Effect of various factors on motor vehicle accident fatality rates

Submitted to:

Lynn A. Agre, MPH, PhD

Statistics 567

August 14, 2019

By: Stefan Maciolek

Abstract:

The link between various factors such as inclement weather, substance impairment, poor road conditions and motor vehicle accident rates has already been well established. However, it still remains unclear as to whether these and other factors contribute to the likelihood that these accidents will result in injury or death. Data were obtained from data.gov of all of the motor vehicle accidents that occurred in the state of Vermont in the year 2016 to serve as a sample for this study. The data set consists entirely of categorical variables, which are replaced with integer values to ease the analysis. Incidents with missing data are discarded, reducing the total number in this analysis to approximately 7400. Univariate methods are used to first analyze the structure of each variable, then bivariate methods are employed to investigate the relationships between pairs of variables. Three hypotheses were outlined and tested: (1) A driver being intoxicated makes the accident more likely to result in an injury or a fatality, (2) accidents are more likely to result in injury or fatalities when they occur late at night or in the early morning hours, (3) Accidents that occur on weekends are more likely to result in injury or death. The analysis is then extended to the multivariate level, and a variety of analysis techniques are performed with the intention of determining which factors are significant, in a mutually confirmatory way.

Introduction:

The relationship between many factors and the occurrence of motor vehicle accidents is already well known. For example, we know that very young or advanced age, driver impairment due to drugs and alcohol, adverse weather conditions, and poor road conditions all increase the probability that an accident will occur. A study conducted by Doherty et Al, found that 16-19 year old drivers were, rather unsurprisingly, involved in accidents at a much higher rate than their 20-24 year old and 25-59 year old counterparts. Furthermore, accident rates were found by the study to be higher at night, on weekends, and when passengers were present in the car (the latter only was significant for the 16-19 year old age group, however). Various public policies have been enacted in order to reduce the accident rate.

Karlaftis and Golias constructed a mathematical model to predict accident rates, and found that roadway geometric design and pavement condition were the most important variables that would predict accident rates. These results were further supported in a study by Anastasopoulos et al. that applied a tobit model to accident data. The study focused on several segments of road, and was able to make accurate analyses, despite the fact that some of the segments analyzed had no recorded accidents at all. They also found that pavement condition, along with a variety of other factors were able to predict accident rates. In fact, the pavement condition was found to be so influential to accident rate, that a later study by Sarwar and Anastasopolous was able to fit a model for pavement deterioration and found that it significantly affected both accident rates, and ensuing injury rates. The policy implications are quite large for this study.

What is less clear is whether any of these factors affect the outcome of an accident once it occurs. For example, it is pretty well understood that driving in snow will make one more likely to get into an accident, but is that accident more likely to be fatal than a fair-weather accident? Of note is a study conducted by Zhou and Sisiopiku that examined the relationship between the ratio of traffic volume and roadway volume and accident rate. It was found that a U-shaped relationship existed, that is that accidents were more likely to occur when the amount of traffic on the road was either very high or very low. However, it was found that fatal accidents decreased with increasing volume to capacity ratio, suggesting that an accident that occurs in a high-traffic environment is less likely to be fatal than one occurring in a low-traffic environment. Li et al. used a data mining technique to determine the significance of several factors on fatality rates. They conclude that environmental factors such as road or weather condition do not strongly affect the fatality rate, whereas "human factors" such as collision type and intoxication do. These results need not necessarily be contradictory to the previously mentioned ones that seem to implicate environmental factors in accident rates – perhaps these environmental factors increase the accident rate, *but not the fatality of those accidents*, whereas the "human factors" do.

Although data on the type of passenger vehicle was not available from the selected data set, previous studies have shown it to be significant in predicting the outcome of an accident. A study by Ulfasson and Mannering investigated the fatality rates of Sport Utility Vehicles (SUVs) as they started to become popular. They found that drivers of SUVs tend to have a higher injury and fatality rate than drivers of sedans and other car body types when only the single vehicle was involved. However, when two cars are involved, sedans tend to have a higher rate of injury and death than SUVs and pickup trucks. A study by Evans et al. generalized this and investigated whether vehicle size or vehicle mass was more likely to predict driver fatality. It was found that the higher the mass of the vehicle, the less likely the driver was to suffer an injury or a fatality, whereas vehicle size mattered little. Rather discouragingly, a large but light car was found to be just as unsafe as a small and light one.

Previous studies have also shown rather surprising and counterintuitive results, which underscores the need for further research to identify these relationships that may not be immediately apparent. For example, a study by Bouaoun et al. looked at the fatality rates in France of various transportation types, including passenger cars, bicycles, and motorized two-wheel vehicles. As expected, the fatality rate was between 2 and 3 times higher for bicycle riders on the basis of deaths per billion kilometers driven. However, it was found that two-wheel vehicle riders had a 20-30 times higher fatality rate than passenger car drivers, which was still much higher than the fatality rate for bicycle riders. Another unexpected relationship was uncovered by Graham et al., when analyzing accident rates in Pennsylvania. It was found that accident rates were significantly higher for counties and years where oil and gas drilling was taking place as compared to counties where it was absent.

A study by Thompson et al. explored the challenges and benefits of using smartphones to detect accidents and reduce the response time of first responders. Smartphones enable drivers to identify where adverse traffic conditions are and avoid them ahead of time, as well as allow them to contact emergency personnel much more quickly than in the past. As these technologies advance and chance, we expect that whether or not the driver has a smartphone and is in possession of it at the time of an accident may become a significant predictor of accident fatality rates in the future. The good news is that the roadways as a whole are becoming safer, but not equally for all drivers. A study by Cicchino found that between 1995-1998 and 2005-2008, fatality rates per vehicle mile travelled (VMT) decreased for all age groups, but this decrease was much more pronounced for older drivers aged 55+ than for middle-aged drivers aged 35-54.

There have been further interesting results found with multivariate methods. De Raedt and Ponjaert-Kristoffersen were able to use discriminant analysis to classify older drivers as either with or without at-fault accidents. Further use of discriminant analysis by Cooper found that older drivers were less likely to be involved in an accident than middle-aged drivers, however, they had higher non-crash incidents, as measured by convictions. A more recent study by Chen used advanced multi-variate data techniques to produce a heatmaps of accident locations and to identify accident-prone areas. Clearly, as new data manipulation and analysis techniques become available, the ability to predict and mitigate them will only increase.

Almjewail et al., used cluster analysis and other data mining techniques to determine the locations where the most accidents occurred in Riyadh. Riyadh is the capital of Saudi Arabia; it is the city with the highest rate of accidents in the country with one of the highest rates of accidents in the world. Ihm and Park used an algorithm known as Apriori to try to find an association rule for accidents in Korea. Finally, Tao et al., developed a questionnaire to determine several personality traits of Chinese drivers, and to see if these traits could predict accident rates. They found that personality traits had a direct effect on risky driving behavior, and that this predicted accident risk. This is an intriguing potential area for future study, albeit the possible implication of driving policy formed around personality screenings is rather dystopian.

Methods:

Pre-analysis:

Data were obtained from data.gov for all of the motor vehicle accidents that occurred in the year 2016 in the state of Vermont, with each motor vehicle incident being the unit of analysis. While there may be variations in the number of accidents with by geographic region (and thus, state) and year, they are assumed to be small relative to the effects being measured. The data set initially contained over 12,000 records, however, many of them had missing values for the variables of interest. Any record with a missing value or a value listed as “unknown” is removed from the analysis, which brings the final total of records down to about 7,400. Even though this was a large number of observations, no further observations are removed– so few records resulted in death that removing some of them randomly could reduce the accuracy of the analysis.

