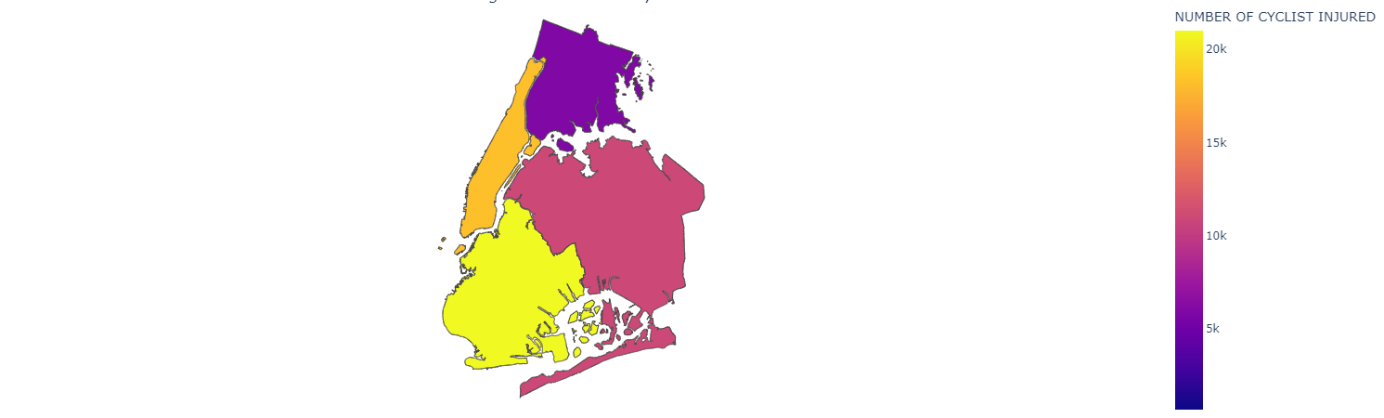
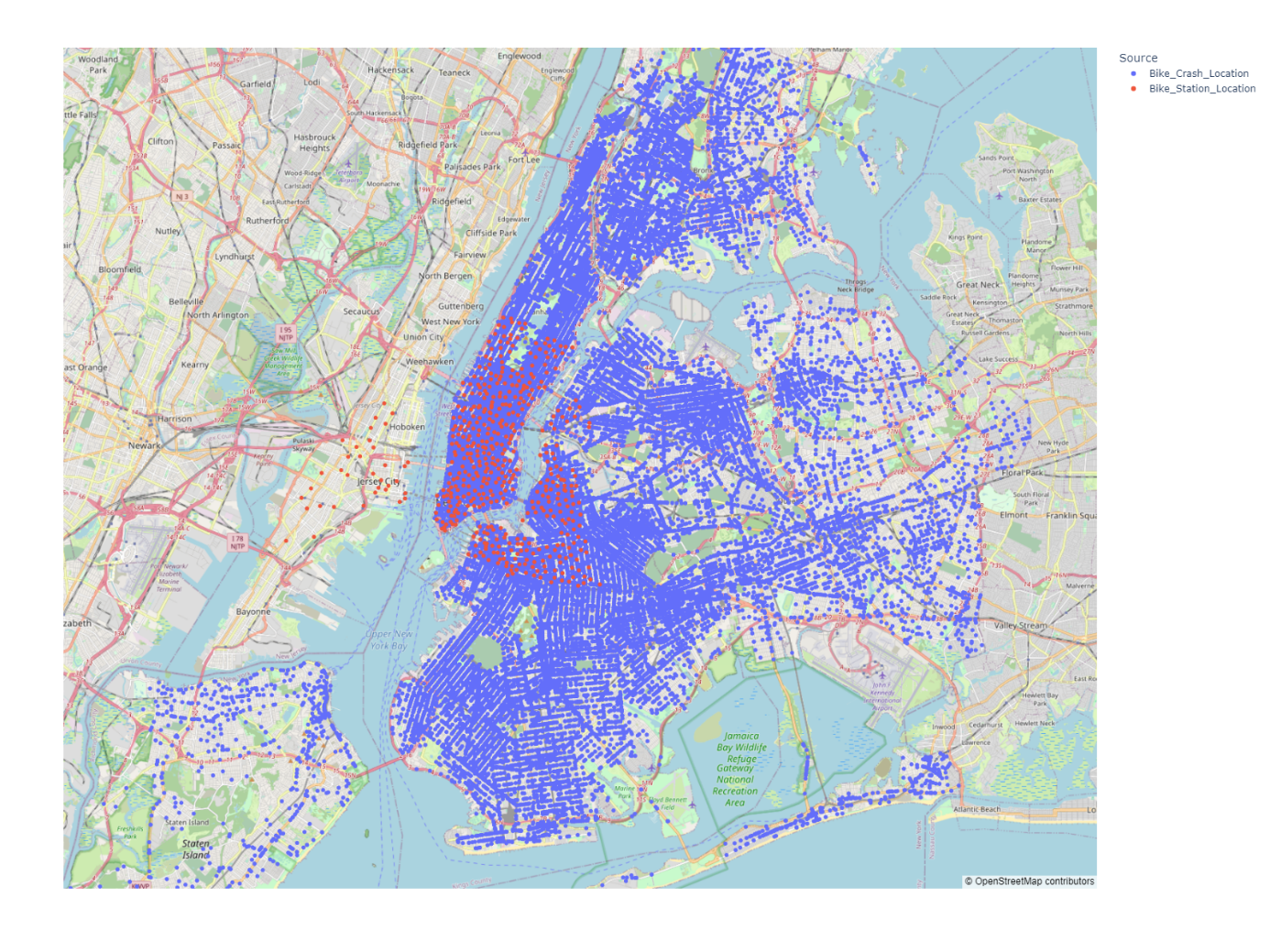
1. **What is the most dangerous NYC borough / area for a bicycle rider?**

