CAMP Project: QA/QC Data Analysis

Md Ismail Hossain

2023-07-03

Data Loading:

```
## Rows: 4082 Columns: 38
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (25): GPSPointLocation, CulvertID, CulvertAccessibility, ReasonforInacc...
## dbl  (10): NumberofCulverts, SpanIn, RiseIn, WidthofMultiCulvertsFt, Degrees...
## lgl   (1): TaskName
## dttm  (2): CreationDateTime, UpdateDateTime
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
## # A tibble: 6 x 38
##   GPSPointLocation CulvertID CulvertAccessibi~ ReasonforInacces~ InletisMDIorSCI
##   <chr>            <chr>     <chr>             <chr>             <chr>
## 1 T-Post Reflector C650      No                Silted Up         No
## 2 Turnouts (Road)  B92       Yes               <NA>              No
## 3 Outlet           A506      Yes               <NA>              No
## 4 Other            C39       Yes               <NA>              No
## 5 Outlet           D463      Yes               <NA>              No
## 6 Inlet            D448      Yes               <NA>              No
## # ... with 33 more variables: Material <chr>, NumberofCulverts <dbl>,
## #   InletEndSectionType <chr>, OutletEndSectionType <chr>,
## #   DropInletDetails <chr>, Culvert_Shape <chr>, PhysicalDamage <chr>,
## #   Corrosion <chr>, SpanIn <dbl>, RiseIn <dbl>, WidthofMultiCulvertsFt <dbl>,
## #   Silting <chr>, Scour <chr>, Erosion_Control <chr>, ChannelType <chr>,
## #   ChannelCondition <chr>, Skew <chr>, DegreesofSkew <dbl>, Comments <chr>,
## #   Collector <chr>, Correctionsource <chr>, Workspace <chr>, DeviceID <chr>,
## #   CorrStatus <chr>, CreationDateTime <dttm>, UpdateDateTime <dttm>,
## #   HorizEstAcc <dbl>, VertEstAcc <dbl>, TaskName <lgl>, Photos_Multiple <chr>,
## #   X <dbl>, Y <dbl>, Z <dbl>
```

```
## Rows: 1422 Columns: 5
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): Culvert ID, Question, Input, Comment_QC, Comment_FO
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
## Warning: package 'tidyverse' was built under R version 4.1.1

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.3     v dplyr   1.0.7
## v tidyr   1.1.3     v forcats 0.5.1

## Warning: package 'ggplot2' was built under R version 4.1.1

## Warning: package 'forcats' was built under R version 4.1.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

## Warning: Values are not uniquely identified; output will contain list-cols.
## * Use `values_fn = list` to suppress this warning.
## * Use `values_fn = length` to identify where the duplicates arise
## * Use `values_fn = {summary_fun}` to summarise duplicates

## Warning: Values are not uniquely identified; output will contain list-cols.
## * Use `values_fn = list` to suppress this warning.
## * Use `values_fn = length` to identify where the duplicates arise
## * Use `values_fn = {summary_fun}` to summarise duplicates

## Warning: Values are not uniquely identified; output will contain list-cols.
## * Use `values_fn = list` to suppress this warning.
## * Use `values_fn = length` to identify where the duplicates arise
## * Use `values_fn = {summary_fun}` to summarise duplicates
```

