US traffic incidents 2016 until 2021

https://www.kaggle.com/code/anjakuchenbecker/eda-on-us-traffic-incidents

Handout

Author: Anja Kuchenbecker in 2022

OVERVIEW OF ANALYSIS AREAS

Overall, seven different analysis areas are identified with in total twenty-eight investigated questions. The following handout depicts each analysis area in detail with their related questions and results:

ANALYSIS AREAS

 Impact Analysis	 Location Analysis	 Time Analysis	 Trend Analysis	 Environment Analysis	 Weather Analysis	 Relationship Analysis
Provides insights about the effects of traffic incidents.	Provides insights about the geographical distribution of traffic incidents.	Provides insights about time-related aspects of traffic incidents.	Provides insights about the development of traffic incidents over an interval of time.	Provides insights about the environment around traffic incidents.	Provides insights about the weather conditions with respect to traffic incidents.	Provides insights about possible relationships between features.

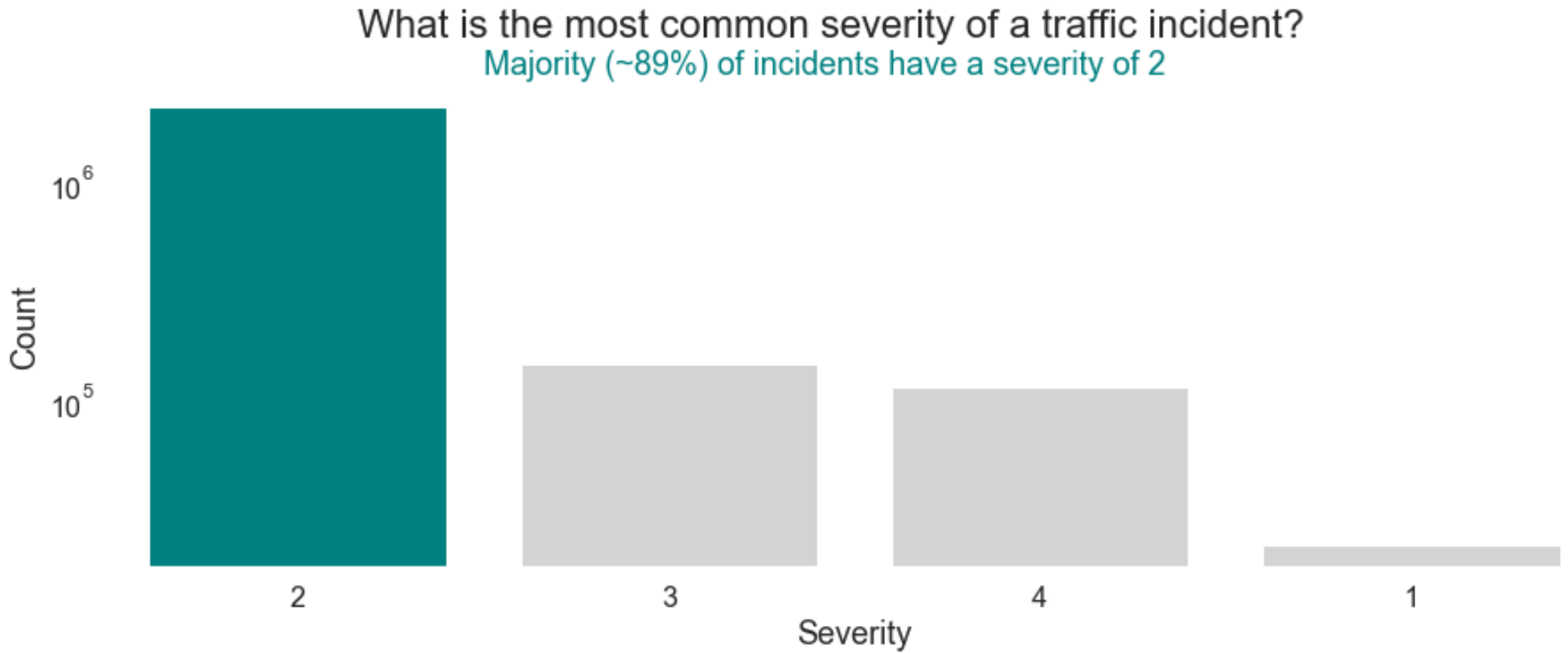
For all details, please refer to the related Jupyter Notebook:
https://www.kaggle.com/code/anjakuchenbecker/eda-on-us-traffic-incidents

Impact Analysis

Provides insights about the effects of traffic incidents.

The following questions will be investigated

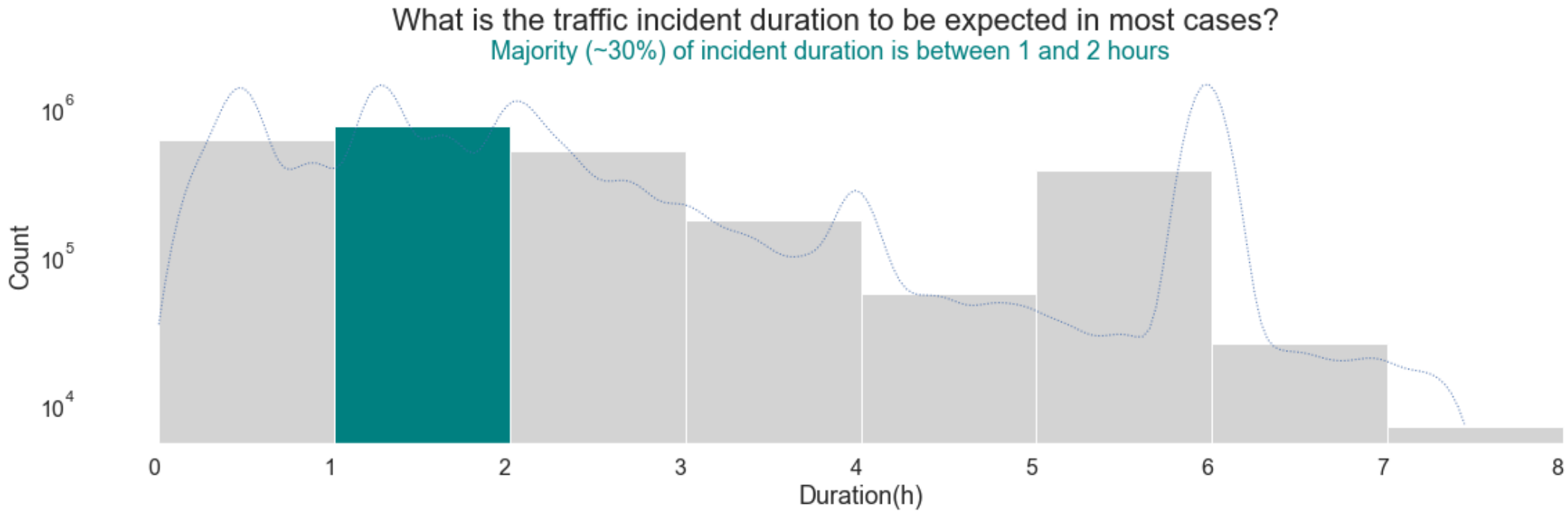
1) What is the most common severity of a traffic incident?
2) What is the traffic incident duration to be expected in most cases?
3) Does the distribution of the traffic incident duration differ with the severity of the incident?
4) Which street length is affected in most cases?
5) Does the distribution of affected road length differ with the severity of the incident?



Result

The typical severity is 2 in case of about 89% of all traffic incidents. Few incidents have a severity of 3 or 4 with about 10%. Very untypical is a severity of 1 with only 1% of all incidents.

IMPACT ANALYSIS – Question 2



Result

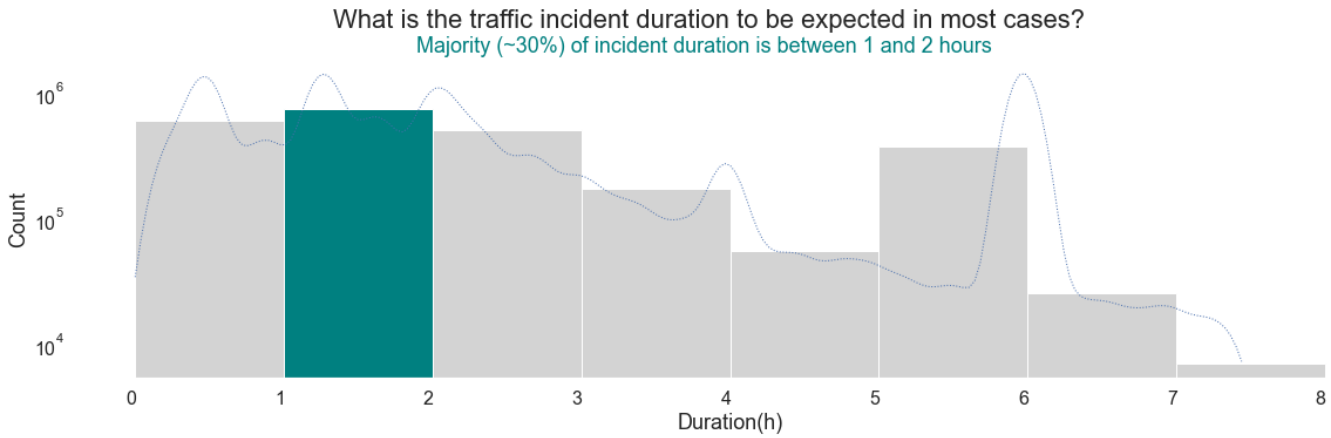
See next slide.

IMPACT ANALYSIS – Question 2 (Result)

Result

Shape

The distribution of the traffic incident duration is moderately skewed right and multimodal. Most accidents cause a lower and fewer a higher traffic incident duration. There are several peaks, the two highest traffic incident durations are concentrated around 1.25 and around 6 hours.



Central Tendency (Location Metrics)

- Mean is ~2.4, but less accurate (due to skewness)
- Median is 1.9 and more accurate

The median duration is 1.9 hours, indicating that the typical traffic incident duration is about 1.9 hours.

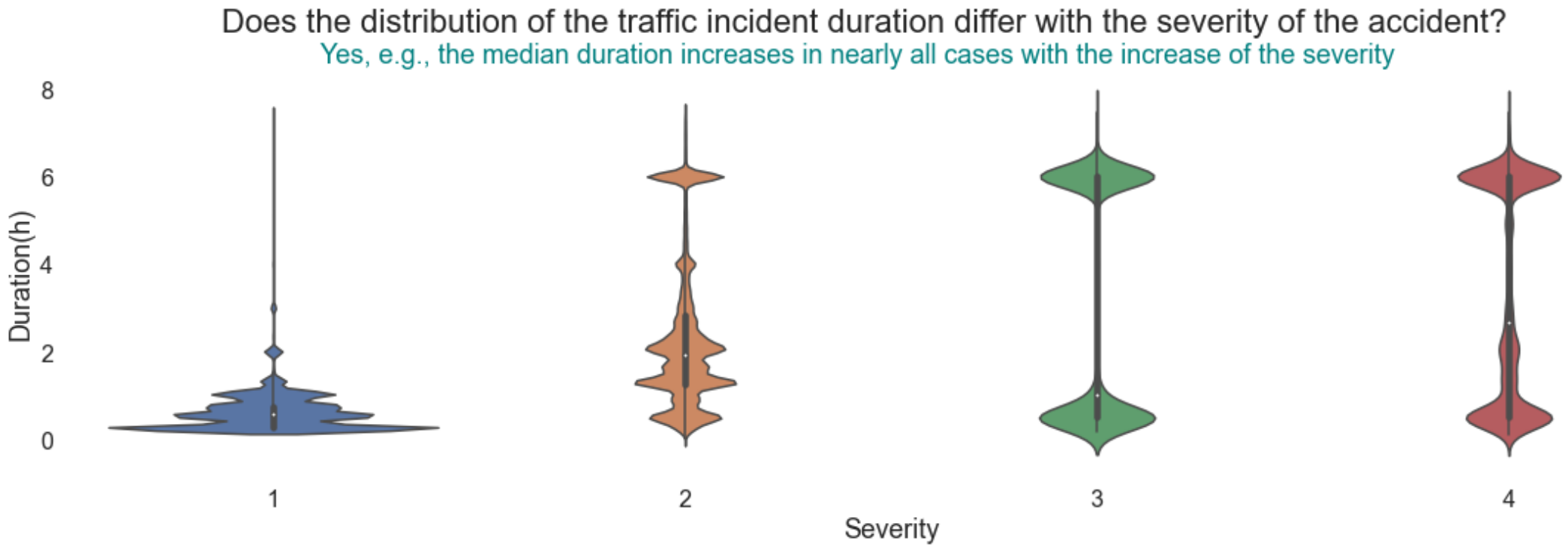
Spread of Data (Variability Metrics)

- Standard Deviation is ~1.85, but less accurate (due to skewness)
- IQR is 1.9 and more accurate

The distribution indicates with an IRQ of 1.9 hours that there is a variation in the traffic incident duration. It ranges from 0.03 to about 7.5 hours. Most (~75%) of the accidents have a traffic incident duration between 0.03 and about 2 hours, fewer (~25%) up to 7.5 hours, suggesting the large differences.

Outliers

There seem to be about 15% outliers according to the 1.5-IQR rule to the far right with a traffic incident duration higher than about 6 hours.



Result

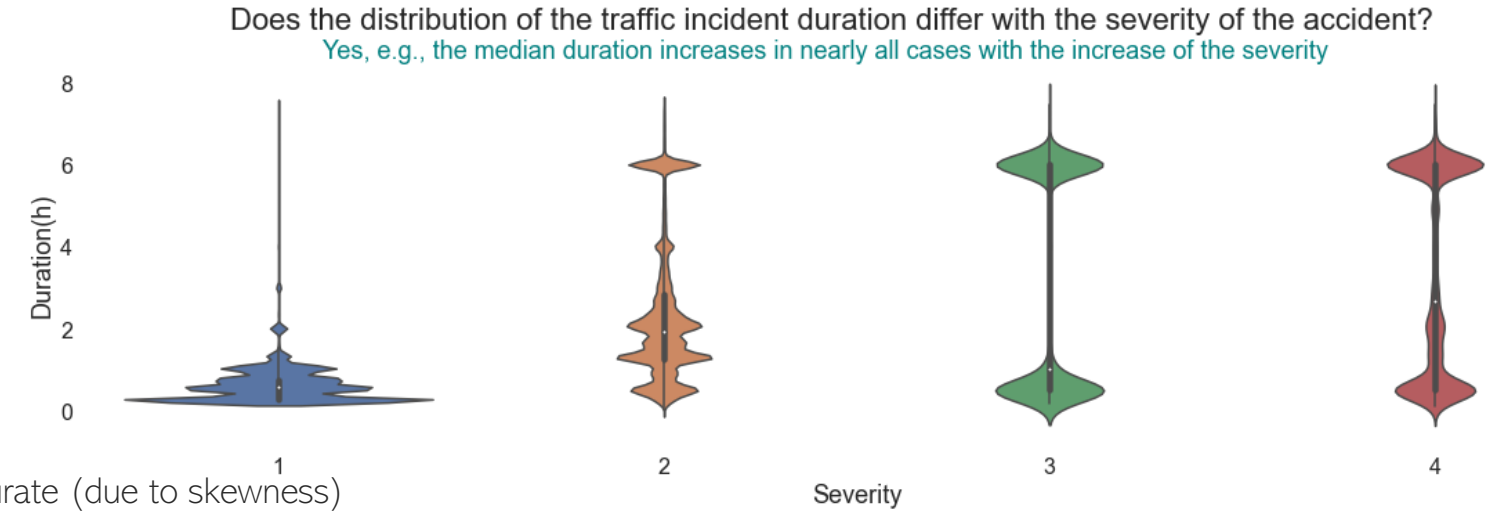
See next slide.

