ANALYSIS OF VEHICLE COLLISION INVETISGATED BY STATE POLICE

BHRUGEN PANDYA

HARIKRISHNA SHIYANI

SARANG KATKAR

SHAUMILKUMAR PATEL

MOTIVATION

- Nearly 1.3 million people die in road crashes each year across the globe, on average of 3,287 deaths a day. An additional of 20-30 million are injured or disabled.

- Road traffic crashes rank as the 9th leading cause of death and account for 2.2% of all deaths globally.

- Road crashes cost USD $518 billion globally, costing individual countries from 1-2% of their annual GDP.

PLAN OF ACTION

- COLLECTION OF DATA
- CLEANING THE DATA
- INITIAL ANALYSIS THROUGH BAR CHARTS
- PERFORMING MULTINOMIAL LOGISTIC REGRESSION & VALIDATING THE RESULTS
- PERFORMING POISSONS REGRESSION
- CLUSTER ANALYSIS
- CONCLUSION
- FUTURE SCOPE

DATA

- We collected the data of Vehicle Collision investigated by the State Police of Maryland State, USA.

- The data was collected for the year 2012 from the website of State of Maryland,

  https://catalog.data.gov/dataset/2012-vehicle-collisions-investigated-by-state-police-4fcd0

- Daily recorded data of Vehicle collision registers the Time, Date, Day, County Name, No. of Vehicles included in the collision, Route, Intersection, Distance from Intersection. It also informs whether the collision has caused the Injury and whether the Vehicle has reached the Proposed Destination.

CLEANING THE DATA

- Few of the cells in our dataset were blank, which did not appear as “NA”. So, the first step was to assign “NA” to the blank cells.

```
df=pd.read_csv('/Users/sarang/Desktop/2012_Vehicle_Collisions_Investigated_by_State_Police.csv', index_col=2, parse_dates=True)
```

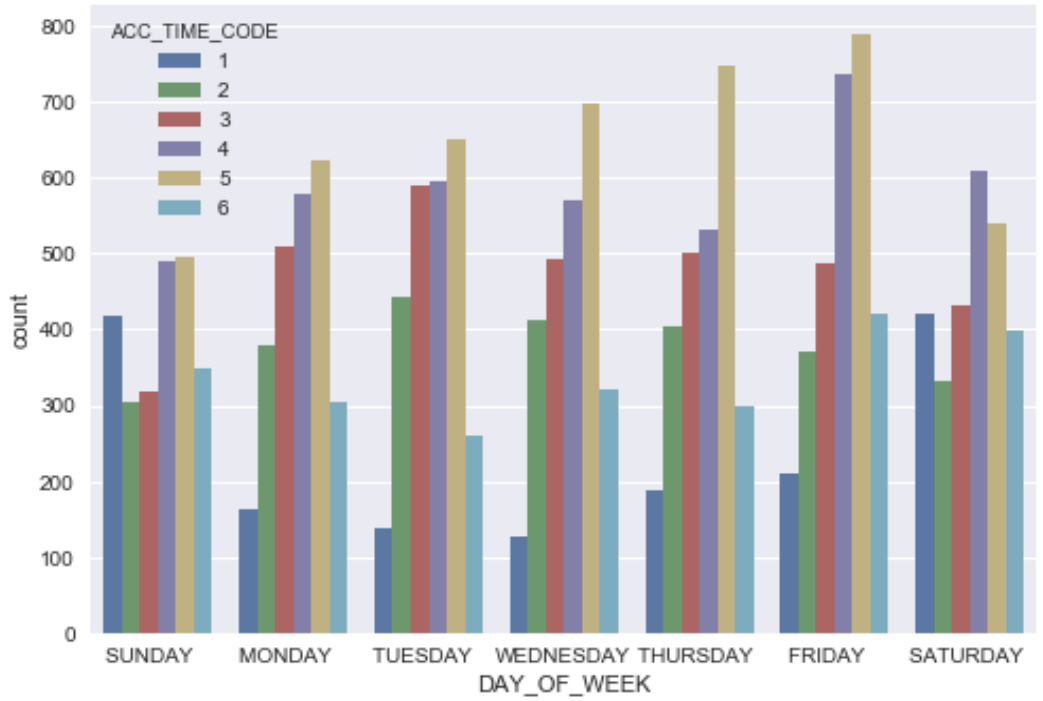
```
df=df.fillna(method='ffill')  with forward fill technique.
```

```
day_of_week=df['DAY_OF_WEEK'].astype('category')
```

