Tanzanian Water wells Status Prediction

In many regions, water pumps suffer from frequent malfunctions, leading to severe consequences on water access and public health. Our goal is to:

Diagnose: Accurately determine the operational status of water pumps.

Classify: Identify whether a pump is fully operational, partially functional, or non functional.

Inform: Provide actionable insights to stakeholders for prioritizing maintenance, thereby reducing downtime.

This isn't merely a machine learning challenge; it's about transforming data into insights that can save lives and improve living conditions.

This problem naturally fits into a multi-class classification framework. The three primary classes are:

Functional (Fully Operational)

Functional Needs Repair (Partially Functional)

Non Functional (Faulty)

Challenges:

Class Imbalance: Often, the number of pumps in one category (say, fully operational) may dominate the dataset, while malfunctioning pumps might be fewer.

Data Quality and Feature Selection: Sensors might provide noisy data, and maintenance logs can be incomplete.

Changing Conditions: Pumps might deteriorate suddenly due to environmental factors, suggesting that a static model might be insufficient. Considering a periodic model update or even a time series approach might be beneficial.

Loading the data

In [1]:
```
#Importing the necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
#Loading the data
values = pd.read_csv("TrainingsetValues.csv")
labels = pd.read_csv("Trainingsetlabels.csv")
test_values = pd.read_csv("Testsetvalues.csv")
```

In [2]:
```
#exploring the datasets
print (values.head())
print(type(values))
```

```
      id  amount_tsh date_recorded        funder  gps_height     installer
\
0  69572      6000.0    14/03/2011         Roman        1390         Roman
1   8776         0.0    06/03/2013       Grumeti        1399       GRUMETI
2  34310        25.0    25/02/2013  Lottery Club         686  World vision
3  67743         0.0    28/01/2013        Unicef         263        UNICEF
4  19728         0.0    13/07/2011   Action In A           0       Artisan

   longitude   latitude             wpt_name  num_private  ... payment_ty
pe  \
0  34.938093  -9.856322                 none            0  ...     annual
ly
1  34.698766  -2.147466             Zahanati            0  ...    never p
ay
2  37.460664  -3.821329          Kwa Mahundi            0  ...   per buck
et
3  38.486161 -11.155298  Zahanati Ya Nanyumbu            0  ...    never p
ay
4  31.130847  -1.825359              Shuleni            0  ...    never p
ay

  water_quality quality_group      quantity  quantity_group  \
0          soft          good        enough          enough
1          soft          good  insufficient    insufficient
2          soft          good        enough          enough
3          soft          good           dry             dry
4          soft          good      seasonal        seasonal

                 source           source_type  source_class  \
0                spring                spring   groundwater
1  rainwater harvesting  rainwater harvesting       surface
2                   dam                   dam       surface
3           machine dbh              borehole   groundwater
4  rainwater harvesting  rainwater harvesting       surface

               waterpoint_type waterpoint_type_group
0           communal standpipe    communal standpipe
1           communal standpipe    communal standpipe
2  communal standpipe multiple    communal standpipe
3  communal standpipe multiple    communal standpipe
4           communal standpipe    communal standpipe

[5 rows x 40 columns]
<class 'pandas.core.frame.DataFrame'>
```

In [3]:
```
print (labels.head())
print(type(labels))
```

```
      id    status_group
0  69572      functional
1   8776      functional
2  34310      functional
3  67743  non functional
4  19728      functional
<class 'pandas.core.frame.DataFrame'>
```

In [4]:
```
print (test_values.head())
print(type(test_values))
```

```
      id  amount_tsh date_recorded                  funder  gps_height  \
0  50785         0.0    04/02/2013                    Dmdd        1996
1  51630         0.0    04/02/2013  Government Of Tanzania        1569
2  17168         0.0    01/02/2013                     NaN        1567
3  45559         0.0    22/01/2013              Finn Water         267
4  49871       500.0    27/03/2013                  Bruder        1260

    installer  longitude   latitude                wpt_name  num_private
\
0        DMDD  35.290799  -4.059696  Dinamu Secondary School            0
1         DWE  36.656709  -3.309214                 Kimnyak            0
2         NaN  34.767863  -5.004344           Puma Secondary            0
3  FINN WATER  38.058046  -9.418672           Kwa Mzee Pange            0
4      BRUDER  35.006123 -10.950412          Kwa Mzee Turuka            0

   ... payment_type water_quality quality_group      quantity  quantity_gr
oup  \
0  ...    never pay          soft          good      seasonal        seaso
nal
1  ...    never pay          soft          good  insufficient    insuffici
ent
2  ...    never pay          soft          good  insufficient    insuffici
ent
3  ...      unknown          soft          good           dry
dry
4  ...      monthly          soft          good        enough          eno
ugh

                 source           source_type  source_class  \
0  rainwater harvesting  rainwater harvesting       surface
1                spring                spring   groundwater
2  rainwater harvesting  rainwater harvesting       surface
3          shallow well          shallow well   groundwater
4                spring                spring   groundwater

      waterpoint_type waterpoint_type_group
0               other                 other
1  communal standpipe    communal standpipe
2               other                 other
3               other                 other
4  communal standpipe    communal standpipe

[5 rows x 40 columns]
<class 'pandas.core.frame.DataFrame'>
```