IMPACT ANALYSIS – Question 3 (Result)

Result

Shape

In case of severity 1 and 2 the distribution of the traffic incident duration is highly skewed right and multimodal.
For severity 3 and 4 the distribution is approximately symmetric and bimodal for the first and multimodal for the latter.



Central Tendency (Location Metrics)

- Mean is ~6.3 (1) / ~2.3 (2) / ~3 (3)/ ~3.3 (4) but less accurate (due to skewness)
- Median is 0.58 (1) / 1.92 (2) / 1.0 (3) / 2.65 (4) and more accurate

The median duration increases in nearly all cases with the increase of the severity. The typical traffic incident duration is about 0.58, 1.92, 1.0 and 2.65 hours regarding severity 1 to 4.

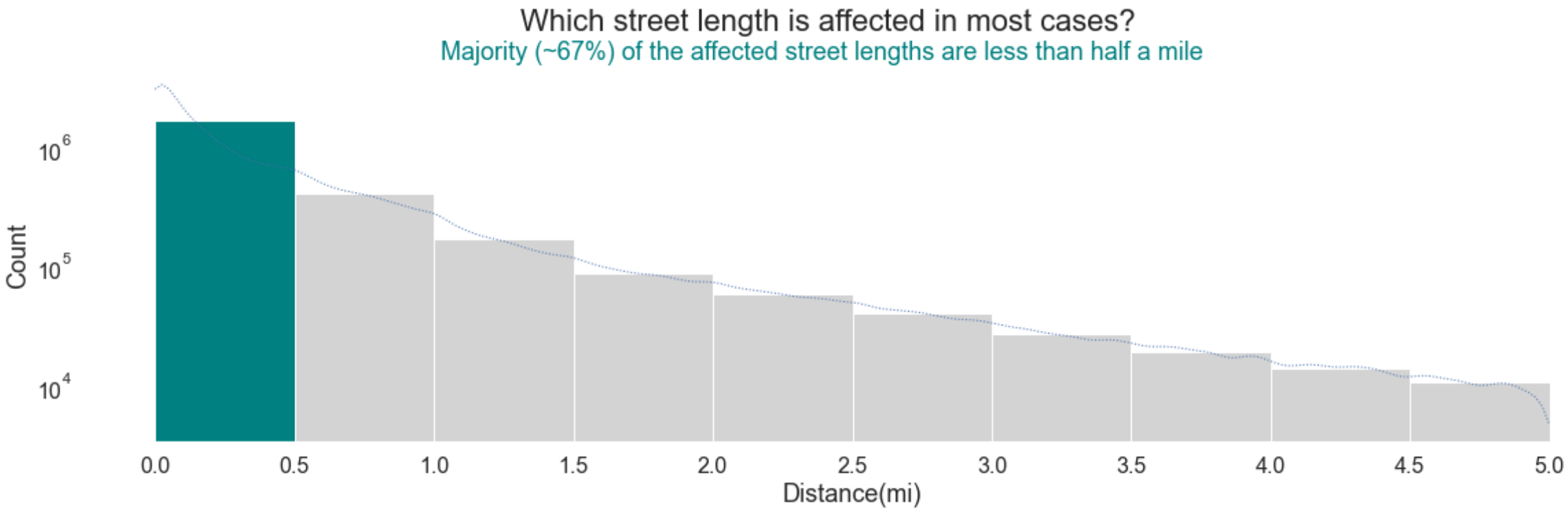
Spread of Data (Variability Metrics)

- Standard Deviation is ~0.5 (1) / ~1.7 (2) / ~2.7 (3) / ~2.5 (4), but less accurate (due to skewness)
- IQR is 0.5 (1) / ~1.6 (2) / ~5.5 (3) / 5.5 (4) and more accurate

The variation in the traffic incident duration increases in nearly all cases with the increase of the severity. It ranges from 0.25, 0.03, 0.15 and 0.11 regarding severity 1 to 4 to about 7.5 hours for all severities.

Outliers

There seem to be about 3% and 13% outliers according to the 1.5-IQR rule to the far right with a traffic incident duration higher than about 1.5 and 5.2 hours regarding severity 1 and 2.

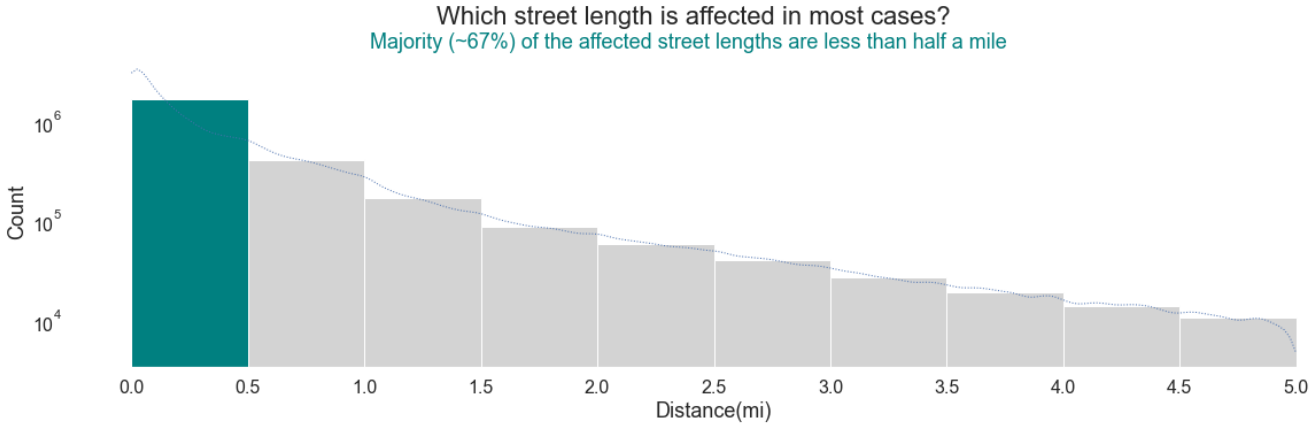


Result

See next slide.

IMPACT ANALYSIS – Question 4 (Result)

Result



Shape

The distribution of affected street length is highly skewed right and unimodal.
Most accidents cause a lower and fewer a higher affected street length.
There is one peek, concentrated around 0.03 miles (~48 meters).

Central Tendency (Location Metrics)

- Mean is ~0.6, but less accurate (due to skewness)
- Median is ~0.2 and more accurate

The median affected street length is about 0.2, indicating that the typical affected street length is about 0.2 miles (~321 meters).

Spread of Data (Variability Metrics)

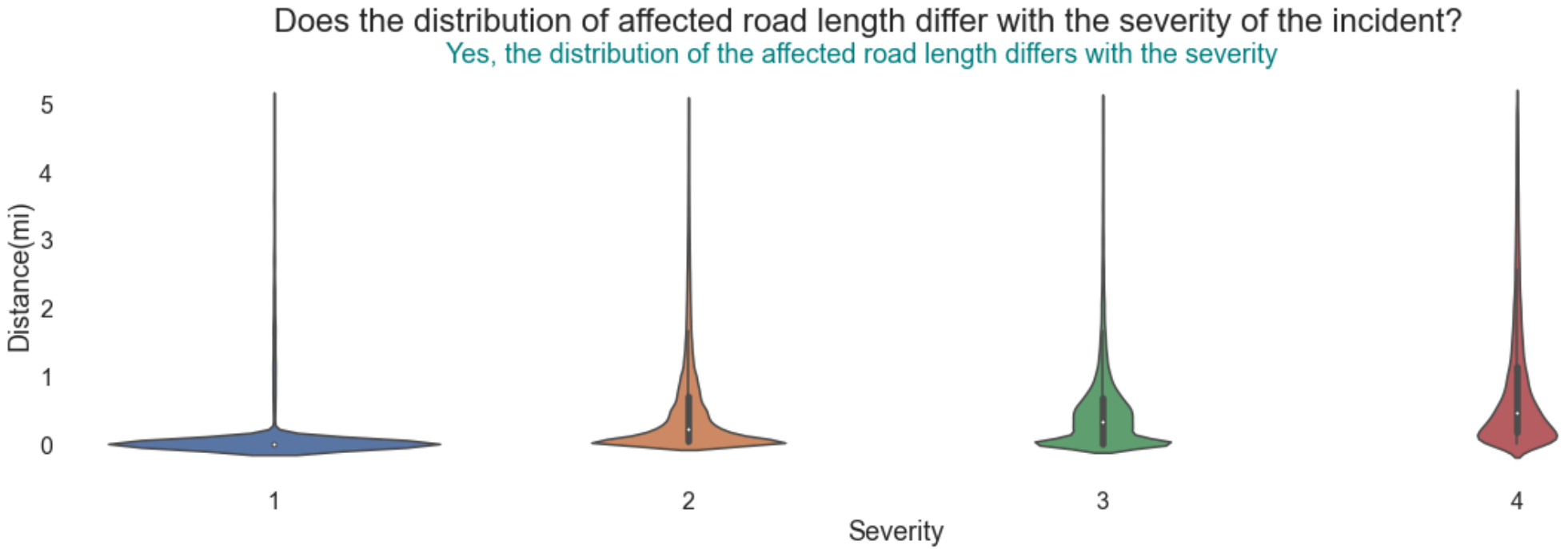
- Standard Deviation is ~0.8, but less accurate (due to skewness)
- IQR is ~0.7 and more accurate

The distribution indicates with an IRQ of about 0.7 miles (~1127 meters) that there is a variation in the affected street length. It ranges from 0.0 to about 5 miles (~8047 meters). Most (~67%) of the accidents have an affected street length between 0.0 and less than 0.5 miles (~805 meters), fewer (~33%) up to about 5 miles (~8047 meters), suggesting the large differences.

Outliers

There seem to be about 9% outliers according to the 1.5-IQR rule to the far right with an affected street length higher than about 1.7 miles (~2736 meters).

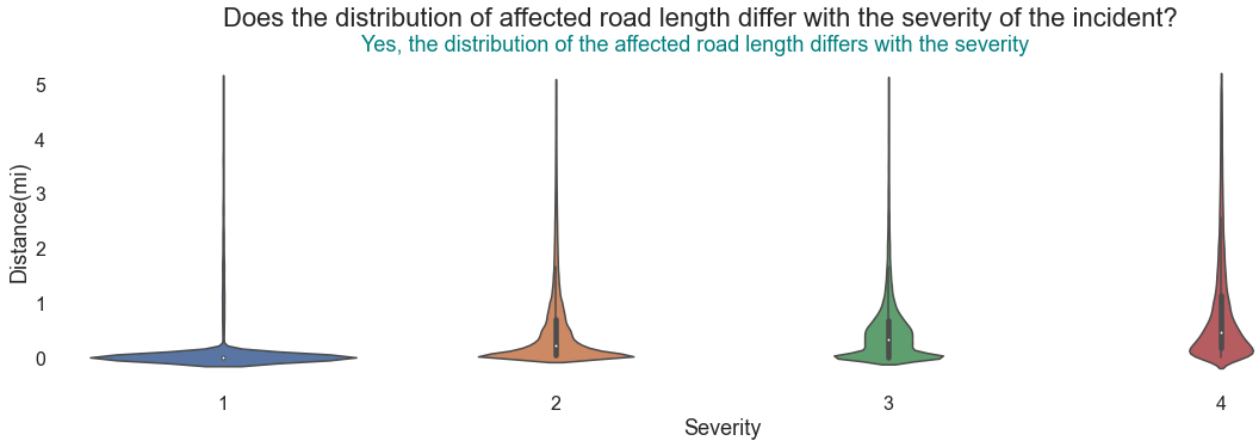
IMPACT ANALYSIS – Question 5



Result

See next slide.

IMPACT ANALYSIS – Question 5 (Result)



Result

Shape

The affected street length is highly skewed right and unimodal for all severities.

Central Tendency (Location Metrics)

- Mean is ~0.2 (1) / ~0.5 (2) / ~0.5 (3)/ ~0.8 (4) but less accurate (due to skewness)
- Median is ~0 (1) / ~0.2 (2) / ~0.3 (3) / ~0.5 (4) and more accurate

The median duration increases in all cases with the increase of the severity. The typical affected street length is about 0, 0.2, 0.3 and 0.5 miles regarding severity 1 to 4.

Spread of Data (Variability Metrics)

- Standard Deviation is ~0.6 (1) / ~0.8 (2) / ~0.7 (3) / ~1 (4), but less accurate (due to skewness)
- IQR is 0 (1) / ~0.6 (2) / ~0.7 (3) / ~1 (4) and more accurate

The variation in the traffic incident duration increases in all cases with the increase of the severity. It ranges from 0 to about 5 miles for all severities.

Outliers

There seem to be about 11%, 9%, 6% and 8% outliers according to the 1.5-IQR rule to the far right with an affected street length higher than about 0, 1.7, 1.7 and 2.6 miles regarding severity 1 to 4.

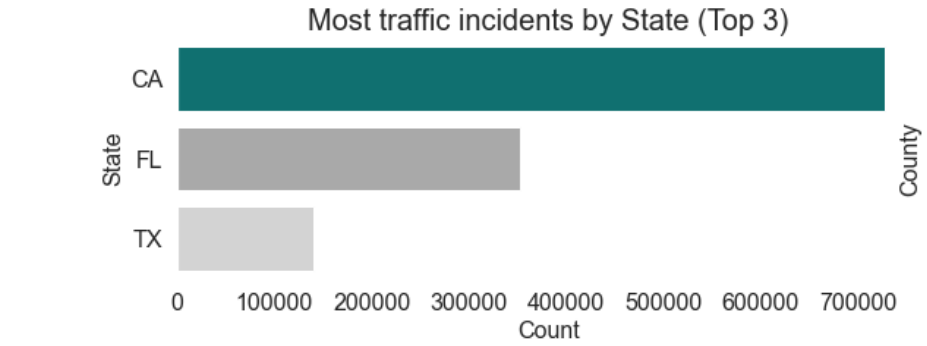
Location Analysis

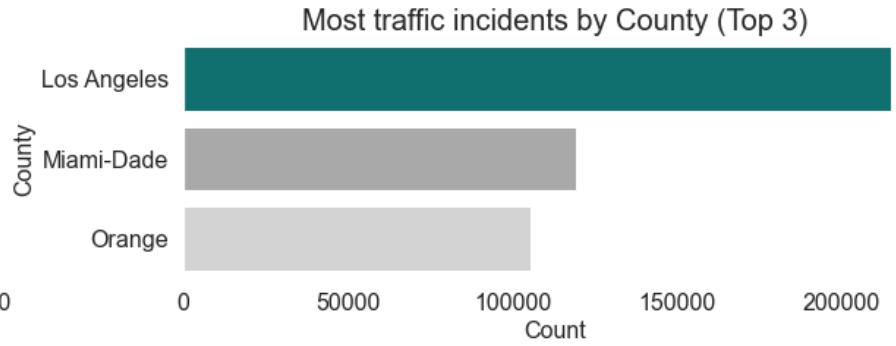
Provides insights about the geographical distribution of traffic incidents.

