FIND YOUR HAVEN

Term Project – Group 7

IS 5126 – Hands on with Business Analytics

APRIL 24, 2015

Ashwini Ravi (A0119980W)
Hushmija Xavier (A0123428L)
Prassanthi Muralidharan (A0123403Y)
Ramki Ravindran (A0123433U)

1. INTRODUCTION

The United States of America remains an admired place of choice for a remarkable number of people especially students and skilled professionals from all over the world to seek a better life. We intend to develop a suggestion system that would help immigrants/migrants in identifying suitable places to live in the United States. The primary objective involves analysing the accident, crime and weather statistics of each city in United States and provide a list of safe cities, which are filtered on the specifics of the individual's profile and requirements. Scores for each city are assigned based on clustering and performing statistical analysis of various factors such as age, alcoholic/non-alcoholic, gender, marital status, preferred driving speed, health issues, preferred area of residence and race etc., The result of this analysis would suggest a prioritised list of places that are conducive for a person to live in United States.

We have also proposed an "*Alcoholic Accidents Eluding System*" that predicts temperature range specific to a particular location during which alcohol consumption would be high and alerts the Highway Patrol Department when that particular temperature range is reached to employ stricter breath analysis testing.

1.1 Why we should care?

"*Prepare and Prevent", don't "repair and repent*"

To influence on the obscured factors that go unnoticed which will bring in an adequate difference in the way safety is perceived. In this project, an *interactive application* is designed to provide a personalized list of suitable counties/cities in United States based on the user's profile. We tend to build a system that can rank the counties and cities in United States that provides the best fit for an individual.

1.2 How does this influence the impact in Business firms?

- Analyzing the range of temperature and accidents caused due to alcoholic driving, we suggest a system called "*Alcoholic Accident Eluding System*" which helps to alert Highway Patrol to increase the checks for alcohol detection when the particular range of temperature reaches. Thereby avoiding alcoholic accidents.

- On an average every year in United States, around 15% of the accidents occur due to driver distraction out of which 12% are caused due to sun glare. In our project, we analyzed how the accident rate in each county can be used to influence the sales of *Polarized Glass.*

2. DATA

- **Accident Data:** City-level accident data is collected from *Fatality Accident Reporting System(FARS)* spanning from 2009 till 2012.
- **Crime Data:** A combination of Python and BeautifulSoup was used to scrape city level crime data across United States from http://www.city-data.com/crime/. The dataset comprised of violent crime(rape, murder, assault) and property crime(arson, burglary, auto theft, theft, robbery).

  Racial crime data is obtained from the open dataset from the *census.gov* website

  Child Abuse data is obtained from the open dataset from *Administration for Children and Families* website
- **Temperature:** Daily temperature data is collected from http://www.usclimatedata.com/climate/united-states/us for every city using *webscraper* plugin.
- **Alcohol Consumption :** State level alcohol consumption data obtained from *cdc.gov* website
- **Polarized glass sales :** State level optical sales establishments data and polarized glass sales data obtained from the Industry Statistics Portal in *census.gov* website.

2.1. Data Preparation

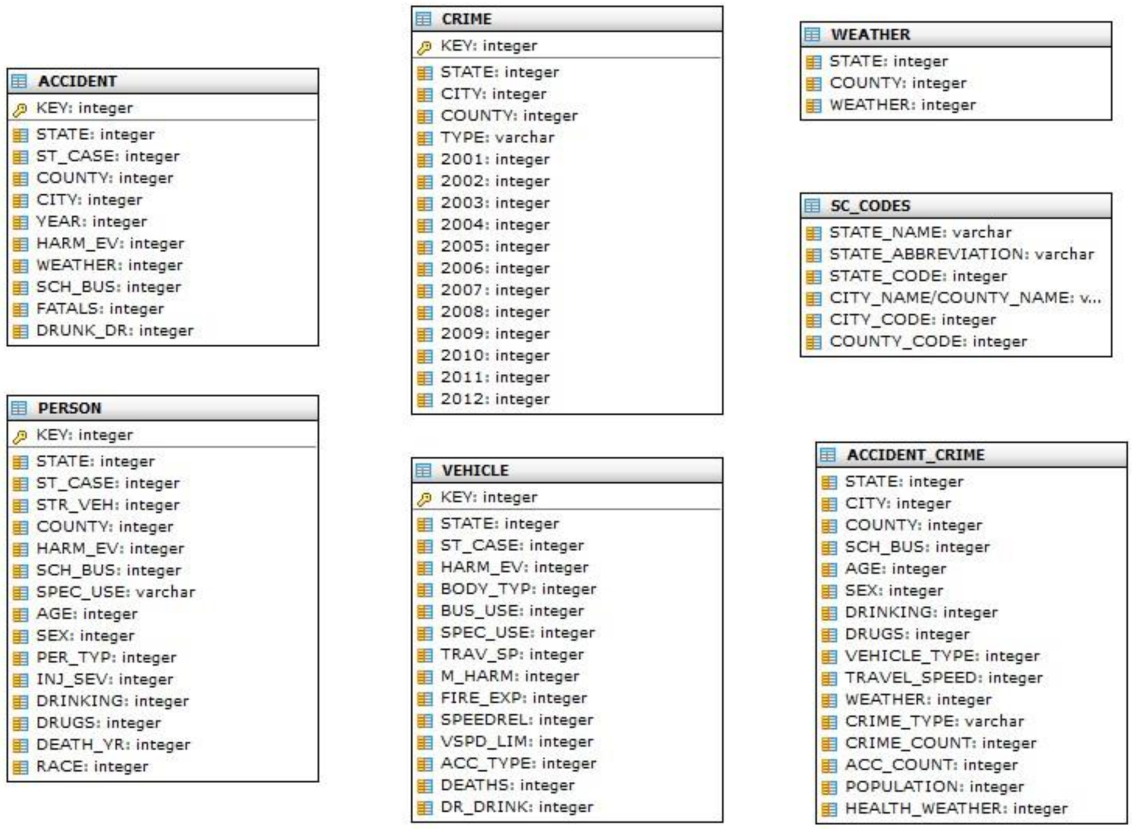
The dataset consisted of heterogeneous types of data making the data cleaning a challenging task. Also, the dataset being huge, the data mashup was tedious as most variables were unrelated. Data cleaning is performed using tools (R/Excel) and by manual checking.