The data are purely categorical, and so in order to do a numerical analysis on them, variables are selected that represent an increasing quantity of some kind, and recoded with numeric values, in order to allow for numerical analysis methods to be used. A total of 6 variables are selected. The response variable is outcome, and it is assigned a value of 0 if only property damage occurred due to the accident, 1 if an injury occurred, and 2 if the accident was fatal. Next is the weather severity variable, coded as “wethsev”. If the weather is listed as clear, it is coded as a 0, if the weather is listed as cloudy, then it is coded as a 1, 2 for wind, 3 for rain, and 4 for freezing precipitation. Timeofday is coded as 1 if the incident occurred between 12AM and 6AM, 2 if it was between 6AM and 12PM, 3 if it was between 12PM and 6PM, and 4 if it was between 6PM and 12AM. Datew represents the day of the week the incident occurred on, with a 1 corresponding to a Sunday, all the way through 7 corresponding to a Saturday. The variable impair is set to 0 if the driver was sober at the time of the accident, 1 if they were impaired one substance, either drugs or alcohol, and 2 if they were impaired on multiple substances. Finally, the road surface condition is given a value of 0 if it was dry and clear, 1 if there was some dirt, sand, or gravel on the surface, 2 if it was wet, and 3 if there was snow or slush present.

Univariate:

For the univariate analysis, means, standard deviations, minimums, and maximums were calculated. Histograms were made for each of the variables. Each variable could only take a small number of discrete integer values, and so boxplots and qq-plots were not suited for this type of analysis. They rely on the data coming from a continuous distribution, rather than a discrete one with a small amount of possible values relative to the total number of observations. Transforming these data was also not possible, again, due to the fact that they only consist of a small amount of possible values. Transforming would have changed those values, but not the frequency of cases which took those values, and thus not affect the distribution.

While no degeneracy between predictors is expected, variance-covariance and correlation matrices are nonetheless produced for all of the variables in order to check for this. The eigenvalues and eigenvectors are also produced for the variance-covariance matrix. Furthermore, traditional scatterplots are not well suited to display bivariate relationships between pairs of variables in this data set. This is because the variables take a small number of discrete variables rather than fall on a continuum, thus any scatter plot would only show each possible level combination of the two variables being analyzed. Instead, "sliced" histograms are produced that display the breakdown of the response variable for each possible level of the predictor. This results in a much better representation of the relationships between these variables.

Bivariate:

In order to test hypotheses (1), (2), and (3), a Chi-squared statistic was selected. This statistic was chosen because it does not depend on an underlying assumption of normality, and because there are more than two groups for each variable, which a t-test would need in order to work. The frequency procedure in SAS was used to construct contingency tables for each variable used in the hypothesis vs the response. The "null" hypothesis being tested was that there was no relationship between the two variables, and a large Chi-Square value implies strong evidence against this hypothesis. It is important to note that the Chi-Square test does not provide any information about the direction of the relationship, *just that there is one.* If the Chi-square test reveals a relationship, inferences about its direction are drawn from the univariate and bivariate analyses.

Multivariate:

In order to test the significance of all the predictors in determining the value of the response, a logistic regression model is used. There are very few observations in the entire set that correspond to a fatality, so any outcomes in which a fatality occurred are recoded to be "1". That way, a 1 represents either an injury or a fatality, and a 0 represents property damage only. This binary response is also necessary for using logistic regression, which is a much better fit to these data than a standard regression model that relies on the assumption of a normal response. Stepwise and backwards model selection are both employed. While either alone would have been sufficient, both are employed to verify that the same model is selected, and thus the same predictors, if there are any, are considered significant.

Next, factor analysis is performed on the five predictors. It is expected that certain variables can be grouped together, for example weather severity and surface condition, and thus their influence on the response can be explained by a single factor. The factors are first found via the principle component method that decomposes the correlation matrix with the communities on the diagonals into eigenvalues and eigenvectors that correspond to the amount of variance explained and the factor loadings on each predictor, respectively. Next, the varimax rotation method is used to maximize the loadings of each variable to a factor. It is important to note that no a priori assumptions regarding the number of factors underlying the predictors are made, therefore the number of factors fit was set to five, the number of predictors, and factors that do not explain much variance are discarded after the fact.

Using the factors that are kept as input, cluster analysis is performed that separates the data into clusters that could then be analyzed separately. The mean outcome (a number between 0 is 1) is computed for each cluster to see if any clusters are significantly different from the others. The number of clusters is chosen such that it is large enough so that there is enough variation that one cluster can show a significantly different outcome mean than another, but low enough so that there is only one cluster with a significantly different outcome mean. Next, ANOVA is performed on the outcome as a function of cluster, to prove that the cluster means are significantly different, with a contrast being used to isolate the cluster that has a different outcome mean that the others.

Finally, means for each of the variables are produced, first for only the observations in the cluster with significantly different outcome means, and then again for the entire dataset. Then t-tests are run for each of the variables, which includes the variance computed from each group, as well as the total number in each group. The t-tests compare the mean of the cluster with higher outcome means to the total dataset mean. Because we expect that the variances might vary widely in each group, and the total number of observations is much lower in the subsetted cluster than in the total group, equality of variances is not a reasonable assumption, and so the Satterthwaite approximation is used for the t-tests. The t-tests that yielded a significant result show that their corresponding variable is significantly different between the cluster with the higher outcome mean and the total group. These are compared to the results from the logistic regression model. This procedure is distinct from the logistic regression procedures used earlier, and as such, any agreement in significant predictors between the two can be taken to be important variables in predicting accident fatality.

Results:

Univariate:

Variable	Mean	Std Dev	Sum	Minimum	Maximum	N
Outcome	0.24798	0.44567	1844	0	2	7436
wethsev	0.81549	1.34744	6064	0	4	7436
surfcond	0.66097	1.12071	4915	0	3	7436

impair	0.06428	0.24582	478	0	2	7436
datew	4.15923	1.89202	30928	1	7	7436
timeofday	2.62803	0.81587	19542	1	4	7436

