Study of the effects of climatic conditions on alcohol consumption and the study of road accident fatalities

Soham Sameer More
School of Computing. National College of Ireland
(of Affiliation)
Dublin, Ireland
x19149240@student.ncirl.ie

***Abstract*— In this paper I study the effects of weather on the alcohol consumption pattern in the state of Iowa, USA. Also, the road accidents in the state of Iowa, USA are analyzed to study the pattern and factors affecting the fatality in an accident. Data Mining and Machine Learning methods like Multiple Linear Regression, Support Vector Regression, Logistic Regression, Naive Bayes and Support Vector Machine are used to conduct this study. It was found that the pattern of alcohol consumption is not related to the weather conditions in the state of Iowa, USA. The fatalities in a road accident were found to be dependent on the circumstances of the accident. The datasets in this study encompass the details of weather, alcohol consumption and road accidents in the state of Iowa, USA.**

Keywords—* *data mining, alcohol consumption pattern, climate, road accidents, fatalities

I. Introduction

A. Climatic effects on alcohol consumption

Alcohol is an integral part of many cultures and traditions across the world. Alcohol is consumed in social settings like weddings, funeral, festivals, etc. or in casual settings where friends and family get together to have a few drinks. Alcohol has many short term and long term effects on the body, especially the thermoregulatory response to cold. The scientific research has a varied opinion on the effects of alcohol consumption on the thermoregulation in cold weather [Alcohol ingestion and temperature regulation during cold exposure]. Haight and Keatinge [Failure of thermoregulation in the cold during hypoglycaemia induced by exercise and ethanol] studied the effects of alcohol (ethanol) consumption in a cold environment by ingesting a study groups with alcohol and increasing the alcohol concentration in successive studies. It was found that at a lower concentration of alcohol (0.34 g per kg body wt.) there is no change in the thermoregulatory mechanism of the body. However at a higher concentration of alcohol (0.79 g per kg body wt.) the thermoregulatory mechanism of the body is affected which increases the surface body temperature and causes sweating.

Beker et al. [Human Physiology in Extreme Heat and Cold] also state that extreme cold stimulates the alpha-adrenoceptors which promotes heat conservation in the body. This is facilitated by cutaneous vascular constriction (narrowing of the blood vessels) and decreased peripheral profusion (decreased blood supply to peripheral organs) which helps in maintaining the core body temperature and blood flow to the vital organs. This results in decreased skin and muscle temperature.

As alcohol has the exact opposite effect on the skin and muscle temperature, it will be interesting to find out if the consumption of alcohol is dependent on the weather conditions.

The objective of this study is to analyse if there is any correlation between cold weather and consumption of alcohol in the state of Iowa, USA. The sales of alcohol is considered as the dependent variable in this study as it is a direct indicator of the consumption of alcohol.

B. Fatalities in road accidents

Road safety and road accidents have been a concern since the advent of motorized vehicles in the early 20th century [Thinking about the history of road safety]. Accidents are caused due to human error as well as unfavorable environment conditions like weather, road condition, lighting conditions, junction type, etc. [Real-time Traffic Accident Severity Prediction using Data Mining Technologies]. Road accident injuries and fatalities are a concern for those involved in the accident as well as the first response teams and emergency medical services (EMS). This study aims at predicting the fatality in an accident based on the circumstances of the accident like the road surface type, light conditions, junction type, drug or alcohol influence, weather type, etc. This prediction can prove to be vital for the EMS in saving lives.
Road accidents data in the state of Iowa, USA for the year 2015 and 2016 is used in this study.

II. Related Work

A. Climatic effects on alcohol consumption

Ventura-Cots et al [5] conducted a study of the relation between colder climates specifically the mean temperature and the number of hours of sunshine and alcohol consumption. Countries were categorized into tropical, dry, temperate, continental and polar in their study and religious subgroups that affected the alcohol consumption were removed from the study to avoid bias. The results confirmed that in the U.S., the colder climate had a positive correlation with alcohol consumption i.e. the colder the climate higher the consumption of alcohol. Hagstr¨om et al [6] studied the effects of fewer sunlight hours and it's the correlation with increased alcohol consumption in Nordic Northern European countries. Their study found alcohol consumption had a negative correlation to the temperature and sunlight hours i.e. the colder the climate lesser the consumption of alcohol. This contrasted with the Southern European countries where they saw a positive correlation between cold climate and alcohol consumption. This shows that the relation between climate and alcohol consumption may be dependent on the geographies. This study intends to find if the study by Ventura-Conts et al [5] about the US also holds true for its state Iowa.

