**Utah Accidents**

Vehicular accidents are unfortunately part of everyone’s reality. Our team sought to use prediction analysis with data that we were able to retrieve, on vehicular accidents. Overall, we all feel that we were very successful with this project and were able to work together to find some of the answers. The largest challenge with this topic proved to be the variety of options and variables we could use in our analysis. There are so many ways to analyze and use the data we had; it was hard just to focus on a few. Our team found the topic to be very interesting and developed a passion for the answer.

The Process Originally we had started with a dataset from https://opendata.utah.gov/Transportation/Crash-data-for-Utah-2014-2018/a64b-mcum/data. This data set was good because it had a large range of dates, many rows, and many variables to work with. Some of the variables were just noise, and after receiving some feedback from peers we realized that driver information would be very valuable for the analysis. Since our main focus was going to be using bad weather occurrences as a predictor for accidents, we also wanted to bring in data from www.NOAA.gov for occurrences of adverse weather. We were able to retrieve all the data from the sources we needed, but in addition to this we sought for more data from the DMV. We had hoped that the DMV would bring in vehicle and driver information that we could add. Although it took a lot longer to get this data (a week before the milestone was due), it proved to be better than everything we had. This new dataset from the DMV not only included driver information, but also detailed crash information, 39 variables, 65k + rows, weather occurrence information, etc. After we had the time to go through both data sets and clean everything, we decided it was best to continue with just the DMV crash data.

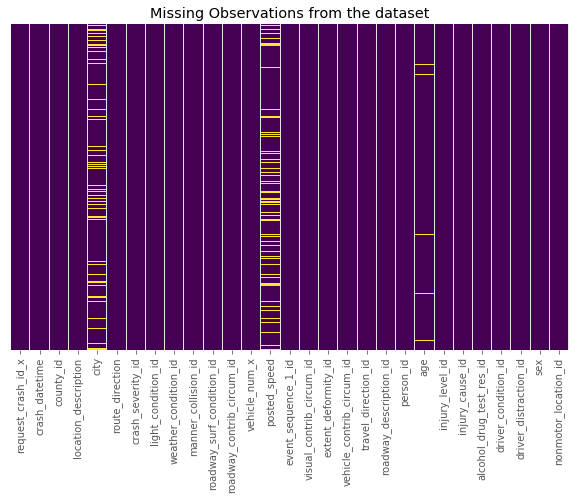
We were able to clean the data easily. This included getting rid of a lot of the blanks, unnecessary variables, and filtering the data for just the drivers. We used a data dictionary to find that the drivers were code one in the variable, so that made it easy to filter out. Age and counties were also variables that we took out some outliers to make the data a bit cleaner. We also decided to take out 99 in the weather condition columns because it stood for “Unknown” and had plenty of data to work with. While working with the data we found charts to be very helpful in seeing the trends of our data. This dataset had so many variables, this was helpful in organizing what we had.

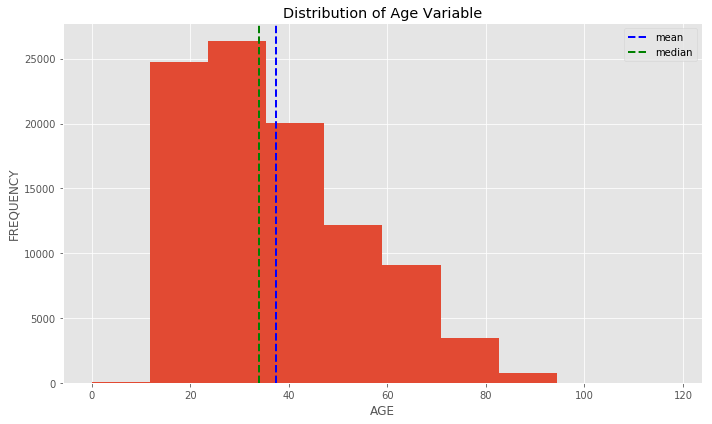
Ethics

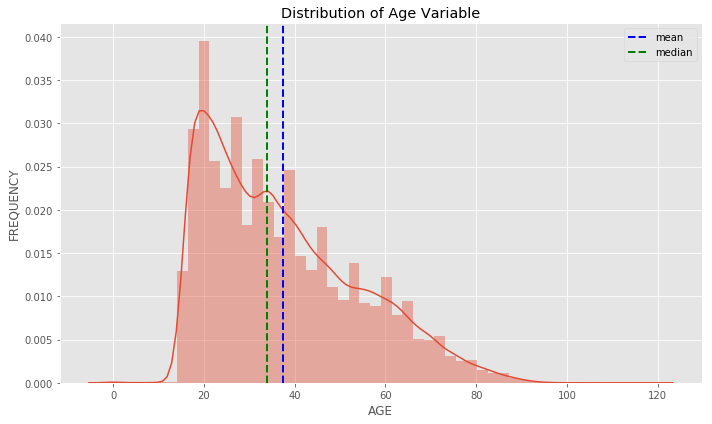
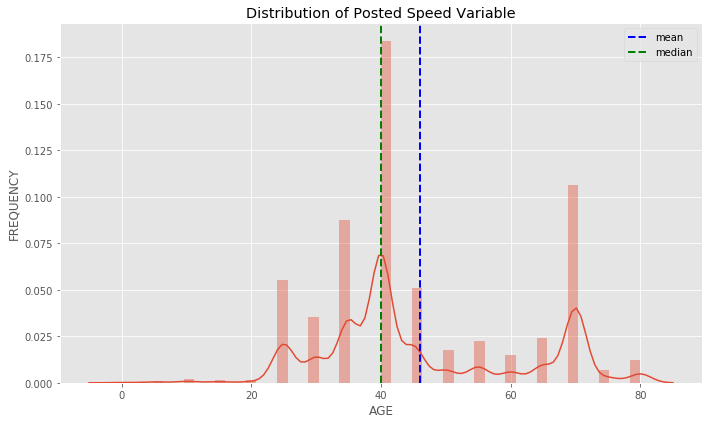
Something that proved to be difficult were the ethical considerations on cleaning the data. There were a couple of times we had to stop and consider if we filtered this out, would we be manipulating the data one way or another to make an argument. Or we asked, if we merge this, will we forgo necessary data points for the analysis? Our team discussed each one of these points thoroughly before making our decisions were successful in the executions.

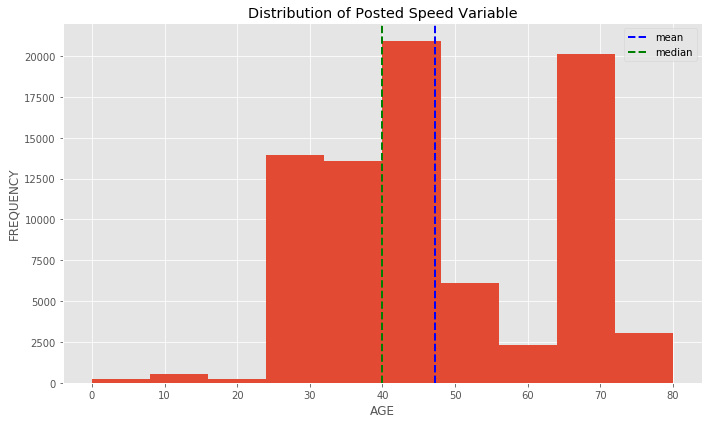
Summary of the findings:

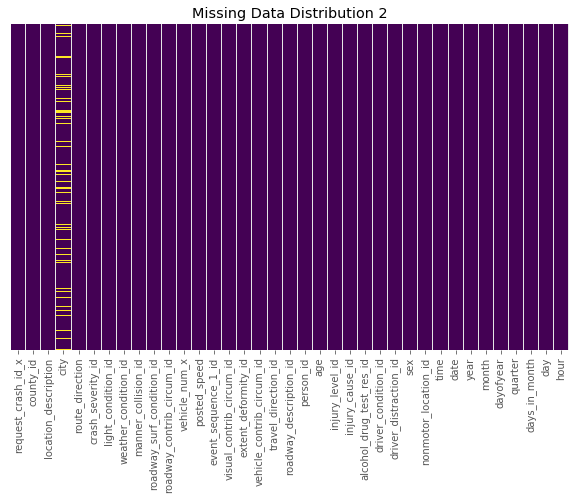
In the next few pages, we will have graphs of interesting observation of our dataset.