Table 1: Univariate statistics of each variable

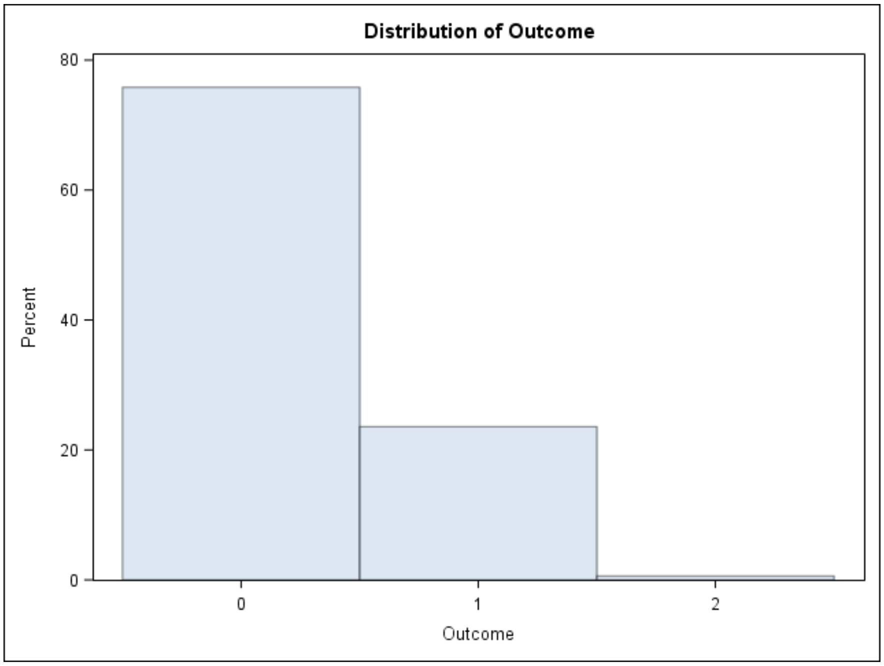

Figure 2: Histogram of outcome

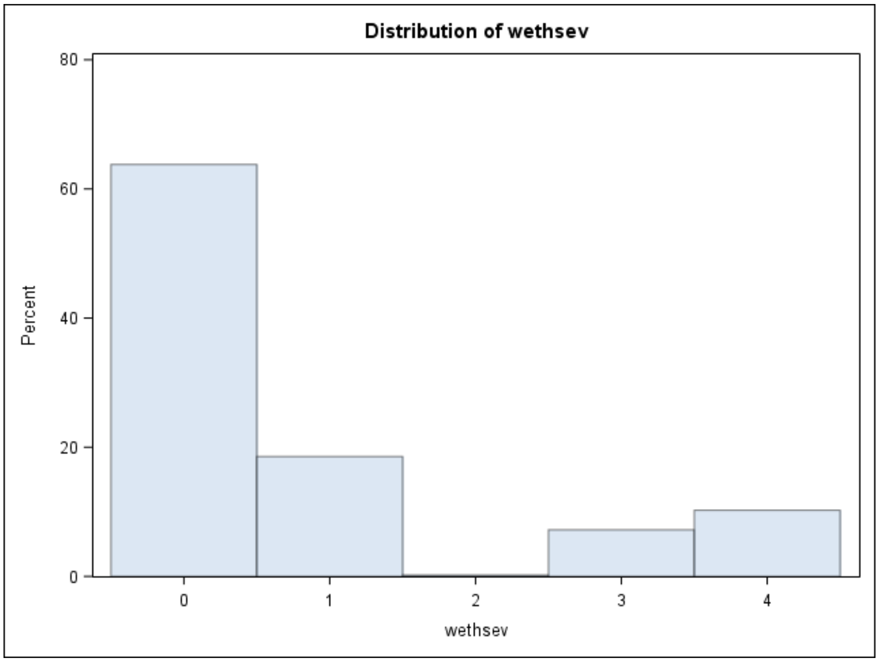

Figure 3: Histogram of weather severity

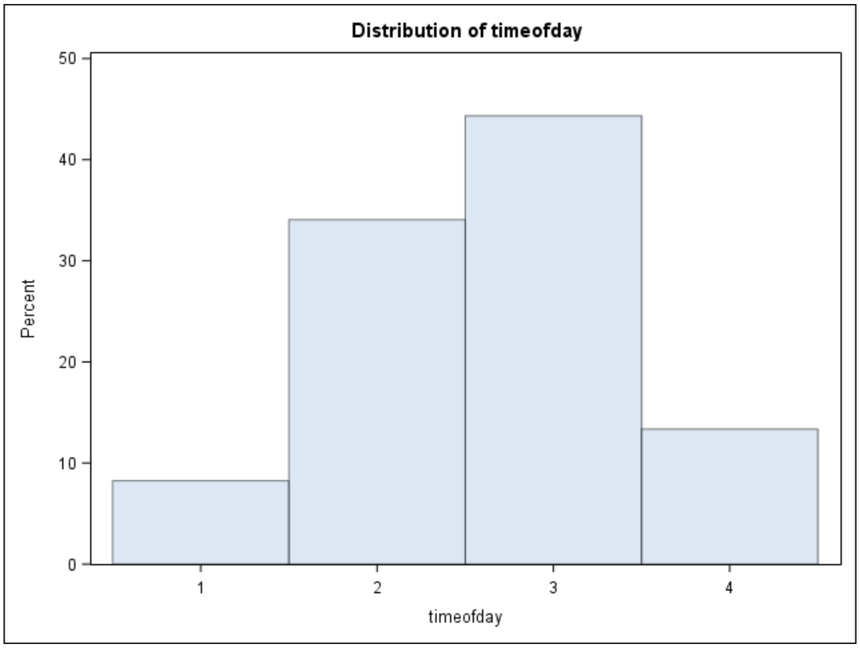

Figure 4: Histogram of time of day

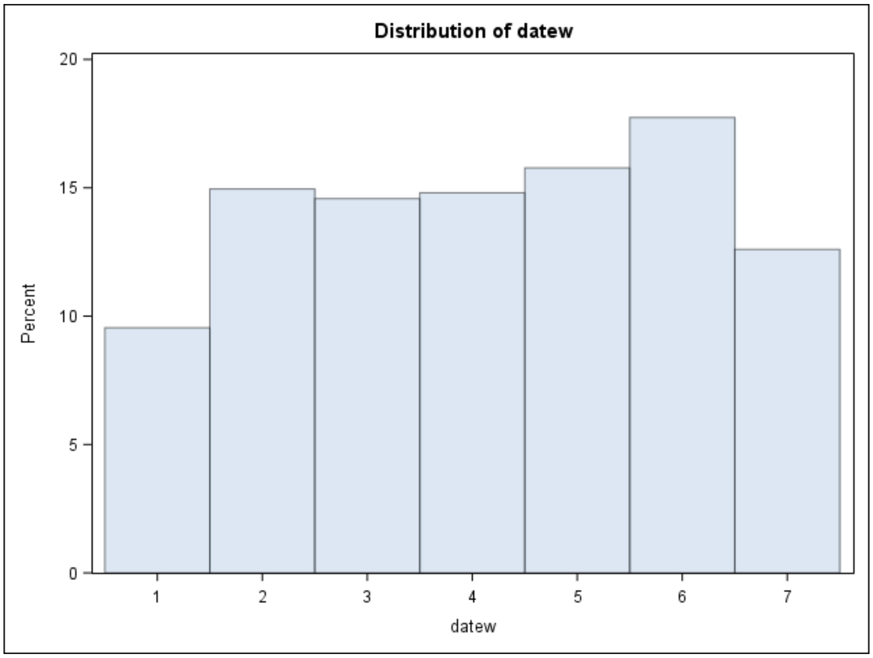

Figure 5: Histogram of day of the week

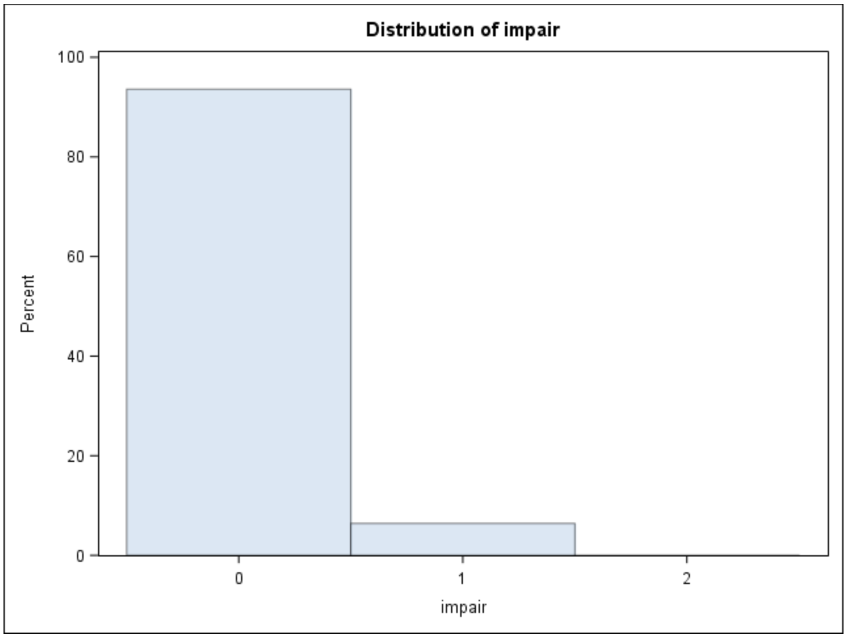

Figure 6: Histogram of Impairment

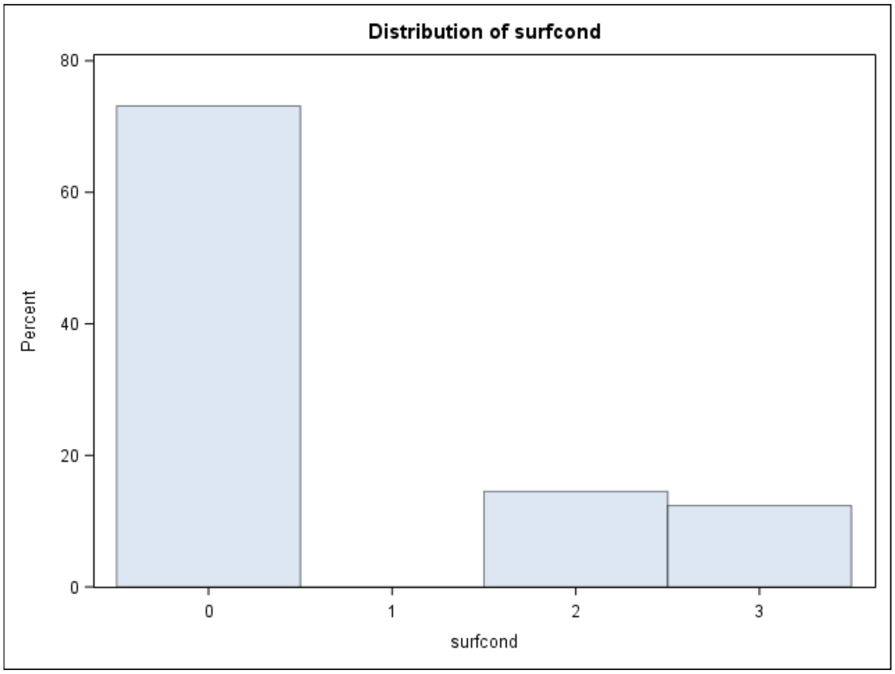

Figure 7: Histogram of road surface condition

Bivariate:

	Outcome	wethsev	surfcond	impair	datew	timeofday
Outcome	0.198622	0.000168	-0.00765	0.020506	-0.00573	0.003352

wethsev	0.000168	1.815595	1.199978	-0.00401	-0.08749	-0.07887
surfcond	-0.00765	1.199978	1.255991	-0.00403	-0.05469	-0.08107
impair	0.020506	-0.00401	-0.00403	0.060427	0.006037	0.009254
datew	-0.00573	-0.08749	-0.05469	0.006037	3.57974	0.024667
timeofday	0.003352	-0.07887	-0.08107	0.009254	0.024667	0.665644