2.2 Data consistency check

Preliminary analysis identified many outliers as below,

- Fatality data with missing data fields.
- Crime data having missing values for certain attributes for a particular year.
- Certain attributes with least significance towards the project scope is removed as they only represent a very minimal percentage of the dataset.
- As multiple datasets are involved identifying the cohesion among the dataset and standardizing these records is of critical importance.
- Unifying the city name to bring in a commonality in a way the data records are identified.
- As the data collected for accidents, crime, weather were all in different units of measure, to stabilize the data between cities, we used the city population under different age groups as a weighting measure and normalized the data to per 10000.

2.3 Database and Schema



4. ANALYSIS AND KEY INSIGHTS

4.1 PERSONALISED SUITABLE PLACES

4.1.1 Motivation

"Safety" is a significant factor considered while migrating to a place. Accident and crime rate have a major influence on the safety aspects of a place. But not all accident and crime types are relevant to an individual. For example, an adult who is single need not worry about the child abuse rate in the place he wants to move in. In existing, various online websites provide accident, crime and weather statistics separately to rank the places based on safety factor. In our growing world of personalization, where everything from an iota to a grandeur is customised, important aspects like safety also has to be personalized according the person's details.

4.1.2 Categorization of Data

Factors considered to personalize crime, accident and weather statistics:

- **Accident**:

  The attributes such as driver age, vehicle speed, driver gender, type of vehicle involved in each accident are considered to personalize it to a specific profile.

- **Crime**:

  The crime rates are grouped based on Violent crime (rape, murder, assault), Property crime (arson, burglary, auto theft, theft, robbery), Racial crimes (Anti-Asian, Anti-Islamic, Anti-Black, Anti-Jewish, Anti-Hispanic or Latino) and Child abuse.

- **Weather**:

  Weather data is categorized based on the health conditions of a person. For instance, a person suffering from lung disease will not prefer cold, rain or snowy weather.

4.1.3 Analysis

A linear model was built by categorizing "*accident_count*" rate of each county/city as dependent variable and attributes such as state, city, county, age, sex, drinking, drugs, weather, crime_type, crime_count and health_weather as independent variables. According to the model output below, the attributes state, city, county, sex, drugs, weather, crime_type, crime_count and health_weather were highly significant. Even though the attributes age and drinking were flagged as non significant, their p values contribute towards predicting the accident count.

```
Call:
lm(formula = ACC_COUNT ~ STATE + CITY + COUNTY + AGE + SEX +
    DRINKING + DRUGS + WEATHER + CRIME_TYPE + CRIME_COUNT + HEALTH_WEATHER,
    data = Result_data1)

Residuals:
    Min      1Q  Median      3Q     Max
-3179.3  -709.4  -123.1   399.6  2647.8

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.890e+03  3.389e+01 114.790  < 2e-16 ***
STATE          -1.507e+01  1.975e+00  -7.630 2.39e-14 ***
CITY            1.476e-01  3.430e-03  43.016  < 2e-16 ***
COUNTY         -1.278e+01  1.836e-01 -69.635  < 2e-16 ***
AGE            -3.900e-01  3.526e+00  -0.111  0.91192
SEX             6.475e+01  1.049e+01   6.174 6.70e-10 ***
DRINKING        1.168e+01  1.192e+01   0.980  0.32707
DRUGS           4.153e+01  1.572e+01   2.642  0.00824 **
WEATHER        -9.596e+00  1.075e+00  -8.929  < 2e-16 ***
CRIME_TYPE     -5.966e+01  2.817e+00 -21.175  < 2e-16 ***
CRIME_COUNT     3.754e-02  4.361e-04  86.086  < 2e-16 ***
HEALTH_WEATHER -8.211e+02  9.965e+00 -82.396  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1086 on 50402 degrees of freedom
Multiple R-squared:  0.3533  Adjusted R-squared:  0.3531
F-statistic:  2503 on 11 and 50402 DF,  p-value: < 2.2e-16
```