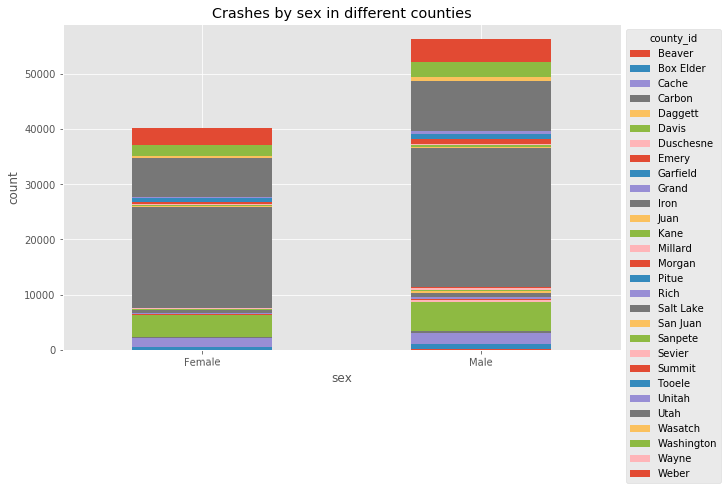


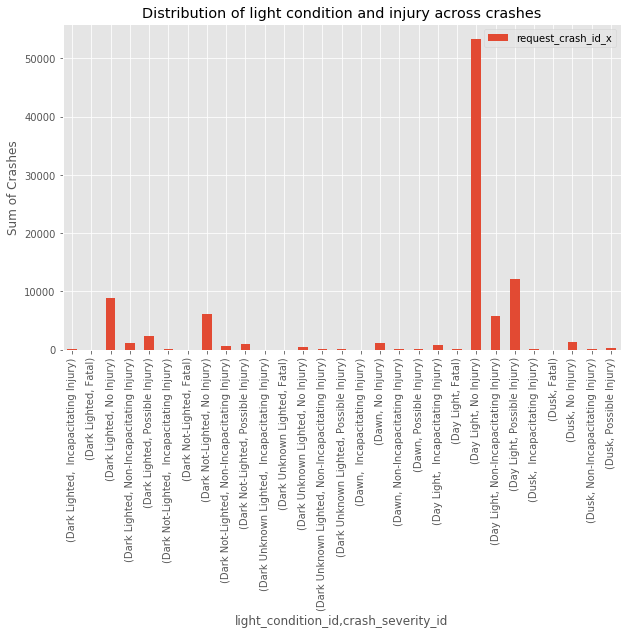




**Interpretation**:

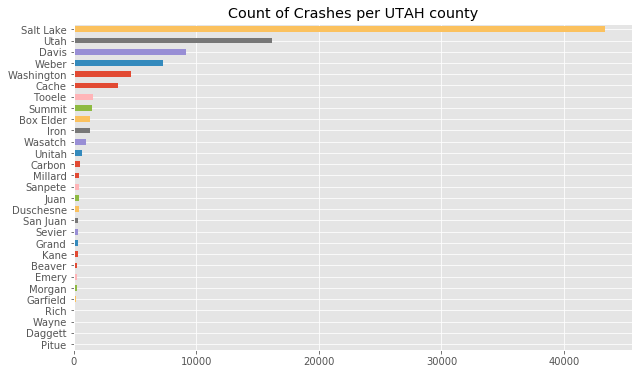
* Above are the steps we took to clean the data.
  1. Got rid of the NAs
  2. Removed the unwanted columns
  3. Age and posted speed Median imputation
  4. Removed Age feature outliers
  5. Dummitized the Sex Variable to numeric
  6. Removed Unkown observations cited for sex ['u'] and other categorical variables ['99']



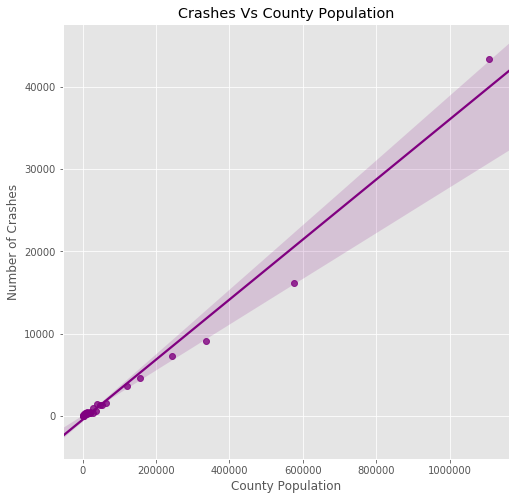
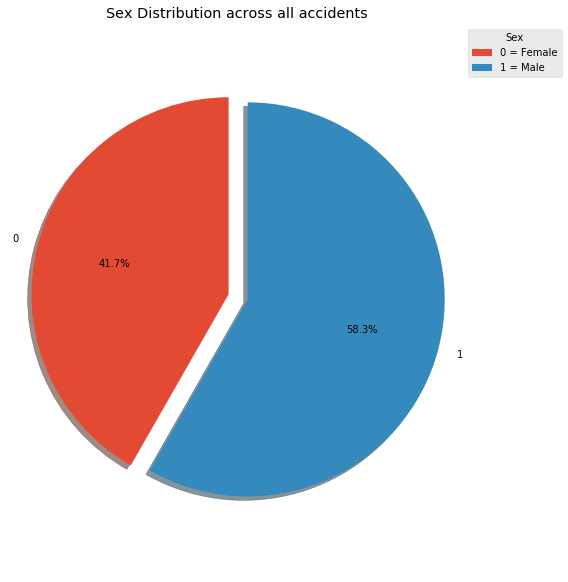


**Interpretation:**

* This graph shows the distribution of count of all crashes in different ranges of light with respect to severity of injuries per crash.
* As seen above most of the crashes with little/no injury occuer during Daylight.

