Synthetic Biology Course

Python Statistics Project

compiled by Emilis Gegevicius

Dataset chosen: US Traffic Accidents (2.25 million records).
Source: https://www.kaggle.com/sobhanmoosavi/us-accidents

Hypothesis: Severity of traffic accidents depends on weather conditions (harsher conditions = higher severity)

The dataset has been chosen as analysis of correlation between severity of accident and weather conditions should be interesting and can be easily logically checked (e.g. lower temperatures should cause more Severe accidents due to icy roads)

Out of 49 columns present in the initial dataset, 7 variables have been chosen and converted to metric scale from scale present in uS (NaN is already provided in place of missing data):

Severity – ordinal variable (from 1 to 4, where 1 indicates the least impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay).
Humidity(%) – continuous variable
Temperature(°C) – continuous variable
Pressure(mbar) – continuous variable
Wind_Speed(kmh) – continuous variable
Precipitation(mm) – continuous variable
Visibility(km) – continuous variable

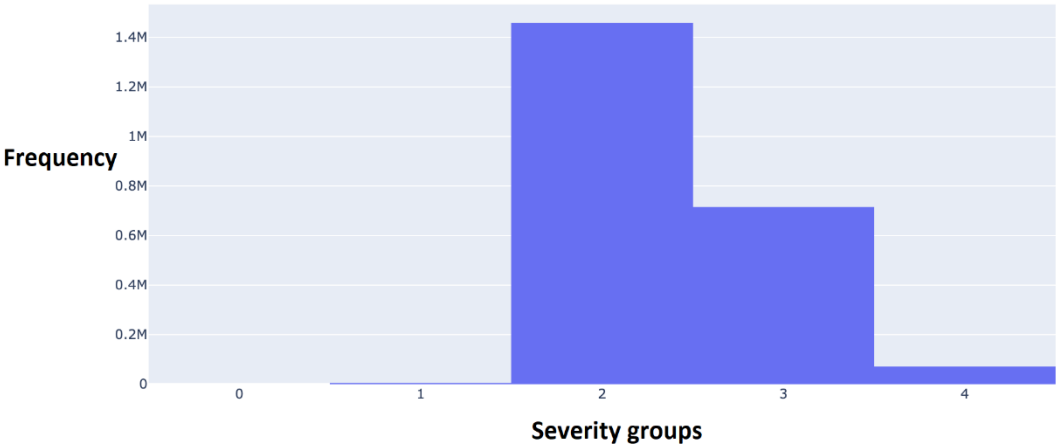
As no further description is present for how exactly Severity is evaluated, it is possible, that there could be bias in this measurement between different people evaluating Severity. Nevertheless, if Severity was precisely described and it's scale evaluated in exact time of traffic impact, bias could be avoided. Another possible bias could occur due to poorer traffic conditions in generally when weather conditions are bad. i.e. even though Severity is not high, long delays could occur.

It is possible to externally validate data by gathering information from different accidents during specific weather conditions (present in current dataset). For internal validation, however, more information on each variable would be needed (e.g. instrument type and measurement method for meteorological information)

There is some missing data in the dataset, to which NaN is assigned. For descriptive analysis, data can be analyzed with NaN in place of missing data, nevertheless, for other tests (such

as normality Anderson-Darling test, ordinal regression, etc.) missing data has to be removed.

There is a total of 2243922 measurements, which are distributed as shown in histogram below:




```
{'Severity0': 17, 'Severity1': 814, 'Severity2': 1455524, 'Severity3': 715582, 'Severity4': 72002}
```

Due to very low number of events of Severity = 0 (17 events, corresponding to 0.0008%), it is dropped out of dataset.

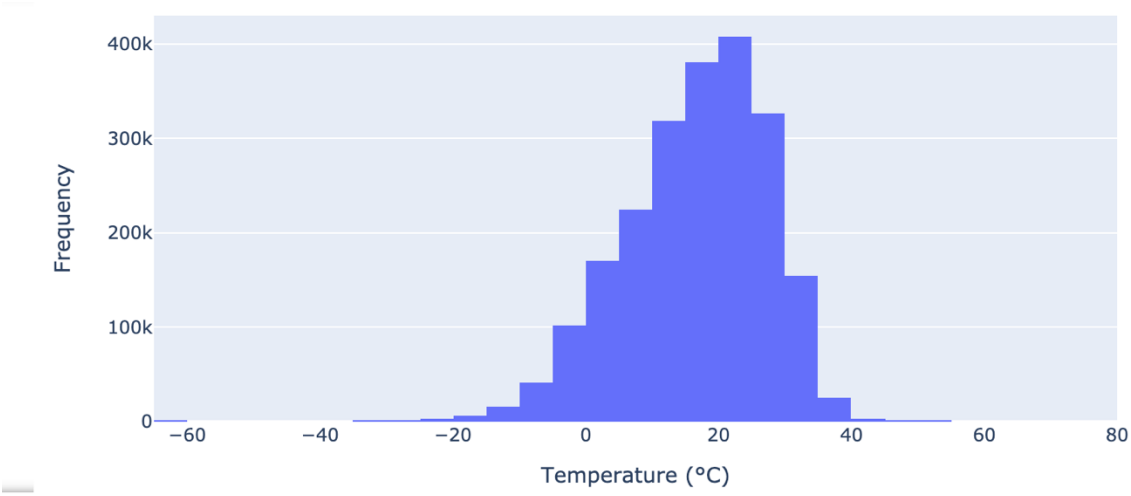
Descriptive Statistics for each of the variable is as follows:

	Severity	Humidity(%)	Temperature(°C)	Pressure(mbar)	Wind_Speed(kmh)	Precipitation(mm)	Visibility(km)
count	2.243922e+06	2.179455e+06	2.181657e+06	2.186642e+06	1.800968e+06	264473.000000	2.172562e+06
mean	2.382710e+00	6.592758e+01	1.624025e+01	1.017185e+03	1.423310e+01	1.535142	1.468380e+01
std	5.487658e-01	2.243017e+01	1.063676e+01	7.677776e+00	8.003607e+00	11.168318	4.806096e+00
min	1.000000e+00	4.000000e+00	-6.100000e+01	0.000000e+00	1.931213e+00	0.000000	0.000000e+00
25%	2.000000e+00	5.000000e+01	9.390000e+00	1.013207e+03	9.334195e+00	0.000000	1.609344e+01
50%	2.000000e+00	6.800000e+01	1.722000e+01	1.016933e+03	1.303569e+01	0.254000	1.609344e+01
75%	3.000000e+00	8.500000e+01	2.439000e+01	1.020996e+03	1.850746e+01	1.016000	1.609344e+01
max	4.000000e+00	1.000000e+02	7.700000e+01	1.118863e+03	1.324168e+03	274.320000	2.253082e+02

Descriptive statistics for each variable is going to be separately presented and commented in further sections.

Normality Analysis

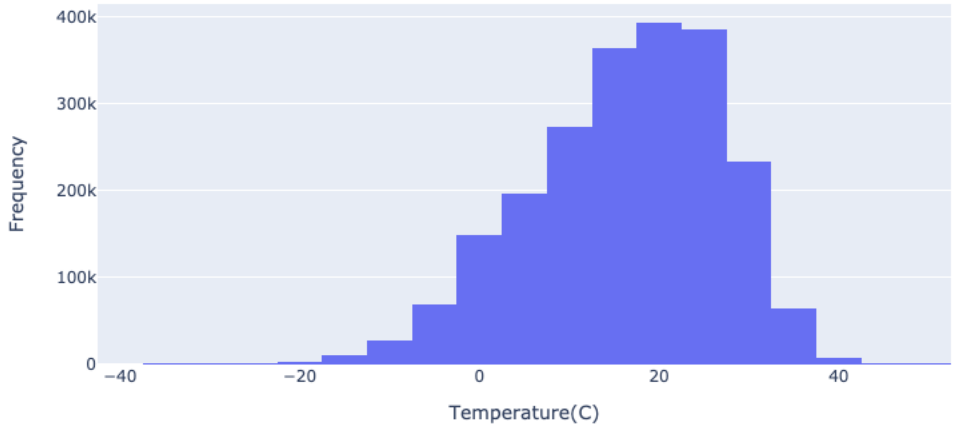
Temperature (ºC) variable



Lowest temperature in dataset is -61.0 °C
Highest temperature in dataset is 77.0 °C

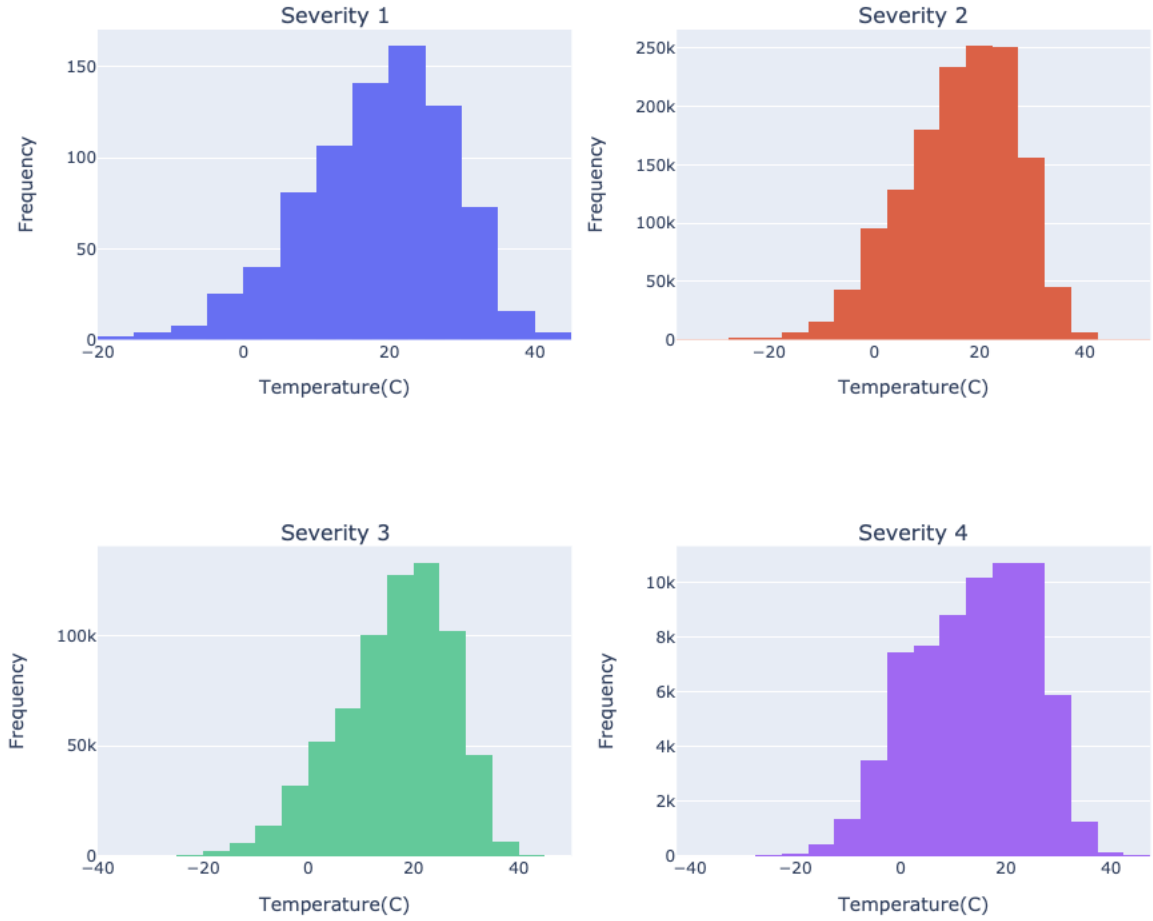
From the initial histogram, it is apparent that extreme temperatures are present in the dataset. Highest recorded temperatures in US ranges from 45-50ºC, lowest -55 - -45ºC. More extreme values could have occurred due to specific temperature measurement errors (e.g. satellite measurements are not precise) or ground temperature readings, which are usually higher than air (no additional information is available on this topic). As these outliers are possible errors in measurement, it is removed from further analysis.

30 measurements of temperature have been removed



Lowest Temperature(C) in dataset is -40.0
Highest Temperature(C) in dataset is 50.0

From the first look in histogram, it is obvious that temperature variable data is not normally distributed and median is shifted towards higher values. The same thing is observed in each Severity groups. Data does not represent Gauss.



	Mean	Median	Standard Deviation	Standard Error
Severity1	18.243005	19.00	10.099228	0.358860
Severity2	16.420982	17.78	10.580025	0.008884
Severity3	16.106557	17.22	10.650338	0.012786
Severity4	13.839206	15.00	11.271631	0.042979

This is proved further with Anderson-Darling test , which is statistical test of whether or not a dataset comes from a certain probability distribution, e.g., the normal distribution. H_0 for this test: The data follow a normal distribution. Shapiro-Wilk test is not suitable as there are more than 500 measurements for the variable.

