US Traffic Accidents

Better Decision-Making Process Before Going on the Road

BUS 462 E100
Group 5
April 11
Presenter: *Davis Gao, Hailey Jun, Katherine Bui*

Meet the Team

Davis Gao
Quote: “I rear-ended a car”

Hailey Jun
Quote: “I hate potholes”

Katherine Bui
Quote: “I don’t have a car”

Table of Contents

Executive Summary

Q: What vehicle body types are associated with the lowest frequency and severity of traffic accidents, and which hours of the day are the most dangerous for driving in the US?

Why is it Interesting?

- Curiosity about the factors that affect accidents on the road, as this can affect ourselves, our families, and our friends.
- To give ourselves a better decision-making process before going on the road by determining the variables that highly correlate with these accidents

About our Data

Initial Data

kaggle

US Traffic Accidents

259K Car Accidents Data

2016 - 2020

15 Files, 1270 Columns

Each Year has 3 Files; acc, veh, pers

Data 1 - veh_20.csv

Information on Vehicle Involved in Accidents

2020

167 Columns

94,700 + Rows

Data 2 - pers_20.csv

Information on Persons Involved in Accidents

2020

104 Columns

132,000 + Rows

About our Data

1. Data Cleaning Process

Removed unnecessary column in Microsoft Excel
- 61 columns from veh_20
- 20 columns from pers_20

Filtered out all person except drivers
- 94,700 rows in veh_20
- 94,500 rows in pers_20

Cleaned rows with inapplicable data points (ex. Unknown or n/a)

2. Data Cleaning Process

Imported dataset into R

Assigned a new primary key to each row in both datasets

Removed uncommon vehicle body types (ex. Off-road vehicles, school buses)

Further narrowed down the range of body types to 9 categories

About our Data

1. Data Cleaning Process

Removed unnecessary column in Microsoft Excel
- 61 columns from veh_20
- 20 columns from pers_20

Filtered out all person except drivers
- 94,700 rows in veh_20
- 94,500 rows in pers_20

Cleaned rows with inapplicable data points (ex. Unknown or n/a)

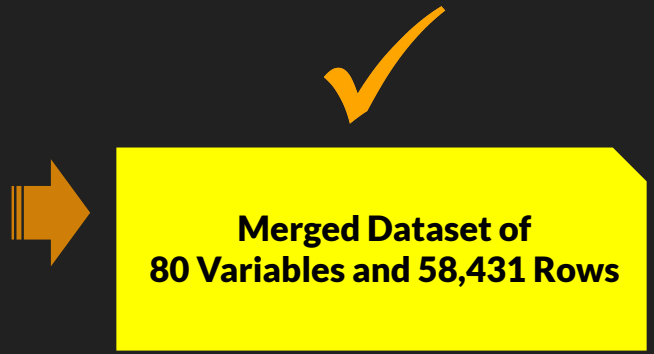
2. Data Cleaning Process

Imported dataset into R

Assigned a new primary key to each row in both datasets

Removed uncommon vehicle body types (ex. Off-road vehicles, school buses)

