Capstone project

Predicting the severity of a car accident

Ferrer García Erick 09/15/2020

Problematic

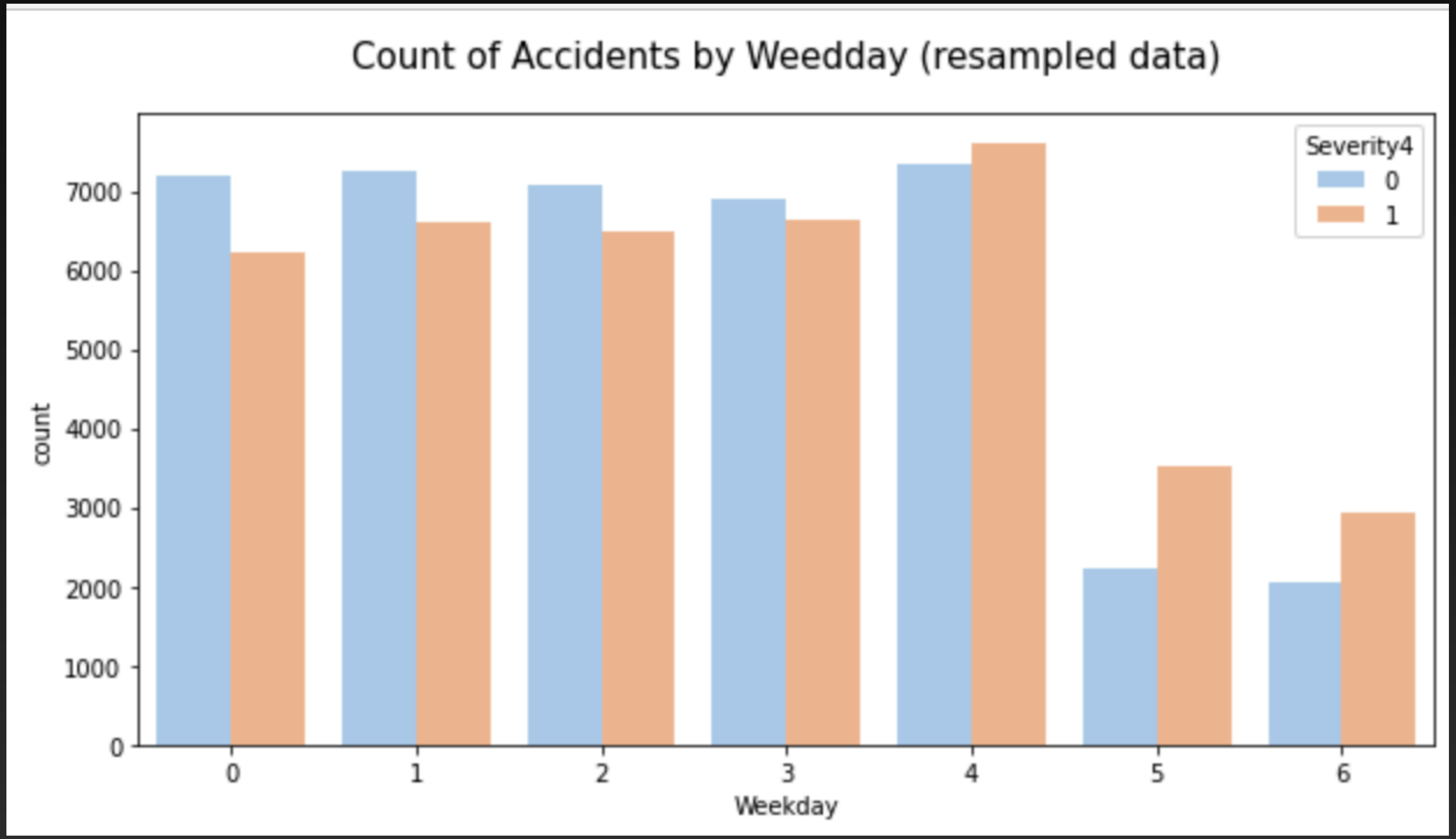
Predicting the severity of an accident with data science