Beau J. Freund et al mention of reports that show a benefit in ingestion of alcohol during cold weather which lead the German and Russian military leaders to recommend alcohol consumption is small doses during World War II. The effects of alcohol consumption in extreme cold are also presumed to protect against cold related injuries like frostbite while also improving sleep and reducing discomfort.

B. Fatalities in road accidents

Liling Li et al [7] studied the road accident fatalities using data mining techniques and with a focus on providing better emergency medical services (EMS) to the victims of road accidents. Their study found that environmental factors like weather, road surface, and light conditions do not affect the fatality rate whereas other factors like driver being drunk or not, type of collision location of the collision have a strong association with fatalities. Pisano et al [8] studied the U.S. highway crashes in adverse weather conditions and found that the weather has a significant impact on road safety. Almost a quarter of road accidents in the US are weather-related. Their study concluded that more research is needed on weather related crashes to help understand the factors affecting the crashes and to take preemptive measures to avoid fatalities.

Xiao-Ling Xia et al. studied the factors affecting the severity of a road accident for the road accidents in Seattle from the year 2004 to 2016. Classifiers like Naïve Bayes, Random Forest, MLP and AdaBoost were used in this study. They built classification models to predict the accident severity into two classes: Property Damage only and Injury Collision. Their models performed well on the dataset after preprocessing with AdaBoost performing the best with a F1 score of 0.86 and AUC of 0.91. While the Naïve Bayes performed the worst among the models chosen with a F1 score of 0.81 and AUC of 0.88.

III. Data Mining Methodology

Both the studies were conducted in accordance to the cross-industry process for data mining (CRISP-DM) methodology as it provides a structured approach to a data mining project. Each step of the CRISP-DM is detailed below for both the studies.

A. Business understanding

1) Climatic effects on alcohol consumption

Objective - The objective of this project is to analyze the alcohol sales data and the weather data in the state of Iowa, USA to find out if there is any correlation amongst them. This information is useful to forecast the alcohol sales based on the weather forecast. Also, this information can prove to be an early warning system for medical services in case of sudden drop in temperature as alcohol consumption in extreme cold can cause hypothermia and prove to be fatal.

Tools and techniques - Statistical computing language R is used to perform data pre-processing and model building. Data mining and machine learning techniques like multiple linear regression and support vector regression are used to predict the alcohol consumption (sales).

2) Fatalities in road

Objective – The objective of this study is to find the relation between a fatality in a road accident and the circumstances of an accident like the weather, road conditions, cause of accident, type of impact, single car or multiple cars etc. This information and prediction can be important for emergency medical services and local police departments.

Tools and techniques - Statistical computing language R is used to perform data pre-processing and model building. Data mining and machine learning techniques like Logistic Regression, Naïve Bayes Classifier and Support Vector Machine are used to predict the alcohol consumption (sales).

B. Data understanding

1) Climatic effects on alcohol consumption

- Data collection -

i. The alcohol sales data was acquired from the Iowa state government website. This dataset provides data for every sale made to a dealer by the state of Iowa. Each sale entry consists of data fields like date and time, store name, store location, item purchased, category of alcohol, quantity, volume and price of alcohol sold. This dataset was in comma separated values (csv) format and had 4464376 rows and 24 columns of data out of which 99348 values for the county column were missing. As the county column was used to combine the climate and alcohol sales data it the missing values populated with values using a python script and FCC census API using the latitude and longitude data available in the dataset.

ii. National Centers for Environmental Information - This dataset gives location and station wise data of climate conditions like temperature, humidity, precipitation, snowfall, etc. Data is available from multiple stations spread over the state of Iowa, USA. This dataset has a lot of missing values as every station reports a particular type of data. This data was consolidated by date and county to get a concise dataset. This raw data was split over two csv files with 10773 and 106654 rows and 62 columns was downloaded from the National Centers for Environmental Information portal.