Further narrowed down the range of body types to 9 categories



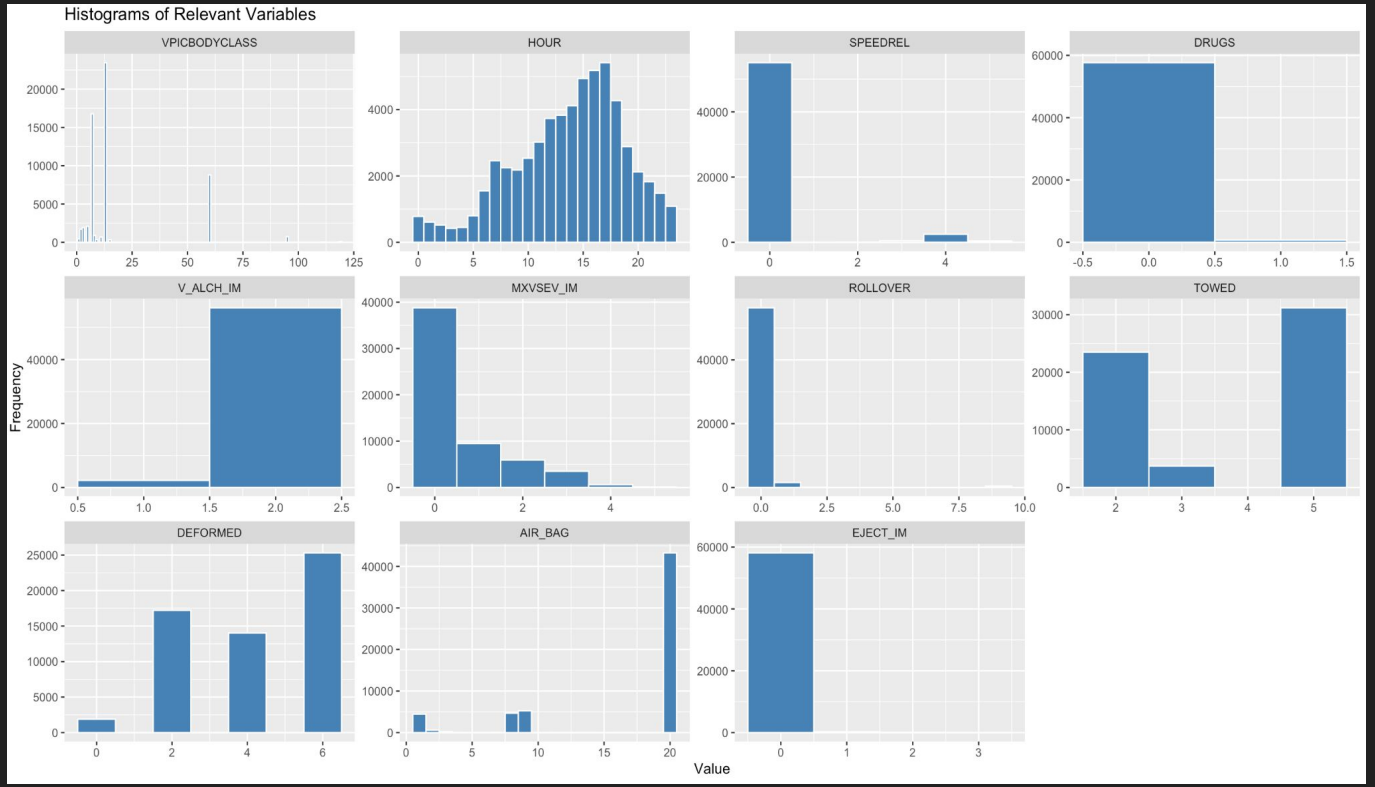
Hypotheses

1
Reckless drivers are more likely to drive coupes or convertibles

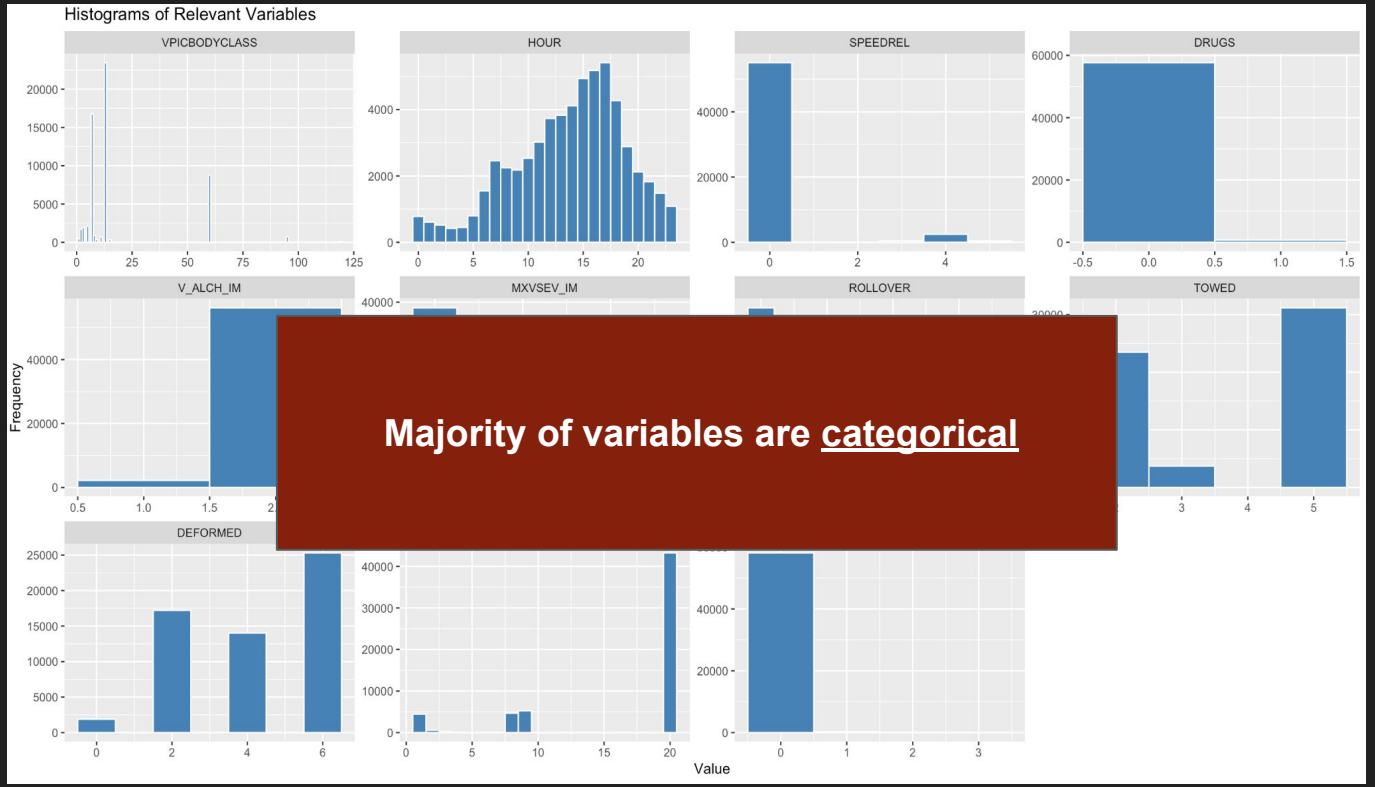
2
SUVs and minivans are less likely to be involved in serious car accidents

3
More severe car accidents occur past midnight (12:00 am - 3:59 am)

Preliminary Analysis

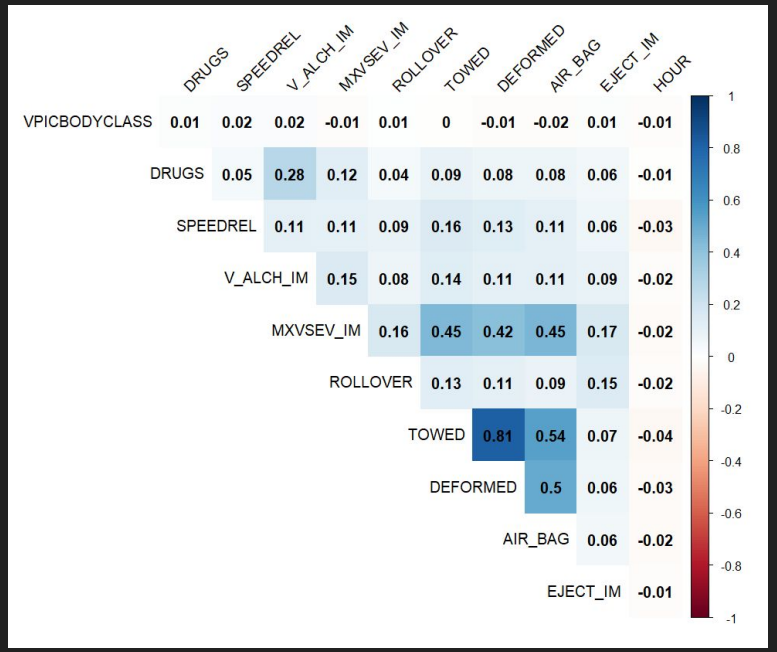


Preliminary Analysis



Correlations

- Positive correlations throughout the IV's
- Indicates as values increase so does the severity of the accident
- Can be used to predict our DV's