Anderson-Darling test (Critical Value = p value) proves, that sample does not look Gaussian as no critical value <0.05 is present.

Title	Sample Size	Statistic	Significance Level	Critical Value	Comment
Severity 1	792	3.8293	15.0	0.573	Sample does not look Gaussian (reject H0)
Severity 1	792	3.8293	10.0	0.653	Sample does not look Gaussian (reject H0)
Severity 1	792	3.8293	5.0	0.783	Sample does not look Gaussian (reject H0)
Severity 1	792	3.8293	2.5	0.913	Sample does not look Gaussian (reject H0)
Severity 1	792	3.8293	1.0	1.087	Sample does not look Gaussian (reject H0)

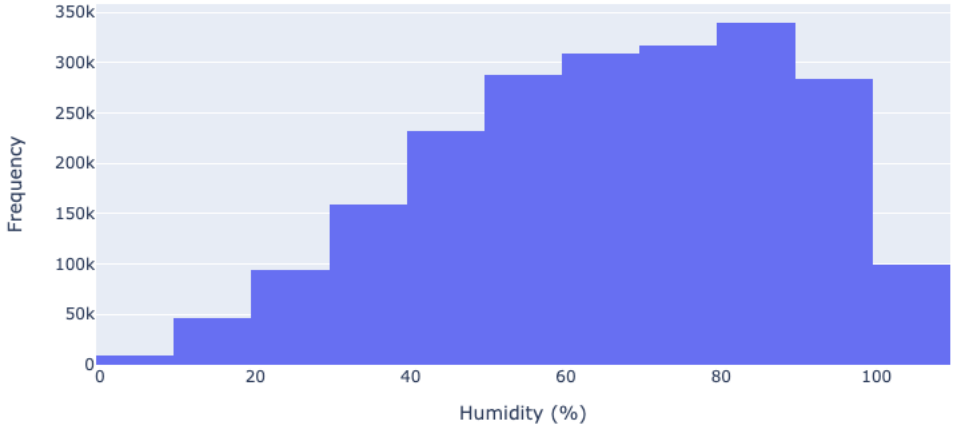
Title	Sample Size	Statistic	Significance Level	Critical Value	Comment
Severity 2	1418248	7030.7939	15.0	0.576	Sample does not look Gaussian (reject H0)
Severity 2	1418248	7030.7939	10.0	0.656	Sample does not look Gaussian (reject H0)
Severity 2	1418248	7030.7939	5.0	0.787	Sample does not look Gaussian (reject H0)
Severity 2	1418248	7030.7939	2.5	0.918	Sample does not look Gaussian (reject H0)
Severity 2	1418248	7030.7939	1.0	1.092	Sample does not look Gaussian (reject H0)

Title	Sample Size	Statistic	Significance Level	Critical Value	Comment
Severity 3	693815	4637.7244	15.0	0.576	Sample does not look Gaussian (reject H0)
Severity 3	693815	4637.7244	10.0	0.656	Sample does not look Gaussian (reject H0)
Severity 3	693815	4637.7244	5.0	0.787	Sample does not look Gaussian (reject H0)
Severity 3	693815	4637.7244	2.5	0.918	Sample does not look Gaussian (reject H0)
Severity 3	693815	4637.7244	1.0	1.092	Sample does not look Gaussian (reject H0)

Title	Sample Size	Statistic	Significance Level	Critical Value	Comment
Severity 4	68779	380.4154	15.0	0.576	Sample does not look Gaussian (reject H0)
Severity 4	68779	380.4154	10.0	0.656	Sample does not look Gaussian (reject H0)
Severity 4	68779	380.4154	5.0	0.787	Sample does not look Gaussian (reject H0)
Severity 4	68779	380.4154	2.5	0.918	Sample does not look Gaussian (reject H0)
Severity 4	68779	380.4154	1.0	1.092	Sample does not look Gaussian (reject H0)

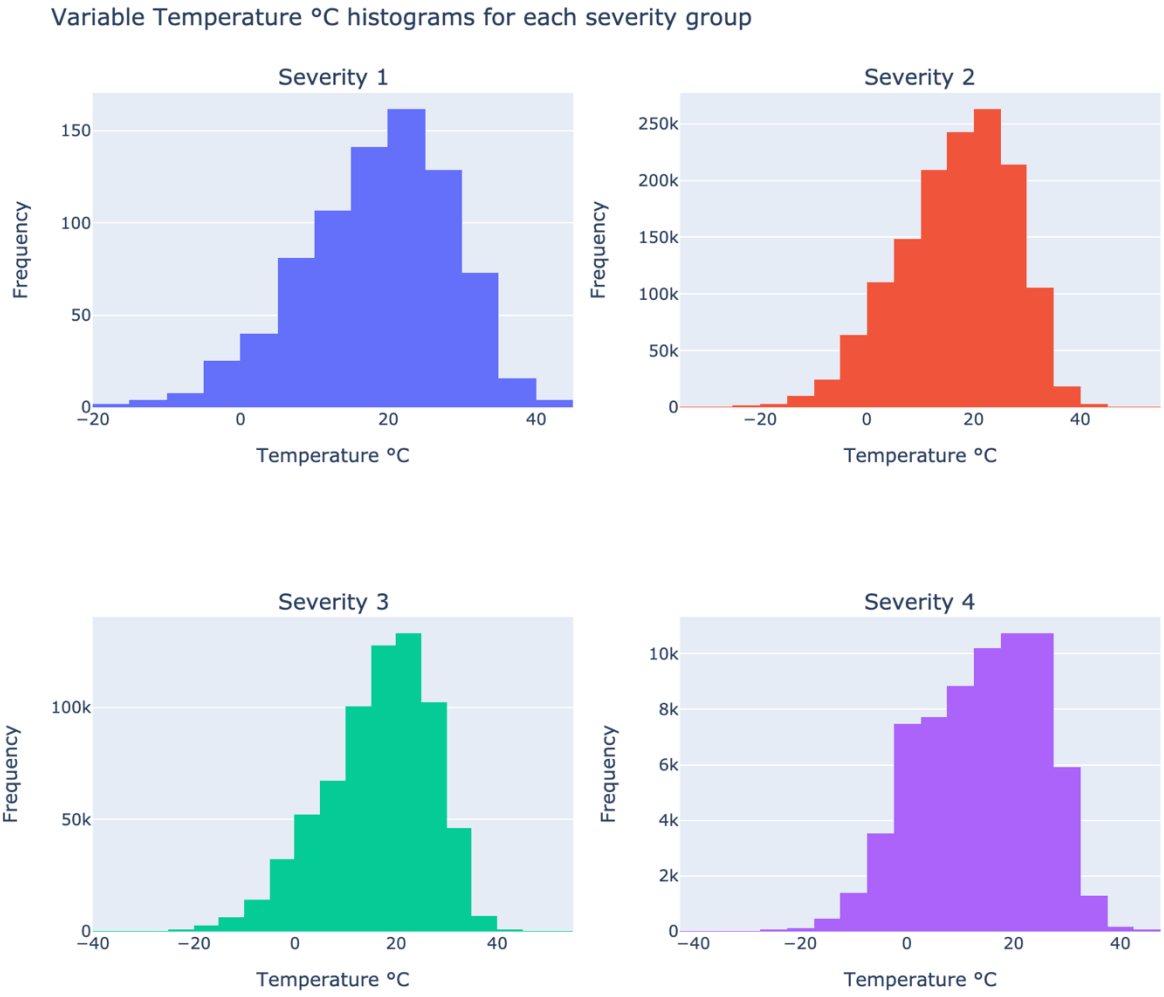
There is a possibility, that these tests could fail due to large amount of data present in the dataset. Nevertheless, even though data failed normality tests, we could assume, that it is more or less normally distributed, in order to use parametric tests.

Humidity(%) variable



Lowest humidity in dataset is 4.0%
Highest humidity in dataset is 100.0%

Humidity variable values are not normally distributed with median shifted to higher values. The same thing is seen from humidity histograms between different Severity groups.



	Mean	Median	Standard Deviation	Standard Error
Severity1	65.200000	68.0	22.390146	0.796606
Severity2	65.694491	68.0	22.541468	0.018938
Severity3	66.150965	68.0	22.196396	0.026659
Severity4	68.489581	71.0	22.293683	0.085072

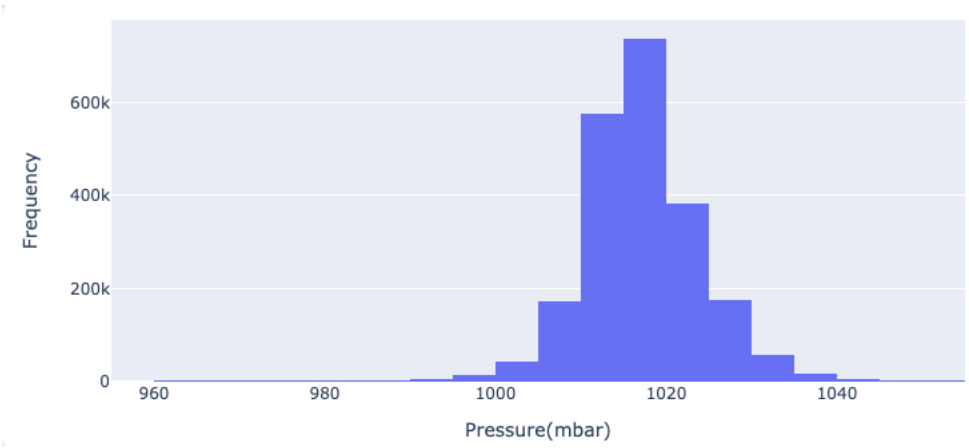
Anderson-Darling test indicates that data does not look normal (simplified less aesthetic table provided to save space)

```
Severity1 Statistic: 5.731
Significance level: 15.000: Critical Value : 0.573, data does not look normal (reject H0)
Significance level: 10.000: Critical Value : 0.653, data does not look normal (reject H0)
Significance level: 5.000: Critical Value : 0.783, data does not look normal (reject H0)
Significance level: 2.500: Critical Value : 0.913, data does not look normal (reject H0)
Significance level: 1.000: Critical Value : 1.087, data does not look normal (reject H0)
Severity2 Statistic: 11541.535
Significance level: 15.000: Critical Value : 0.576, data does not look normal (reject H0)
Significance level: 10.000: Critical Value : 0.656, data does not look normal (reject H0)
Significance level: 5.000: Critical Value : 0.787, data does not look normal (reject H0)
Significance level: 2.500: Critical Value : 0.918, data does not look normal (reject H0)
Significance level: 1.000: Critical Value : 1.092, data does not look normal (reject H0)
Severity3 Statistic: 5747.477
Significance level: 15.000: Critical Value : 0.576, data does not look normal (reject H0)
Significance level: 10.000: Critical Value : 0.656, data does not look normal (reject H0)
Significance level: 5.000: Critical Value : 0.787, data does not look normal (reject H0)
Significance level: 2.500: Critical Value : 0.918, data does not look normal (reject H0)
Significance level: 1.000: Critical Value : 1.092, data does not look normal (reject H0)
Severity4 Statistic: 768.965
Significance level: 15.000: Critical Value : 0.576, data does not look normal (reject H0)
Significance level: 10.000: Critical Value : 0.656, data does not look normal (reject H0)
Significance level: 5.000: Critical Value : 0.787, data does not look normal (reject H0)
Significance level: 2.500: Critical Value : 0.918, data does not look normal (reject H0)
Significance level: 1.000: Critical Value : 1.092, data does not look normal (reject H0)
```

Nevertheless, we are going to assume normal distribution for this variable as well.

