CAMP Project: QA/QC Data Analysis

Md Ismail Hossain

2023-07-03

Exploratory Analysis of Main Data:

Data collection starting from 2023-05-22 and we are using the data till 2023-06-28.

```
## [1] "Number of Rows: 4082"
```

```
## [1] "Number of unique culverts: 4074"
```

So, there are 4082 rows in the data and 4074 unique culvert inspected. Let's find the culverts which are not uniquely entered.

```
## # A tibble: 8 x 1
##   CulvertID
##   <chr>
## 1 A257
## 2 B709
## 3 D873
## 4 C781
## 5 D241
## 6 D814
## 7 D234
## 8 C752
```

So, this 8 culvert have entered more than one times. We need to fix the entry for them.

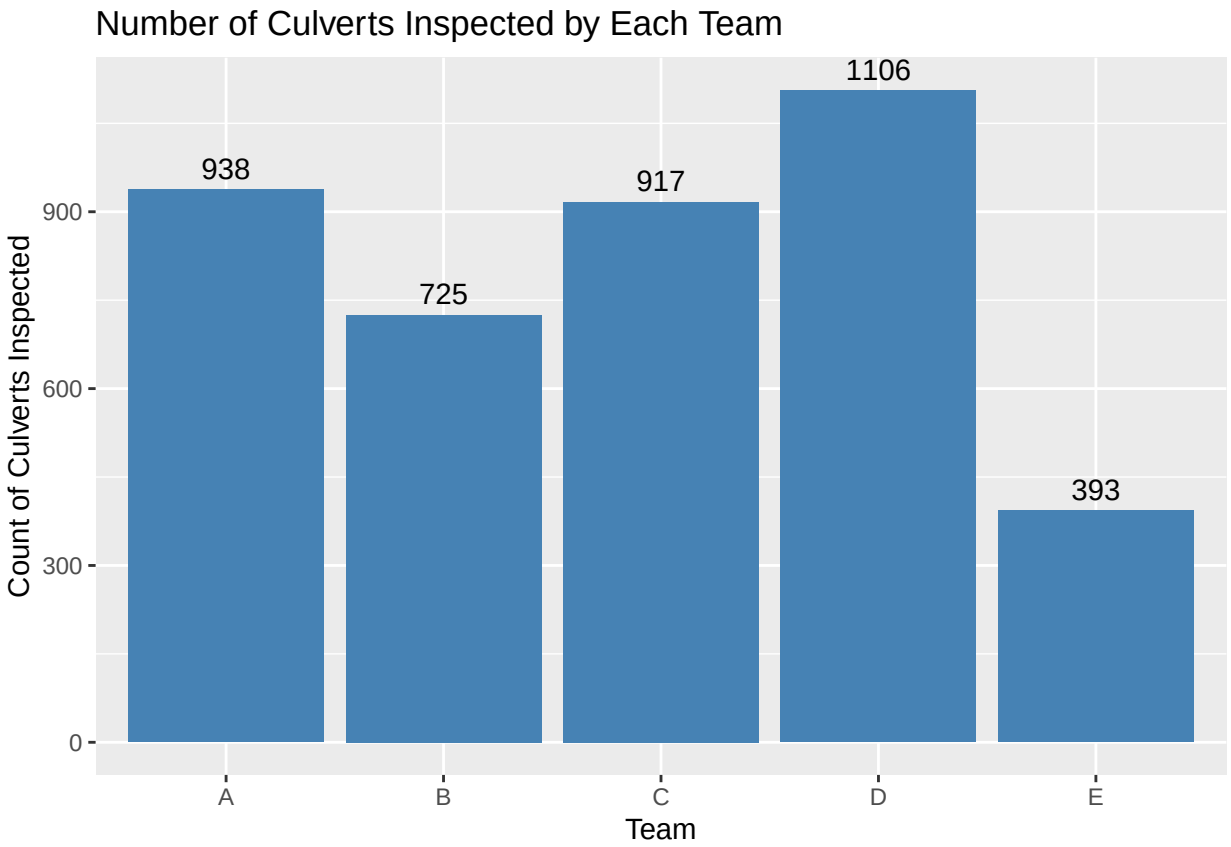
Let's find the number of culvert inspected by team:

```
## # A tibble: 8 x 2
##   Team  CulvertCount
##   <chr>        <int>
## 1 8                1
## 2 9                1
## 3 A              938
## 4 B              725
## 5 C              917
## 6 D             1106
## 7 E              393
## 8 <NA>             1
```