Hypotheses

1
Reckless drivers are more likely to drive coupes or convertibles

2
SUVs and minivans are less likely to be involved in serious car accidents

3
More severe car accidents occur past midnight (12:00 am - 3:59 am)

Hypothesis 1 - Variables Used

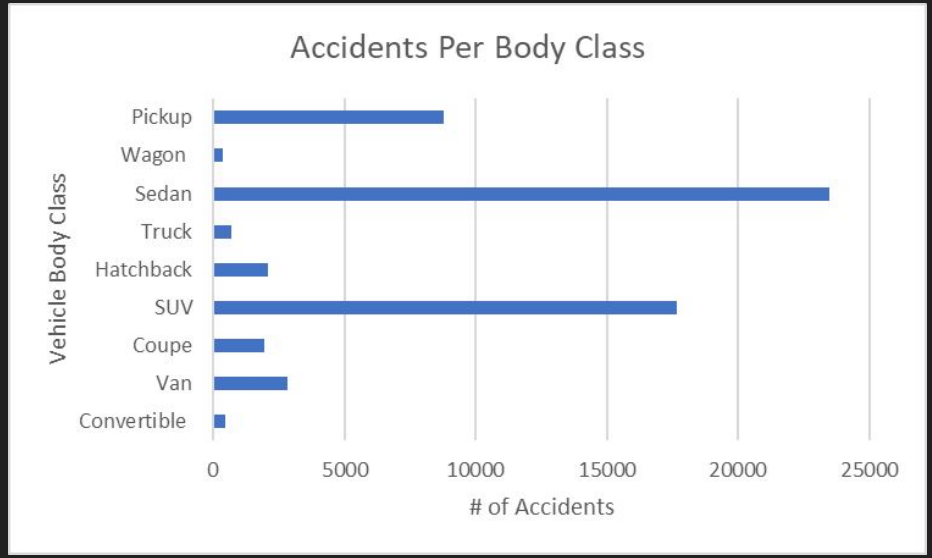
Dependent Variable

Vehicle Body Type

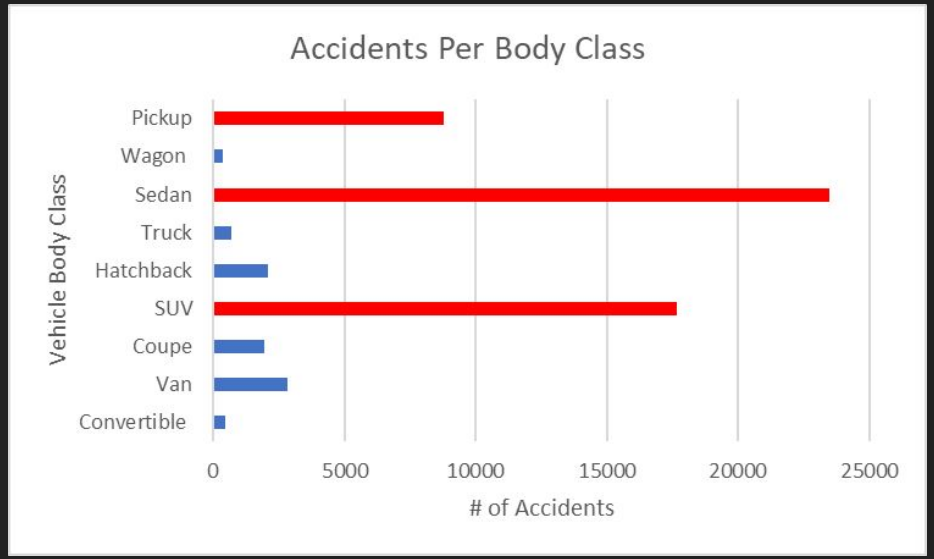
Independent Variables

- **Drugs**
- **Alcohol**
- **Speeding**

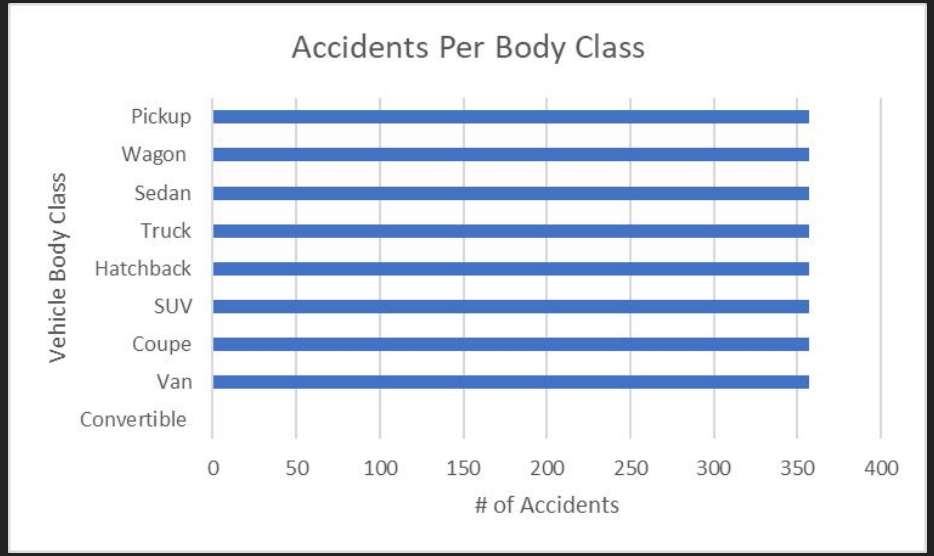
Hypothesis 1 - Downsampling



Hypothesis 1 - Downsampling



Hypothesis 1 - Downsampling



Hypothesis 1 - Modelling Approach

LOGIT

- DV has 9 categories
- GLM not best suited to be the model

```
Call:
glm(formula = VPICBODYCLASS ~ DRUGS + SPEEDREL + V_ALCH_IM, data = h1data_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.1257  -1.9963   0.0037   2.0037   4.0915

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.82062    0.56822   8.484   <2e-16 ***
DRUGS        0.10424    0.48782   0.214    0.831
SPEEDREL     0.11300    0.23681   0.477    0.633
V_ALCH_IM    0.08783    0.28706   0.306    0.760
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 6.701509)

    Null deviance: 15047  on 2248  degrees of freedom
Residual deviance: 15045  on 2245  degrees of freedom
AIC: 10667

Number of Fisher Scoring iterations: 2
```

Hypothesis 1 - Modelling Approach

Multinomial Logistic Regression vs. **Ordinal Logistic Regression**

Hypothesis 1 - Modelling Approach

	Multinomial Logistic Regression	vs.	Ordinal Logistic Regression
AIC:	9,885		9,904
Residual Deviance:	9,821		9,882
McFadden	6.17e-03		1.76e-05

Hypothesis 1 - Modelling Approach

	Multinomial Logistic Regression
AIC:	9,885
Residual Deviance:	9,821
McFadden	6.17e-03

Hypothesis 1 - Prediction

IV Prediction Set:

Drug - “Yes”

Alcohol - “Yes”

Speeding - “Yes”

Top Values

Coupes - 23.19%

Pickups - 17.75%

Hatchbacks - 17.73%

Convertibles - 12.69%

Hypothesis Values

Coupes - 23.19%

Convertibles - 12.69%

Hypothesis 1 - Prediction

Top Brands

Top Values

Coupes - 23.19%

Pickups - 17.75%

Hatchbacks - 17.73%

Convertibles - 12.69%

Hypothesis Values

Coupes - 23.19%

Convertibles - 12.69%

Hypothesis 1 - Variables Added

Independent Variables

- Drugs
- Alcohol
- Speeding

+

Control Variables

- Age
- Sex

Interaction Variable

- Speeding*Urban|Rural

Hypothesis 1 - Variables Added

Independent Variables

- Drugs
- Alcohol
- Speeding

+

Control Variables

- Age
- Sex

Interaction Variable

- Speeding*Urban|Rural

AIC: 9,885 → **9,511**

Hypothesis 1 - Prediction

IV Prediction Set:

Drug - “Yes”

Alcohol - “Yes”

Speeding - “Yes”

Age - 67

Sex - Male

Area - Rural

Hypothesis 1 - Prediction

IV Prediction Set:

Drug - “Yes”

Alcohol - “Yes”

Speeding - “Yes”

Age - 67

Sex - Male

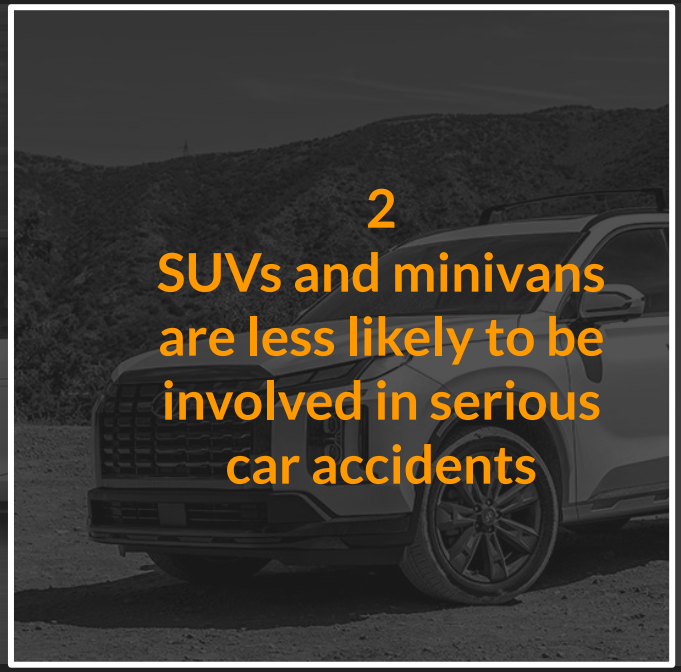
Area - Rural

Top Predicted Value

Pickups- 28.97%

Hypotheses

1
Reckless drivers are more likely to drive coupes or convertibles



2
SUVs and minivans are less likely to be involved in serious car accidents

3
More severe car accidents occur past midnight (12:00 am - 3:59 am)

Level of Accidents Severity for Hypothesis 2 & 3

Level 0	No injury, no rollover, not towed, no damage, not ejected, airbag not deployed
Level 1	Possible injury, rollover (tripped), towed (not due to disabling damage), minor damage, airbag not deployed, or any combination of these factors
Level 2	Suspected minor injury, rollover (untripped), towed (due to disabling damage), functional damage, partially ejected, airbag deployed (any type), or any combination of these factors
Level 3	Suspected serious injury, rollover (unknown type), towed (unknown reason), disabling damage, totally ejected, airbag deployed (any type), or any combination of these factors
Level 4	Fatal injury, injured (severity unknown), airbag deployed (any type), or any combination of these factors

Hypothesis 2 - Modelling Approach

Variables	Type	Description	Class
VPICBODYCLASS	**DV**	**Body type of vehicles**	**integer**
MXVSEV_IM	IV	The severity of injury	integer
ROLLOVER	IV	Whether or not the vehicle rolled over	integer
TOWED	IV	Whether or not the vehicle was towed	integer
DEFORMED	IV	The condition of the car involved in the accident	integer
AIR_BAG	IV	Whether or not the airbag deployed	integer
EJECT_IM	IV	Whether or not the driver was ejected from the vehicle	integer

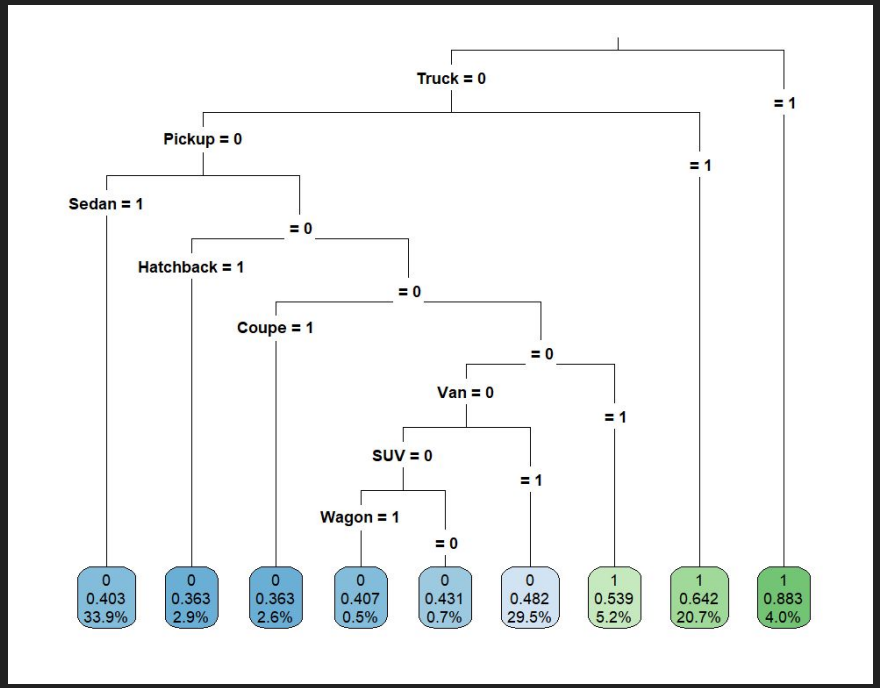
Modelling Approach

CART

- ✓ Categorical variables
- ✓ Easier to interpret
- ✓ Visualize interactions

Hypothesis 2 - Modelling Analysis

Severity Level 0




```
Variable importance
  Pickup     Truck     Sedan       SUV    Coupe Hatchback       Van
      43        40         6         5        2         2         1
```