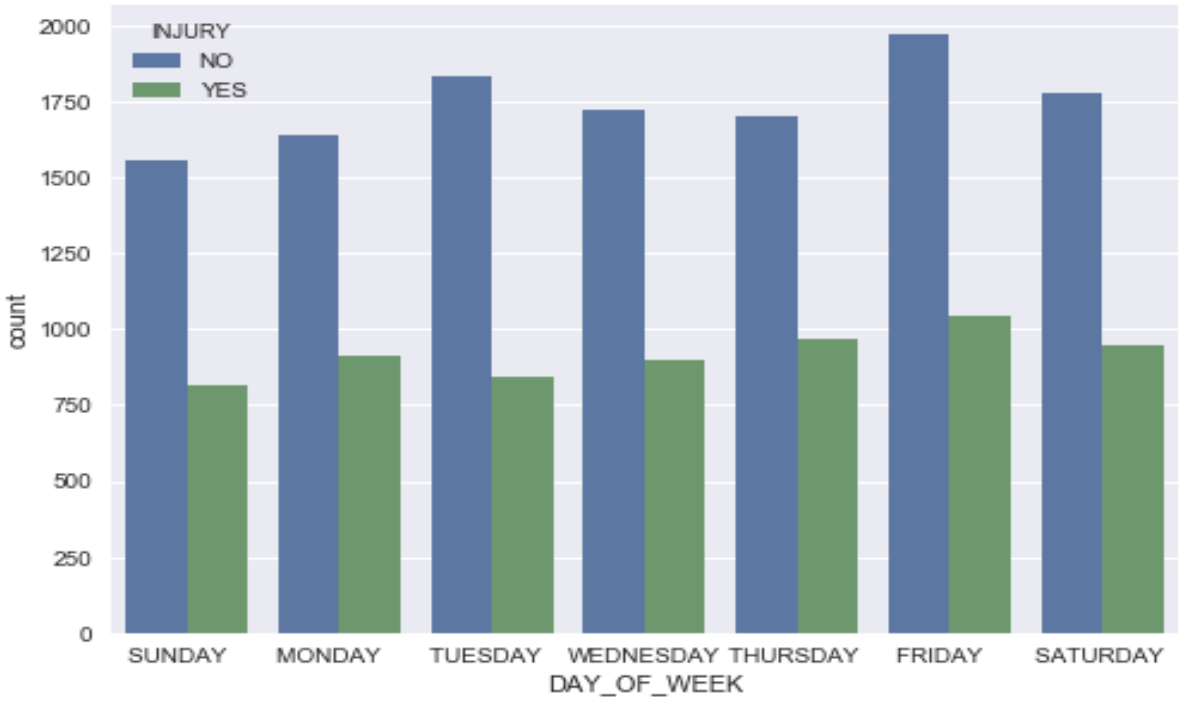
- Removing rows with NAs (missing values) in data frame using na.omit() funtion.

