UGent MASTAT program

Course: Statistical Computing

SAS Project 2020-2021

Exploratory data analysis of car accidents in Pennsylvania (USA)

Group 20: Deogracious Luyirika (DL), Dereje Mengist Belete (DB),
Echo Capwell Forbang (EF) and Tom Coussement (TC)

Contribution:

DL,DB,EF: questions 1-11 + draft report
TC: questions 1-6,12 + conclusion + reporting in ODS

1. Introduction

The United States of America is among the nations in the world with the highest rate of traffic-related fatalities per million count. Taking a case study of the state of Pennsylvania to have an in-depth knowledge of traffic accidents, a car accident dataset of the state was collected from the Kaggle website which contained a countrywide car accident dataset collected from February 2016 to June 2020. The goal of this data analysis was to understand and clean data for proper analysis and relevant inferences. In addition, to see how accidents varied across the various counties and to have an insight on the factors that influenced traffic accidents and the level of severity in the State of Pennsylvania.

2. Data Preparation and Manipulation

2.1 Data Manipulation

For the purpose of this data analyses, library named SASproj was created for referencing the data set. An appropriate contents procedure was used to explore the 23 variables name, type, and length of the data set, see Table I.b.

In addition, the weather variables in the data set were renamed, given permanent labels and stored as a temporary data set called weathernew.

We decided to restrict our analysis to the top 10 counties with the highest number of accidents reported. Therefore we subsetted the temporary data set weathernew using the data set top10 containing the names of the 10 counties and stored it to the permanent data set named top10. The top 10 counties are listed in Table I.a here below.

Table I.a: Top 10 counties with highest number of accidents

Obs	*County*	*COUNT*
1	Montgomery	34506
2	Lancaster	11689
3	York	9899
4	Allegheny	9131
5	Philadelphia	7883
6	Chester	3510
7	Delaware	2854
8	Bucks	2551
9	Dauphin	2177
10	Lehigh	1988

Variables presented in not user-friendly formats such as Start_Time were formatted by extracting all relevant parts for convenient use.

To simplify interpretation of results we created a new binary variable Severity4 based on the ordinal variable Severity with four levels.

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
9	City	Char	26	$26.	$26.
10	County	Char	20	$20.	$20.
7	Description	Char	309	$309.	$309.
4	End_Time	Num	8	DATETIME.	ANYDTDTM40.
16	Humidity___	Num	8	BEST12.	BEST32.
1	ID	Char	8	$8.	$8.
21	Precipitation_in_	Num	8	BEST12.	BEST32.
17	Pressure_in_	Num	8	BEST12.	BEST32.
2	Severity	Num	8	BEST12.	BEST32.
5	Start_Lat	Num	8	BEST12.	BEST32.
6	Start_Lng	Num	8	BEST12.	BEST32.
3	Start_Time	Num	8	DATETIME.	ANYDTDTM40.
11	State	Char	2	$2.	$2.
8	Street	Char	47	$47.	$47.
23	Sunrise_Sunset	Char	5	$5.	$5.
14	Temperature_F_	Num	8	BEST12.	BEST32.
18	Visibility_mi_	Num	8	BEST12.	BEST32.
22	Weather_Condition	Char	28	$28.	$28.
13	Weather_Timestamp	Num	8	DATETIME.	ANYDTDTM40.
15	Wind_Chill_F_	Num	8	BEST12.	BEST32.
19	Wind_Direction	Char	8	$8.	$8.
20	Wind_Speed_mph_	Num	8	BEST12.	BEST32.
12	Zipcode	Char	10	$10.	$10.

2.2 Investigation of the Missing Values of Weather Variables

A macro variable containing the weather variables was used for the analysis of missing values. Using the mi procedure on the permanent top10 data set, we found 26680 non missing data values and the missing data values for weather variables were summed across the rows.