Table 8: Variance-Covariance matrix of the variables

	Outcome	**wethsev**	**surfcond**	**impair**	**datew**	**timeofday**
Outcome	1	0.00028	-0.01531	0.18718	-0.0068	0.00922
wethsev	0.00028	1	0.79464	-0.0121	-0.03432	-0.07174
surfcond	-0.01531	0.79464	1	-0.01462	-0.02579	-0.08866
impair	0.18718	-0.0121	-0.01462	1	0.01298	0.04614
datew	-0.0068	-0.03432	-0.02579	0.01298	1	0.01598
timeofday	0.00922	-0.07174	-0.08866	0.04614	0.01598	1

Table 9: Correlation matrix of the variables

Eigenvalues	***Eigenvectors:***					
3.593068	*-0.00149*	*-0.00215*	*0.008*	*-0.05776*	*0.98795*	*-0.14337*
2.760821	*-0.1001*	*0.77547*	*0.06678*	*-0.61882*	*-0.0349*	*0.00192*
0.660312	*0.00192*	*-0.00198*	*0.01515*	*-0.00616*	*0.14313*	*0.98957*
0.303333	*0.99205*	*0.12558*	*-0.00643*	*-0.00387*	*0.00133*	*-0.00179*
0.201185	*0.01314*	*-0.05158*	*0.9976*	*0.0413*	*-0.0078*	*-0.01401*
0.057308	*-0.07506*	*0.61661*	*0.000847*	*0.78228*	*0.04688*	*-0.00054*