EXPLORATORY DATA ANALYSIS

Number of Accidents

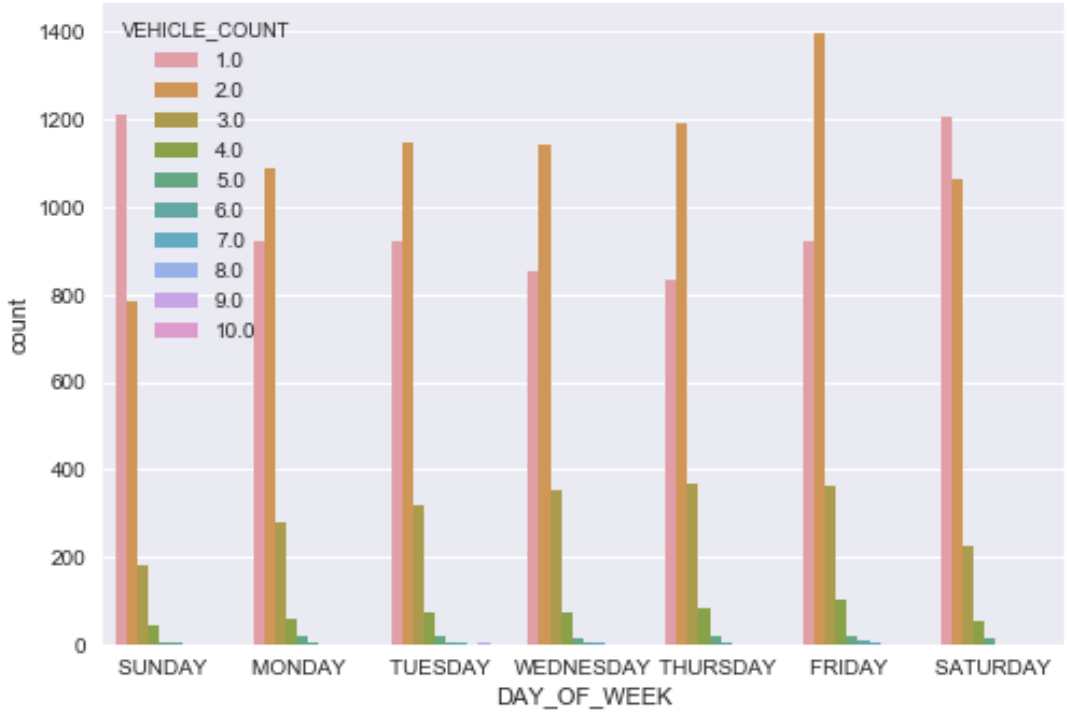


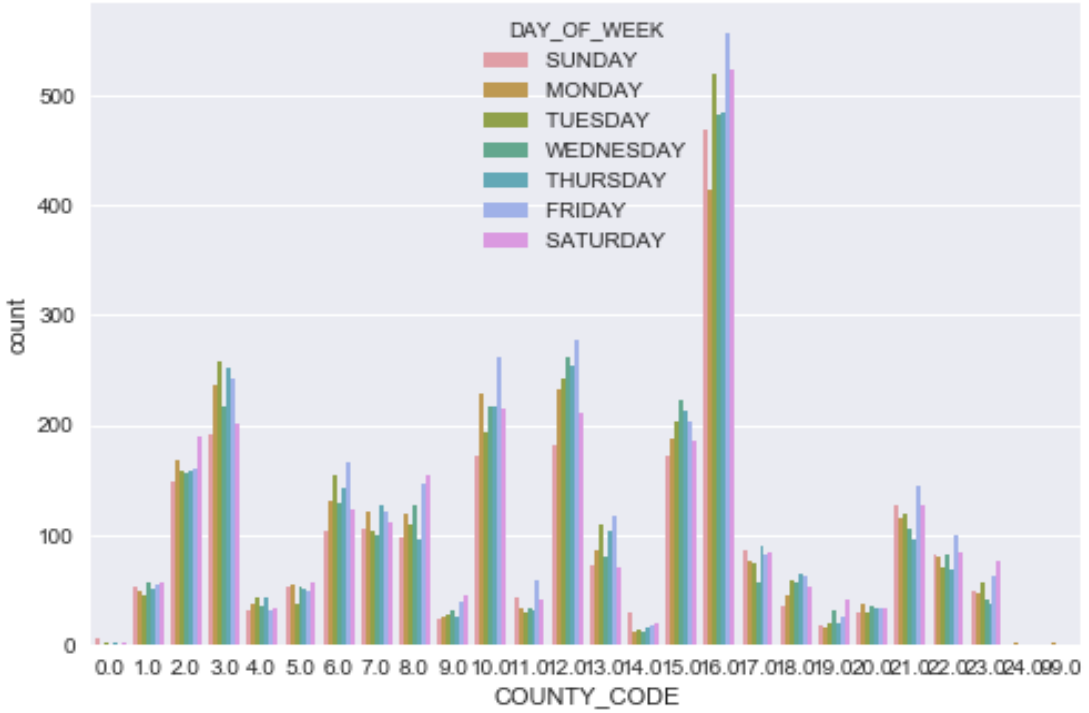
Day of week



>sns.countplot(x='DAY_OF_WEEK',y=None,hue='ACC_TIME_CODE', data=df)

>sns.countplot(x='DAY_OF_WEEK',y=None,hue='INJURY', data=df)






```
>sns.countplot(x='DAY_OF_WEEK',y=None,hue='VEHICLE_COUNT', data=df)
```

```
>sns.countplot(x='COUNTY_CODE',hue='DAY_OF_WEEK', data=df)
```

MULTINOMIAL LOGISTIC REGRESSION

- Logistic regression is a regression model where the dependent variable is categorical. We'll be using binary logistic model to estimate the probability of a binary response based on one or more predictor (or independent) variables (features)

- Multivariate analysis can be used to identify the effects of several factors on the causes of a vehicle collision compared with uni-variate analysis. The proposed multinomial logistic regression (MLR) model, to examine whether the person is injured considering various parameters.

- Multinomial logistic regression was used to investigate (1) INJURY and (2)PROP_DEST, which included Region, Time of Collision, Road, Intersection of Road, County Code, Vehicle Count, and Object 1 and Object 2 of Collision.

QUESTIONS THAT CAN BE ANSWERED WITH LOGISTIC REGRESSION

- Can the categories be correctly predicted given a set of predictors?
- What is the relative importance of each predictor?
- How good is the model at classifying cases for which the outcome is known?

Logistic regression taking Injury as Response variable

```
>fit<-glm(main$Inj~main$time+main$day+main$x_int+main$road+main$county+main$collision1+main$collision2+main$intersect)
> summary(fit)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4971  -0.3703  -0.2847   0.5995   0.8523

Coefficients:
                 Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)     1.197e+00   3.080e-02   38.853   < 2e-16 ***
main$time       7.228e-03   2.496e-03    2.896   0.00378 **
main$day       -2.854e-03   1.819e-03   -1.569   0.11674
main$x_int      5.671e-03   1.563e-02    0.363   0.71672
main$road       3.626e-05   7.868e-06    4.609   4.08e-06 ***
main$county     1.346e-03   6.140e-04    2.192   0.02839 *
main$collision1 2.707e-02   1.958e-03   13.829   < 2e-16 ***
main$collision2 -5.786e-03  3.799e-03   -1.523   0.12769
main$intersect  - 1.685e-05  2.551e-06   -6.607   4.03e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.2223702)

Null deviance: 3683.7  on 16262  degrees of freedom
Residual deviance: 3614.4  on 16254  degrees of freedom
AIC: 21713
Number of Fisher Scoring iterations: 2
```