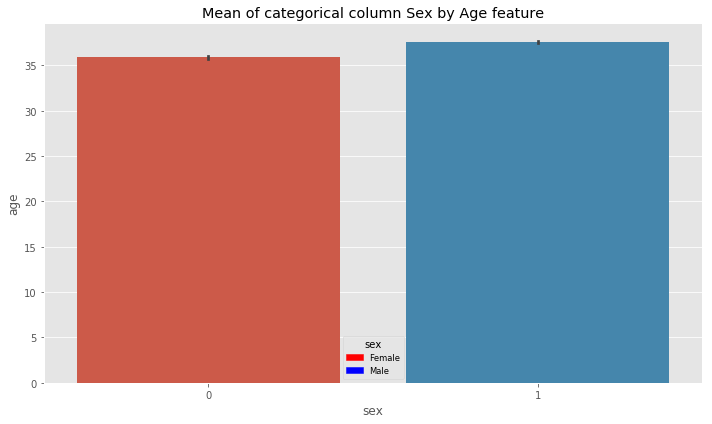


**Reading in Country Population data**

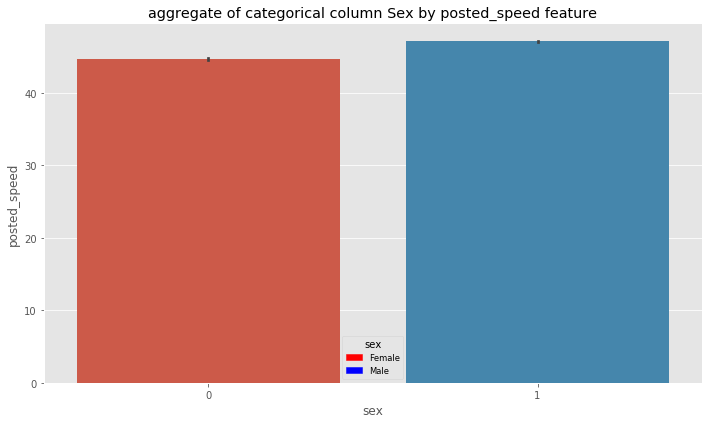
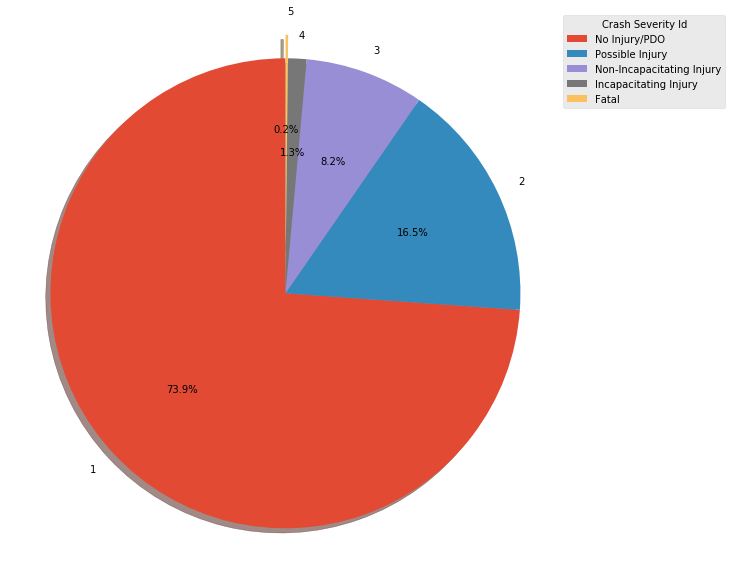
* Here we calculated the percentage of crashes with respect to each county's population.
* As seen below there is not a significant difference between counties.Therefore we can conclude [From the below linear regression graph] that there is linear relationship between the population and total of crashes across all UTAH counties
* 
* 

**Interpretation**

* Percentage of crashes among male and female drivers .

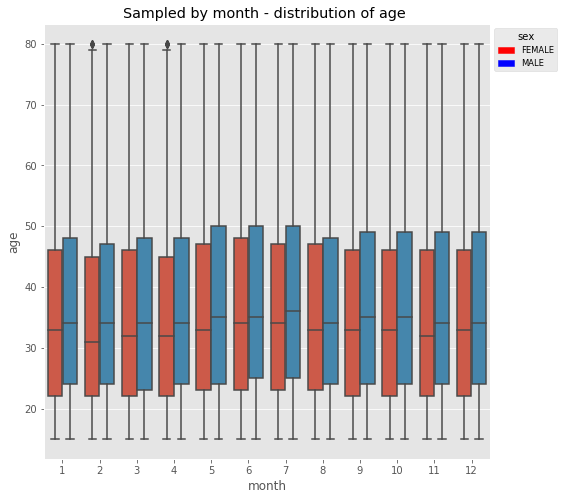


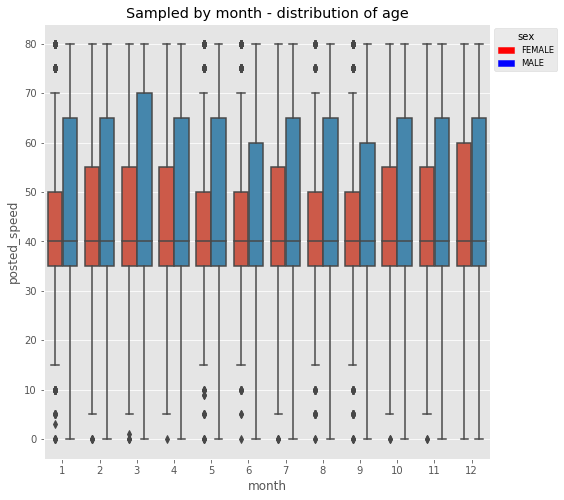
**Interpretation**

* In this graph we show that the mean age of drives across male and female are very close.
* 
* 

**interpretation**

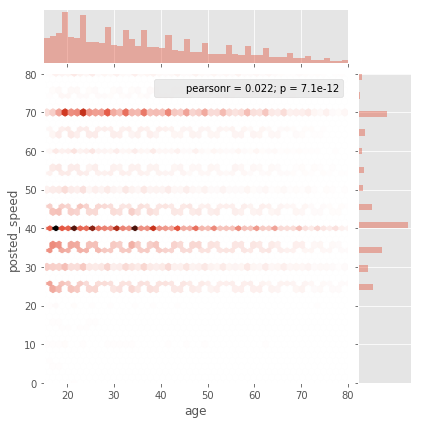
* Pie distribution of crash severity across all accidents.



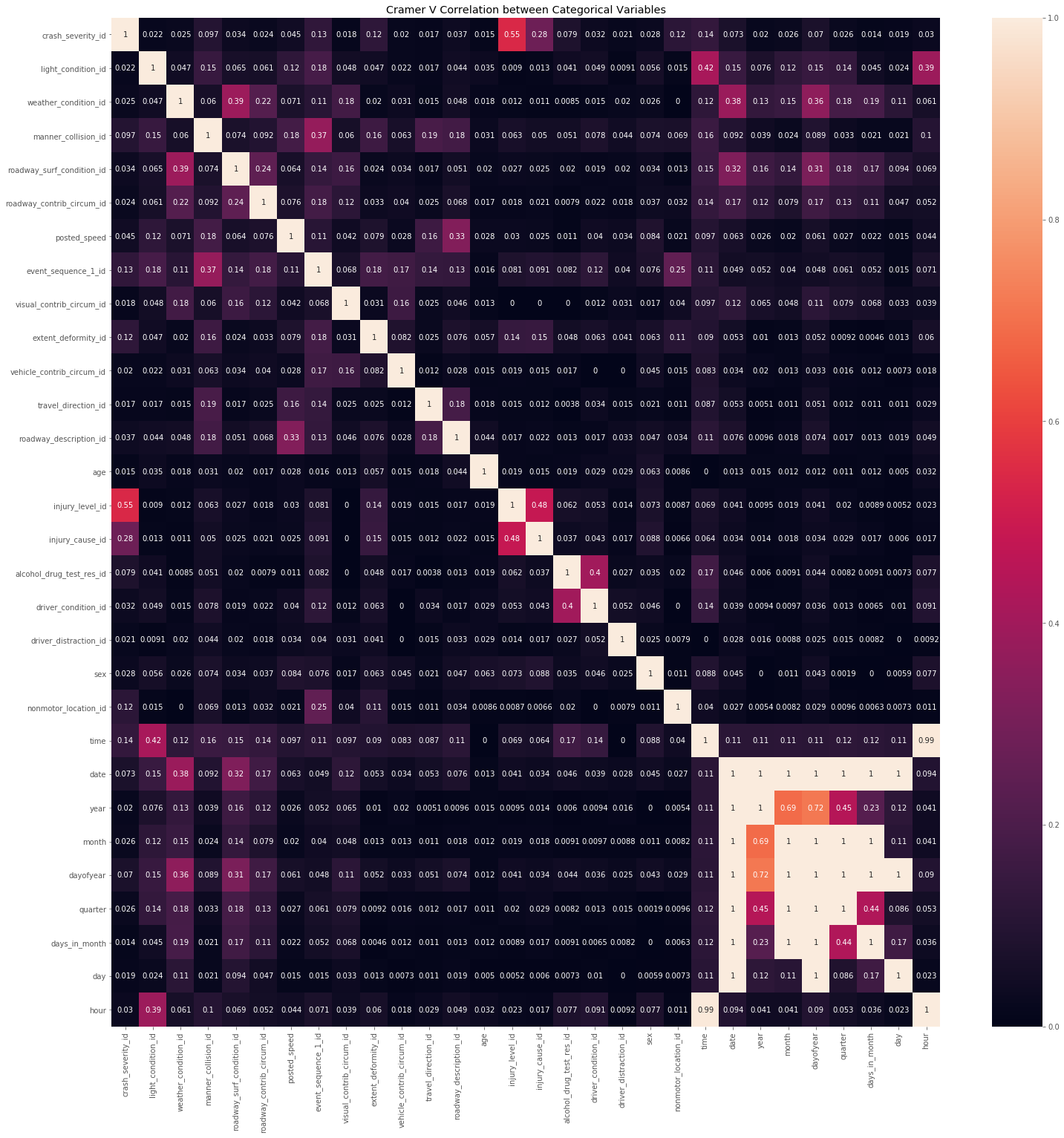
* **Interpretation**:
* - In this graph we show the distribution of crashes between both sexes across 12 months of the year
* - As seen above the median of age in females is always lower than males.
* - In particular months such as february and April this difference is wider.
* 