Most Involved ⟷ Least Involved

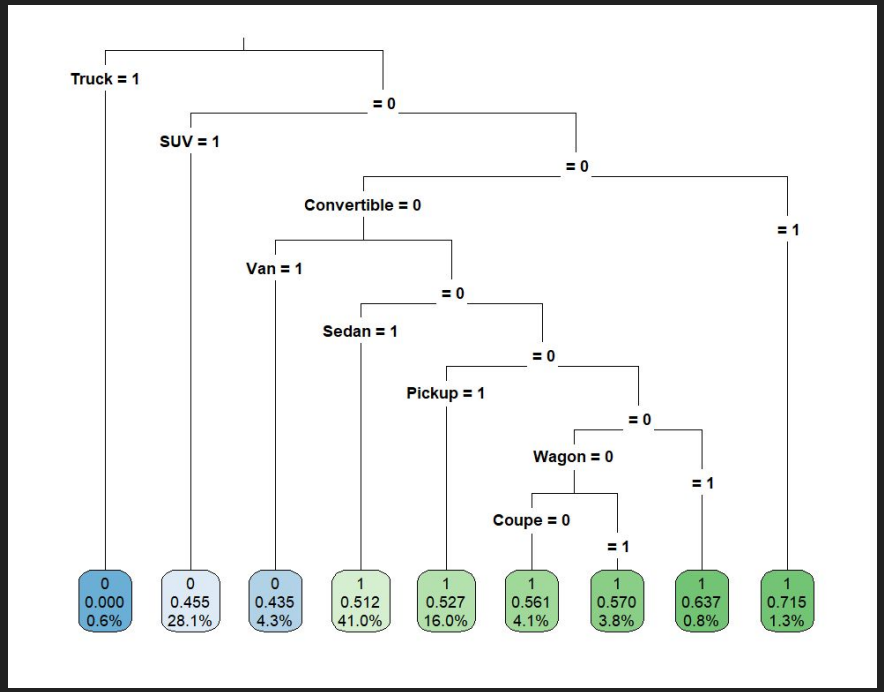
Truck Pickup **Van** Convertible **SUV** Wagon Sedan Hatchback Coupe

Accuracy Test
0.782079139981702

Confusion Matrix		
Severity Level 0	**0**	**1**
0	**13477**	**3500**
1	**311**	**200**

Hypothesis 2 - Modelling Analysis

Severity Level 4




```
Variable importance
      Truck         SUV Convertible         Van
         46          28          16          10
```

Most Involved ⟷ Least Involved

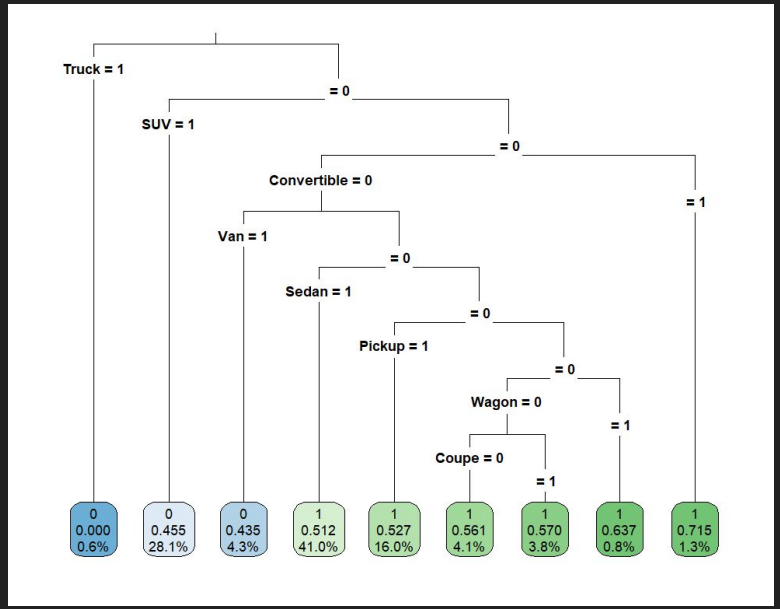
Convertible Wagon Hatchback Coupe Sedan Pickup **Van** **SUV** Truck

Accuracy Test
0.366193961573651

Confusion Matrix		
Severity Level 4	**0**	**1**
0	**6250**	**11022**
1	**62**	**154**

Hypothesis 2 - Conclusion

Severity Level 4



Most Involved

Least Involved

Convertible Wagon Hatchback Coupe Sedan Pickup **Van** **SUV** Truck

Hypotheses

1
Reckless drivers are more likely to drive coupes or convertibles

2
SUVs and minivans are less likely to be involved in serious car accidents

3
More severe car accidents occur past midnight (12:00 am - 3:59 am)

Hypothesis 3 - Modelling Approach

Variables	Type	Description	Class
HOUR	DV	The time range of the accident occured	integer
MXVSEV_IM	IV	The severity of injury	integer
ROLLOVER	IV	Whether or not the vehicle rolled over	integer
TOWED	IV	Whether or not the vehicle was towed	integer
DEFORMED	IV	The condition of the car involved in the accident	integer
AIR_BAG	IV	Whether or not the airbag deployed	integer
EJECT_IM	IV	Whether or not the driver was ejected from the vehicle	integer