In [5]:
```
test_values.describe
```

Out[5]:
```
<bound method NDFrame.describe of           id  amount_tsh date_recorded
funder  gps_height  \
0      50785         0.0    04/02/2013                    Dmdd        1996
1      51630         0.0    04/02/2013  Government Of Tanzania        1569
2      17168         0.0    01/02/2013                     NaN        1567
3      45559         0.0    22/01/2013              Finn Water         267
4      49871       500.0    27/03/2013                  Bruder        1260
...      ...         ...           ...                     ...         ...
14845  39307         0.0    24/02/2011                  Danida          34
14846  18990      1000.0    21/03/2011                    Hiap           0
14847  28749         0.0    04/03/2013                     NaN        1476
14848  33492         0.0    18/02/2013                 Germany         998
14849  68707         0.0    13/02/2013  Government Of Tanzania         481

        installer  longitude   latitude                wpt_name  num_priv
ate  \
0            DMDD  35.290799  -4.059696  Dinamu Secondary School
0
1             DWE  36.656709  -3.309214                  Kimnyak
0
2             NaN  34.767863  -5.004344           Puma Secondary
0
3      FINN WATER  38.058046  -9.418672           Kwa Mzee Pange
0
4          BRUDER  35.006123 -10.950412          Kwa Mzee Turuka
0
...           ...        ...        ...                      ...
...
14845          Da  38.852669  -6.582841                Kwambwezi
0
14846        HIAP  37.451633  -5.350428         Bonde La Mkondoa
0
14847         NaN  34.739804  -4.585587                  Bwawani
0
14848         DWE  35.432732 -10.584159                 Kwa John
0
14849  Government  34.765054 -11.226012         Kwa Mzee Chagala
0

       ... payment_type water_quality quality_group      quantity  \
0      ...    never pay          soft          good      seasonal
1      ...    never pay          soft          good  insufficient
2      ...    never pay          soft          good  insufficient
3      ...      unknown          soft          good           dry
4      ...      monthly          soft          good        enough
...    ...          ...           ...           ...           ...
14845  ...    never pay          soft          good        enough
14846  ...     annually         salty         salty  insufficient
14847  ...    never pay          soft          good  insufficient
14848  ...    never pay          soft          good  insufficient
14849  ...    never pay          soft          good           dry

       quantity_group                source           source_type  \
0            seasonal  rainwater harvesting  rainwater harvesting
1        insufficient                spring                spring
2        insufficient  rainwater harvesting  rainwater harvesting
3                 dry          shallow well          shallow well
4              enough                spring                spring
...               ...                   ...                   ...
14845          enough                 river            river/lake
14846    insufficient          shallow well          shallow well
```

```
14847    insufficient                   dam                   dam
14848    insufficient                 river            river/lake
14849             dry                spring                spring

       source_class     waterpoint_type waterpoint_type_group
0           surface               other                 other
1       groundwater  communal standpipe    communal standpipe
2           surface               other                 other
3       groundwater               other                 other
4       groundwater  communal standpipe    communal standpipe
...             ...                 ...                   ...
14845       surface  communal standpipe    communal standpipe
14846   groundwater           hand pump             hand pump
14847       surface  communal standpipe    communal standpipe
14848       surface  communal standpipe    communal standpipe
14849   groundwater  communal standpipe    communal standpipe

[14850 rows x 40 columns]>
```

Data Preprocessing

In [7]:
```
#Merging values and labels on ID
data = pd.merge(values, labels.copy(), on='id',how ='inner')
```

In [8]:
```
data.head()
print(type(data))
```

```
<class 'pandas.core.frame.DataFrame'>
```

In [9]:
```
data.info
```

Out[9]:
```
<bound method DataFrame.info of          id  amount_tsh date_recorded
funder  gps_height  \
0      69572      6000.0    14/03/2011            Roman        1390
1       8776         0.0    06/03/2013          Grumeti        1399
2      34310        25.0    25/02/2013     Lottery Club         686
3      67743         0.0    28/01/2013           Unicef         263
4      19728         0.0    13/07/2011      Action In A           0
...      ...         ...           ...              ...         ...
59395  60739        10.0    03/05/2013  Germany Republi        1210
59396  27263      4700.0    07/05/2011      Cefa-njombe        1212
59397  37057         0.0    11/04/2011              NaN           0
59398  31282         0.0    08/03/2011            Malec           0
59399  26348         0.0    23/03/2011       World Bank         191

          installer  longitude   latitude              wpt_name  num_priva
te  \
0             Roman  34.938093  -9.856322                  none
0
1           GRUMETI  34.698766  -2.147466              Zahanati
0
2      World vision  37.460664  -3.821329           Kwa Mahundi
0
3            UNICEF  38.486161 -11.155298  Zahanati Ya Nanyumbu
0
4           Artisan  31.130847  -1.825359               Shuleni
0
...             ...        ...        ...                   ...
...
59395           CES  37.169807  -3.253847   Area Three Namba 27
0
59396          Cefa  35.249991  -9.070629     Kwa Yahona Kuvala
0
59397           NaN  34.017087  -8.750434               Mashine
0
59398          Musa  35.861315  -6.378573                Mshoro
0
59399         World  38.104048  -6.747464       Kwa Mzee Lugawa
0

       ... water_quality quality_group      quantity  quantity_group  \
0      ...          soft          good        enough          enough
1      ...          soft          good  insufficient    insufficient
2      ...          soft          good        enough          enough
3      ...          soft          good           dry             dry
4      ...          soft          good      seasonal        seasonal
...    ...           ...           ...           ...             ...
59395  ...          soft          good        enough          enough
59396  ...          soft          good        enough          enough
59397  ...      fluoride      fluoride        enough          enough
59398  ...          soft          good  insufficient    insufficient
59399  ...         salty         salty        enough          enough

                     source           source_type source_class  \
0                    spring                spring  groundwater
1      rainwater harvesting  rainwater harvesting      surface
2                       dam                   dam      surface
3               machine dbh              borehole  groundwater
4      rainwater harvesting  rainwater harvesting      surface
...                     ...                   ...          ...
59395                spring                spring  groundwater
59396                 river            river/lake      surface
```

```
59397           machine dbh              borehole  groundwater
59398          shallow well          shallow well  groundwater
59399          shallow well          shallow well  groundwater

                   waterpoint_type waterpoint_type_group    status_group
0               communal standpipe    communal standpipe      functional
1               communal standpipe    communal standpipe      functional
2      communal standpipe multiple    communal standpipe      functional
3      communal standpipe multiple    communal standpipe  non functional
4               communal standpipe    communal standpipe      functional
...                            ...                   ...             ...
59395           communal standpipe    communal standpipe      functional
59396           communal standpipe    communal standpipe      functional
59397                    hand pump             hand pump      functional
59398                    hand pump             hand pump      functional
59399                    hand pump             hand pump      functional

[59400 rows x 41 columns]>
```