Significant factors of this model are Time of the collision, Road, County , Object 1 of the Collision and Intersection point

Using highly correlated or significant predictors and developing better model

```
> summary(glm(main$Inj~main$time+main$road+main$county+main$collision1+main$intersect))
```

```
Call:
glm(formula = main$Inj ~ main$time + main$road + main$county + main$collision1 + main$intersect)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4891  -0.3713  -0.2847   0.6005   0.8552

Coefficients:
                  Estimate    Std. Error  t value     Pr(>|t|)
(Intercept)     1.166e+00    1.898e-02    61.449     < 2e-16 ***
main$time       7.120e-03    2.494e-03     2.855     0.00431 **
main$road       3.632e-05    7.863e-06     4.619     3.88e-06 ***
main$county     1.380e-03    6.137e-04     2.249     0.02456 *
main$collision1 2.655e-02    1.934e-03    13.727     < 2e-16 ***
main$intersect -1.697e-05    2.550e-06    -6.654     2.94e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.222397)

Null deviance: 3683.7  on 16262  degrees of freedom
Residual deviance: 3615.5  on 16257  degrees of freedom
AIC: 21712
Number of Fisher Scoring iterations: 2
```

Model Testing for Accuracy

```
# Dividing the dataset into TRAIN DATA and TEST DATA
>trainrows=sample(nrow(main),0.7*nrow(main))
>datatrain=main[trainsrows,]
>datatest=main[-trainrows,]

>actual=main$Inj
>predictmodel=predict(fit,main,type="response")
>predict=rep("Actual False", 4860)
>predicted[predictmodel>0.5]="Actual True"
>table(actual,predict)
                predict
 actual         Actual False   Actual True
   False.           3102             155
   True.             972             276
>conf_mtx=data.frame(table(actual,predicted))
>accuracy=sum(diag(conf_mtx))/n
 [1] 0.6950617
```

Logistic regression taking Proposed Destination as Response variable

```
>fitnew=glm(x$PROP_DEST~main$time+main$day+main$x_int+main$road+main$Inj+main$county+main$collision1+main$collision2+main$intersect,family="binomial")
> summary(fitnew)
```

```
Call:
glm(formula = x$PROP_DEST ~ main$time + main$day + main$x_int + main$road + main$Inj + main$county + main$collision1 + main$collision2 + main$intersect, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6829  -0.1609   0.2923   0.3746   3.0843
Coefficients:
                  Estimate        Std. Error      z value     Pr(>|z|)
(Intercept)      9.904e+00       3.197e-01       30.975      < 2e-16 ***
main$time        5.690e-02       2.607e-02        2.183       0.0291 *
main$day         2.693e-03       1.873e-02        0.144       0.8857
main$x_int      -1.299e-01       9.909e-02       -1.311       0.1897
main$road        5.565e-05       8.258e-05        0.674       0.5004
main$Inj        -6.941e+00       1.173e-01      -59.156      < 2e-16 ***
main$county      4.728e-03       6.402e-03        0.739       0.4602
main$collision1 -1.583e-01       2.192e-02       -7.223      5.07e-13 ***
main$collision2  9.791e-02       4.031e-02        2.429       0.0152 *
main$intersect  -1.753e-06       2.668e-05       -0.066       0.9476
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 21611.1  on 16262  degrees of freedom
Residual deviance:  5659.7  on 16253  degrees of freedom
AIC: 5679.7
Number of Fisher Scoring iterations: 7
```

Using highly correlated or significant predictors and developing better model

```
> summary(glm(main$Prop~main$Inj+main$time+main$collision1))

Call:
glm(formula = main$Prop ~ main$Inj + main$time + main$collision1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.97181  -0.00835   0.04758   0.06713   0.99886

Coefficients:
                    Estimate       Std. Error      t value      Pr(>|t|)
(Intercept)        2.8856053       0.0071684      402.546      < 2e-16 ***
main$Inj          -0.9221195       0.0033998     -271.227      < 2e-16 ***
main$time          0.0024021       0.0010825        2.219       0.0265 *
main$collision1   -0.0060892       0.0008453       -7.203     6.14e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.04200767)
    Null deviance: 3834.4  on 16262  degrees of freedom
Residual deviance:  683.0  on 16259  degrees of freedom
AIC: -5393.7
Number of Fisher Scoring iterations: 2
```

POISSON'S REGRESSION

- A Poisson regression model allows you to model the relationship between a Poisson distributed response variable and one or more explanatory variables. It is suitable for modelling the number of *events* that occur in a given time period or area.

- Since poissions regression is useful in the scenario where the response is the number counts we used the two significant factors day of week and accident time code in analyzing how these two affect the respponse that is the number of counts involved in the accident.