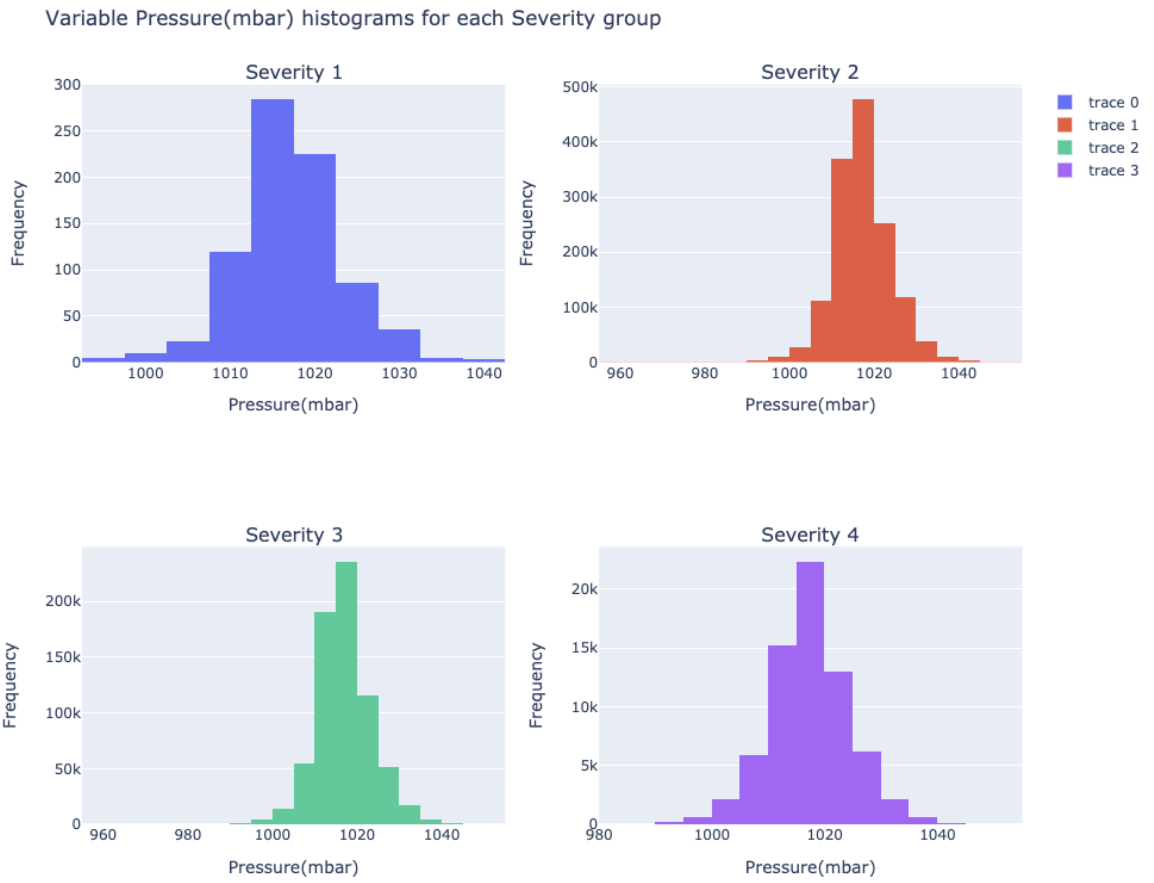
Pressure(mbar) variable

Pressure variable contains some extreme data points, which could have occurred due to errors in measurements. Highest ever recorded atmospheric pressure: 1083.3 mbar; Lowest - 870 mbar. There are 68 extreme values in this dataset, which was removed. Also, as there were only 7 pressure values present up to 950mbar (compared to 2 million), it was removed from this variable and 950-1055 mbar range has been chosen for further analysis.



Lowest Pressure(mbar) in dataset is 955.6388817333427
Highest Pressure(mbar) in dataset is 1054.860069666677

Once again, it is obvious that data is not normally distributed, nevertheless, when ploted group by group, Severity 1 and Severity 4 are very close to normal distribution.



	Mean	Median	Standard Deviation	Standard Error
Severity1	1017.069676	1016.932517	6.235253	0.221142
Severity2	1017.293983	1016.932517	6.640754	0.005570
Severity3	1017.003160	1016.593878	6.605718	0.007921
Severity4	1017.384161	1017.271155	7.205315	0.027432

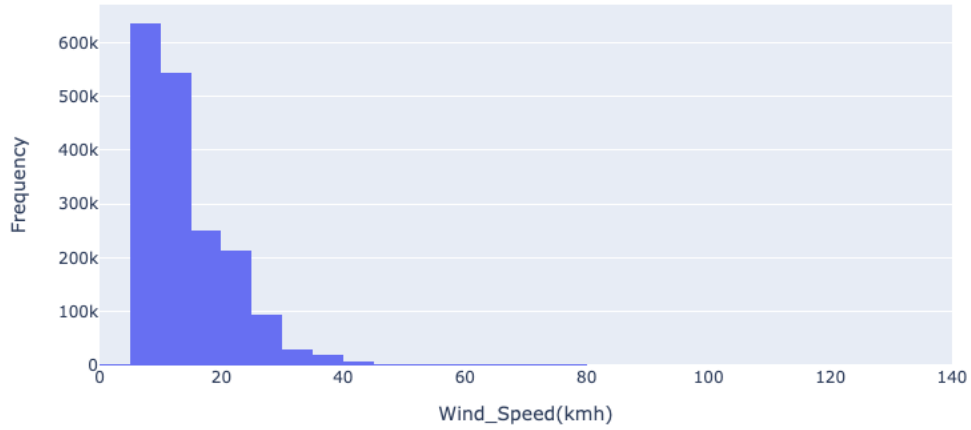
Anderson-Darling indicates, that data does not look normal.

```
Severity1 Statistic: 2.906
Significance level: 15.000: Critical Value : 0.573, data does not look normal (reject H0)
Significance level: 10.000: Critical Value : 0.653, data does not look normal (reject H0)
Significance level: 5.000: Critical Value : 0.783, data does not look normal (reject H0)
Significance level: 2.500: Critical Value : 0.913, data does not look normal (reject H0)
Significance level: 1.000: Critical Value : 1.087, data does not look normal (reject H0)
Severity2 Statistic: 4379.956
Significance level: 15.000: Critical Value : 0.576, data does not look normal (reject H0)
Significance level: 10.000: Critical Value : 0.656, data does not look normal (reject H0)
Significance level: 5.000: Critical Value : 0.787, data does not look normal (reject H0)
Significance level: 2.500: Critical Value : 0.918, data does not look normal (reject H0)
Significance level: 1.000: Critical Value : 1.092, data does not look normal (reject H0)
Severity3 Statistic: 2862.272
Significance level: 15.000: Critical Value : 0.576, data does not look normal (reject H0)
Significance level: 10.000: Critical Value : 0.656, data does not look normal (reject H0)
Significance level: 5.000: Critical Value : 0.787, data does not look normal (reject H0)
Significance level: 2.500: Critical Value : 0.918, data does not look normal (reject H0)
Significance level: 1.000: Critical Value : 1.092, data does not look normal (reject H0)
Severity4 Statistic: 154.910
Significance level: 15.000: Critical Value : 0.576, data does not look normal (reject H0)
Significance level: 10.000: Critical Value : 0.656, data does not look normal (reject H0)
Significance level: 5.000: Critical Value : 0.787, data does not look normal (reject H0)
Significance level: 2.500: Critical Value : 0.918, data does not look normal (reject H0)
Significance level: 1.000: Critical Value : 1.092, data does not look normal (reject H0)
```

This variable is going to be assumed to be distributed normally as well.

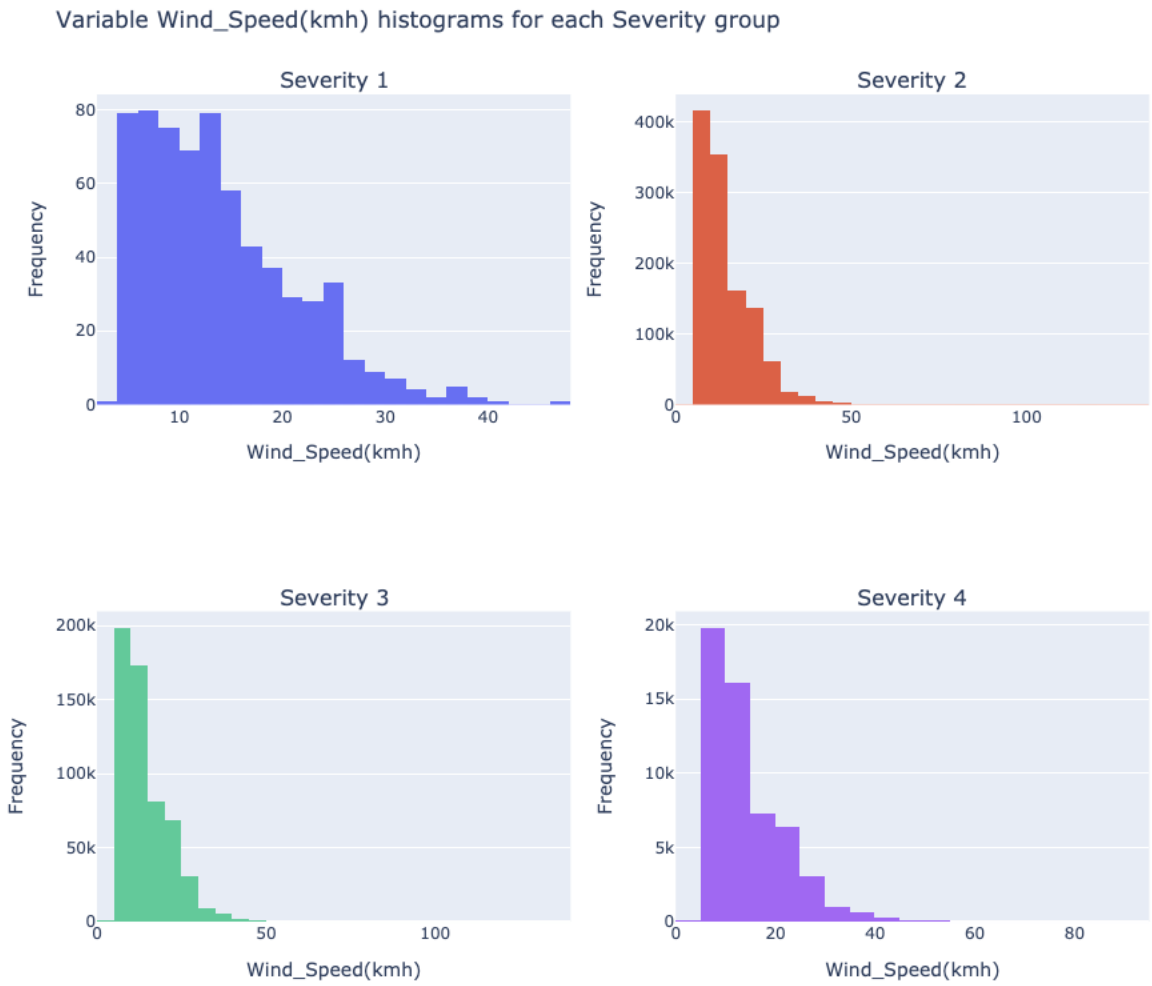
Wind Speed (kmh) variable

Once again, some extreme values are present in the dataset. Highest ever recorded wind speed in US 372km/h. There are 17 temperature values higher than this number due to measurement error, which are removed. There are only 39 values higher than 150 km/h, which is omitted, thus, selecting 0-150km/h wind speed interval.



Lowest Wind_Speed(kmh) in dataset is 1.9312128
Highest Wind_Speed(kmh) in dataset is 137.1161088

Data is obviously not normally distributed, with highest chance of value being in the first part of the interval.

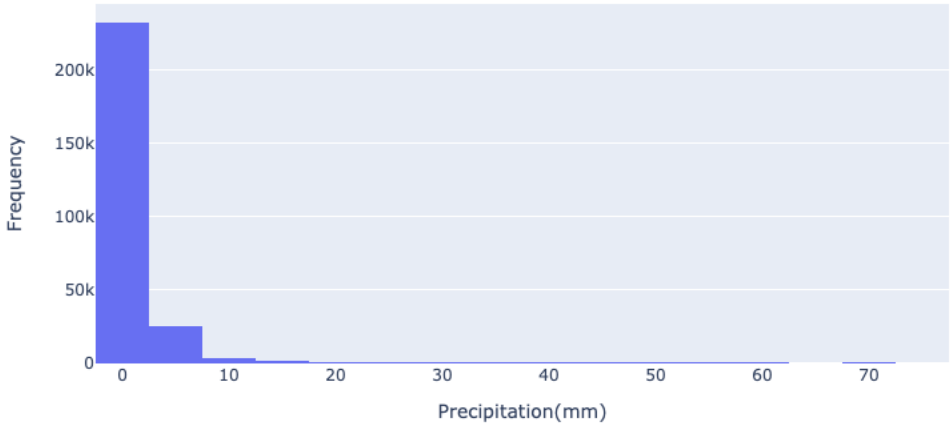


	Mean	Median	Standard Deviation	Standard Error
Severity1	14.090127	13.035686	7.248740	0.283448
Severity2	14.170377	13.035686	7.275209	0.006721
Severity3	14.299521	13.035686	7.300550	0.009643
Severity4	14.443565	13.035686	7.917426	0.033684

