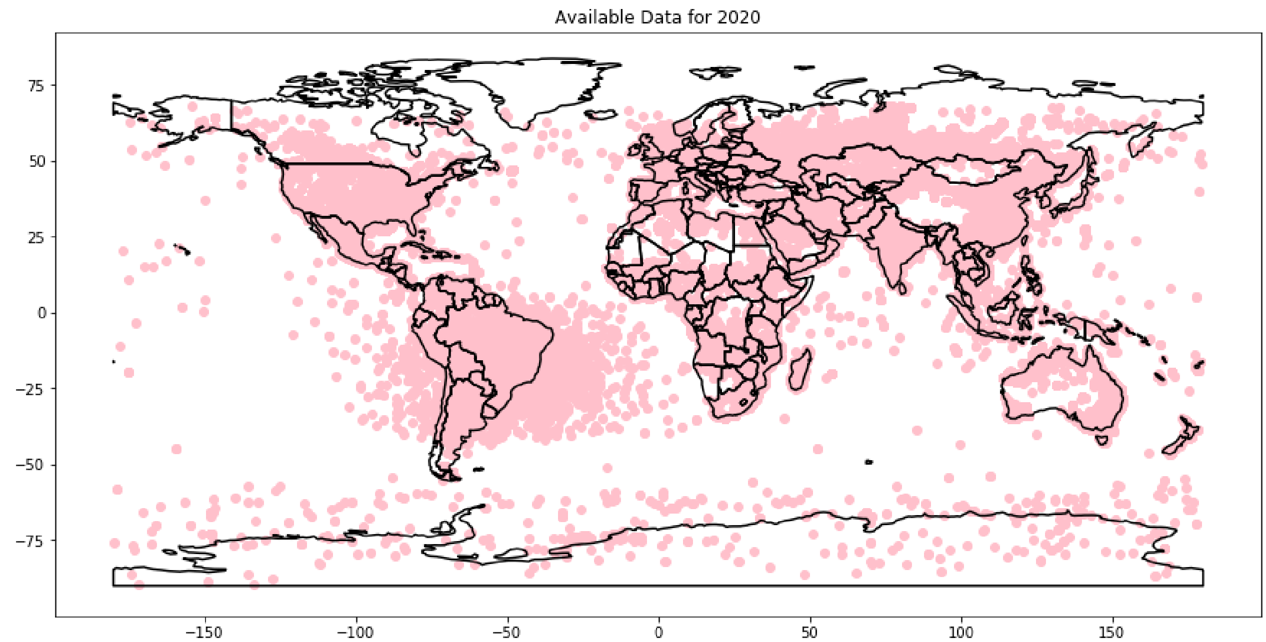
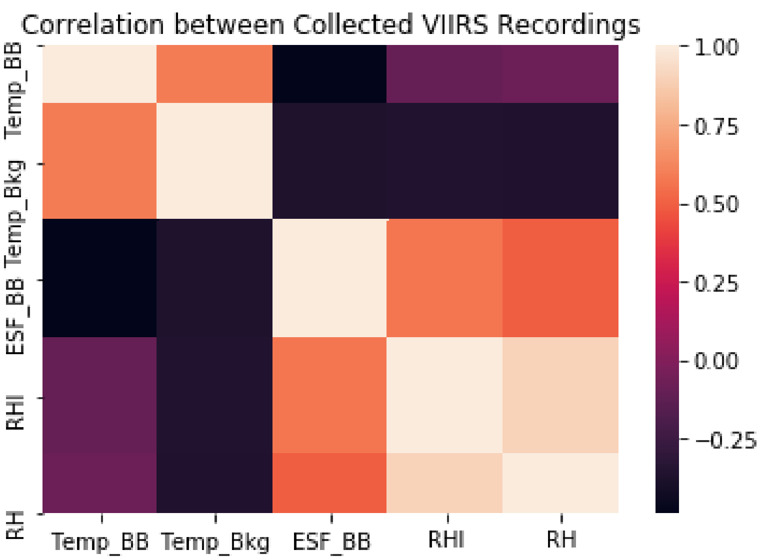
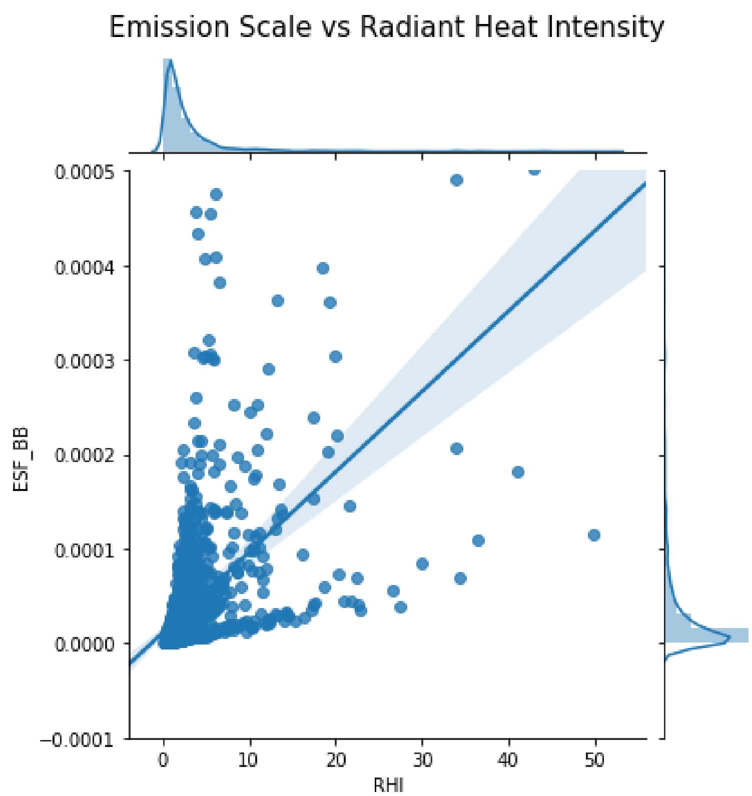
# Anomaly Detection of Radiance Emissions utilizing Machine Learning

