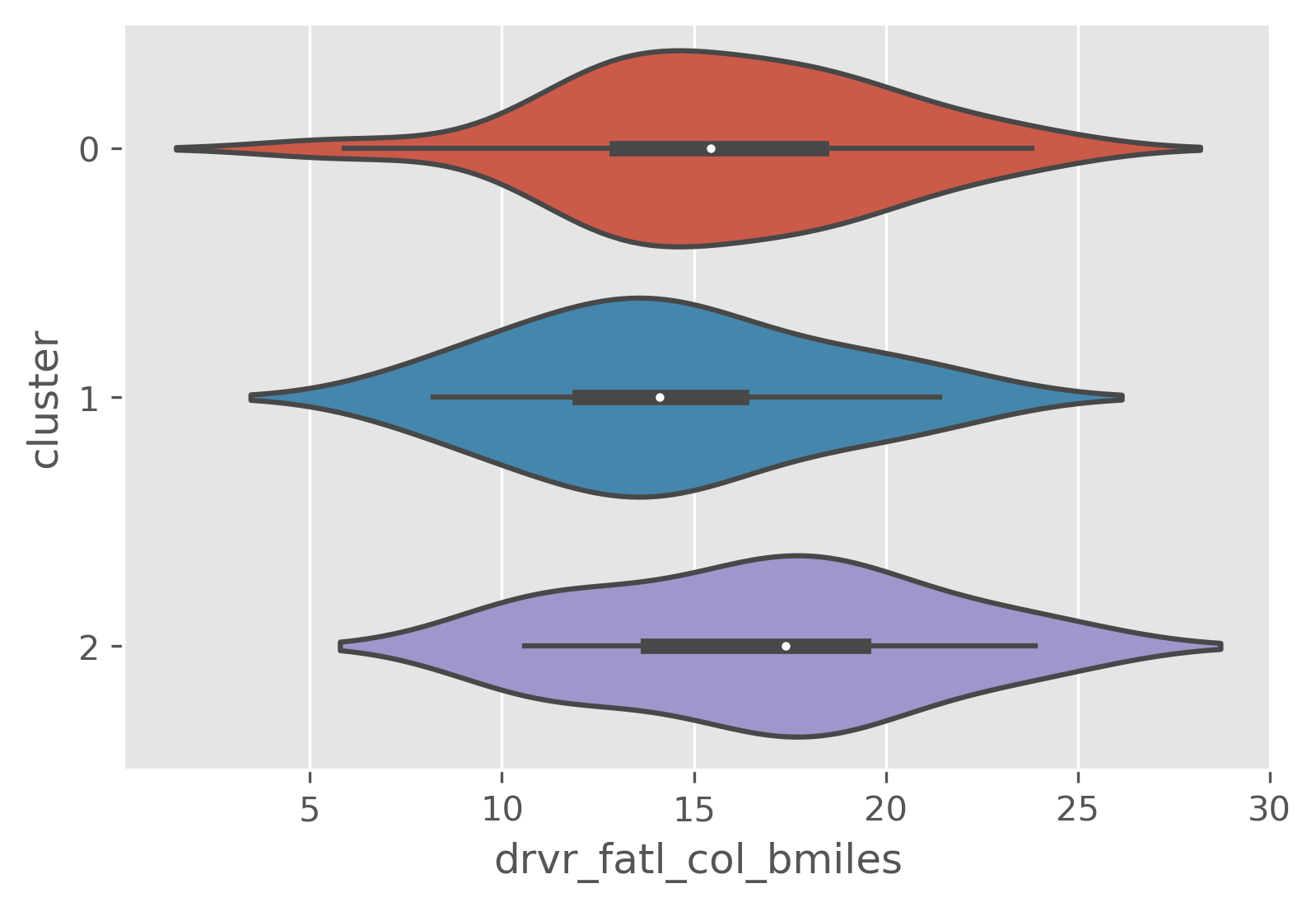
### Next was to determine patterns using Unsupervised Machine Learning

Fatalities per billion miles were set as target variable and the rest features were used as independent features.

2 components were selected out of the 3 using Principal Component Analysis. K-Means Clustering algorithm was used for clustering the datapoints generated into groups. Optimum number of clusters was found using the Elbow Method which turned out to be 3. Clusters were visualized on scatterplot.