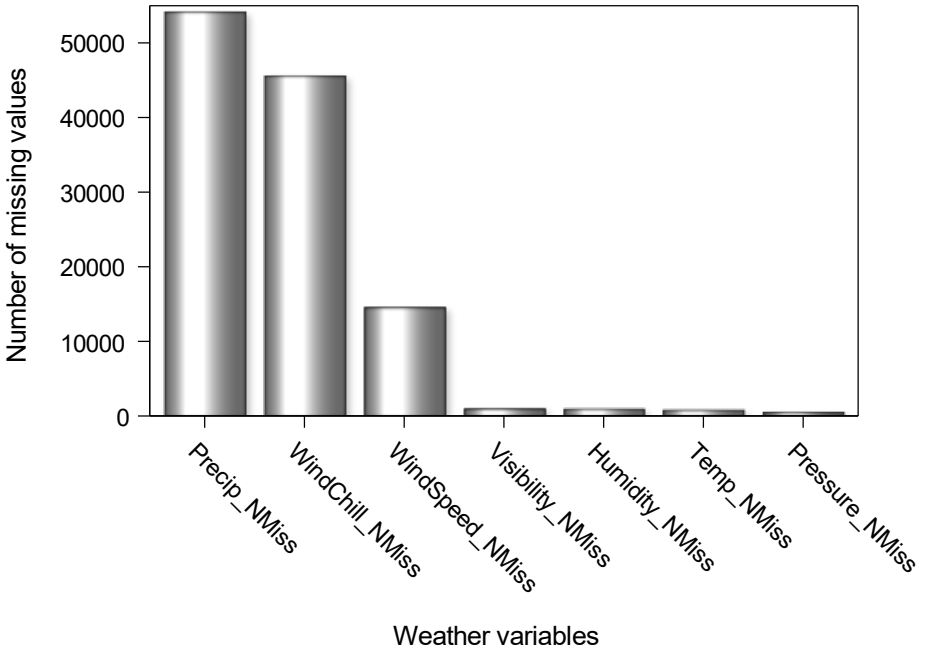
Alternatively, the means procedure, specifying nmiss as statistic of interest indicated the number of missing values for each variable. Precipitation has the highest number of missing values with 54122 and Pressure the least with 562 missing values. The table containing the number of missing values was stored in a temporary data set called cmissing of which a print is shown in Table II.

Table II: Number of missing values for all weather variables

N obs	Humidity	Temp	WindChill	Pressure	Visibility	Windspeed	Precip
86188	983	818	45536	562	1017	14588	54122

The cmissing table was then transposed, stored and the number of missing values replaced with macro variables nmiss1 to nmiss7. After transposing the cmissing data set, we plotted the missing values for the weather variables in a bar chart shown in Figure I.

Figure I: Number of missing values for all weather variables



3. Data Analysis

3.1 Descriptive Statistics for Weather Variables

Table III shows the mean and variance for various weather variables on the severity of accidents. The severity which explains the length of delay of traffic caused by accidents had four levels; level 1 which indicated the least delay and level 4, the longest delay. For instance, higher humidity has resulted on average in the longest traffic delay (severity level 4), while precipitation showed no impact on severity levels. Whereas the variance shows how the weather variables deviate from the mean across the different severity levels.

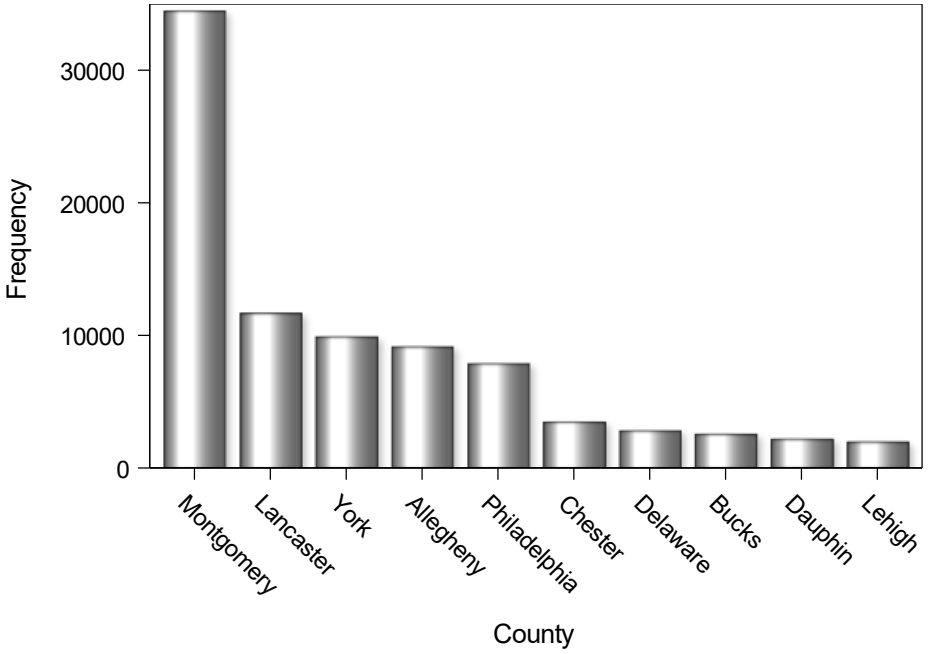
Table III: Descriptive Statistics of Numeric Variables