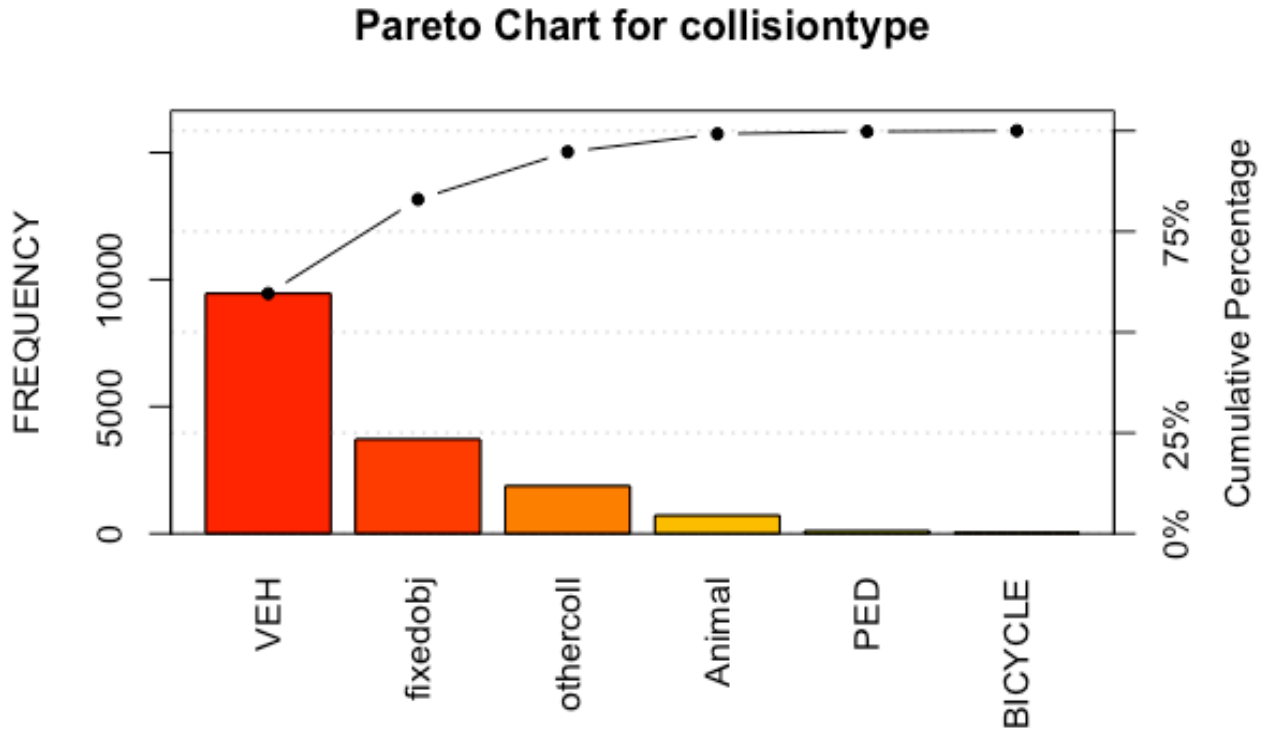
- **CODE USED:**

```
>r<-glm(VEHICLE_COUNT~DAY_OF_WEEK+ACC_TIME_CODE, data, family=poisson)
> summary(r)
```

```
r<-glm(VEHICLE_COUNT~DAY_OF_WEEK+ACC_TIME_CODE, family=poisson, data=data)
> summary(r)

Call:
glm(formula = VEHICLE_COUNT ~ DAY_OF_WEEK + ACC_TIME_CODE, family = poisson,
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8021  -0.6133   0.0571   0.2086   4.0256

Coefficients:
                                Estimate   Std. Error   z value   Pr(>|z|)
(Intercept)                     0.535843   0.020376     26.298    < 2e-16 ***
DAY_OF_WEEKMONDAY               -0.052468  0.020536     -2.555     0.0106 *
DAY_OF_WEEKSATURDAY             -0.117415  0.020498     -5.728    1.02e-08 ***
DAY_OF_WEEKSUNDAY               -0.171958  0.021694     -7.926    2.26e-15 ***
DAY_OF_WEEKTHURSDAY              0.006197  0.019914      0.311     0.7556
DAY_OF_WEEKTUESDAY              -0.021517  0.020100     -1.070     0.2844
DAY_OF_WEEKWEDNESDAY            -0.009448  0.020100     -0.470     0.6383
ACC_TIME_CODE                   0.027617   0.003801      7.265    3.73e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 6124.0  on 17386  degrees of freedom
Residual deviance: 5946.7  on 17379  degrees of freedom
  (1251 observations deleted due to missingness)
AIC: 48431

Number of Fisher Scoring iterations: 4
```

