SAS Project Report - Group 16

Nurefsan Davulcu & Antoniya Antonova

Introduction

Traffic accidents may not be as publicly covered as airplane crashes but they certainly charge a heavy toll on the economy since they have repercussions on the productivity, medical costs, legal and court costs, emergency services costs, insurance administration costs, congestion costs, property damage, and workplace losses.

A report from the U.S. Department of Transportation from 2015 gives, to name but a few examples, the following :The $242 billion cost of motor vehicle crashes represents the equivalent of nearly $784 for each of the 308.7 million people living in the United States, and 1.6 percent of the $14.96 trillion real U.S. Gross Domestic Product for 2010.

In this report we study the data provided in the US Accidents data set which was collected over the period of 52 months starting February 2016 until June 2020. Data was collected through different channels including two APIs, over 49 states, and contains approximately 3.5 million records to this day. The focus of our report was the state of New York where we looked at the number of accidents per country, the severity of the accidents and the weather conditions during the accidents

Data Contents

The contents of the data set were explored as follows.

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat	Label
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.	Humidity(_%)
1	ID	Char	8	$8.	$8.	
21	Precip	Num	8	BEST12.	BEST32.	Precipitation(in)
17	Pressure	Num	8	BEST12.	BEST32.	Pressure(in)
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	Temp	Num	8	BEST12.	BEST32.	Temperature(F)
18	Visibility	Num	8	BEST12.	BEST32.	Visibility(mi)
22	Weather_Condition	Char	28	$28.	$28.	
13	Weather_Timestamp	Num	8	DATETIME.	ANYDTDTM40.	
15	WindChill	Num	8	BEST12.	BEST32.	Wind Chill(F)
20	WindSpeed	Num	8	BEST12.	BEST32.	Wind Speed (mph)
19	Wind_Direction	Char	8	$8.	$8.	
12	Zipcode	Char	10	$10.	$10.	

Exploration: Weather Conditions

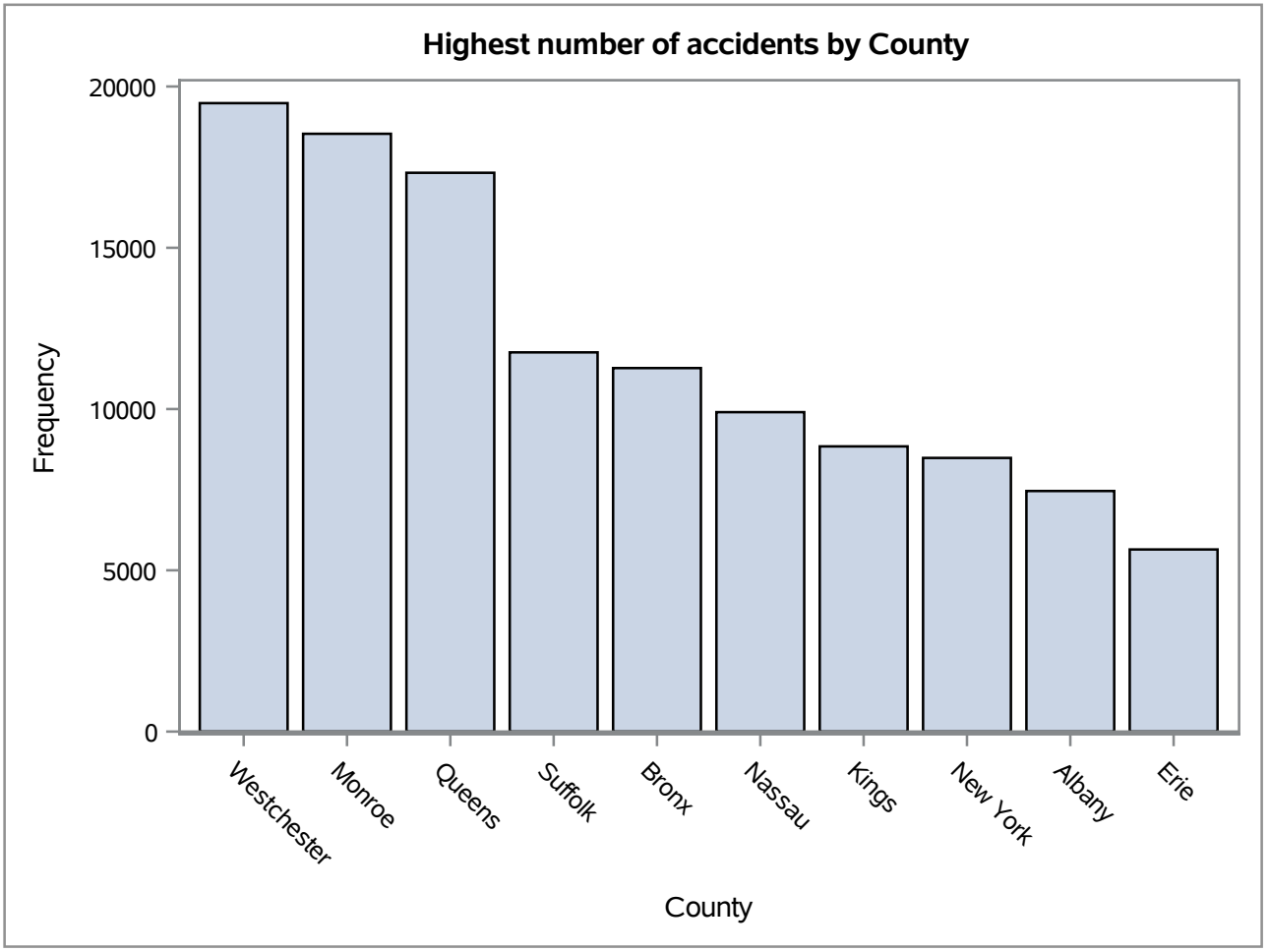
The weather variables relating to weathers conditions (humidity, temperature, windchill, pressure, visibility, windspeed, and precipitation) were explored for each level of severity. The mean level of humidity was significantly lower for the least severe accidents compared to the other levels. For the least severe accidents, the mean for humidity is lower, and the temperature and windchill are higher in comparison to the other levels. The mean does not vary much across severity for the remaining weather conditions. For humidity, temperature, windchill and precipitation there are differences in variability across levels of severity.