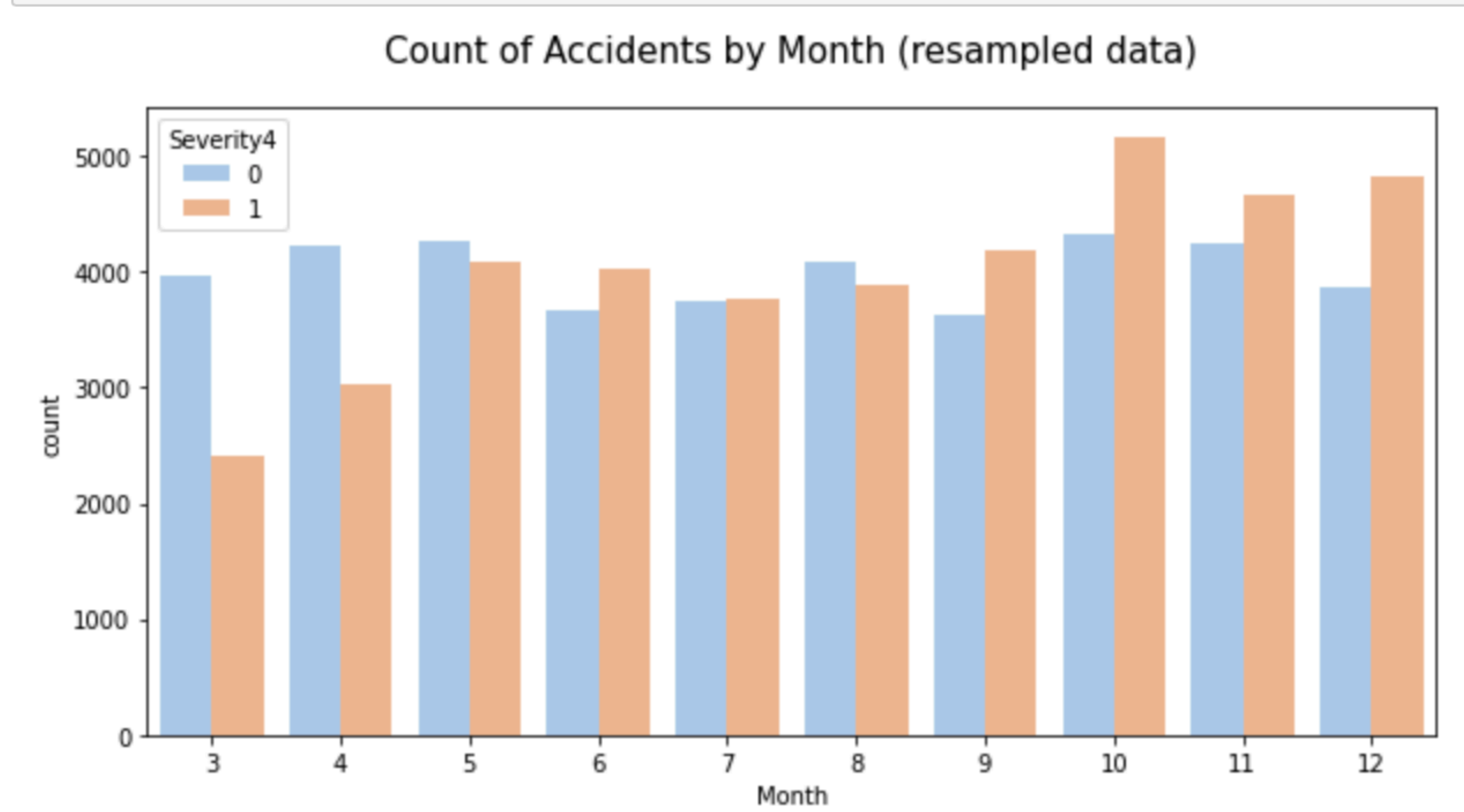
Data acquisition and cleaning

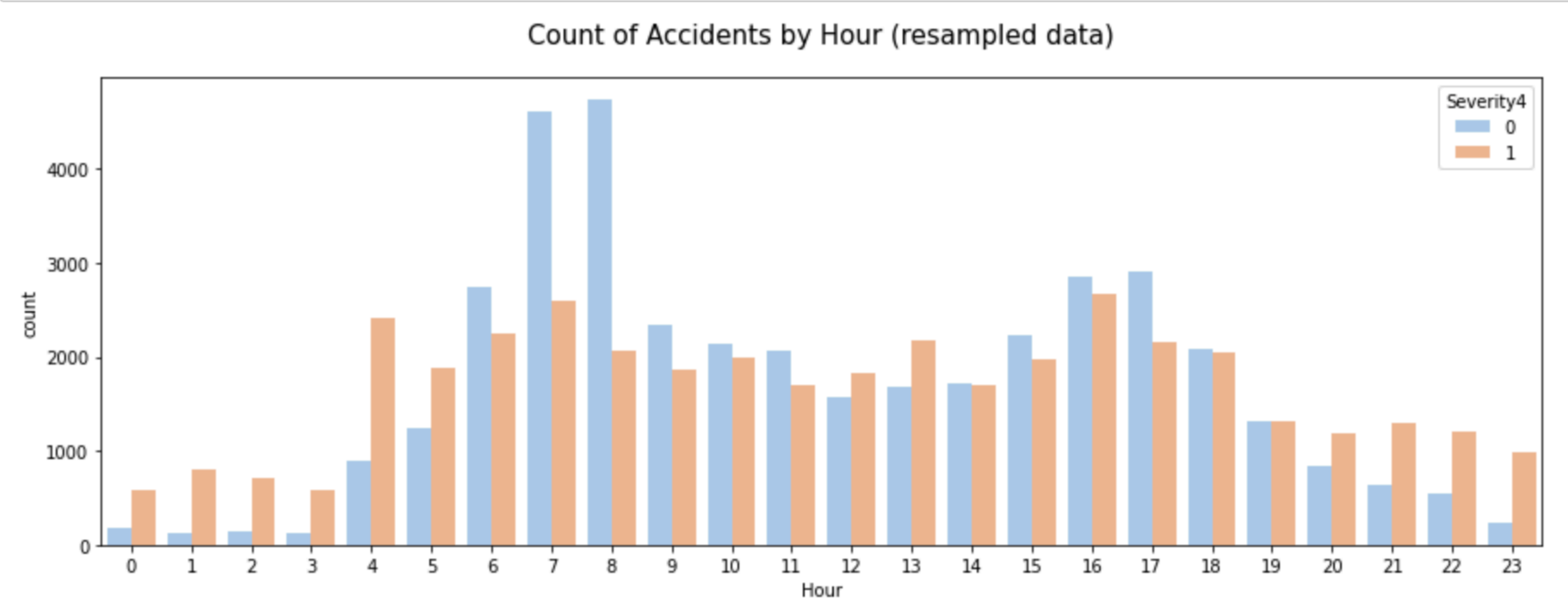
- This is a countrywide car accident dataset, which covers 49 states of the USA. The accident data are collected from February 2016 to June 2020, using two APIs that provide streaming traffic incident (or event) data.
- Came from 2 sources, MapQuest and Bing.
- The shape of the data is 3,513,617.
- Duplicate, and empty rows were drop, same with the data that doesn’t give a lot of info, or redundant data .
- Normalice the date time, and adjust the wind and the weather conditions.
- And do an one hot encoding, to some features.