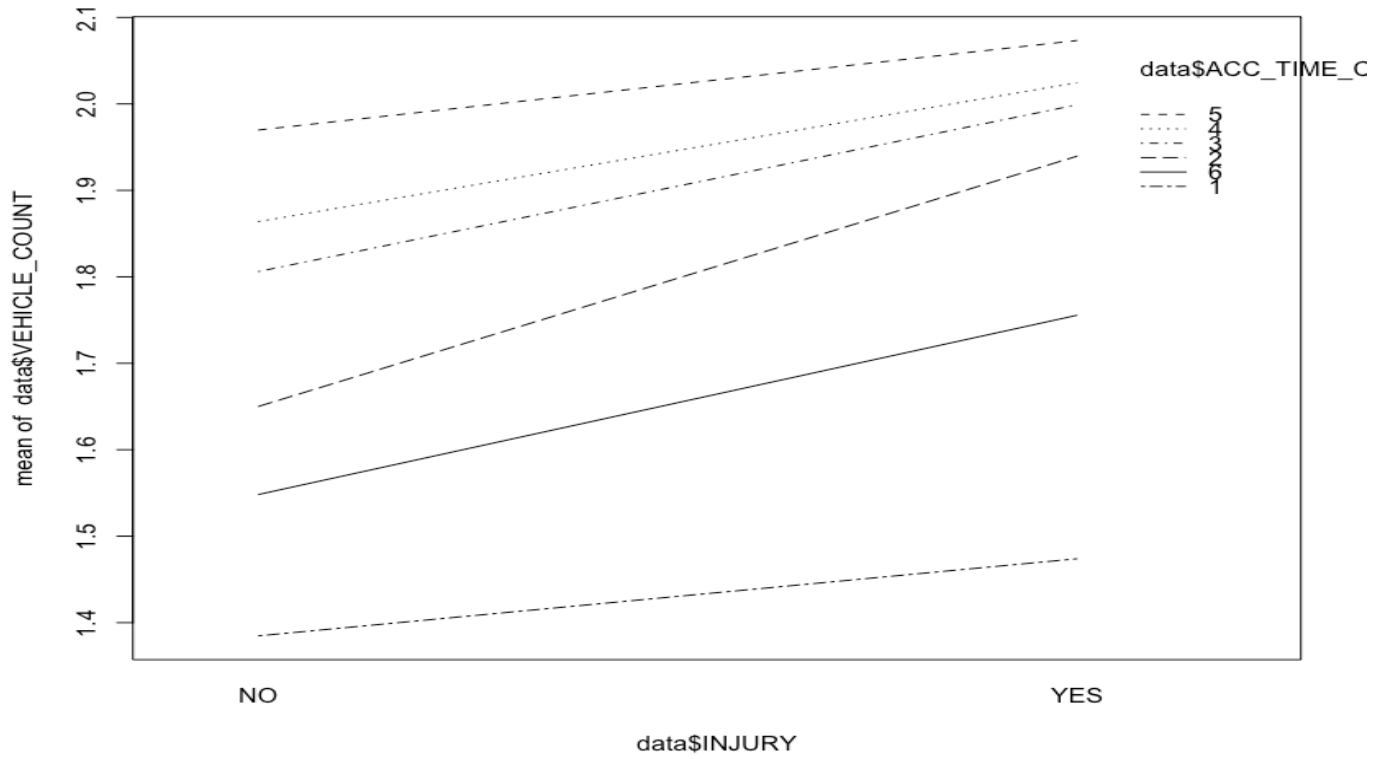
PARETOCHART.

```
>pareto.chart(collisiontype,  ylab = "FREQUENCY", col=heat.colors(length(collisiontype)))
```



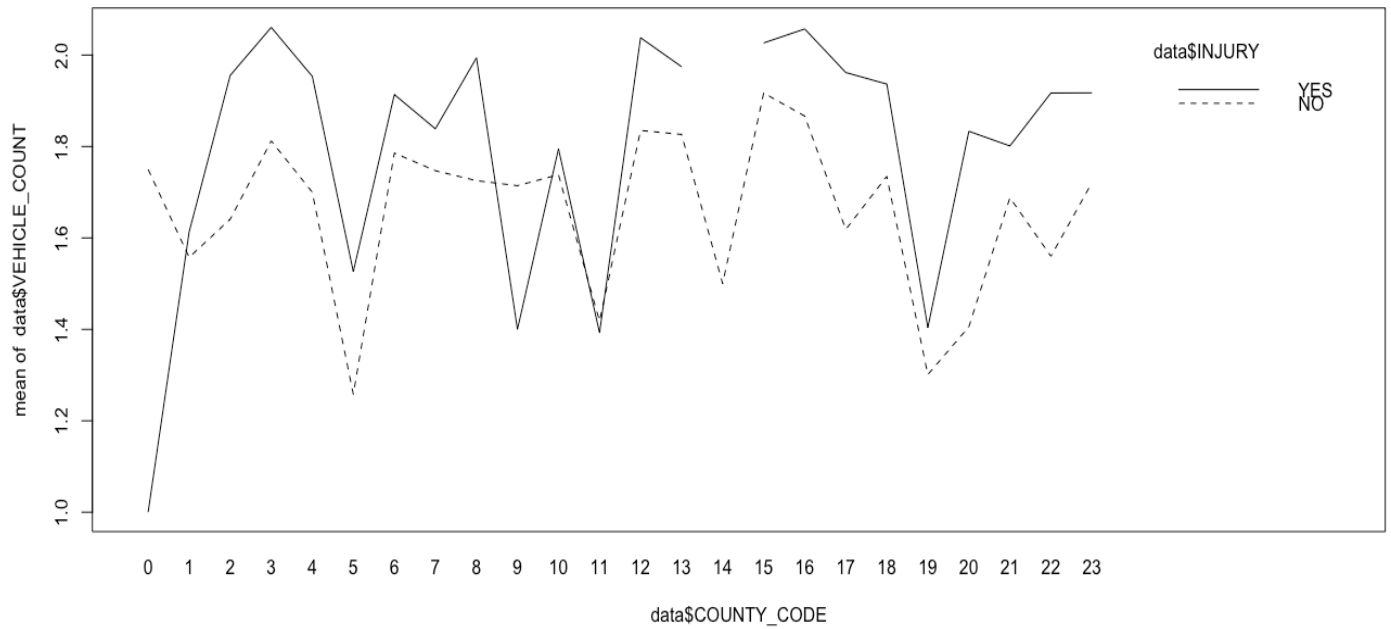
RELATIONSHIP BETWEEN INJURY VS ACCIDENT TIME , VEHICLE COUNT

```
>interaction.plot(data$INJURY, data$ACC_TIME_CODE, data$VEHICLE_COUNT)
```