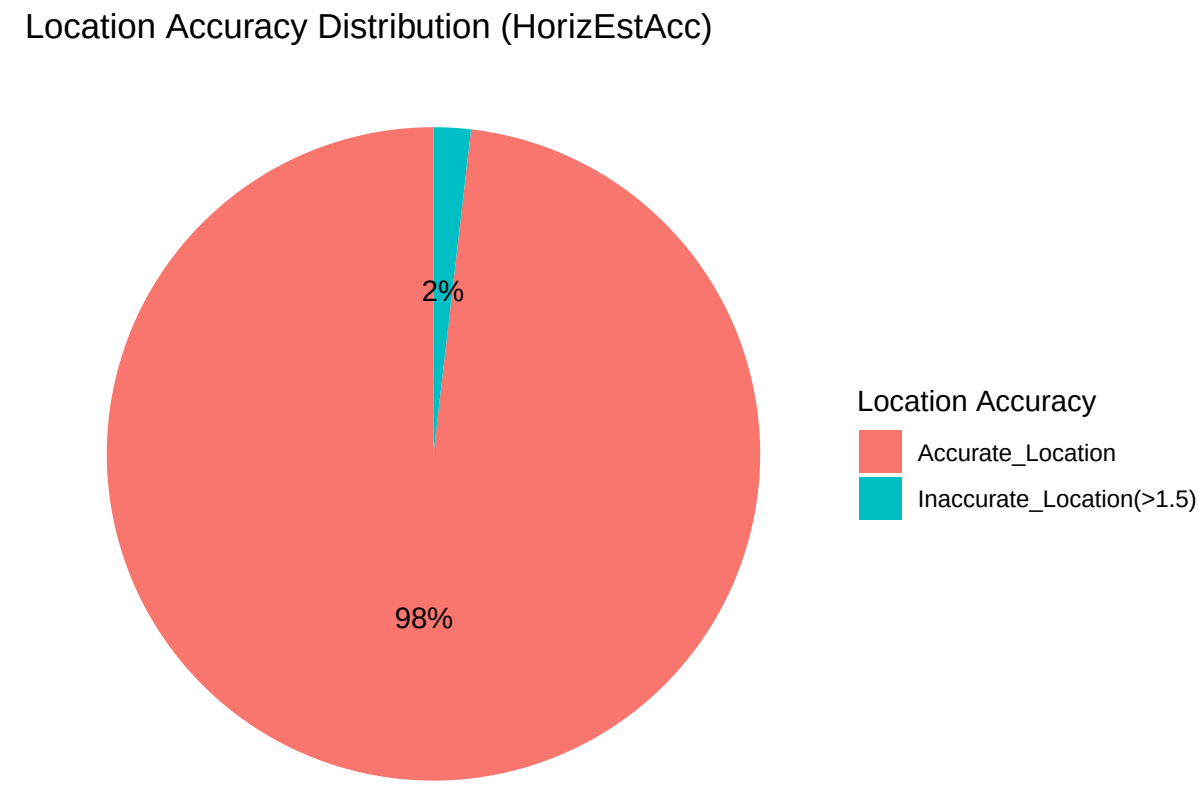
Let's find the rows where the Team ID have discrepancies.

```
## # A tibble: 3 x 2
##   CulvertID Collector
##   <chr>     <chr>
## 1 822       culvertam1@npe.nmt.edu
## 2 943       culvertam1@npe.nmt.edu
## 3 <NA>      culvertam1@npe.nmt.edu
```

Let's remove this


Number of Culverts Inspected by Each Team
Count of Culverts Inspected
0
300
600
900
938
725
917
1106
393
A
B
C
D
E
Team