4.1.4 Data Visualization and App Development:

An interactive application is designed using Javascript,Php and Sqlite to provide user friendly interface that lists top counties preferences for a specific profile.[Refer Appendix Fig]

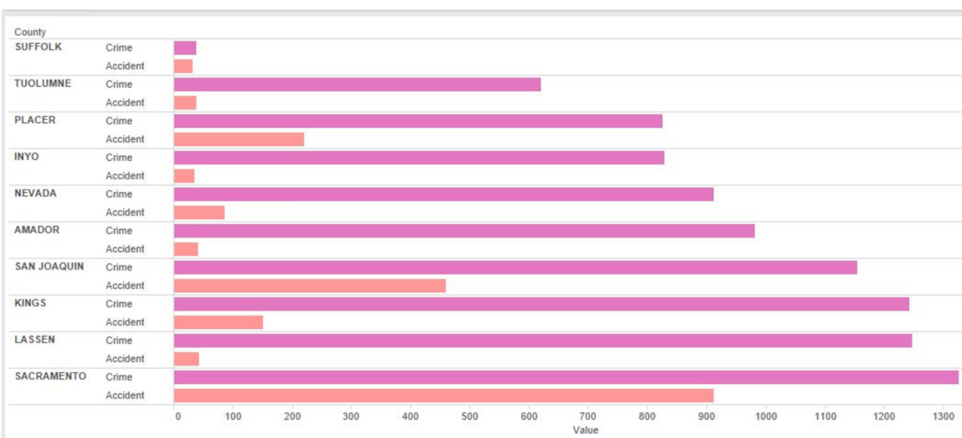


Fig 1: Data Visualization

4.2 ALCOHOLIC ACCIDENTS ELUDING SYSTEM :

4.2.1 Motivation:

*“DRINK AND **D**r**I**v**E**”*

Alcohol being a depressant has a strong influence on human stimulus and slowing down the brain activity. Drunk driving has been a serious safety epidemic for years. Drunk driving accidents contribute to nearly 31% of total accidents in United States.

4.2.2 Data

The data of fatalities caused due to alcohol involvement of drivers was collected along with the temperature and weather at that particular city on the date of accident.

4.2.3 Analysis

First we started correlating the *alcohol consumption* of the city to the number of *accidents caused by alcohol* in that cities of New York and California, and found a significant relationship among the two attributes. According to the US National Library of Medicine, annual seasonality in alcohol is high in summer. Considering this we tried to find a relationship by combining the *weather* and *alcohol consumption* of the city with the number of alcoholic accidents occurred. But from this study we found that most of the alcoholic accidents took place during clear sky. In order to make our analysis more intuitive and detailed, we decided to include the *temperature of a city* on the day of accident. Further we are trying to find out a range of temperature which induces a person to consume alcohol than usual leading to more accidents as a result of it, by clustering the existing alcoholic accident fatals, alcohol consumption rate and weather data with temperature data. Based on our cluster results we found that people tend to drink more when the temperature lies between *67 to 64°F*.

```
Call:
lm(formula = FATALS ~ DRINKING, data = data1)

Residuals:
    Min      1Q  Median      3Q     Max
-0.7008 -0.0364 -0.0364 -0.0364  3.2992

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.7595240  0.0308370   24.63   <2e-16 ***
DRINKING    0.0061521  0.0004752   12.95   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.397 on 860 degrees of freedom
Multiple R-squared:  0.1631, Adjusted R-squared:  0.1621
F-statistic: 167.6 on 1 and 860 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = FATALS ~ DRINKING + WEATHER + TEMP, data = data1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.31635 -0.03251 -0.03130 -0.03026  2.54232

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.306e-02  4.870e-02   1.705 0.088480 .
DRINKING     2.116e-02  6.450e-04  32.801  < 2e-16 ***
WEATHER2    -1.714e-01  5.109e-02  -3.355 0.000829 ***
WEATHER3    -3.167e-01  1.453e-01  -2.180 0.029507 *
WEATHER4    -6.036e-01  2.057e-01  -2.934 0.003431 **
WEATHER5    -6.982e-01  1.045e-01  -6.683 4.22e-11 ***
WEATHER6    -9.852e-01  2.072e-01  -4.755 2.33e-06 ***
WEATHER8    -1.364e+00  2.085e-01  -6.543 1.04e-10 ***
WEATHER10   -1.623e+00  5.920e-02 -27.413  < 2e-16 ***
TEMP        -6.128e-05  5.284e-04  -0.116 0.907698
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2894 on 852 degrees of freedom
Multiple R-squared:  0.5592, Adjusted R-squared:  0.5545
F-statistic: 120.1 on 9 and 852 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = DRINKING ~ WEATHER + TEMP, data = data1)

Residuals:
     Min       1Q   Median       3Q      Max
-127.089   -4.419   -3.937   -3.183  104.847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 51.13566    1.90200  26.885  < 2e-16 ***
WEATHER2     7.88461    2.69836   2.922 0.003570 **
WEATHER3     9.37515    7.70361   1.217 0.223948
WEATHER4    22.25007   10.89177   2.043 0.041376 *
WEATHER5    27.06261    5.46717   4.950 8.94e-07 ***
WEATHER6    39.90364   10.91203   3.657 0.000271 ***
WEATHER8    58.91281   10.87954   5.415 7.97e-08 ***
WEATHER10   77.15840    1.70189  45.337  < 2e-16 ***
TEMP        -0.03012    0.02803  -1.075 0.282802
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.36 on 853 degrees of freedom
Multiple R-squared:  0.7115, Adjusted R-squared:  0.7087
F-statistic: 262.9 on 8 and 853 DF,  p-value: < 2.2e-16
```