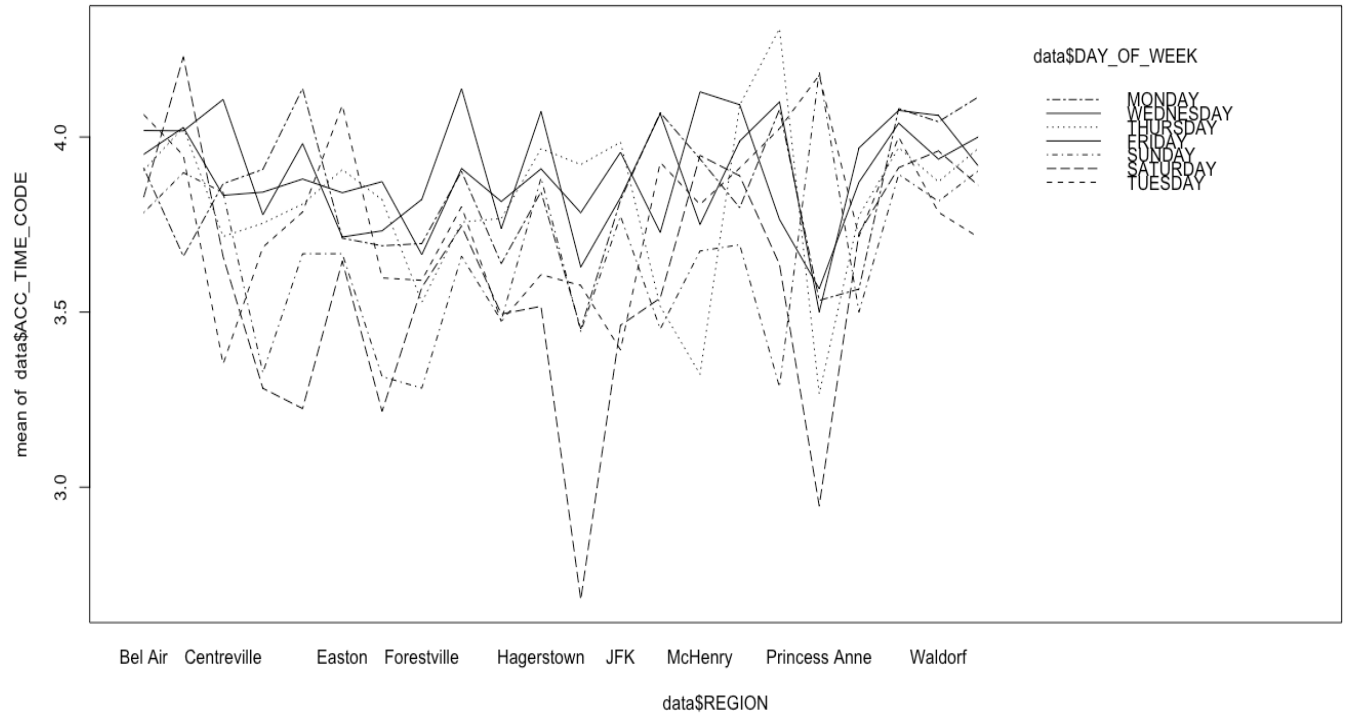
INTERACTION PLOT BETWEEN THE TYPE OF COUNTY AND INJURY CONDITION ON NUMBER OF VEHICLES COUNT

```
>interaction.plot(data$COUNTY_CODE, data$INJURY, data$VEHICLE_COUNT)
```