Table 10: Eigenvalue and Eigenvectors of the variance-covariance matrix

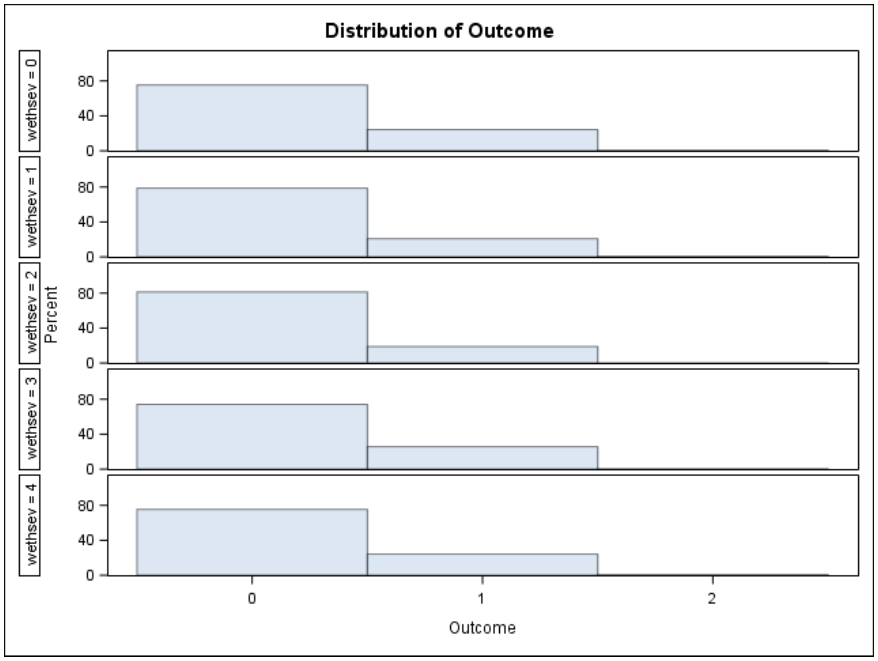

Figure 11: Sliced histogram of outcome by levels of weather severity

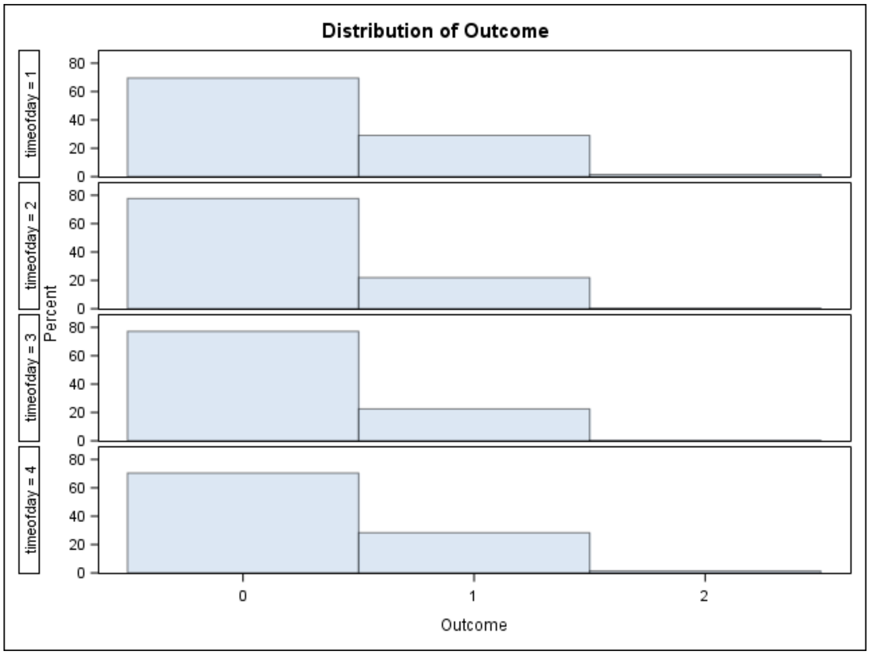

Figure 12: Sliced histogram of outcome by levels of time of day

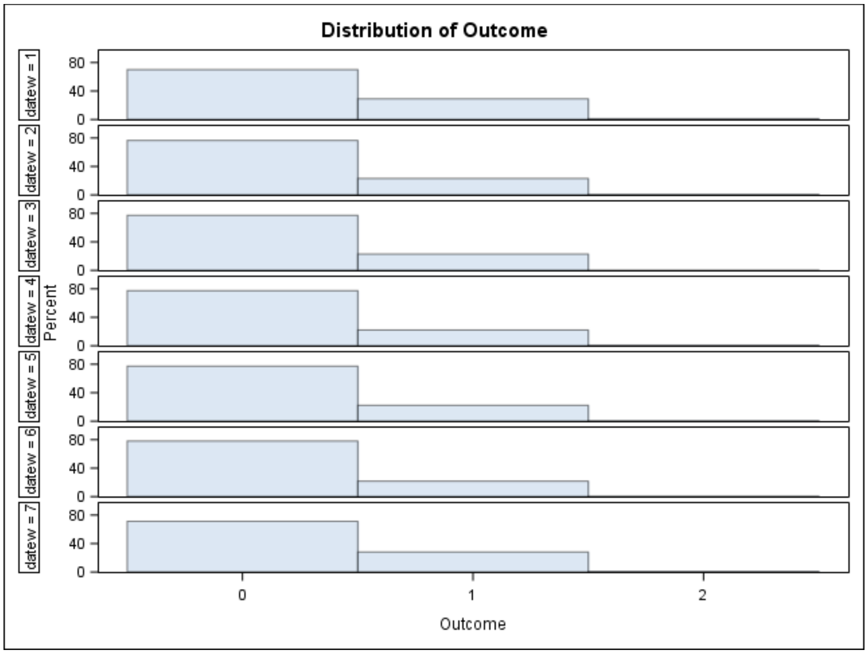

Figure 13: Sliced Histogram of outcome by levels of day of the week

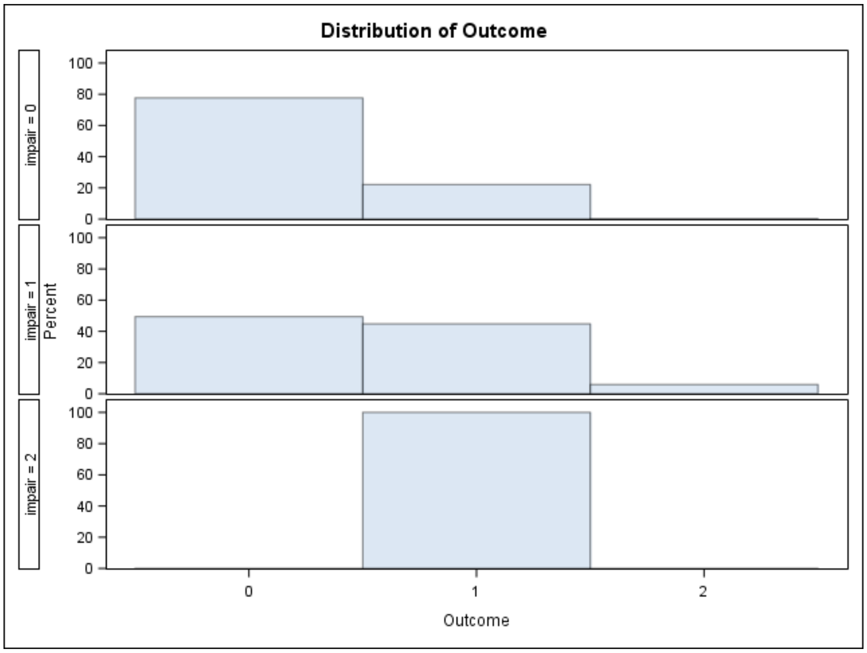

Figure 14: Sliced histogram of outcome by levels of impairment

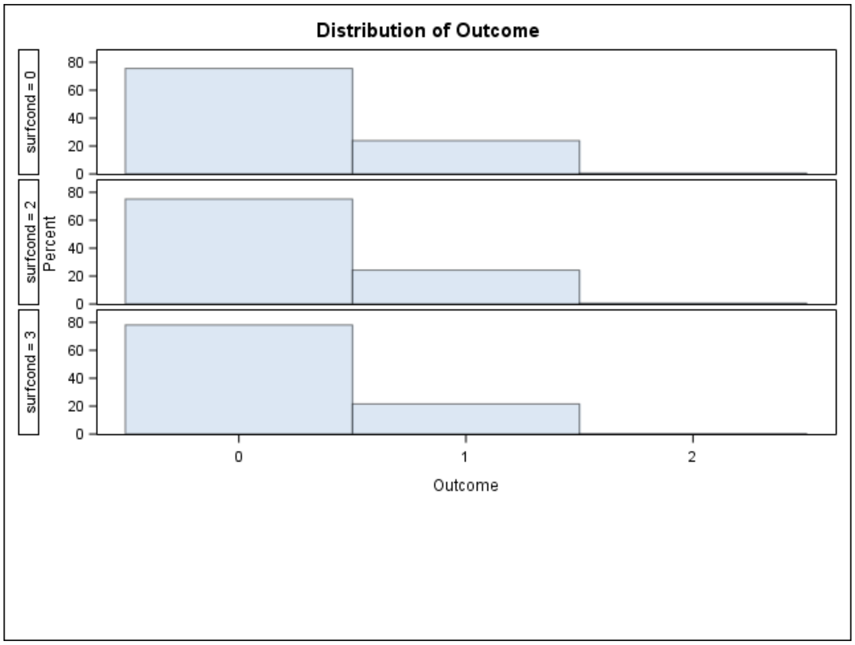

Figure 15: Sliced histogram of outcome by levels of road surface condition

```
Frequency|
Expected |       0|       1|       2|  Total
---------+--------+--------+--------+
       0 |   5402 |   1540 |     17 |   6959
         | 5275.4 | 1641.5 | 42.113 |
---------+--------+--------+--------+
       1 |    235 |    213 |     28 |    476
         | 360.84 | 112.28 | 2.8806 |
---------+--------+--------+--------+
       2 |      0 |      1 |      0 |      1
         | 0.7581 | 0.2359 | 0.0061 |
---------+--------+--------+--------+
Total        5637     1754       45     7436
            75.81    23.59     0.61   100.00
```