1. **Business and Data Understanding**
   a. Business needs addressed:
      i. To discover anomalies in the oil industry (production, extraction, recovery) via change detection in gas flares and biomass burning activity in the continents of Africa and Asia.
   b. Objectives:
      i. The machine learning algorithm will train on sensor data in order to predict the radiant heat intensity. This prediction will then be compared with the most recent data from the area collected by the VIIRS sensor in order to spot differences based on a decided upon difference (Back-tracking).
      Part A: Supervised Learning
      Part B: Anomaly Detection
   c. Data
      i. **Available** - VIIRS (Visible Infrared Imaging Radiometer Suite) is a sensor on the NASA operated SNPP Satellite, which can observe the radiant emissions from gas flares, biomass burning, industrial sites and volcanoes worldwide with at least one coverage every 24 hours.
      ii. **Size** - There is 2.8 MB (7000 rows) of data for each day. - about 4.7 GB of data for the past 5 years. This can still be reduced if constraining to a specific geo-location zone.
      DATA: https://eogdata.mines.edu/download_viirs_fire.html
      iii. **Main features** - The data contains geolocation (lang/lat), Radiant heat intensity, Earth temperature estimate, Source temperature estimate, and Methane detection estimation among others.
      iv. **Other potential data sets online** - This could be enriched with another dataset derived from the VIIRS satellite, or a dataset that is geolocated and can be merged with the existing dataset (for example, C02/other gas emissions in that area, weather, etc.).