*Brooklyn is the most dangerous NYC Borough in terms of number of vehicle accidents from 2013 to 2021.*



1. **What would you change in the locations of bike stations to increase safety?**

*As shown in the plot below, the bike stations are evenly spread over New York city, instead, they are mostly concentrated in the Manhattan and a small area in Brooklyn, where bike accidents used to occur rather a lot. If we zoom in the plot, we can identify some bike stations located too near bike accidents locations. It would be advisable to relocate some of the bike stations into the rest of NYC.*



1. **Where can an accident occur and how close is this from the nearest bike station? Create a predictive model.**

*I use the dataset “Motor\_Vehicle\_Collisions” from NYPD, which has about 1.67 million accidents in general, 57384 Bike related accidents in particular, to build a binary classification mode in order to predict the probability of bike accident occurring at a specific location in NYC. Since the dataset is imbalanced, I perform undersampling on the number of no\_bike-related accidents from the “Motor\_Vehicle\_Collisions” dataset.Iin the dataset used for predictive modelling, there are 57384 Bike related accidents and 57384 non-Bike related accidents.*

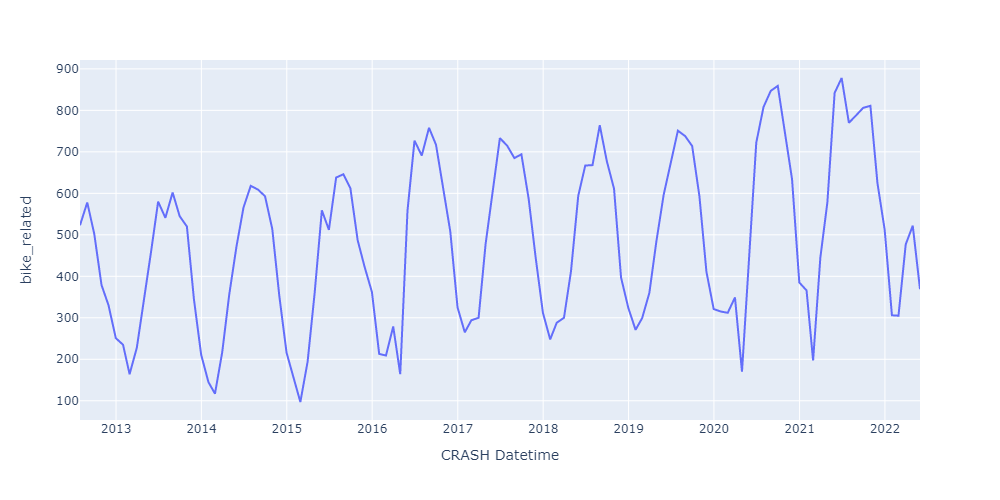
*This dataset is used to train a binary classification model (random forest) to predict whether a bike accident can occur at a location in New york City (defined by LATITUDE and LONGITUDE). In this a binary classification model, LATITUDE and LONGITUDE are used as input features and whether the accident is bike related is the label. To train the model, I use randomized hyperparameter tuning.*