	Pressure(in)		Humidity(_%)		Temperature(F)		Wind Chill(F)		Visibility(mi)		Wind Speed (mph)		Precipitation(in)	
	Mean	Var	Mean	Var	Mean	Var	Mean	Var	Mean	Var	Mean	Var	Mean	Var
Severity														
1	29.55	0.12	58.85	531.70	60.63	246.60	58.86	317.13	9.63	6.80	9.36	32.85	0.01	0.00
2	29.90	0.13	68.00	395.54	53.16	357.81	42.25	486.28	8.92	8.35	9.46	28.99	0.06	0.47
3	29.95	0.11	65.25	408.07	54.67	385.91	40.87	515.71	8.98	8.52	9.78	29.95	0.09	0.77
4	29.89	0.16	66.79	410.79	53.59	369.69	42.94	524.53	9.09	7.98	9.54	30.66	0.03	0.17

Exploration: Frequency of Accidents

The top 10 counties for the number of accidents is reported and graphed below. Westchester was found to have the highest number of accidents. The frequency distribution shows how the frequency of accidents differ in the top 3 counties in comparison to the others. Westchester, Monroe, and Queens have two - three times the number of accidents of the others. The top county Westchester has three times the number of accidents as the lowest county, Erie.

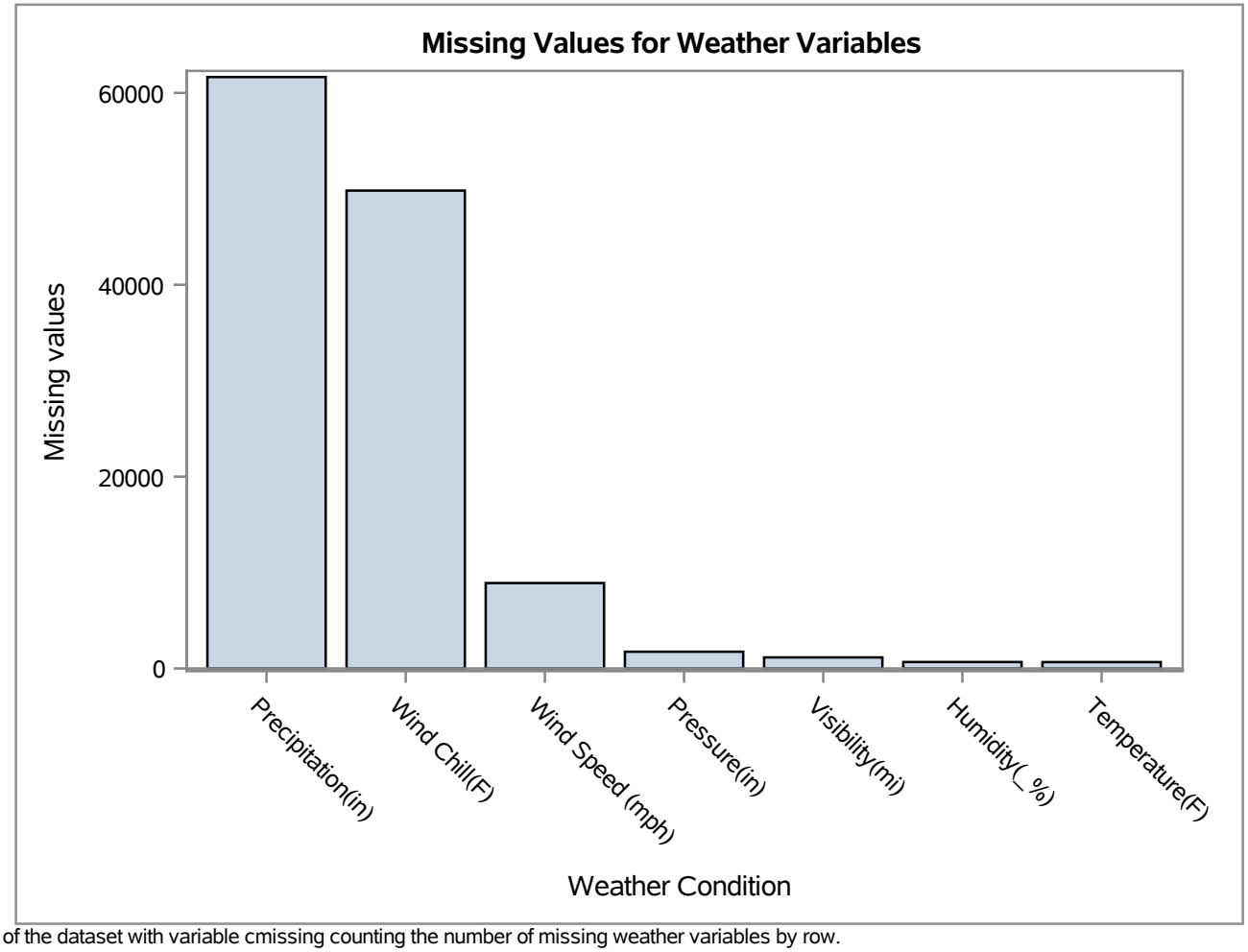
Obs	County	_FREQ_
1	Westchester	19484
2	Monroe	18534
3	Queens	17325
4	Suffolk	11757
5	Bronx	11270
6	Nassau	9903
7	Kings	8842
8	New York	8484
9	Albany	7458
10	Erie	5646