Table 16: Hypothesis 1: Contingency table showing observed and expected counts, with levels of impairment making up the rows and levels of outcome making up the columns. Chi-Square = 380.8155, p < .0001

```
Frequency|
Expected |       0|       1|       2|  Total
---------+--------+--------+--------+
       1 |    426 |    178 |      9 |    613
         |  464.7 | 144.59 | 3.7097 |
---------+--------+--------+--------+
       2 |   1968 |    554 |     11 |   2533
         | 1920.2 | 597.48 | 15.329 |
---------+--------+--------+--------+
       3 │    2544 │     741 │      12 │   3297
         │  2499.4 │  777.69 │  19.952 │
─────────┼─────────┼─────────┼─────────┤
       4 │     699 │     281 │      13 │    993
         │  752.76 │  234.23 │  6.0093 │
─────────┼─────────┼─────────┼─────────┤
Total         5637      1754        45     7436
             75.81     23.59      0.61   100.00
```

Table 17: Hypothesis 2: Contingency Table showing observed and expected counts, with times of day making up the rows and the levels of outcome making up the columns. Chi-Square = 51.0723, p < .0001

```
Frequency│
Expected │
Percent  │       0│       1│       2│  Total
─────────┼─────────┼─────────┼─────────┤
       1 │     497 │     205 │       8 │    710
         │  538.23 │  167.47 │  4.2967 │
─────────┼─────────┼─────────┼─────────┤
       2 │     850 │     255 │       7 │   1112
         │  842.97 │   262.3 │  6.7294 │
─────────┼─────────┼─────────┼─────────┤
       3 │     837 │     246 │       1 │   1084
         │  821.75 │  255.69 │    6.56 │
─────────┼─────────┼─────────┼─────────┤
       4 │     851 │     242 │       8 │   1101
         │  834.63 │   259.7 │  6.6629 │
─────────┼─────────┼─────────┼─────────┤
       5 │     906 │     260 │       7 │   1173
         │  889.21 │  276.69 │  7.0986 │
─────────┼─────────┼─────────┼─────────┤
       6 │    1029 │     284 │       6 │   1319
         │  999.89 │  311.13 │  7.9821 │
─────────┼─────────┼─────────┼─────────┤
       7 │     667 │     262 │       8 │    937
         │  710.31 │  221.02 │  5.6704 │
─────────┼─────────┼─────────┼─────────┤
Total         5637      1754        45     7436
             75.81     23.59      0.61   100.00
```

Table 18: Hypothesis 3: Contingency Table showing observed and expected counts, with days of the week making up the rows and levels of outcome making up the columns. Chi-square = 38.4154, p = .0001

Multivariate:

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	p - value
Intercept	1	1.2442	0.0288	1871.0445	<.0001
impair	1	-1.273	0.0958	176.5151	<.0001

Table 19: Parameter estimates and standard error for logistic regression model

Odds Ratio Estimates			
Effect	**Point Estimate**	**Lower Odds Ratio CI**	**Upper Odds Ratio CI**
impair	0.28	0.232	0.338

Table 20: Odds Ratio point estimate and Confidence Interval

Rotated Factor Pattern					
	Factor1	**Factor2**	**Factor3**	**Factor4**	**Factor5**
wethsev	0.95349	-0.02942	-0.01793	-0.00485	-0.2994
timeofday	-0.04493	0.99869	0.00756	0.02295	-0.00245
datew	-0.01678	0.00753	0.99981	0.00635	0.00048
impair	-0.0069	0.02287	0.00635	0.99969	-0.00035
surfcond	0.93887	-0.04547	-0.00982	-0.00692	0.34105

Table 21: Factor loadings for each predictor, after rotation

Variance Explained by Each Factor				
Factor1	**Factor2**	**Factor3**	**Factor4**	**Factor5**
1.7929725	1.0009046	1.0001364	1.0000271	0.2059594

Table 22: Variance explained by each factor

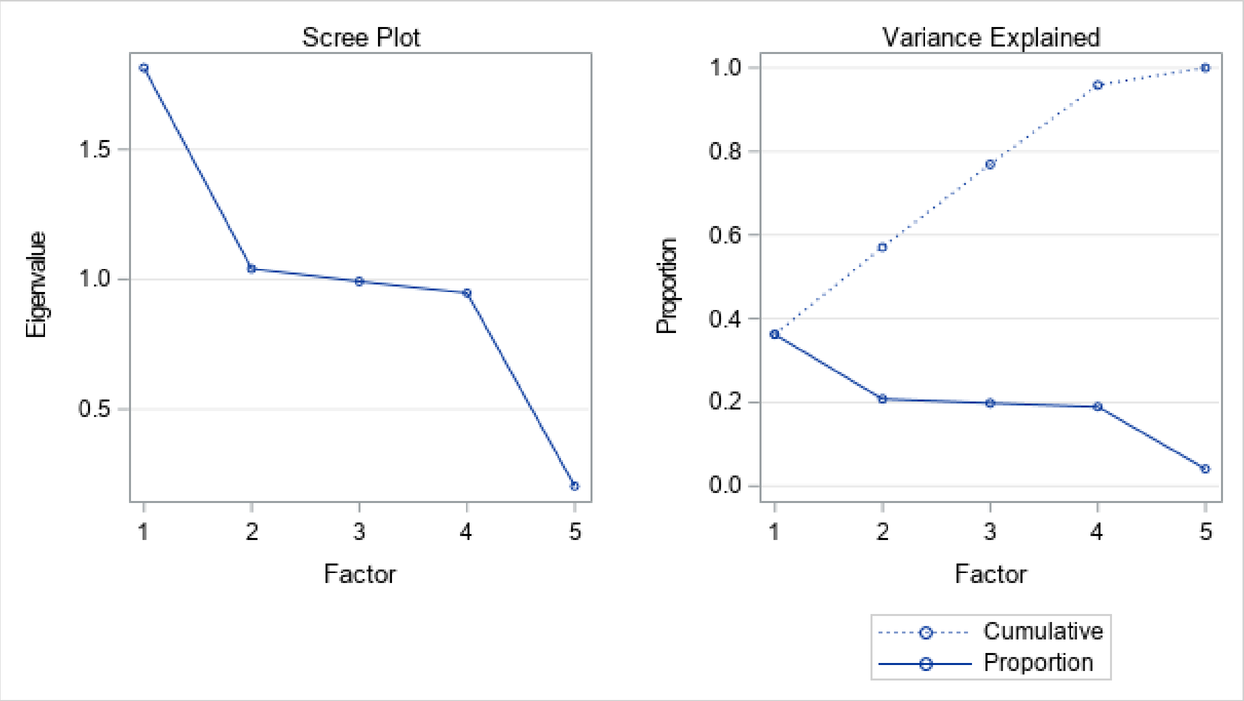

Figure 23: Scree plot of the rotated factors

Cluster	N	Outcome Mean	Outcome Std Dev
1	1708	0.2265808	0.4187417
2	1501	0.2178548	0.4129257
3	2008	0.2241036	0.4170945
4	1742	0.2256028	0.4180987
5	477	0.5073375	0.500471