The application which we are trying to develop will incorporate this feature and will notify highway patrol when the temperature in the particular region reaches 67 to 64°F. This situation might vary at different places depending on the climate or weather condition of that place. To develop a fine grained model, this idea can be extended to other states by performing more elaborate analysis on other factors like climate, precipitation, air density data collected for each state in city level.

4.2.4 Concept

On analysis, we found that at a particular range (64 - 67°F) the tendency to consume alcohol is more. In order to avoid crash due to alcohol impaired driving, we could design a system to enhance the ***alcoholic breath analyser patrol system.***

Monitor the hourly climate data, when the temperature reaches the particular range when alcohol consumption will be more, send an alert message to the police patrol to employ stricter breath analyser alcohol testing. Thereby, stopping drunk drivers from driving and averting further accidents. ***This system can be an effective measures that can help prevent injuries and deaths from alcohol-impaired driving.***

4.3. POLARIZED SUNGLASS VS ACCIDENTS

4.3.1 Motivation

Sun is a cohesive part of the environment that drivers interact with every minute. The number of fatalities caused by sun glare would be minimal when compared to other major influencers like alcohol when the statistics is taken into account. Glare is not listed as a category for checking off in most investigations. However, as visibility being one of the important factors that influence driving, any obstruction to visibility would cause fatal damages. So, the sun glare as a cause of accident has to be given importance.

Sun Glare will reflected on the road will get concentrated horizontally. Human eye cannot tolerate this vertical light and hence the visibility is obstructed. Using Polarized sunglass will block the vertical light from disturbing the visibility.

4.3.2 Data

The data of fatalities caused due to sun glare was collected along with the temperature and weather at that particular city on the date of accident.

4.3.3 Analysis

We started correlating the *temperature* of the city to the number of *accidents caused by sun glare* in that city. A significant positive relationship was found between the two. To bring in the sales of polarized glass into picture, the mean temperature of the city along the sales data of polarized glasses and the number of establishments selling those is considered for New York and California.

We found the correlation between the sales of polarised glass and number of accidents in that region due sun glare and found it to be negatively correlated, which means that the sales is less in the region where there are more accidents due to sun glare. Similarly the correlation between the number of polarised glass manufacturer’s establishments (ESTAB) in a city and the temperature of the city on the day of accident was found to be positively correlated. The temperatures taken was the average of high temperature in that city in a particular month in the year 2012.This positive correlation denotes that there are more establishments in a city where the temperature is quite high during most of the year.