In [10]:
```
print (data.describe)
```

```
<bound method NDFrame.describe of          id  amount_tsh date_recorded
funder  gps_height  \
0      69572      6000.0    14/03/2011            Roman        1390
1       8776         0.0    06/03/2013          Grumeti        1399
2      34310        25.0    25/02/2013     Lottery Club         686
3      67743         0.0    28/01/2013           Unicef         263
4      19728         0.0    13/07/2011      Action In A           0
...      ...         ...           ...              ...         ...
59395  60739        10.0    03/05/2013  Germany Republi        1210
59396  27263      4700.0    07/05/2011      Cefa-njombe        1212
59397  37057         0.0    11/04/2011              NaN           0
59398  31282         0.0    08/03/2011            Malec           0
59399  26348         0.0    23/03/2011       World Bank         191

          installer  longitude   latitude              wpt_name  num_priva
te  \
0             Roman  34.938093  -9.856322                  none
0
1           GRUMETI  34.698766  -2.147466              Zahanati
0
2      World vision  37.460664  -3.821329           Kwa Mahundi
0
3            UNICEF  38.486161 -11.155298  Zahanati Ya Nanyumbu
0
4           Artisan  31.130847  -1.825359               Shuleni
0
...             ...        ...        ...                   ...
...
59395           CES  37.169807  -3.253847   Area Three Namba 27
0
59396          Cefa  35.249991  -9.070629     Kwa Yahona Kuvala
0
59397           NaN  34.017087  -8.750434               Mashine
0
59398          Musa  35.861315  -6.378573                Mshoro
0
59399         World  38.104048  -6.747464       Kwa Mzee Lugawa
0

       ... water_quality quality_group      quantity  quantity_group  \
0      ...          soft          good        enough          enough
1      ...          soft          good  insufficient    insufficient
2      ...          soft          good        enough          enough
3      ...          soft          good           dry             dry
4      ...          soft          good      seasonal        seasonal
...    ...           ...           ...           ...             ...
59395  ...          soft          good        enough          enough
59396  ...          soft          good        enough          enough
59397  ...      fluoride      fluoride        enough          enough
59398  ...          soft          good  insufficient    insufficient
59399  ...         salty         salty        enough          enough

                     source           source_type source_class  \
0                    spring                spring  groundwater
1      rainwater harvesting  rainwater harvesting      surface
2                       dam                   dam      surface
3               machine dbh              borehole  groundwater
4      rainwater harvesting  rainwater harvesting      surface
...                     ...                   ...          ...
59395                spring                spring  groundwater
59396                 river            river/lake      surface
```

```
59397           machine dbh              borehole  groundwater
59398          shallow well          shallow well  groundwater
59399          shallow well          shallow well  groundwater

                   waterpoint_type waterpoint_type_group    status_group
0               communal standpipe    communal standpipe      functional
1               communal standpipe    communal standpipe      functional
2      communal standpipe multiple    communal standpipe      functional
3      communal standpipe multiple    communal standpipe  non functional
4               communal standpipe    communal standpipe      functional
...                            ...                   ...             ...
59395           communal standpipe    communal standpipe      functional
59396           communal standpipe    communal standpipe      functional
59397                    hand pump             hand pump      functional
59398                    hand pump             hand pump      functional
59399                    hand pump             hand pump      functional

[59400 rows x 41 columns]>
```

In [11]:
```
print(data.isna().sum())
```

```
id                         0
amount_tsh                 0
date_recorded              0
funder                  3637
gps_height                 0
installer               3655
longitude                  0
latitude                   0
wpt_name                   2
num_private                0
basin                      0
subvillage               371
region                     0
region_code                0
district_code              0
lga                        0
ward                       0
population                 0
public_meeting          3334
recorded_by                0
scheme_management       3878
scheme_name            28810
permit                  3056
construction_year          0
extraction_type            0
extraction_type_group      0
extraction_type_class      0
management                 0
management_group           0
payment                    0
payment_type               0
water_quality              0
quality_group              0
quantity                   0
quantity_group             0
source                     0
source_type                0
source_class               0
waterpoint_type            0
waterpoint_type_group      0
status_group               0
dtype: int64
```

In [12]:
```
print(data['status_group'].value_counts())
```

```
status_group
functional                 32259
non functional             22824
functional needs repair     4317
Name: count, dtype: int64
```

In [13]:
```
missing_v = data.isna().mean()*100
missing_v
```

Out[13]:
```
id                        0.000000
amount_tsh                0.000000
date_recorded             0.000000
funder                    6.122896
gps_height                0.000000
installer                 6.153199
longitude                 0.000000
latitude                  0.000000
wpt_name                  0.003367
num_private               0.000000
basin                     0.000000
subvillage                0.624579
region                    0.000000
region_code               0.000000
district_code             0.000000
lga                       0.000000
ward                      0.000000
population                0.000000
public_meeting            5.612795
recorded_by               0.000000
scheme_management         6.528620
scheme_name              48.501684
permit                    5.144781
construction_year         0.000000
extraction_type           0.000000
extraction_type_group     0.000000
extraction_type_class     0.000000
management                0.000000
management_group          0.000000
payment                   0.000000
payment_type              0.000000
water_quality             0.000000
quality_group             0.000000
quantity                  0.000000
quantity_group            0.000000
source                    0.000000
source_type               0.000000
source_class              0.000000
waterpoint_type           0.000000
waterpoint_type_group     0.000000
status_group              0.000000
dtype: float64
```

Handling missing Values

In [15]:
```
print(type(data))
```

```
<class 'pandas.core.frame.DataFrame'>
```

In [16]:
```
#The scheme_name column has close to half of its values missing. it has to be dropped.
data =data.drop("scheme_name", axis =1)
data.head()
```

Out[16]:

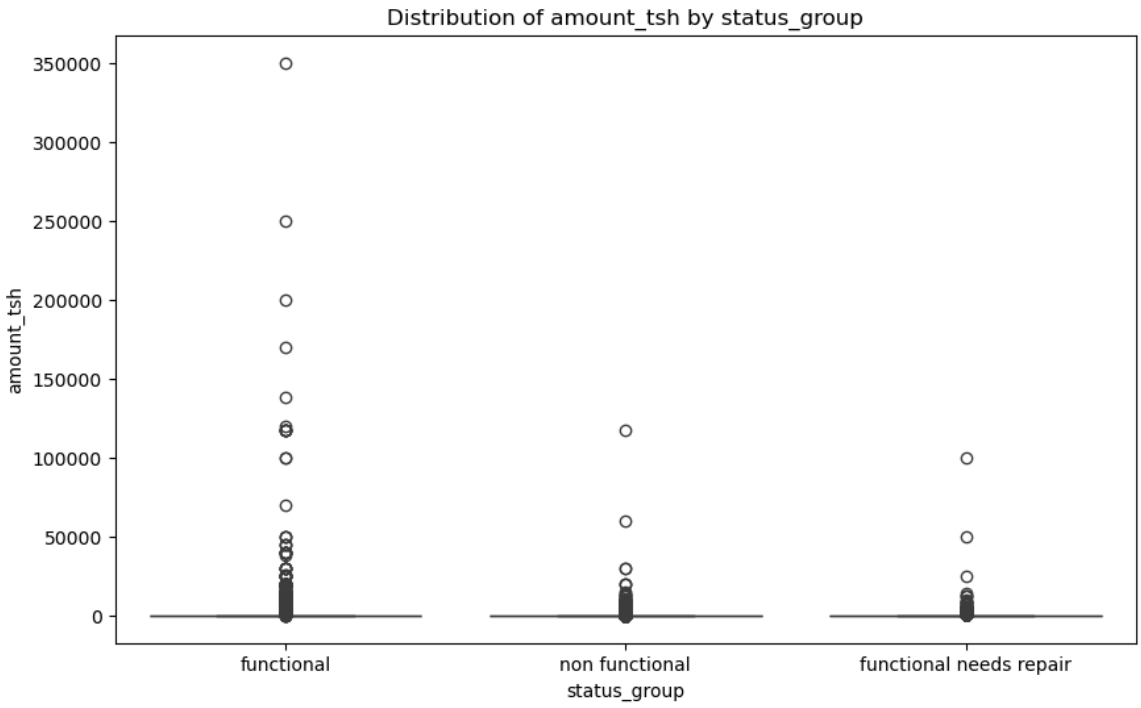
	id	amount_tsh	date_recorded	funder	gps_height	installer	longitude	latitude
0	69572	6000.0	14/03/2011	Roman	1390	Roman	34.938093	-9.856322
1	8776	0.0	06/03/2013	Grumeti	1399	GRUMETI	34.698766	-2.147466
2	34310	25.0	25/02/2013	Lottery Club	686	World vision	37.460664	-3.821329
3	67743	0.0	28/01/2013	Unicef	263	UNICEF	38.486161	-11.155298
4	19728	0.0	13/07/2011	Action In A	0	Artisan	31.130847	-1.825359