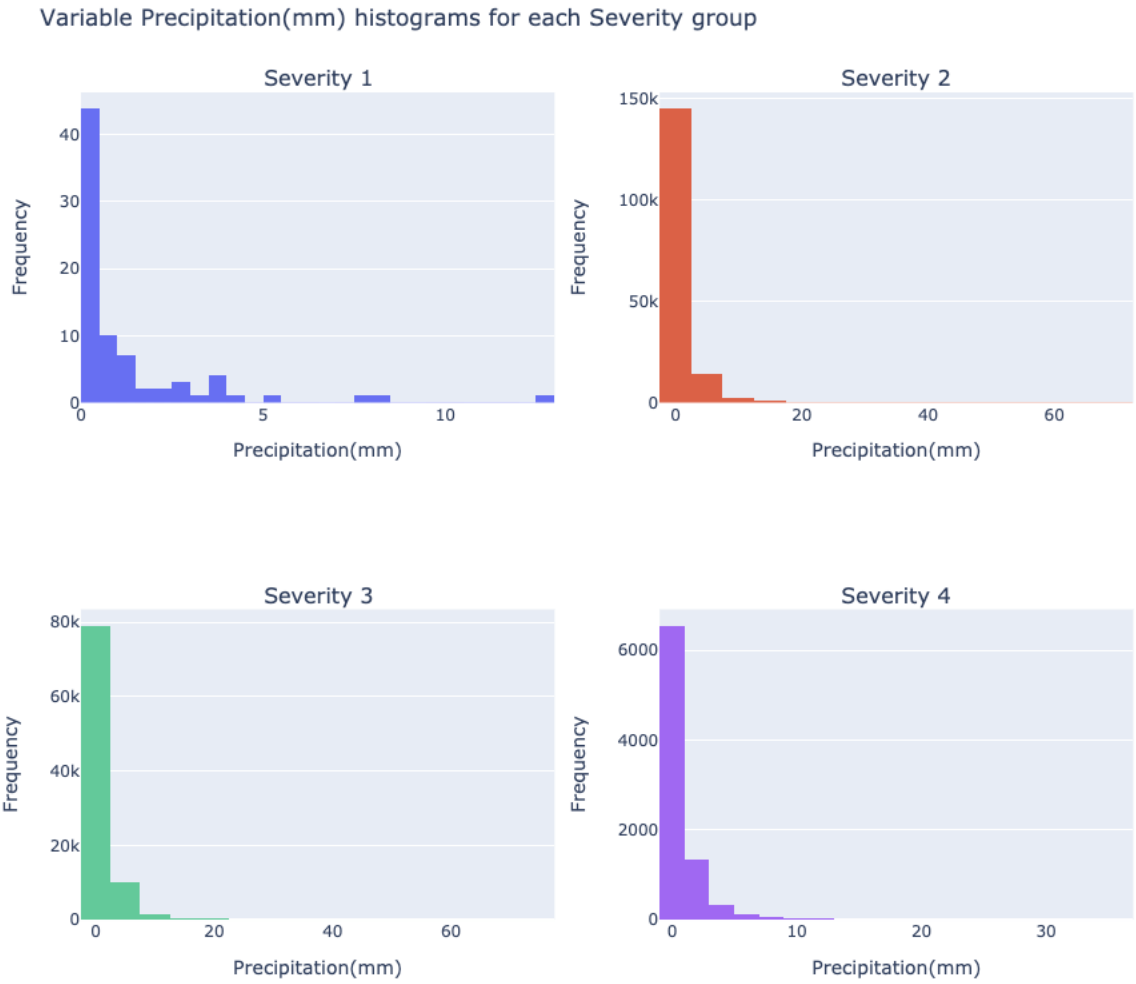
For such obviously non normally distributed data, Anderson-Darling normality test will not be conducted.

Precipitation (mm) variable

The highest category of rainfall, termed “Violent shower” is when precipitation per hour is greater than 50mm. There are 497 measurements of precipitation above 80mm, which can be classified as super extreme weather and must be removed from the dataset. It was decided to keep values between 50-80mm as a buffering zone, not to drop important values in between.



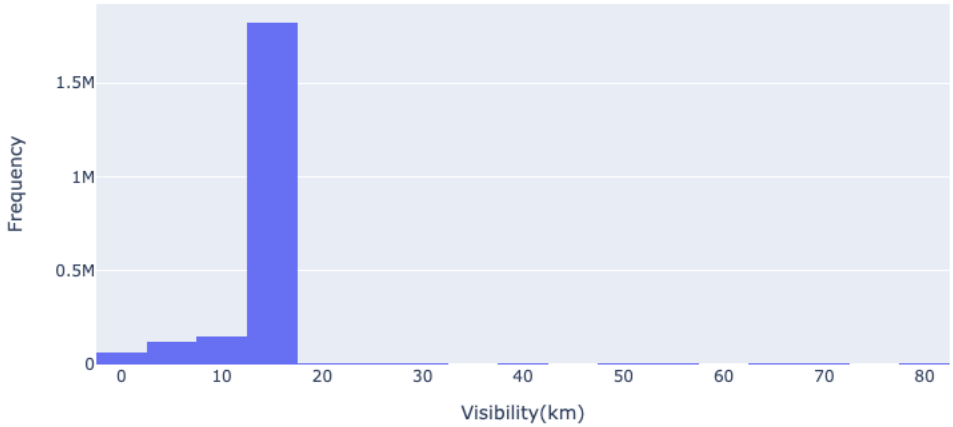
Lowest Precipitation(mm) in dataset is 0.0
Highest Precipitation(mm) in dataset is 74.67599999999999



Data is not normally distributed.

Visibility (km) variable

Even though, when visibility is reported to be 16km (10 miles), it usually means visibility is “unlimited”, we take 85km visibility as a higher upper limit (dropping 404 measurements) as there are significant amount of data present in this interval.

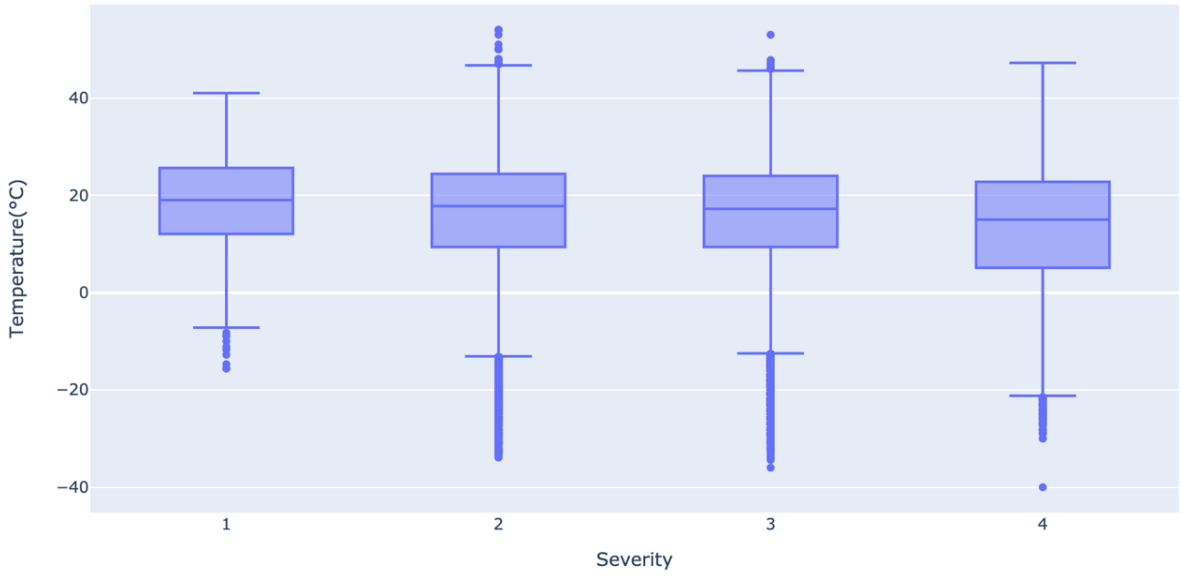


Lowest Visibility(km) in dataset is 0.0
Highest Visibility(km) in dataset is 80.4672

Data is not normally distributed

Box plot analysis

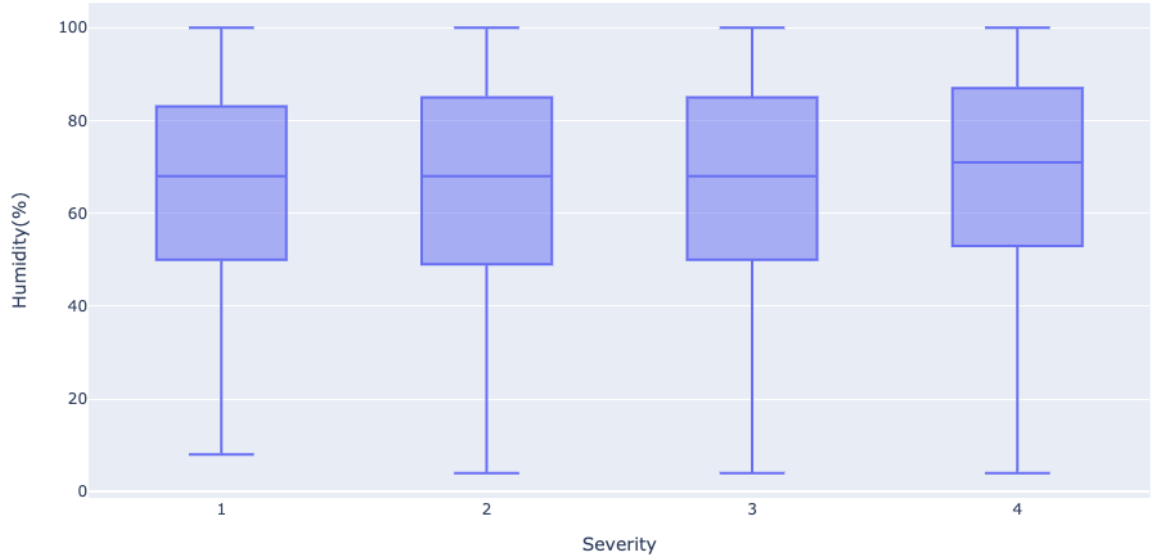
1) Temperature(C)



	Mean	Median	Standard Deviation	Standard Error
Severity1	18.243005	19.00	10.099228	0.358860
Severity2	16.420982	17.78	10.580025	0.008884
Severity3	16.106557	17.22	10.650338	0.012786
Severity4	13.839206	15.00	11.271631	0.042979

From the first look, Severity 2 and 4 box plots are obviously skewed (median is shifted towards second half of the data). The spread of data in different severity groups are similar, nevertheless, wider whiskers and higher number of extreme outliers in 2 and 3 severity indicates, that deviations from minimal and maximal temperature values (min = Q1 – 1.5IQR; max = Q3 + 1.5IQR) causes more severe accidents (Severity > 1, usually Severity 2 and 3, not the highest one). Medians and means seems to decrease with higher Severity levels, which could indicate relationship between them.

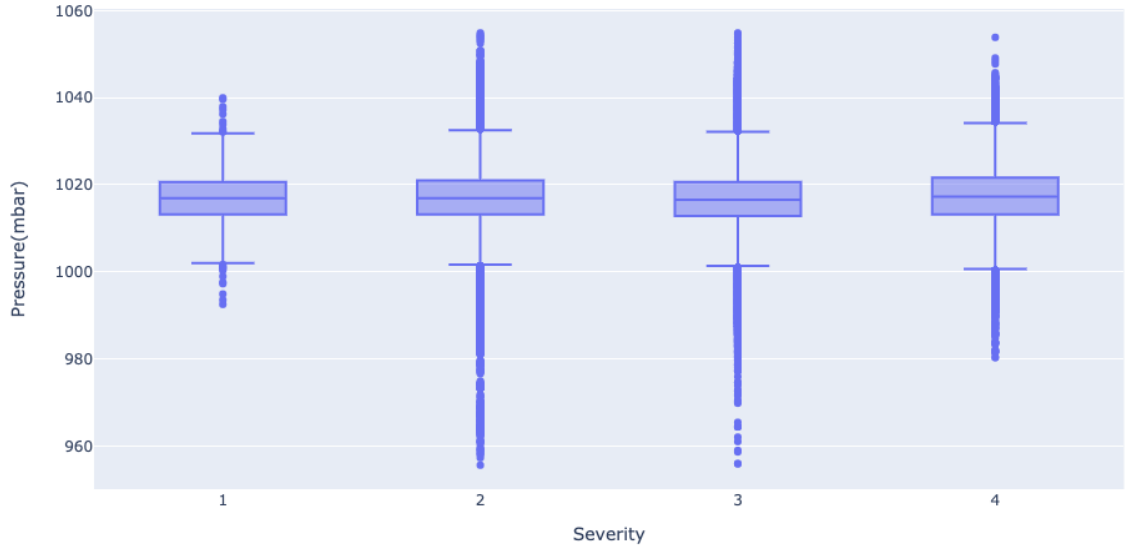
2) Humidity (%)



	Mean	Median	Standard Deviation	Standard Error
Severity1	65.200000	68.0	22.390146	0.796606
Severity2	65.694491	68.0	22.541468	0.018938
Severity3	66.150965	68.0	22.196396	0.026659
Severity4	68.489581	71.0	22.293683	0.085072

From the box plot, data looks more or less normally distributed (not skewed) with the same spread and medians (only median of Severity 4 is somewhat higher). The spread of data is very similar. Means are increasing in higher Severity groups, which could indicate a certain relationship.