Exploration: Missing Data for Weather Conditions

Precipitation, Wind Chill and Wind speed have the highest number of missing values.

Obs	_NAME_	_LABEL_	nmiss
1	Pressure_NMiss	Pressure(in)	1747
2	Humidity_NMiss	Humidity(_%)	673
3	Temp_NMiss	Temperature(F)	666
4	WindChill_NMiss	Wind Chill(F)	49818
5	Visibility_NMiss	Visibility(mi)	1154
6	WindSpeed_NMiss	Wind Speed (mph)	8904
7	Precip_NMiss	Precipitation(in)	61653



A snippet of the dataset with variable cmissing counting the number of missing weather variables by row.

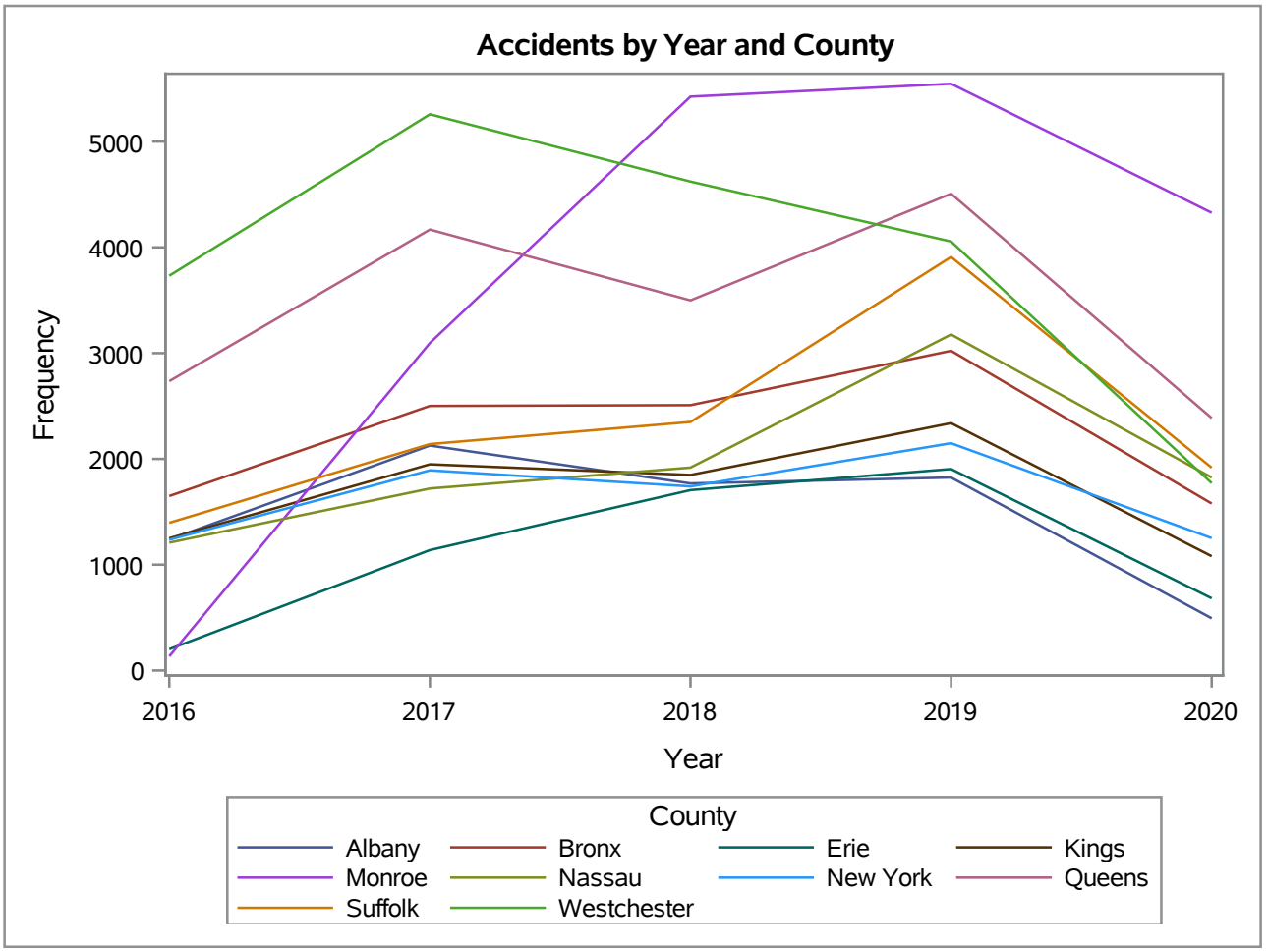
Obs	Pressure	Humidity	Temp	WindChill	Visibility	WindSpeed	Precip	cmissing
1	29.79	89	51.1	.	10	5.8	0	1
2	29.79	89	51.1	.	10	5.8	0	1
3	29.75	82	51.8	.	10	11.5	0	1
4	29.76	86	51.1	.	10	13.8	0	1
5	29.76	89	50	.	10	15	0	1
6	29.75	89	50	.	10	12.7	0	1
7	29.55	87	48.2	.	10	4.6	0.01	1
8	29.58	71	48.9	.	10	11.5	0	1
9	29.58	71	48.9	.	10	11.5	0	1
10	29.63	68	45	41.8	10	5.8	0	0