- Data properties -

i. The state of Iowa, USA is an alcohol beverage control state, which means that the state holds a monopoly on wholesaling alcohol in the state. The alcohol sales data contains the details of each, and every liquor sale made by the state of Iowa to liquor

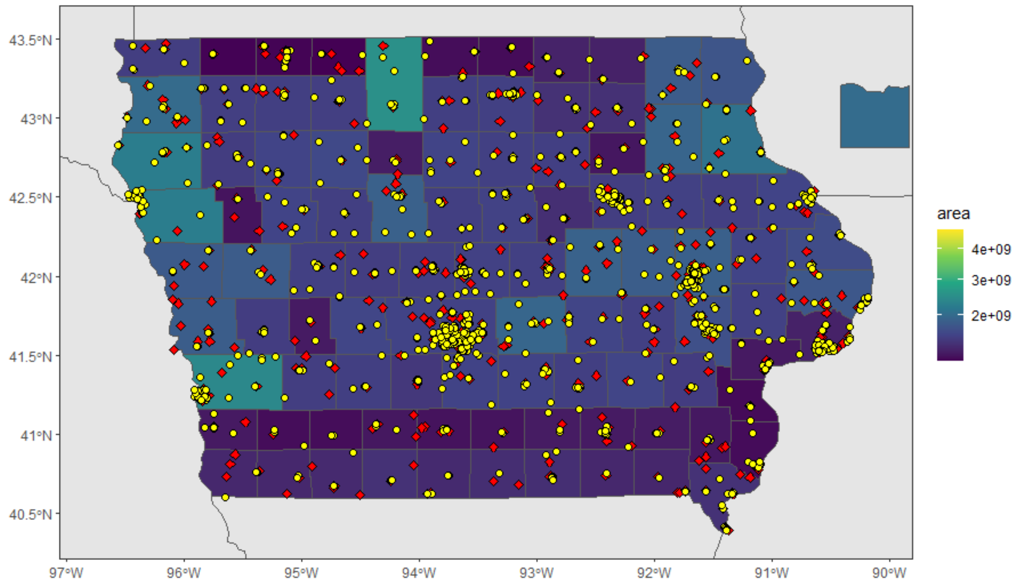


Fig: Map of Iowa with red dots representing locations of weather station in red and alcohol stores in yellow.

stores. It contains columns such as Invoice Number, Date, Store Number, Location and Address, County, Alcohol Category, Item Number and Description, Number of Bottles sold, Cost and Retail value of a bottle, Volume of Liquor sold and the Sale in US Dollars.

ii. The weather datasets follow the Global Historical Climatology Network (GHCN) - daily documentation format. This is a well-defined documentation method and is used by GHCN which is an integrated database of climate summaries from land surface stations across the globe. Every data column has an associated attributes column which contains information about the data collection time and quality of data. The column names are in GHCN abbreviation format. The documentation from GHCN was used to understand values in each column. Factors like minimum and maximum temperature, precipitation and snowfall were used in this study.

2) Fatalities in road accidents

- Data collection -
  The road accidents data was acquired from the Iowa state government website. This dataset has a record of every accident in Iowa state. This dataset was in comma separated values (csv) format and had 1048575 rows and 37 columns. The details and circumstances of every accident are recorded in this dataset.

- Data properties -
  The accidents dataset contains details like Department of Transportation (DOT) case number, Crash Date, Location information, Major Cause, Crash Manner, Crash Severity, Surface Conditions, Drug or Alcohol involvement, Light Conditions, Weather, Roadway Junction, First Harmful Event, Number of Vehicles and Occupants etc.

C. Data preparation

1) Climatic effects on alcohol consumption

i. Cleaning the alcohol data
The date column was first converted to a standard format using the anytime package. The dataset was then filtered to use the data for the years 2015 and 2016. This data had 99348 missing county values. County values play an important part in this study as the alcohol and weather datasets are combined using Date and County information. These missing values were populated using the FCC census API using the latitude and longitude information present in the data.

API Request example:
https://geo.fcc.gov/api/census/area?lat=42.63390&lon=-96.29007&format=json

API Response example:

```
{
  "input": {
    "lat": 42.6339,
    "lon": -96.29007
  },
  "results": [
    {
      "block_fips": "191499706003014",
      "bbox": [
        -96.290259,
        42.633361,
        -96.285638,
        42.648959
      ],
      "county_fips": "19149",
      "county_name": "Plymouth",
      "state_fips": "19",
      "state_code": "IA",
      "state_name": "Iowa",
      "block_pop_2015": 0,
      "amt": "AMT004",
      "bea": "BEA117",
      "bta": "BTA421",
      "cma": "CMA427",
      "eag": "EAG005",
      "ivm": "IVM427",
      "mea": "MEA021",
      "mta": "MTA032",
      "pea": "PEA252",
      "rea": "REA003",
        "rpc": "RPC003",
        "vpc": "VPC004"
      }
    ]
}
```

An aggregate dataset was created from this dataset. Sales of all the stores in a county for a date were added to give a single row which gives the total sales of liquor in a county by date.