		Severity				
		1	*2*	*3*	*4*	*Total*
Humidity (%)	*Mean*	63.2	70.0	67.9	71.1	69.6
	Variance	445.1	415.0	444.6	417.8	422.3
Temperature (F)	*Mean*	57.9	55.0	55.8	53.0	55.1
	Variance	199.3	354.3	366.7	352.4	356.8
Wind Chill (F)	*Mean*	56.6	42.7	46.0	42.4	43.5
	Variance	249.7	429.8	509.7	506.3	455.0
Pressure (in)	*Mean*	29.6	29.9	29.8	29.8	29.9
	Variance	0.1	0.2	0.3	0.2	0.2
Visibility (mi)	*Mean*	8.7	8.7	8.6	8.6	8.6
	Variance	6.7	7.4	8.2	8.9	7.6
Wind Speed (mph)	*Mean*	8.3	7.8	8.6	8.3	8.0
	Variance	26.4	21.7	23.9	26.0	22.6
Precipitation (in)	*Mean*	0.0	0.0	0.0	0.0	0.0
	Variance	0.0	0.0	0.0	0.0	0.0

3.2 Number of Accidents per County

The top 10 counties with highest number of accidents were computed. Among these top10 counties, Montgomery registered the highest number of accidents (34506) within the years 2016-2020 while Lehigh County registered the lowest (1988) number of accidents. This was represented on a bar graph (Figure II), which further portrayed Montgomery having more than three times the number of accidents of the second county Lancaster.

Figure II: Top 10 counties with highest number of accidents



3.3 Factors Affecting the Severity of the Accidents

3.3.1 Association Between Time Factors and Severity of Accidents

Using the permanent data set top10_cmiss and based on the ordinal variable severity we created the new binary variable severity4 with levels 0 (less and moderate traffic delay) and 1 (highest traffic delay).

Less traffic delay was witnessed across the years with the highest number of accidents registered in 2017 (22127) while 2020 (11918) had the lowest number of accidents. On the other hand, high traffic delay was observed with the highest number of accidents registered in 2020 (606) while 2019 (405) had the lowest number of accidents (Table IV). While the less severe car accidents seem to decrease starting from 2017, the more severe car accidents seem to increase starting from 2019 after 4 years of steady decrease

Table IV: Frequency and severity of accidents per year

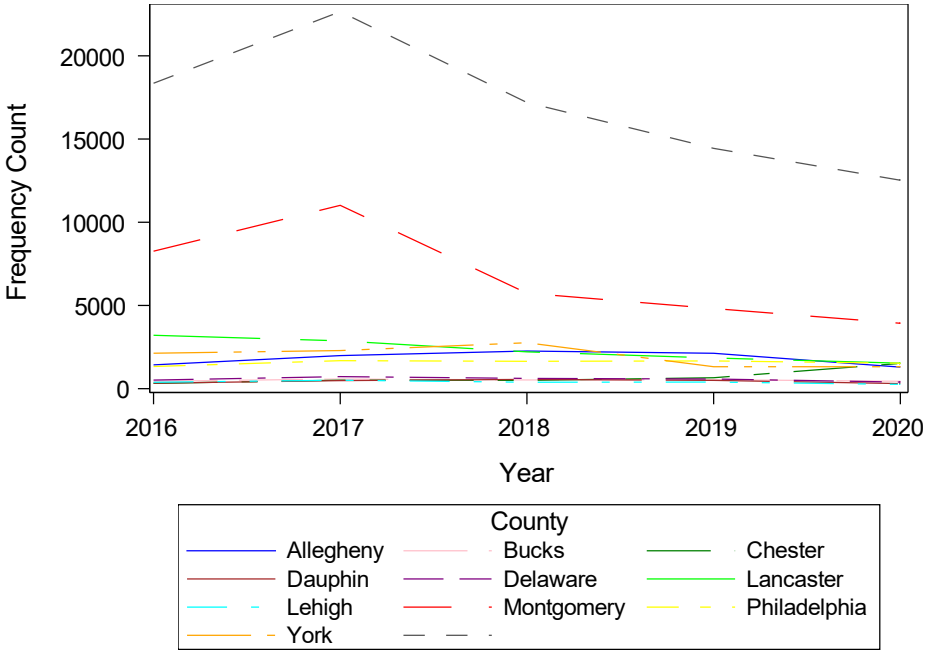
Frequency

Table of Year by severity4

	severity4		
Year	*0*	*1*	*Total*
2016	17827	525	18352
2017	22127	527	22654
2018	16693	501	17194
2019	14035	405	14440
2020	11918	606	12524
Total	82600	2564	85164

A line plot showcasing the number of accidents across the years for the various counties indicated a generally decreasing pattern in the number of accidents from 2016-2020 for most counties. An exception in this pattern is witnessed in Montgomery which displayed a sharp increase in number of accidents from 2016-2017, a sharp decease to 2018 and a steady decrease up to and including 2020 (Figure III).