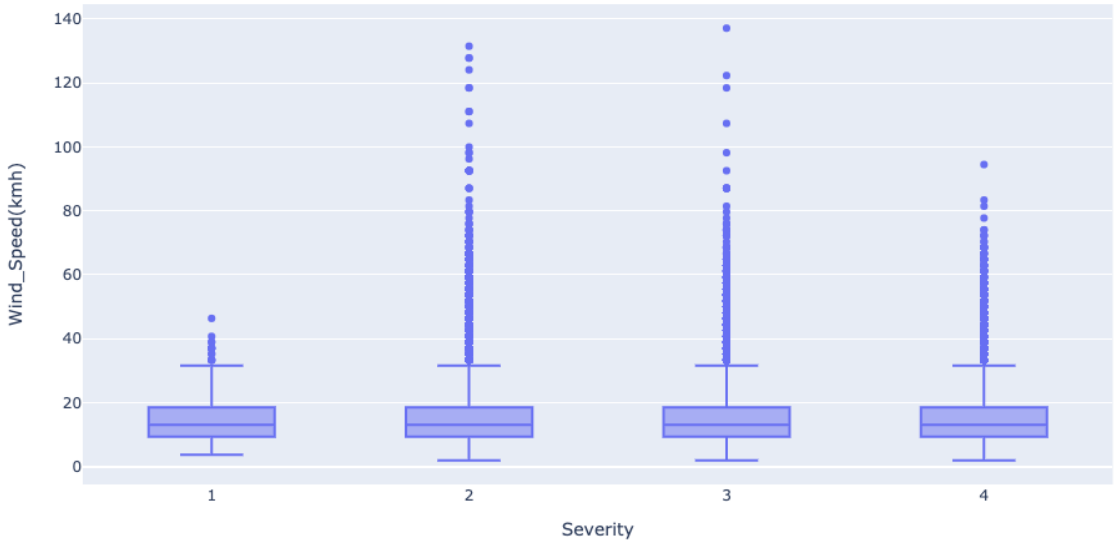
3) Pressure(mbar)



	Mean	Median	Standard Deviation	Standard Error
Severity1	1017.069676	1016.932517	6.235253	0.221142
Severity2	1017.293983	1016.932517	6.640754	0.005570
Severity3	1017.003160	1016.593878	6.605718	0.007921
Severity4	1017.384161	1017.271155	7.205315	0.027432

Median of pressure is the same throughout all the groups. Pressure values varies less than other variables. Once again, higher number of extreme outliers in 2 and 3 severity indicates, that deviations from minimal and maximal temperature values (min = Q1 – 1.5IQR; max = Q3 + 1.5IQR) causes more severe accidents (Severity > 1, usually Severity 2 and 3 , not the highest one).

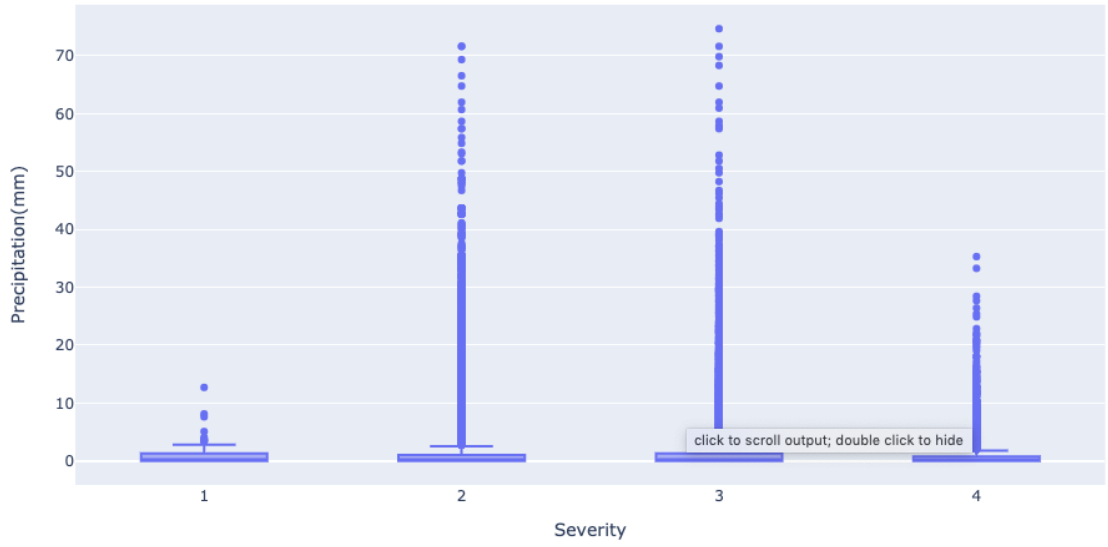
4) Wind_Speed(kmh)

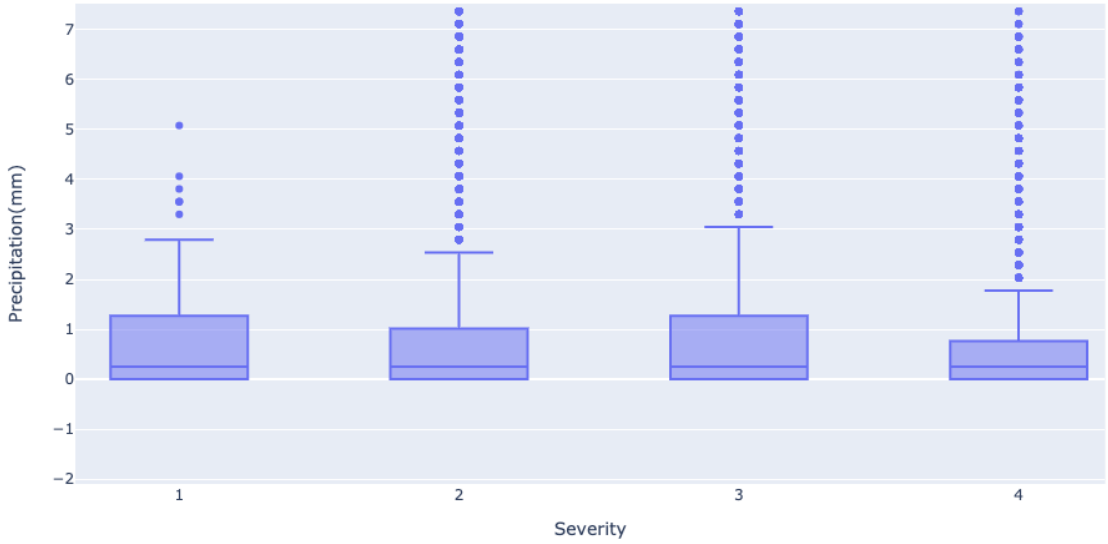


	Mean	Median	Standard Deviation	Standard Error
Severity1	14.090127	13.035686	7.248740	0.283448
Severity2	14.170377	13.035686	7.275209	0.006721
Severity3	14.299521	13.035686	7.300550	0.009643
Severity4	14.443565	13.035686	7.917426	0.033684

Median of wind speed is the same throughout all the groups. The same tendency of deviations from minimal and maximal temperature values (min = Q1 – 1.5IQR; max = Q3 + 1.5IQR) causing more severe accidents (usually Severity 2 and 3) is valid.

5) Precipitation(mm)

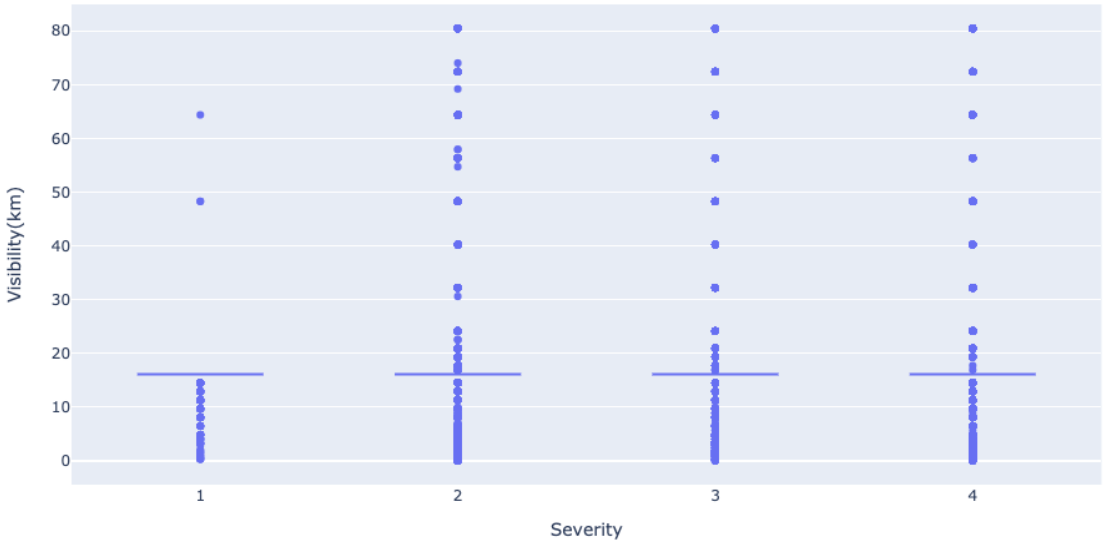




	Mean	Median	Standard Deviation	Standard Error
Severity1	1.133231	0.254	2.102799	0.238095
Severity2	0.999982	0.254	2.515988	0.006226
Severity3	1.199481	0.254	2.776146	0.009152
Severity4	0.891496	0.254	2.099970	0.022646

Not much can be seen from Precipitation(mm) box plot. Median is the same throughout all the groups, distribution is not normal. Once again the same tendency of deviations from minimal and maximal temperature values causing more severe accidents is valid.

6) Visibility(km)



	Mean	Median	Standard Deviation	Standard Error
Severity1	14.965875	16.09344	4.105253	0.146430
Severity2	14.702780	16.09344	4.544215	0.003822
Severity3	14.597476	16.09344	4.723405	0.005687
Severity4	14.595590	16.09344	5.177352	0.019808

Even less can be told from visibility box plot. Median is the same throughout all the groups, IQR is very narrow. Nevertheless, from this box plot higher visibility values (outliers) seem to be in higher Severity groups, which could indicate, that there is no apparent relationship between better visibility and lower severity.

Significant statistical differences analysis (ANOVA and Kruskal-Wallis test)

In order to test, whether all these groups are statistically different, several different tests can be made. ANOVA (and non-parametric ANOVA alternative Kruskal-Wallis test) is the most suitable test for this dataset as there are multiple groups to compare. Multiple t-tests (difference between means of compared groups) is not recommended when more than 2 groups are compared (the more hypothesis tests are used the bigger the risk of making a type I error “rejection of true null hypothesis).

```
Temperature(C) variable ANOVA value F is equal to 1353.641061453785
Temperature(C) variable ANOVA value p is equal to 0.0

Humidity(%) variable ANOVA value F is equal to 373.0323012288522
Humidity(%) variable ANOVA value p is equal to 3.021436404746539e-242

Pressure(mbar) variable ANOVA value F is equal to 315.3889881999925
Pressure(mbar) variable ANOVA value p is equal to 9.488169783036232e-205
```

The F ratio is the ratio of two mean square values. The output of the test indicates large F ration, which can be seen both when the null hypothesis is wrong (the data are not sampled from populations with the same mean) and when random sampling happened to end up with large values in some groups and small values in others. p value for ANOVA test was <0.05, which indicates, that data is not sampled from populations with the same mean, thus, it is statistically significantly different.

Kruskal–Wallis one-way analysis of variance for non-parametric variables (wind speed, precipitation and visibility)

```
Wind_Speed(kmh) variable Kruskal-Wallis Statistics value is equal to 159.28554656850787
Wind_Speed(kmh) variable Kruskal-Wallis p value is equal to 2.61404210235802e-34

Precipitation(mm) variable Kruskal-Wallis Statistics value is equal to 532.6760367643442
Precipitation(mm) variable Kruskal-Wallis p value is equal to 3.952327829835457e-115

Visibility(km) variable Kruskal-Wallis Statistics value is equal to 356.85887755841463
Visibility(km) variable Kruskal-Wallis p value is equal to 4.8806713085368856e-77
```

p value for Kruskal-Wallis test test was <0.05, which indicates, that at least one of the samples (groups) comes from a different population than the others.