Time Factors affecting Severity

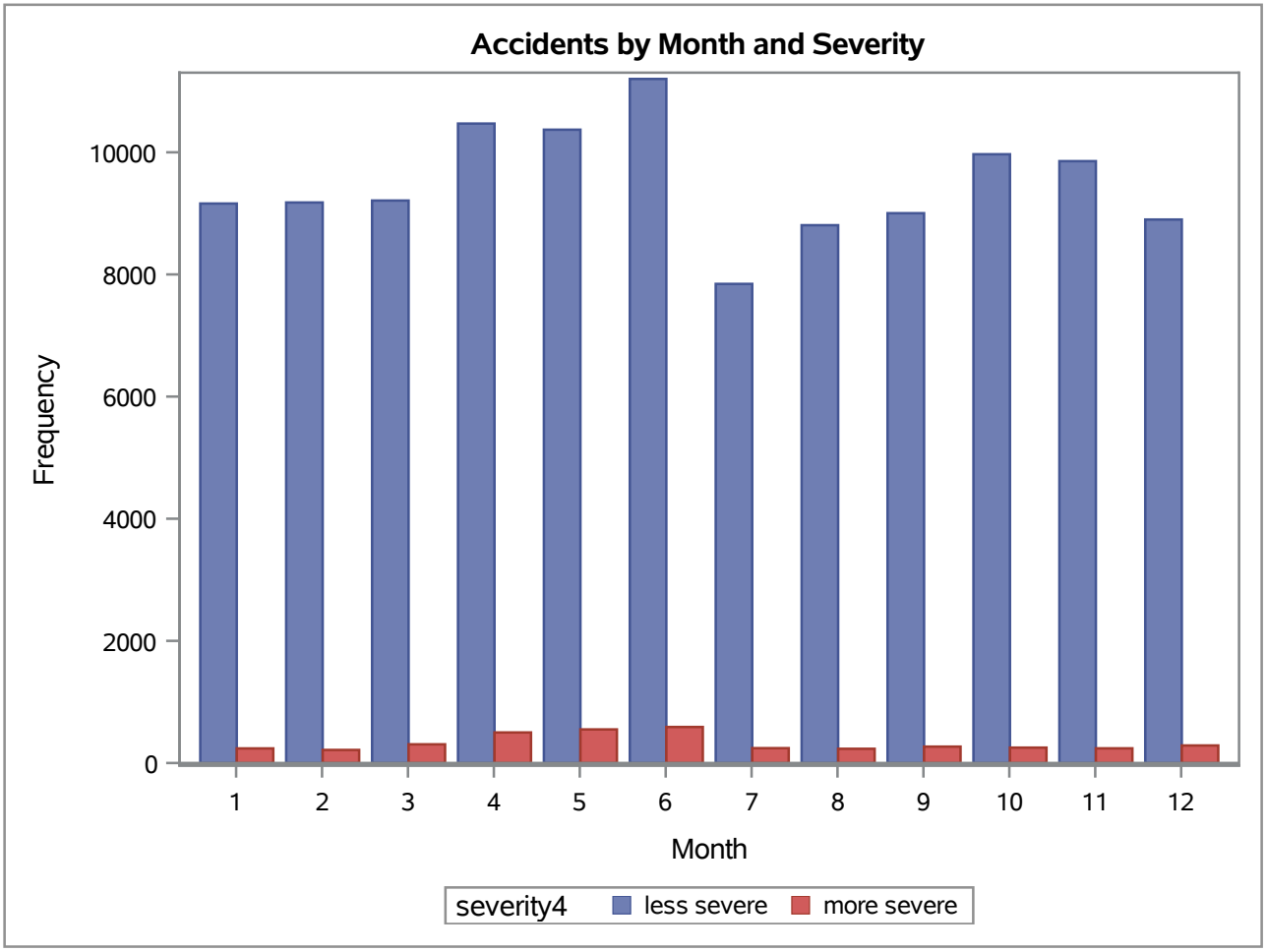
The time factors years, months, days, and hours were investigated to see how they affected severity.