Further to substantiate the idea, linear regression was used to compute how much each attribute relates or contributes to one another.

```
> cor(mydata$ACCIDENTS,mydata$Sales)
[1] -0.01489318
> cor(mydata$ESTAB,mydata$TEMP)
[1] 0.03696834
>
```

```
Call:
lm(formula = ACCIDENTS ~ TEMP, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-5680.0 -1622.2  -588.4   897.2  9646.1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1384.32    1626.52  -0.851  0.40034
TEMP          103.79      33.39   3.108  0.00367 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3003 on 36 degrees of freedom
Multiple R-squared:  0.2116, Adjusted R-squared:  0.1897
F-statistic:  9.66 on 1 and 36 DF,  p-value: 0.003667
```

```
Call:
lm(formula = Sales + ESTAB ~ TEMP + ACCIDENTS, data = mydata)

Residuals:
   Min     1Q Median     3Q    Max
-13498  -9795  -5294   3352  73306

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 22487.5156 10550.5742   2.131   0.0402 *
TEMP         -296.7765   241.5217  -1.229   0.2274
ACCIDENTS       0.7765     1.0704   0.725   0.4730
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19280 on 35 degrees of freedom
Multiple R-squared:  0.04221, Adjusted R-squared:  -0.01252
F-statistic: 0.7712 on 2 and 35 DF,  p-value: 0.4701
```

```
Call:
lm(formula = ACCIDENTS ~ ESTAB + TEMP + Sales, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-3862.8 -1099.6  -317.0   609.6  7588.7

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.329e+03  1.268e+03  -2.625 0.012874 *
ESTAB        1.572e+00  2.662e-01   5.905 1.14e-06 ***
TEMP         1.029e+02  2.444e+01   4.212 0.000176 ***
Sales        2.012e-02  1.884e-02   1.068 0.293138
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2165 on 34 degrees of freedom
Multiple R-squared:  0.6127, Adjusted R-squared:  0.5786
F-statistic: 17.93 on 3 and 34 DF,  p-value: 3.751e-07
```

Based on the linear regression results, it is found that the number of establishments and temperature in that region are significantly related to the accident number.

To get a fine grained analysis on accidents due to sun glare, further analysis has to be done considering the position of the Sun relative to the vehicle at the time of accident (zenith angle).

4.3.4 Concept

This analysis can be used as a marketing strategy by the optical industry to promote the sales of polarized glass by educating the people on the statistics of accidents caused by sun glare and how it can be averted by using polarized glass. Geographical conditions of the location has also be taken into account while doing targeted marketing.

5. Tools and Techniques used

- SQLite
- R
- Excel
- Python

6. Conclusion

“Little drops make the mighty ocean”