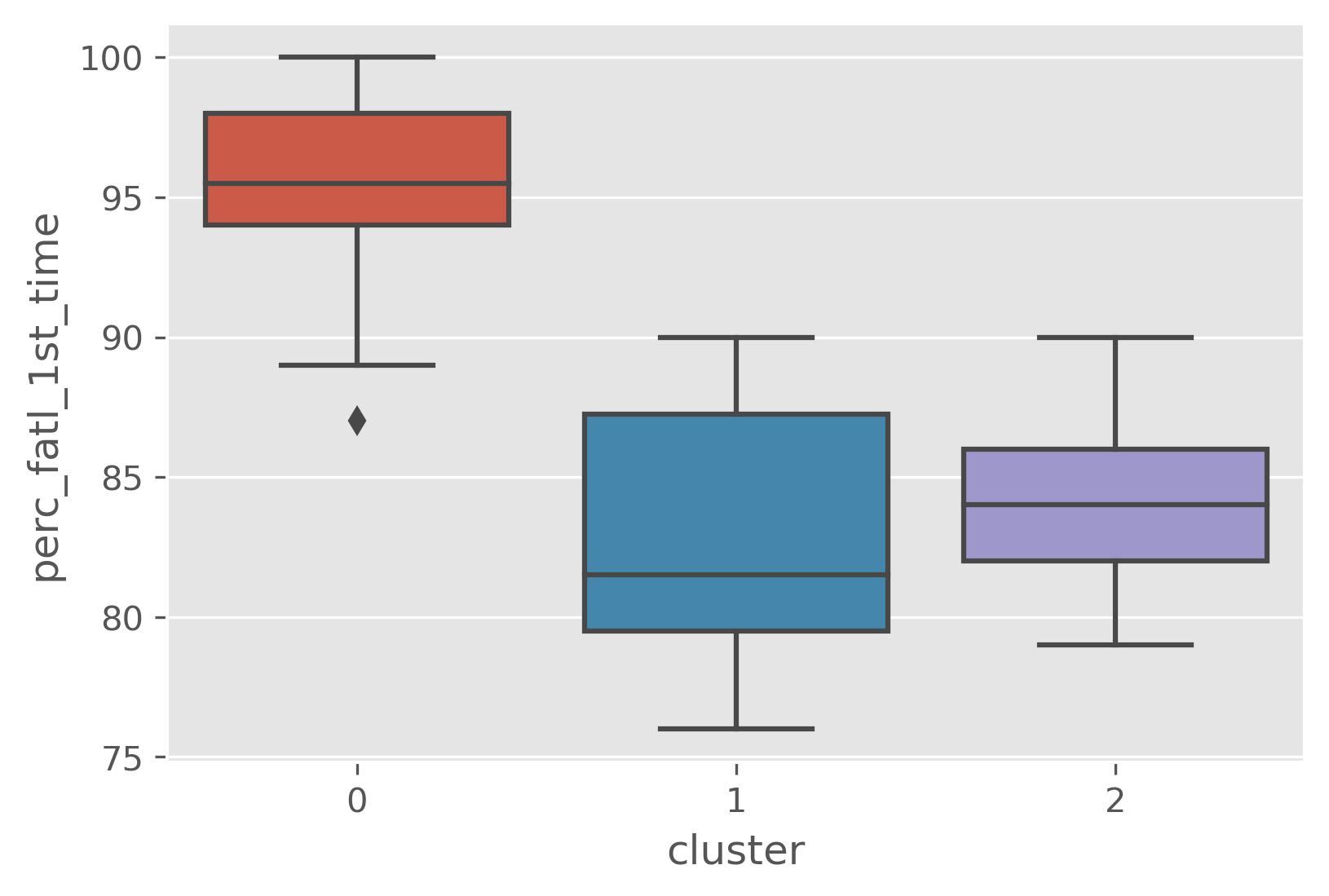