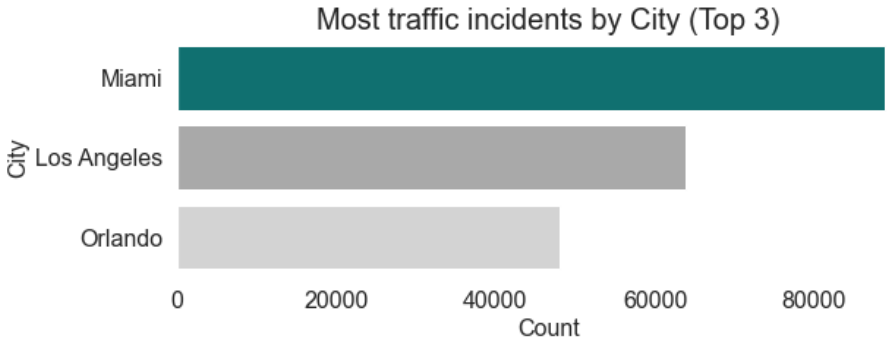
The following questions will be investigated

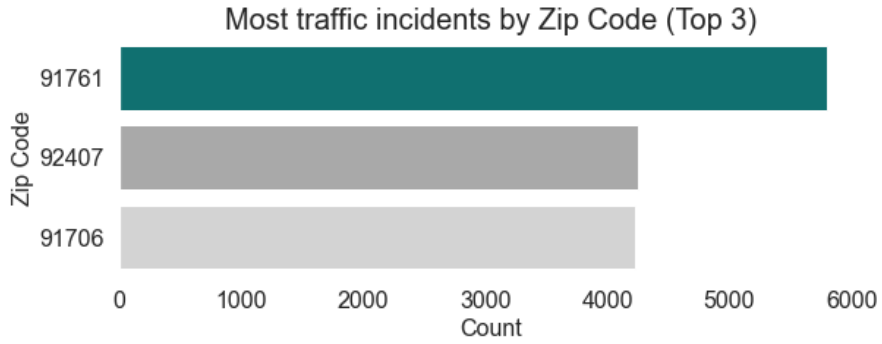
1) Which places have the most traffic incidents (Top 3)?
2) Which states have the least traffic incidents (Top 3)?
3) What street type is affected in most cases (Top 5)?
4) How are traffic incidents geographically distributed regarding their severity?

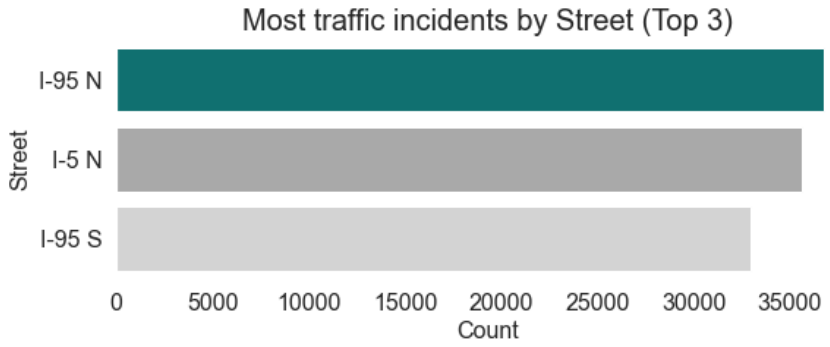
LOCATION ANALYSIS – Question 1











Result

See next slide.

LOCATION ANALYSIS – Question 1 (Result)

Result

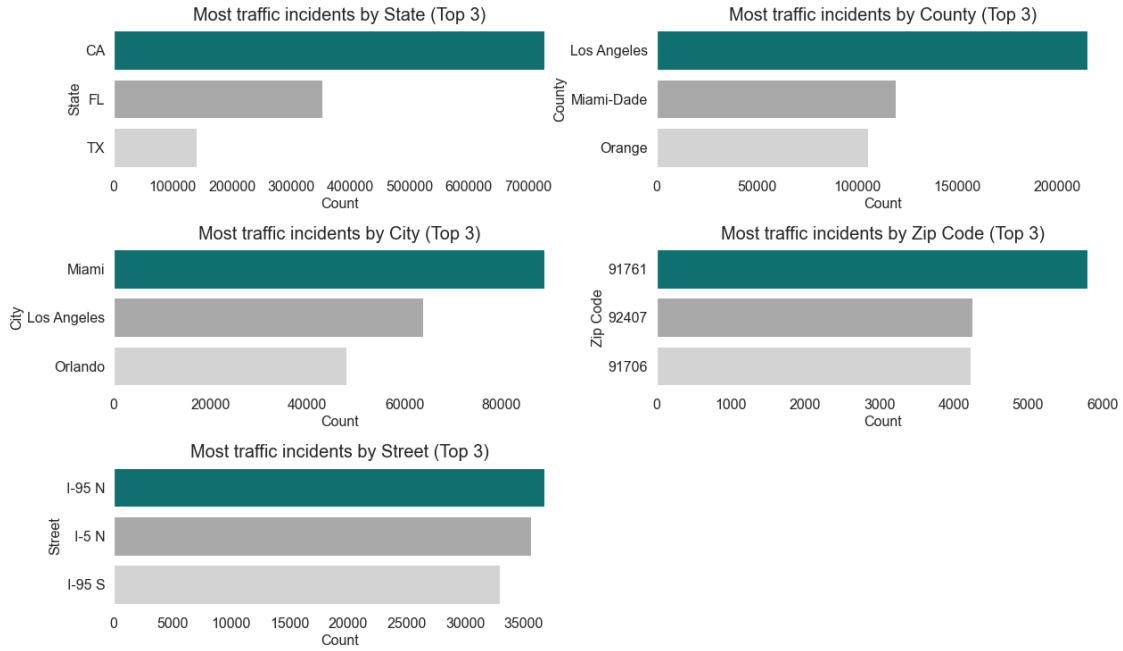
State
Most traffic incidents regarding states occur in *California* with about 30%, followed by *Los Angeles* and *Orlando* with about 14% and 6% respectively.

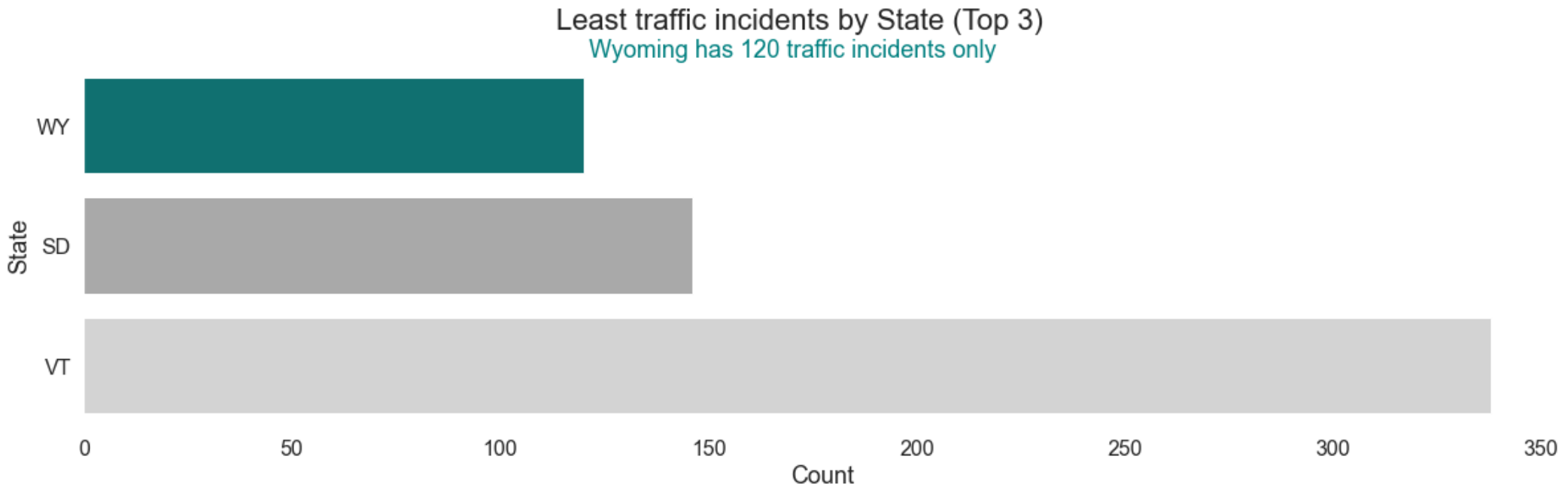
County
Most traffic incidents regarding counties occur in *Los Angeles* with about 9%, followed by *Miami-Dade* and *Orange* with about 5% and 4% respectively.

City
Most traffic incidents regarding cities occur in *Miami* with about 4%, followed by *Los Angeles* and *Orlando* with about 3% and 2% respectively.

Zip Code
Most traffic incidents regarding zip codes occur in *91761 (Ontario, California)* with about 0.2%, followed by *92407 (San Bernardino, California)* and *91706 (Baldwin Park, California)* with about 3 % and 2 % respectively.

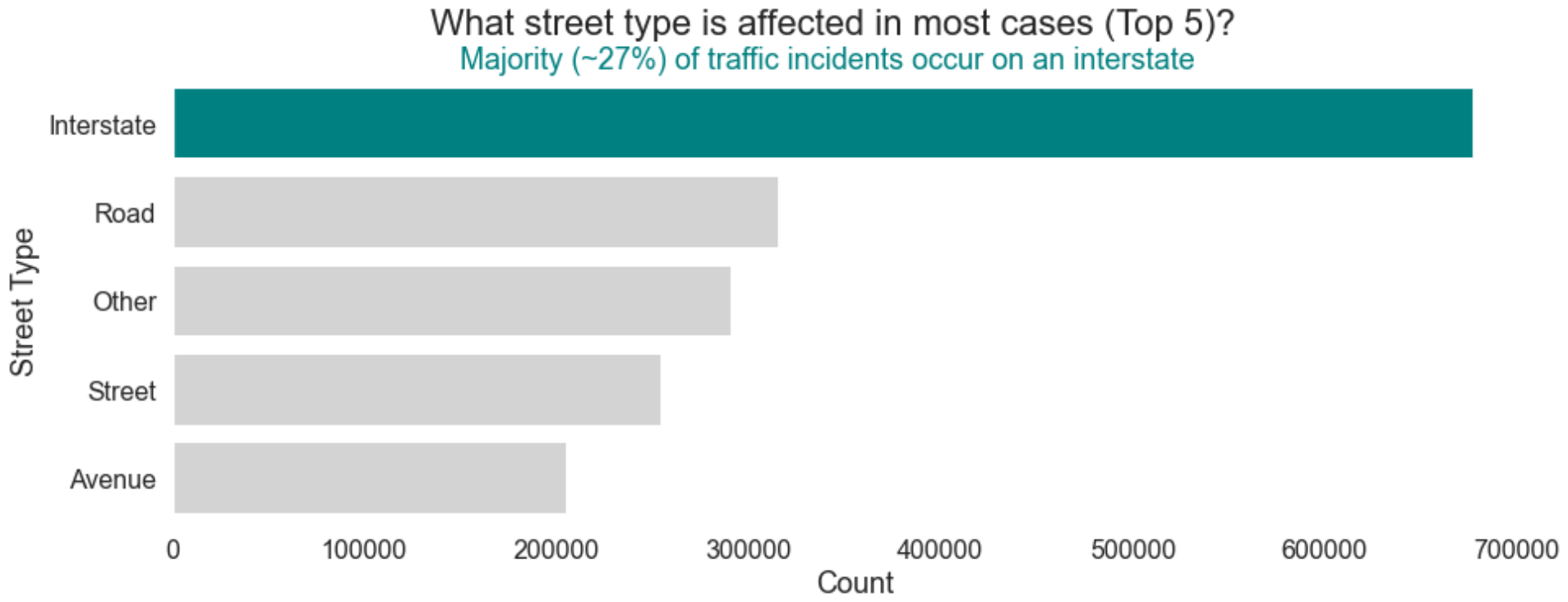
Street
Most traffic incidents regarding streets occur on the *I-95 N* with about 2%, followed by *I-5 N* and *I-95 S* with about 1% and 1% respectively.





Result

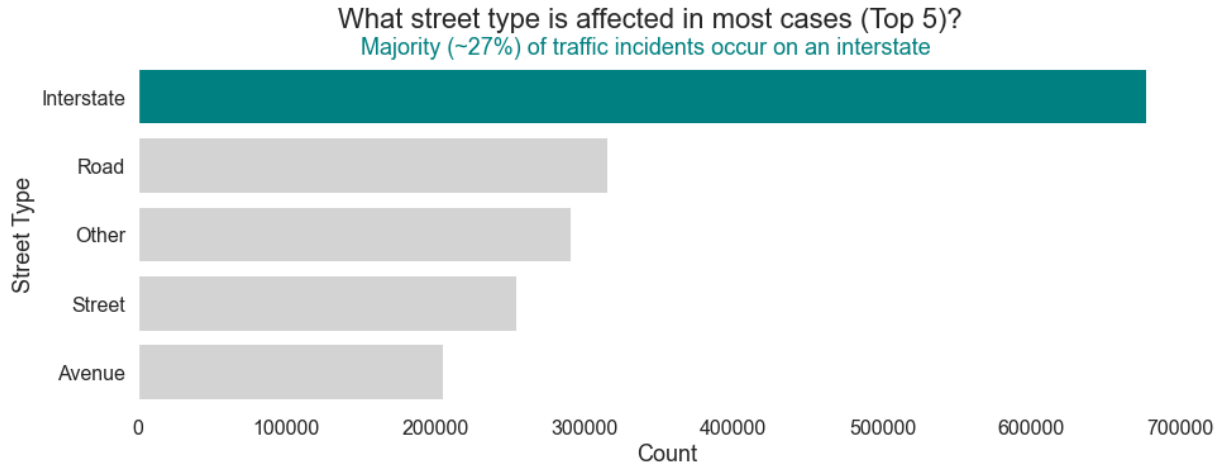
With 120 in total the state *Wyoming* has the least traffic incidents, followed by *South Dakota* and *Vermont* with 146 and 338 incidents, respectively.