Modelling Approaches

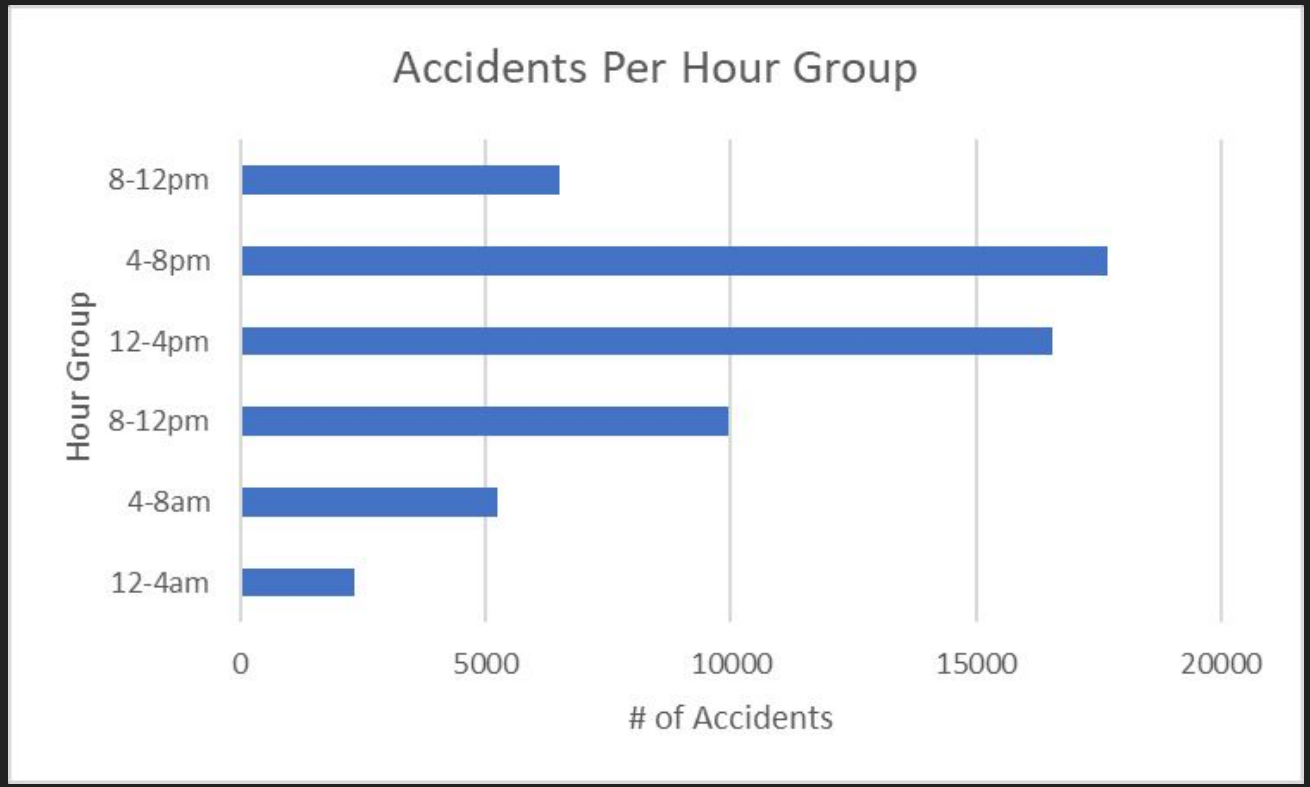
1. Ordered Logistic Regression Model

2. Multinomial Logistic Regression Model

Reason: HOUR is a categorical response variable with more than two categories

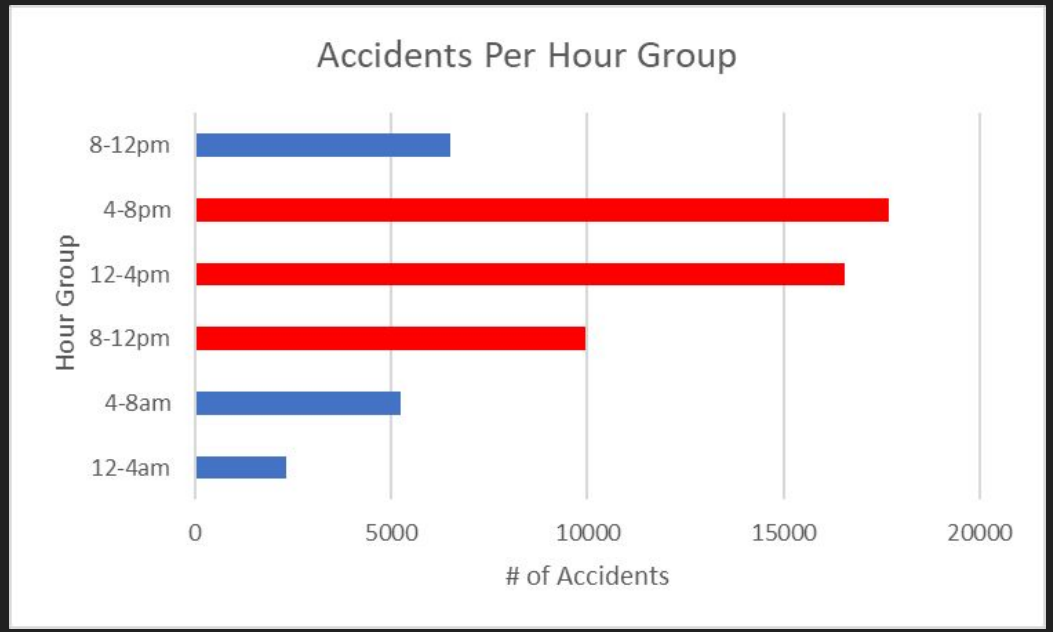
Hypothesis 3 - Downsampling

Before



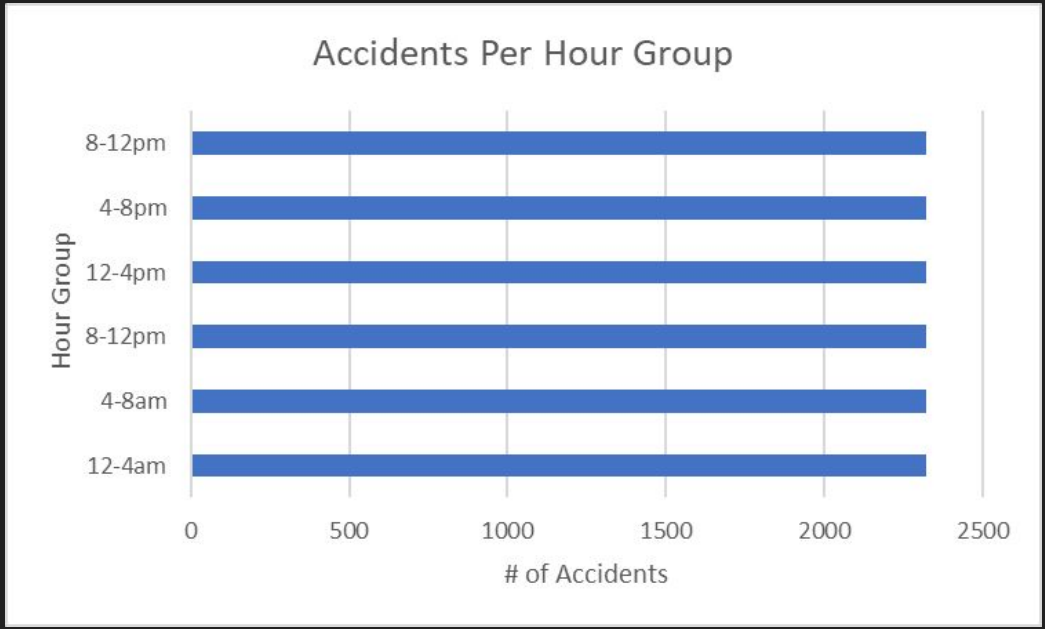
Hypothesis 3 - Downsampling

Before



Hypothesis 3 - Downsampling

After



Hypothesis 3 - Modelling Analysis

Ordered Logistic Regression Model

```
Coefficients:
                   Value Std. Error t value
Severity_Score1 -0.02421    0.09068  -0.267
Severity_Score2 -0.42542    0.11049  -3.850
Severity_Score3 -0.29803    0.08821  -3.379
Severity_Score4 -0.64074    0.15647  -4.095

Intercepts:
    Value    Std. Error t value
1|2  -1.8427   0.0882   -20.8854
2|3  -0.9192   0.0869   -10.5810
3|4  -0.2214   0.0865    -2.5593
4|5   0.4740   0.0866     5.4704
5|6   1.3899   0.0878    15.8381

Residual Deviance: 49834.78
AIC: 49852.78
```