Figure III: Frequency of accidents per county over the years



Overall, more accidents were recorded with less traffic delay compared to the longest delay for all the hours, days of the week, and months of the year. In relation with short delay in traffic, the highest number of accidents were observed in the month of October whereas the lowest number of accidents were recorded in the months of July. On the other hand, with longest traffic delay the highest and lowest number of accidents were recorded in May and July respectively (Figure IV).

Looking at the days of the week, more accidents occurred but with less impact on the delay of traffic. Although fewer accidents caused more delay in traffic, we realized that, towards the weekends such as Saturday's, a higher number of accidents were recorded with the most delay in traffic caused probably due to lifestyle and extravagance during weekends (Figure V).

More accidents were observed to have occurred during the morning hours (7am and 8am) of the day, causing less traffic delay. For long traffic delay, however, the hours seemed to depict the same impact. (Figure VI).

Figure IV: Number of accidents across months for both levels of severity

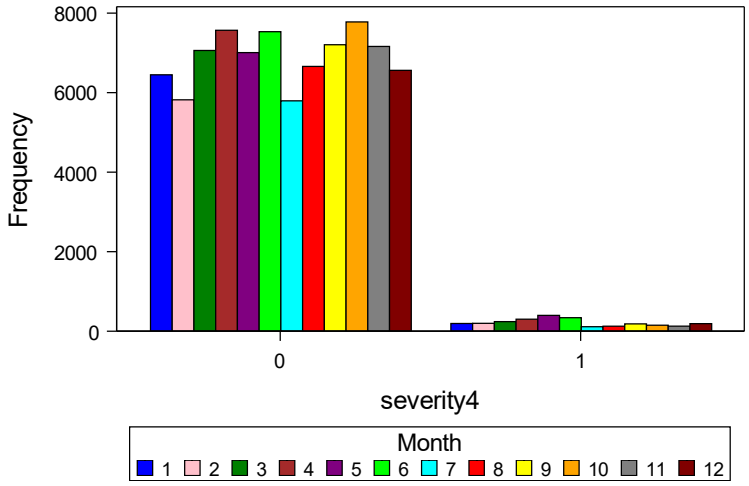


Figure V: Number of accidents across days of the week for both levels of severity

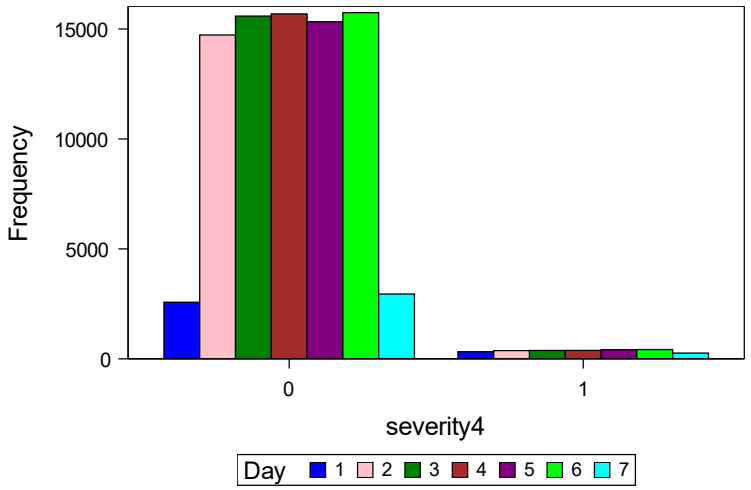
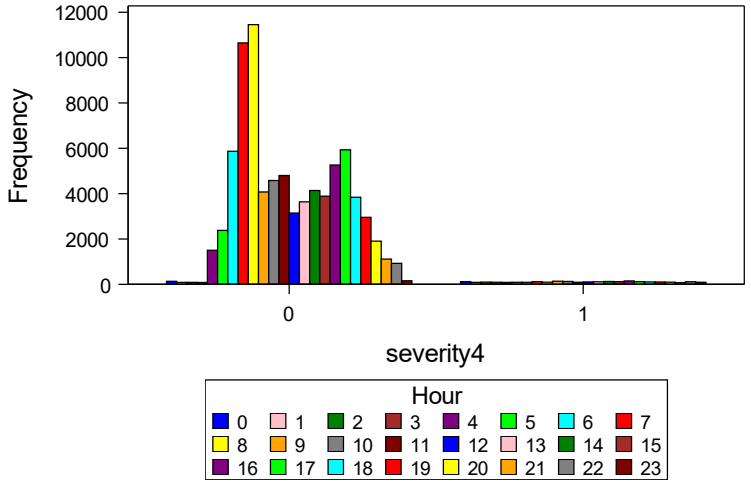