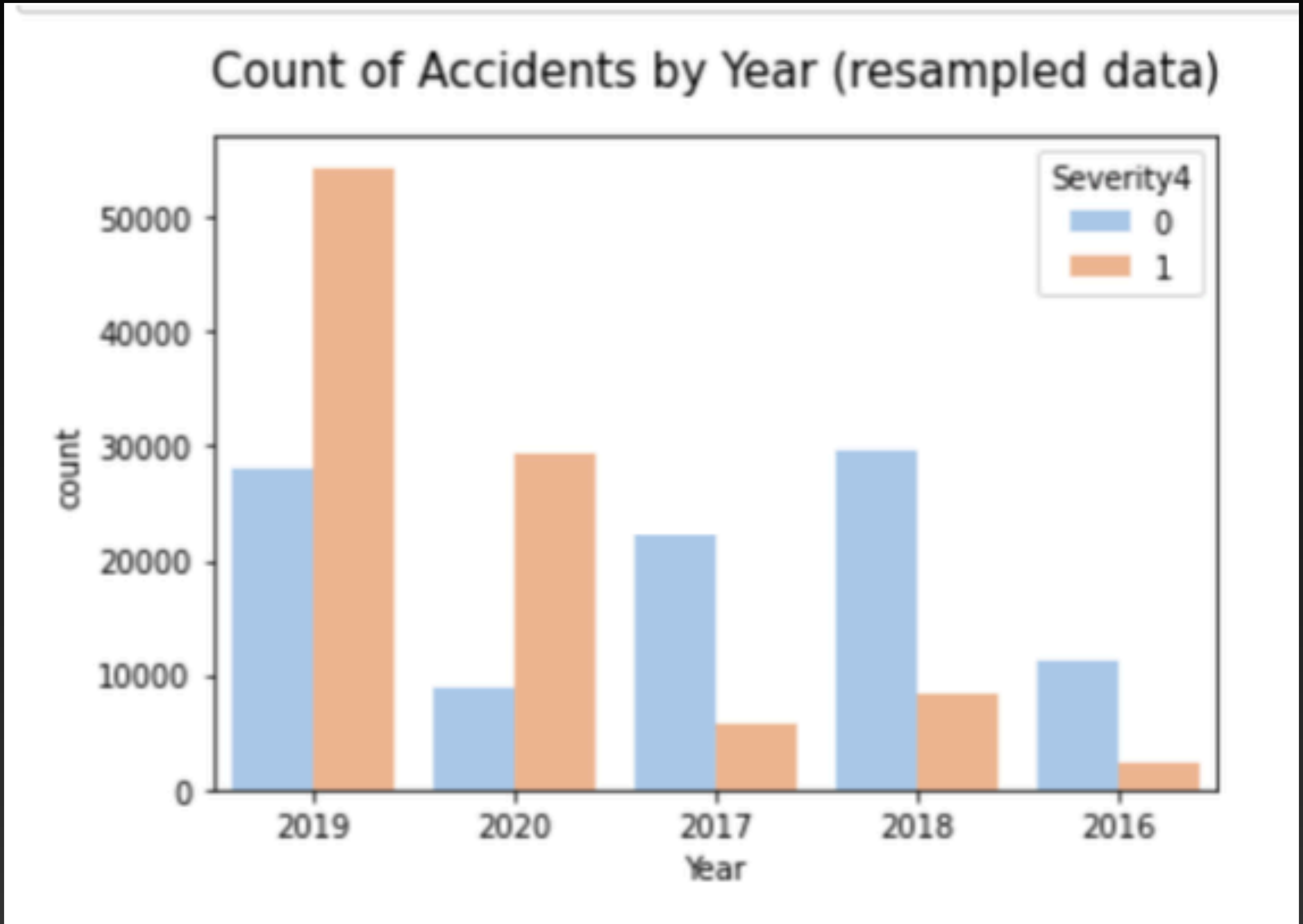
Data Analysis

Correlation with time features

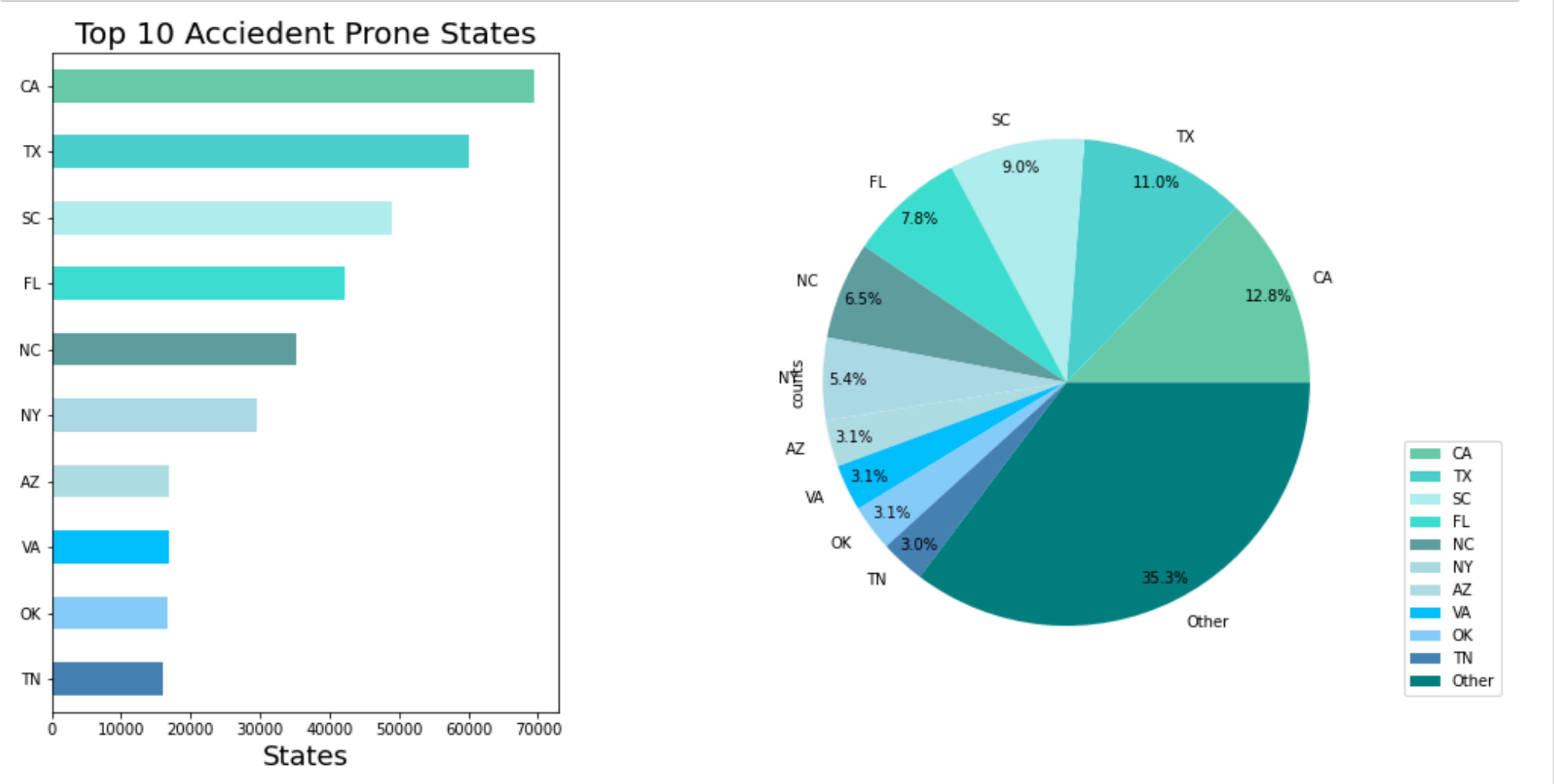




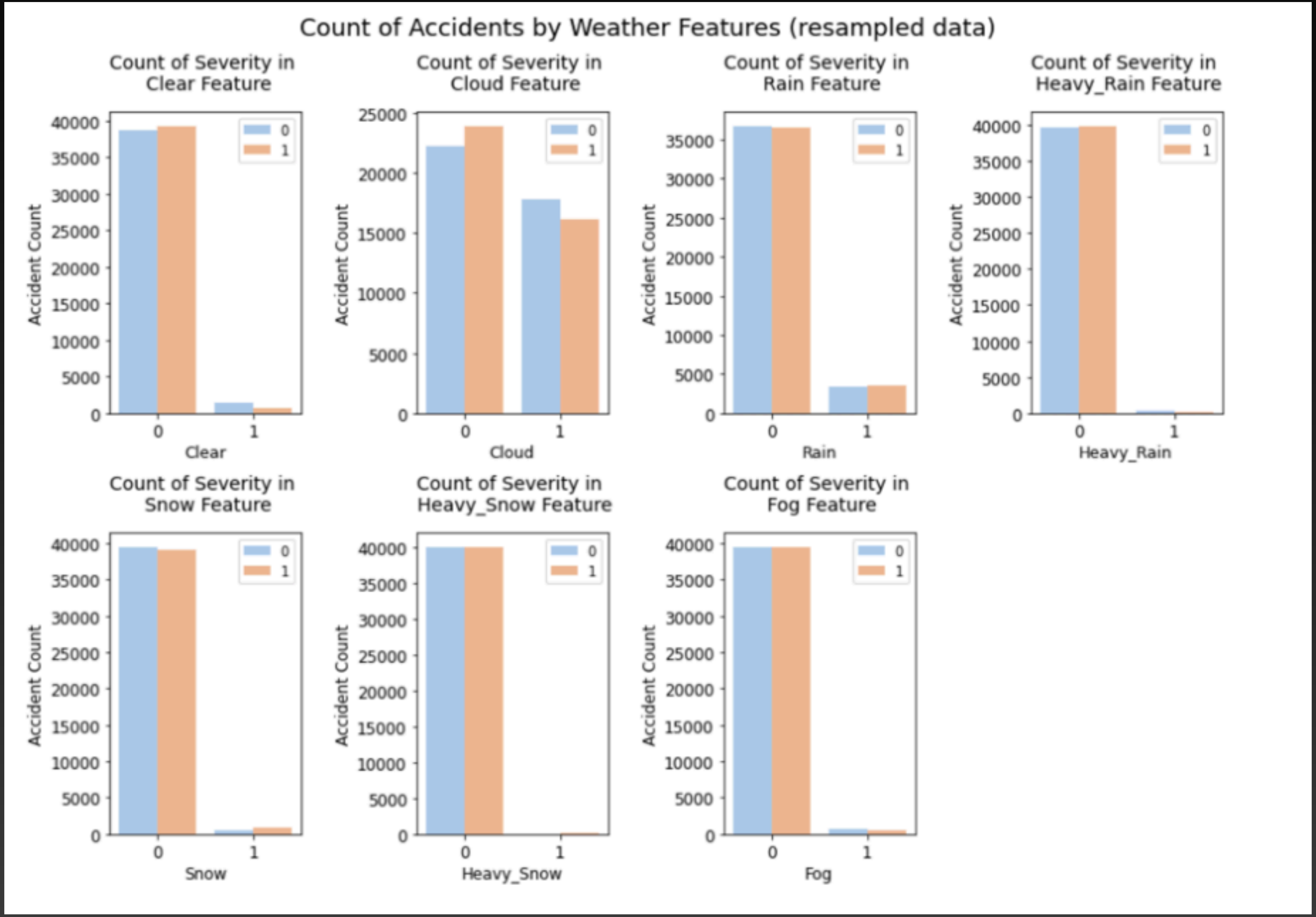

Count of Accidents by Hour (resampled data)
Severity4
0
1
count
4000
3000
2000
1000
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Hour


Count of Accidents by Year (resampled data)
Severity4
0
1
count
50000
40000
30000
20000
10000
0
2019
2020
2017
2018
2016
Year