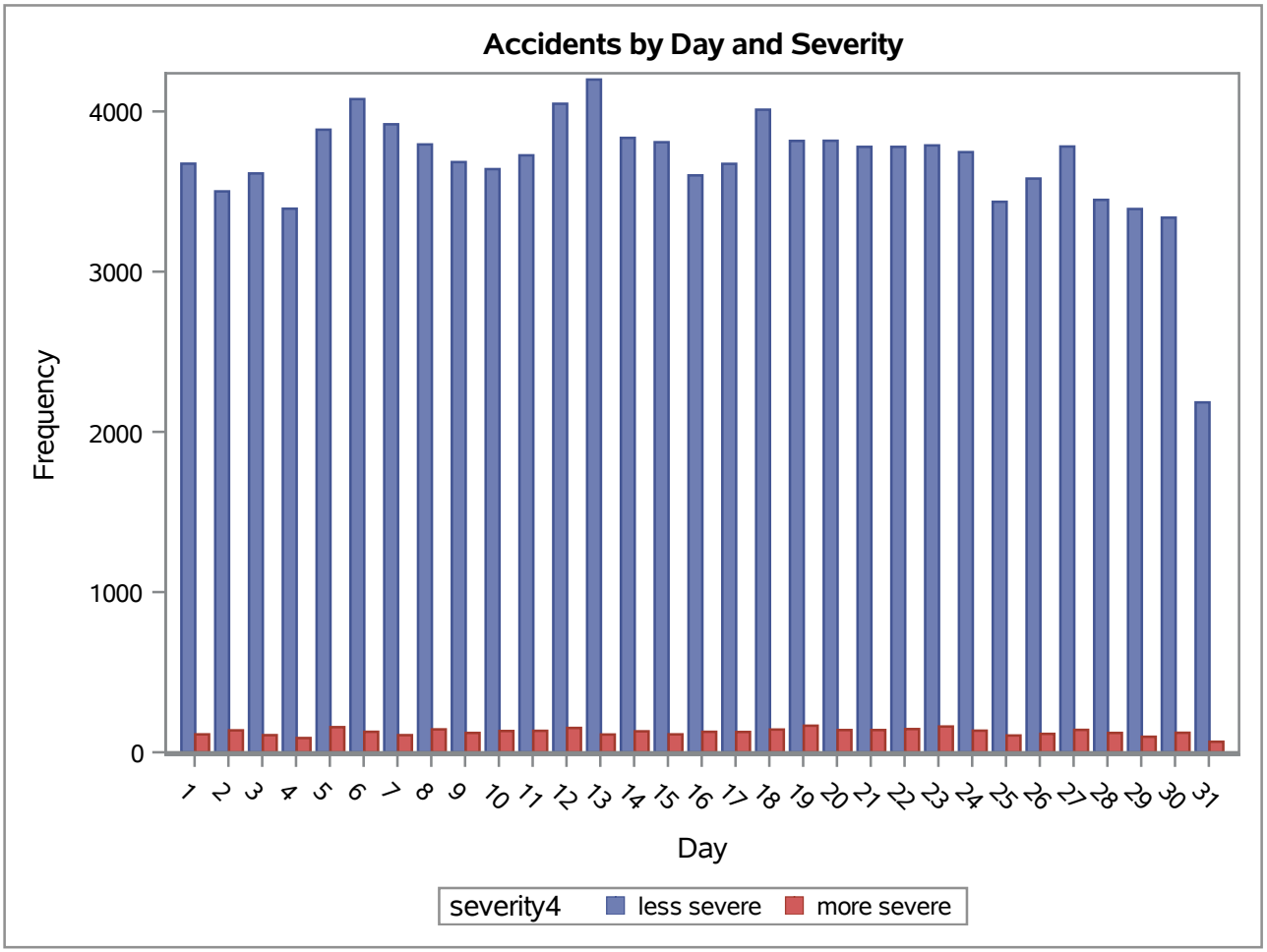
Frequency Percent Row Pct Col Pct	Table of severity4 by Year					
	Year					
severity4	**2016**	**2017**	**2018**	**2019**	**2020**	**Total**
less severe	14266 12.10 12.52 96.52	25249 21.42 22.15 97.16	26691 22.64 23.42 97.48	31679 26.87 27.80 97.69	16081 13.64 14.11 92.91	113966 96.67
more severe	514 0.44 13.11 3.48	739 0.63 18.84 2.84	691 0.59 17.62 2.52	750 0.64 19.12 2.31	1228 1.04 31.31 7.09	3922 3.33
Total	14780 12.54	25988 22.04	27382 23.23	32429 27.51	17309 14.68	117888 100.00