Multinomial Logistic Regression Model

```
Coefficients:
  (Intercept) Severity_Score1 Severity_Score2 Severity_Score3 Severity_Score4
2   1.0200536      -0.4902999      -0.9356488      -1.161990      -1.858341
3   1.3158935      -0.5358340      -1.4806961      -1.553073      -2.441939
4   1.2399030      -0.4402939      -1.5248840      -1.470687      -2.599376
5   1.2659108      -0.4676819      -1.5508712      -1.501663      -2.351000
6   0.8010263      -0.3329187      -1.0043195      -0.904138      -1.193017

Std. Errors:
  (Intercept) Severity_Score1 Severity_Score2 Severity_Score3 Severity_Score4
2   0.2487118       0.2583528       0.2778682       0.2511365       0.3265294
3   0.2401236       0.2492926       0.2740318       0.2427678       0.3359212
4   0.2421202       0.2511619       0.2779585       0.2447335       0.3532162
5   0.2414216       0.2504924       0.2773498       0.2440496       0.3343590
6   0.2566549       0.2662245       0.2892680       0.2589575       0.3152546

Residual Deviance: 49504.37
AIC: 49554.37
```

llh	llhNull	G2	McFadden	r2ML	r2CU
-2.491739e+04	-2.496279e+04	9.080559e+01	1.818819e-03	6.496577e-03	6.682193e-03

<

llh	llhNull	G2	McFadden	r2ML	r2CU
-2.475219e+04	-2.496279e+04	4.212111e+02	8.436779e-03	2.978090e-02	3.063178e-02

Hypothesis 3 - Modelling Analysis

Modelling Approach

~~Ordered Logistic Regression Model~~

⬇

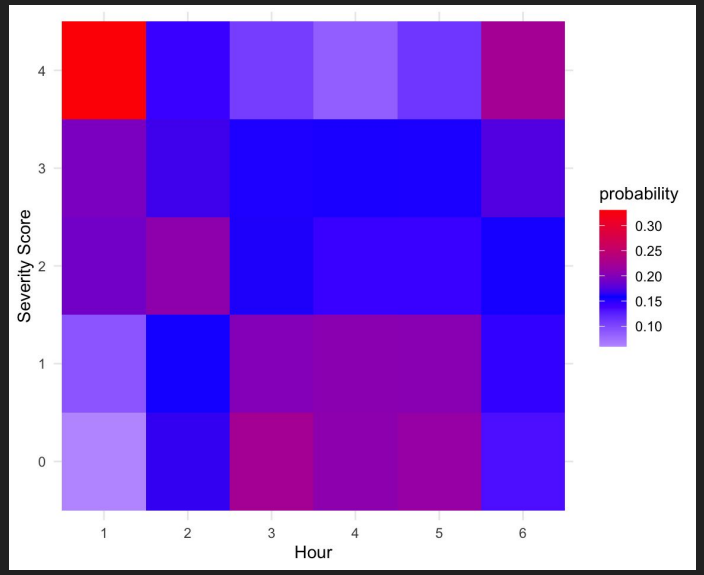
Multinomial Logistic Regression Model

Hypothesis 3 - Modelling Analysis

P-value Matrix for **Multinomial Logistic Regression Model**

	(Intercept)	Severity_Score1	Severity_Score2	Severity_Score3	Severity_Score4
2	4.107513e-05	0.05772344	7.592478e-04	3.711356e-06	1.261574e-08
3	4.251651e-08	0.03160127	6.539928e-08	1.580867e-10	3.610445e-13
4	3.038827e-07	0.07959722	4.111071e-08	1.862798e-09	1.849632e-13
5	1.574998e-07	0.06189453	2.247906e-08	7.597909e-10	2.045253e-12
6	1.802226e-03	0.21111016	5.167233e-04	4.804012e-04	1.541432e-04

Hypothesis 3 - Modelling Analysis



Severity_Score	1	2	3	4	5	6
3	0.19446022	0.1687286	0.1533999	0.15438415	0.1536189	0.1754082
1	0.09154693	0.1554937	0.1997190	0.20366182	0.2033809	0.1461976
2	0.19025000	0.2070052	0.1613436	0.14307343	0.1430764	0.1552514
4	0.33034464	0.1428577	0.1071351	0.08483121	0.1116141	0.2232172
0	0.05976983	0.1657623	0.2228267	0.20652139	0.2119630	0.1331568

Interpretation

- Regarding Severity Level 4, there is highest probabilities in HOUR category 1 (12:00am-3:59am)

12:00am-3:59am is the most dangerous hour

Key Takeaway

1

Reckless drivers are more likely to drive coupes or convertibles

2

SUVs and minivans are less likely to be involved in serious car accidents

3

More severe car accidents occur past midnight (12:00 am - 3:59 am)


Thank you
Q&A

Appendix - Summary Statistics

H1

Statistic	N	Mean	St. Dev.	Min	Max
VPICBODYCLASS	58,293	5.892	2.072	1	9
DRUGS	58,293	0.013	0.113	0	1
SPEEDREL	58,293	0.057	0.232	0	1
V_ALCH_IM	58,293	1.962	0.191	1	2
AGE_IM	58,293	40.134	17.457	8	99
URBANICITY	58,293	1.253	0.435	1	2
VPICMAKE	58,293	594.756	3,279.607	441	99,999

H2

Statistic	N	Mean	St. Dev.	Min	Max
VPICBODYCLASS	58,293	5.892	2.072	1	9
MXVSEV_IM	58,293	0.596	0.981	0	5
ROLLOVER	58,293	0.095	0.779	0	9
TOWED	58,293	3.665	1.447	2	5
DEFORMED	58,293	4.148	1.845	0	6
EJECT_IM	58,293	0.006	0.097	0	3
AIR_BAG	58,293	16.360	6.398	1	20

H3

Statistic	N	Mean	St. Dev.	Min	Max
HOUR	58,293	4.056	1.279	1	6
MXVSEV_IM	58,293	0.596	0.981	0	5
ROLLOVER	58,293	0.095	0.779	0	9
TOWED	58,293	3.665	1.447	2	5
DEFORMED	58,293	4.148	1.845	0	6
EJECT_IM	58,293	0.006	0.097	0	3
AIR_BAG	58,293	16.360	6.398	1	20

Appendix - VPICBODYCLASS

1	**Convertible**
2	**Van**
3	**Coupe**
4	**Hatchback**
5	**SUV**
6	**Truck**
7	**Sedan**
8	**Wagon**
9	**Pickup**