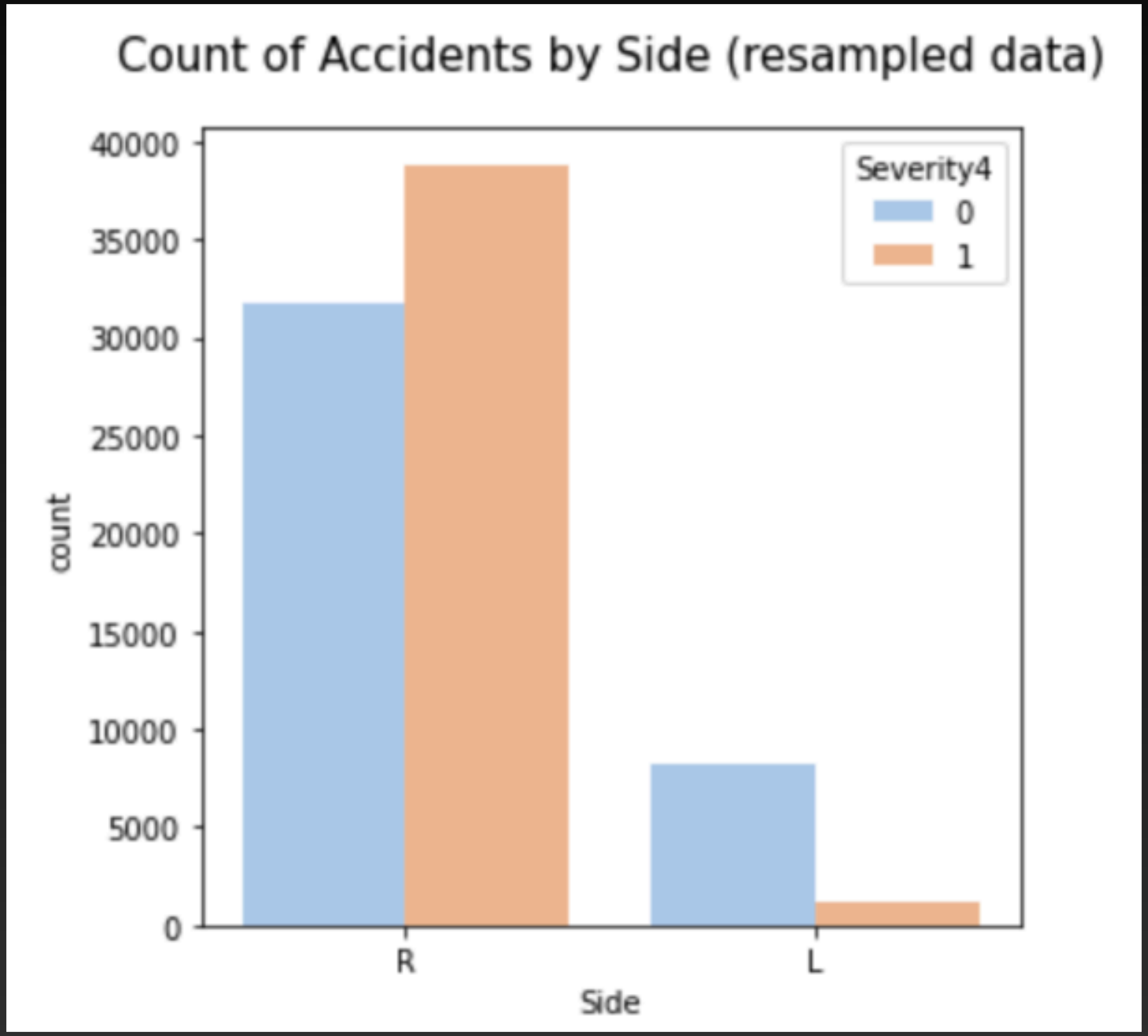
Correlation with the state



Correlation with the weather

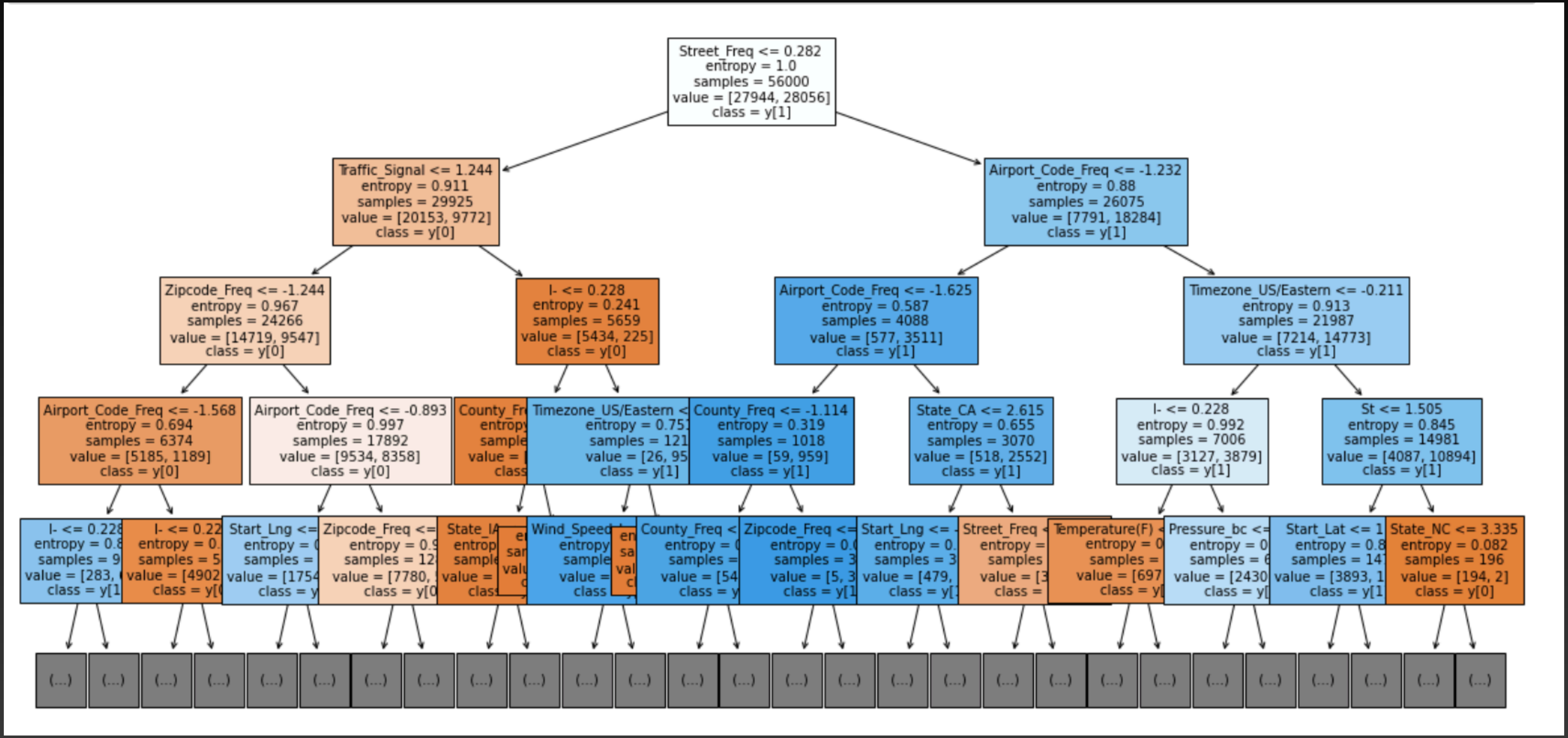


Correlation with the side

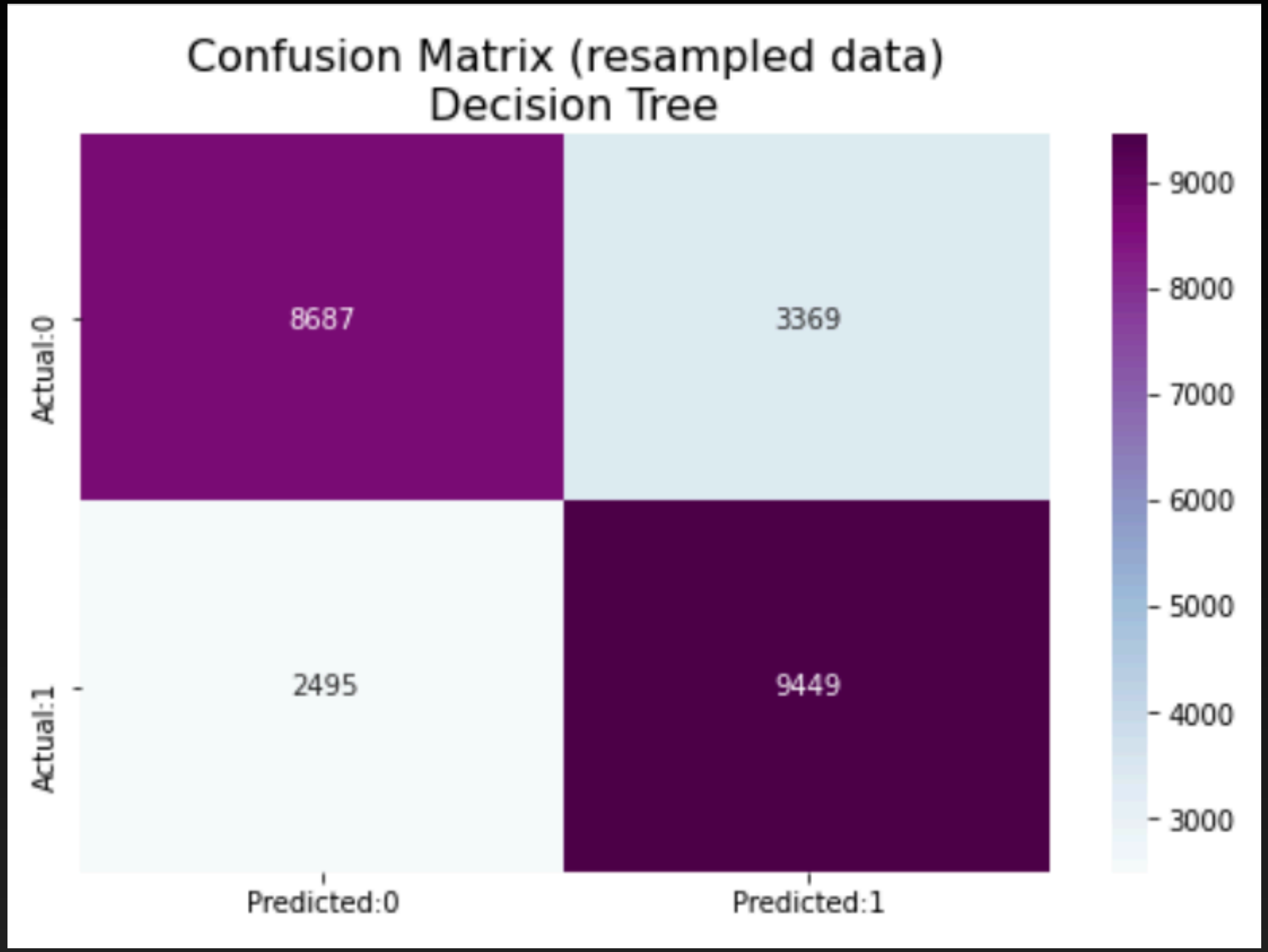


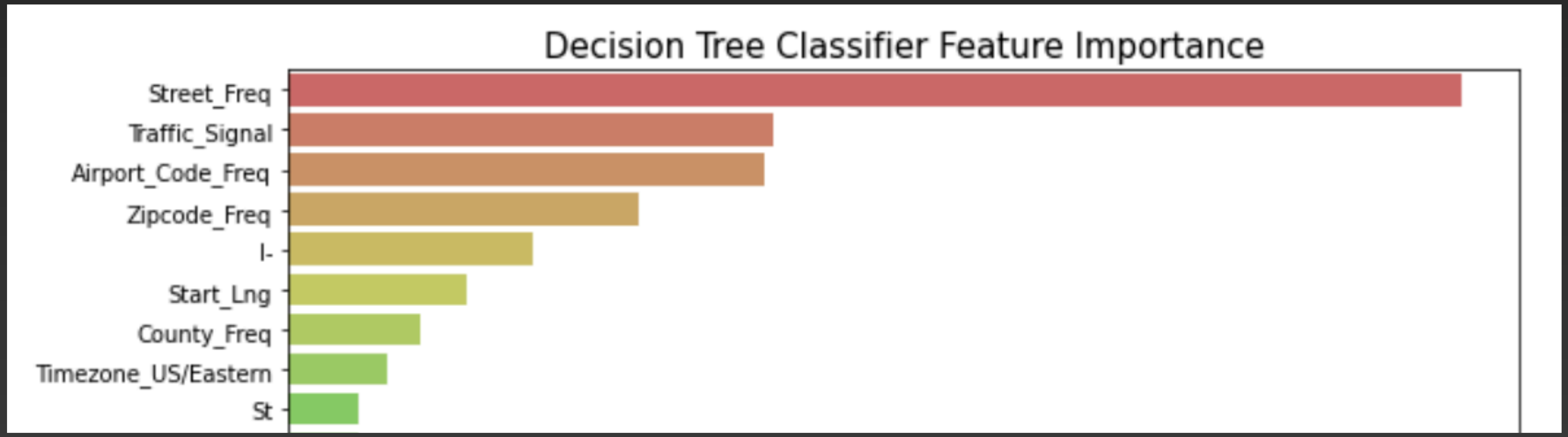
Modeling

Results of decision tree