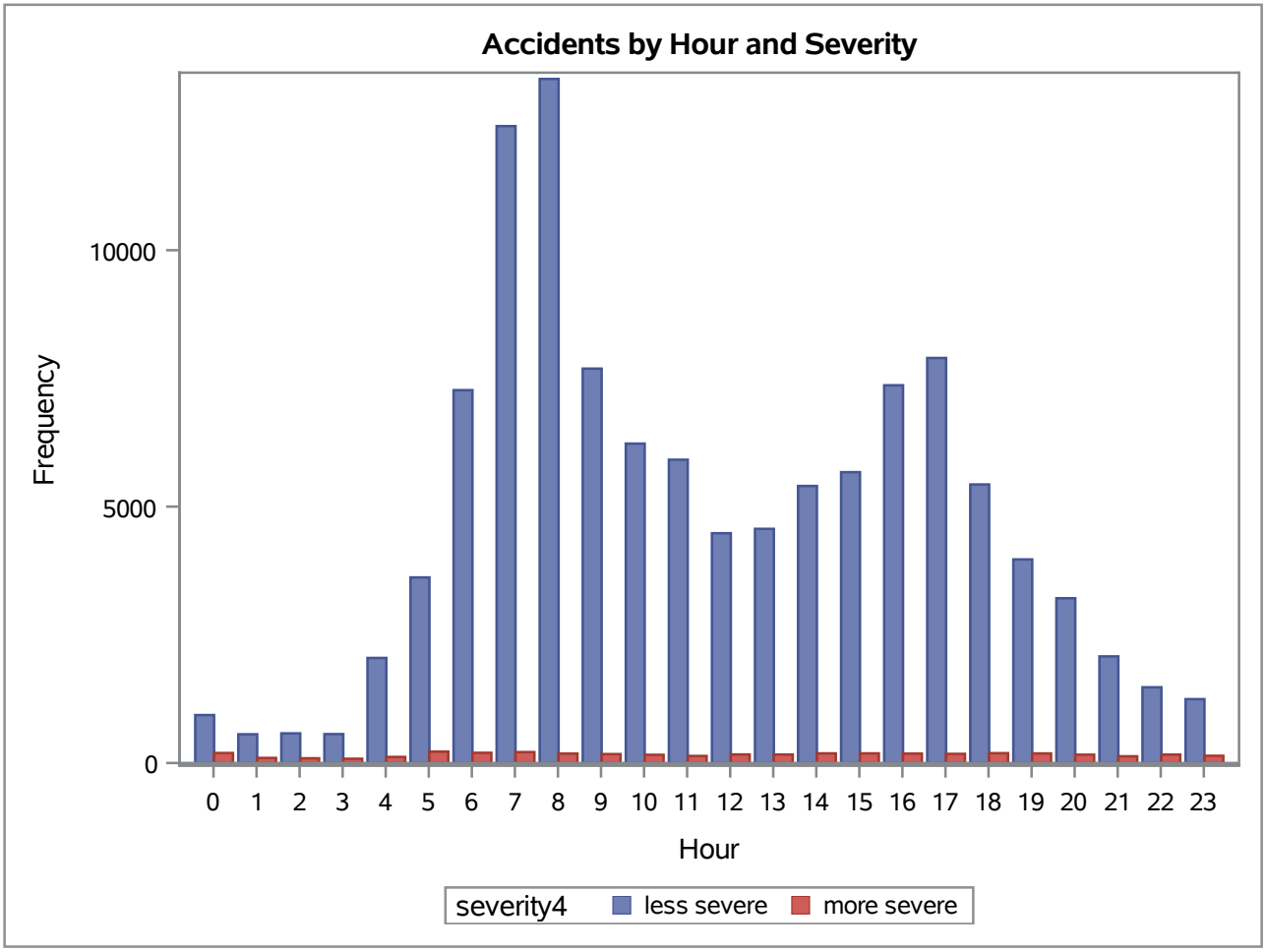
The frequency table by year shows an increase in the number of less severe accidents from 2016-2019, with the proportion of total accidents being the highest in 2019 at %26.87. For the most severe accidents, there was an increase for the whole time period from 2016-2020.



There was a large increase in the number of accidents in Monroe from 2016-2020. In contrast, the accidents in Westchester halved from 2016-2020. The remaining eight counties show a steady trend over the four year period.

