0.1 Processing and Cleaning the Data

In order to make predictions the authors decided that the PdDistrict would suffice as location information, so the address could be dropped. The description and resolution of the crime and arrest were also unnecessary to make a description of the type of crime. The authors changed the date into numbers and splitted into separate columns consisting of Year, month, day, hour and DayOfWeek.

While plotting the cooridinates, some outliers were noticeable.

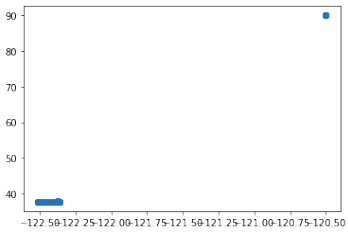

Abbildung 1: Plot of Cooridinates.

Since San Francisco has the latitude of circa 37, all coordinates with latitude under 40 were kept and the rest deleted, there were only 76 coordinates to be deleted.

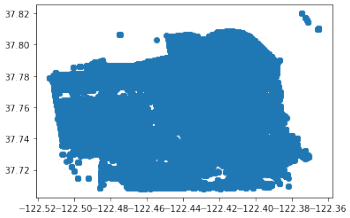

Abbildung 2: Plot of Cooridinates cleaned up.

0.1.1 Binning

To make the patterns more noticeable the authors decided to use Data binning. The Hours were binned in an equal-width size 6 bin, the divisions represent Early morning, morning, noon, afternoon, evening and night. Additionaly the longitude and latitude were binned separately in size 80 bins.