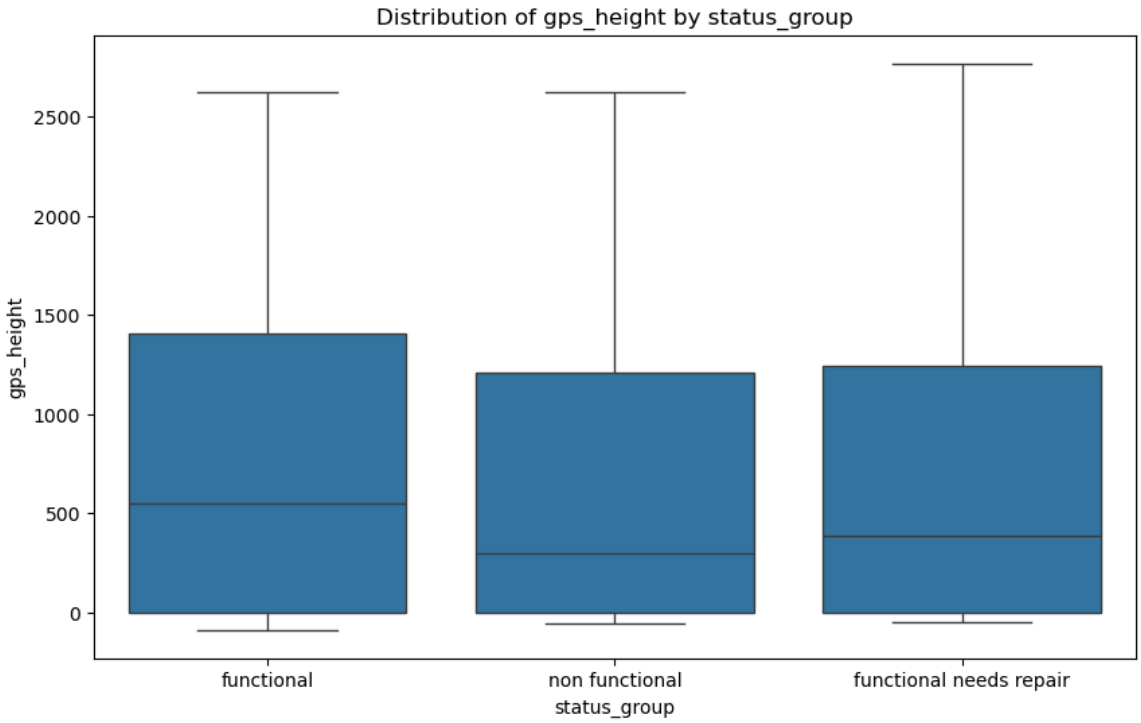
5 rows × 40 columns

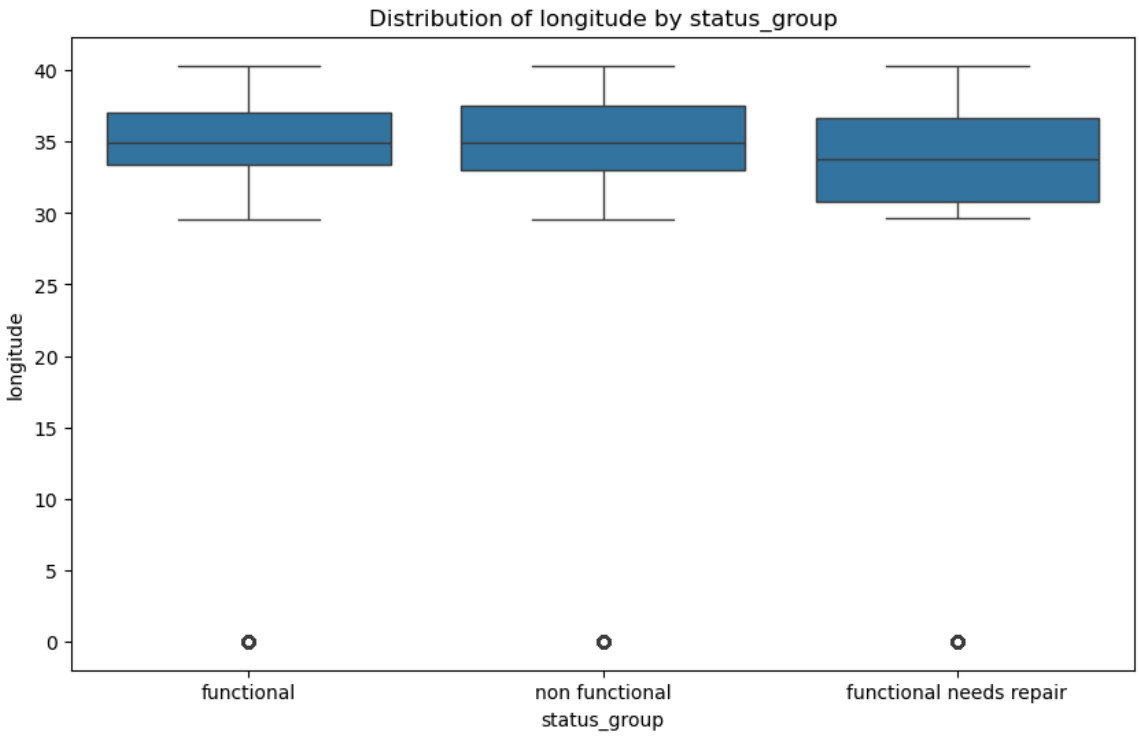
In [17]:
```
from sklearn.impute import SimpleImputer
imputed_data = data.copy()
cat_imputer = SimpleImputer(strategy='most_frequent')
imputed_data[['permit', 'scheme_management', 'public_meeting',
              'subvillage', 'funder', 'installer', 'wpt_name']] = \
    cat_imputer.fit_transform(imputed_data[['permit', 'scheme_management',
'public_meeting',
                                            'subvillage', 'funder', 'insta
ller', 'wpt_name']])
print(imputed_data.isna().sum())
```

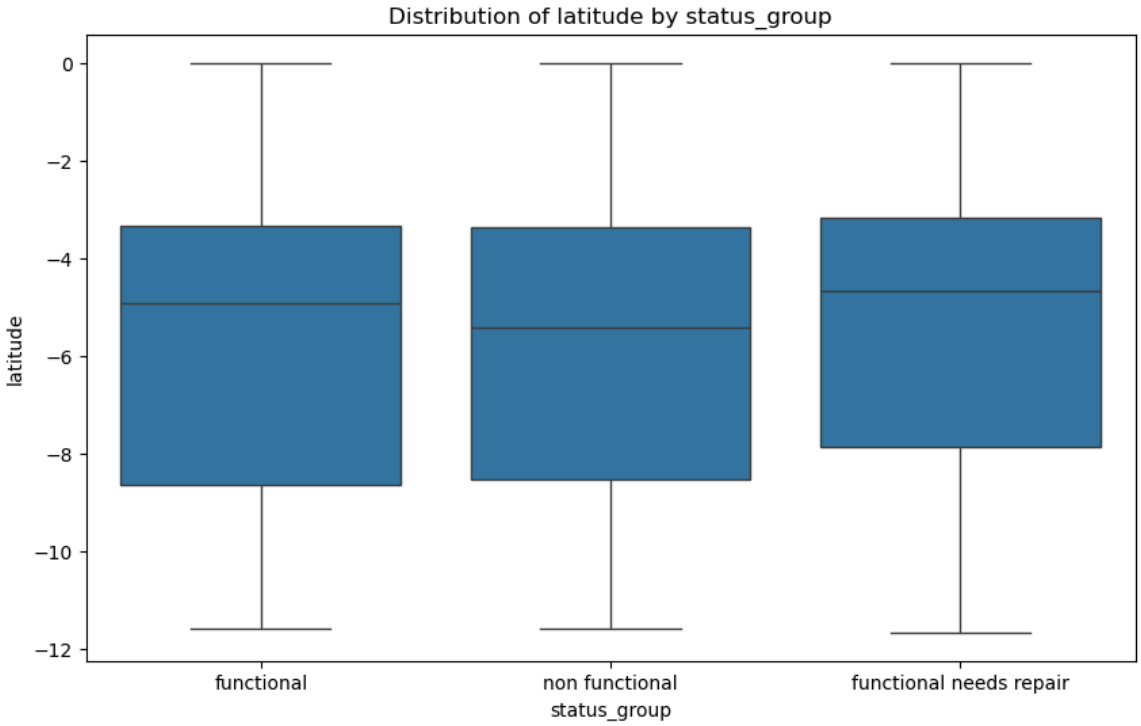
```
id                       0
amount_tsh               0
date_recorded            0
funder                   0
gps_height               0
installer                0
longitude                0
latitude                 0
wpt_name                 0
num_private              0
basin                    0
subvillage               0
region                   0
region_code              0
district_code            0
lga                      0
ward                     0
population               0
public_meeting           0
recorded_by              0
scheme_management        0
permit                   0
construction_year        0
extraction_type          0
extraction_type_group    0
extraction_type_class    0
management               0
management_group         0
payment                  0
payment_type             0
water_quality            0
quality_group            0
quantity                 0
quantity_group           0
source                   0
source_type              0
source_class             0
waterpoint_type          0
waterpoint_type_group    0
status_group             0
dtype: int64
```