Table 24: N and outcome mean and standard deviation for each of the clusters

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	35.975322	8.993831	50.33	<.0001
Error	7431	1327.790546	0.178683		
Corrected Total	7435	1363.765869			
Contrast	1	35.93015854	35.93015854	201.08	<.0001

Table 25: General ANOVA and contrast output for the difference in outcome means across the clusters

	total mean	subset mean	95% Confidence interval	t-value	p-value
Outcome:	**0.2419**	**0.5073**	**(0.2193,0.3115)**	**11.32**	**<.0001**

wethsev:	0.8155	0.7547	(-0.1802,0.0587)	-1	0.3181
timeofday:	**2.628**	**2.7757**	**(0.0366,0.2587)**	**2.61**	**0.0093**
datew:	4.1592	4.2474	(-0.1115,0.2878)	0.87	0.3862
impair:	**0.0643**	**1.0021**	**(0.9309,0.9448)**	**265.03**	**<.0001**
surfcond:	0.661	0.5996	(-0.1538,0.355)	-1.24	0.214

*Table 26: T tests for all variables for total sample vs cluster 5 only (denoted subset). Variables with a significant group mean difference are shown in **bold.***

Discussion:

Univariate Analysis:

As can be seen from Figure 2, the most common outcome (thankfully) is property damage only. Injuries make up a minority of the outcomes, and fatalities occurred for only a tiny fraction of the incidents recorded. It is, however, important to note that these data did not make note of the severity of the injuries or property damage sustained. Therefore, an "easy fix" and a car being totaled are recorded in the same category, as are a minor injury with a life-altering one. This is definitely a limitation of this study, and a potential future area of study would include recording these severities and including them in the analysis.

We can see from Figure 3 that the weather was recorded as "clear" for the vast majority of incidents, and "cloudy" made up the next highest category. "Windy" was barely recorded at all, but it is unclear as to whether this is a function of people avoiding driving under these conditions, or officers rarely choosing to use this option when recording an accident. Finally, we can see the percentage of accidents increases when it's raining, and increases even further for freezing precipitation. This is likely due to the loss of vehicle control associated with these conditions, but because the total ratio of drivers in clear weather to drivers in rain or snow is not known, it is unclear what the severity of this effect is. Colloquial experience suggests that there are less drivers present on the road under these conditions.

From Figure 4, we can see that most accidents occur between 6 AM and 6 PM – if the accident rate is higher late at night, it is not enough to offset the effect of more cars being on the road during the day. In Figure 5, a spike in accidents is noted on Fridays, with a decrease on the weekends. This is probably explained by more intoxicated and late night drivers on the road on Friday, and this interaction is a key area of study for future multivariate analysis of these data. The decrease on the weekends is probably explained by there being less cars on the road in general, with the impairment and late night driving effects somewhat offsetting this decrease on Saturdays but not Sundays. The density of traffic also varies with time of day, and so it would be expected that the density of traffic would affect accident survivability.

In Figure 6, we see that the vast majority of incidents occurred when the driver was sober. A small amount were intoxicated on one substance, and only a tiny amount had two or more substances in their systems at the time of the accident. Finally, Figure 7 shows us the distribution of road surface conditions. Clear roads were, again, by far the most common roads at the time of the accident. We can see that muddy or dirty roads were exceedingly uncommon, and this would make sense anecdotally, since we rarely see roads that are so dirty that they would affect the accident rate. Finally, wet and snowy roads both make up a minority of incidents, but a significant one at that.

Bivariate Analysis:

In Figure 11, we can see that injury as an outcome is slightly more common when the weather is listed either as rain or freezing precipitation. This would seem to suggest that not only do adverse weather conditions make accidents more likely, but they also make them more likely to result in injury. In Figure 10, we can see that the injury rate is slightly higher for both the 12AM to 6AM and 6PM to 12AM times. While we cannot say definitively, it is likely that this is related to reduced visibility due to darkness and impairment. The interaction of all three of these factors holds potential for future multivariate investigation.

In Figure 13, it seems that the fatality rate is slightly higher for weekends – that is both on Saturdays and Sundays. It is suspected that this is because there are less cars on the road in general on the weekends (as shown by Figure 5), and thus the accidents that occur are probably high speed accidents that are more likely to occur when traffic density is lower, as opposed to the low speed, less dangerous “fender benders” that occur in dense traffic. Figure 14 shows a striking increase in both the injury and death rate when the driver was impaired on one substance. Interestingly, the correlation between these two variables is only .046. When the driver was impaired on two or more substances, the outcome was only injury, but the number of cases where this condition was met is so low that it is essentially meaningless. No real change in the injury or fatality rate by surface condition is shown in Figure 15.

Hypotheses:

All of the Chi-Square tests yielded statistically significant results. Impairment had the largest Chi-square value, of 380.8155, and a p-value of less than .0001. The effect of impairment is clearly significant, and we can infer from Figure 14 that this effect is an increased rate of injury or death when the driver is impaired. Hypothesis 2 was that accidents that occur late at night or in the early morning hours are more likely to result in injury or death. Its Chi-square value was 51.0723, which corresponded to a p-value of less than .0001, which is also clearly significant. However, from table 17, we can see that there is no pattern between the expected and observed numbers of accidents that occurred in each time of day and what their outcome was. However, Figure 12 does show that the amount of accident that occurred during these times was higher, so we can conclude that this hypothesis is also correct. Finally, the hypothesis that accidents that occur on weekends were more likely to be fatal. Again, the chi-square test produced a significant result, with a value of 38.4154 and a p-value of under .0001. We can conclude that the day of the week does have a significant effect on the injury and fatality rate of an accident, referencing Figure 13, we do see a higher rate of injuries on weekends, so we can conclude that this hypothesis is correct, as well.

Multivariate Analysis:

The stepwise and backwards regression both resulted in the same model. From table 19, we can see that the only factor that is significant at the .05 level is impairment. The parameter estimate of -1.273 is harder to interpret intuitively than a standard regression parameter estimate, so instead we turn to table 20, and see the odds ratio estimate of .28. While this is not a perfect model due to the fact that the impairment score is not continuous or normally distributed, from this value we can determine approximately that for every increase in the impairment score of one unit, we can expect the odds ratio (the quotient of the probability of an accident being fatal and the probability against this) to increase by .28. 95% confidence intervals were also constructed for this odds ratio, and are (.232, .338), which means this effect is highly significant (an odds ratio containing 1 is not considered significant).

The factor analysis grouped weather severity and surface condition into one factor, as seen by their high loadings on Factor 1 in Table 21. It would make sense intuitively that these two predictors would be grouped into one factor. Examining the other factors, time of day loads heavily onto Factor 2, day of the week onto Factor 3, and impairment onto Factor 4. None of the predictors loaded heavily onto Factor 5. Examining the variance explained by each factor, we see that Factor 1 explains the highest amount of variance at 1.793, and Factors 2, 3, and 4 each explained about 1. Factor 5 only explained around .205 of the variance, and as such is was discarded. Figure 23 is a scree plot of the variance explained, and Factor 5 is clearly not needed, and as such it is discarded.