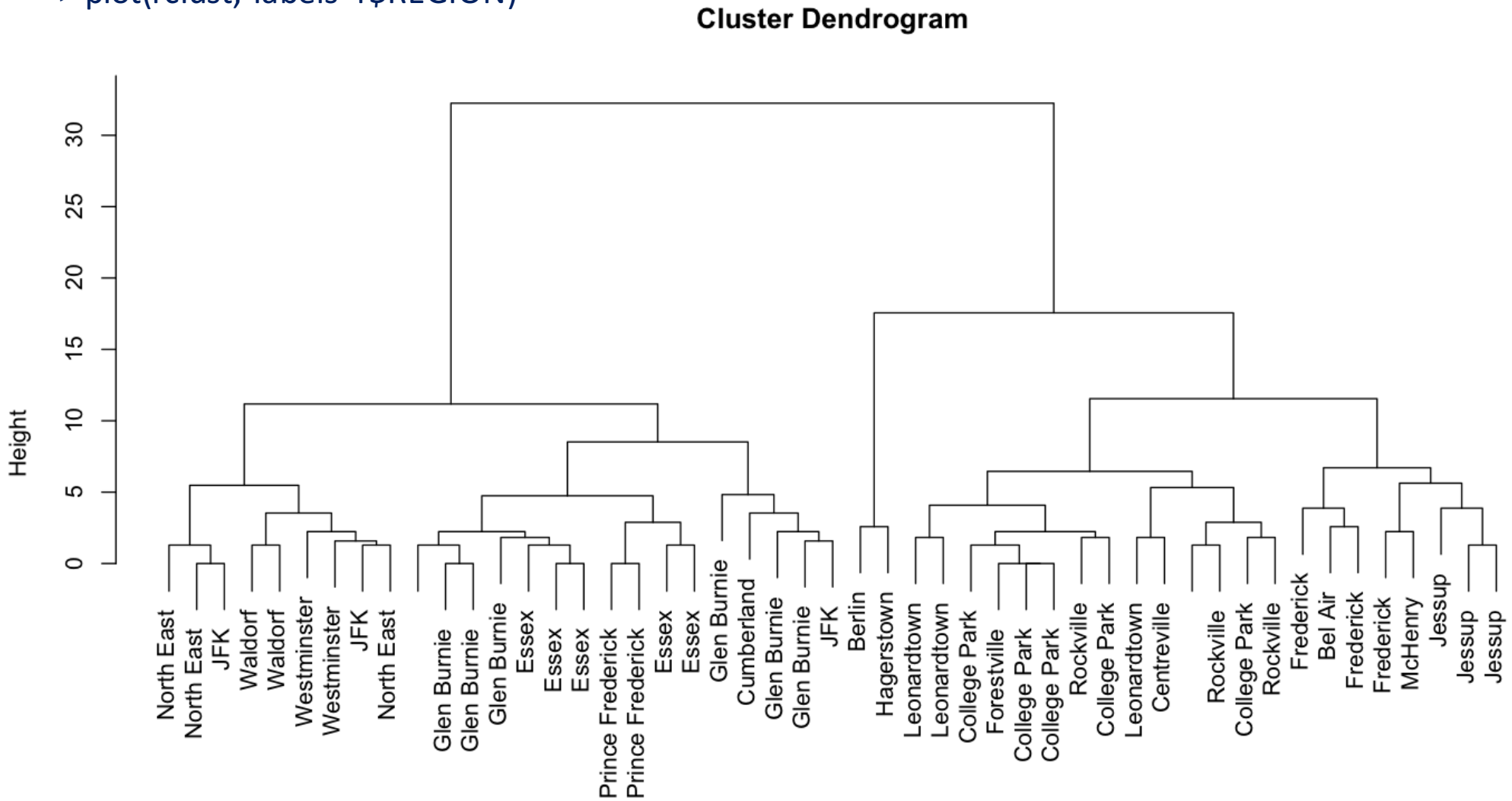
RELATIONSHIP BETWEEN TIME OF ACCIDENT VS REGION-DAY OF WEEK

```
>interaction.plot(data$REGION, data$DAY_OF_WEEK, data$ACC_TIME_CODE)
```



CLUSTER ANALYSIS

```
>rclust<-hclust(dist(r[-1]))
> plot(rclust, labels=r$REGION)
```



CONCLUSIONS

From the analysis performed, we can draw following conclusions:

- The factors Time of the accident, Road of the commute, county, Object 1 of the Collision and Point of Intersection plays a major role in causing the Injury to the person.

- The factors Injury caused to the person, Time of the accident and Object 1 of the Collision has significant impact of decision of person's ability to reached the Proposed Destination after Collision.

FUTURE SCOPE

- MACHINE LEARNING MODELS AND TECHNIQUES CAN BE DEVELOPED BASED ON THE APPROACH FOR ACCURATE RESULTS AND ENHANCED PREDICTING.

- PREDICTION RESULTS ACQUIRED FROM MACHINE LEARNING MODELS CAN HELP IN DEVELOPING BLACK SPOT ANALYSIS WHICH CAN FURTHER REDUCE THE COLLISION BY IMPLEMENTING SAFETY MEASURES.

- MODEL CAN BE FRUITFUL IN AVOIDING THE VEHICLE COLLISION BY CHANNELIZING THE TRAFFIC AWAY FROM BLACK SPOT AND PREVENTING ECONOMIC AND CASUALITY.

THANK YOU