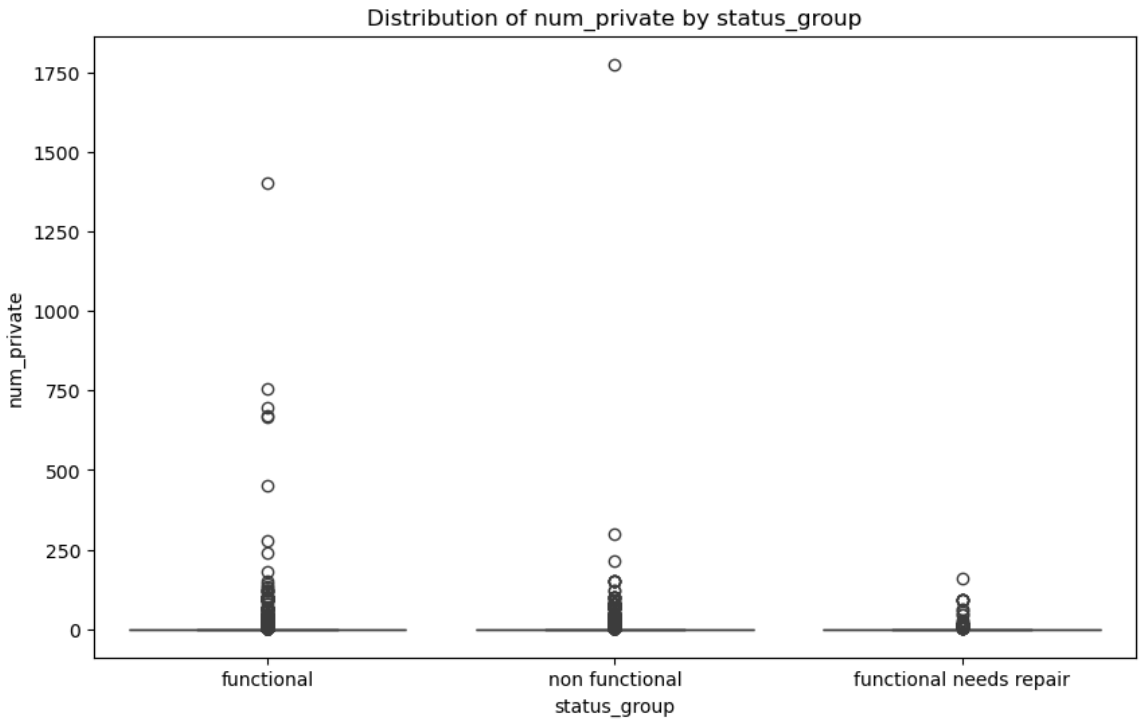
In [18]:
```
for col in imputed_data.select_dtypes(include=[np.number]).columns:
    if col != 'status_group' and col != 'id':
        plt.figure(figsize=(10, 6))
        sns.boxplot(x='status_group', y=col, data=imputed_data)
        plt.title(f'Distribution of {col} by status_group')
        plt.show()
```

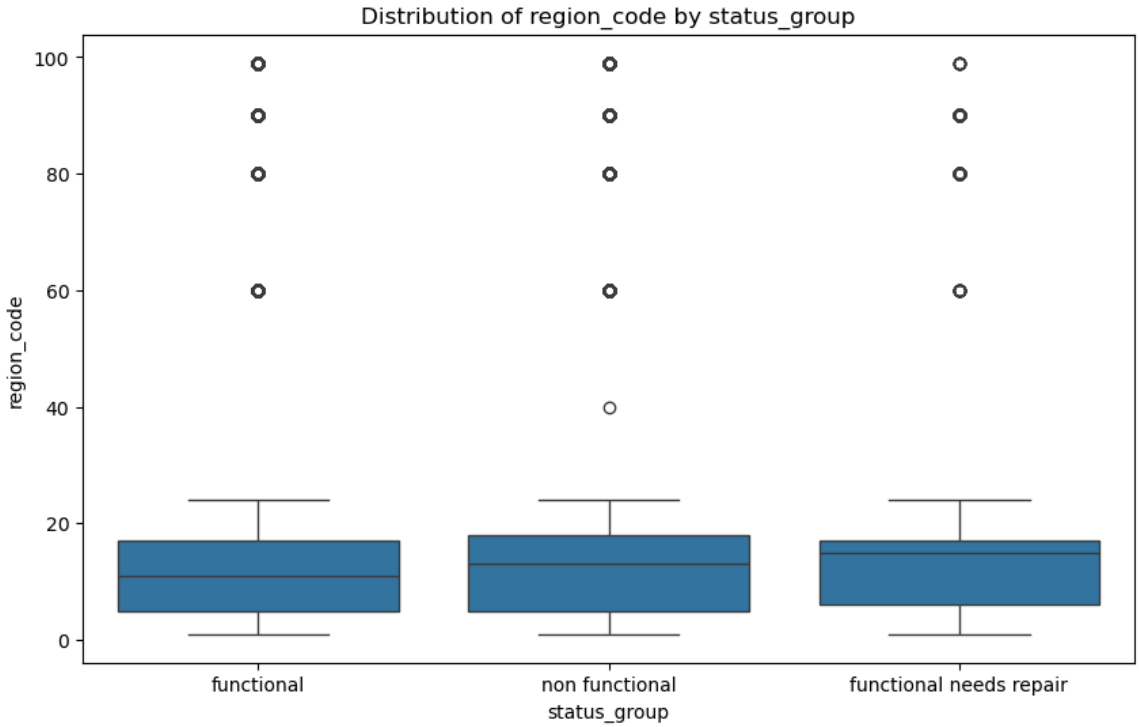

Distribution of amount_tsh by status_group
amount_tsh
350000
300000
250000
200000
150000
100000
50000
0
functional
non functional
functional needs repair
status_group