The data are then clustered, using Factors 1 – 4 as input. The outcome means are then computed for each of the clusters. The intended goal of the cluster analysis is to isolate cases that share similarities that are more likely to result in an accident fatality. As such, the number of clusters is modified – too few, and the clusters contain too many cases to show a large difference in mean outcome. Too many, and more than one will show this difference. The optimal number was found to be five. As can be seen from table 24, cluster 5 has a much higher outcome mean than the rest of the clusters, at almost twice as high. The cluster analysis did not take the outcome into account at all, what this means is that the cases in this cluster were determined to be similar based purely from the values of their predictor variables, and so happened to have a much higher mean outcome. Of further note is that the number of cases in this cluster is much lower than the other clusters, at only 477. This further suggests that they are separate and somehow distinct from the rest of the cases.

While group 5 clearly has some characteristics that are correlated with higher injury and fatality rates, ANOVA is run on the cluster outcome means to prove that they have a statistically significant difference. First, general ANOVA is run that has a p-value below .0001. This proves that the means are different, but does not prove anything beyond that. Next, a contrast is applied to the cluster means with the coefficient vector being (1,1,1,1,4). This is again significant, with a p-value below .0001. This analysis proves that the outcome mean for cluster 5 is significantly higher than clusters 1-4. See Table 25 for the general ANOVA and contrast output.

With cluster 5 having a much higher, statistically significant outcome mean, it was compared against the total group. T-tests are used to compare the mean of all of the variables in the total data set against those in cluster 5, and the results are shown in table 26 (cluster 5 is denoted as "subset"). The first significant difference is outcome, but this was already proven in the ANOVA. Next, time of day is significant, with a p-value of .0093, and impairment is significant, with a p-value of below .0001. Both the logistic regression and the t-tests agree that impaired driving is a significant variable in predicting the injury and fatality probability of a motor vehicle accident. They disagree about time of day – the logistic regression does not result in a significant result for it, but the t-test does. We note, however, that the t-test p-value for it is still much higher than for impaired driving.

Conclusions and Broader Implications:

From these data, we can see that first and foremost, driver impairment makes both injury and death much more likely to occur in an accident. Currently, there is a huge societal push against impaired driving, and these data only further support it. When driving sober, the chance of injury or death when an accident occurs is small, but when the driver is intoxicated, this chance increases to over 50%. Hypothesis 1 proved that impaired driving is more likely to result in injury or death when considered against outcome alone. Further multivariate tests that took every variable into account (and their interactions in the case of the logistic regression) further proved this correlation. There was not a single statistical test performed in this study that yielded a negative result in examining impairment's effect on motor vehicle injury and fatality rates; as a result, **this finding is probably the single most important takeaway from this entire study**.

We can also see that injuries are more likely to occur when an accident happens at night – the low correlation between this and impairment seem to suggest that this is a separate phenomenon from impairment, even though colloquially it is known that majority of impaired drivers drive at night. This is shown by the results of hypothesis 2. These findings suggest that there may be some truth to the old adage “nothing good happens after midnight”. This is further supported by the results of the t-test, that found a significant effect between the effects of time of day in cluster 5, or the “high injury/fatality group”, and the total group. However, the logistic regression failed to find this factor significant using both stepwise and backwards variable selection techniques, therefore, while there is statistical significance, the “real-world” implications of these findings seem to be somewhat weaker than with driver impairment.

Hypothesis 3 showed that there was a significant effect on outcome by day of the week. However, this finding was not replicated by either of the multivariate tests, and so the evidence for its “real-world” importance is weaker still. No significant interaction was found between time of day and day of the week either, which is rather surprising. From a policy perspective, a recommendations for drivers to stay off of the road during the late night hours could be beneficial to public health, however the association is weaker and does not warrant an outright ban like with impairment. A recommendation for drivers to stay off to the road during the weekend hours also could have some effect, but the evidence for the importance of this is tenuous at best.

References:

Jove Graham, Jennifer Irving, Xiaoqin Tang, Stephen Sellers, Joshua Crisp, Daniel Horwitz, Lucija Muehlenbachs, Alan Krupnick, & David Carey (2015). Increased traffic accident rates associated with shale gas drilling in Pennsylvania. *Accident Analysis and Prevention,* 74, 203-209.

Matthew G. Karlaftis, Ioannis Golias (2002). Effects of road geometry and traffic volumes on rural roadway accident rates. *Accident Analysis and Prevention,* 34, 357-365.

Gudmundur F. Ulfarsson, Fred L. Mannering (2004). Differences in male and female injury severities in sport-utility vehicle, minivan, pickup and passenger car accidents. *Accident Analysis and Prevention,* 36, 135-147.

Panagiotis Ch. Anastasopoulos, Andrew P. Tarko, & Fred L. Mannering (2008). Tobit analysis of vehicle accident rates on interstate highways. *Accident Analysis and Prevention,* 40, 768-775.

Liacine Bouaoun, Mohamed Mouloud Haddak, & Emmanuelle Amoros (2015). Road crash fatality rates in France: A comparison of road user types, taking account of travel practices. *Accident Analysis and Prevention*, 75, 217-225.

Jessica B. Cicchino (2015). Why have fatality rates among older drivers declined? The relative contributions of changes in survivability and crash involvement. *Accident Analysis and Prevention*, 83, 67-73.

Sean T. Doherty, Jean C. Andrey, & Carolyn MacGregor (1998). The situational risks of young drivers: the influence of passengers, time of day, and day of week on accident rates. *Accid. Anal. and Prev.,* 30(1), 45-52.

Md Tawfiq Sarwar, Panagiotis Ch. Anastasopoulos (2017). The effect of long term non-invasive pavement deterioration on accident injury-severity rates: A seemingly unrelated and multivariate equations approach. *analytic Methods in Accident Research,* 13, 1-15.

Leonard Evans, DPhil and Michael C. Frick (1992). Car Size or Car Mass: Which Has Greater Influence on Fatality Risk? *American Journal of Public Health,* 82, 1105-1112.

Chris Thompson, Jules White, Brian Dougherty, Adam Albright, and Douglas C. Schmidt (2010). Using Smartphones to Detect Car Accidents and Provide Situational Awareness to Emergency Responders. *Lecture Notes of the Institute for Computer Sciences,* 48, 29-42.

Min Zhou, Virginia P. Sisipiku (1997). Relationship between Volume-to-Capacity Ratios and Accident Rates. *Transportation research records,* 1581, 47-52.

Liling Li, Sharad Shrestha, & Gongzhu Hu (2017). *IEEE computer society.* 363-370

Almjewail A., Almjewail A., Alsenaydi S., ALSudairy H., Al-Turaiki I. (2018) Analysis of Traffic Accident in Riyadh Using Clustering Algorithms. In: Alenezi M., Qureshi B. (eds) 5th International Symposium on Data Mining Applications. Advances in Intelligent Systems and Computing, vol 753. Springer, Cham

Chen Chen (2017). Analysis and Forecast of Traffic Accident Big Data. ITM Web Conf. DOI: 10.1051/itmconf/20171204029

Ihm Sun-Young, Park Young-Ho. Brief Paper: Analysis of Traffic Accident using Association Rule Model. J Multimed Inf Syst 2018;5(2):111-114.

Da Tao, Rui Zhang, Xingda Qu (2017). The role of personality traits and driving experience in self-reported risky driving behaviors and accident risk among Chinese drivers. *Accident Analysis and Prevention*, 99, A, 228-235

Rudi De Raedt, Ingrid Ponjaert-Kristoffersen (2001). Predicting at-fault car accidents of older drivers*. Accident Analysis and Prevention*, 33, 809–819

Peter J.Cooper (1990). Differences in accident characteristics among elderly drivers and between elderly and middle-aged drivers. *Accident Analysis & Prevention* 22, 5, 499-508