For the sake of curiosity, Z-tests between different curiosity groups have been conducted and output indicated, that there are some groups, between which no significant difference exists in a set of given observations (colored in red below), i.e. p value higher than 0.05. Z test is an alternative for t-test, when there are >30 values of specific variable.

a) Temperature:

	Z Statistics	Z P Value
Severity1 vs Severity2	4.845285	1.264302e-06
Severity1 vs Severity3	5.642464	1.676336e-08
Severity1 vs Severity4	10.944719	7.043597e-28
Severity2 vs Severity3	20.240697	4.290592e-91
Severity2 vs Severity4	62.305259	0.000000e+00
Severity3 vs Severity4	52.968722	0.000000e+00

All Severity populations are different between each other.

b) Humidity

	Z Statistics	Z P Value
Severity1 vs Severity2	-0.616411	5.376230e-01
Severity1 vs Severity3	-1.203494	2.287851e-01
Severity1 vs Severity4	-4.123511	3.731405e-05
Severity2 vs Severity3	-13.885494	7.756333e-44
Severity2 vs Severity4	-31.750287	3.146428e-221
Severity3 vs Severity4	-26.326120	9.635388e-153

There are no significant differences between Severity 1 and 2 and Severity 1 and 3.

c) Pressure

	Z Statistics	Z P Value
Severity1 vs Severity2	-0.952141	3.410255e-01
Severity1 vs Severity3	0.283773	7.765840e-01
Severity1 vs Severity4	-1.225361	2.204392e-01
Severity2 vs Severity3	29.979375	1.822848e-197
Severity2 vs Severity4	-3.469099	5.222079e-04
Severity3 vs Severity4	-14.327896	1.464920e-46

There are no significant differences between Severity 1 and 2 and Severity 1 and 3 and Severity 1 and 4.

Mann Whitney U is alternative to t-test (Z-test) for non-parametric variables

d) Wind Speed

	Statistics	P Value
Severity1 vs Severity2	3.798376e+08	3.481675e-01
Severity1 vs Severity3	1.835947e+08	1.821214e-01
Severity1 vs Severity4	1.782598e+07	2.785747e-01
Severity2 vs Severity3	3.319018e+11	9.207091e-37
Severity2 vs Severity4	3.221478e+10	2.689148e-02
Severity3 vs Severity4	1.572853e+10	5.229570e-03

There are no significant differences between Severity 1 and 2 and Severity 1 and 3 and Severity 1 and 4.

e) Precipitation

	Statistics	P Value
Severity1 vs Severity2	6.005649e+06	1.808997e-01
Severity1 vs Severity3	3.562418e+06	4.535190e-01
Severity1 vs Severity4	3.078770e+05	9.409840e-02
Severity2 vs Severity3	7.137136e+09	6.156368e-107
Severity2 vs Severity4	6.841963e+08	1.530392e-05
Severity3 vs Severity4	3.659400e+08	2.092560e-33

There are no significant differences between Severity 1 and 2 and Severity 1 and 3 and Severity 1 and 4.

f) **Visibility**

	Statistics	P Value
Severity1 vs Severity2	5.438619e+08	7.575248e-02
Severity1 vs Severity3	2.623049e+08	1.500365e-02
Severity1 vs Severity4	2.618183e+07	4.597128e-02
Severity2 vs Severity3	4.818769e+11	2.276448e-79
Severity2 vs Severity4	4.805512e+10	1.905542e-03
Severity3 vs Severity4	2.341202e+10	7.602462e-05

There are no significant differences between Severity 1 and 2

It seems that Severity 1 group in several variables (Humidity, Pressure, Wind Speed, Precipitation) are not statistically different between other Severity groups. Thus, majority of differences emerge when Severity 2, 3, 4 groups are compared.

Ordinal Regression

As we are dealing with ordinal variable (Severity), Ordinal Regression model has to be fitted, in order to predict severity outcome based on weather conditions.

There are only 2 packages for Ordinal Regression available in python environment, both of which are not described thoroughly. The chosen model (*bevel*, from Shopify team, available from: https://github.com/Shopify/bevel.git) outputs beta coefficient values, nevertheless, it is not entirely clear, whether it is a pure value of coefficient or log odds.

Despite the fact, that it is not entirely clear, how to precisely indicate relationship between Severity and other variables, sign before beta coefficient (colored in red) is a good indicator of qualitative relationship. Positive β means that with the values for all other *weather conditions* variables fixed, an increase in *specific weather condition* is associated with a stochastic increase in the distribution of Severity.

Somers D (Somers' Delta) value is more important, which is a measure of agreement, telling how two pairs of variables are connected. It measures how much the prediction for dependent variable improves, based on knowing a value of independent variable. It varies between -1 and 1 (-1 indicating that all pairs disagree). In this analysis Somers D (colored in blue) varies 0.007 and 0.029, which is small, but significant (p values below 0.05 – colored in green).

a) **Temperature** – β = -0.066 (as temperature decreases, there is a higher probability of Severity to end up in higher severity group), p < 0.05

```
n=1783885
                    beta  se(beta)      p  lower 0.95  upper 0.95
attribute names
Temperature(C)   -0.0066    0.0001 0.0000     -0.0069     -0.0063  ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Somers' D = 0.024
```

b) **Humidity** – β = 0.0020 (as humidity increases, there is a higher probability of Severity to end up in higher severity group), p < 0.05

```
n=1782339
                   beta  se(beta)      p  lower 0.95  upper 0.95
attribute names
Humidity(%)      0.0020    0.0001 0.0000      0.0018      0.0021  ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Somers' D = 0.017
```

c) **Precipitation** – β = -0.0042 (as pressure decreases, there is a higher probability of Severity to end up in higher severity group), p < 0.05

```
n=1783885
                    beta  se(beta)      p  lower 0.95  upper 0.95
attribute names
Pressure(mbar)  -0.0042    0.0002 0.0000     -0.0047     -0.0038  ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Somers' D = 0.011
```

d) **Wind Speed** – β = 0.0027 (as wind speed increases, there is a higher probability of Severity to end up in higher severity group), p < 0.05

```
n=1783885
                   beta  se(beta)      p  lower 0.95  upper 0.95
attribute names
Wind_Speed(kmh) 0.0027    0.0002 0.0000      0.0023      0.0031  ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Somers' D = 0.007
```

e) **Precipitation** – β = 0.0028 (as precipitation increases, there is a higher probability of Severity to end up in higher severity group), $p < 0.05$

```
n=236007
                      beta  se(beta)      p  lower 0.95  upper 0.95
attribute names
Precipitation(mm) 0.0028    0.0005 0.0000      0.0018      0.0037  ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Somers' D = 0.029
```

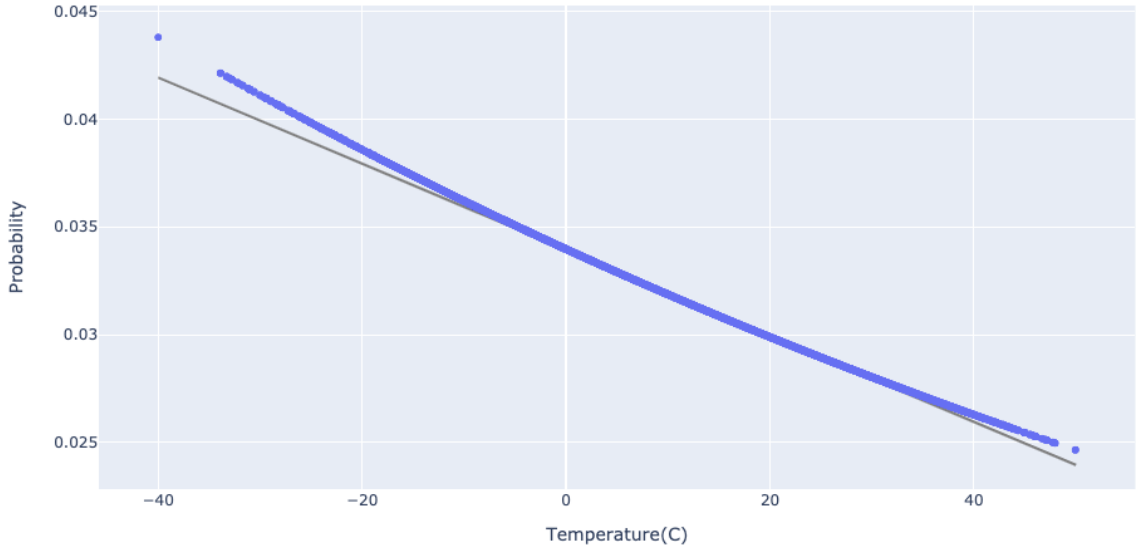
f) **Visibility** – β = 0.0028 (as visibility decreases, there is a higher probability of Severity to end up in higher severity group), $p < 0.05$

```
n=1783885
                     beta  se(beta)      p  lower 0.95  upper 0.95
attribute names
Visibility(km)  -0.0096    0.0004 0.0000    -0.0104     -0.0089  ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Somers' D = 0.015
```

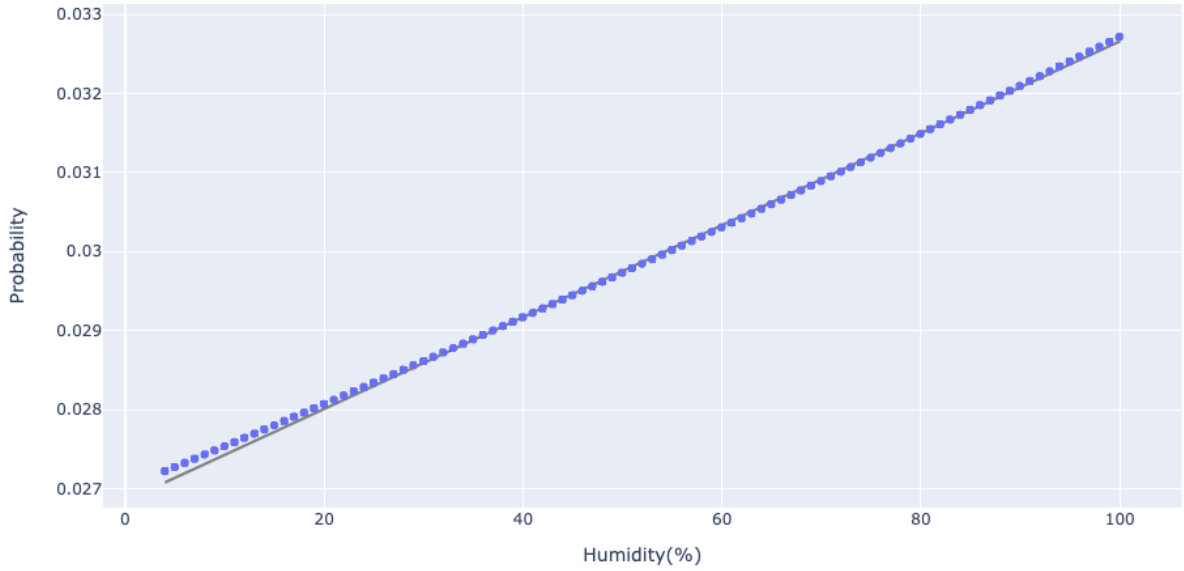
Even though, no precise assumptions can be made, I have come across, that model also outputs probability of each value being in each of different Severity groups. Thus, in this way it is possible to analyze whether higher/lower variable values corresponds to higher/lower probability of being in Severity 4 and whether there are any relation between these groups. For this, probabilities of being in group Severity 4 was extracted with corresponding values, plotted and linear regression model has been fit. This resulted in linear relationships for each variables (R^2 value ranging from 0.993 to 0.999) with correctly predicted relationships (same as in ordinal regression model)

Temperature – lower temperature values corresponds to higher probability of being in Severity 4 group (R^2 = 0.998)



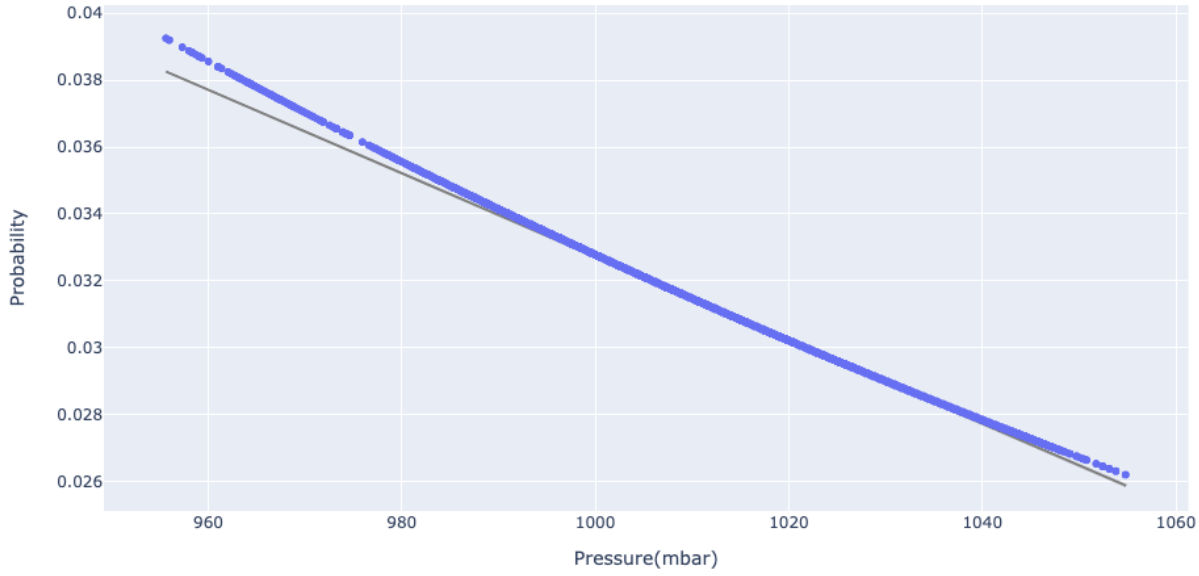

```
Coefficient of determination (R^2 value): 0.9981132169514044
Slope : [-0.00019969]
Intercept : 0.03393778261541452
```