**Interpretation :**

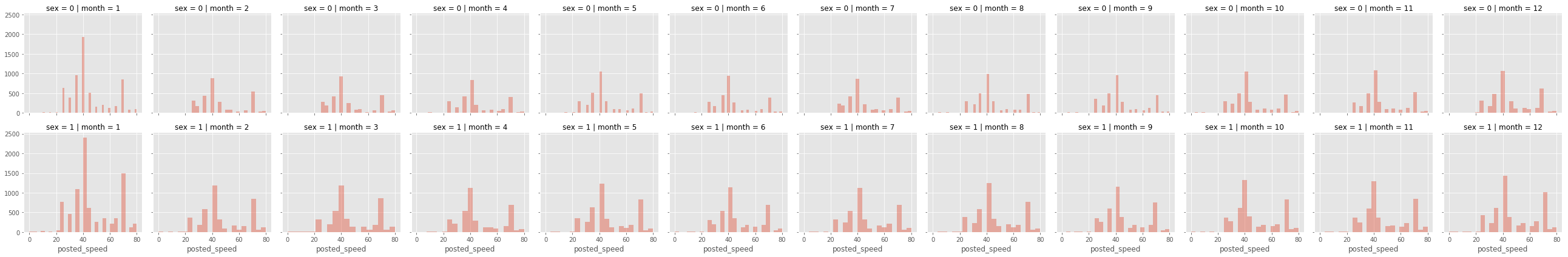
* In this graph we show the distribution of crashes between both sexes across 12 months of the year
* As seen above the median of posted speed is the same for both the sexes as 40 miles/hr.
* Men drive faster than women.

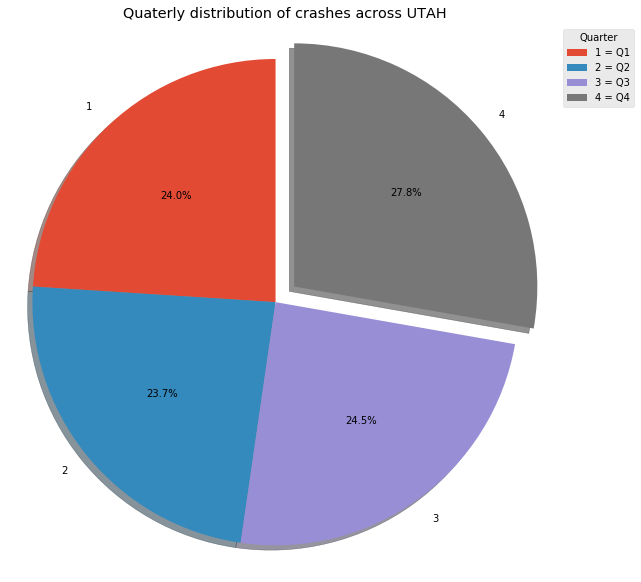


**Interpretation**:

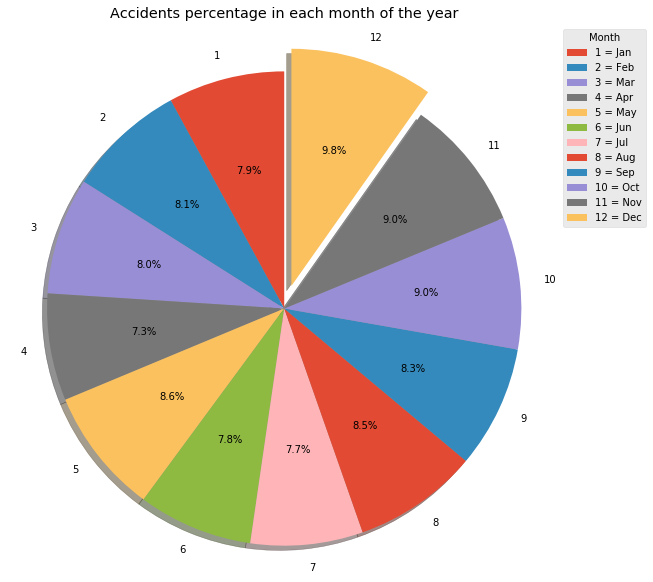
* In this graph we see that the accidents at posted speed of 40 miles/hr is the highest across all the ages.
* We also have similar results at 70 miles/hr
* We particularly observe the surge in accidents at the above speed for yourger age between 18 - 28 appx.
* We also observe that there are less drivers with age ranging from 50 to 80.
* 

**Interpretation**

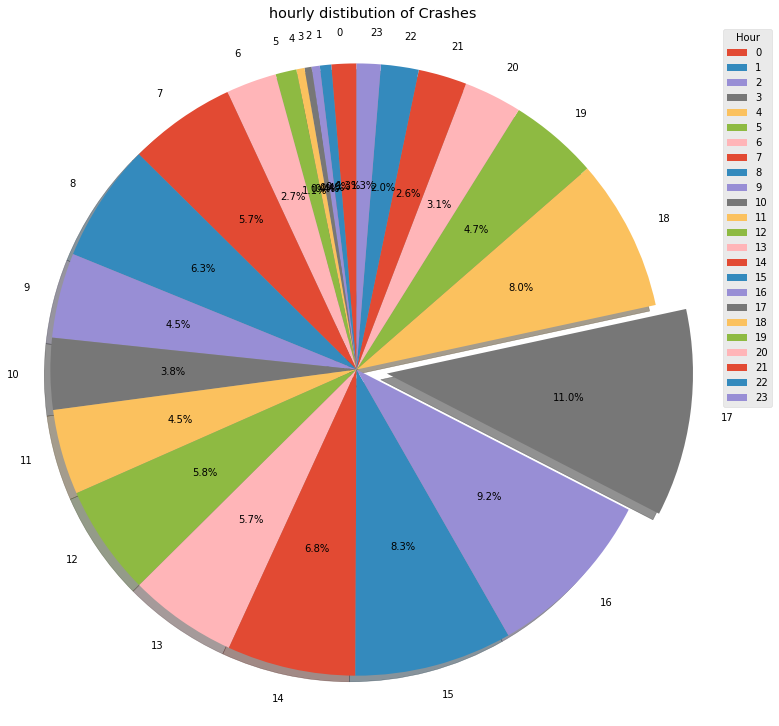
* Above plot explains correlation of both categorical and numeric variables.
* Injury\_level\_Id correlates good with crash\_severity\_id
* Injury\_level\_Id correlates good with injury\_cause\_id.
* 