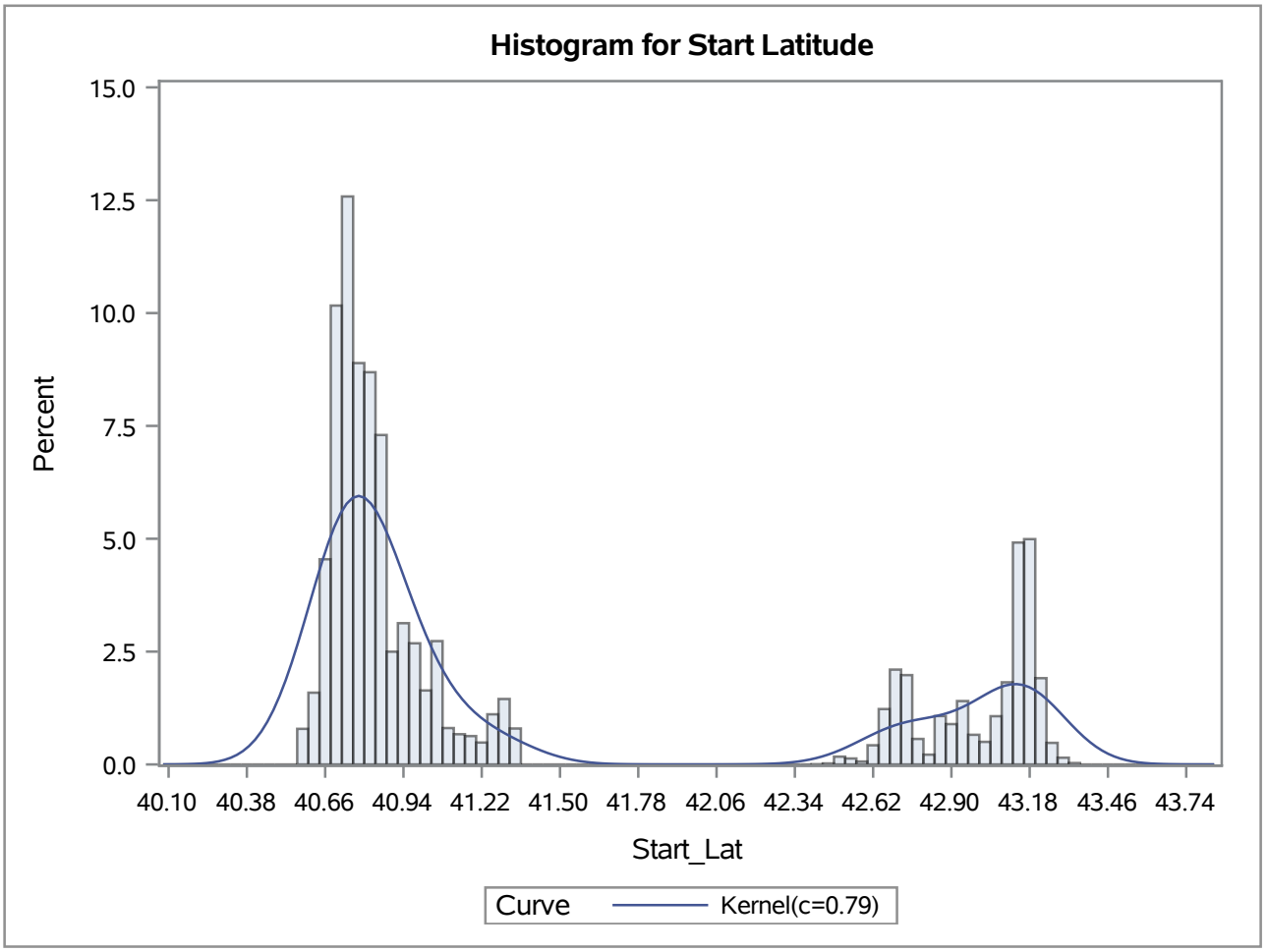

Accidents by Month and Severity
Frequency
0
2000
4000
6000
8000
10000
1
2
3
4
5
6
7
8
9
10
11
12
Month
severity4
less severe
more severe