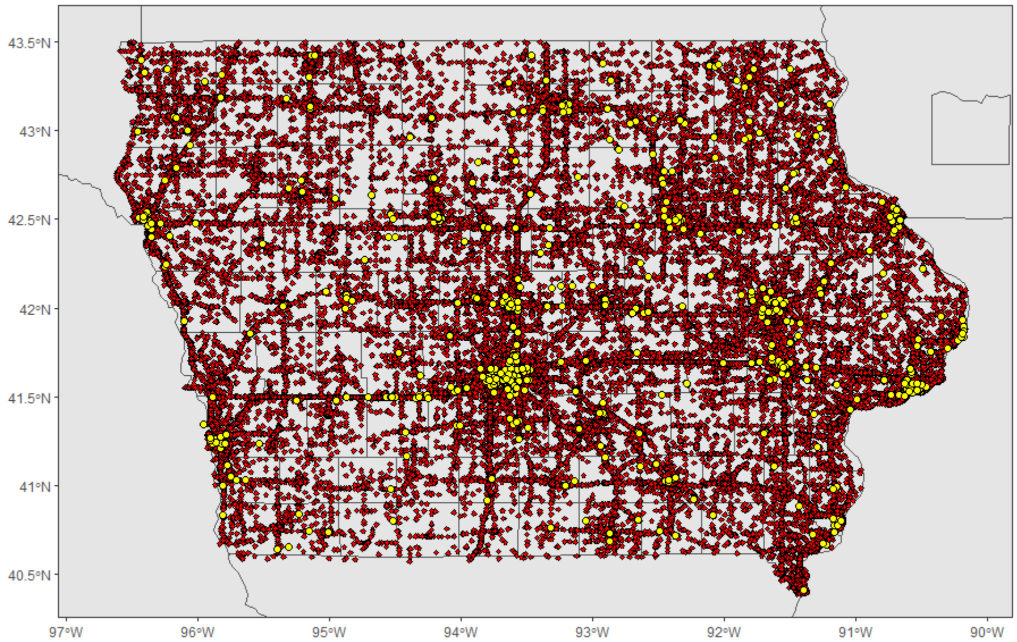


Fig: A map representing accident locations in red dots and fatal accidents in yellow dots.

ii. Cleaning the weather data

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Max Temperature	-17.8	8.6	20.85	17.34	27.2	35.6
Min Temperature	-30	0	9.4	7.315	16.1	24.4
Precipitation	0.02	0.825	3.571	7.808	10.425	117.1
Alcohol Sales	4.4	434.9	2127.6	21114.5	13149.2	755335

Challenges in getting the correct weather data.

Weather dataset contained 56 columns with data readings regarding weather and a few more columns specifying the location, date and station name of the station where the reading was recorded. These 56 columns are 26 pairs (values, attributes) of readings. Most of these columns contained NA values and these columns were not used in this study. As a result, these columns were deleted from the dataset. Columns such as Weather Type, Daily percent of possible sunshine, Minimum soil temperature, Maximum soil temperature, Daily total sunshine, etc. were removed from this dataset as majority of these columns contained NA values. The columns that were used in the final study are Maximum temperature, Minimum temperature, Snowfall, Snow depth, Precipitation. Similar to the alcohol sales dataset, this dataset too was reduced to one row per date per county. This was achieved by taking the mean of all the readings for a county on any day. For example, if there are 10 readings for maximum temperature for county Dallas then the maximum temperature is taking as the mean of all these 10 readings.

iii. Combining the alcohol and weather data

Both the datasets had two common columns, namely Date and County. The datasets were combined using a condition similar to inner join used in relational database systems. The non-matching rows in both the datasets were ignored.

The basic statistics for temperature, precipitation and alcohol sales can be seen in the table above.

2) *Fatalities in road accidents*

The road accidents dataset had very less noise and missing values. The Occupants column had a few outliers that seemed like errors. These rows were not included in the study to reduce erroneous variance in the model.

One additional column for Fatality was added to the data which takes on a binary value of 0 or 1. The value of this column is 0 if there is no death in an accident and it is 1 if there is a death in an accident.