Humidity – larger humidity values corresponds to higher probability of being in Severity 4 group (R^2 = 0.9995)



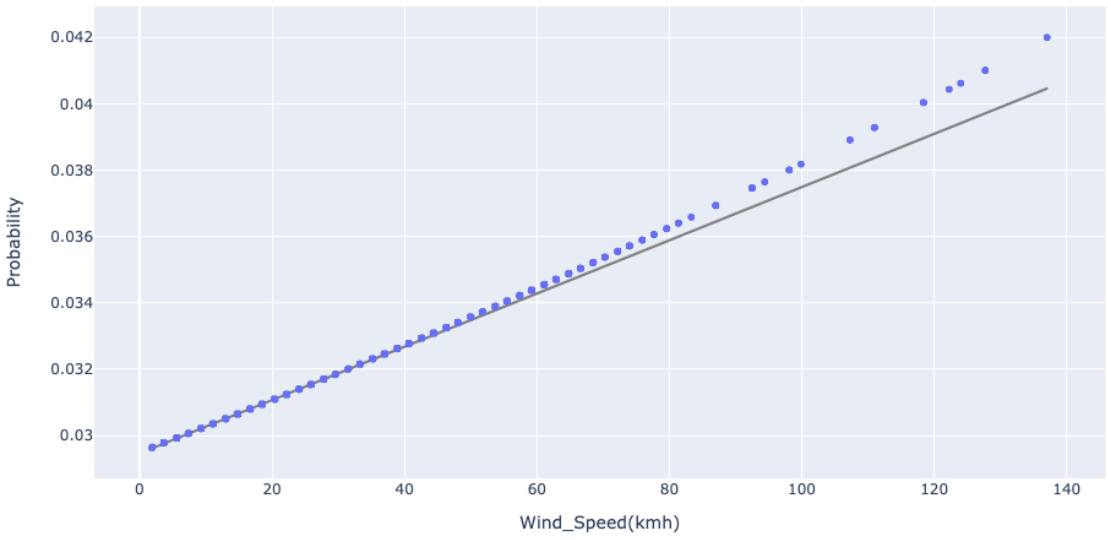

```
Coefficient of determination (R^2 value): 0.9995056882956769
Slope : [5.81609431e-05]
Intercept : 0.02684421635167199
```

Pressure – lower pressure values corresponds to higher probability of being in Severity 4 group (R^2 = 0.9994)



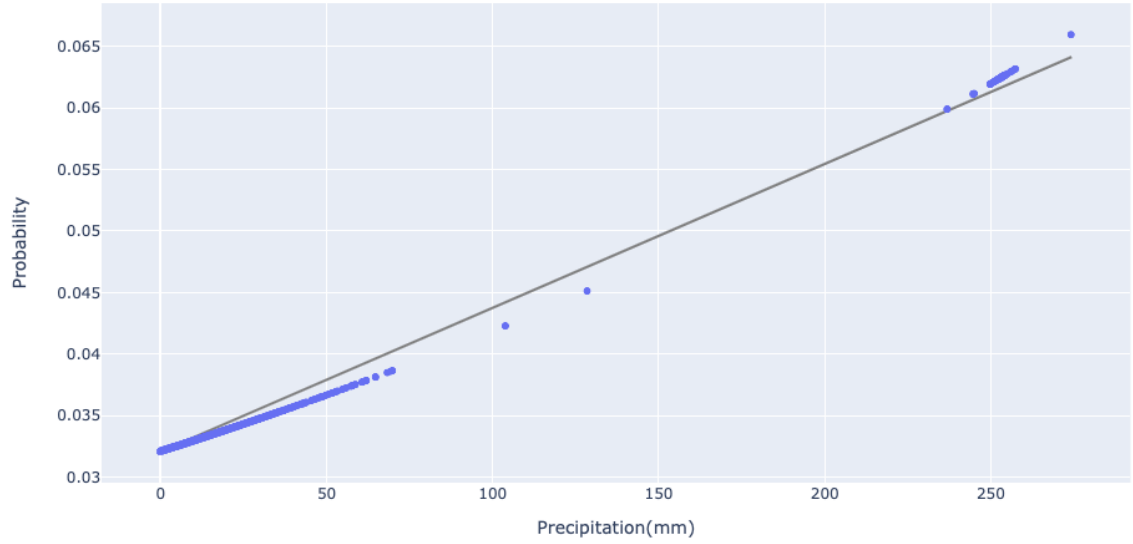

```
Coefficient of determination (R^2 value): 0.9993883523601378
Slope : [-0.00012491]
Intercept : 0.15762343222749434
```

Wind Speed – larger wind speed values corresponds to higher probability of being in Severity 4 group (R^2 = 0.9997)



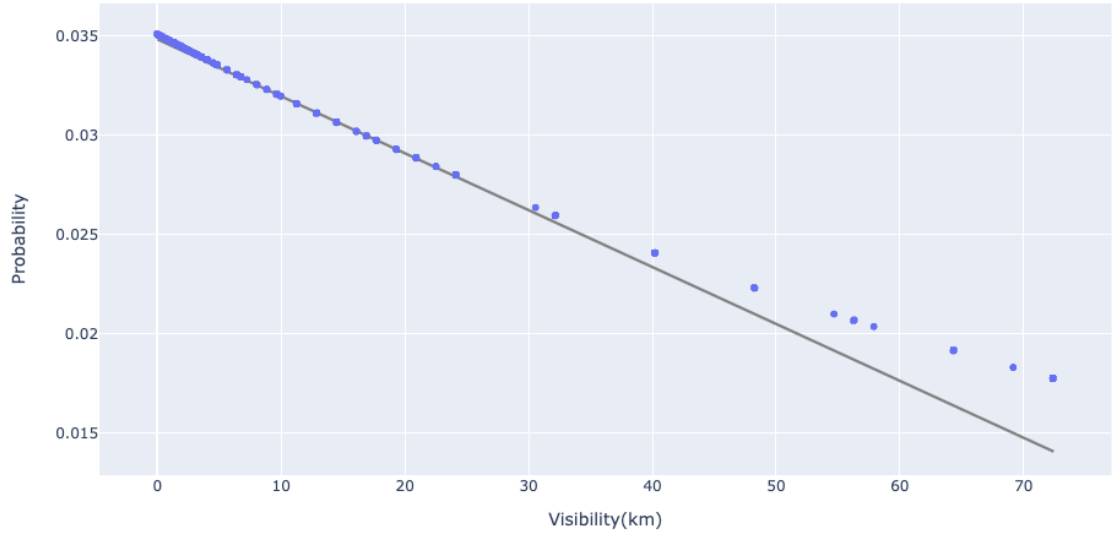

```
Coefficient of determination (R^2 value): 0.99974098665163
Slope : [8.02253211e-05]
Intercept : 0.029466529991104728
```

Precipitation – larger precipitation values corresponds to higher probability of being in Severity 4 group (R^2 = 0.9995) [scarce data between 100 and 230mm could give inaccurate model]




```
Coefficient of determination (R^2 value): 0.9932869187659628
Slope : [0.00011694]
Intercept : 0.03206324610554409
```

Visibility – larger visibility values corresponds to higher probability of being in Severity 4 group (R^2 = 0.9995)



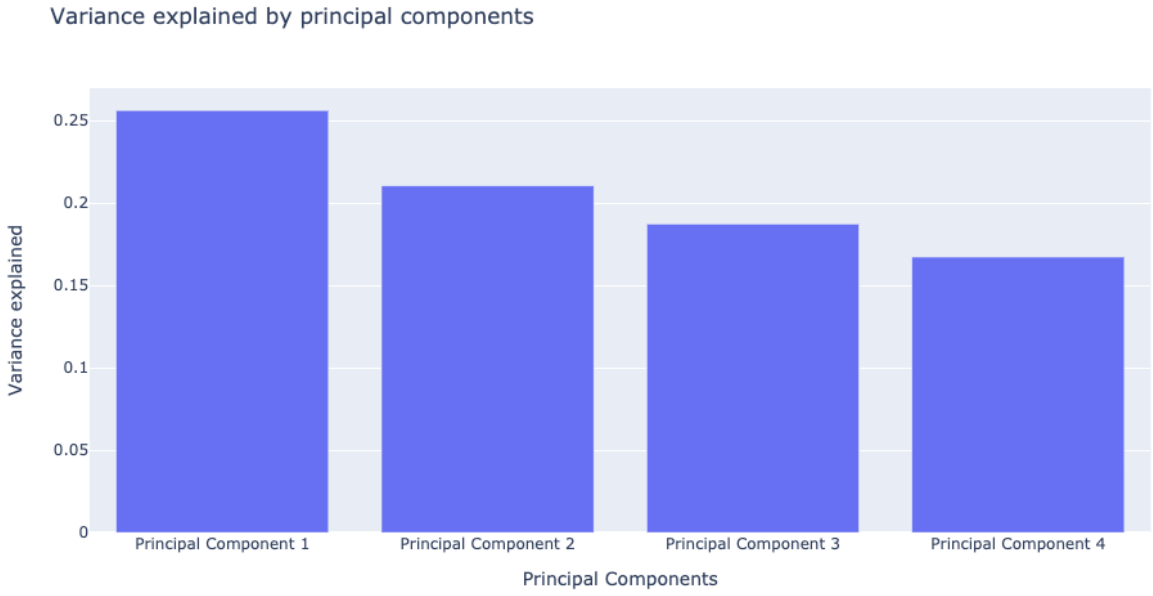

```
Coefficient of determination (R^2 value): 0.9926518565643312
Slope : [-0.00028648]
Intercept : 0.0348209086014382
```

Thus, comparison between variable values and probabilities of being in Severity group 4, indicated the same relationship between Severity groups and weather conditions that was predicted with Ordinal Regression model. It would be beneficial to analyze mentioned relationships throughout all severity groups, nevertheless, for the sake of the size of this report, it is not going to be conducted (please contact me if you think that this analysis is an absolute necessity for this report and I will be glad to add it! 😊)

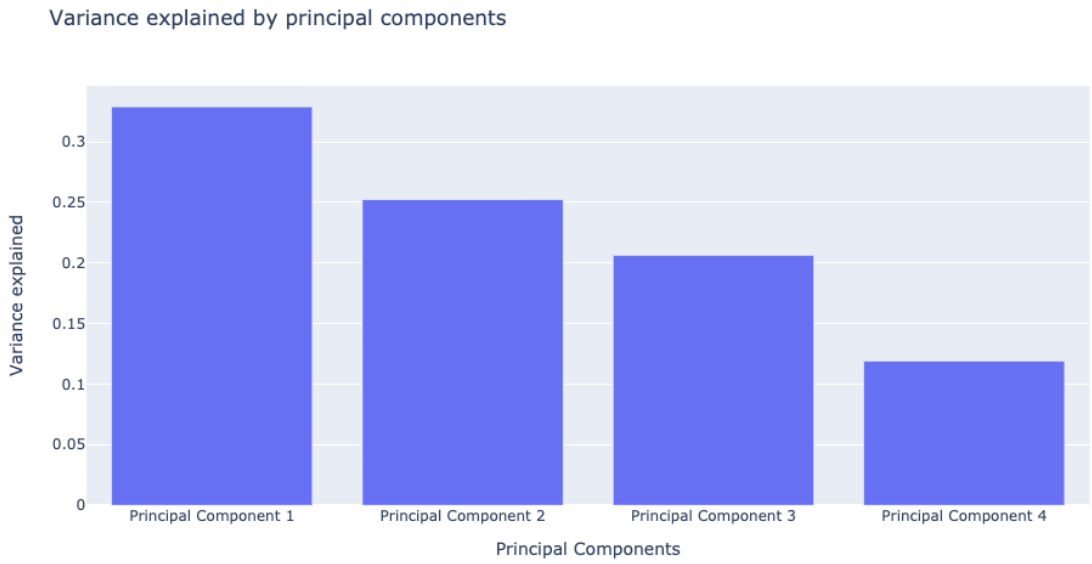
PCA analysis

In order to check, whether all weather conditions combined correlate with higher severity level in accidents, PCA analysis have been conducted.

