**Data Analytics broad plan**

1. US ACCIDENTS dataset is single dataset.
2. **Introduction:** This public dataset has car accidents (in USA) data since Feb, 2016 and going till Dec 2019 and covers most of the officially reported accidents in 49 states of USA (except Alaska and Hawaii but including Washington DC) categorized by severity level. This data is non-personal data used for community. This data is collected automatically through API calls to MapQuest and Bing database systems and merging them together in one file. This single dataset has approx. 3 million (though I got only 1,048,576 on my local machine) rows of car accident records but it keeps growing. It has 49 different columns all of which are scalar data including accident description in plain English etc. It begins with unique id (called ID) for each row, source of data which may be ‘MapQuest’ or ‘Bing’ or ‘MapQuest-Bing’ meaning both. It has Traffic Message Control code (TMC) type with which type of this accident can be tracked in various systems in the nation. The structure of TMC is generally complex implying the section of road, direction of traffic, zone and unique identifier together. But in our case, this appears to be like a category of accident. There is a severity column that encodes levels of severity in the accident which starts from 0 and goes upto 4 (meaning high severity). There are many columns like the address of accident site, start\_location (Start\_longitude, Start\_lattitude), end\_location (longitude, latitude) (in case it spread far), incident time (start\_time, end\_time) (date & time), distance covered within zone of accident (in case it spread far), its’ impact on traffic flow, the lane side (right or left) where accident landed, address, daytime/nighttime, weather condition, amenities, infra-structure traffic constraints (like traffic light or intersection or round-about or traffic calmer etc).
3. The web-link is : <https://www.kaggle.com/sobhanmoosavi/us-accidents>

In order to download, create a Kaggle account and download the dataset into S3 bucket.

1. In Athena, check if we can query the data in bucket. Also check if you can use Quicksight.
2. The detailed description of the 49 columns of data can be found in either the web-site above or another one as <https://smoosavi.org/datasets/us_accidents> . Please learn them.
3. **Broad steps of data processing:**
4. **Data format:**

ID varchar

Source varchar

TMC float64

Severity int64

Start\_Time DateTime

End\_Time DateTime

Start\_Lat float64

Start\_Lng float64

End\_Lat float64

End\_Lng float64

Distance(mi) float64

Description varchar

Number float64

Street varchar

Side varchar

City varchar

County varchar

State varchar

Zipcode varchar

Country varchar

Timezone varchar

Airport\_Code varchar

Weather\_Timestamp DateTime

Temperature(F) float64

Wind\_Chill(F) float64

Humidity(%) float64

Pressure(in) float64

Visibility(mi) float64

Wind\_Direction varchar

Wind\_Speed(mph) float64

Precipitation(in) float64

Weather\_Condition varchar

Amenity bool

Bump bool

Crossing bool

Give\_Way bool

Junction bool

No\_Exit bool

Railway bool

Roundabout bool

Station bool

Stop bool

Traffic\_Calming bool

Traffic\_Signal bool

Turning\_Loop bool

Sunrise\_Sunset varchar

Civil\_Twilight varchar

Nautical\_Twilight varchar

Astronomical\_Twilight varchar

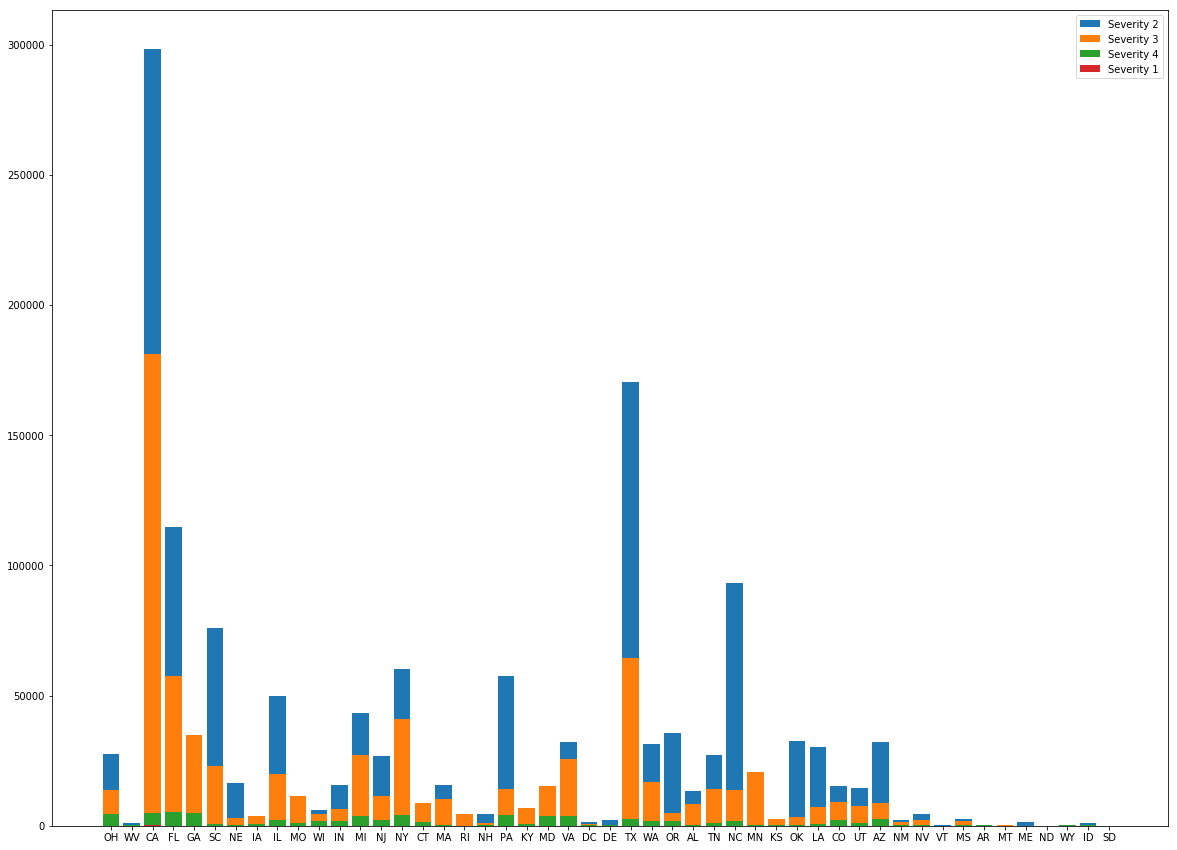
1. **Initial check:** We need to check some very basic things like: State has only 49 distinct non-null values; Country is non-null and ‘US’. Start\_Time of accident occurrence is not null and a valid datetime. Zipcode is not null. Start\_Lattitude, Start\_Longitude is non-null and > 0. If there is any data in violation, delete those rows. If we have severity other than 0-5, we can delete the data. You should delete all data with severity 0.



1. **Clean-up of data:** If there are Timezones (for incident-site; start) with more than 5 distinct values (Esatern, Central, Mountain, Pacific, Hawai), then update them to the standard 5 values. If Start\_Time is not a valid datetime then delete those data. If Side is neither null nor ‘R’ nor ‘L’, then make it null. Distance, Temperature, Wind-Chill, Humidity, Pressure, wind\_Speed, Precipitation should be >= 0 or null. Amenity, Bump, Crossing, Give\_way, Junction, No\_Exit, Railway, RoundAbout,Station, Stop, Traffic\_Calming, Traffic\_Signal, Turning\_Loop should be TRUE or FALSE or null. Change other values to null. Sunrise\_Sunset, Civil\_Twilight, Nautical\_Twilight, Astronomical\_Twilight should have allowed values (‘Night’, ‘Day’, null). Change other values to null.
2. **Queries:** This table is very simple with simple columns. Only complex columns can be Description and Weather (and may be a few more). We will have simple queries. But we will pay attention to picking majority. For example, what is the majority severity in state of Texas? Like this. In addition, we may need to calculate various co-relation coefficients and save on tables. We will have about 5-6 such problems.
3. **Data visualization:** Later we will move onto presenting result data in colored visual forms using graphs and charts for each of the saved result tables. Various visual forms will be used like histogram graphs, plots, pie-charts and if possible colored matrix squares. We will try to use Quickstart tool in AWS for this purpose.

That was just a broad plan where our objective is to draw the diagrams. Now we enumerate each problem.

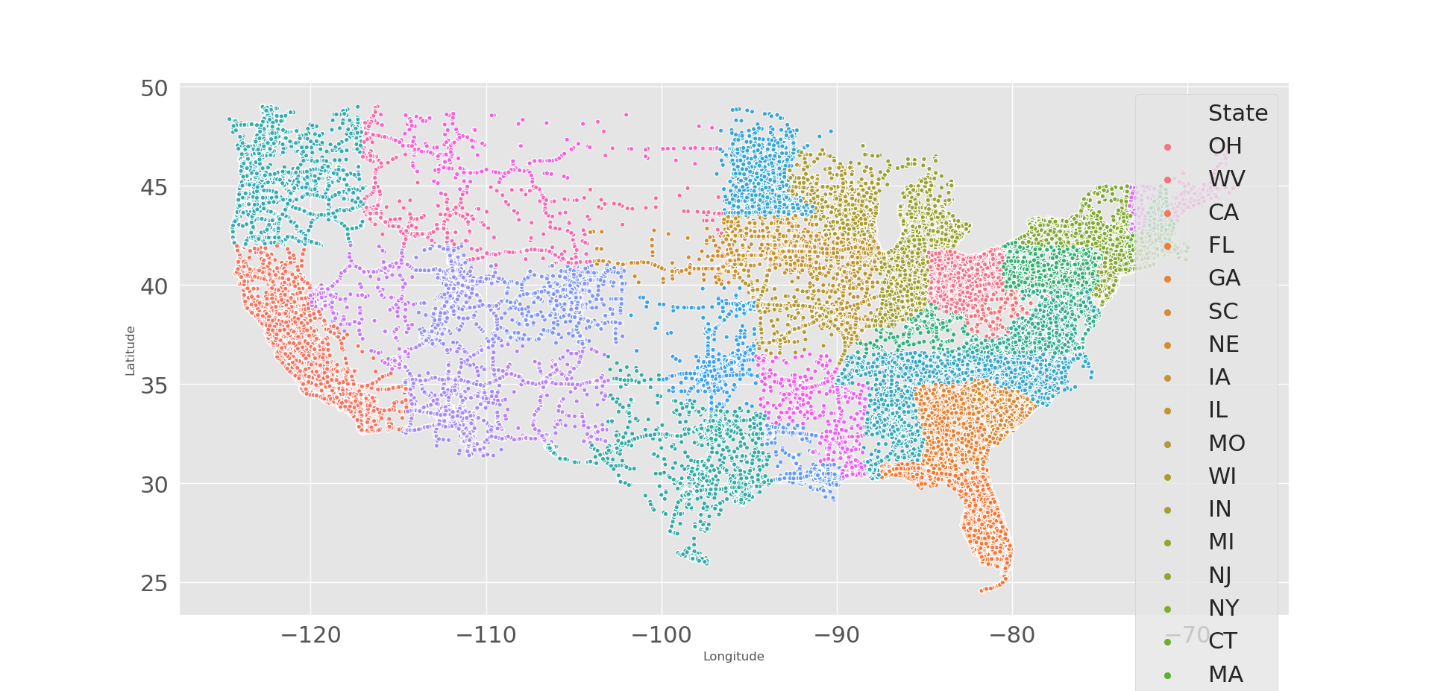
**Query & Visual 1.** We assume there is no severity 0 data. Using vertical histogram bars, draw vertical rectangles to show total no. of accidents for each state (in sequence of total no. of accidents). Use size 16 (horizontal) X 10 (vertical) aspect ratio to printable on 8.5 X 11 inches paper. In X-dimension plot states (showing state code) and in Y-dimension, draw the rectangle (say each fixed width like 0.15 inches or so) with height as total no. of accidents in the respective state. Inside each vertical rectangle, use four colors to show the strength of each severity (or percentage) in this sequence: Green (on the base) for severity 4 (most severe), Orange (above base) for severity 3 (less severe), Blue (above severity 3) for severity 2 (less severe), Red (on top) for severity 1 (least severe). The histogram plot should look like fig 1. The sequence of states on X -direction is dictated by y value the state for highest y-value should show first and then the state with next highest y-value etc.



49 States in sequential order of accidents

Fig 1

**Query & Visual 2.** Simple Scatterplot: First take a map of USA with gray shade and gray border lines and with bold state border lines. Generate a scatterplot of accidents (ignore severity 0) as points with longitude and latitude upon this map as shown in figure 2. Break the map into rectangles each of 10 degrees longitude (-130 to -70) and 5 degrees latitude (+25 to 50). So we separate these 42 rectangles with gray lines. This is second level. Now take 20 different colors and assign one to these rectangles such that no neighboring rectangles have same color. Now generate the scatterplots of accidents within these rectangles with their colors. The plot should look like fig 2.



Scatterplots of accidents by start\_location

**Query & Visual 3.** Simple Pie-chart: Let us define duration of accidents as end\_time – start\_time in minutes & create 5 categories of durations as: 1) minutes (<= 59 minutes), 2) hours (between 60 minutes and 239 minutes), 3) 4hours\_16hours (between 240 minutes and 959 minutes), 4) 16hours\_24hours (between 960 minutes and 1,439 minutes), 5) days (>= 1440 minutes). Ignore severity 0 and count total\_no\_accidents by duration and draw a full-size pie-chart with parts of pies proportional to total\_no\_accidents for 5 different categories with 5 different contrasting colors. The figure should somewhat look like fig 3 (no years).

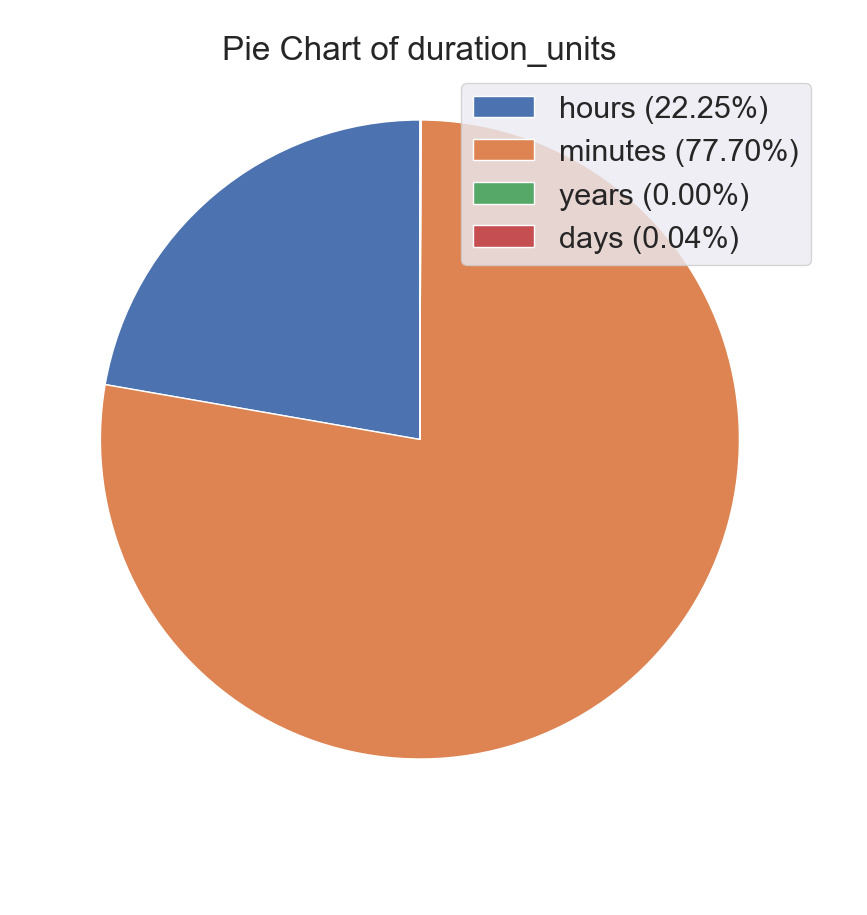


Fig 3: Pie-chart of no\_of\_accidents with duration.

**Query & Visual 4.** Relations: If we look for number of accidents against various weather conditions, we notice most accidents happen in clear or good weather. But suppose we look at breakdown of accidents within each weather condition and then we acquire further insights. Say we count total no of accidents within each weather condition and call it weather\_accidents[i]. Then we calculate percentage\_of\_accidents within each weather condition by calculating percentage\_of\_severity\_1\_accidents[i] as no\_of\_severity\_1\_accidents[i]/weather\_accidents[i] and like that for severity 2, severity 3 and severity 4. Now if we plot percentage\_of\_severity\_1\_accidents[i] , percentage\_of\_severity\_2\_accidents[i], percentage\_of\_severity\_3\_accidents[i], percentage\_of\_severity\_4\_accidents[i] (together) against I then we look for contribution of weather on severity of weather. The bar plot should look somewhat like fig 4 .



Fig 4: On x direction each weather condition and on y direction total percentage of severity\_of\_accidents

1. **Query & Visual 5.** Relations with parameters : We want to observe in bad weather conditions (meaning leaving out fair weather conditions like fair, clear day, overcast, mostly cloudy, partly cloudy, light rain, scattered rain etc.), if there is any relation between some natural parameters like temperature, wind-chill, humidity, pressure, visibility, wind-speed, precipitation, infra-structure parameters like Bump, Crossing, Give\_Way, Junction, No\_Exit, Railway, Roundabout, Station, Stop, Traffic\_Calming, Traffic\_Signal, Turning\_Loop and accidents at each severity. We ask this question? Within bad weather conditions (non-good weather conditions), is there any relation between any of the 7 infra-structure parameters as crossing, give\_way, junction, no\_exit, roundabout, stop, traffic\_signal and number of accidents (for each severity) or 7 natural parameters (like temperature, wind-chill, humidity, pressure, visibility, wind-speed, precipitation) and number of accidents (for each severity)? That means for each day, we want to plot curves for each of the 7 + 7 parameters vs number of accidents (without it and with it) at each severity. Show 7+ 7 frames each containing 4 graphs for 4 severities. We may extend this problem by adding another constraint like odd time (outside normal working hour period) in future. Let us take temperature ranges like ((-30 - -21), (-20 - -11), (-10 - -1), (0 - 9), (10 - 19), (20 - 29), (30 - 39), (40 - 49), (50 - 59), (60 - 69), (70 - 79), (80 - 89), (90 - 99), (100 – 130). Let us produce a result table which looks like below:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| date | Temperature (F)  (range) | Bad\_  wthr\_  sev4\_  acc | Good\_  wthr\_  sev4\_  acc (non-bad) | Bad\_  wthr\_  sev3\_  acc | Good\_  wthr\_  sev3\_  acc | Bad\_  wthr\_  sev2\_  acc | Good\_  wthr\_  sev2\_  acc | Bad\_  wthr\_  sev1\_  acc | Good\_  wthr\_  sev1\_  acc |
| 03/03/2019 | -30 - -21 | 100 | 110 | 26 | 30 | 16 | 12 | 13 | 11 |
| 03/03/2019 | -20 - -11 |  |  |  |  |  |  |  |  |

In the above figure, we mean to have rows for all existing ranges and then go for the next date. This way, we need to build all 14 result tables for all 14 parameters. Wind-chill should have ranges like (-70 - -61), (-60 - -51), (-50 - -41), (-40 - -31), (-30 - -21), (-20 - -11), (-10 - -1), (0 - 9), (10 - 19), (20 - 29), (30 - 39), (40 - 49), (50-59), (60 - 69), (70 - 79), (80 - 89), (90 - 99), (100-109), (110 - 119), (120-129), (130 -139), (140-149), (150 - 159), (160 -169) etc.. Humidity (%) ranges may be (1-5), (6-10), (11-15), (16-20), (21-25), (26 - 30), (31-35), ….., (91 – 95), (95 – 100). Pressure ranges may be like (0 -3), (4 - 7), (8 - 11), (12 - 15), (16 - 19), (20 - 23), (24 - 27), (28 - 31), (32 - 35). Visibility(mi) ranges may be (0 – 0.06), (0.06001 – 0.1), (0.10001 (means greater than 0.1) – 0.12), (0.12001 (means greater than 0.12) – 0.19), (0.19001 (means greater than 0.19) – 0.25), (0.25001 (means greater than 0.25) – 0.30), (0.30001 (means greater than 0.3) – 0.38), (0.38001 (means greater than 0.38) – 0.4), (0.40001 (means greater than 0.4) – 0.5), (0.50001 (means greater than 0.5) – 0.6), (0.60001 (means greater than 0.6) – 0.7), (0.70001 (means greater than 0.7) – 0.8), (0.80001 (means greater than 0.8) – 0.9), (0.90001 (means greater than 0.9) – 1.0), (1.00001 (means greater than 1.0) – 1.2), (1.20001 (means greater than 1.2) – 1.4), (1.40001 (means greater than 1.4) – 10), (10.0001 (means greater than 10) – 50), (50.00001 (means greater than 50) – 150). For wind\_speed, take 44 ranges like (0 – 1), (1.00001-2), (2.00001-3), (3.00001-4), (4.00001-5), ……. (41.00001 – 42), (42.00001 – 43) and above 43.00001. For precipitation, keep 25 ranges like (0 – 1), (1.00001-2), (2.00001-3), (3.00001-4), (4.00001-5), ……. (24.00001 – 25). For each of the seven infra-structure params, it is Boolean value true or false. The line plot should look somewhat like fig 5 .

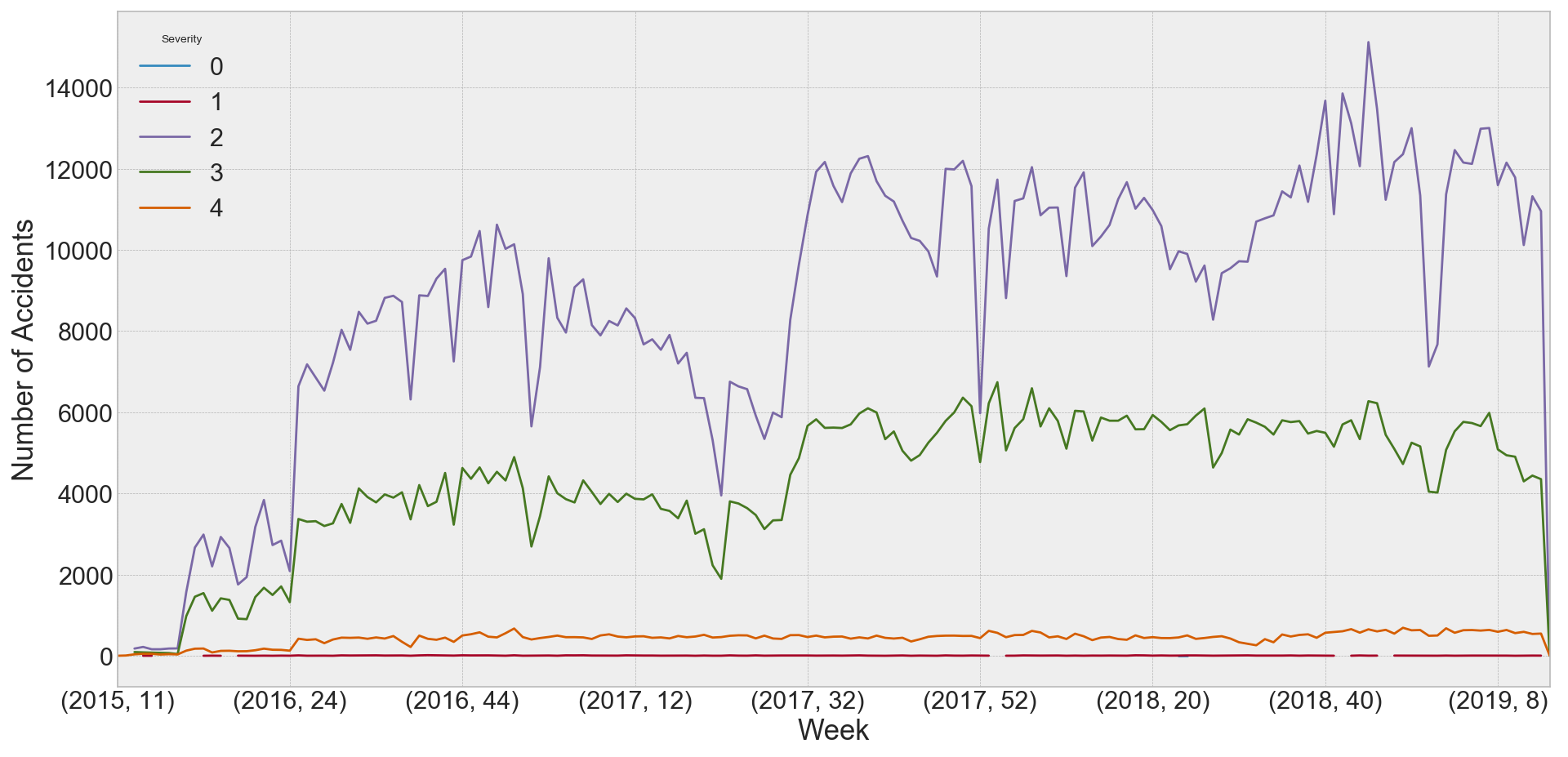


Fig 5: showing number of accidents (good weather and bad weather -2 graphs) for various ranges for one parameters.

**Query & Visual 6.** Trends presentation (Continuous plot): Objective is to watch the plot for total no of accidents vs the hour of the day. Count total number of accidents that occurred during each hour (0 -th hour, 1-st hour,… 23-rd hour) and plot that using line plots (on grid) along x-axis of hours as in fig 6.

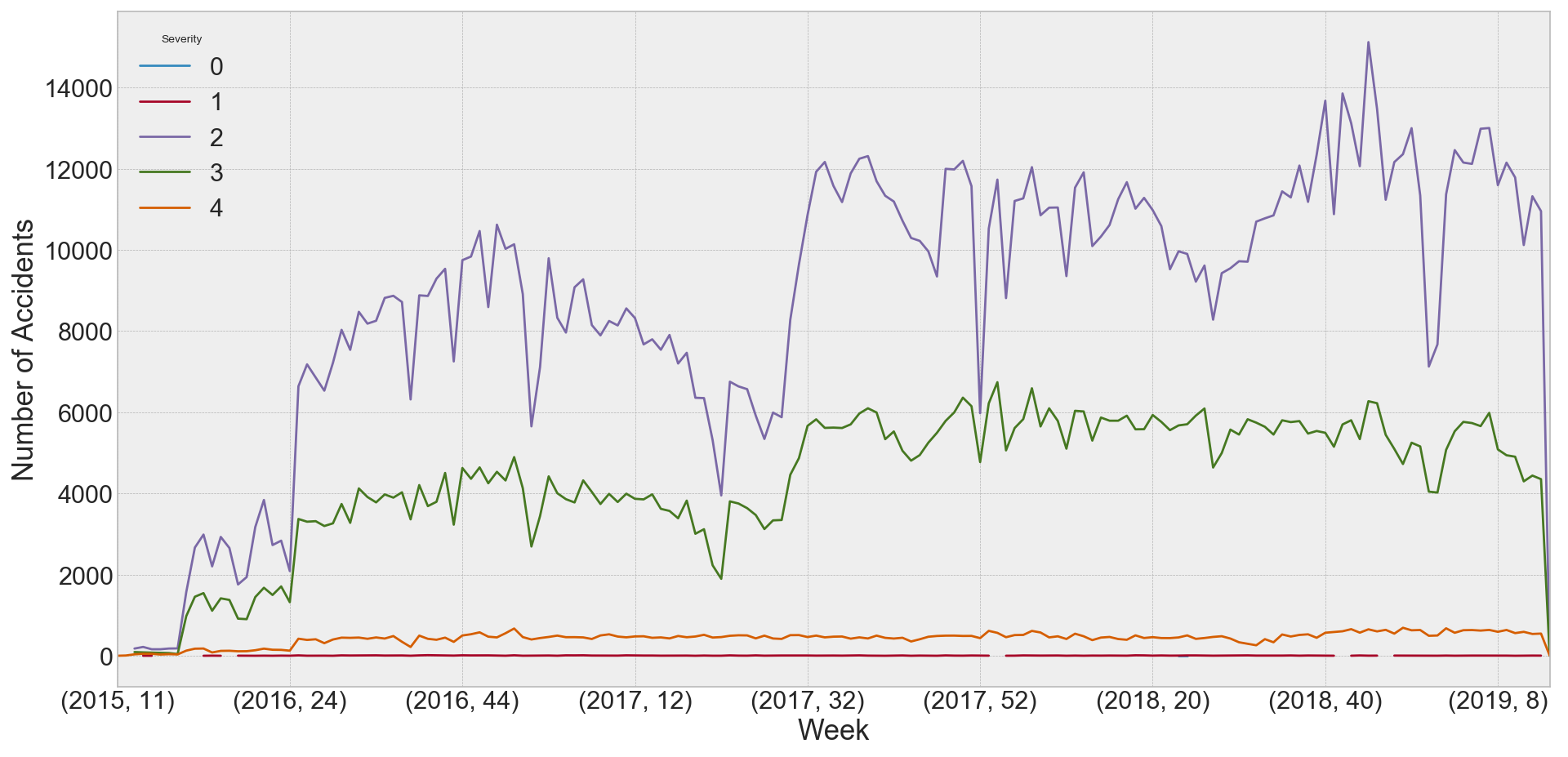


Fig 6: showing number of accidents for each severity over each week

**Query & Visual 7.** Average distanceTrends presentation for each day (Histogram plot): May we show the average distance (& max and stddev) of traffic affected by accidents of each severity? This will result in 4 triples of histograms (1 for each severity) of 3 items (avg, max and std-dev with 3 distinct colors). The visual may be somewhat like shown in fig 7.



Fig 7 : consecutive triples of 3 colors for each severity; totally 12

**Query & Visual 8.** Distance affected by severity of accidents (Histogram plot): Objective is to find the relations between each severity of accidents and average distance affected over given severity. It may look like Fig 7.

Fig 6: total\_number\_of\_accidents vs hour of day.

**Next level:**

**Query & Visual 9.** Relation between side of road and number of accidents: Objective is to answer questions like which side of the road is more prone to accidents for each severity. For each seHere relations between calculate co-relations between some attributes. To keep things simple, we will ignore all text attributes. For Boolean values, assign 1 to true and 0 to false. Now compute the co-relation matrix between following columns: TMC, Severity, Start\_Lat, Start\_Lng, End\_Lat, End\_Lng, Distance(mi), Temperature(F), Wind\_Chill(F), Humidity(%), Pressure(in), Visibility(mi), Wind\_Speed(mph), Precipitation(in), Amenity, Bump, Crossing, Give\_Way, Junction, No\_Exit, Railway, Roundabout, Station, Stop, Traffic\_Calming, Traffic\_Signal, Turning\_Loop. Use the formulae for Pearson co-relation coefficients (PCC). Use this formula:

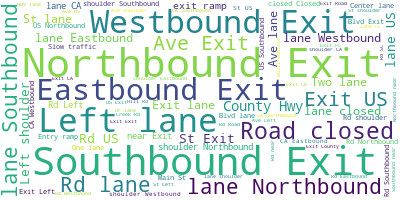
R = (n(∑xy) – (∑x) (∑y)) / ([n∑x2 - (∑x)2].[n∑x2 - (∑x)2]). Here n indicates ranging over all individual data points (n is not a multiple).

/

Fig 7: Severity of accidents vs Average Distance (in mi) affected

**Query & Visual 10.** Concepts in Description: The description column contains a free-flowing text description on the accident. Can we find any small phrase concepts in that field which repeat in many accidents (meaning they are statistically precious)? Do we need to put some stop words to avoid noise or to set direction in our search (otherwise stop words will flood our analysis)? We may use Athena to build the query to build a table of concepts and frequency. The concepts and frequencies may be represented using a wordCloud diagram as in fig 10.

Fig 10: Concepts in a tagCloud



**Query & Visual 13.** Simple aggregate function: Enough of individual rows! We want to try generating a very simple summary report on state, date, precipitation and a running total over date for each state. It can be easily done using “**group by**” clause but we want to take the alternative approach. Can we build this report from table avg\_precipitation (output of problem 14) without using “group by” clause in Athena? The state, date, precipitation graph may be displayed for top 5 states using 5 graphs like in fig 13.

**Hint:** Use Window function.

Fig 13: Graph plot for precipitation vs dates for each state (5 states)