In summary, we have used the fatality data of United States and have emphasized on the minor factors that when looked at for the first time will not look huge. But these minor factors put together will bring in a major devastating effect.

7. Future Work

- In our current application, minimal emphasis is given for health . Medical expenses being more costly in United States, in future work high emphasis would be placed on including health insurance,health care facilities data in our analysis.
- To emphasize more on the personalisation, certain other aspects like property price (to take the economic condition of an individual into account) , tax rate (useful for working professionals), topography (to match the individual’s interest)
- Climate of a particular location is dependent on multiple factors such as temperature, precipitation, air density, etc., For the Alcoholic Accidents Eluding System, more fine grained analysis have to be done by analysing the aforementioned factors.

The source code and dataset can be found in the below link : https://www.dropbox.com/home/Group-7_HowBA_Deliverables

APPENDIX

R-CODE for Clustering and Linear Regression:

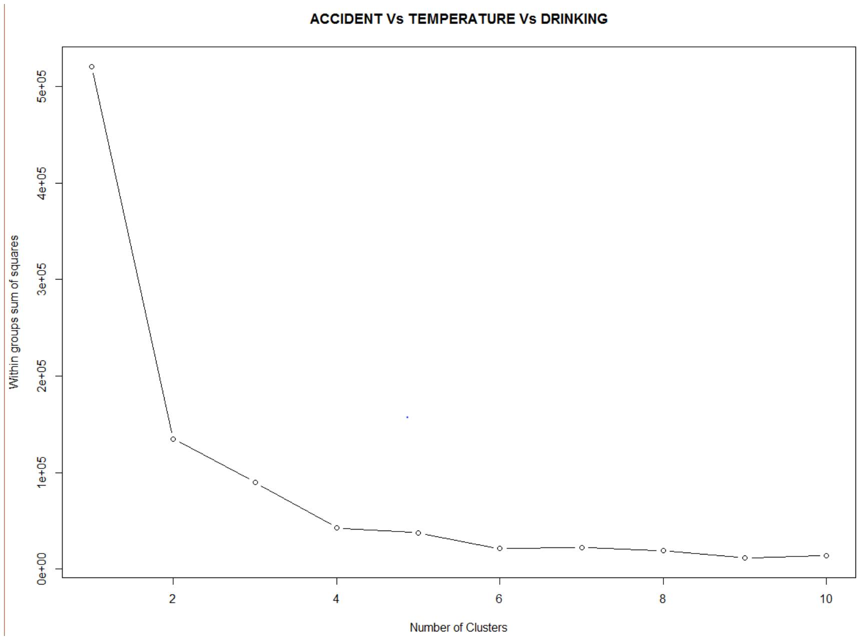
```
##################ACCIDENT VS WEATHER VS DRINKING##################
data1<-read.csv("Temperature_weather.csv", header=TRUE)
data1$TEMP_TYPE<-as.factor(data1$TEMP_TYPE)
data1$WEATHER<-as.factor(data1$WEATHER)
data<-data1
data[is.na(data)] <- 0
data$TEMP = NULL
#data1$TEMP_TYPE<-as.numeric(data1$TEMP_TYPE)
#data1$WEATHER<-as.numeric(data1$WEATHER)
is.numeric(data$TEMP_TYPE)
##########################################################
##########Finding the number of clusters#################
wss <- (nrow(data[,3:7])-1)*sum(apply(data[,3:7],2,var))
for (i in 2:10) wss[i] <- sum(kmeans(data[,3:7], centers=i)$withinss)
plot(1:10, wss, type="b", xlab="Number of Clusters", ylab="Within groups sum of squares",
     main="ACCIDENT VS TEMPERATURE VS DRINKING")

kdata<- kmeans(data[,3:7], 4,nstart=20)
plot(data1$TEMP,c(data1$DRINKING+data1$FATALS),col=kdata$cluster)
KMClusterMeans = aggregate(data1[,3:9],by=list(kdata$cluster),FUN=mean)
player.clustered <- data.frame(data1, kdata$cluster)
head(KMClusterMeans)
######################################################
############LINEAR REGRESSION###########################

LinearM1 <- lm(FATALS~DRINKING,data = data1 )
LinearM2<-lm(DRINKING~WEATHER+TEMP,data=data1)
LinearM3<-lm(FATALS~DRINKING+WEATHER+TEMP,data=data1)
summary(LinearM1)
summary(LinearM2)
summary(LinearM3)
scatterplot(FATALS~DRINKING, reg.line=lm, smooth= TRUE, data=data1,main="FATALITY VS TEMPERATURE_DRINKING",
            xlab="Fatality", ylab="DRUNK_DRIVE ", pch=16)
#Plot graph for Linear Regression:
residualPlots(LinearM1)
avPlots(LinearM1, id.n=2, id.cex=0.7)
qqPlot(LinearM1, id.n=3)
outlierTest(LinearM1)
influenceIndexPlot(LinearM1, id.n=3)
###############################################
```