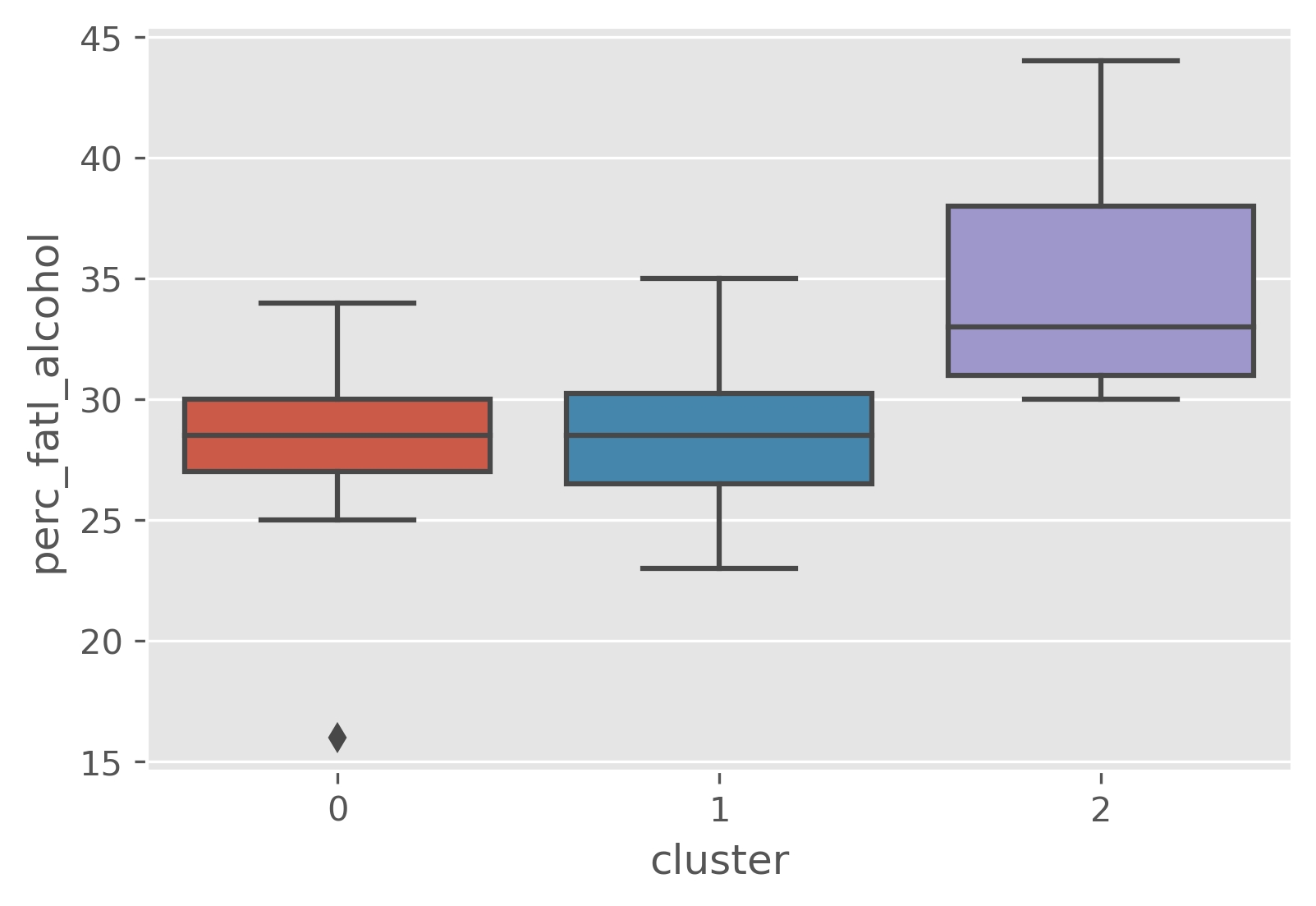