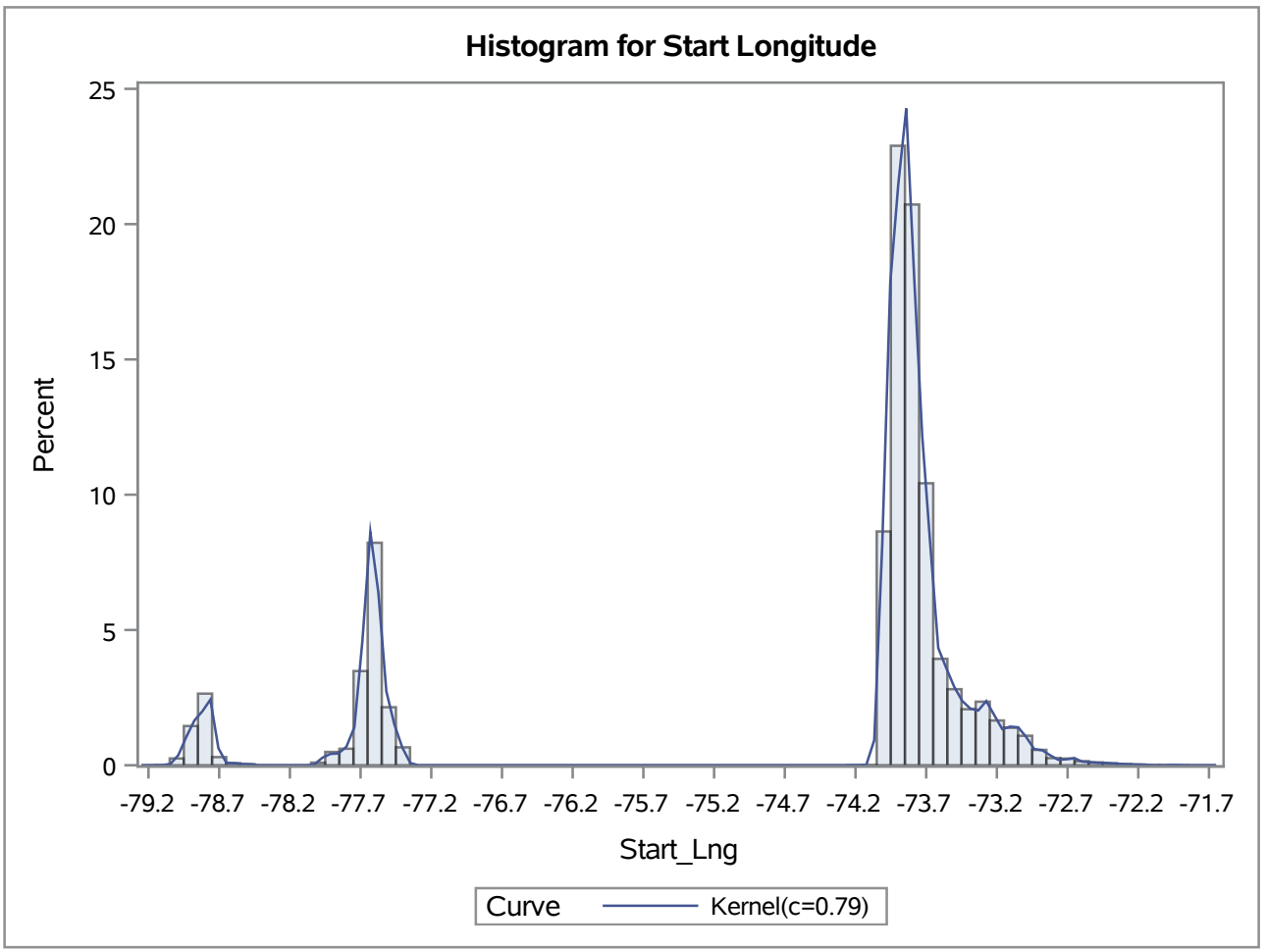

Accidents by Day and Severity
Frequency
0
1000
2000
3000
4000
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
Day
severity4
less severe
more severe



We can observe a drop in accidents in July which can maybe be explained due to school closure and holiday time. Looking at days, there is a somewhat clear seven-day trend. The drop in day 31 is not meaningful as it does not exist in half the months of the year. Furthermore, the distribution of accidents by hour is bimodal and clearly reflects the morning and evening rush hours; with the highest number of accidents during these periods. This trend is more present in the less severe accidents. In general, the trend between the less and more severe accidents does not differ much but

Histogram for Location Variables





Inference: Weather Conditions affecting Severity

The weather conditions that are the most correlated with severity are windchill, temperature, and visibility.

1 With Variables:	severity4
7 Variables:	Pressure Humidity Temp WindChill Visibility WindSpeed Precip

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	WindChill	Temp	Visibility
severity4	0.03853	0.01708	0.01462
	<.0001	<.0001	<.0001
	68885	117869	117374

The distribution of wind chill, temperature and visibility for both level of severity was assessed. Visually, the distribution of all three weather variables did not vary across severity. In addition, a t-test was performed totest if the weather condition on average was significantly different between more and less severe accidents. It was found for all three variables that the equality of variancesacross severity was rejected (p-value <0.01). Hence, we look at the Satterthwaite testwhich showed a significant difference between the weather conditions and theseverity (p-value <0.001). However, the QQ plots of the weather variables across both levels indicated clear violations of normality hence these conclusions may not be reliable. A transformation of the data or other methods may be used.

Conclusion

In this analysis, we collected information on the counties most responsible for the accidents in New York. Furthermore, we saw that different time factors had an influence on the number of accidents. A conclusion on if weather conditions were affecting severity could not be made due to a strong violation of normality in the data, however visually the distribution of the weather conditions that were most correlated with severity did not differ.

References

Blincoe, L. J., Miller, T. R., Zaloshnja, E., & Lawrence, B. A. (2015, May). The economic and societal impact of motor vehicle crashes, 2010. (Revised) (Report No. DOT HS 812 013). Washington, DC: National Highway Traffic Safety Administration.