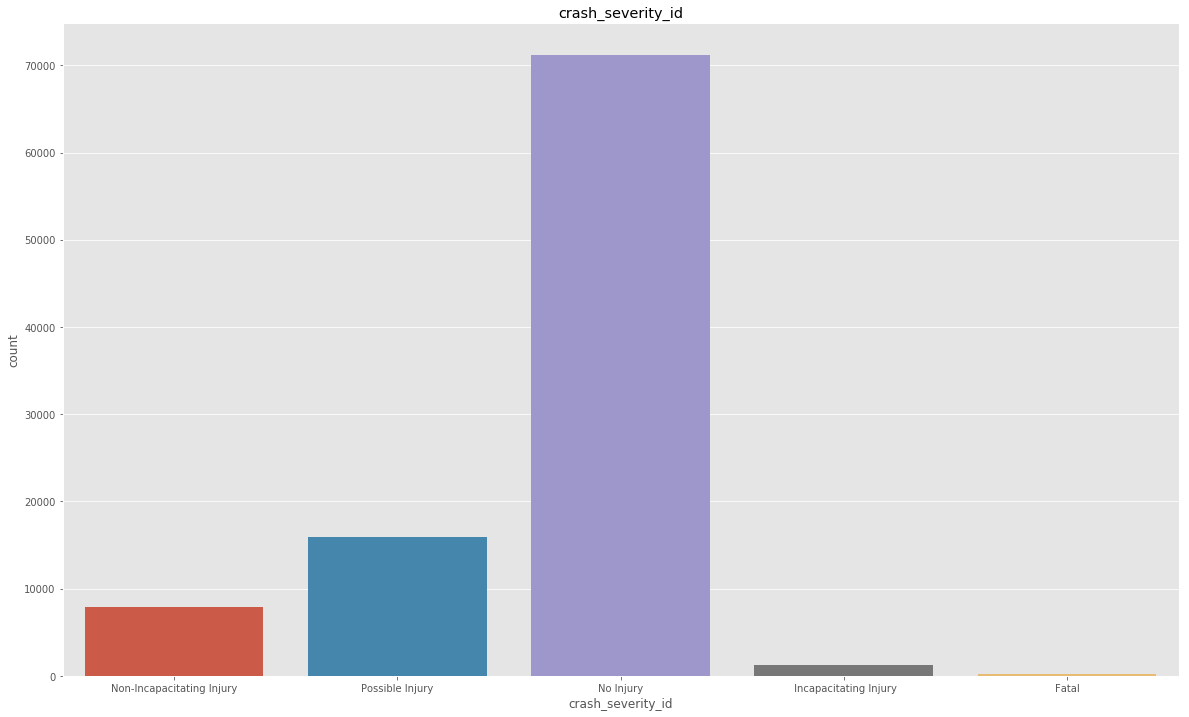
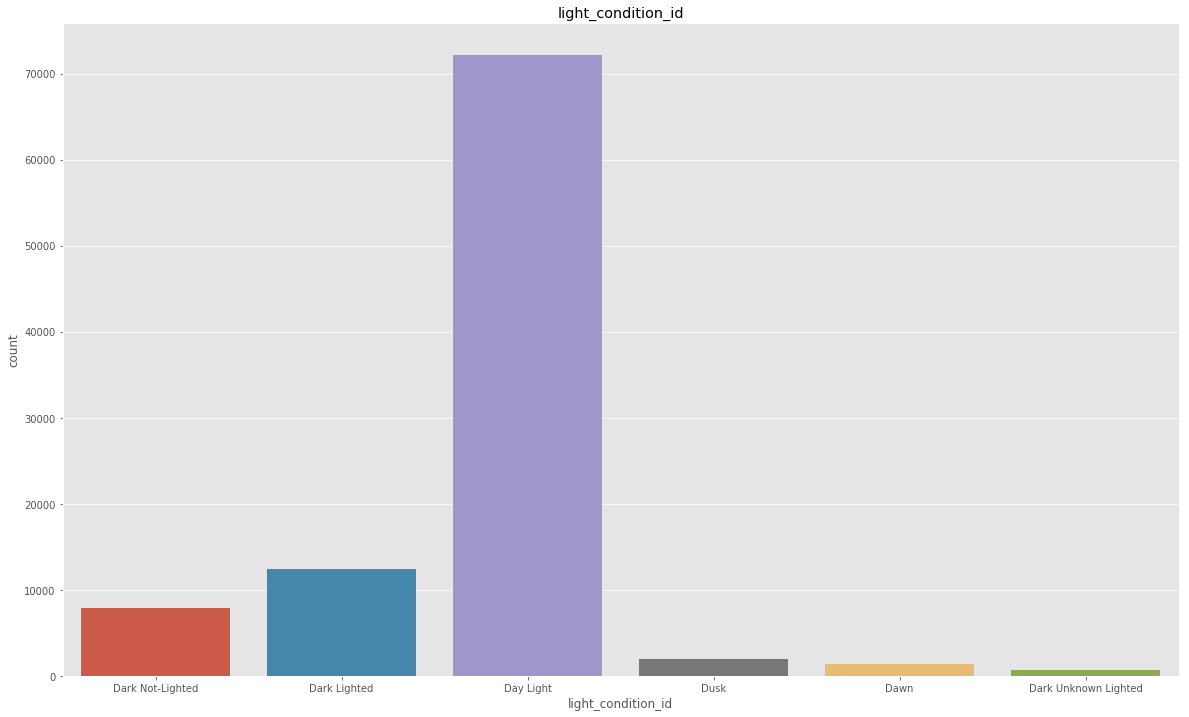
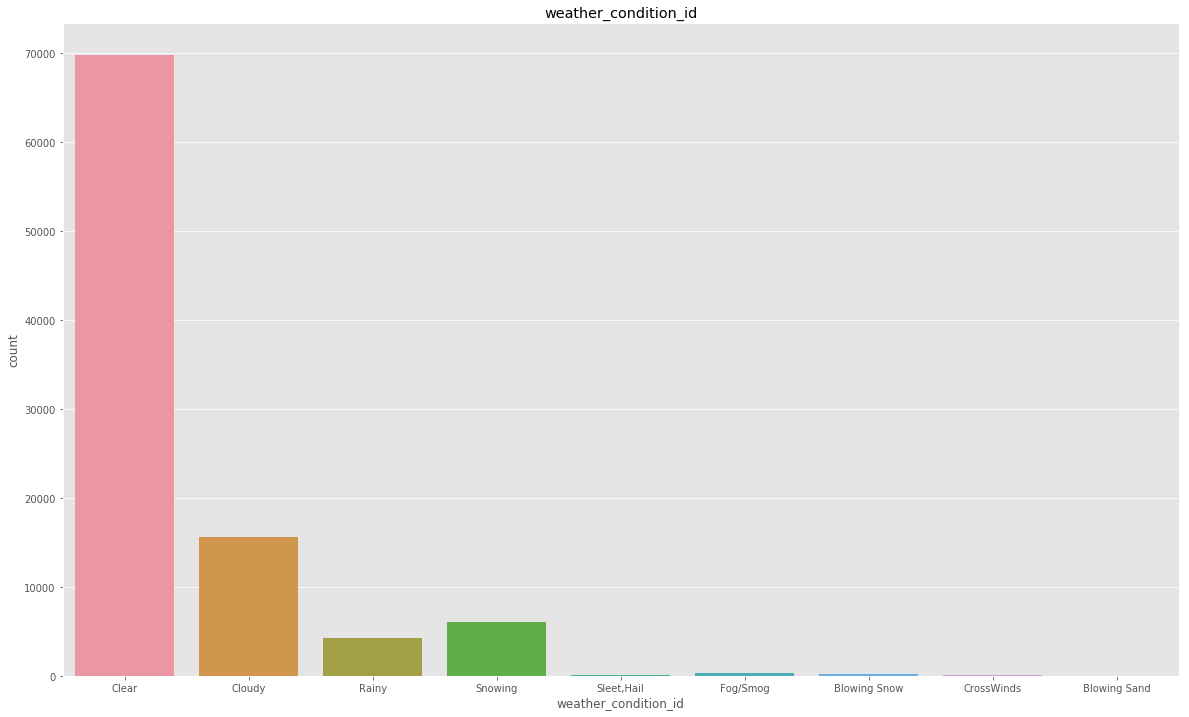
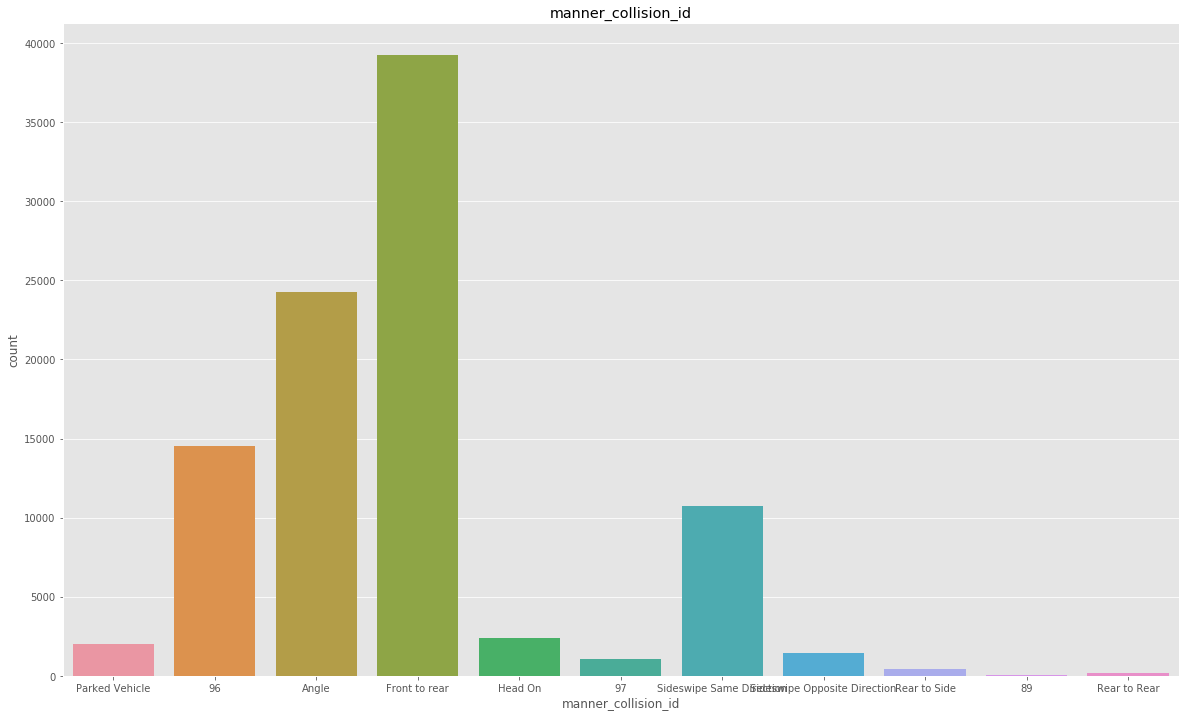