D. *Modeling*

1) *Climatic effects on alcohol consumption*

Regression analysis was conducted in this study to predict the alcohol consumption based on the weather conditions like temperature, snowfall and precipitation. The following regression models were implemented:

a) Multiple Linear Regression

Multiple linear regression is a supervised parametric model that tries to fit a straight line that explains the variance in the dependent variable based on the variance in the independent variables. The equation of the line is as follows in this study:

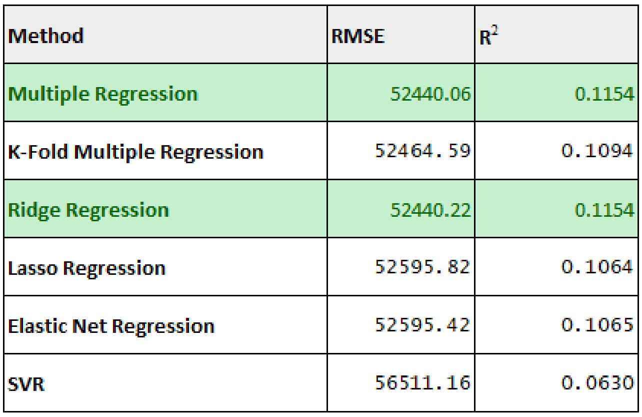
Yi (Alcohol Sales in USD) = B0 + B1 (precipitation) + B2(Snowfall) + B3(Snow depth) + B4(Minimum Temperature) + B5(Maximum Temperature) + B6 (wind speed) + e (error)

Ridge, Lasso and Elastic Net regularization techniques were also applied in this study to reduce overfitting and to add an additional penalty parameter that aims to minimize complexity or reduce the number of features used in the final model.

Multiple regression was chosen for this study as it offers easy and fast implementation and fast training of the model. Regularization parameters can also be added to the model to avoid overfitting.

b) Support Vector Regression

In contrast to ordinary least squares (OLS), the objective function of support vector regression (SVR) is to minimize the coefficients — more specifically, the L2-norm (L2 norm is calculated as the square root of the sum of the squared vector values) of the coefficient vector — not the squared error.



Method	RMSE	R^2
Multiple Regression	52440.06	0.1154
K-Fold Multiple Regression	52464.59	0.1094
Ridge Regression	52440.22	0.1154
Lasso Regression	52595.82	0.1064
Elastic Net Regression	52595.42	0.1065
SVR	56511.16	0.0630

SVR is a non-linear regression model whereas Multiple Linear Regression is a linear regression method. SVR was chosen for this study in order to check how a non-linear regression method performs in contrast to a linear regression method.

2) Fatalities in road accidents

This study was conducted to predict road accident fatalities based on the accident circumstances and the environment of the accident. This dataset is unbalanced as the number of accidents with fatalities is very small as compared to the number of accidents without fatalities.

a) Logistic Regression

The target variable is a binary variable, Fatality which takes a value of 1 in case of an accident related death and 0 if there was no death in the accident. Logistic regression predicts the probability of occurrence of an outcome. The probability threshold of 0.5 is usually used to separate the outcome. However, since predicting a fatality in an accident is more important than predicting a non-fatal accident, a probability threshold of 0.4 was used in this study.

Also, to remedy the unbalanced dataset, weights were used while building the logistic regression model.

Logistic regression was chosen in this study as it is designed to work on data where the dependent variable is binary in nature.

b) Naive Bayes

Naive Bayes classifier usually performs at par with other classifiers and it is a very robust model that can be used in most of the classification problems. Naive Bayes is a generative model (it considers the conditional probability of the observations given that the target variable holds true) in contrast to Logistic Regression which is a discriminative model (it considers the conditional probability of the target variable given an observation holds true).

Naïve Bayes was chosen in this study as it works best when the independent variables are not dependent on each other and because it is not sensitive to irrelevant data.

c) Support Vector Machine

SVM uses an error margin threshold to calculate the hyperplane that will divide the data into distinct classes.

Both Linear and Radial kernels were used to build SVM models. Both these models were evaluated to check which one performs better in this study.

SVM was chosen for this study as it uses kernel trick to solve complex data problems and because outliers can be handled with a soft margin.