Main Data

```
## Warning: package 'lubridate' was built under R version 4.1.1

##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

Exploratory Analysis:

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.

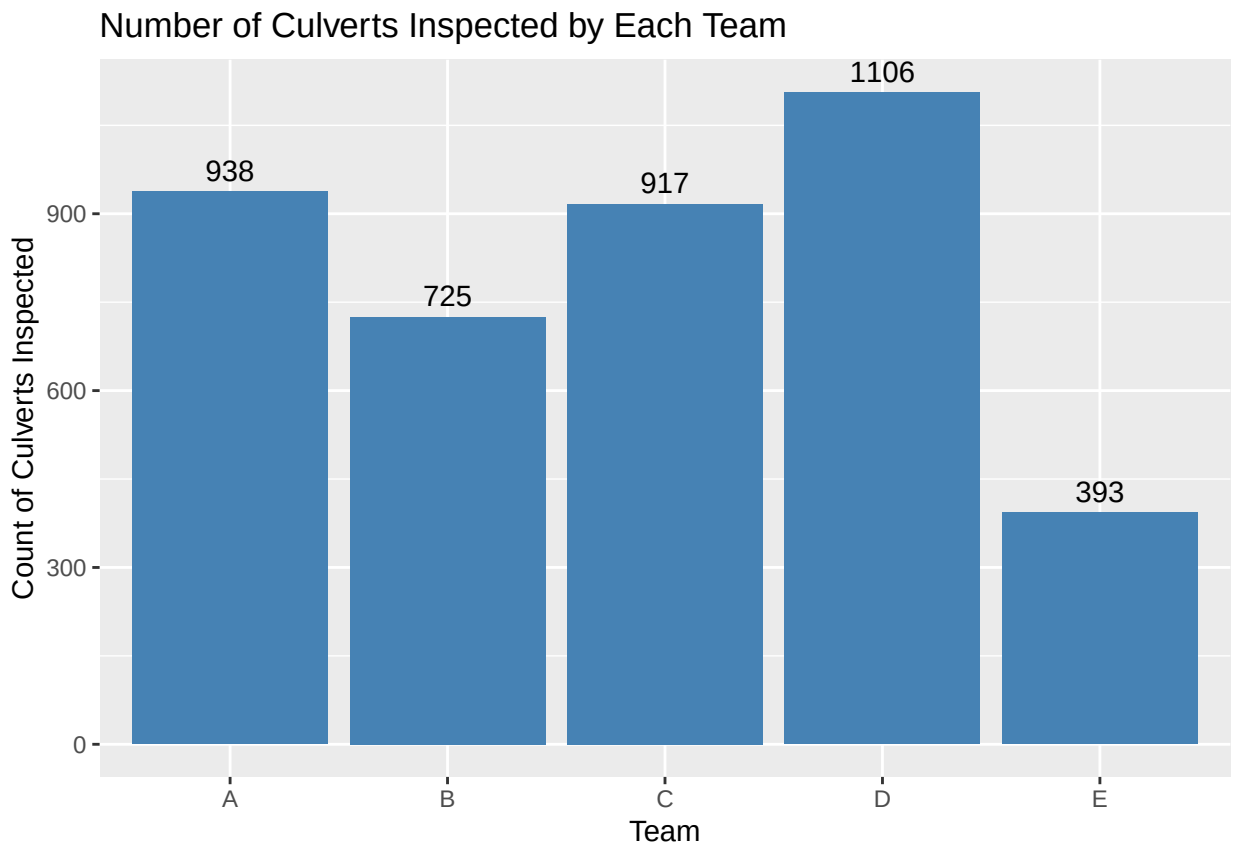
Let's find the number of culvert inspected by team:

```
## # 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
```

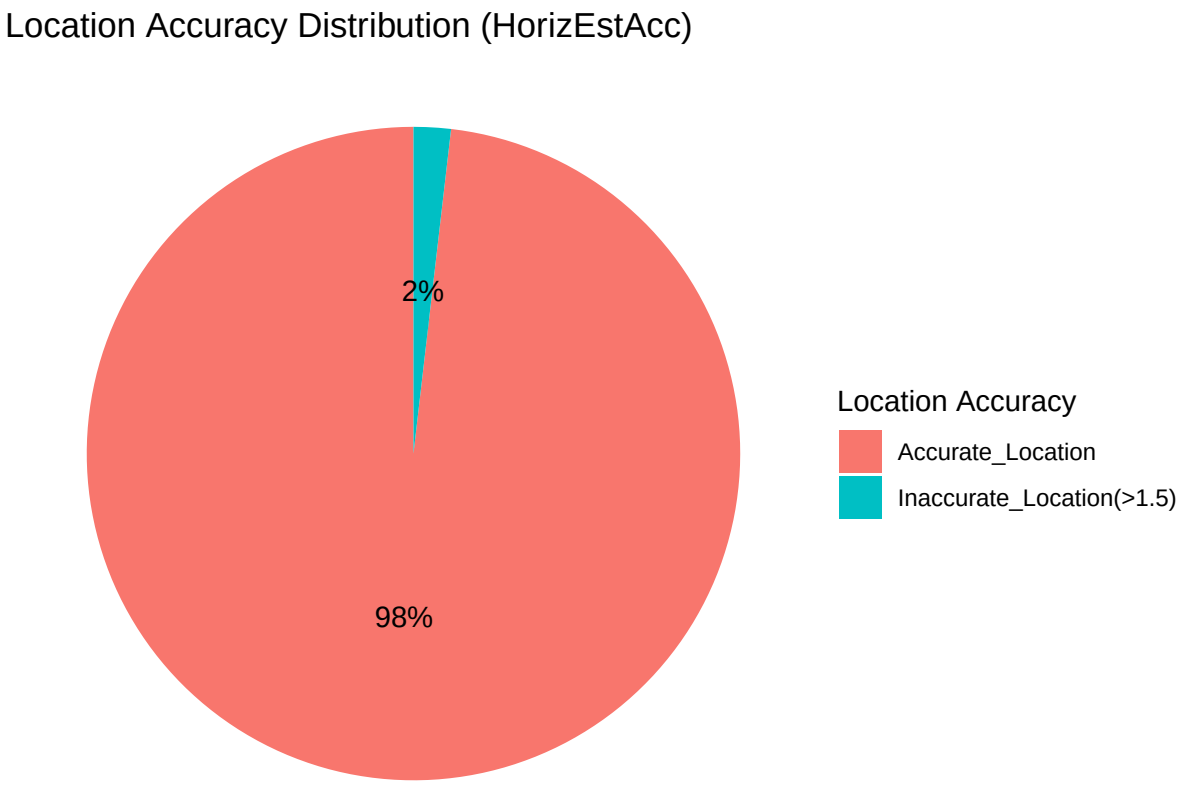
Let's find the rows where the Team ID have discrepancies.

```
## # 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



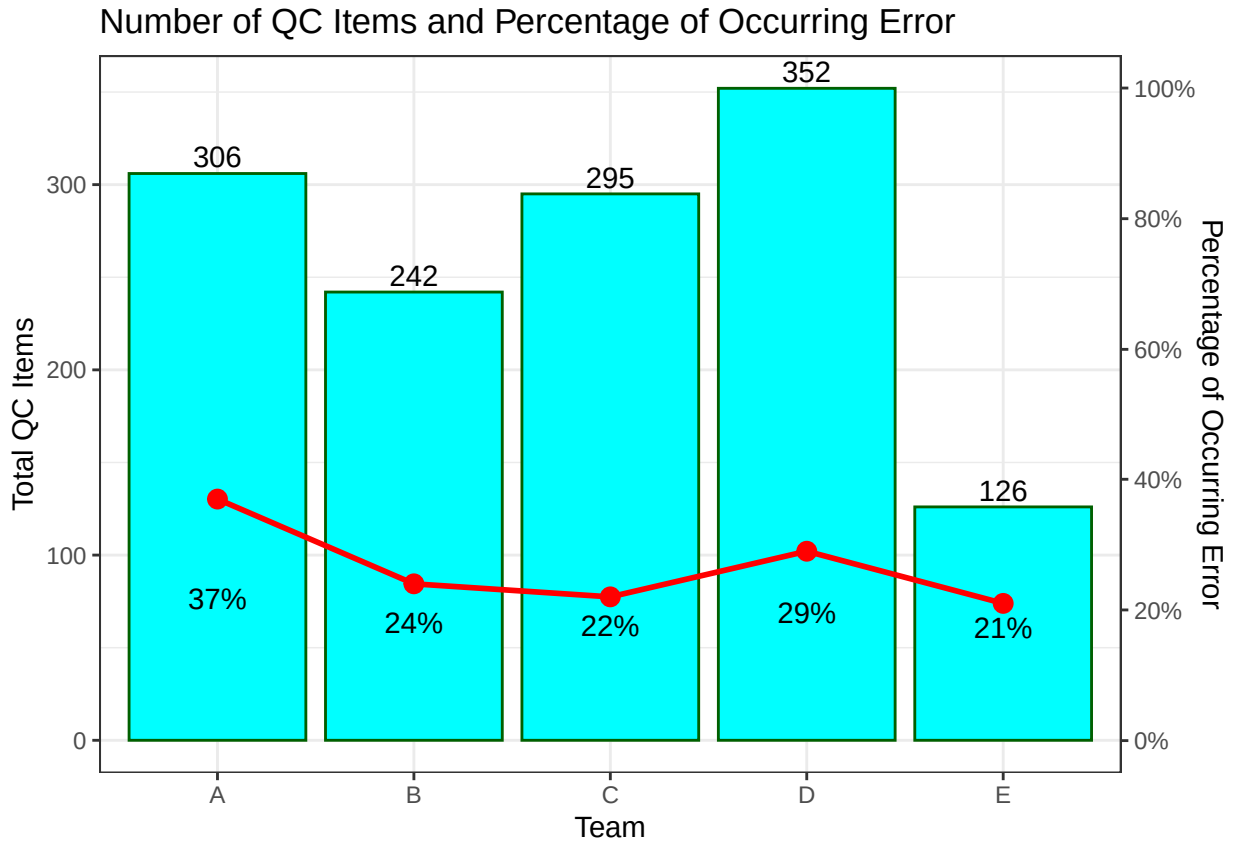
GPS coordinate location accuracy:



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.

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):



Error Analysis:

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   3.426   3.000  15.000
```

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

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

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

Word cloud about the QC comments and FO response:

```
## Loading required package: NLP
```

```
##
## Attaching package: 'NLP'
```

```
## The following object is masked from 'package:ggplot2':
##
##     annotate
```

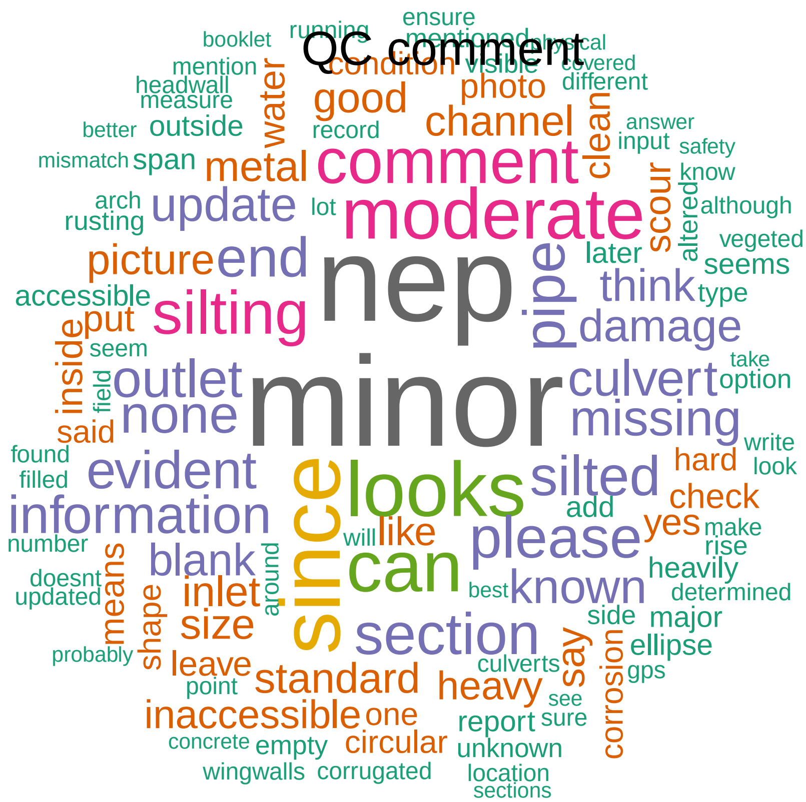
```
## Loading required package: RColorBrewer
```

```
## Warning in tm_map.SimpleCorpus(corpus, content_transformer(tolower)):
## transformation drops documents
```

```
## Warning in tm_map.SimpleCorpus(corpus, removePunctuation): transformation drops
## documents
```

```
## Warning in tm_map.SimpleCorpus(corpus, removeNumbers): transformation drops
## documents
```

```
## Warning in tm_map.SimpleCorpus(corpus, removeWords, stopwords("english")):
## transformation drops documents
```

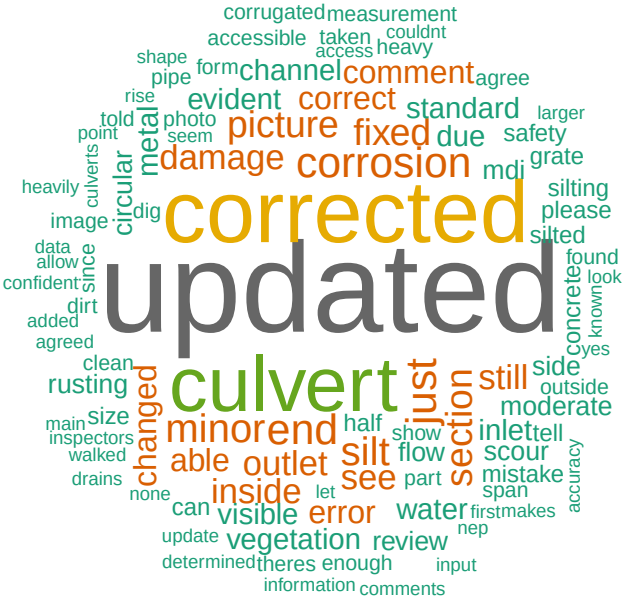



```
## Warning in tm_map.SimpleCorpus(corpus, content_transformer(tolower)):
## transformation drops documents
```

```
## Warning in tm_map.SimpleCorpus(corpus, removePunctuation): transformation drops
## documents
```

```
## Warning in tm_map.SimpleCorpus(corpus, removeNumbers): transformation drops
## documents
```

```
## Warning in tm_map.SimpleCorpus(corpus, removeWords, stopwords("english")):
## transformation drops documents
```



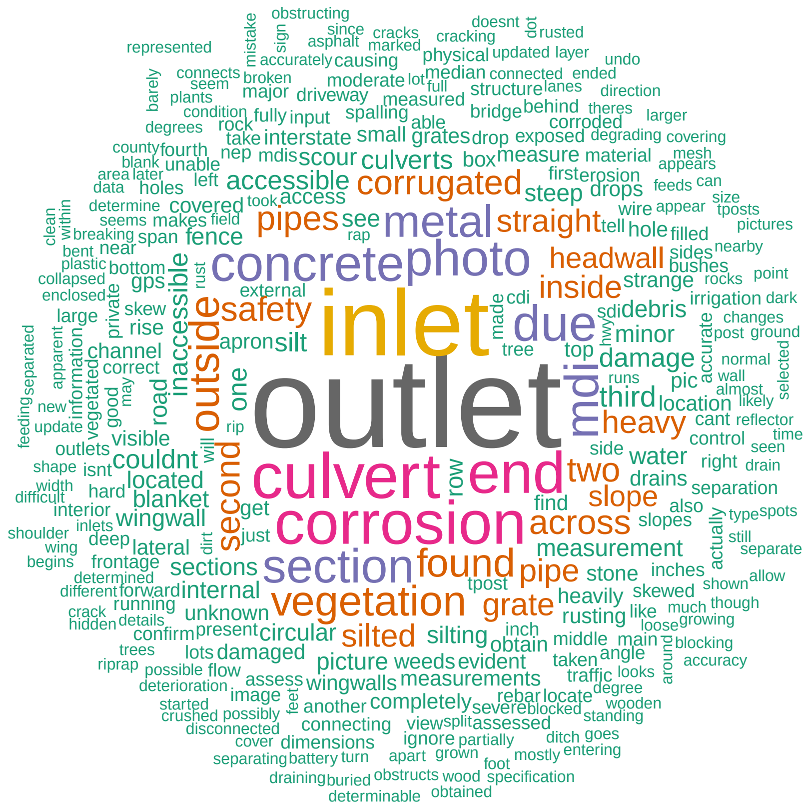
Word cloud about comments they made about the culverts for all the data collected:

```
## Warning in tm_map.SimpleCorpus(corpus, content_transformer(tolower)):
## transformation drops documents
```

```
## Warning in tm_map.SimpleCorpus(corpus, removePunctuation): transformation drops
## documents
```

```
## Warning in tm_map.SimpleCorpus(corpus, removeNumbers): transformation drops
## documents
```

```
## Warning in tm_map.SimpleCorpus(corpus, removeWords, stopwords("english")):
## transformation drops documents
```


outlet
inlet
culvert
corrosion
end
section
concrete
photo
metal
due
mdi
vegetation
found
outside
second
inaccessible
corrugated
pipes
straight
headwall
inside
safety
heavy
two
slope
across
pipe
grate
silted
accessible
culverts
scour
damage
third
debris
minor
water
drains
measurement
sections
internal
unknown
circular
silting
picture
weeds
evident
measurements
wingwalls
completely
wingwall
blanket
located
couldnt
visible
channel
apron
silt
one
road
fence
covered
steep
drops
erosion
first
measure
material
small
grates
interstate
drop
exposed
corroded
bridge
behind
structure
moderate
driveway
measured
spalling
physical
median
causing
obstructing