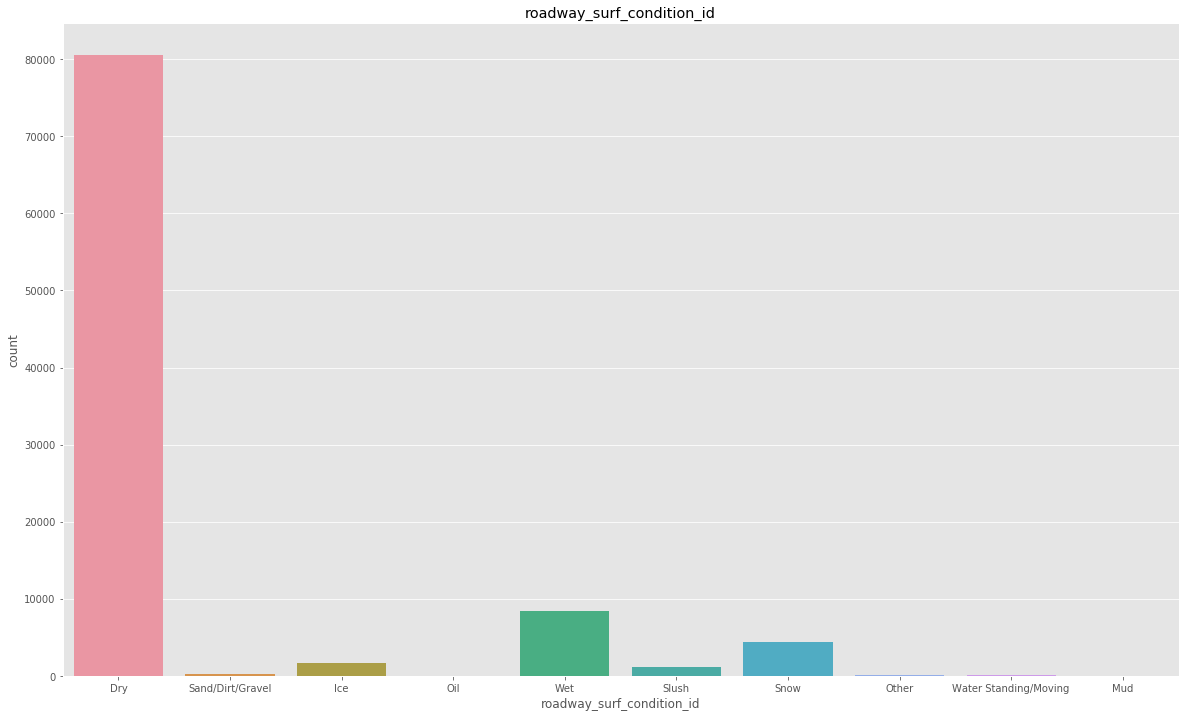
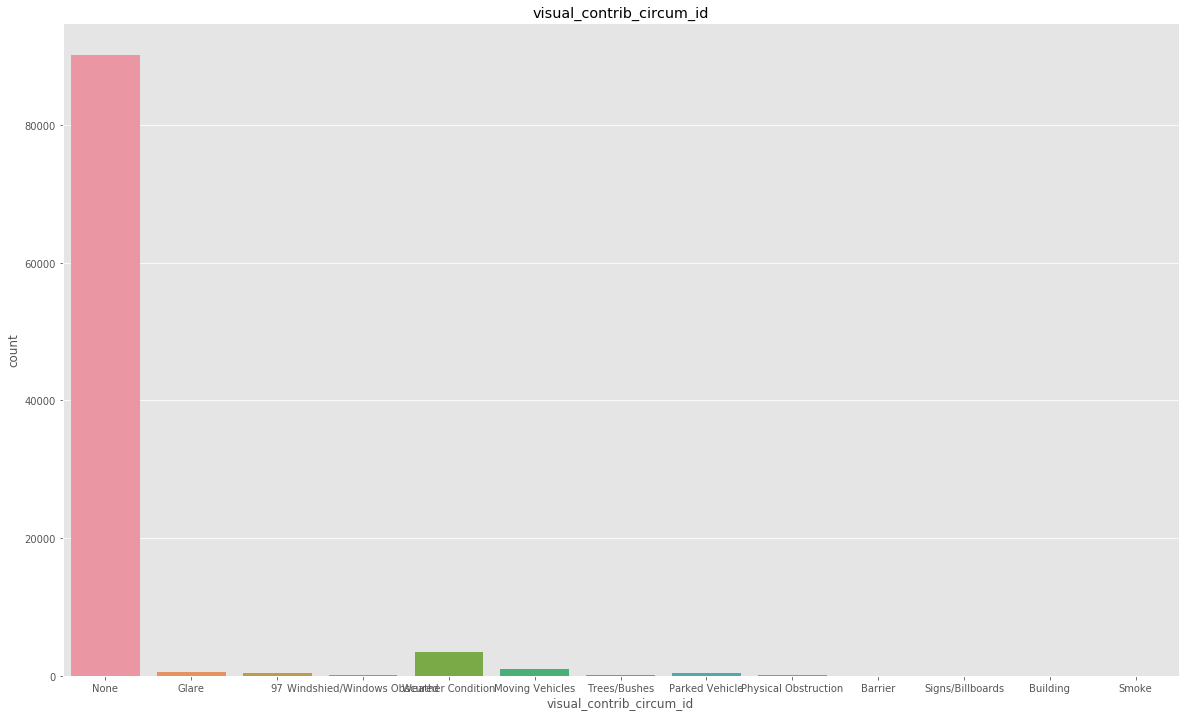
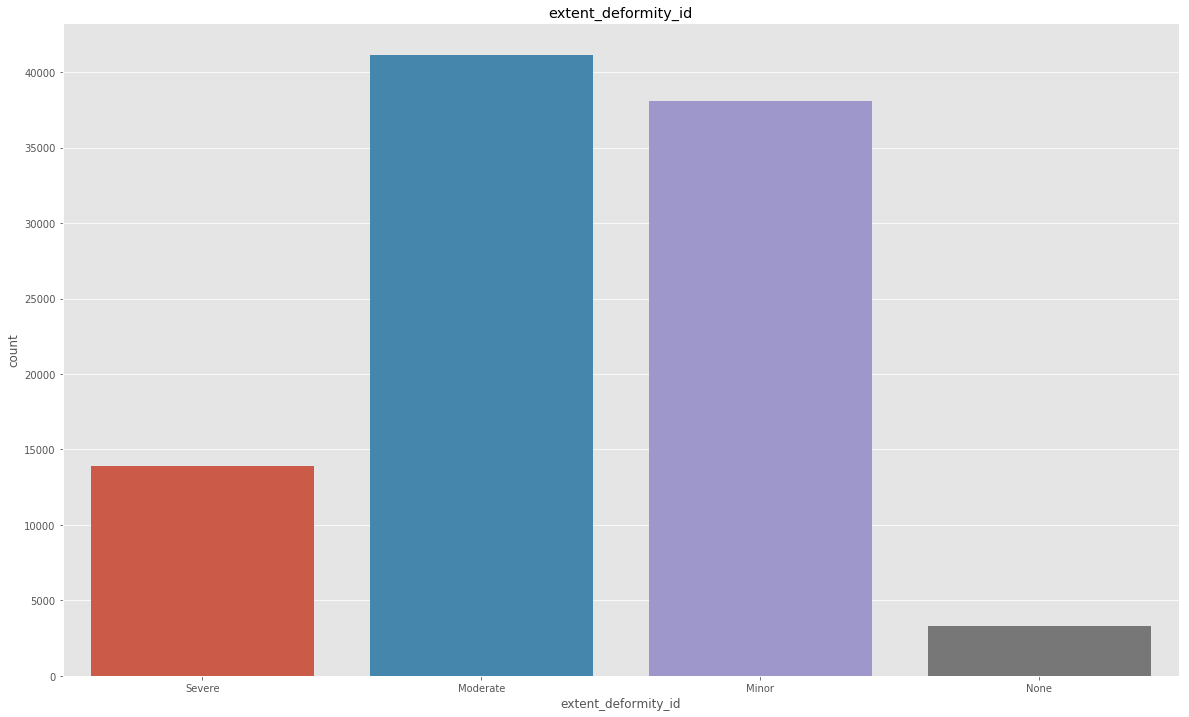