*To use the model for prediction, one has to give the values for the two input features: LATITUDE and LONGITUDE of a location in NYC, then the model will produce a probability about how likely will a bike accident occur at this location.*

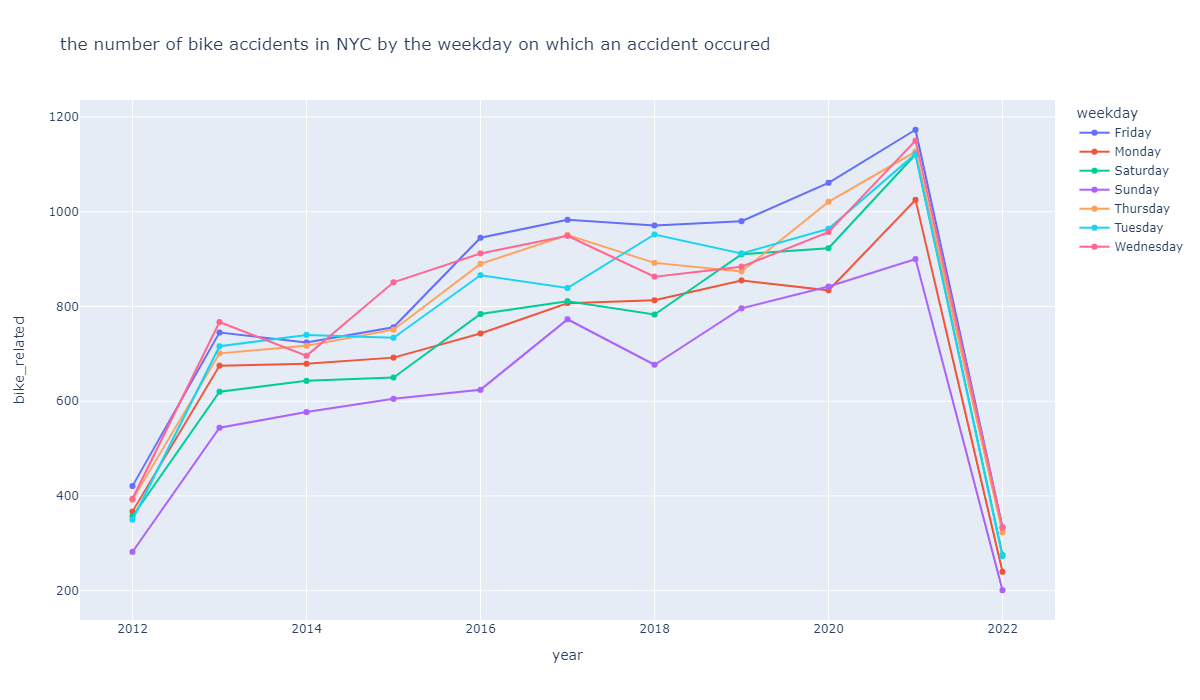
*I write a function* ***nearest\_station(lat1,lon1)*** *to find the nearst bike station given a location in NYC. The function take two parameters latitude and longitude of a location in NYC as input and returns the bike station id and the distance from the location and the bike station.*