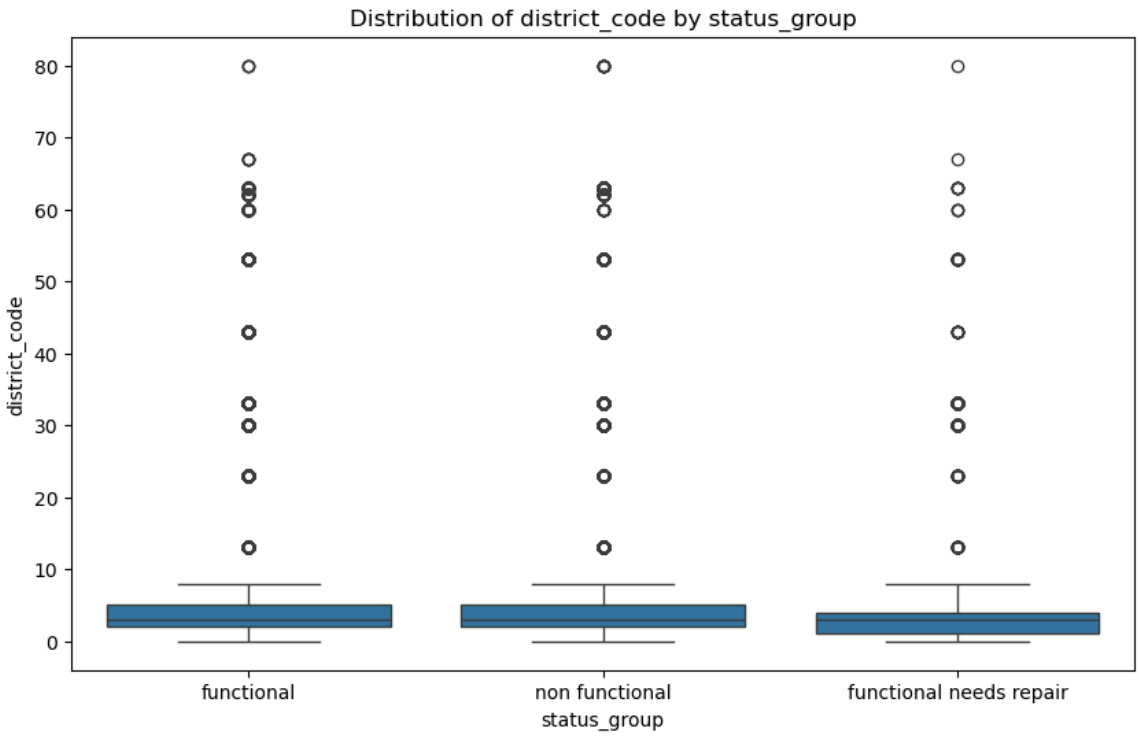

Distribution of gps_height by status_group
gps_height
2500
2000
1500
1000
500
0
functional
non functional
functional needs repair
status_group

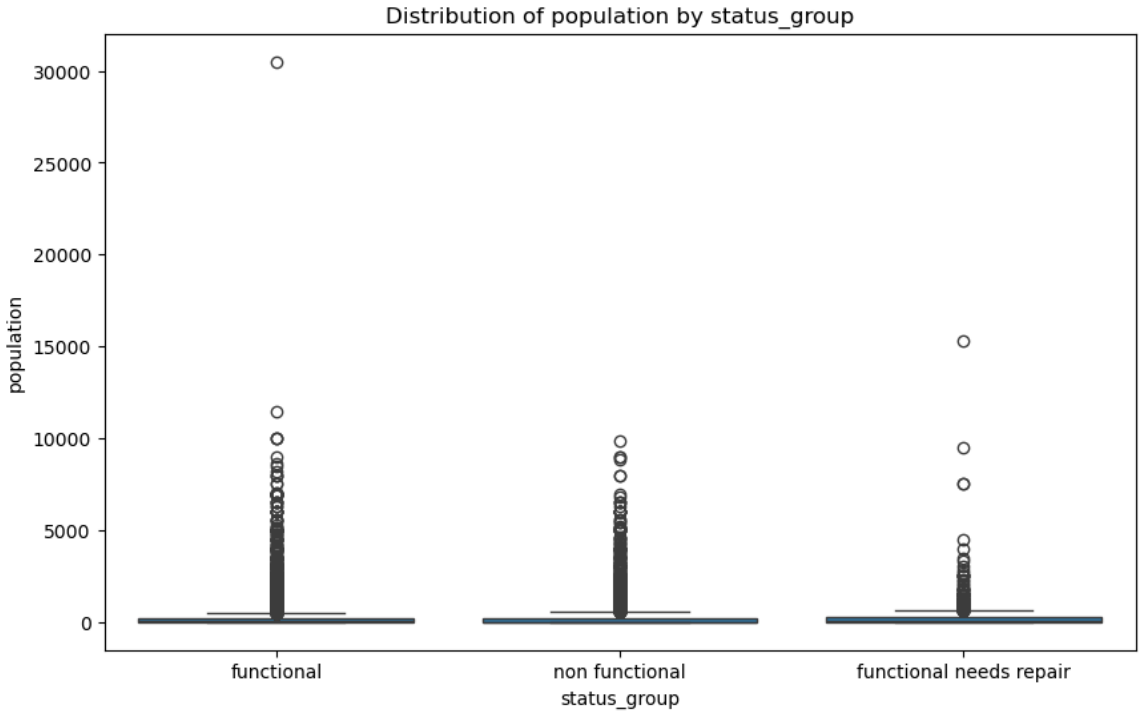

Distribution of longitude by status_group
longitude
40
35
30
25
20
15
10
5
0
functional
non functional
functional needs repair
status_group