CLUSTERING RESULTS:

Average value of attributes in each cluster

Group.1	DRUNK_DR	FATALS	DRINKING	TEMP
1	1	1.109375	59.8828125	64.14323
2	1	1.143791	57.44117647	66.67647
3	1	1.121951	55.09756098	60.81707
4	1	1.066667	57.6	59.33333

Calculating the number of clusters

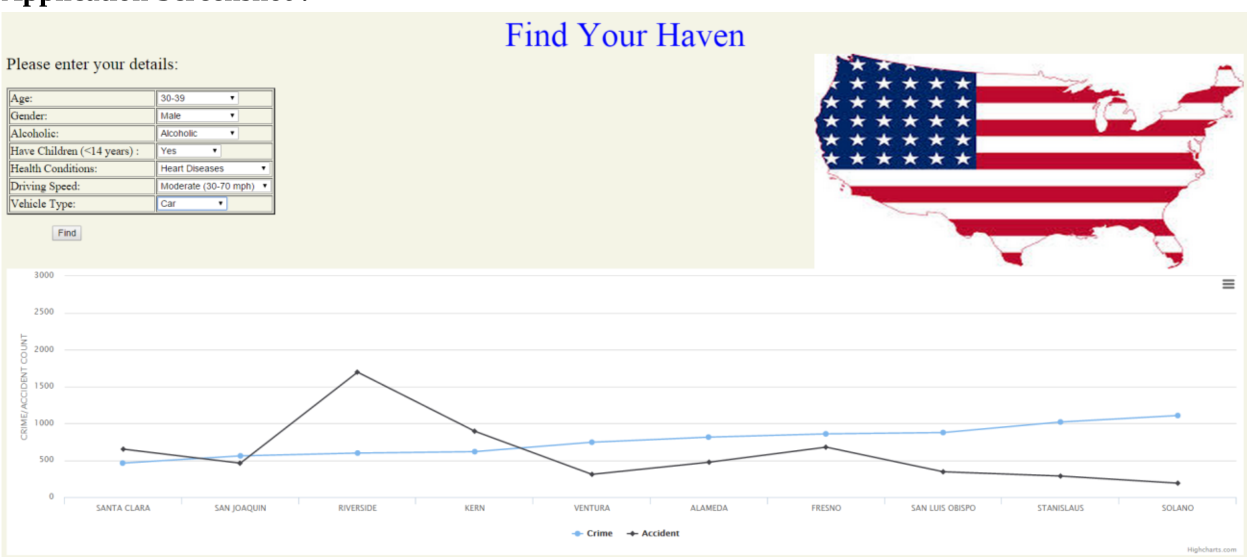


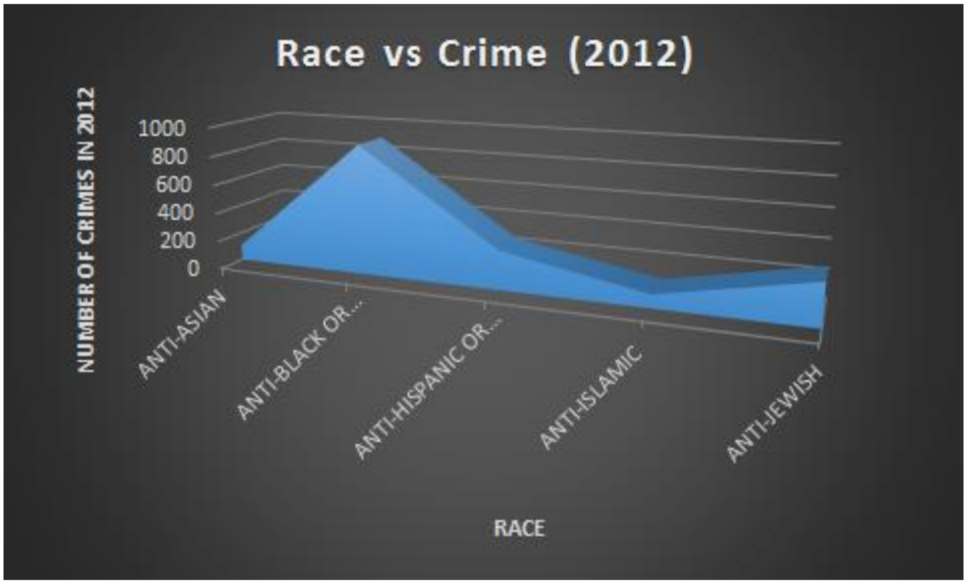
Health conditions vs Climate

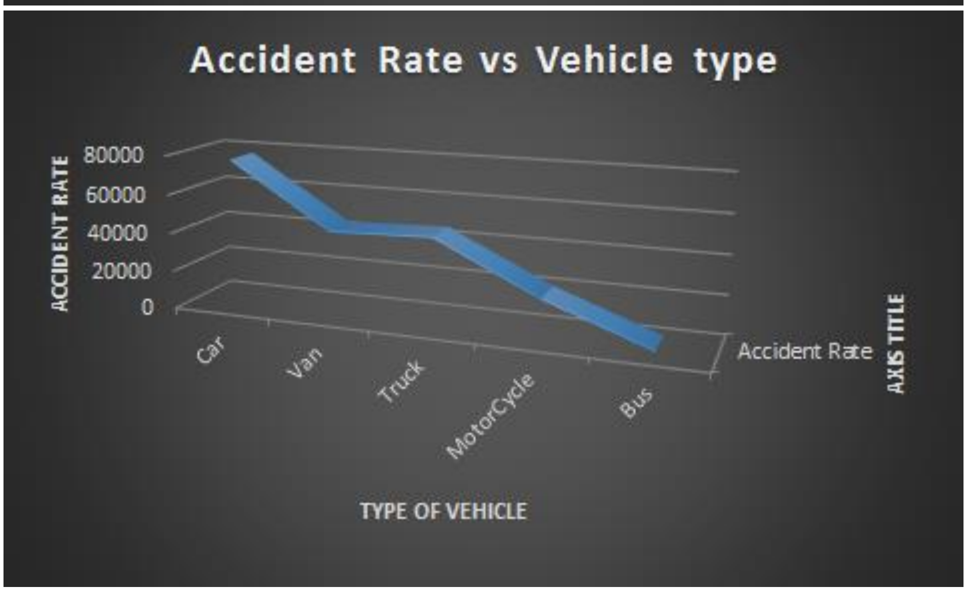
Health Condition	**Reason**	**Unfavorable Climate**
Heart attacks	For each 1.8F temp drops, heart attack rates increases	Cold,Rain,Snow
Obesity	In cold weather, Good brown fat activates thus burning the calories	Hot
Diabetes	Low pressure and cold weather affects blood sugar rate	Cold,Rain,Snow
Joint Pain	Cold weather increases the painful changes in the join fluid thickness	Cold,Rain,Snow
Headache	Sudden changes in Barometric pressure may cause headache	other than Best
Blood pressure	When atmospheric presssure decreases BP drops. BP is lower in summer	other than Best
COPD and Lung diseases	Hot humid weather will make	Hot