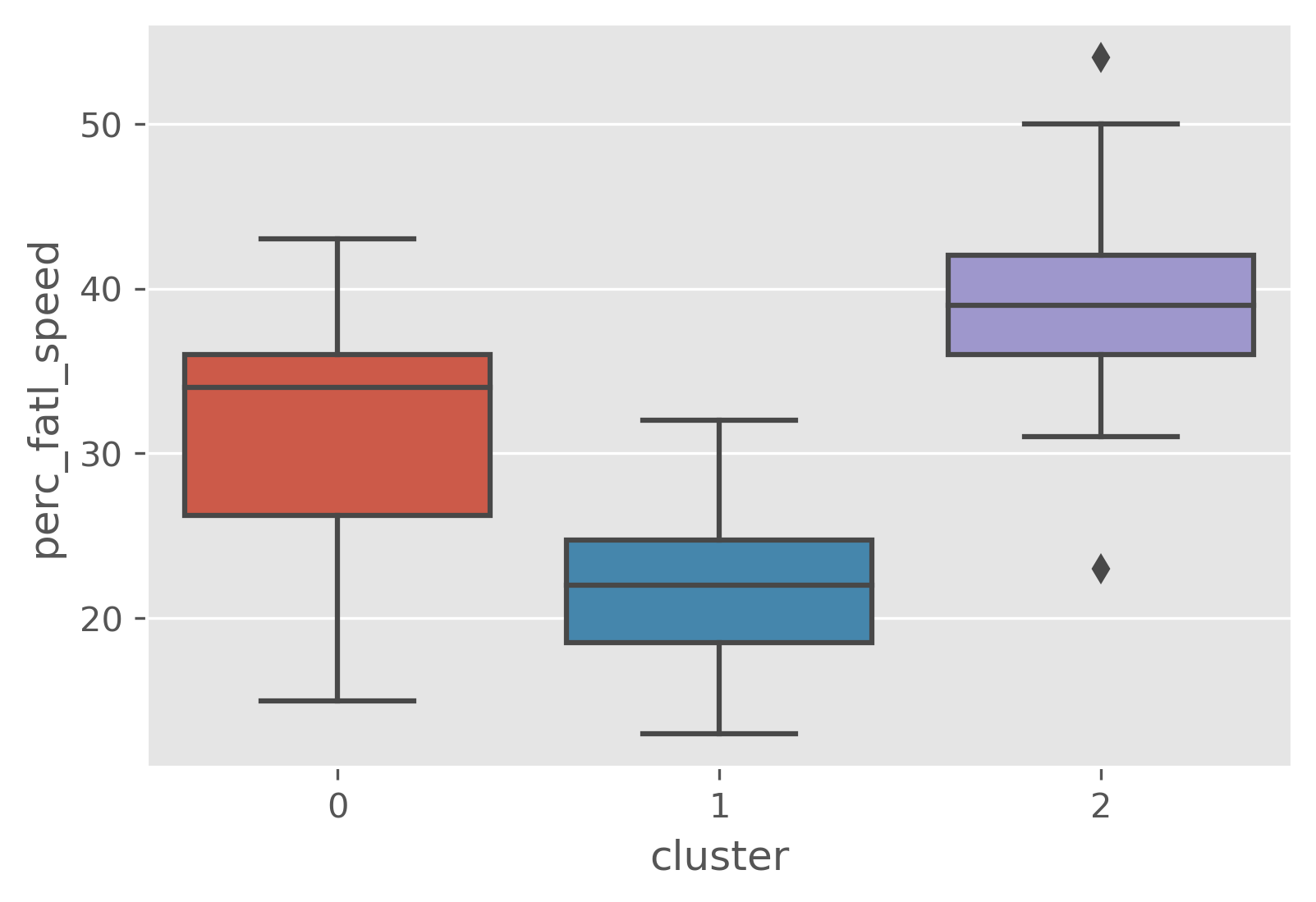