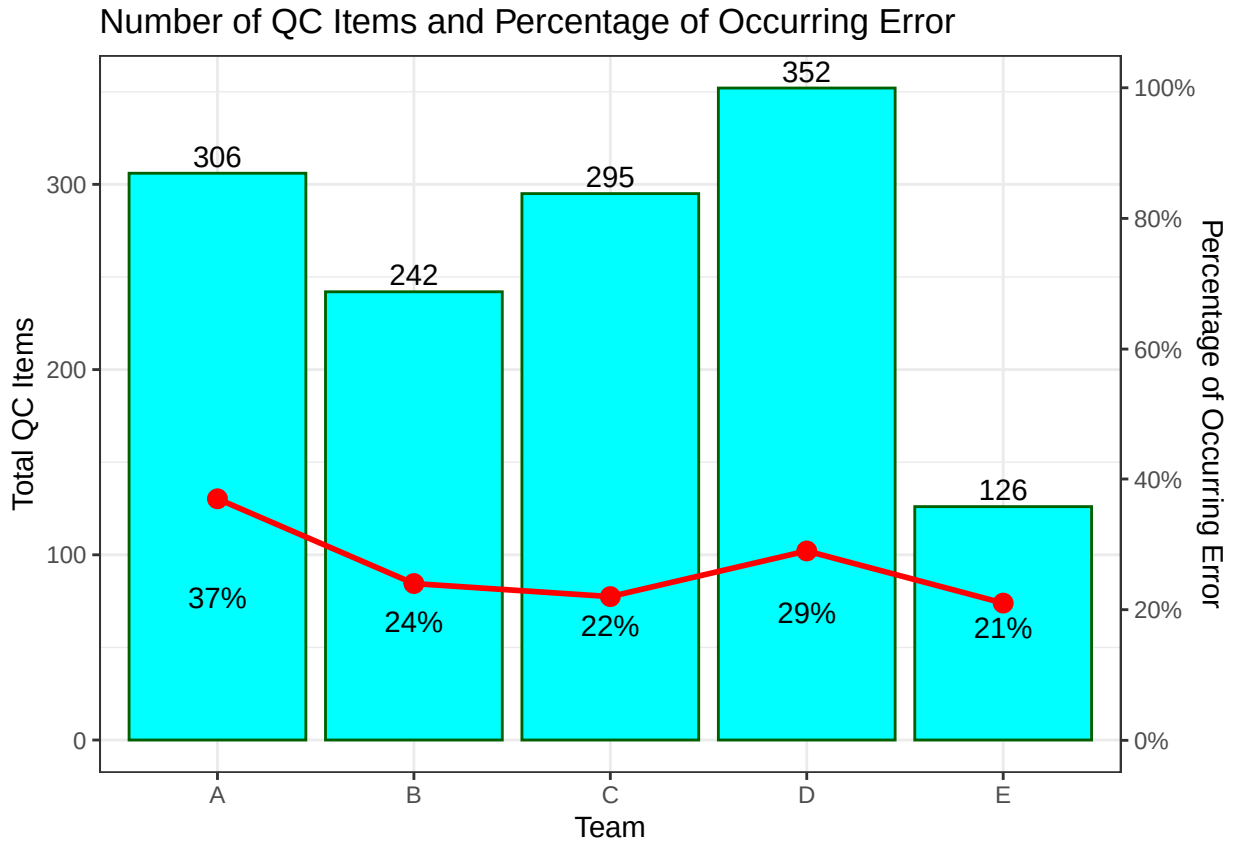
GPS coordinate location accuracy:



Now we have a hypothesis that if there is an error occurred in any column then there is high possibility that there will be another error made on measuring the location accuracy. We will test this hypothesis in the later section.

Exploratory analysis of QC Data:

Total QC item is 1321 among 4074 culvert collected till 2023-06-28, which is 32% overall. Let's observe the team wise inspection and the % of error (at least one error):



A closer look on error data:

So, in total 364 culvert id have at least one issue.

```
##
## > One_Error   One_Error
##          78         286
```

When an error is made in more than one entry or exactly in one entry, this tally will indicate the corresponding number of culverts.

After merging the main data with QC data (infected id's only) we are observing some id mismatch. This happens may be that id data removed from the database or something else happened. This are the id's (total 14 culvert id) which are not matching with the main data set although we did the QC for them:

A60, A19, A55, A45, A49, A36, A66, A61, A32, C806, A822, D970, A1020, A1015

Hypothesis testing:

Let's find number of culvert where there is also location error:

```
##
##               Accurate_Location Inaccurate_Location(>1.5)
##   > One_Error                75                         0
##   One_Error                 270                         5
```

Now we have the hypothesis that if an error occurs in any column, there is a high likelihood that another error will occur when measuring the location accuracy. This hypothesis will be tested in a subsequent section.

The above contingency table demonstrates that only 5 of the QC items with the error also have an inaccurate location issue. We can conduct a statistical test of the following hypothesis:

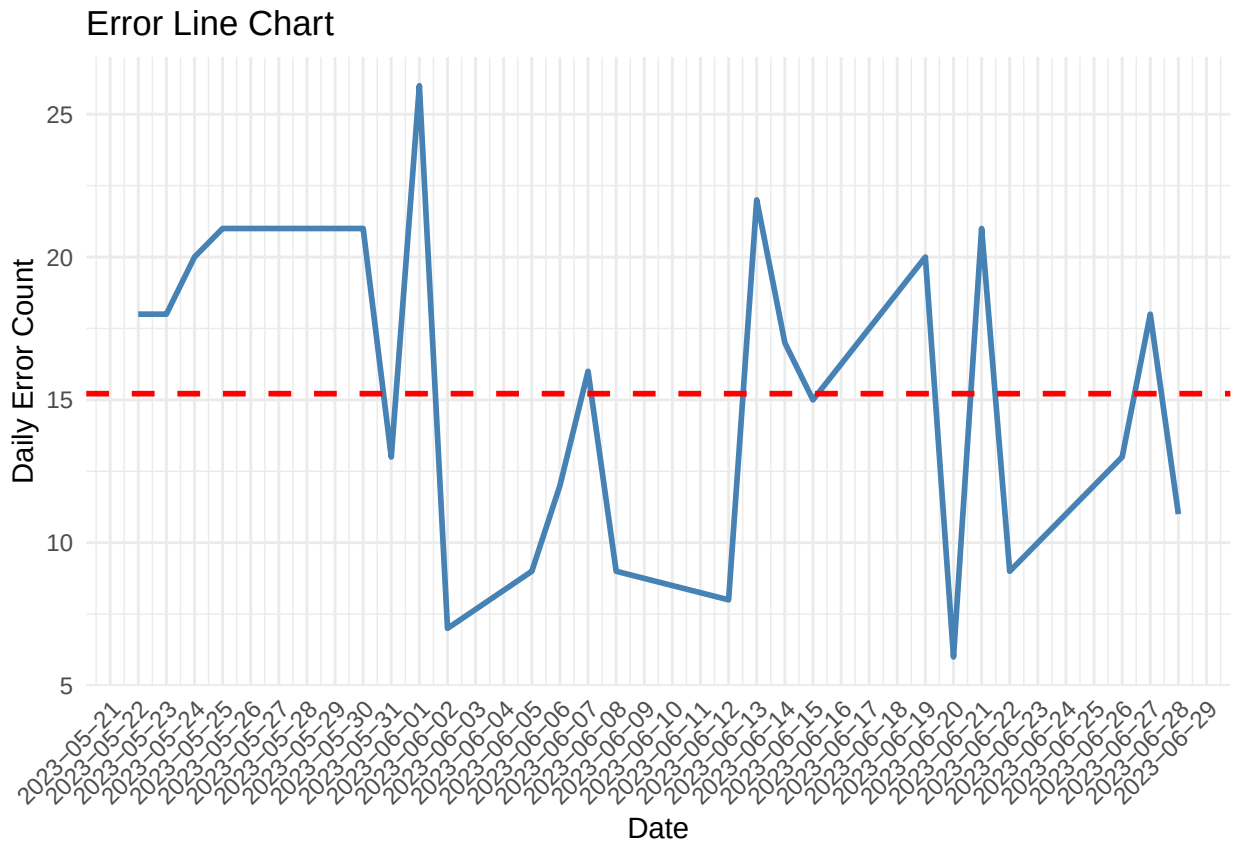
H_0**:** Error counts per culvert and GPS location error are independent.

H_1**:** Error counts per culvert and GPS location error are dependent.

```
##
##  Pearson's product-moment correlation
##
## data:  df_merged$non_missing_count and df_merged$HorizEstAcc
## t = -1.2461, df = 348, p-value = 0.2135
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.17029201  0.03844642
## sample estimates:
##         cor
## -0.06665204
```

The p-value calculated is greater than 0.05. Therefore, at a 5% level of significance, we cannot reject the null hypothesis or conclude that error counts per culverts are independent of GPS location (HorizEstAcc) error.

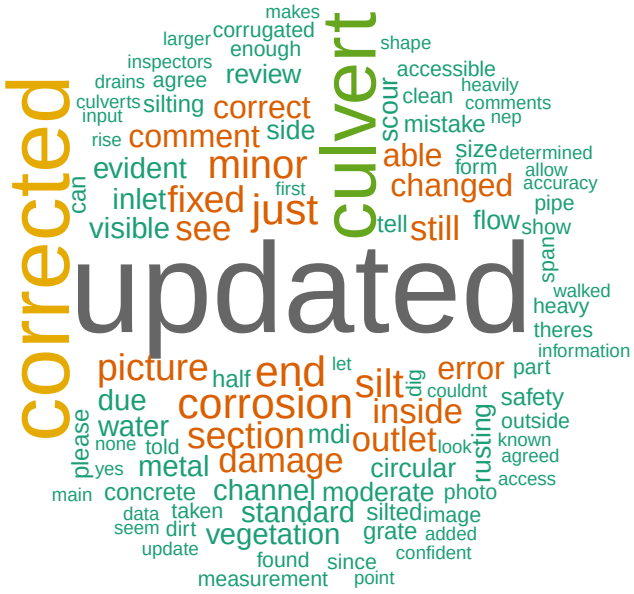
Error Trend (at least one error per culvertID):



Daily average count of the error is 15.22 which presented using red dashed line. Let's observe few analysis (word cloud) using the comments made by Field officer and QA/QC analyst:

Word cloud of QC comments:

Word cloud of FO comments:



Word cloud of comments (main data):