1. **What other interesting patterns do you find in the data? What other data sources would be interesting to correlate?**

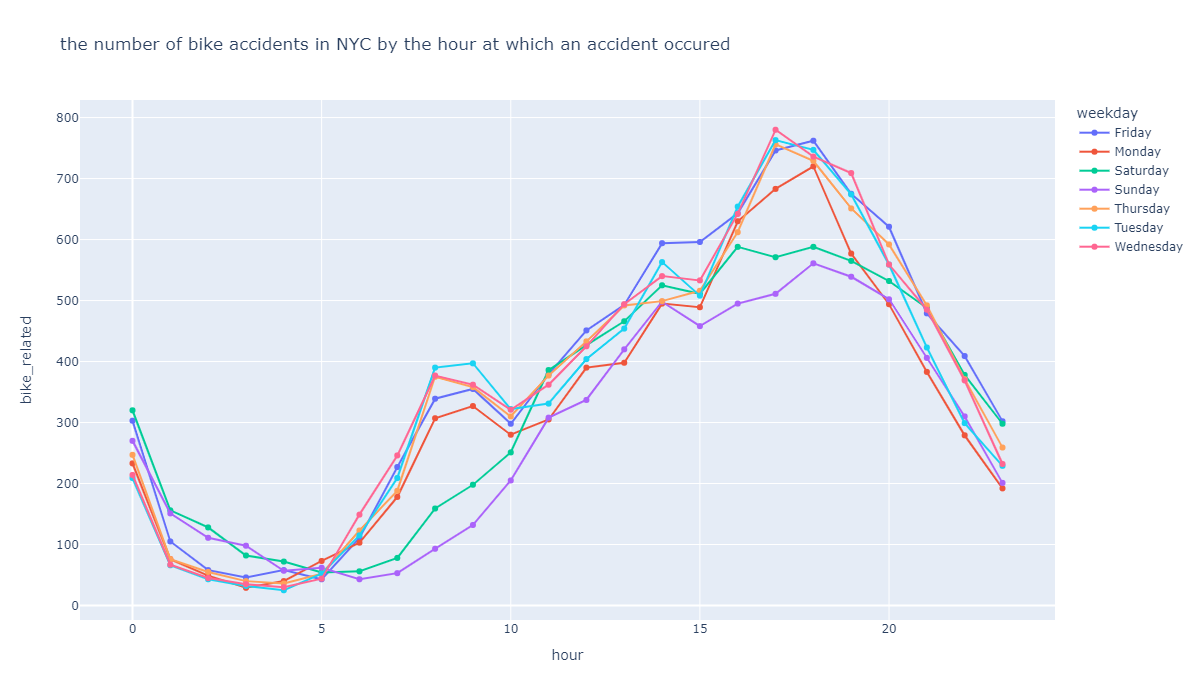
*4.1 as shown in the following graph, the temporal development of the number of bike accidents shows seasonally cyclical pattern.  From June to October, roughly the summer time, the number of bike accidents reach high levels, while in Winter months, from December to March, the number of bike accidents are relatively few.*

**

*4.2 as shown in the following graph, Bike accidents tend to occur most on Fridays, while they occur least on Sundays. From 2020 to 2021, we see a relatively sharp increase in bike accidents for almost all weekdays, it may be due to coronavirus in NYC, which causes the New Yorkers to use less public transport and bike more.*

**

*4.3 as shown in the following graph, Bike accidents tend increase from 10:00 to 17:00, the number of bike accidents reaches its peak between 17:00-18:00 on weekdays in NYC. Those hours correspond to the end of a workday, when more people ride a bike to go home. There is a second peak of bike accidents between 8:00 and 9:00, which may be due to people riding to work by bike in the morning.*



* 1. ***additional data***

It would be helpful to have additional data about the population density for different neighborhoods in New York City. This might correlate positively the total number of bike accidents in that neighborhoods.

Additional data about the weather conditions for the day when the collision occurred would also be helpful. Bad weather, for example, rain or snow, could cause a rise in the number of bike accidents.