Result

See next slide.

LOCATION ANALYSIS – Question 3 (Result)



Result

Background Knowledge

- *Interstate:* A large, typically federally funded network of roads states, but it doesn't have to.
- *Road:* The most general category, this is a way that connects two points.
- *Street:* A public way with buildings on both sides. Often, it runs perpendicular to an avenue.
- *Avenue:* A public way often in a city, usually with trees or buildings on the side. Frequently, it runs perpendicular to a street.

Result interpretation

The typical traffic incident occurs on interstates in case of about 27% of all traffic incidents with respect to the top 5 street types. The rest occurs on street type *road*, *others*, *street* and *avenues* in descending order.

LOCATION ANALYSIS – Question 4

How are traffic incidents geographically distributed regarding their severity?

Traffic incidents are differently distributed regarding their severity.

Result

See next slide.

LOCATION ANALYSIS – Question 4 (Result)

Result

Traffic incidents are differently distributed regarding their severity.

Traffic incidents with severity 1 have the lowest counts and occurs mainly in the margin of east and west.

Severity 2 incidents are most common and distributed in north, east, south and west.

Incidents with severity 3 occur secondly common and are more distributed in the margin of east and west, and less in north and south.

Severity 4 traffic incidents occur third common and are mainly concentrated in the east and in the margin of west.

How are traffic incidents geographically distributed regarding their severity?

Traffic incidents are differently distributed regarding their severity.

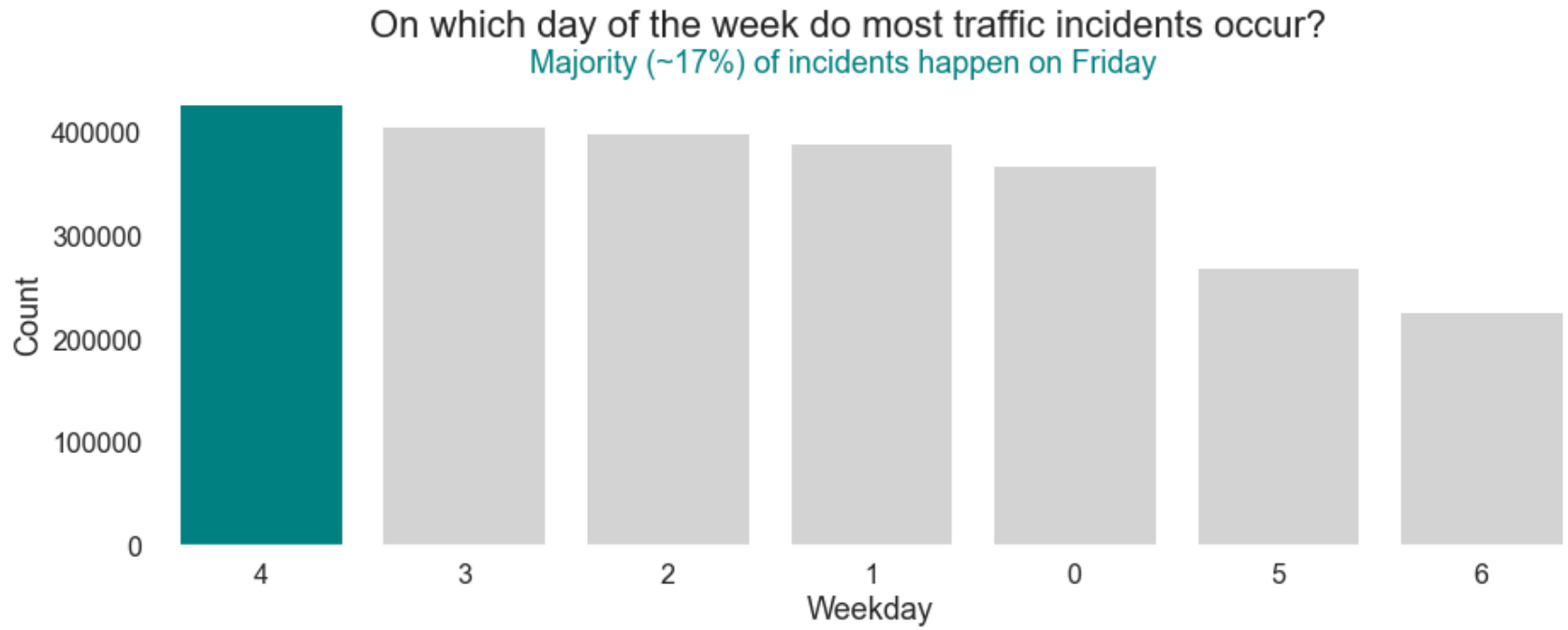
Time Analysis

Provides insights about time-related aspects of traffic incidents.

The following questions will be investigated

1) On which day of the week do most traffic incidents occur?
2) At which hour do most of the traffic incidents occur?
3) On which month do most of the traffic incidents occur?
4) Do traffic incidents occur mostly by day, twilight or night?

TIME ANALYSIS – Question 1

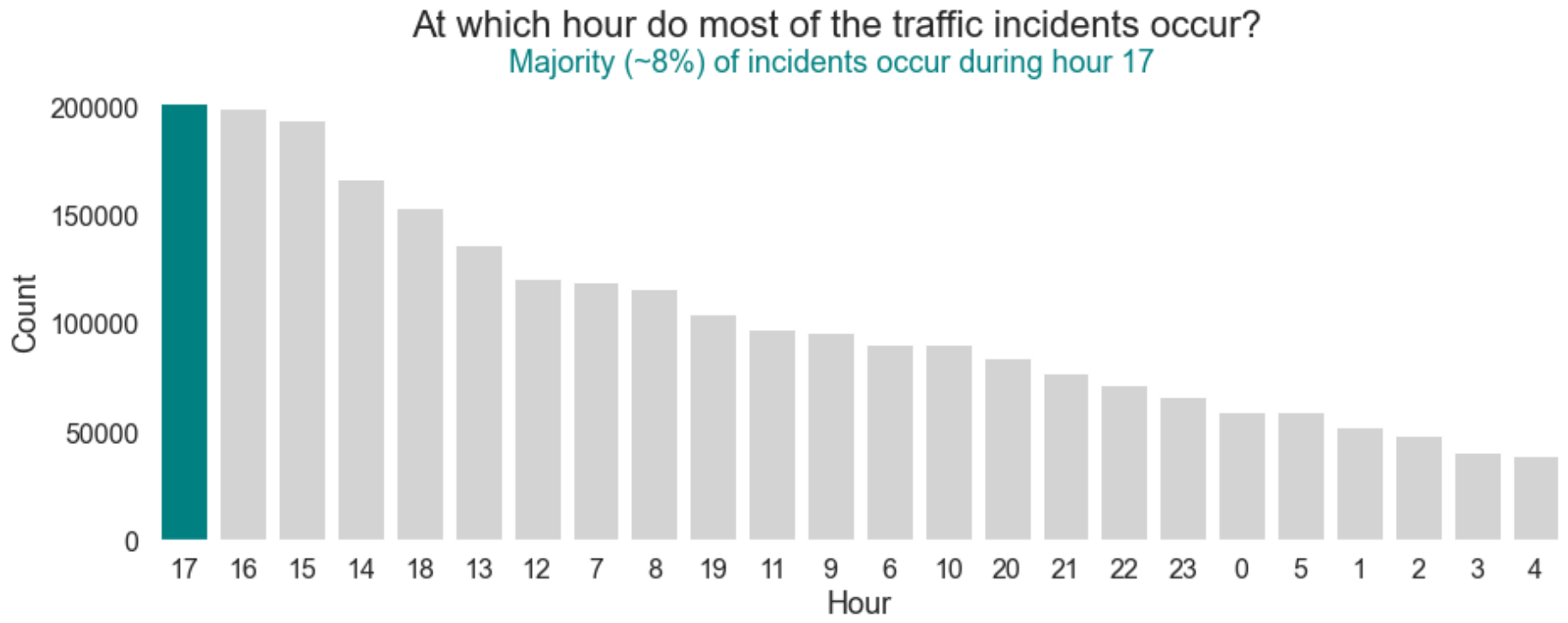


Result

The typical traffic incident occurs on Friday in case of about 17% of all incidents. The rest occurs on Thursday, Wednesday, Tuesday, Monday, Saturday and Sunday in descending order.

Only about 20% of traffic incidents happen during weekend and about 80% during the week.

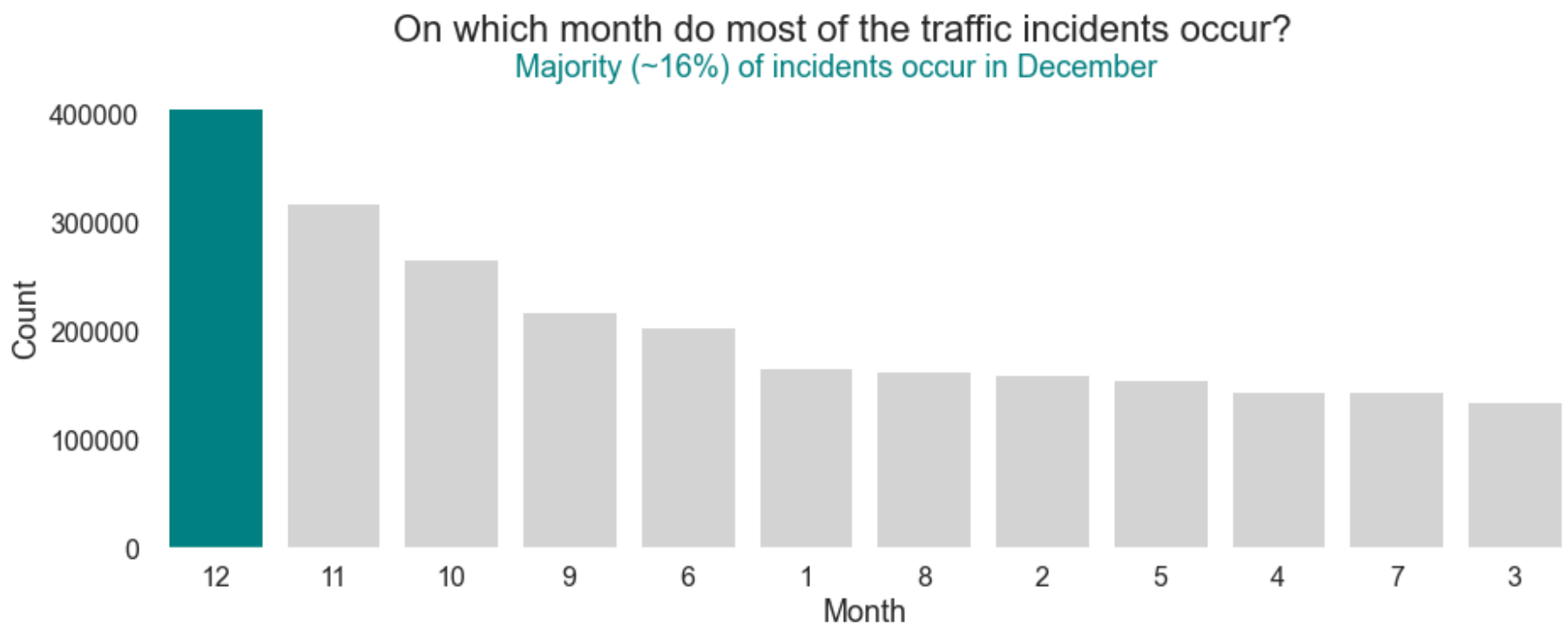
TIME ANALYSIS – Question 2



Result

The typical traffic incident occurs during hour 17 in case of about 8% of all incidents, directly followed by hour 16 and 15 both also with about 8%.

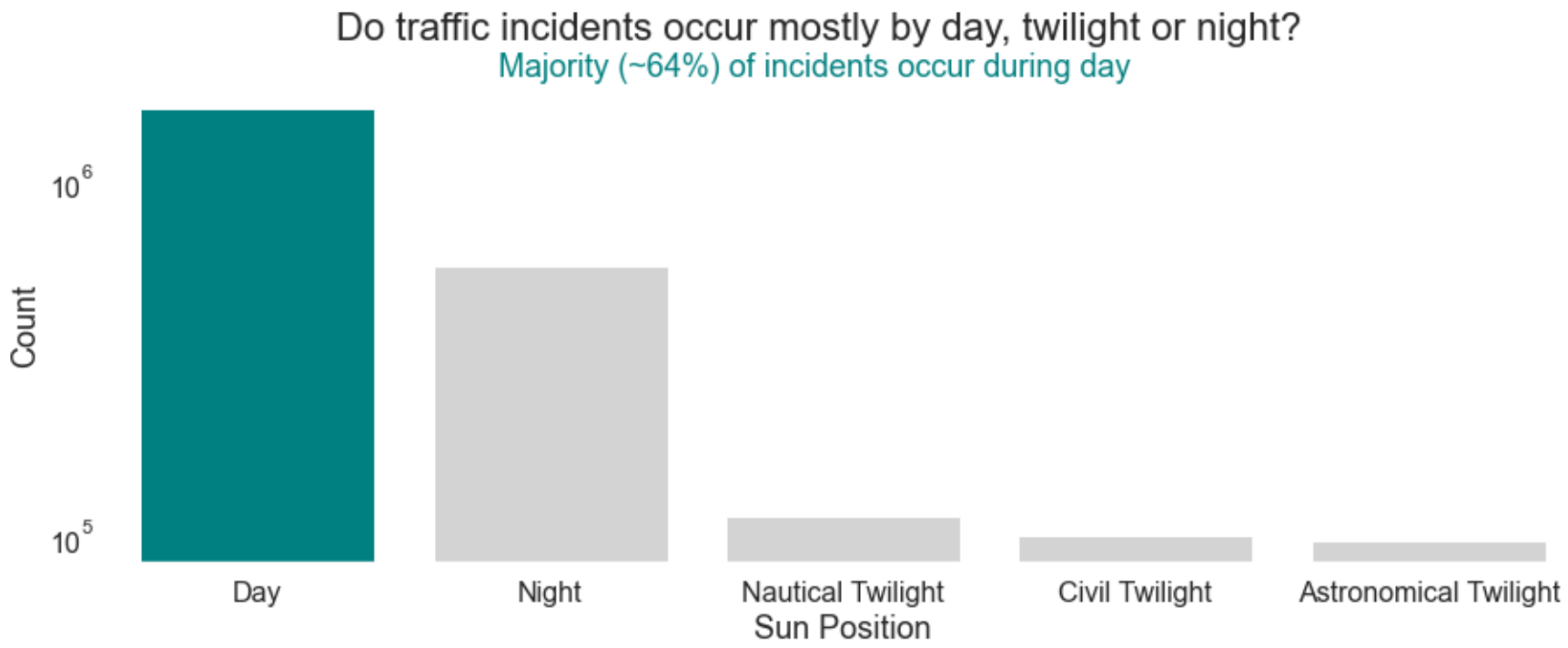
TIME ANALYSIS – Question 3



Result

The typical traffic incident occurs in December in case of about 16% of all incidents, directly followed by November and October with about 13% and 11% respectively.

TIME ANALYSIS – Question 4



Result

The typical traffic incident occurs during day in case of about 64% of all incidents, followed by night with about 23%.

The overall twilight proportion is only about 12%, with about 4% per twilight type.

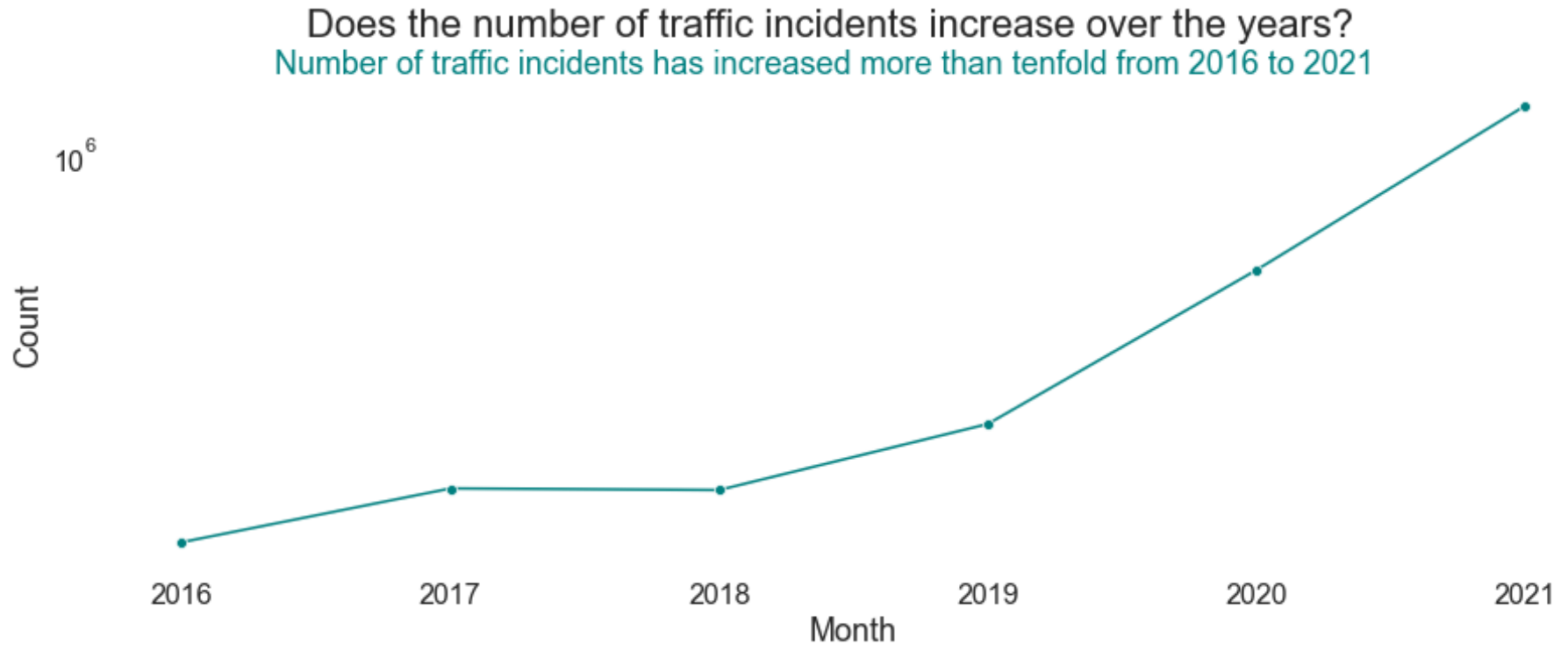
Trend Analysis

Provides insights about the development of traffic incidents over an interval of time.

The following questions will be investigated

1) Does the number of traffic incidents increase over the years?
2) Is there a change of the proportion of the severity levels overs the years?
3) Does the traffic flow impact duration increase over the years?

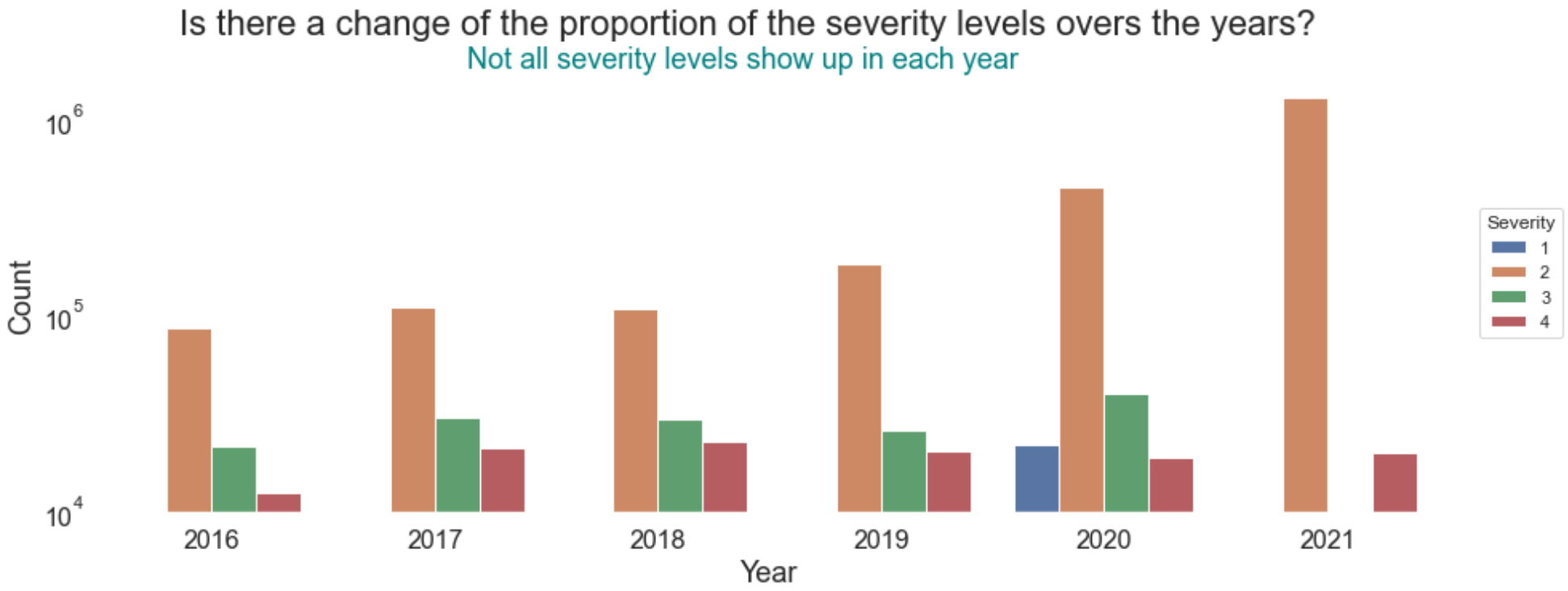
TREND ANALYSIS – Question 1



Result

The number of traffic incidents has increased more than tenfold from 2016 to 2021.

TREND ANALYSIS – Question 2

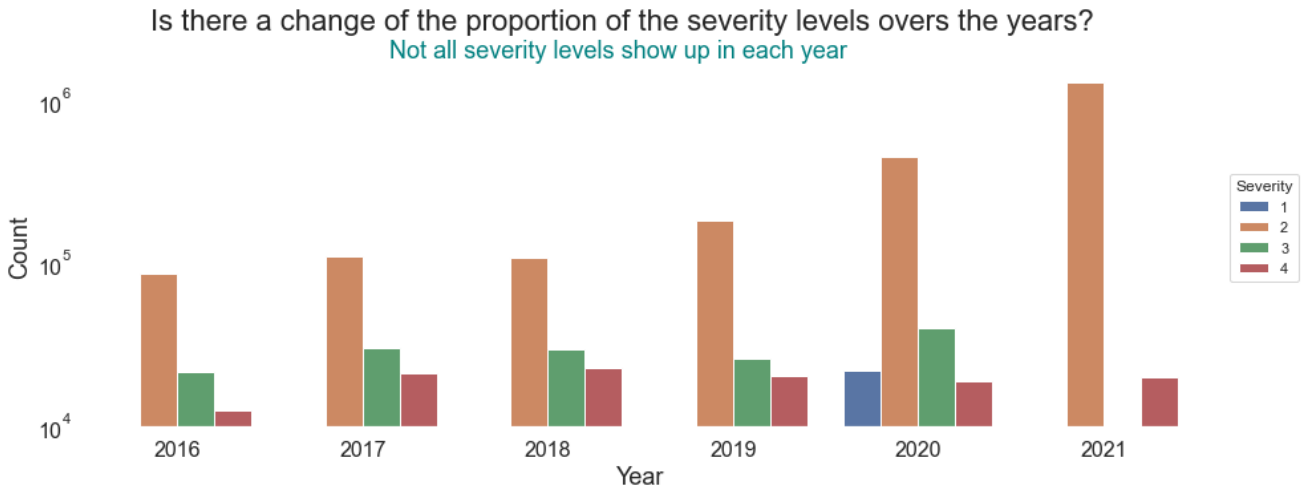


Result

See next slide.

TREND ANALYSIS – Question 2 (Result)

Result



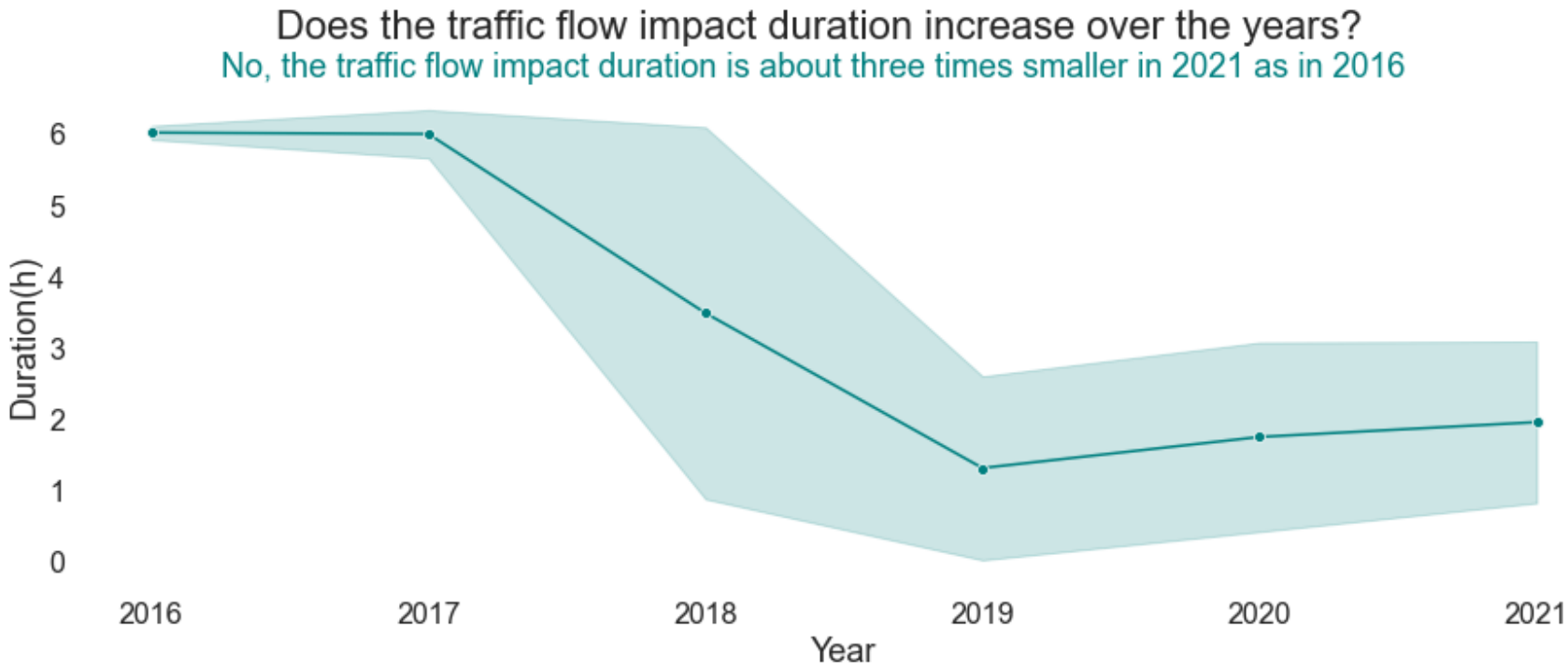
Not all severity levels show up in each year, meaning severity level of 1 appears only in year 2020 and severity level 3 is missing in year 2021.

Severity level 2 has remained at almost the same level for 2016 to 2018 but continuously increased from 2019 to 2021. In 2021 it is nearly fifteen times higher than in 2016.

Severity level 3 increased from 2016 to 2017, remains at nearly the same level in 2018 and 2019 and has been about doubled from 2016 to 2020.

Severity level 4 has been about doubled from 2016 to 2017 but has remained at almost the same level from 2017 to 2021.

TREND ANALYSIS – Question 3



Result

The traffic flow impact duration is about three times smaller in 2021 than in 2016.

In 2016 as well as in 2017 the typical traffic flow impact duration was about 6 hours (+/- 0.1 respectively 0.34 hours) and decreased in 2018 to about 3.5 hours (+/- 2.61 hours). Since 2019 it decreased again to 1.3 hours (+/- 1.29 hours) with a little increase in 2020 and 2021 to about 1.7 hours (+/- 1.33 hours) and about 2 hours (+/- 1.13 hours) respectively.

Environment Analysis

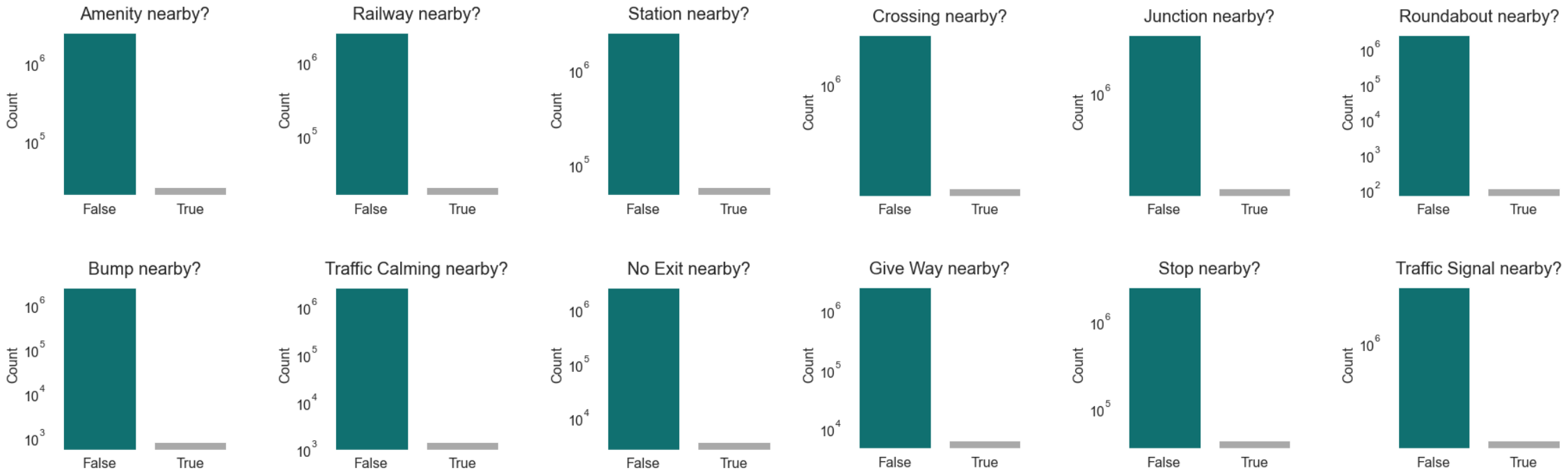
Provides insights about the environment around traffic incidents.

The following questions will be investigated

1) Do traffic incidents happen more often near certain points of interest?
2) On which relative streetside of the road do most traffic incidents happen?

Do traffic incidents happen more often near certain points of interest?

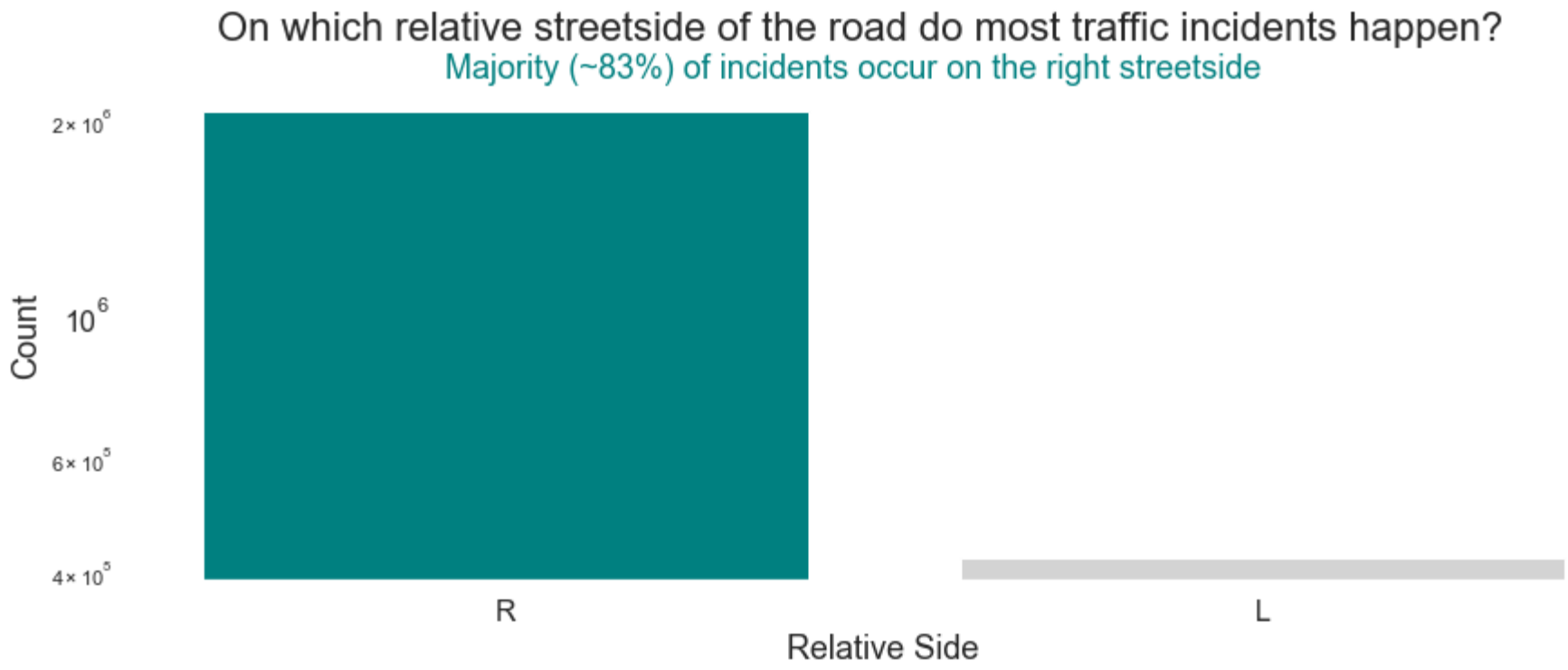
No, traffic incidents do not happen more often nearby such POIs



Result

The majority of traffic incidents do not happen more often nearby certain points of interest, most of them area about less than 2% true.

Only the POI types *Crossing*, *Junction* and *Traffic Signal* show a little bit more positive proportion with about 7%, 11% and 10% respectively.



Result

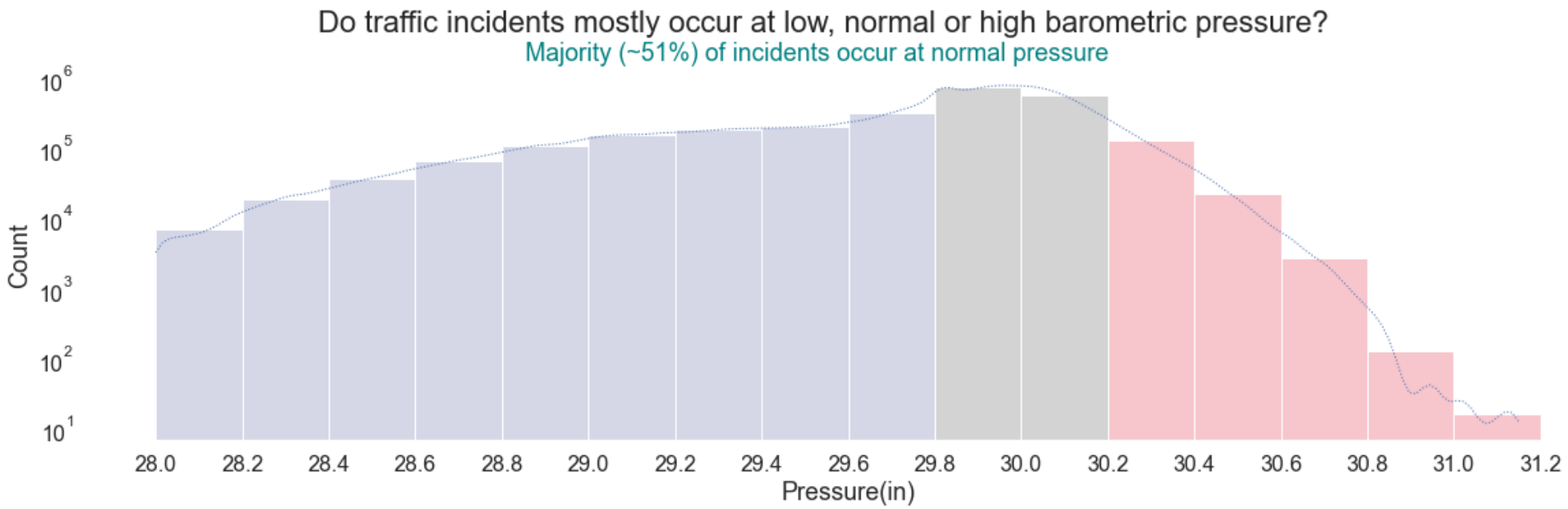
A typical traffic incident occurs on the right relative streetside with about 83% percent and only 17% on the left.

Weather Analysis

Provides insights about the weather conditions with respect to traffic incidents.

The following questions will be investigated

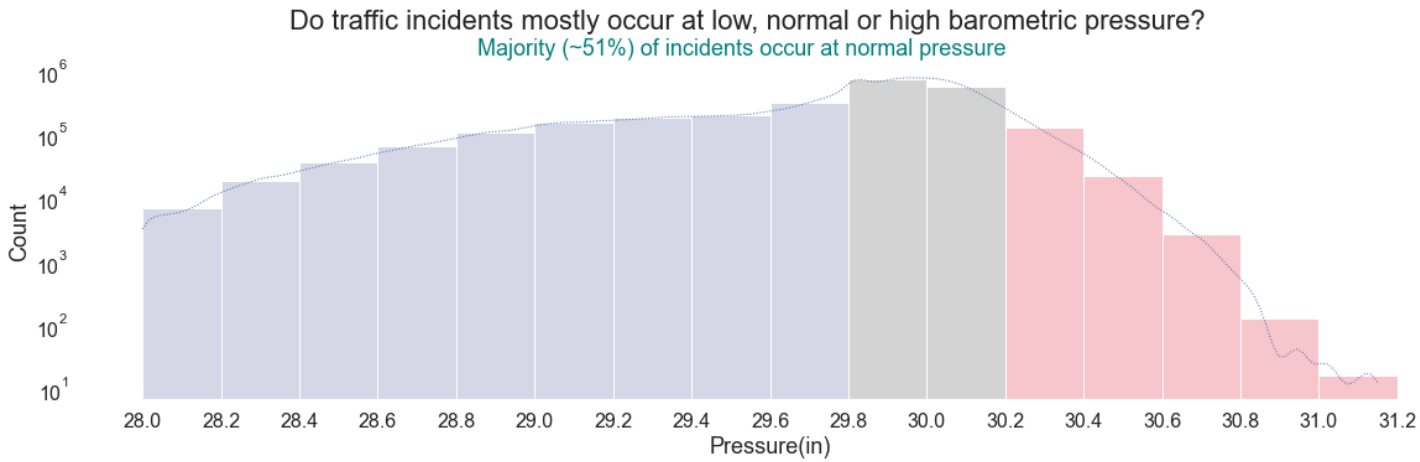
1) Do traffic incidents mostly occur at low, normal or high barometric pressure?
2) At which thermal sensation do most of the traffic incidents occur?
3) At which wind force do most of the traffic incidents occur?
4) At which rain grade do most of the traffic incidents occur?
5) At which visibility grade do most of the traffic incidents occur?
6) Do traffic incidents happen more often by rain, fog or snow?



Result

See next slide.

WEATHER ANALYSIS – Question 1 (Result)



Result

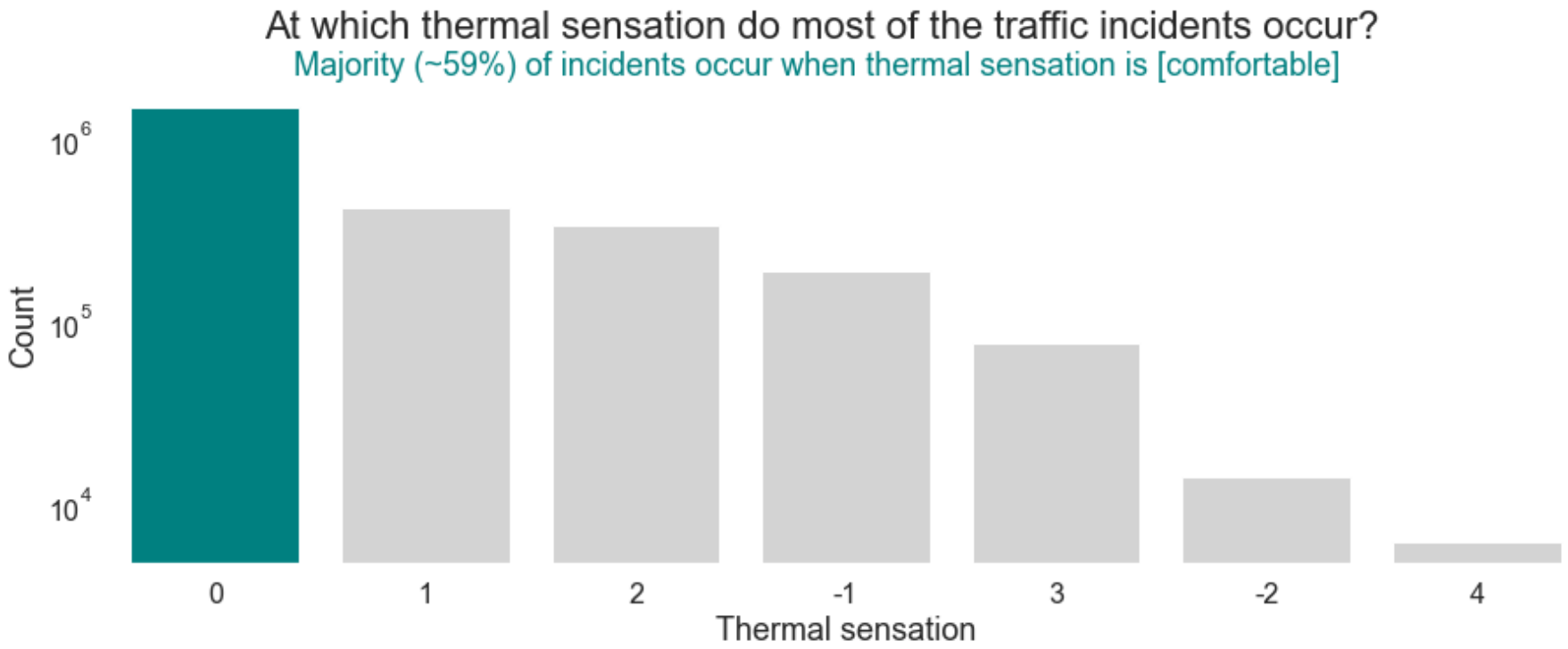
Background Knowledge

- *Normal range:* The normal range of barometric pressure is denoted between 29 inHg and 31 inHg.
- *High Pressure:* A barometric reading over 30.20 inHg is generally considered high, and high pressure is associated with clear skies and calm weather.
- *Normal Pressure:* A barometric reading in the range of 29.80 and 30.20 inHg can be considered normal, and normal pressure is associated with steady weather.
- *Low Pressure:* A barometric reading below 29.80 inHg is generally considered low, and low pressure is associated with warm air and rainstorms.

Result Interpretation

About 51% (the majority) of incidents occur at normal barometric pressure, followed by low pressure with about 43%. The taillight is the high barometric pressure, with only a proportion about 6%.

A typical traffic incident happens at normal barometric pressure with about 29.84 inHg.



Result

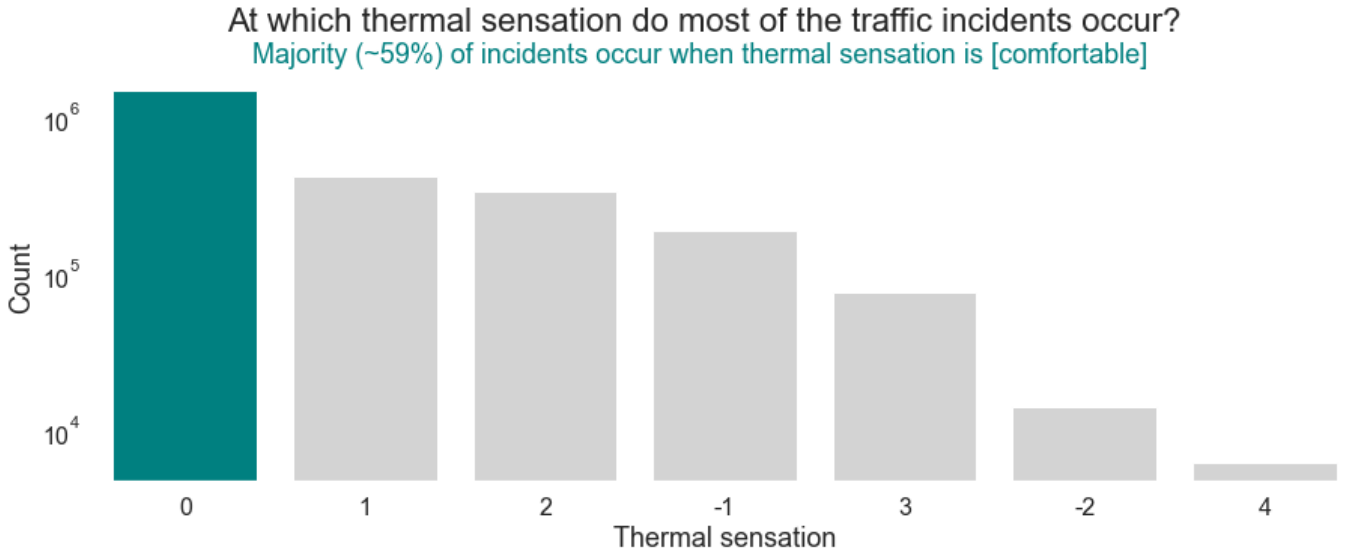
See next slide.

WEATHER ANALYSIS – Question 2 (Result)

Result

Background Knowledge

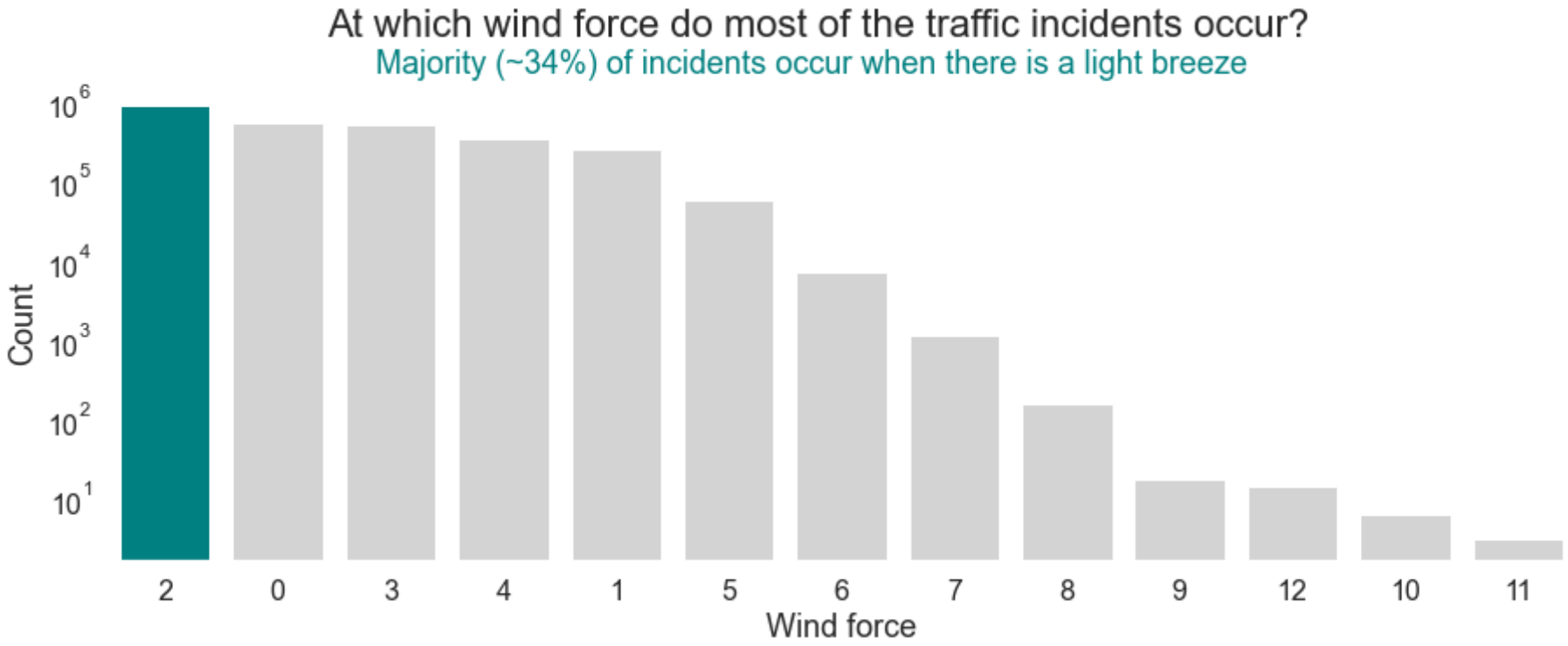
Level	Description	Feels-like Temperature (Fahrenheit)
-4	very cold	<= -38.2
-3	cold	> -38.2 and <= -14.8
-2	cool	> -14.8 and <= 8.6
-1	slightly cool	> 8.6 and <= 32
0	comfortable	> 32 and <= 68
1	slightly warm	> 68 and <= 78.8
2	warm	> 78.8 and <= 89.6
3	hot	> 89.6 and <= 100.4
4	very hot	> 100.4



Result Interpretation

About 59% (the majority) of incidents occur when thermal sensation is *comfortable*, followed by *slightly warm* with about 17%.

The rest is distributed in descending order among thermal sensation of *warm*, *slightly cool*, *hot, cool*, and *very hot* with about 13%, 7%, 3%, 0.6% and 0.3% respectively.



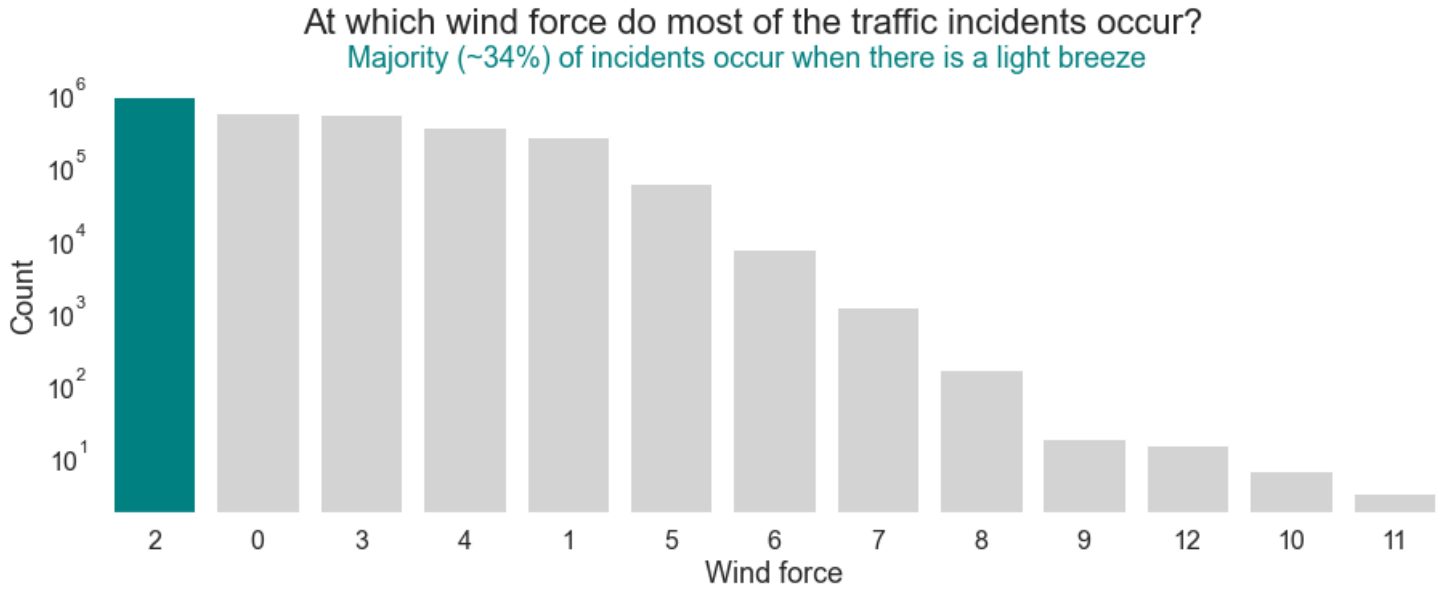
Result

See next slide.

Result

Background Knowledge

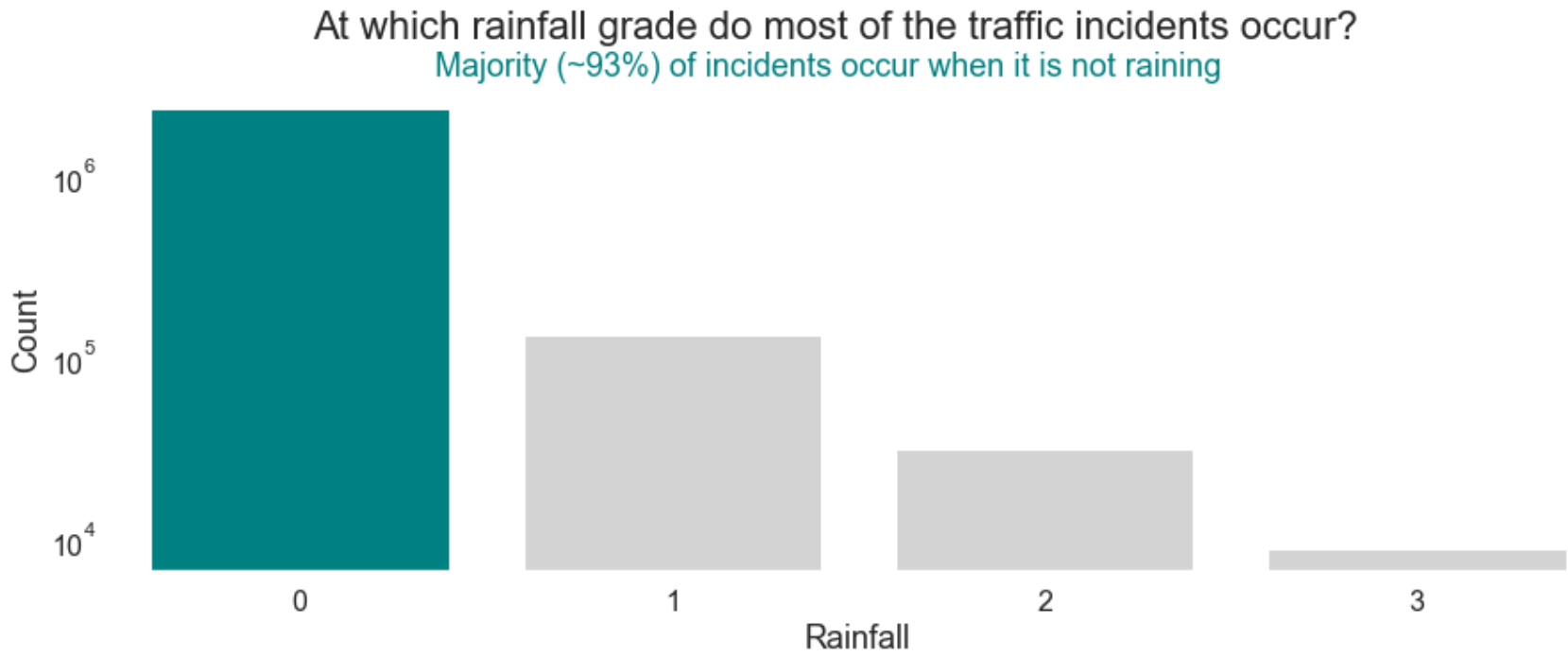
Wind Force	Description	Wind Speed (mph)
0	Calm	< 1
1	Light Air	>= 1 and <= 3
2	Light Breeze	>= 4 and <= 7
3	Gentle Breeze	>= 8 and <= 12
4	Moderate Breeze	>= 13 and <= 18
5	Fresh Breeze	>= 19 and <= 24
6	Strong Breeze	>= 25 and <= 31
7	Near Gale	>= 32 and <= 38
8	Gale	>= 39 and <= 46
9	Strong Gale	>= 47 and <= 54
10	Storm	>= 55 and <= 63
11	Violent Storm	>= 64 and <= 72
12	Hurricane	>= 73



Result Interpretation

About 34% (the majority) of incidents occur when wind force is *light breeze*, directly followed by *calm* with about 21%.

Traffic incidents during *strong gale*, *storm*, *violent storm* or *hurricane* are rather untypical.



Result

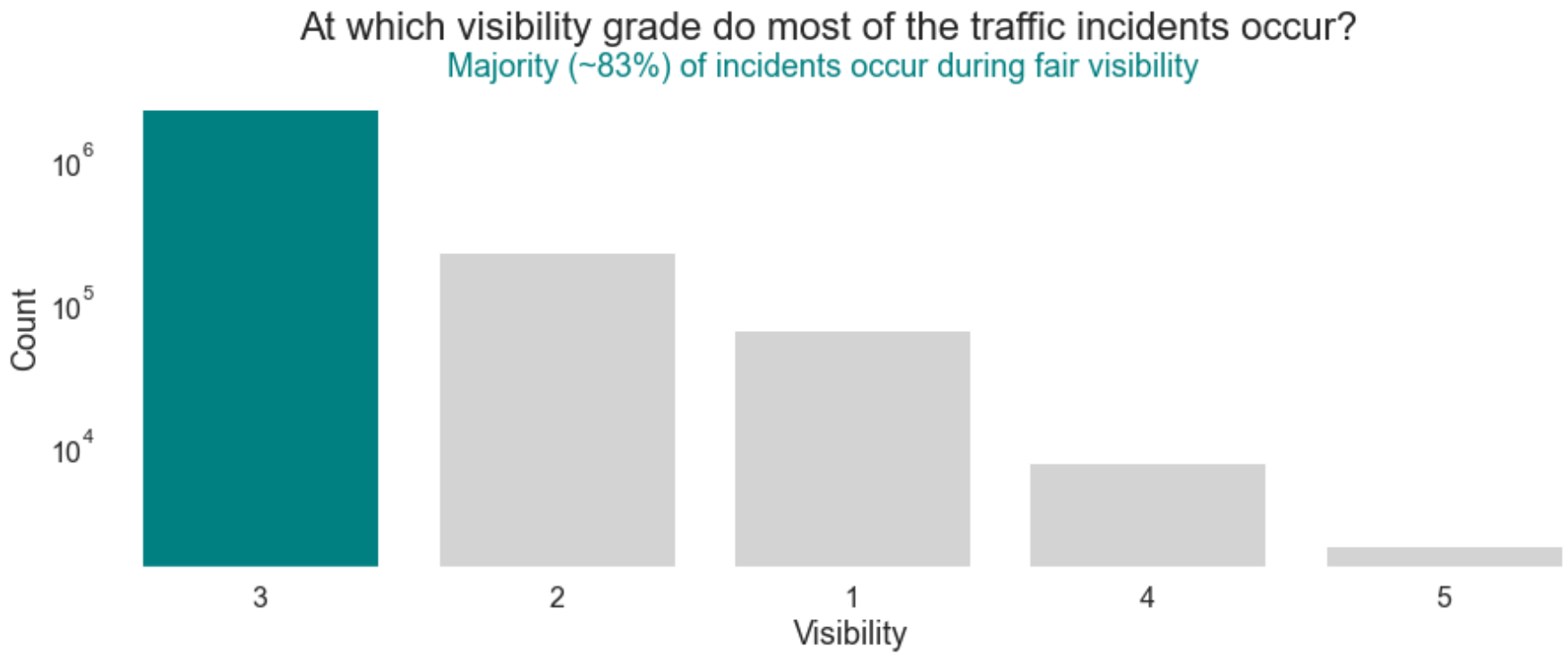
Background Knowledge

Grade	Description	Precipitation(in)
0	No rain	= 0
1	Light rain	> 0 and < 0.10
2	Moderate rain	>= 0.10 and < 0.30
3	Heady rain	>= 0.30

Result Interpretation

About 93% (the majority) of traffic incidents occur when it is not raining, directly followed by light and moderate rainfall with about 5% and 1% respectively.

Traffic incidents during heavy rainfall are rather untypical.



Result

Background Knowledge

Grade	Description	Visibility (mi)
1	Excellent visibility	> 31
2	Good visibility	> 12 and <= 31
3	Fair visibility	> 6 and <= 12
4	Poor visibility	> 1 and <= 6
5	Bad visibility	> 31

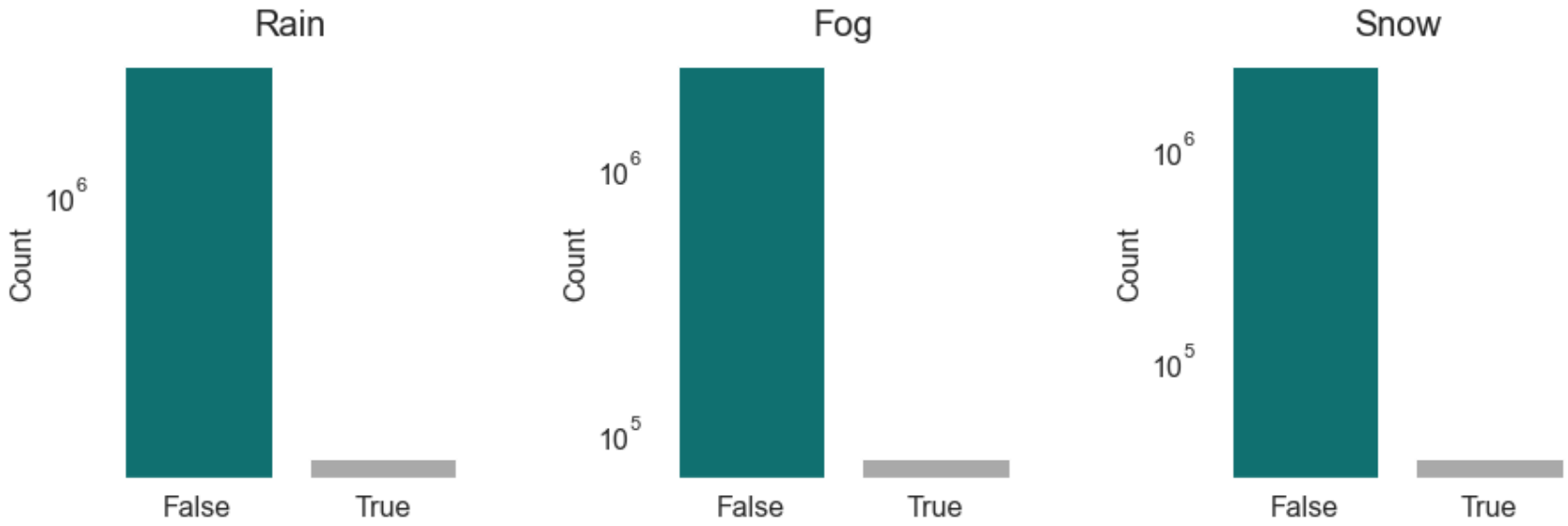
Result Interpretation

About 88% (the majority) of traffic incidents occur during fair visibility, directly followed by good and excellent visibility with about 9% and 3% respectively.

Traffic incidents during bad visibility are rather untypical.

Do traffic incidents happen more often by rain, fog or snow?

No, traffic incidents do not happen more often during such conditions



Result

The majority of traffic incidents do not happen more often by rain, fog or snow. Only about 7% happens during rain, about 3% during fog and about 1% during snow.

Relationship Analysis

Provides insights about possible relationships between features.

The following questions will be investigated

1) What relations exist between the features (Covariance)?
2) What relations exist between the features (Correlation / Pearson)?
3) Which features have a positive relationship (Correlation / Pearson)?
4) Which features have a negative relationship (Correlation / Pearson)?

RELATIONSHIP ANALYSIS – Question 1

What relations exist between the features (Covariance)?

There are 70 positive and 46 negative covariances

Total positive covariances: 70

	Feature_1	Feature_2	Cov_Value	Cov_Type
32	Start_Lng	Second	79.98000	Positive
21	Start_Lng	Humidity(%)	62.59000	Positive
51	Temperature(F)	Hour	17.41000	Positive
60	Temperature(F)	Thermal_Sensation	13.19000	Positive
15	Start_Lat	Second	8.19000	Positive
...	...	...	...	...
43	Temperature(F)	Pressure(in)	0.18000	Positive
83	Year	Day	0.17000	Positive
14	Start_Lat	Minute	0.17000	Positive
100	Day	Thermal_Sensation	0.15000	Positive
45	Temperature(F)	Station	0.15000	Positive

70 rows × 4 columns

Total negative covariances: 46

	Feature_1	Feature_2	Cov_Value	Cov_Type
42	Temperature(F)	Humidity(%)	-139.80000	Negative
8	Start_Lat	Temperature(F)	-40.74000	Negative
67	Humidity(%)	Hour	-28.44000	Negative
6	Start_Lat	Start_Lng	-10.52000	Negative
87	Year	Second	-7.24000	Negative
...	...	...	...	...
62	Humidity(%)	Traffic_Signal	-0.18000	Negative
112	Second	Fog	-0.17000	Negative
24	Start_Lng	Junction	-0.16000	Negative
57	Temperature(F)	Fog	-0.16000	Negative
70	Humidity(%)	Duration(h)	-0.15000	Negative

46 rows × 4 columns

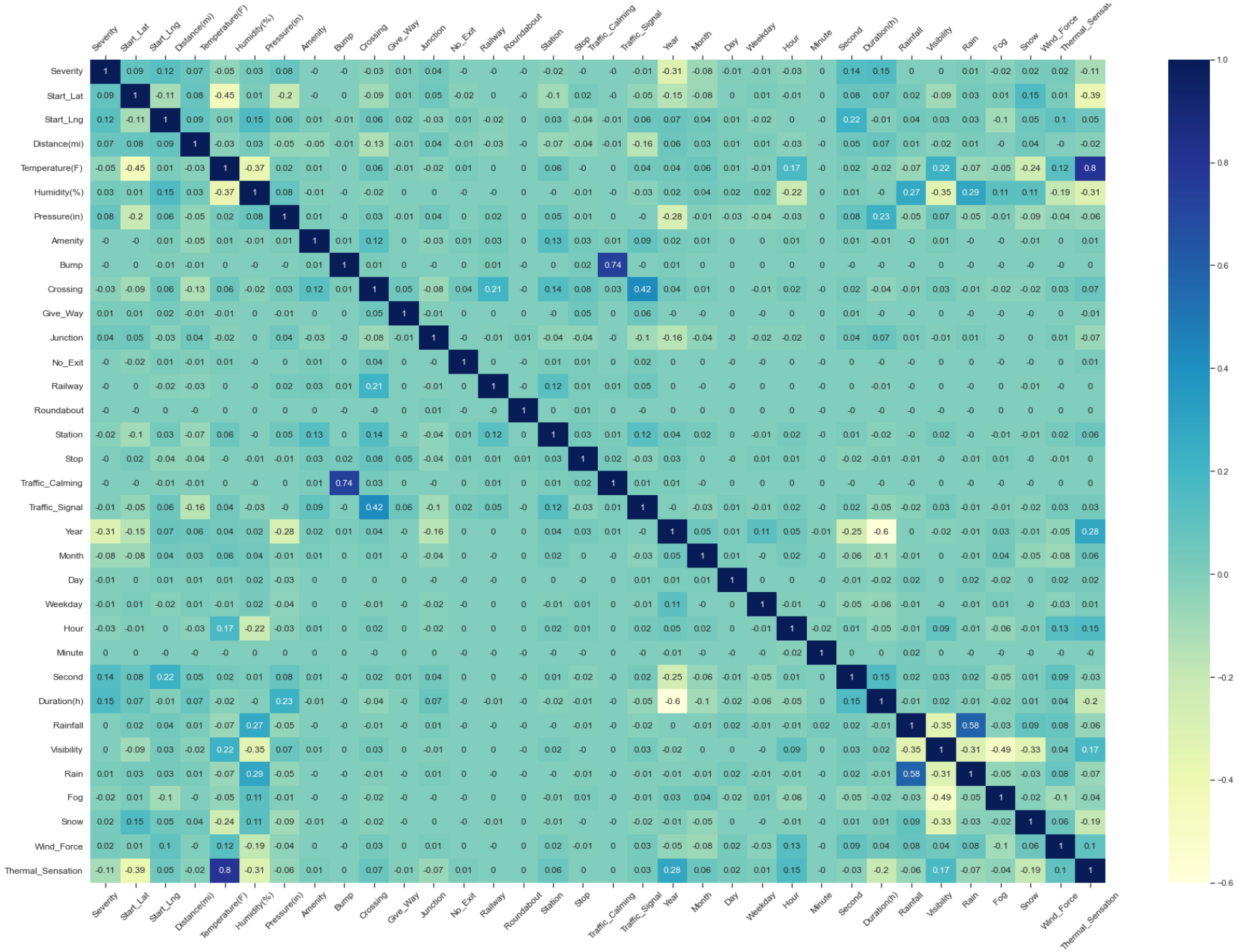
Result

The pairwise covariance data frame reveals that several positive linear relationships (about 70, greater than 0) as well as negative linear relationships (about 46, smaller than 0) between two features exist. No linear relationships are equal to zero (keep in mind that nevertheless non-linear relationships might exist).

The magnitude of the covariance is not easy to interpret because it is not normalized and hence depends on the magnitudes of the features.
As a result, the further relationship analysis will be performed with support of the normalized version of the covariance, the Pearson Correlation Coefficient.

(With covariance threshold of at least / less than 0.15)

RELATIONSHIP ANALYSIS – Question 2



What relations exist between the features (Correlation / Pearson)?

Result

The correlation heatmap reveals that several positive linear relationships (blue gradient) as well as negative linear relationships (yellow gradient) between two features exist.

No linear relationships are colored green (keep in mind that nevertheless non-linear relationships might exist).

⚠ Important

Correlation does not imply causation! A correlation does not mean that the changes in one feature actually *cause* the changes in the other feature.

However, a suspected causal and linear relationship between two variables can be ruled out if the correlation coefficient is close to zero.

RELATIONSHIP ANALYSIS – Question 3

Which features have a positive relationship (Correlation / Pearson)?

List of high degree / strong correlations (>= 0.50)

Feature 1	Feature 2	Correlation Value
Temperature(F)	Thermal_Sensation	0.80
Bump	Traffic_Calming	0.74
Rainfall	Rain	0.58

List of moderate degree / medium correlations (>= 0.30 and < 0.50)

Feature 1	Feature 2	Correlation Value
Crossing	Traffic_Signal	0.42

List of low degree / small correlations (< 0.30)

Feature 1	Feature 2	Correlation Value
Humidity(%)	Rain	0.29
Year	Thermal_Sensation	0.28
Humidity(%)	Rainfall	0.27
Pressure(in)	Duration(h)	0.23
Start_Lng	Second	0.22
Temperature(F)	Visibility	0.22
Crossing	Railway	0.21
Visibility	Thermal_Sensation	0.17
Temperature(F)	Hour	0.17
Start_Lat	Snow	0.15
Hour	Thermal_Sensation	0.15
Second	Duration(h)	0.15
Start_Lng	Humidity(%)	0.15
Severity	Duration(h)	0.15

Result

With correlation threshold of at least 0.15 there are 18 total positive correlations, meaning when one feature increases the other increases (the larger A, the larger B).

As the magnitude is a measure of strength there are weaker and stronger positive correlations: There are 3 correlations of high degree, 1 of moderate degree and 14 with low degree.

Which features have a negative relationship (Correlation / Pearson)?

List of high degree / strong correlations (>= -0.50)

Feature 1	Feature 2	Correlation Value
Year	Duration(h)	-0.60

List of moderate degree / medium correlations (>= -0.30 and < -0.50)

Feature 1	Feature 2	Correlation Value
Visibility	Fog	-0.49
Start_Lat	Temperature(F)	-0.45
Start_Lat	Thermal_Sensation	-0.39
Temperature(F)	Humidity(%)	-0.37
Humidity(%)	Visibility	-0.35
Rainfall	Visibility	-0.35
Visibility	Snow	-0.33
Severity	Year	-0.31
Visibility	Rain	-0.31
Humidity(%)	Thermal_Sensation	-0.31

List of low degree / small correlations (< -0.30)

Feature 1	Feature 2	Correlation Value
Pressure(in)	Year	-0.28
Year	Second	-0.25
Temperature(F)	Snow	-0.24
Humidity(%)	Hour	-0.22
Start_Lat	Pressure(in)	-0.20
Duration(h)	Thermal_Sensation	-0.20
Humidity(%)	Wind_Force	-0.19
Snow	Thermal_Sensation	-0.19
Distance(mi)	Traffic_Signal	-0.16
Junction	Year	-0.16
Start_Lat	Year	-0.15

Result

With correlation threshold of less than -0.15 there are 22 total negative correlations, meaning when one feature decreases the other increases (the larger A, the smaller B).

As the magnitude is a measure of strength there are weaker and stronger negative correlations: There are 1 correlation of high degree, 10 of moderate degree and 11 with low degree.

For all details, please refer to the related Jupyter Notebook:
https://www.kaggle.com/code/anjakuchenbecker/eda-on-us-traffic-incidents