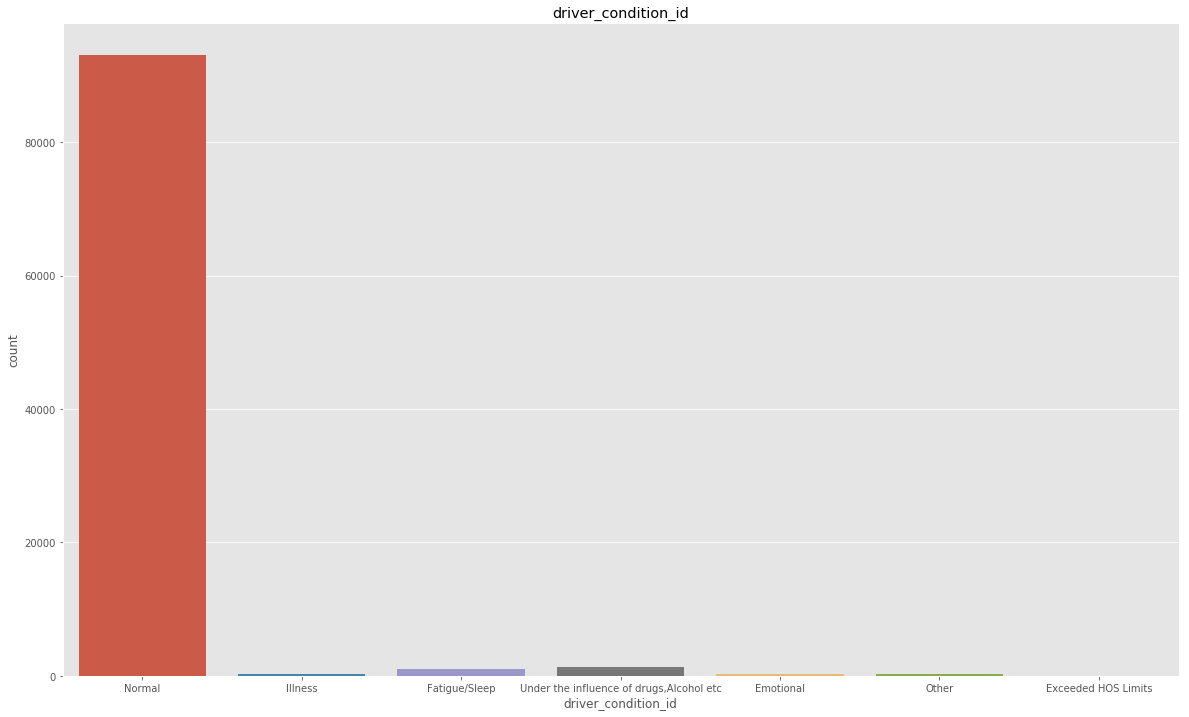
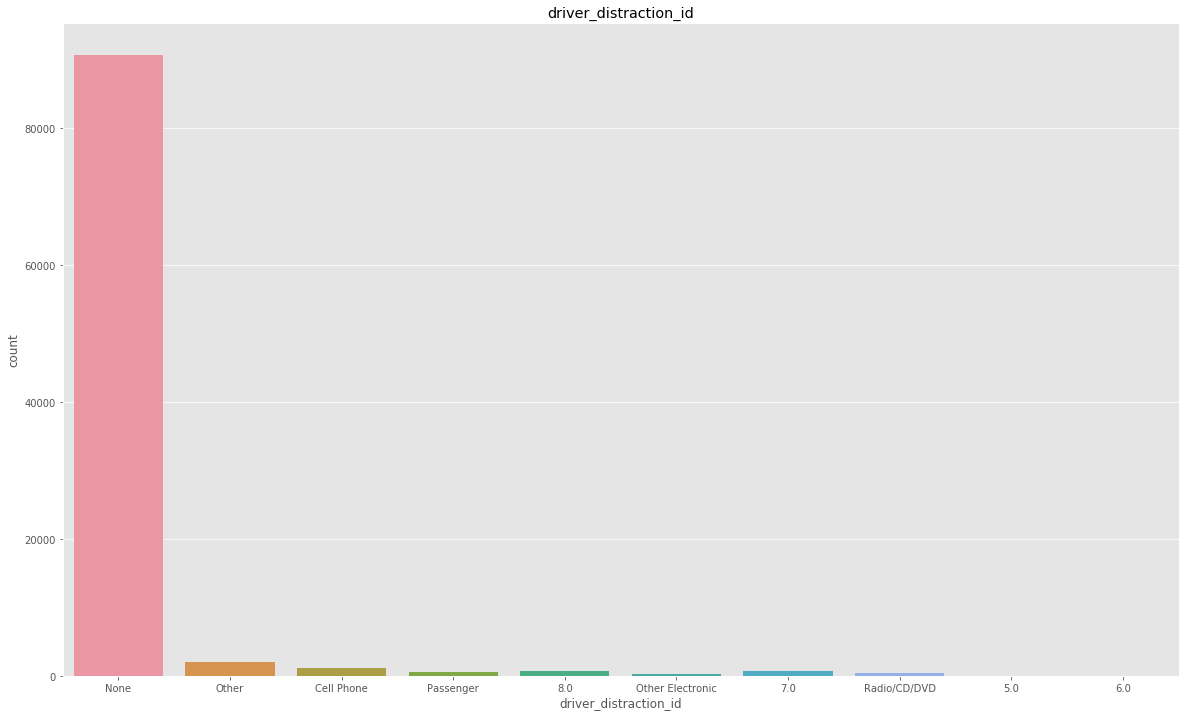
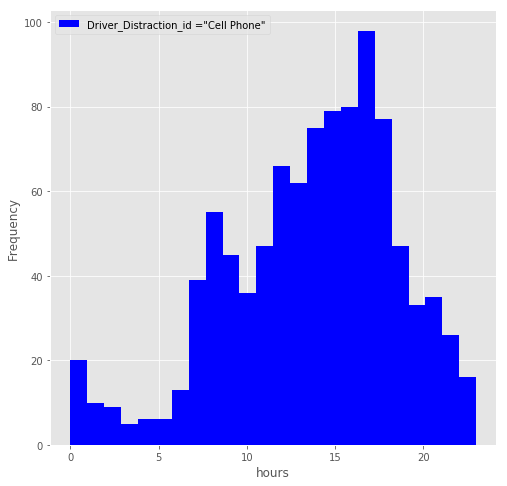
**interpretataion**