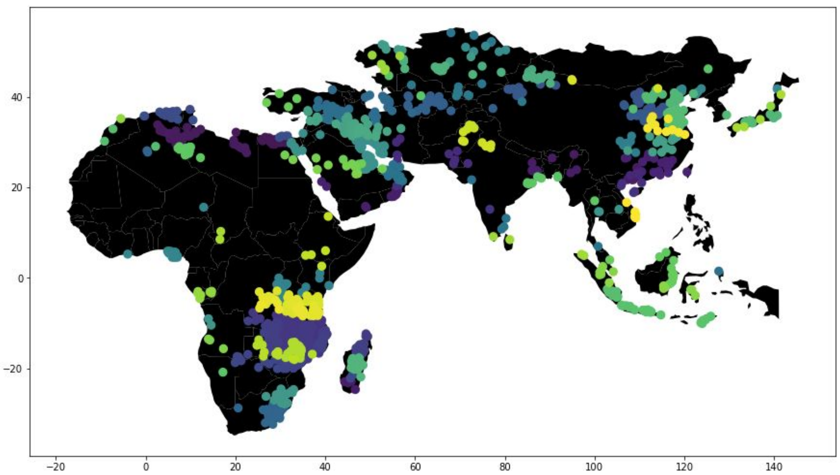
2. **Data Analysis:**


Available Data for 2020
75
50
25
0
−25
−50
−75
−150
−100
−50
0
50
100
150


Correlation between Collected VIIRS Recordings
Temp_BB
Temp_Bkg
ESF_BB
RHI
RH
1.00
0.75
0.50
0.25
0.00
−0.25

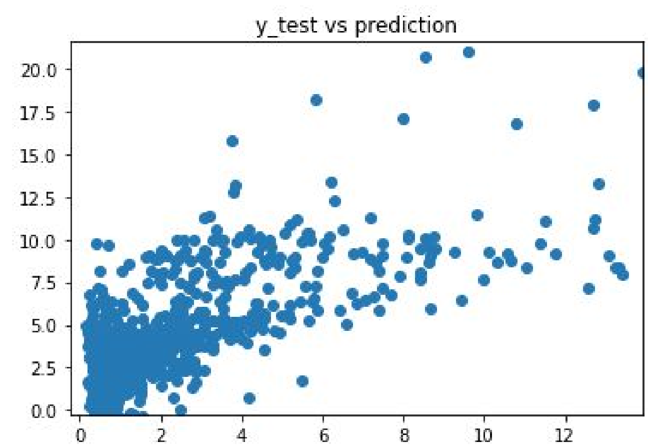

Emission Scale vs Radiant Heat Intensity
ESF_BB
RHI
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
−0.0001
0
10
20
30
40
50

3. **Data Preparation**
    a. The data was cleaned of Null data along with any filler temperature/sensor values (defined usually as 9999). A histogram of the data was performed in order to see any significant outliers which are then subsequently removed. The data was geospatially refined to the continents of africa and asia. In order to simplify the geospatial location for the machine learning model, the data was then aggregated through spatial modular clustering in which points are grouped iteratively by location to create “hotspots” of data.




4. **Modeling**
    **a.** RH (radiant heat or the radiation emitted by hot objects) is set as the predictor y value. X is set as the rest of the **numerical** values in the set (values such as object IDs, latitude and longitude, and spectral pixel location are removed). A linear regression model is used in order to predict the changing value of RH.

5. **Evaluation**
    **a.** The model itself has an r-squared score of .806 for the test data.




    b. The model is then used to predict a set number of values for the most recent day, and compared to the real days results. The comparison between these two values are calculated and then binned based on how large the gap is. A map based on key colouring shows **anomalies in red**, which indicate radiation emission in unexpected areas.

6. **Summary**
    a. In short, this model can be used to track and investigate anomalies in high radiant emissions that are often associated with gas flares, biomass burning, volcanoes, and industrial sites. This could indicate suspicious activity, wildfires, changes in the oil industry and more. Regarding the model itself, it would be worthwhile to continue feature extraction and incorporating a larger set of data with more years of the satellite data, as well as consistently incorporating the newer data to the model.