The datapoints (States) were mapped to the cluster they belong to and dataframe was updated with a cluster column.



Fatalities per billion miles don't show much variation in the probability distribution as depicted by the violin plot above. Therefore the 3 independent features grouped by clusters and visualized using boxplots.

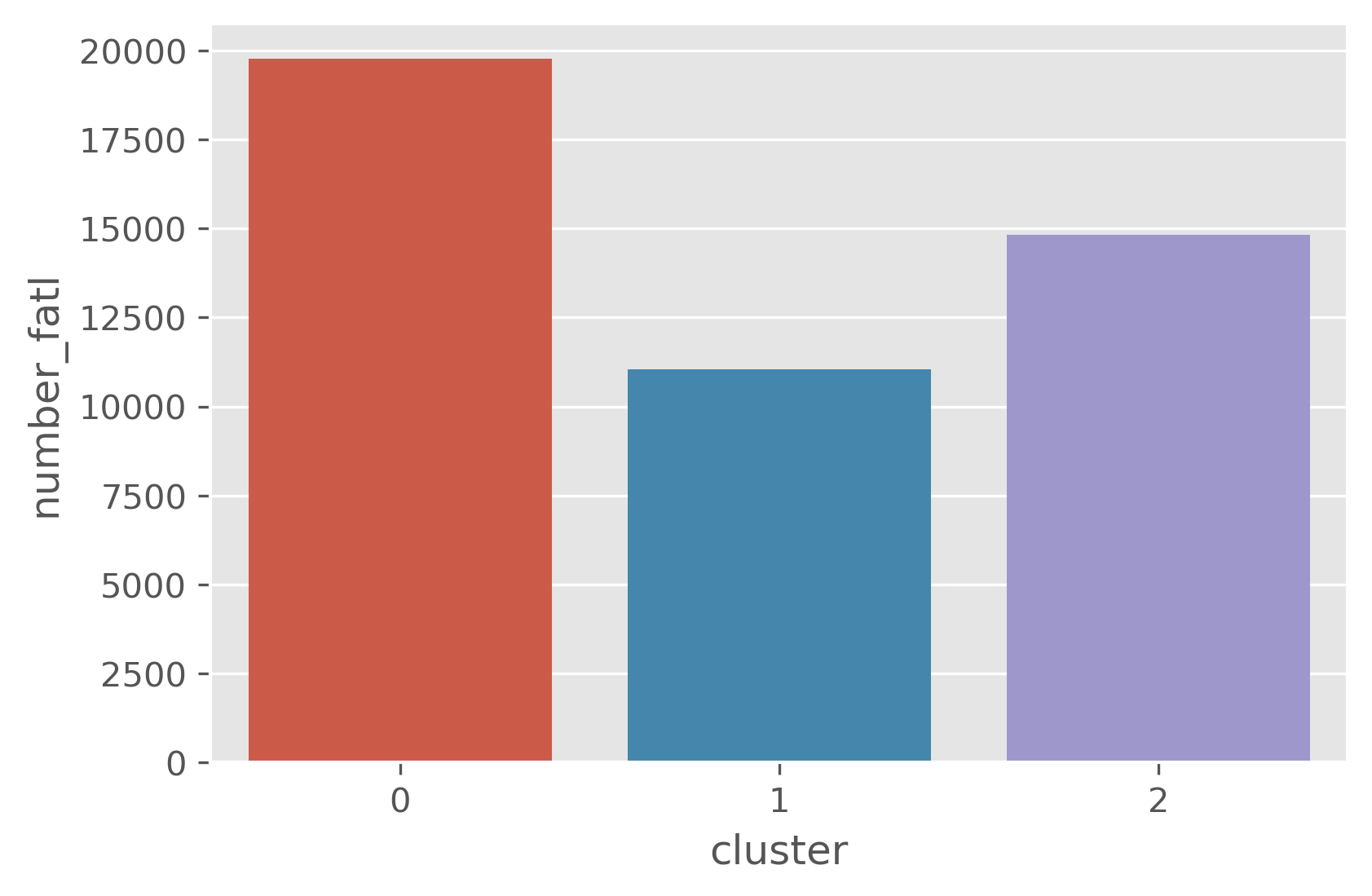
All 3 features showed variation in frequency distribution when grouped by clusters separately.



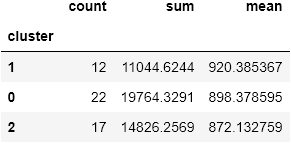
**Results-**

* **Cluster 2 requires less attention to speeding policies as compared to others clusters**
* **All the clusters require nearly equal attention to policies regarding the alcohol consumption while driving**
* **Cluster 0 shows much more fatality for first time drivers indicating to the need to establish policies to monitor first timers on road**

Next, the total number of fatalities was calculated using the miles.csv data.



**Cluster 0 shows highest number of fatalities but cluster size should be considered too for proper depiction of results. Therefore calculating mean number of fatalities in each cluster.**



#### Therefore it can be implied from the above analysis that Cluster 1 needs highest attention in terms policies to tackle traffic mortality. The following states fall under this cluster-

* **Iowa**
* **Kansas**
* **Kentucky**
* **Massachusetts**
* **Michigan**
* **Minnesota**
* **Nebraska**
* **New Jersey**
* **New York**
* **Ohio**
* **Tennessee**
* **Virginia**

**The administration can now move towards policy formulation according to the results analyzed above.**