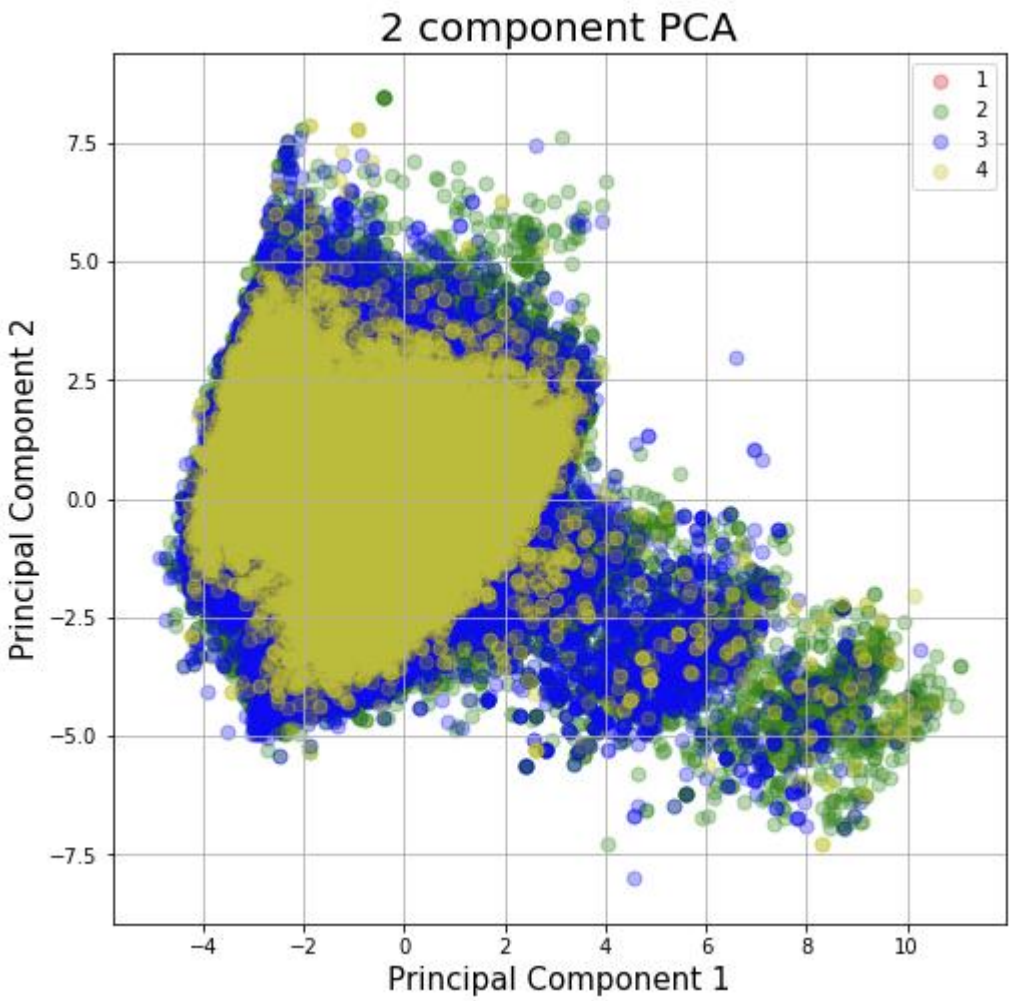
Precipitation(mm) variable only contains ~250k measurements and after PCA with this variable (if all NaN is removed and only values with data for all 6 variables are kept), first two components explain only ~46,6% of variance.



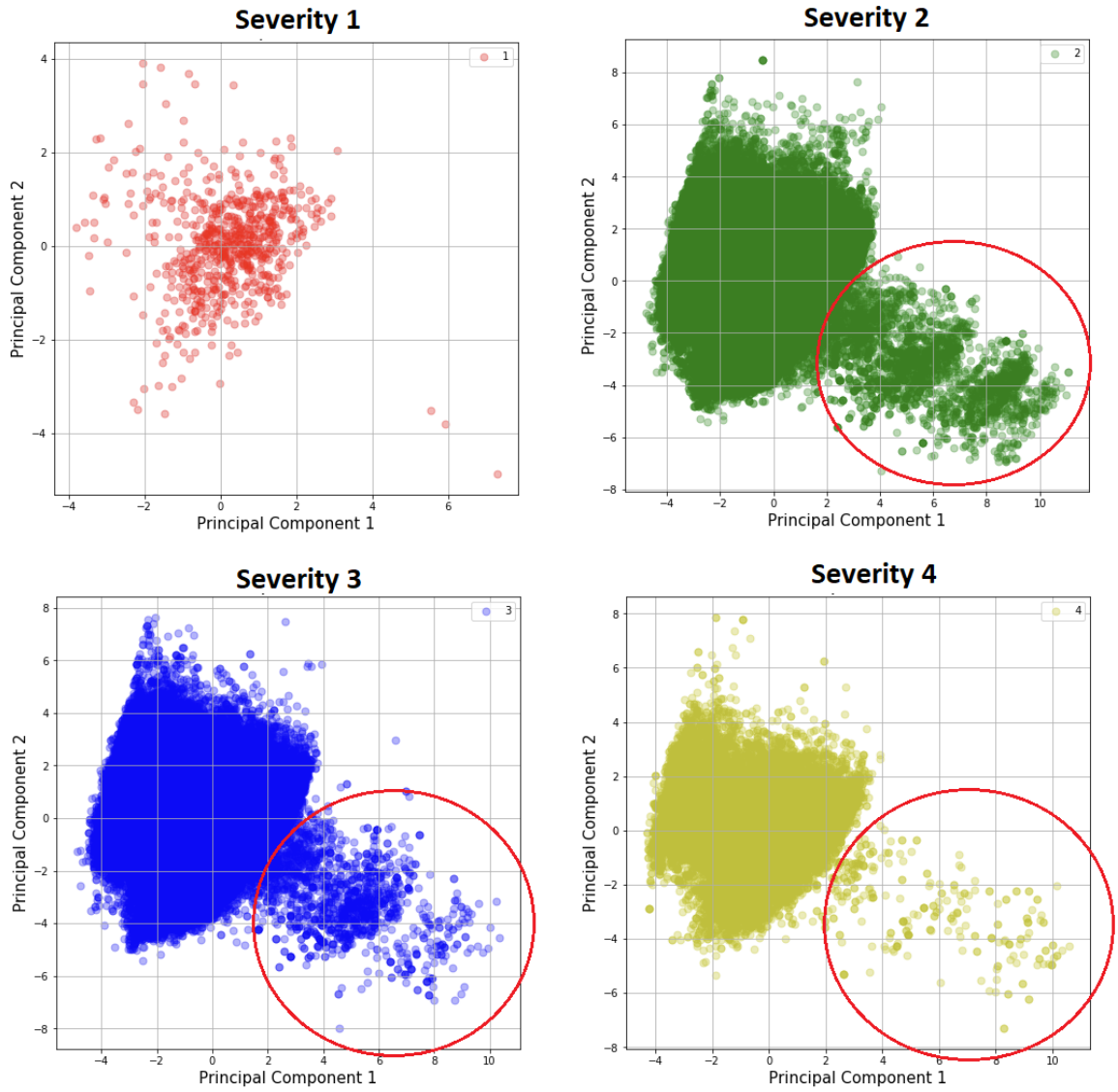
If precipitation variable is removed and PCA done on 5 variables, first two components explain 58%, which is significantly better.



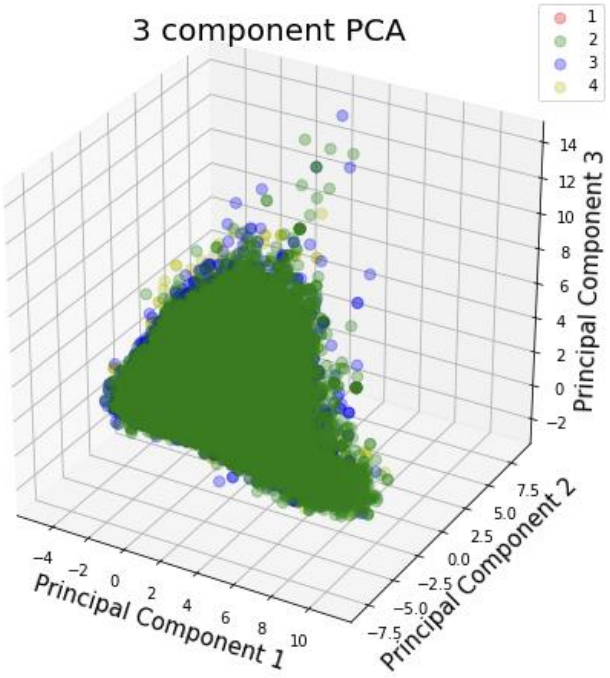
After analyzing first vs second principal component, no apparent separate clusters are visible

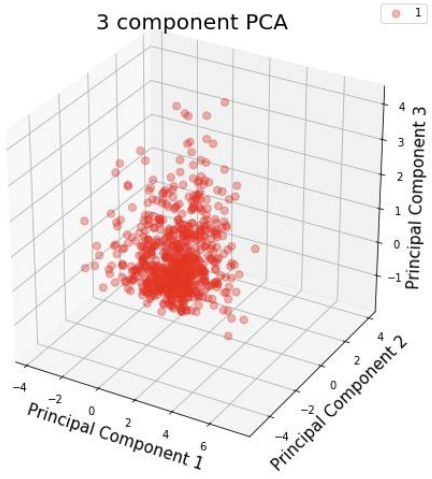


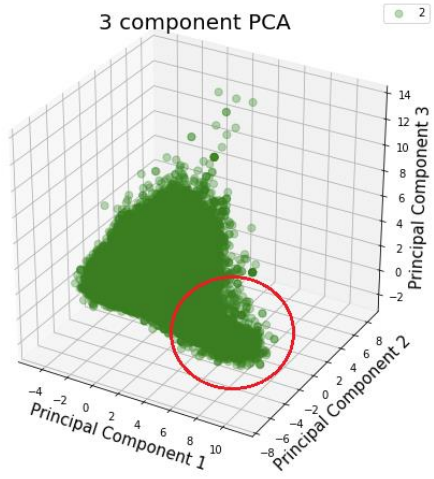
Nevertheless, a difference between Severity 1 and rest of the severity groups can be spotted on plot below (circled in red color). Data (weather conditions) falling under this particular “tail” explains severity which is higher than 1.

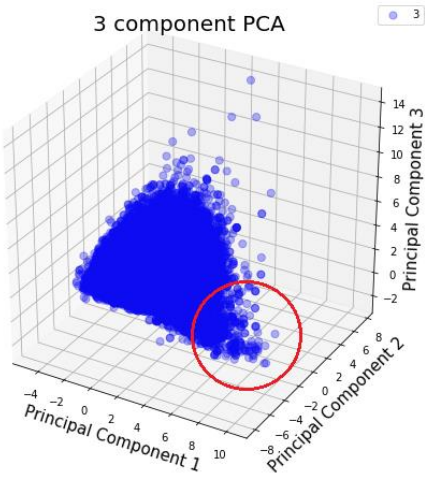


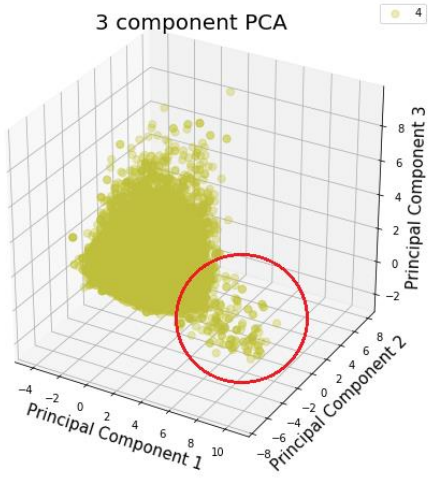
No additional insights can be made after adding third principal component (the same tail, though less apparent, can be distinguished (marked in red))


3 component PCA
1
2
3
4
Principal Component 1
Principal Component 2
Principal Component 3


3 component PCA
1
Principal Component 1
Principal Component 2
Principal Component 3


3 component PCA
2
Principal Component 1
Principal Component 2
Principal Component 3


3 component PCA
3
Principal Component 1
Principal Component 2
Principal Component 3


3 component PCA
4
Principal Component 1
Principal Component 2
Principal Component 3

Summary

The initial aim of the project was to analyze, whether weather conditions affect the severity of traffic accidents.

- Temperature, Humidity and Pressure variables assumed to be distributed normally, while Wind speed, Precipitation and Visibility was distributed non-normally.
- Box plot analysis indicated, that Temperature means and medians are decreasing in higher Severity groups, while Humidity – increasing. No other variables showed any significant changes.
- In Severity groups higher than 1, more outliers were generally seen, indicating, that bigger deviations from mean/median causes higher Severity accidents.
- ANOVA and Kruskal-Wallis tests indicated, that there are significant differences between all compared Severities, nevertheless individual Z-tests showed, that differences usually occur when Severity 2 and higher groups are compared.
- Ordinal regression with additional analysis on probability of being in specific Severity group showed logically correct relationships between Severity and weather conditions (specific relations are listed in Ordinal Regression section)
- Even though no separate clusters were present after PCA analysis, a “tail” of data present in higher groups explains higher Severity than 1.

All in all, it is apparent that Severity of traffic accidents depends on weather conditions. However, in order to determine more precise relationships, more advanced modelling and analysis methods are required.