Appendix - H1

Original Multinom

```
Coefficients:
  (Intercept)       DRUGS    SPEEDREL   V_ALCH_IM
2  -3.2711944   0.4156899 -0.78058004  1.66687994
3   1.2858742   0.1317961 -0.14221620 -0.67242064
4  -1.0590018   0.4935900 -0.83192816  0.54506942
5  -0.1542064   0.6513828 -0.23317042  0.07038061
6 -26.6672329 -11.1929944 -1.02535533 13.38181124
7   0.1224945   0.4391444 -0.59836074 -0.06242951
8  -1.0148930   0.3211321 -0.06349786  0.51506829
9   0.6624659   0.1080505 -0.08770694 -0.34726239

Std. Errors:
  (Intercept)        DRUGS  SPEEDREL  V_ALCH_IM
2  1.59714314 9.736429e-01 0.4123049 0.80090376
3  0.82488191 7.910670e-01 0.3346977 0.41773570
4  1.09142478 8.770114e-01 0.4132766 0.54944387
5  0.94424441 7.943370e-01 0.3457750 0.47622003
6  0.01839329 3.408904e-05 0.4490299 0.03678599
7  0.93188756 8.243042e-01 0.3797287 0.47048935
8  1.04998807 8.780974e-01 0.3346570 0.52848127
9  0.86912815 8.238488e-01 0.3326560 0.43936986

Residual Deviance: 9821.129
AIC: 9885.129
```

llh	llhNull	G2	McFadden	r2ML	r2CU
-4.910565e+03	-4.941396e+03	6.166333e+01	6.239464e-03	2.704565e-02	2.738377e-02

Original POLR

```
Call:
polr(formula = VPICBODYCLASS ~ DRUGS + SPEEDREL + V_ALCH_IM,
    data = h1data_train, Hess = TRUE)

Coefficients:
            Value Std. Error  t value
DRUGS     0.01259     0.3471  0.03626
SPEEDREL  0.04121     0.1650  0.24982
V_ALCH_IM -0.05459     0.2172 -0.25135

Intercepts:
    Value   Std. Error t value
1|2 -2.1659  0.4350    -4.9794
2|3 -1.3441  0.4325    -3.1077
3|4 -0.7933  0.4319    -1.8367
4|5 -0.3291  0.4320    -0.7617
5|6  0.1137  0.4322     0.2631
6|7  0.5914  0.4325     1.3676
7|8  1.1422  0.4330     2.6378
8|9  1.9788  0.4350     4.5484

Residual Deviance: 9882.619
AIC: 9904.619
```

llh	llhNull	G2	McFadden	r2ML	r2CU
-4.941310e+03	-4.941396e+03	1.735111e-01	1.755689e-05	7.714735e-05	7.811183e-05

Appendix - H1

New Multinom

```
Call:
multinom(formula = VPICBODYCLASS ~ DRUGS + SPEEDREL + V_ALCH_IM +
    AGE_IM + SPEEDREL * URBANICITY + SEX_IM, data = h1data_train)

Coefficients:
  (Intercept)        DRUGS    SPEEDREL    V_ALCH_IM       AGE_IM URBANICITY      SEX_IM SPEEDREL:URBANICITY
2  -3.53700848   0.40475749 -1.95372265  1.698261674  0.001919356  0.3142271 -0.18971852           0.8728274
3   2.33234302   0.02600161 -0.59083277 -0.674773065 -0.029095153  0.1881404 -0.09478349           0.1819558
4  -1.87879464   0.40612127 -2.44779273  0.427860187 -0.008372396  0.4913267  0.54520238           1.1397601
5  -0.34587962   0.62961857  1.39872589 -0.004642867 -0.014277072  0.1386783  0.52037422          -1.4913746
6 -20.86044690 -10.96837883 -0.15876324 11.566936610  0.002264696  0.4526000 -2.37296163          -0.8173415
7  -0.02984737   0.37669166 -0.08336706 -0.147547738 -0.018817600  0.3037064  0.48962496          -0.4927594
8  -1.94440642   0.35661774 -0.32293474  0.433030619  0.005894621  0.1208203  0.47130707           0.2857494
9   1.29814961  -0.09184923  0.27332866 -0.135757315 -0.002064988  0.9829408 -1.80593897          -0.3643118

Std. Errors:
  (Intercept)        DRUGS SPEEDREL V_ALCH_IM      AGE_IM URBANICITY    SEX_IM SPEEDREL:URBANICITY
2  1.65287144 9.839063e-01 1.277680 0.8054780 0.005052536  0.2261144 0.1856034           0.9026108
3  0.93563473 7.987529e-01 1.034960 0.4251861 0.005591701  0.2345863 0.1879279           0.8008584
4  1.16622326 8.858760e-01 1.336379 0.5525486 0.005179493  0.2226376 0.1828333           0.9024754
5  1.03635990 8.021605e-01 1.357764 0.4819224 0.005273900  0.2336154 0.1826283           1.1972375
6  0.09735865 4.511538e-05 1.511521 0.1947538 0.005153486  0.2273268 0.2966586           1.2450480
7  1.02390189 8.309716e-01 1.221546 0.4755982 0.005352320  0.2285026 0.1832291           0.9838067
8  1.13214856 8.873650e-01 1.036081 0.5338116 0.005017945  0.2323885 0.1811168           0.7995966
9  0.98829438 8.367974e-01 1.026242 0.4486105 0.005207828  0.2196749 0.2522421           0.7791805

Residual Deviance: 9383.609
AIC: 9511.609
```

Appendix - H1

Confusion Matrix

h1predict

	1	2	3	4	5	6	7	8	9
1	2	1	23	24	16	34	3	0	12
2	0	2	8	30	2	44	3	0	11
3	0	0	22	13	12	28	6	0	19
4	1	1	9	34	11	28	17	1	8
5	1	4	21	22	9	24	12	3	10
6	2	0	2	2	2	77	0	0	28
7	1	0	16	20	14	35	8	0	9
8	2	1	15	26	9	30	5	4	11
9	1	1	14	6	1	52	2	2	35

New Variable Prediction

1	2	3	4	5	6	7	8	9
1.507406e-01	5.146161e-02	1.220381e-01	1.178462e-01	3.487605e-02	1.067030e-11	5.363651e-02	1.844270e-01	2.849739e-01

Original Variable Prediction

1	2	3	4	5	6	7	8	9
1.269041e-01	1.771172e-02	2.319368e-01	5.411767e-02	1.772955e-01	1.064948e-12	1.149250e-01	9.960785e-02	1.775014e-01

Appendix - H3

```
cleaned_data$HOUR[cleaned_data$HOUR %in% c(0,1,2,3)] <- 1
cleaned_data$HOUR[cleaned_data$HOUR %in% c(4,5,6,7)] <- 2
cleaned_data$HOUR[cleaned_data$HOUR %in% c(8,9,10,11)] <- 3
cleaned_data$HOUR[cleaned_data$HOUR %in% c(12,13,14,15)] <- 4
cleaned_data$HOUR[cleaned_data$HOUR %in% c(16,17,18,19)] <- 5
cleaned_data$HOUR[cleaned_data$HOUR %in% c(20,21,22,23)] <- 6
```