Figure VI: Number of accidents across hours of the day for both levels of severity



The odds ratio was estimated to be 2.8 with a 95% confidence interval of 2.56 to 3.04 indicating a clear association between the variables severity4 and sunrise_sunset. This means that severe accidents are 2,6 to 3,0 times more likely during nighttime than during daytime. Looking at the relative risk measures we notice that the risk for less severe accidents is the same for nighttime and daytime, but the risk for severe accidents is 0.4 times lower during daytime than during nighttime. We refer to Appendix I for detailed output of these association measures.

3.3.2 Association Between Weather Variables and Severity of the Accidents

Daily, several weather conditions could be experienced, such as clear, overcast, fair, cloudy, light rain just to name a few. Amongst these weather conditions, the most severe accidents were recorded during the clear weather condition. This is however strange but could be related to the high-speed driving of car owners during such conditions, as they turn to be more cautious during cloudy or rainy weather conditions.

Though there is no strong correlation between weather variables and severity, the result showed that some weather variables are more associated with the severity or delay of traffic resulting from accidents. Pressure was negatively correlated with high severity, meaning an increase in pressure resulted in decreasing intensity of level severity. Contrarily, windchill and windspeed were positively associated with high severity levels but with windchill having a higher association (Table V).

In order to know how a specific weather variable is distributed and to check for possible outliers, we created a macro that provides a boxplot and histogram for a given weather variable for both levels of severity4, together with a statistical test to check if there is a significant difference in mean value between the levels of severity4. We ran this macro on the three weather variables that have the biggest correlation with severity4. We refer to our code for details of the output.

Table V: Correlation of weather variables with severity4

	Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations						
	Humidity	*Temp*	*WindChill*	*Pressure*	*Visibility*	*WindSpeed*	*Precip*
severity4	0.00328 0.3386 84963	-0.00832 0.0152 85121	0.02709 <.0001 40652	-0.03363 <.0001 85155	-0.00385 0.2618 84871	0.00973 0.0093 71457	-0.00355 0.5247 32045

4. Conclusion

Our exploratory data analysis indicated that most accidents were recorded in the county of Montgomery. We restricted our analysis to the top 10 counties with highest number of accidents reported. We checked if there was a time effect on the severity of accidents and this on yearly, weekly, daily and hourly basis as well as the difference between night and day. We found significant evidence that the likelihood for severe accidents is about three times higher during nighttime compared to daytime. We noticed that over five years the number of less severe accidents decreased overall by 33% but severe accidents actually increased by 15% which is due to the big increase in the year 2020 (50% more than in 2019). Most less severe accidents are reported during the week and especially during rush hours. The same is not true for severe accidents, where no clear time trend can be noticed except that the least number of severe accidents are reported during the summer holidays

We also checked for effects of the weather on the severity of accidents. We noticed that the weather variables Precip, WindChill and WindSpeed have a lot of missing values, so one should be careful drawing conclusions because of possible selection bias. The weather variables seemed to be fairly similarly distributed over the four levels of severity. The highest number of severe accidents were reported during clear weather conditions. The weather variables Pressure, WindChill and WindSpeed showed a significant but very light correlation with severity4

APPENDIX I: Association measures for effect of sunrise_sunset on severity4

Frequency Row Pct	Table of Sunrise_Sunset by severity4		
		severity4	
Sunrise_Sunset	0	1	Total
Day	64952 97.81	1454 2.19	66406
Night	17648 94.08	1110 5.92	18758
Total	82600	2564	85164

Statistics for Table of Sunrise_Sunset by severity4

Statistic	DF	Value	Prob
Chi-Square	1	696.1192	<.0001
Likelihood Ratio Chi-Square	1	595.1974	<.0001
Continuity Adj. Chi-Square	1	694.8431	<.0001
Mantel-Haenszel Chi-Square	1	696.1110	<.0001
Phi Coefficient		0.0904	
Contingency Coefficient		0.0900	
Cramer's V		0.0904	

Odds Ratio and Relative Risks			
Statistic	Value	95% Confidence	Limits
Odds Ratio	2.8097	2.5940	3.0433
Relative Risk (Column 1)	1.0396	1.0357	1.0435
Relative Risk (Column 2)	0.3700	0.3428	0.3994

Sample Size = 85164