IV. Evaluation

A. Climatic effects on alcohol

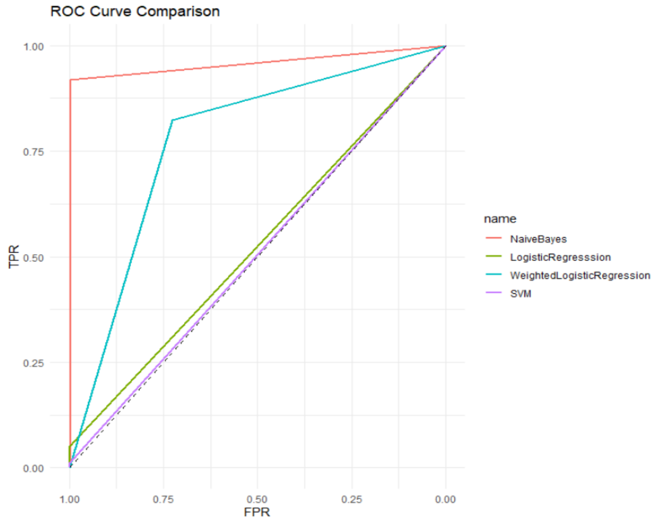
Multiple Regression with regularization techniques and cross validation was performed in this study. The weather parameters like temperature, wind, snow and precipitation do not show any significant relation to alcohol consumption in the state of Iowa, USA as can be seen from the evaluation metrics below.

This means that either the alcohol sales in not related to weather conditions or the weather conditions in this specific geographic region do not influence alcohol consumption.

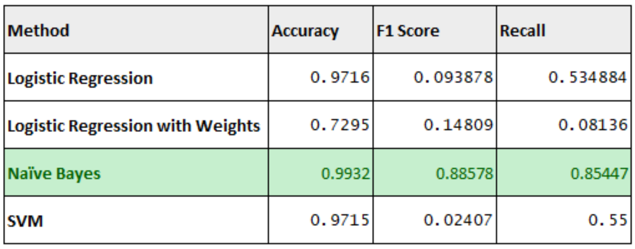
B. Fatalities in road accidents

The dependent variable in this study is Fatality. 1 represents the accidents in which there was a fatality and 0 represents an accident where there are no fatalities.

This dataset is highly unbalanced, and hence we cannot use accuracy to evaluate the model performance. Here predicting a death in a road accident is more important than predicting the accidents which do not have death. That is why Recall and F1 score along with Area Under Receiver Operating Characteristics (AUROC) are used to evaluate the model performance.



In the study of road accident fatalities, the correct detection of a fatality is more important than the correct detection of a non-fatal accident. To correctly evaluate the models for this criterion, I used the recall for detection of a fatality in the test dataset as a metric of evaluation. Three classification models were used in this study, Logistic Regression, Naïve Bayes and Support Vector Machine. From the evaluation metrics mentioned above we can see that the Naïve Bayes model outperforms rest of the models.



Method	Accuracy	F1 Score	Recall
Logistic Regression	0.9716	0.093878	0.534884
Logistic Regression with Weights	0.7295	0.14809	0.08136
Naïve Bayes	0.9932	0.88578	0.85447
SVM	0.9715	0.02407	0.55

The recall for detection of a fatality in an accident is 85.45% in Naïve Bayes classifier. This means that the model detects a fatality in an accident correctly 85.45% of the time. Which seems like a reasonably well performing model.

V. Conclusions and Future Work

A. Climatic effects on alcohol

This paper looked at the effects of weather conditions like temperature, wind, precipitation and snow on the consumption of alcohol in the state of Iowa, USA. The analysis concluded that there is no significant relation between weather conditions and the pattern of alcohol consumption.

The main problem during data gathering and cleaning was with the unavailability of a proper source of weather data. The weather data from the National Centers for Environmental Information portal had many parameters that were not usable due to lot of NA values.

Future studies on this topic can refer to a more reliable weather data source that provides details of other weather parameters like humidity, sunshine time, weather type, etc. The alcohol stores and weather stations are not evenly distributed in the state of Iowa as can be seen from figure. Weather and alcohol sales data can be appropriately weighted to provide a standardized measure across the whole state. Also, future studies can check if there exists a relation in weather with the type of alcohol like Whiskey, Beer, Rum, etc.

B. Fatalities in road accidents

The results of this study determine that there is a relation between the circumstances and type of collision in an accident to whether there will be a fatality in the accident. The Naïve Bayes classification outperforms other classification models of Logistic Regression and SVM.

Future studies on this topic can take into consideration the geographic accident information as well as the geo-hotspots of accidents as can be seen from the accidents map in figure. Other features like day and time can be used to find a pattern in the accidents over a week, month or a year. Future studies can also take into consideration the Emergency Medical Services (EMS) and study the relation between a road accident fatality and the EMS quick response process.

VI. References