**Identifying Anomalies in CH4 Sensor Data Recordings**

**By Rupali Roy(rur73)**

Date: Thursday, December 10, 2020

Course: Data Storytelling

Contents

[Introduction 1](#_Toc30435930)

[Description of partnering organization and their needs 1](#_Toc30435931)

[Description of audience 1](#_Toc30435932)

[Design requirements based on user needs analysis 1](#_Toc30435933)

[Story/exploration experience for the audience 1](#_Toc30435934)

[Visual encoding choices 2](#_Toc30435935)

[Data provenance 2](#_Toc30435936)

[Data dictionary 2](#_Toc30435937)

[Data transformation and cleaning 2](#_Toc30435938)

[Calculations 2](#_Toc30435939)

## Introduction

## Description of partnering organization and their needs

The Bureau of Economic Geology is the oldest research unit at The University of Texas at Austin. The Bureau is the State Geological Survey of Texas and has been an integral part of the development of the state’s economic success through the years. I am working closely with a Senior Research Scientist at Bureau of Economic Geology, Austin. His area of interest is into Geosciences and Natural Sciences Data. I am happy to be able to work with him again after my Summer 2020 internship. I did my internship under his guidance and worked on project related to real time anomaly detection for earthquake.

## Description of audience

This kind of data analysis is used in industrial sector. Since the major goal of this project is to build a dashboard for monitoring the anomalies in the CH4 levels. The target audience are people employed into the risk assessment and monitoring department of any company. The oil and gas industries would be most likely interested in such data analysis and dashboard creation. Additionally, research scientist and students who are working in the natural sciences and geosciences sector may be highly interested in this dashboard.

## Design requirements based on user needs analysis

The main requirement was to create a dashboard to enable users identity anomalies in CH4 sensor level readings and possibly the cause of the peak and the time for the rise in sensor reading levels. Try to find correlated features and identify dependencies in the dataset.

## Story/Exploration experience for the audience

The audience can use this dashboard to visualize and identify the timings of the day when the CH4 level crosses the average levels and identify the cases of anomalies. After they identify the time and the anomalous CH4 methane level recorded by the sensor, they can check other parameters and try to corelate the peak in CH4 levels with the other parameters in the dashboard such as wind direction and wind speed. The dashboard also provides the possible emission event that might have occurred during the peak CH4 range.

## Visual encoding choices

For the dashboard, I chose red-blue diverging color palette to highlight the variation in the parameters. I personally feel that this color palette really helped in highlighting the anomalies in the data for me and making it more visually appealing. The anomalies are highlighted in the red shade. I also used a reference line in the graphs to highlight the average values for important parameter in our analysis.

Since most of the data shared by the client was time series data, the best graph was to use a line chart or a scatter plot to observe the trends and patterns in the data with respect to time. I have used scatter plot with circles to show the distribution of CH4 levels and the wind direction. The windspeed is shown using a line chart, this helps us in clearly visualizing the changes in the speed recordings. A stacked bar chart is also shown to understand the distribution of CH4 emission event type and BFE (Bio-mass based fuel ethanol) levels. It shows the different event types that is responsible for the emission of gas.

## Data provenance

The dataset that I am using was provided by a research scientist at Bureau of Economic Geology. The data is mostly sensor recordings for the CH4 levels and parameters impacting the CH4 levels like wind speed, wind direction, BFE levels and emission event type etc. This data was provided to him by the company PSIcorp <http://www.psicorp.com/> , it’s a company focused on research in the physical and geosciences field.

## Data dictionary

There were three different data sheets provided by the company. I had a discussion with my guide to understand the major factors that needs to be considered while analyzing the data and creating the dashboard using those important features. Here’s a list of the important field names and data description:

|  |  |  |
| --- | --- | --- |
| Field name | Data type | Definition |
| CH4 Average | Integer | Average Methane Gas Level measured in PPM |
| CH4 Peak | Integer | CH4 peak level |
| Wind Direction | Integer | Recorded in degrees (0-360) |
| Wind Speed | Integer | Coded from 0-11 (11 being the highest recorded speed) |
| Emission Category | String | CH4 Emission categories- Fugitive(Leaks) and Vent |
| Emission Event Type | String | Different Emission Events |
| BFE | Integer | Bio mass fuel Ethanol |

## Data transformation and cleaning

The main task was to select the important features that affect the CH4 Emission Levels. I tried to segregate the data and understand these factors by having constant communication with my project guide. After getting an idea bout the features that affect the methane levels, I tried to restrict to those fields and created the dashboard using just the fields shared by them. I made sure there were no null values while creating the dashboard and the time units were constant across the data sheet.

## Calculations

N/A