* As seen in the above graph in the last quarter of the year the number of accidents are the highest.
* 

**Interpretation :**

* From the above graph we can see that the closer we get to the holiday season the percentage of accidents rise which is perhaps to rise in travelling when the roads are most dangerous.
* 

**Interpretation**:

* From our point of view this is one of the most important graphs as it describes how the number of accidents rise with the progressing of the day.
* We think this is due to both the rush hour/traffic and people's tiredness during the afternoon hours.
* 
* 
* 
* 
* 
* 
* 
* 
* 
* 

4 Models Used:

INPUTS for all models:

light\_condition\_id

weather\_condition\_id

manner\_collision\_id

roadway\_surf\_condition\_id

roadway\_contrib\_circum\_id

event\_sequence\_1\_id

visual\_contrib\_circum\_id

extent\_deformity\_id

vehicle\_contrib\_circum\_id

travel\_direction\_id

roadway\_description\_id

injury\_level\_id

injury\_cause\_id

alcohol\_drug\_test\_res\_id

driver\_condition\_id

driver\_distraction\_id

nonmotor\_location\_id

Name: 79895, dtype: float64

OutCome Variable :

Crash\_severity\_id

**METHOD 1 : Using KNN**[**¶**](#METHOD-1-:-Using-KNN)

Accuracy on training data = 0.8059314564214237

Accuracy on testing data = 0.8059314564214237

[[14201 18 11 1 0]

[ 2281 866 67 0 0]

[ 860 242 469 4 0]

[ 98 35 90 7 1]

[ 10 4 16 5 1]]

True: [2 1 1 1 2 3 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1 2 1 1 1 3 1 1 1 1 1 1 2

1 1 1 1 1 3 2 1 1 1 1 1 1]

Predicted: [1 1 1 1 2 3 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 3 1 1 1 1 1 1 1

1 1 1 1 1 1 1 1 1 1 1 1 1]

Cross Validation:

0.7351765368493078

## METHOD 2 : Using Naive Bayes

Accuracy on training data:

0.7990459771604859

Accuracy on test data:

0.7950951418053611

Cross Validation:

0.7981004816905118

Confusion matrix

[[13841 0 1 109 280]

[ 1956 916 163 45 134]

[ 754 79 549 31 162]

[ 79 16 77 11 48]

[ 5 1 8 4 18]]

True: [2 1 1 1 2 3 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1 2 1 1 1 3 1 1 1 1 1 1 2

1 1 1 1 1 3 2 1 1 1 1 1 1]

Predicted: [1 1 1 1 2 3 1 1 1 1 1 1 1 1 1 1 5 2 1 1 1 1 1 1 1 1 1 1 1 3 1 1 1 1 1 1 2

1 1 1 1 1 1 1 1 1 1 1 1 1]

## METHOD 3 : Using Decision Tree

Accuracy on training data= 0.8451527603147239

Accuracy on test data= 0.8342406802509462

Cross Validation:

0.8392786709669121

Confusion Matix:

[[14164 35 27 5 0]

[ 2003 1157 49 5 0]

[ 776 120 672 4 3]

[ 80 27 32 89 3]

[ 7 3 8 10 8]]

True: [2 1 1 1 2 3 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1 2 1 1 1 3 1 1 1 1 1 1 2

1 1 1 1 1 3 2 1 1 1 1 1 1]

Predicted: [1 1 1 1 2 3 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 3 1 1 1 1 1 1 2

1 1 1 1 1 1 1 1 1 1 1 1 1]

## METHOD 4 : Using Random Forest

Accuracy on training data= 0.8788416918350681

Accuracy on test data= 0.8297298698605278

Cross Validation:

0.8231847334417044

Confusion Matrix

[[13843 284 92 11 1]

[ 1881 1263 58 10 2]

[ 741 140 692 2 0]

[ 78 24 36 91 2]

[ 9 3 7 8 9]]

## METHOD 5 : Logistic Regression

Accuracy on training data= 0.583146460653039

Accuracy on test data= 0.580702027272256

Confusion Matrix:

1 11200

0 8087

True: [0 0 0 0 0 1 1 0 1 0 0 0 1 0 1 1 0 1 1 1 1 1 1 1 0]

Pred: [1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]

Analysis:

Our initial goal of doing this project was to see if weather condition would be an accurate predictor of Utah accidents. After evaluating the data that came from DMV, we quickly realized that our data set was consumed by bunch of categorical columns! It did not take us long before we realized that weather could not be a sole or even an important factor in predicting any kind of accident! However, the richness and diversity of our data along with the initial basic statistical and visual graphs gave our analysis depth and meaningful insights which eventually led us to desirable predictive results. The best part of our analysis was running 4 different classification models including; KNN, Naïve Bayes, Decision Tree and Random forest which all gave us very consistent and robust results for the final conclusion. We chose “Crash Severity” as the predictive outcome of our accidents model. While initially we considered only 4-5 different features for our models, after many trial and errors we realized that considering a number of features would be a more feasible modeling approach which would direct us toward a more accurate result. Above we are showing our accuracy of both our training and test data set for each model. We also tried our Decision tree for different tree dept but decided that for the sake of our graph we only use depth of three. We are very happy with our results and are hoping that in the near future can dive even more into this ocean of data.

END