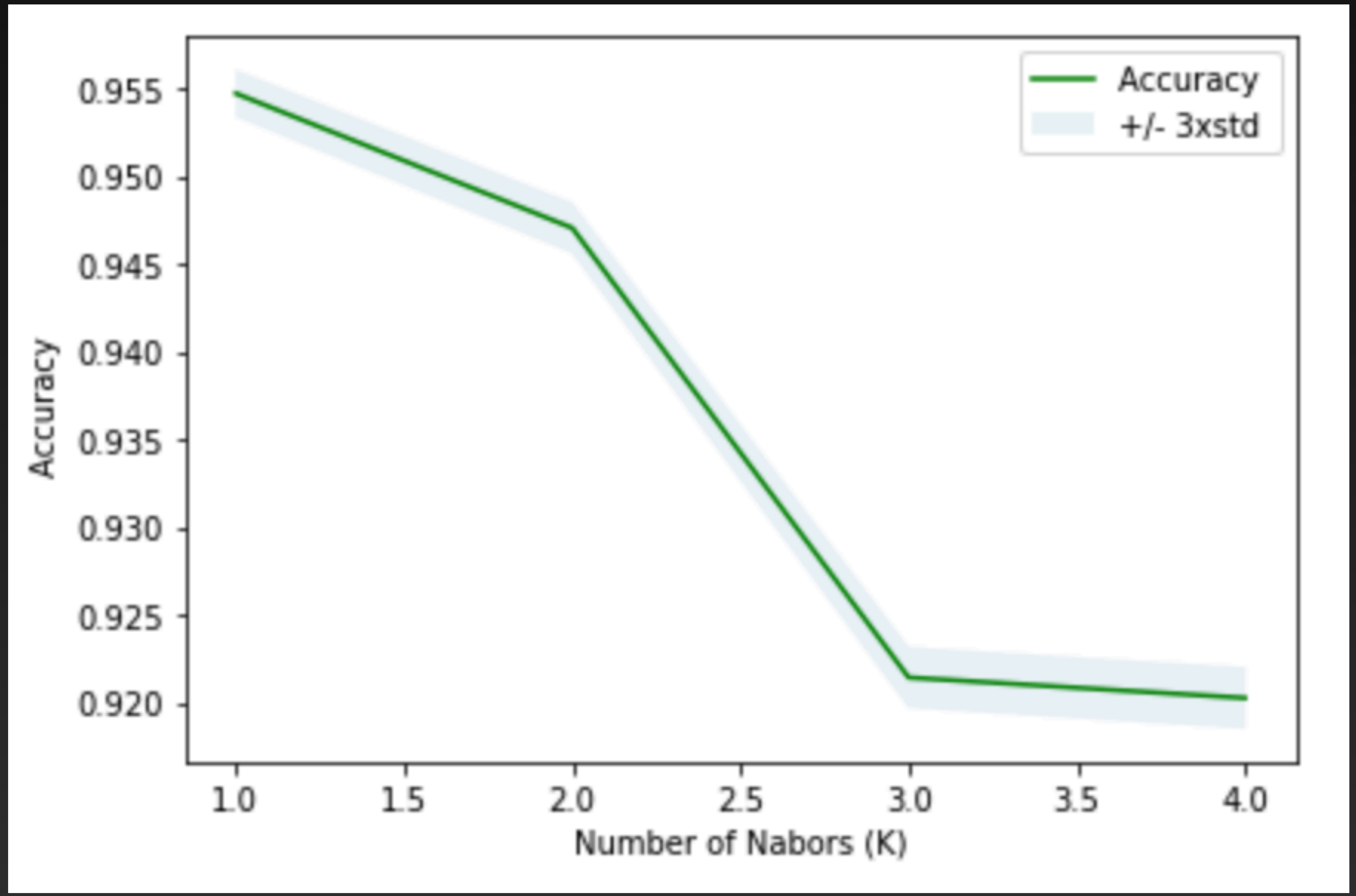
- The accuracy of the prediction with a decision tree is 76%
- We can see that the 5 features more important in the modeling.
- The data is already adjusted so this is why we don't do more in this step.
- The Street Frequency is the most importan, i guess because with more cars more accidents.

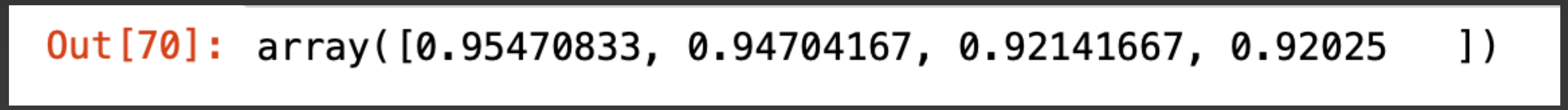




Results on KNN

- In this algorithm we can see that the best K is 1.
- This maybe is a bad idea to use it.
- The process is too slow to do it with more than 5 k's.
- the accuracy drop in k = 2, so lets not choose this algorithm.