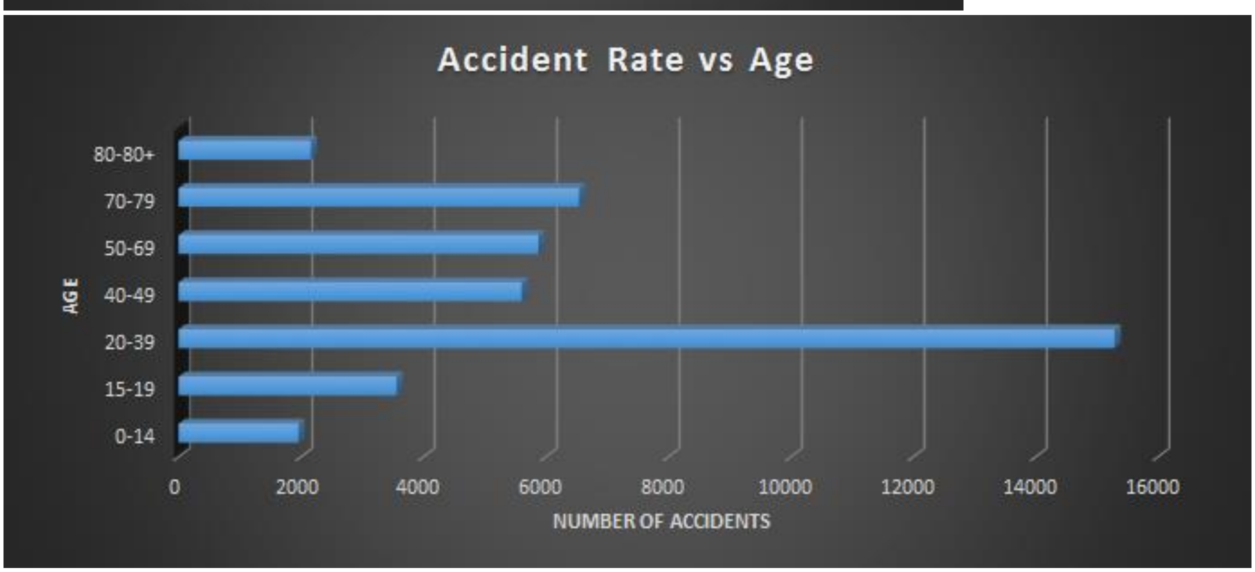
	breathing difficult	
Mental Disorders	May lead to death in hot weather	Hot
Cold and Flu	Cold virus transmits better in cold air	Cold,Snow,Rain
Attention Deficit Hyperactivity Disorder	Sunny regions will have more ADHD patients	Hot
Skin allergies	More prevalent in cold weathers and dry weathers	cold,dry
Eye problems	UV rays tend to reflect on the snow and affects eye	Snow

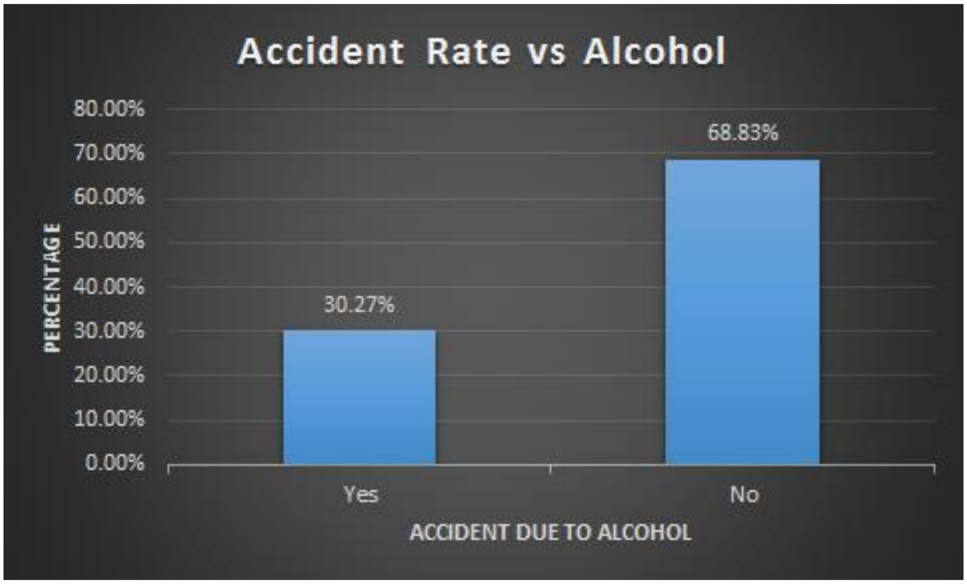
Application Screenshot :




Race vs Crime (2012)
NUMBER OF CRIMES IN 2012
1000
800
600
400
200
0
ANTI-ASIAN
ANTI-BLACK OR...
ANTI-HISPANIC OR...
ANTI-ISLAMIC
ANTI-JEWISH
RACE


Accident Rate vs Vehicle type
ACCIDENT RATE
80000
60000
40000
20000
0
Car
Van
Truck
MotorCycle
Bus
Accident Rate
AXIS TITLE
TYPE OF VEHICLE


Accident Rate vs Age
AGE
80-80+
70-79
50-69
40-49
20-39
15-19
0-14
0
2000
4000
6000
8000
10000
12000
14000
16000
NUMBER OF ACCIDENTS



Schema for Polarized glass sales vs Accidents: