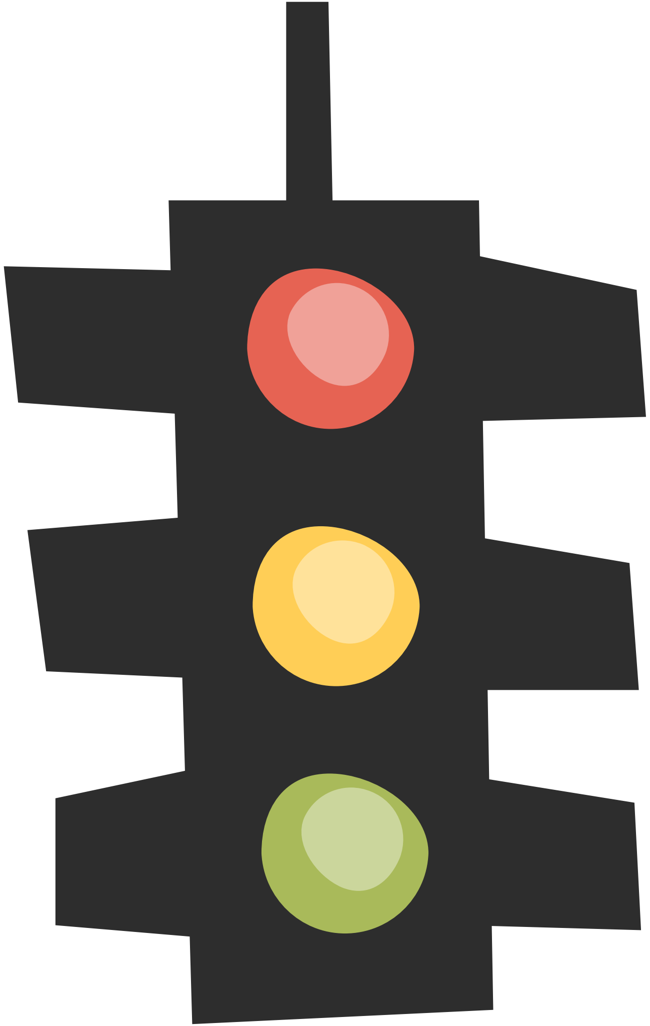
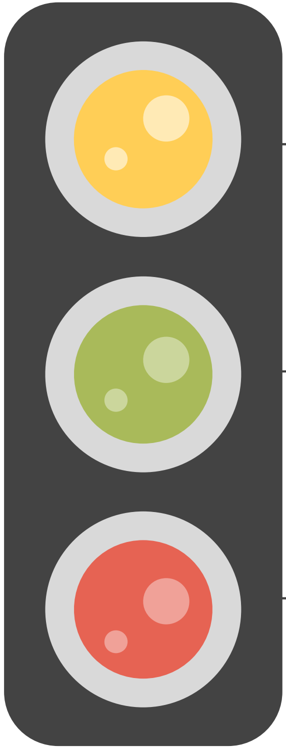


# Accident Severity & DST

Jeffrey Floyd, Mary Schindler, & Curtis HopeHill

# Presentation Outline



| | |
|---|---|
| Problem | Problem Statement & Stakeholders |
| Background | Prior research on Daylight Savings Time (DST) & Accidents |
| Cleaning | “mrclean” function to clean data & further fine tuning with feature engineering |
| EDA | Exploring data distribution & correlations of features |
| Modeling | Modeling to predict accident severity, & hypothesis testing. |
| Streamlit | Real time severity predictions. |

# Problem Statement

Rideshare Company ® has contracted with Sigmoid Data Science to investigate the impacts of the Spring and Fall time changes caused by Daylight Savings Time (DST). We were asked to determine if the average severity of accidents in the United States increased in the week of the time change and the week following the change. They would like to lobby to recommend eliminating DST and would like traffic accident data to support their recommendation.

# Daylight Savings Time

- Daylight Savings Time (DST) is annual time change that occurs in the spring in nearly 70 countries across the globe.

- In the US the change occurs on the second Sunday in March at 2:00 AM
  - At 2:00 AM the time skips ahead an hour to 3:00 AM

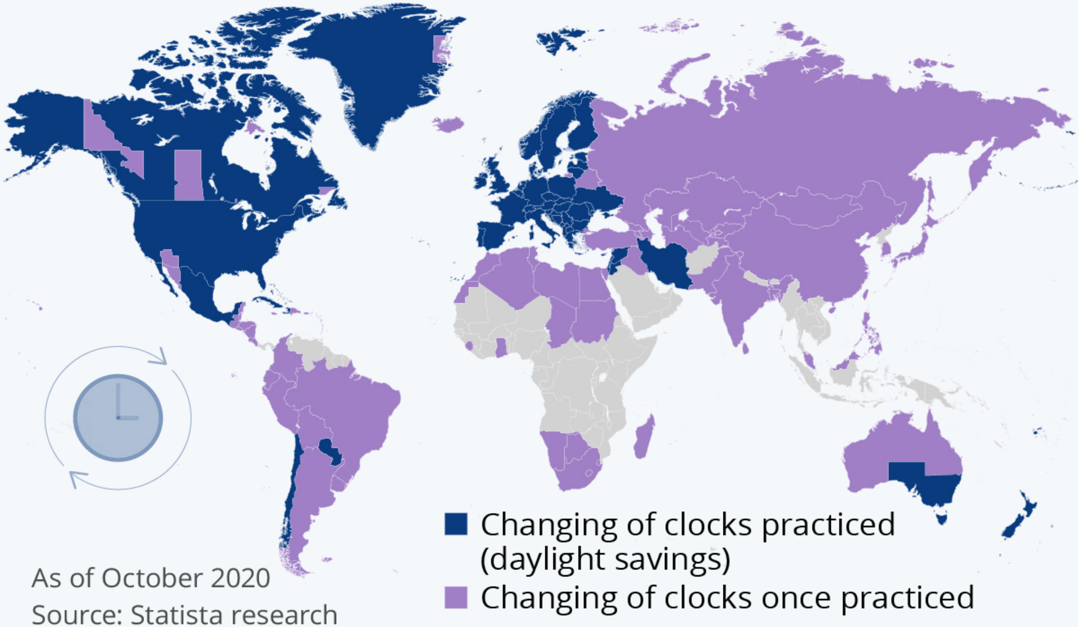
- It ends annually on the first Sunday in November at 2:00 AM
  - At 2:00 AM the time falls back an hour to 1:00 AM

# Background Research

- Medical and behavioral research suggests that the DST time changes lead to numerous negative outcomes.
- Spring change:
  - Increase in cardiovascular issues (stroke, heart attack, etc.)
  - Acute increase in mood disorders
  - Stock Volatility
  - Sleep loss
  - Traffic Accidents
- Fall Change
  - Sleep disruption
  - Mood disturbances & Suicides

# Which Countries Change the Clock?

Countries and regions which practice time change and those who have done so in the past


🟦 Changing of clocks practiced (daylight savings)
🟪 Changing of clocks once practiced

As of October 2020
Source: Statista research

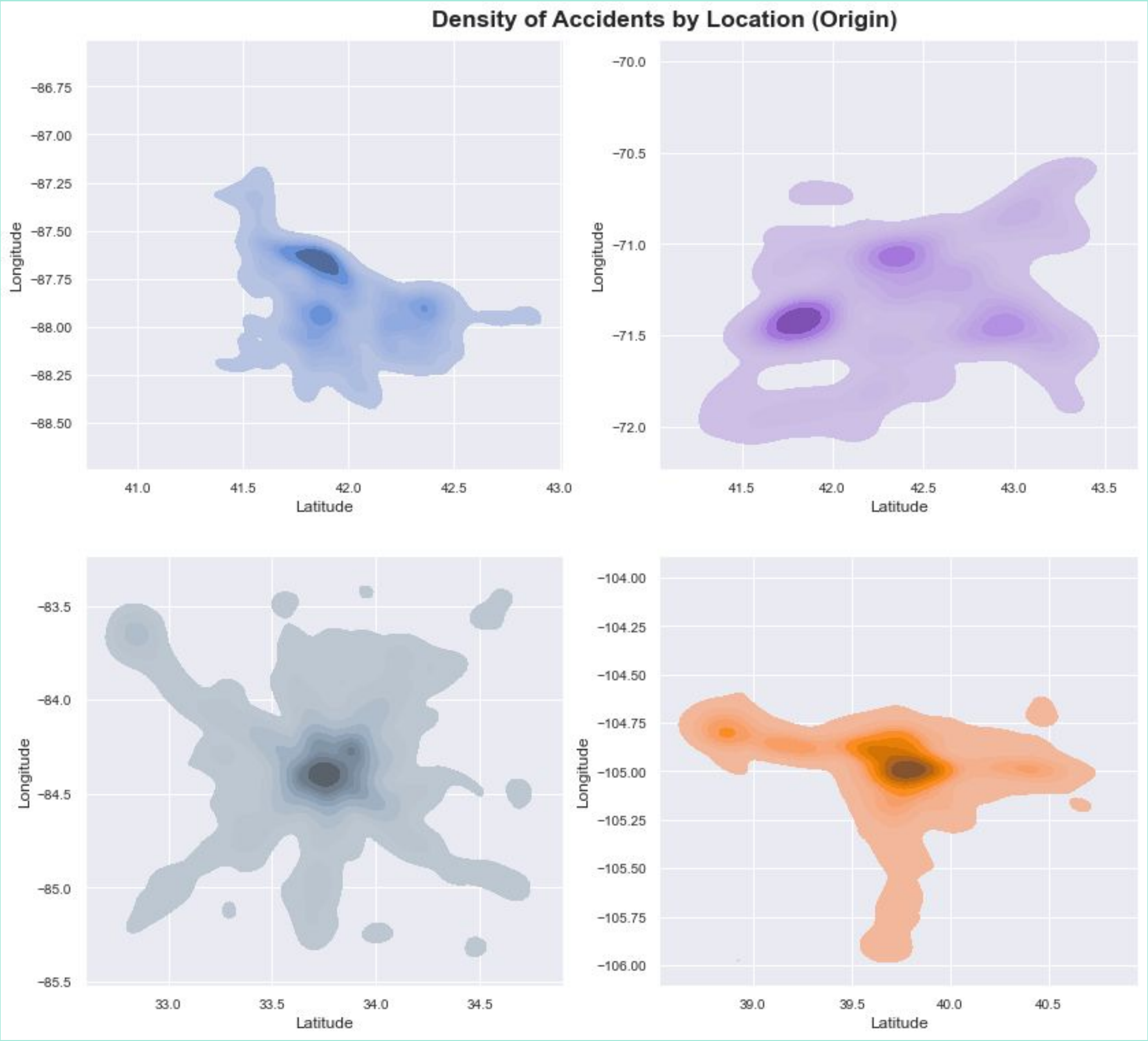
# Data Cleaning & EDA

Sobhan Moosavi’s US Accidents dataset (via kaggle https://www.kaggle.com/sobhanmoosavi/us-accidents) had over 1,516,064 accident entries from 2016 - 2020 and ranked the traffic impact of each accident recorded on a scale of 1 to 4. This was too much data to consider given our time and resource constraints, so we narrowed our scope to a few cities’ data (based on a range of latitude and longitude).

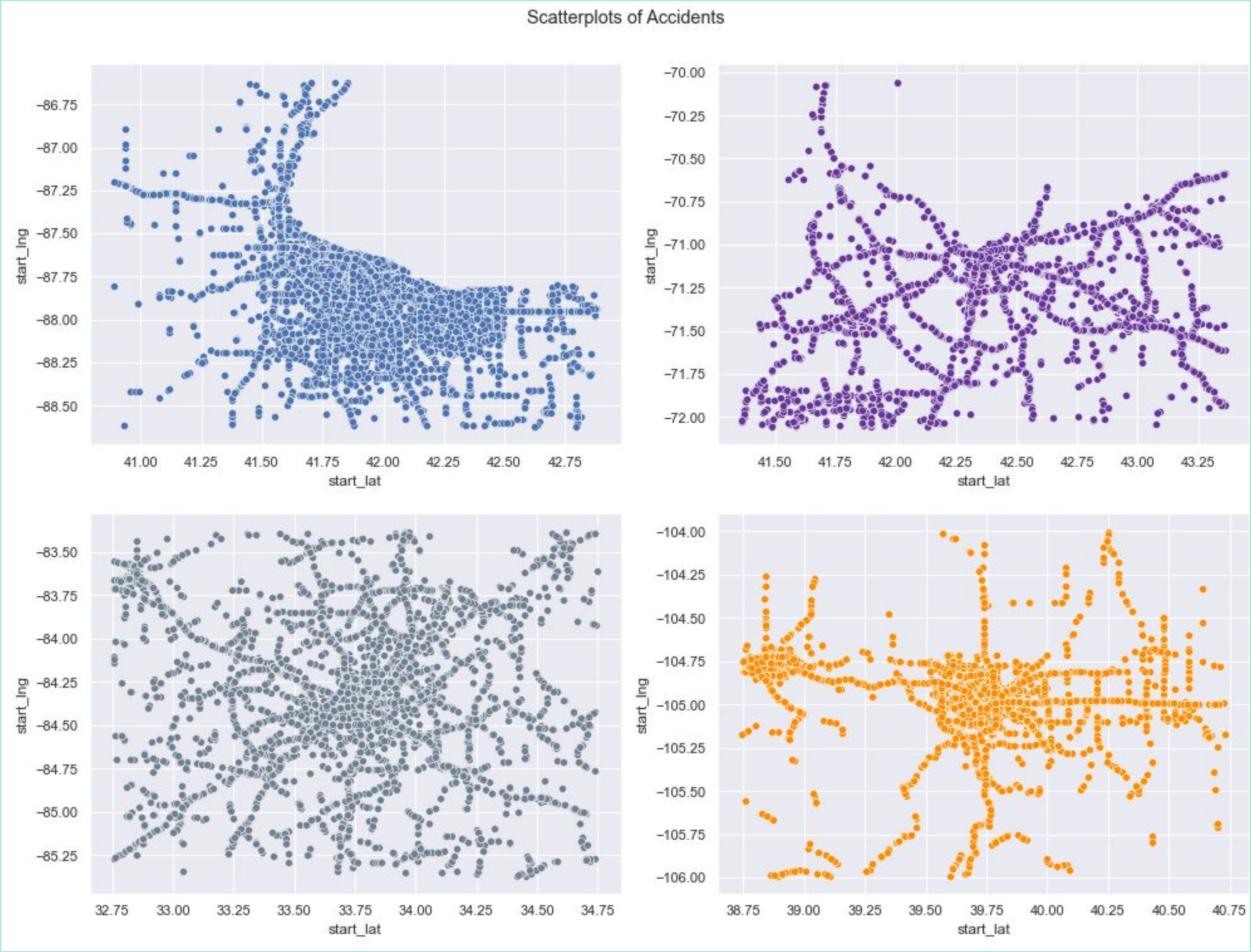
We considered the following cities:

- Atlanta
- Boston
- Chicago
- Denver

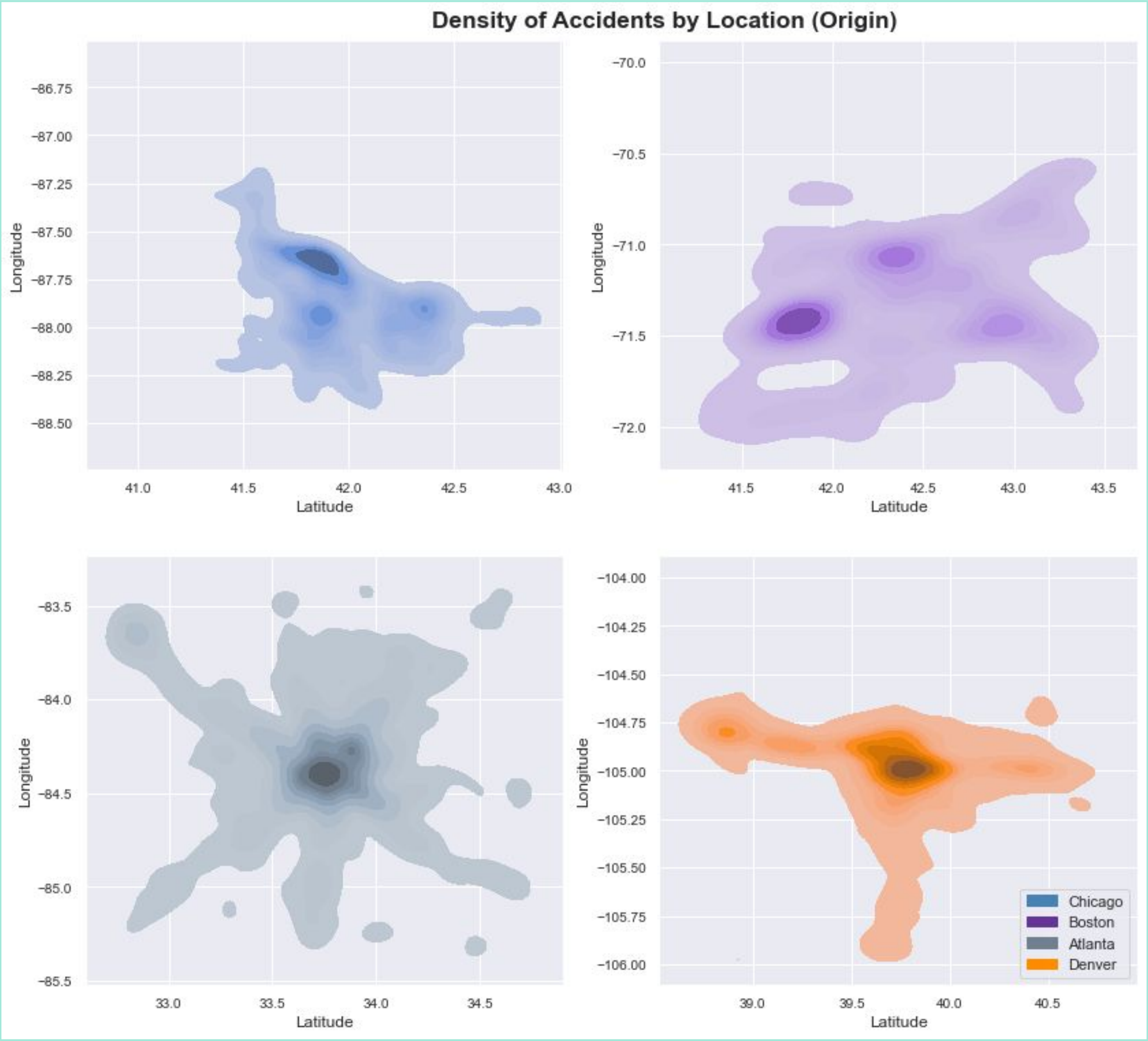
Chicago had the most available data so we decided to focus our efforts there. We built several functions in order to clean our data that could be extrapolated to the larger dataset.



Density of Accidents by Location (Origin)

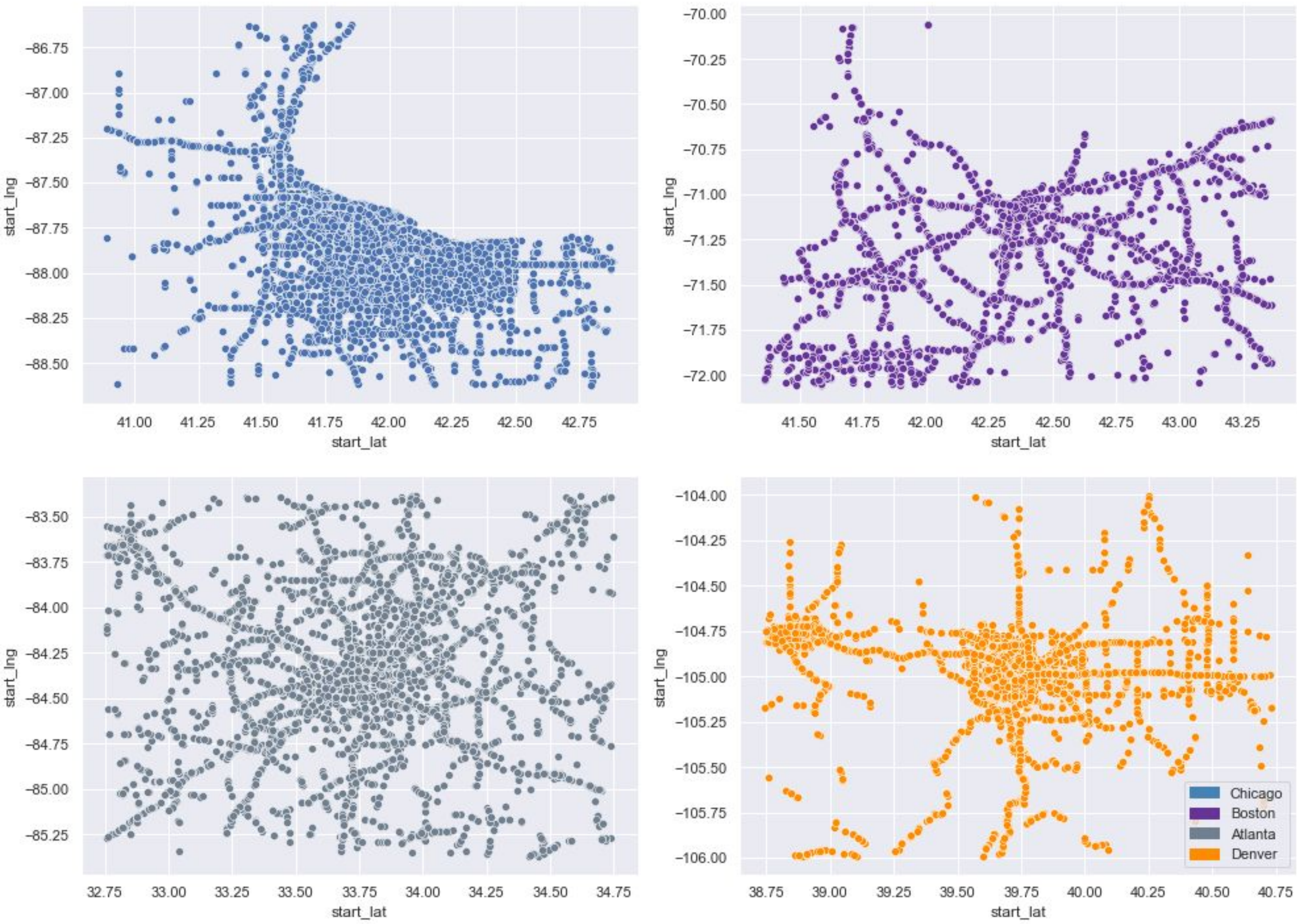


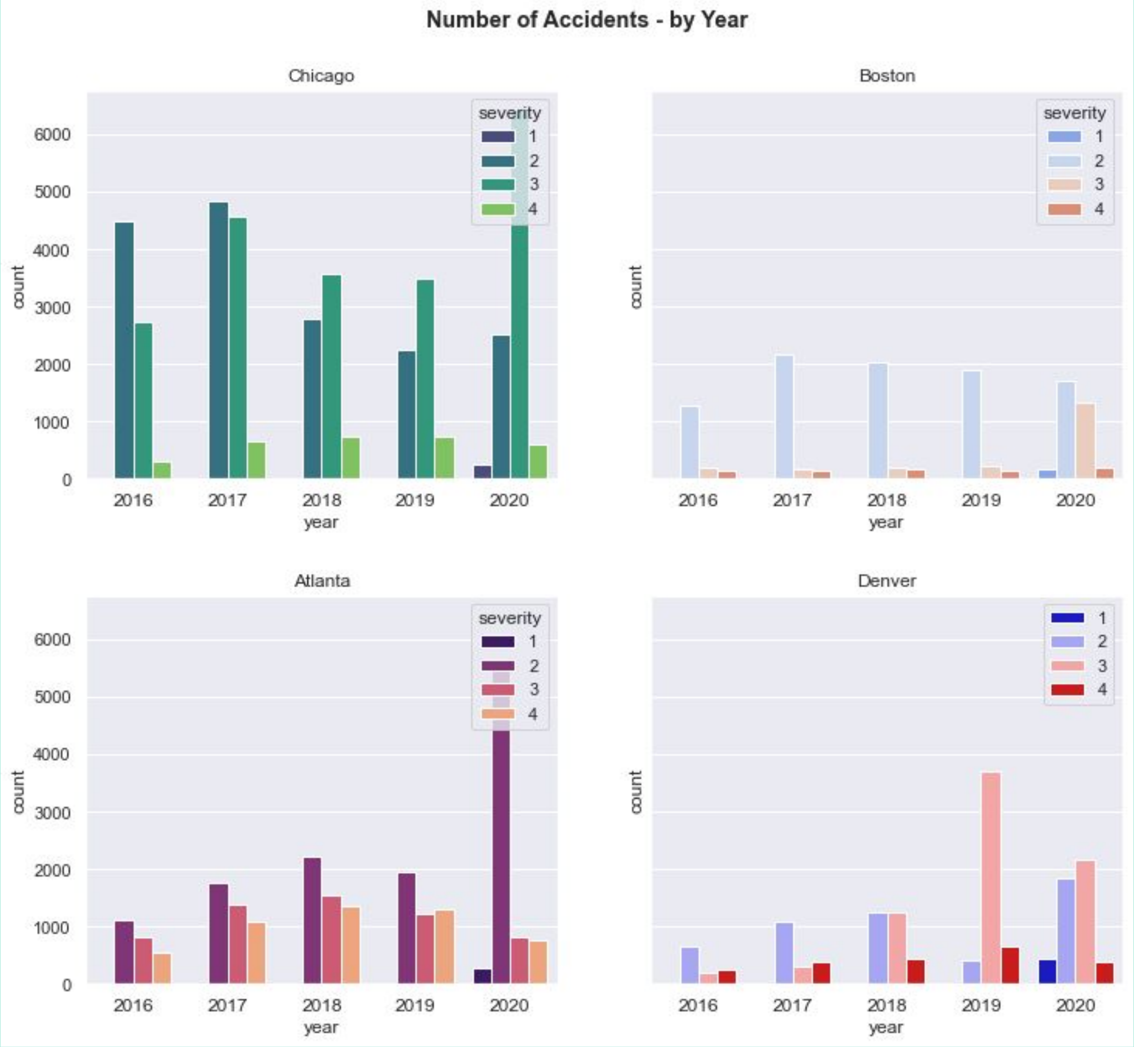
Scatterplots of Accidents



Density of Accidents by Location (Origin)

Scatterplots of Accidents




Number of Accidents - by Year
Chicago
Boston
Atlanta
Denver
severity
1
2
3
4
count
0
1000
2000
3000
4000
5000
6000
2016
2017
2018
2019
2020
year

# Data Cleaning & EDA

We observed our data types and instances of null values. The following columns were dropped outright and not used in modeling:

- 'id'
- 'description',
- 'number'
- 'street'
- 'city'
- 'county'
- 'state'
- ‘zipcode'
- 'country’
- 'timezone'
- 'airport_code'

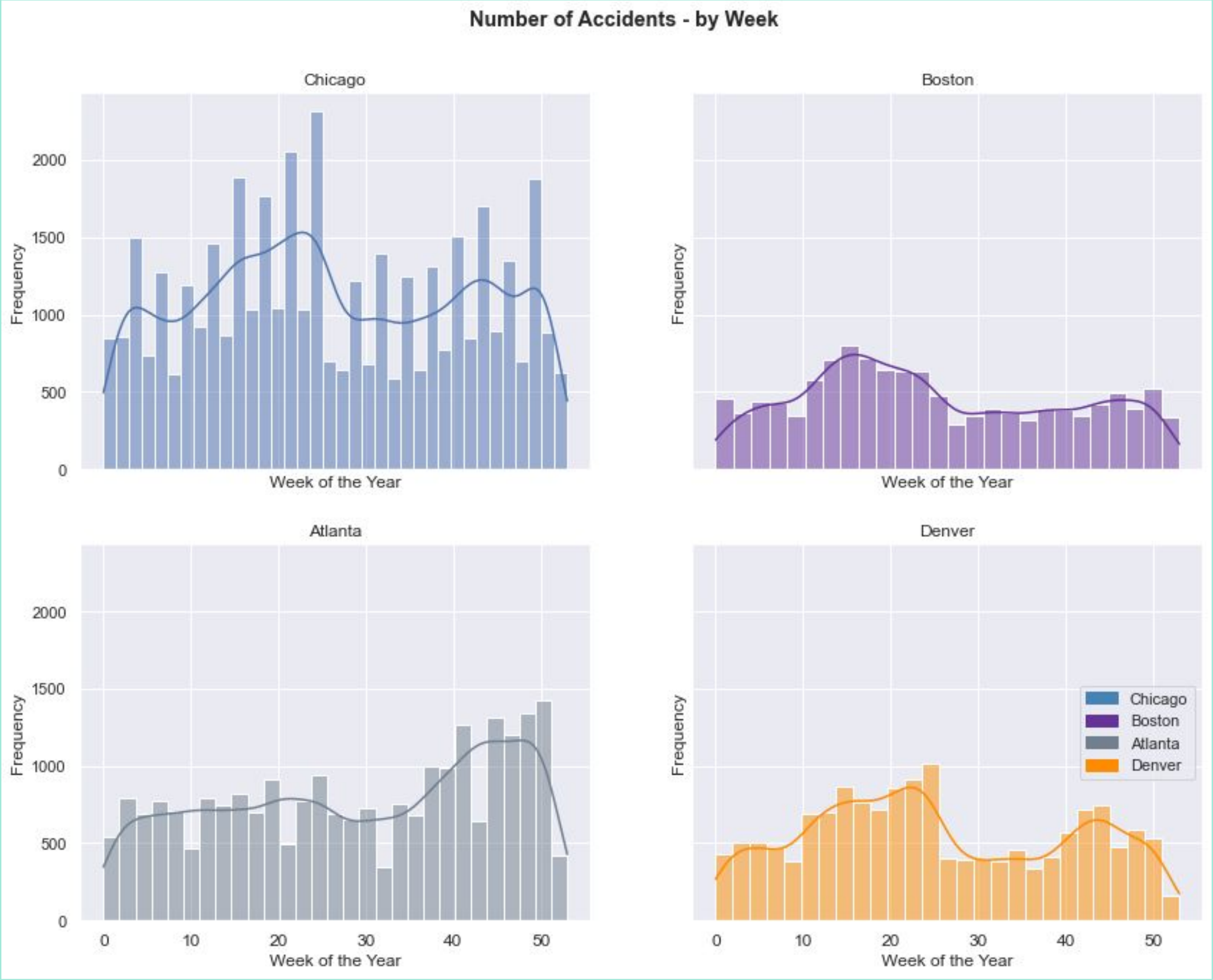
Note: ‘number’ represents street address and was missing 26593 values out of 42472 entries for Chicago.

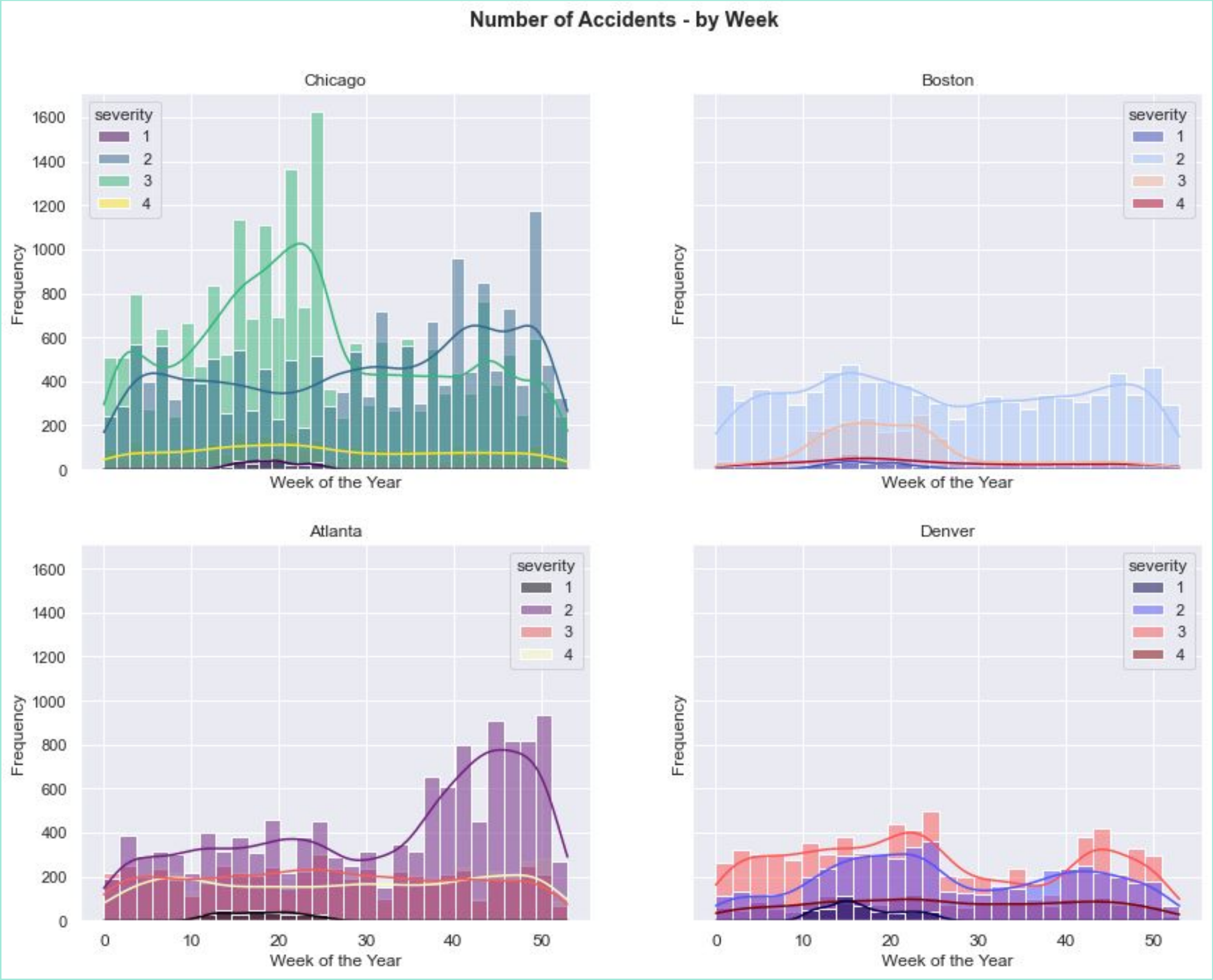
# Data Cleaning & EDA

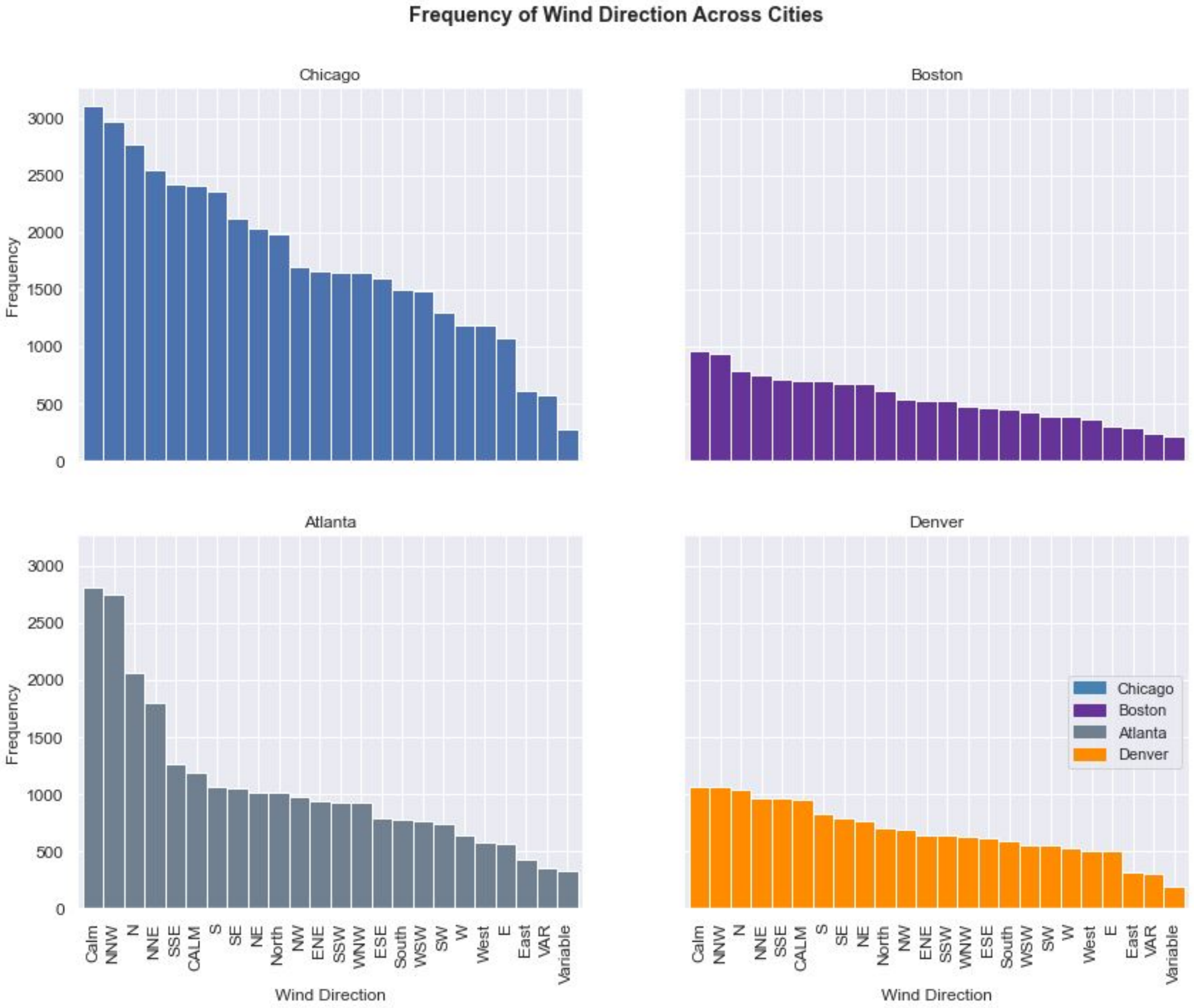
Using columns that were “datetime” type, the following columns were engineered:

- year
- month
- week
- day as a categorical variable
- hour as a categorical variable
- accident_duration as the difference of end_time and start_time
- start_time_ep (start_time in epoch format)
- end_time_ep (end_time in epoch format)
- weather_timestamp_ep (weather_timestamp in epoch format)

Any instances of null values in “datetime” entries were dropped.


Number of Accidents - by Week
Chicago
Boston
Atlanta
Denver
Frequency
Week of the Year
0
500
1000
1500
2000
0
10
20
30
40
50
Chicago
Boston
Atlanta
Denver


Number of Accidents - by Week
Chicago
Boston
Atlanta
Denver
severity
1
2
3
4
Frequency
0
200
400
600
800
1000
1200
1400
1600
Week of the Year
0
10
20
30
40
50


Frequency of Wind Direction Across Cities
Chicago
Boston
Atlanta
Denver
Frequency
Frequency
0
500
1000
1500
2000
2500
3000
Wind Direction
Wind Direction
Calm
NNW
N
NNE
SSE
CALM
S
SE
NE
North
NW
ENE
SSW
WNW
ESE
South
WSW
SW
W
West
E
East
VAR
Variable
Chicago
Boston
Atlanta
Denver

# Data Cleaning & EDA

The column wind_direction had values which represented the same thing and needed to be replaced:

- North and N
- South and S
- West and W
- East and E
- Calm and CALM
- Variable and VAR

This column was then dummified.

# Data Cleaning & EDA

Like with wind_direction, the column weather_condition had values which represented the same thing and needed to be replaced.

Regrettably, not all cities represent all possible weather_condition values so this was an involved process.

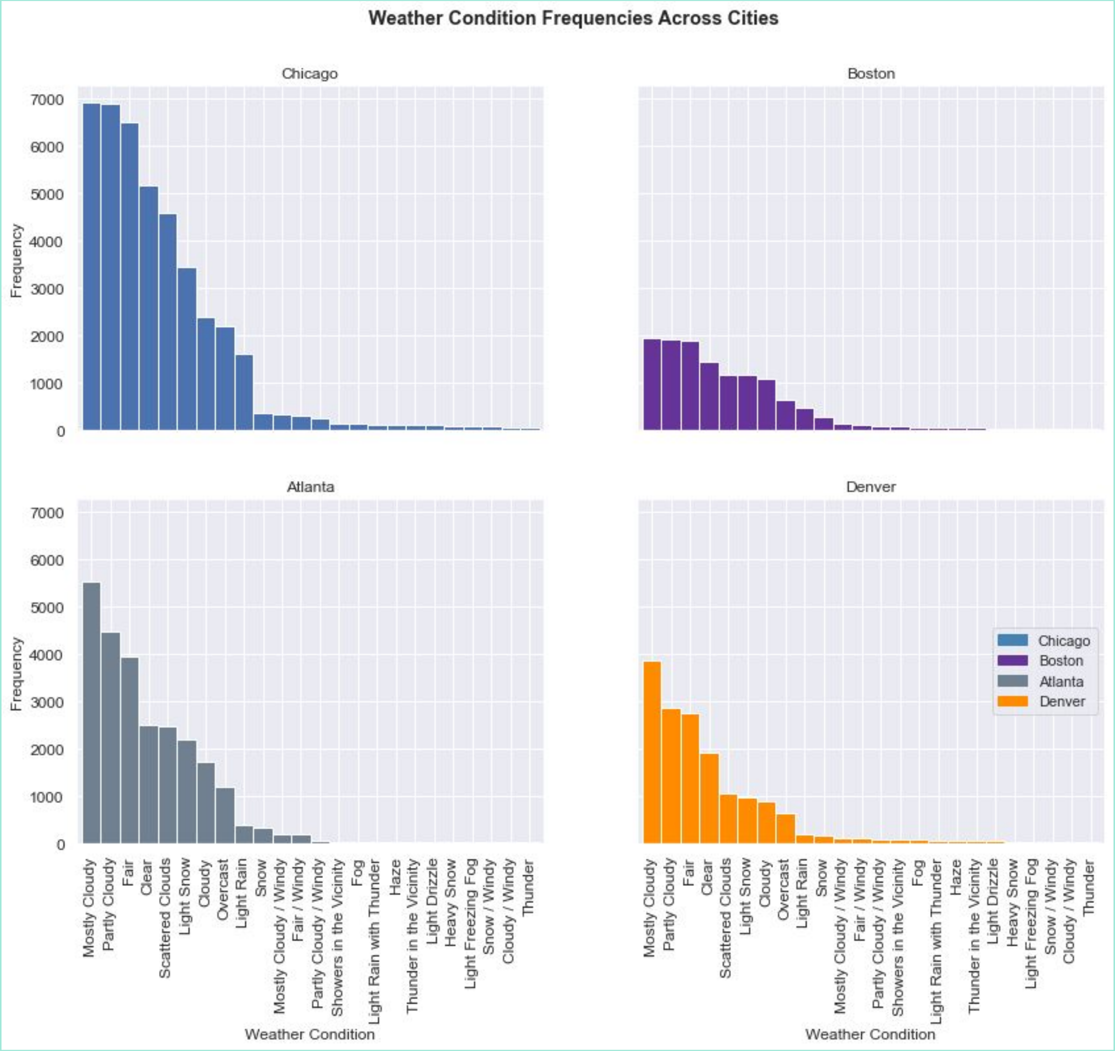
We created several functions to accomplish this goal ...

# Data Cleaning & EDA

We created functions which did the following:

1. Built a list of all possible weather_condition values in each city.
2. Built a function that created columns for each value of weather_condition not represented and merged all similar weather values (eg. Cloudy, Cloudy / Windy, Mostly Cloudy, and Mostly Cloudy / Windy) into one column (eg. Cloudy).

These functions were called in “mrclean” after the column was dummified (NaNs included).


Weather Condition Frequencies Across Cities
Chicago
Boston
Atlanta
Denver
Frequency
0
1000
2000
3000
4000
5000
6000
7000
Weather Condition
Mostly Cloudy
Partly Cloudy
Fair
Clear
Scattered Clouds
Light Snow
Cloudy
Overcast
Light Rain
Snow
Mostly Cloudy / Windy
Fair / Windy
Partly Cloudy / Windy
Showers in the Vicinity
Fog
Light Rain with Thunder
Haze
Thunder in the Vicinity
Light Drizzle
Heavy Snow
Light Freezing Fog
Snow / Windy
Cloudy / Windy
Thunder
Chicago
Boston
Atlanta
Denver

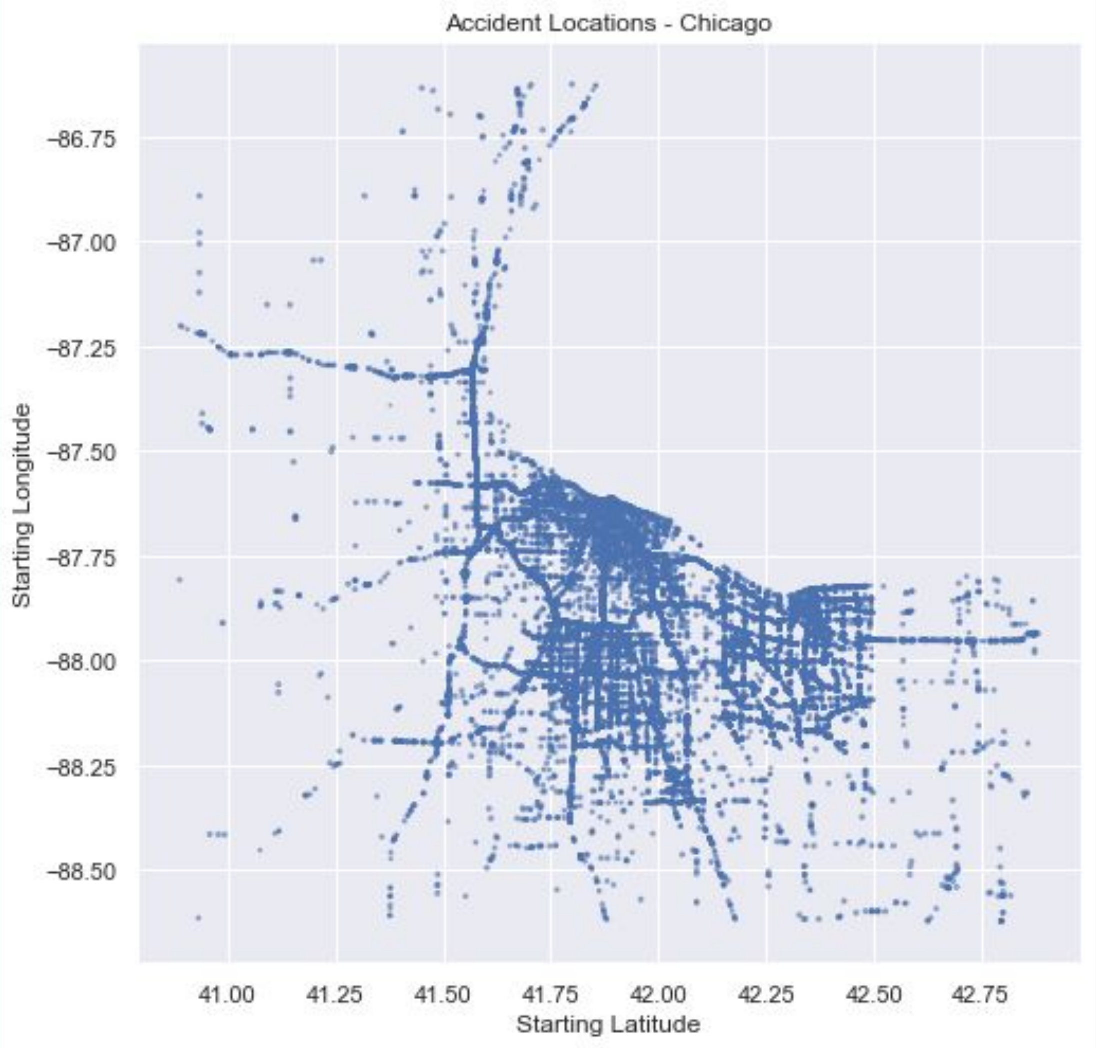
# Data Cleaning & EDA

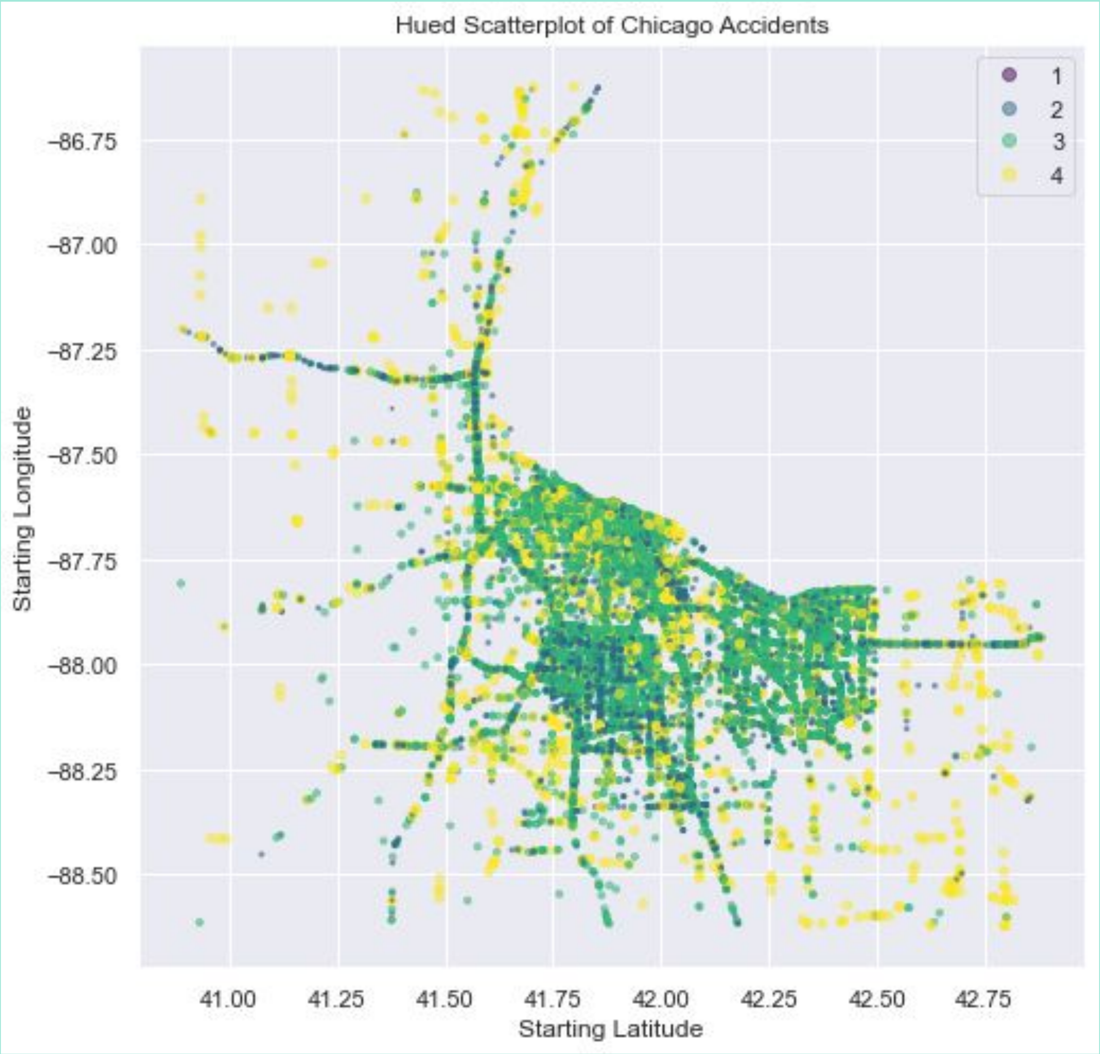
When completing EDA on cleaned data, further feature engineering was done to create a new column is_DST.

We assigned 0s and 1s to our entries based on whether or not the accident took place the week following the change to DST.

There were a total of 9 (nine) weeks of accidents to consider:

- November 6 - 12, 2016
- March 12 - 18, 2017 and November 5 - 11, 2017
- March 11 - 17, 2018 and November 4 - 10, 2018
- March 10 - 16, 2019 and November 3 - 9, 2019
- March 8 - 14, 2020 and November 2 - 8, 2020


Accident Locations - Chicago
Starting Longitude
-86.75
-87.00
-87.25
-87.50
-87.75
-88.00
-88.25
-88.50
41.00
41.25
41.50
41.75
42.00
42.25
42.50
42.75
Starting Latitude


Hued Scatterplot of Chicago Accidents
1
2
3
4
-86.75
-87.00
-87.25
-87.50
-87.75
-88.00
-88.25
-88.50
Starting Longitude
41.00
41.25
41.50
41.75
42.00
42.25
42.50
42.75
Starting Latitude

# Modeling

We wanted to look at the feature importance of the is_DST feature to determine how impactful Daylight Savings could be when predicting accident severity. We focused in on the Chicago data for modeling purposes since it was our largest data set among the cities we observed.

We tried a few different model types to see which produced the best results:

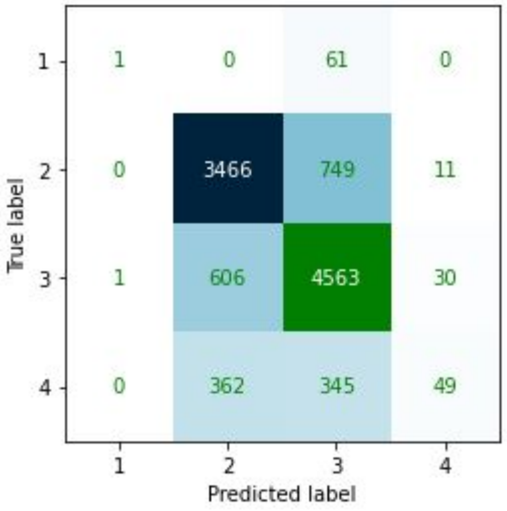
- Decision Tree
- Random Forest
- BAG
- XGBoost

# Baseline Scores

Severity label 3 is our majority class at 50.76%. We'll use this as our baseline score to beat when evaluating model performance but also will keep an eye on precision in relation to severity 3 and 4 accidents.

| Severity Label | Percentage of all labels |
|---|---|
| 1 | 0.61% |
| 2 | 41.25% |
| 3 | 50.76% |
| 4 | 7.38% |

# Decision Tree Model



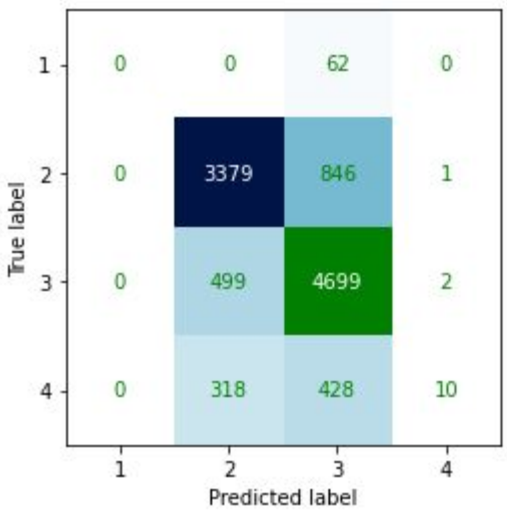
| True label \ Predicted label | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 1 | 0 | 61 | 0 |
| 2 | 0 | 3466 | 749 | 11 |
| 3 | 1 | 606 | 4563 | 30 |
| 4 | 0 | 362 | 345 | 49 |

| | |
|---|---|
| is_DST coefficient | 0.00000 |

| Severity | True Distribution | Predicted Distribution |
|---|---|---|
| 1 | 0.61% | 0.001% |
| 2 | 41.25% | 43.28% |
| 3 | 50.76% | 55.82% |
| 4 | 7.38% | 0.85% |

| | |
|---|---|
| Training Accuracy | 0.799 |
| Testing Accuracy | 0.789 |
| Precision (Severity 3) | 0.80 |
| Precision (Severity 4) | 0.55 |

# Random Forest Model



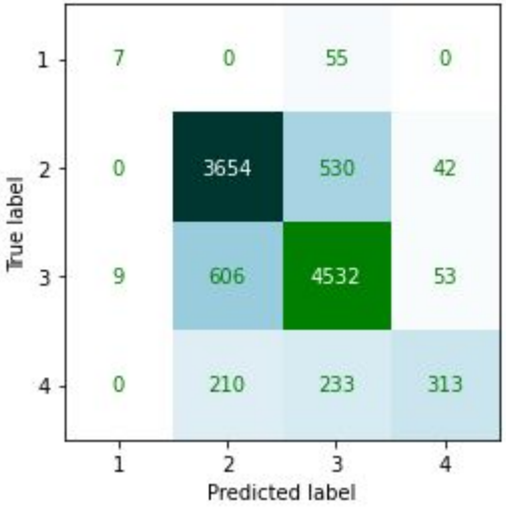
|  | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 0 | 0 | 62 | 0 |
| 2 | 0 | 3379 | 846 | 1 |
| 3 | 0 | 499 | 4699 | 2 |
| 4 | 0 | 318 | 428 | 10 |

|  |  |
|---|---|
| is_DST coefficient | 0.00058 |

| Severity | True Distribution | Predicted Distribution |
|---|---|---|
| 1 | 0.61% | 0.00% |
| 2 | 41.25% | 40.96% |
| 3 | 50.76% | 58.91% |
| 4 | 7.38% | 0.13% |

|  |  |
|---|---|
| Training Accuracy | 0.800 |
| Testing Accuracy | 0.788 |
| Precision (Severity 3) | 0.78 |
| Precision (Severity 4) | 0.77 |

# BAG Model



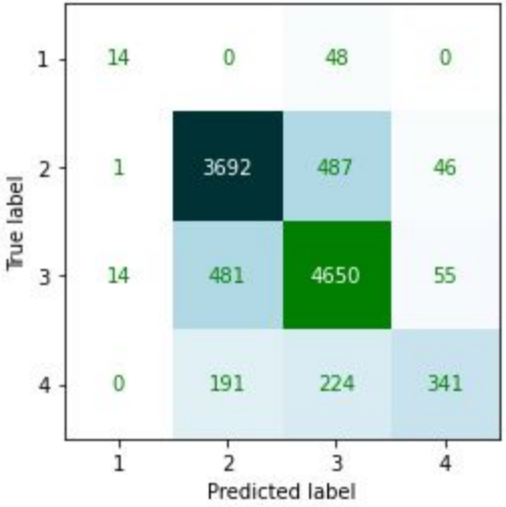
| True label \ Predicted label | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 7 | 0 | 55 | 0 |
| 2 | 0 | 3654 | 530 | 42 |
| 3 | 9 | 606 | 4532 | 53 |
| 4 | 0 | 210 | 233 | 313 |

| | |
|---|---|
| is_DST coefficient | 0.00614 |

| Severity | True Distribution | Predicted Distribution |
|---|---|---|
| 1 | 0.61% | 0.16% |
| 2 | 41.25% | 43.85% |
| 3 | 50.76% | 51.85% |
| 4 | 7.38% | 4.13% |

| | |
|---|---|
| Training Accuracy | 0.991 |
| Testing Accuracy | 0.827 |
| Precision (Severity 3) | 0.84 |
| Precision (Severity 4) | 0.78 |

# XGBoost Model



| True label \ Predicted label | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 14 | 0 | 48 | 0 |
| 2 | 1 | 3692 | 487 | 46 |
| 3 | 14 | 481 | 4650 | 55 |
| 4 | 0 | 191 | 224 | 341 |

| Severity | True Distribution | Predicted Distribution |
|---|---|---|
| 1 | 0.61% | 0.02% |
| 2 | 41.25% | 43.28% |
| 3 | 50.76% | 55.82% |
| 4 | 7.38% | 0.88% |

| | |
|---|---|
| Training Accuracy | 0.913 |
| Testing Accuracy | 0.849 |
| Precision (Severity 3) | 0.86 |
| Precision (Severity 4) | 0.77 |

# Research Hypotheses

- $H_0$: There is not a statistically significant difference in the average severity of accidents in Chicago in the weeks before, during, and after the spring time change.
- $H_a$: There is a statistically significant difference in the average severity of accidents in Chicago in the weeks before, during, and after the spring time change.
- The hypotheses were the same for the fall time change.

# ANOVA on Chicago Accident Severity for Spring Time Change

- To test our hypotheses on the impact of DST on accident severity we ran an ANOVA on week 10 (before DST), week 11 (DST), and week 12 (after DST).
- The results of the ANOVA gave strong evidence that there is a significant difference in the severity of traffic accidents in the week before DST, the week of DST, and after the week of DST.
- There were no significant differences for the fall time change.

| | **Sum_sq** | **df** | **F stat** | **PR(>F)** |
|---|---|---|---|---|
| **Week** | 6.234790 | 2.0 | 8.688793 | 0.000174 |
| **Residual** | 778.201214 | 2169.0 | NaN | NaN |

# Comparisons of each week for Spring Time Change in Chicago

The results of the multiple comparisons tests suggest a significant difference in the severity scores:

- The week before DST and the week of DST
- The week of DST and the week after DST.

| Group 1 | Group 2 | Stat | pval | pval_corr | Reject H0 |
|---|---|---|---|---|---|
| Week Before | Week of Spring DST | 4.0866 | 0.0 | 0.0001 | True |
| Week Before | Week After | 1.7884 | 0.0739 | 0.2058 | False |
| Week of Spring DST | Week After | -2.4028 | 0.0164 | 0.0483 | True |

# Conclusions and Recommendations

- The end of Daylight Savings in the Fall has no statistically relevant effect on average severity of traffic impact in any cities we observed
- The start of Daylight Savings in the Spring has a large effect on the average severity of traffic impact in all cities we observed

Because the impact is so significant at the start of Daylight Savings Sigmoids Data Science can confidently recommend that Rideshare Company ® begin lobbying to abolish Daylight Savings altogether.

# Streamlit App

We created an app that will accept gps coordinates of a reported accident. It will then get the current local weather data and make a prediction on how severely the accident will impact traffic!

Let’s take a look!

# Limitations

- Individual computing power prevented processing of full data set
- Accidents of severity 1 are likely under reported and under represented. This limited our ability to model this label as accurately
- Individual cities see impacts at different rates making it more difficult to paint a broader picture of Daylight Savings impact
- Dependency incompatibility for certain python packages that could increase streamlit user-friendliness

# Future Directions

- NLP on the dropped feature “description”  to see if a determination could be made on accident locations. For example, a service road can be beneath a highway and accidents could happen in both locations but it’s more likely for an accident to occur at higher speeds on the highway.
- Further ANOVA to include the frequency of accidents and impact of Day/Night on accident severity during the week of DST.
- Additional modeling on cities other than Chicago to see if the models’ performance holds across geographic locations.
- Contact the creator of the dataset and ask what the quality of “accident severity” measures (how long the duration of each severity is).
- Integrate GeoPandas into the Streamlit app in order to have ‘drag-and-drop’ functionality with the map.

# References

- Data used in project from Sobhan Moosavi's US Accidents Dataset
- Infographics are from SlidesGo Traffic Light Infographics
- Statista "Which countries change the clock?" Chart
- Latitude and Longitude Finder https://www.latlong.net/

# Questions?