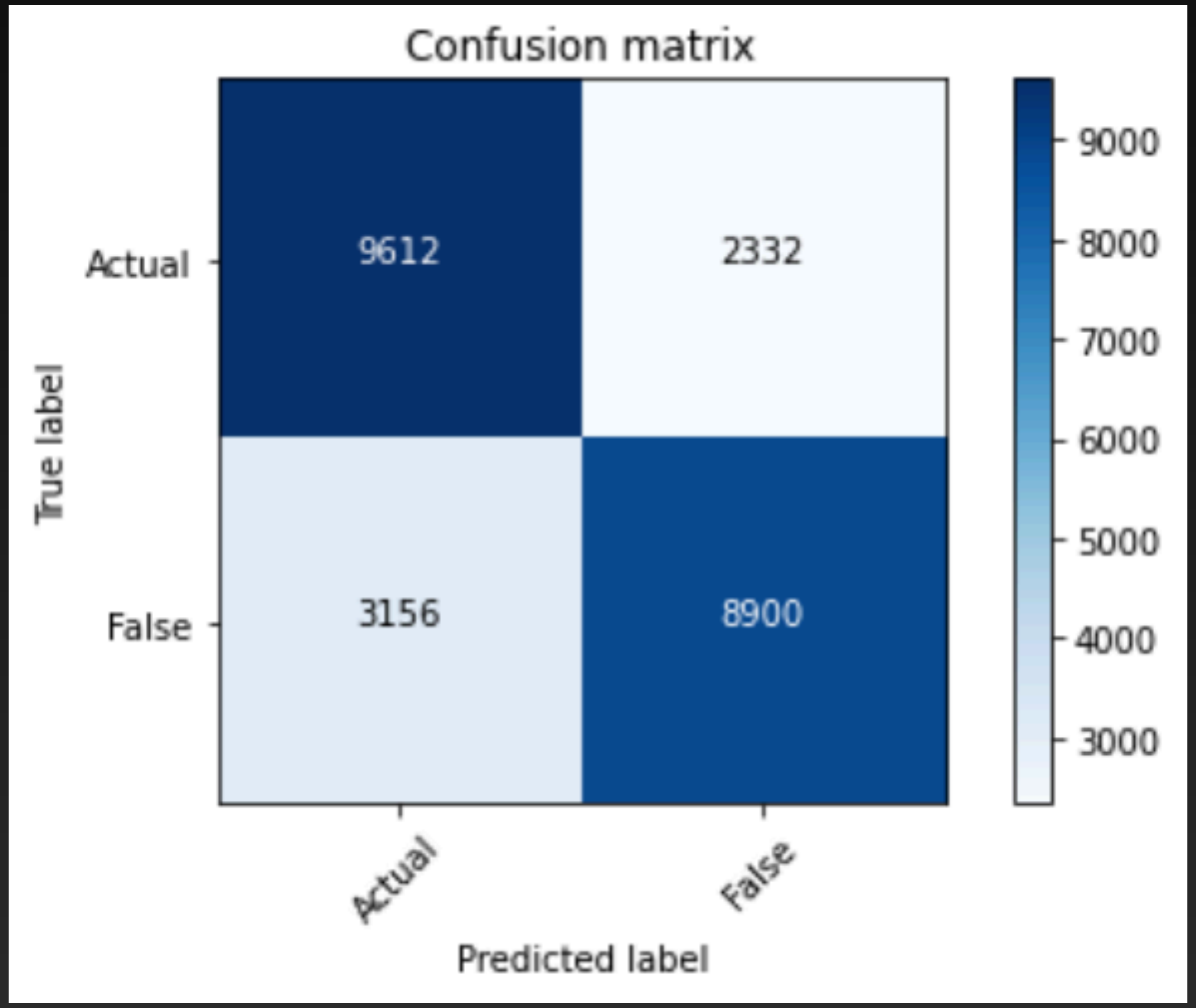


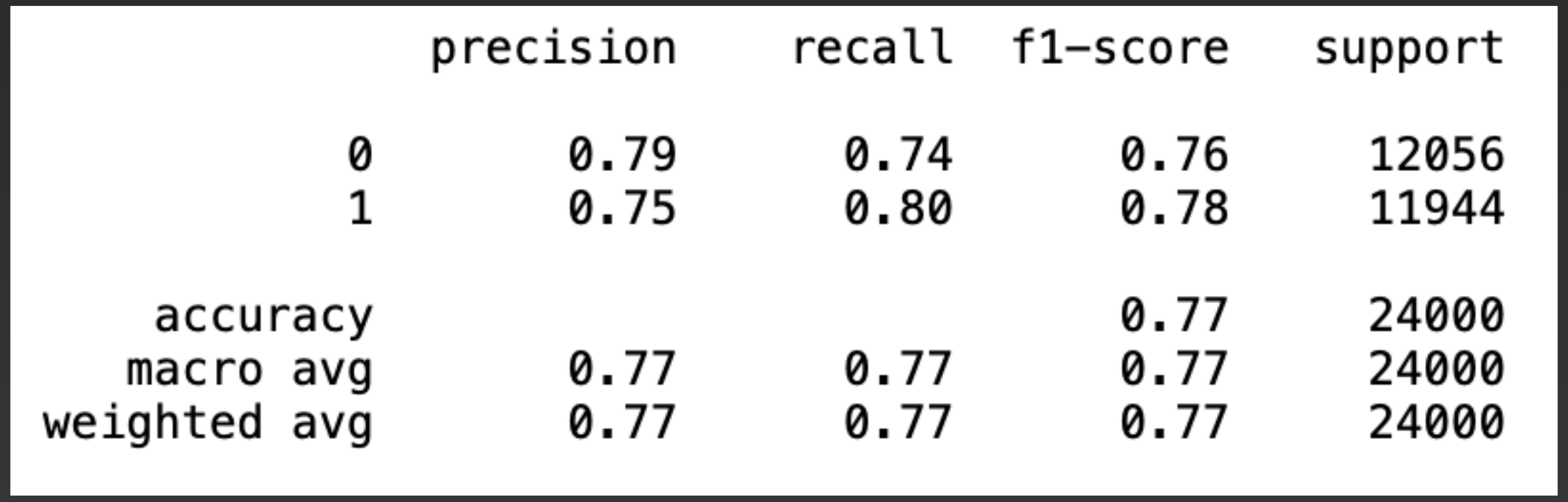


```
Out[70]: array([0.95470833, 0.94704167, 0.92141667, 0.92025    ])
```

Results on Logistic Regression

- The accuracy on this algorithm is 77% when it's time to predict.
- This was the fastest algorithm.





```
              precision    recall  f1-score   support

           0       0.79      0.74      0.76     12056
           1       0.75      0.80      0.78     11944

    accuracy                           0.77     24000
   macro avg       0.77      0.77      0.77     24000
weighted avg       0.77      0.77      0.77     24000
```

Results and conclusion

As we can see both os them are almost the same, but in this process i will choose decision tree, because it give me more information such as the top 5 features with most impact of an accident like:

- The street frequency
- Airport code
- Traffic signal
- Zipcode
- The kind of street in this case I

So for now, i decide stay with decision tree algorithm, the logistic regression was algo a good algorithm but the information that i can plot about it is less graphic, and the result was almost the same as the decision tree, now on the KNN i already said that i don't think is a good algorithm for this case because even when we normalize the data the result about the best K is still 1, this and the time to process the algorithm give reasons to avoid it.

Thank you for the attention