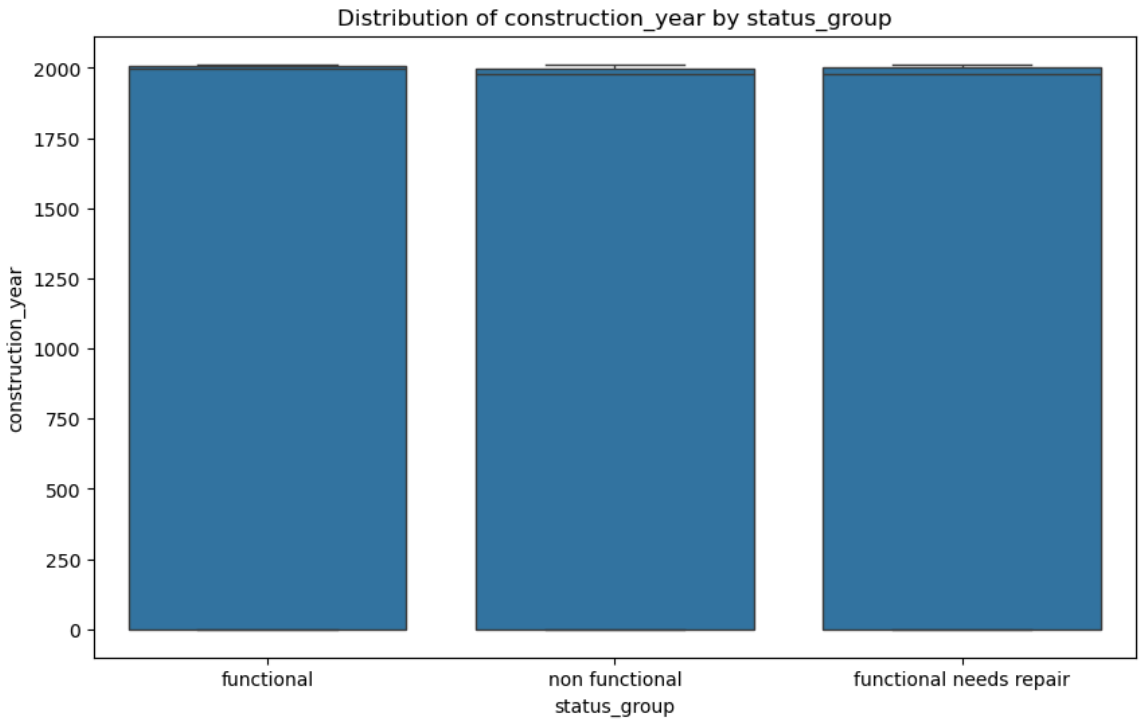

Distribution of latitude by status_group
latitude
0
−2
−4
−6
−8
−10
−12
functional
non functional
functional needs repair
status_group


Distribution of num_private by status_group
num_private
1750
1500
1250
1000
750
500
250
0
functional
non functional
functional needs repair
status_group

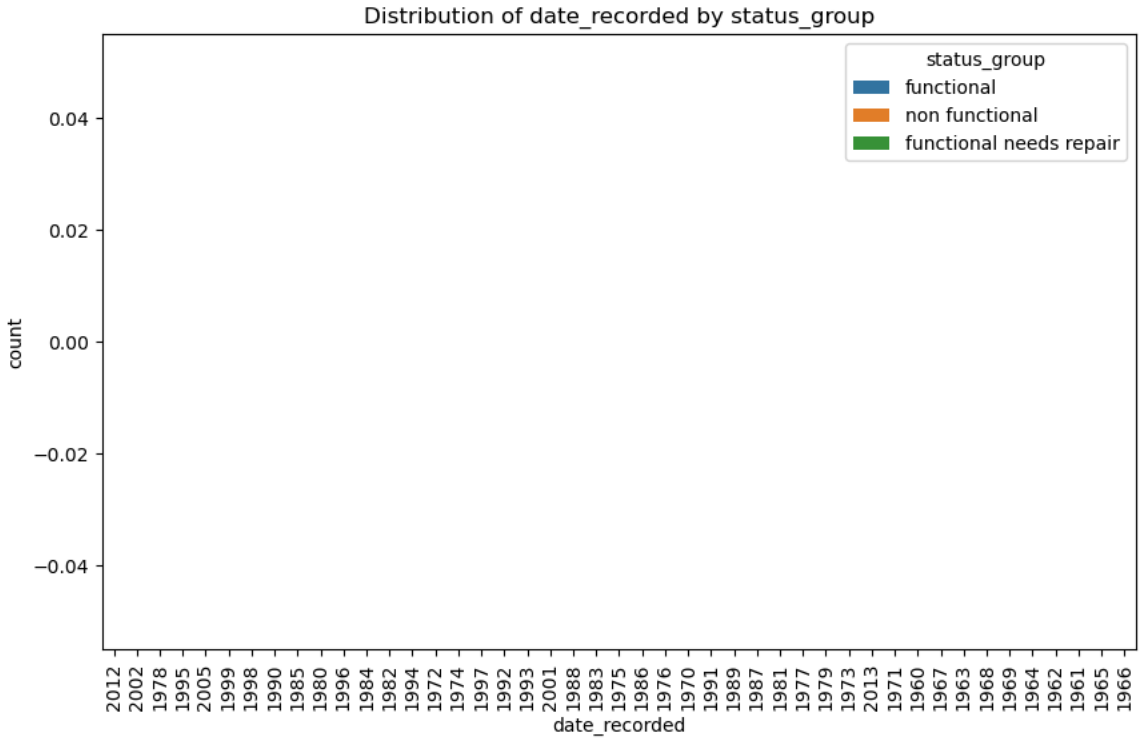

Distribution of region_code by status_group
region_code
100
80
60
40
20
0
functional
non functional
functional needs repair
status_group

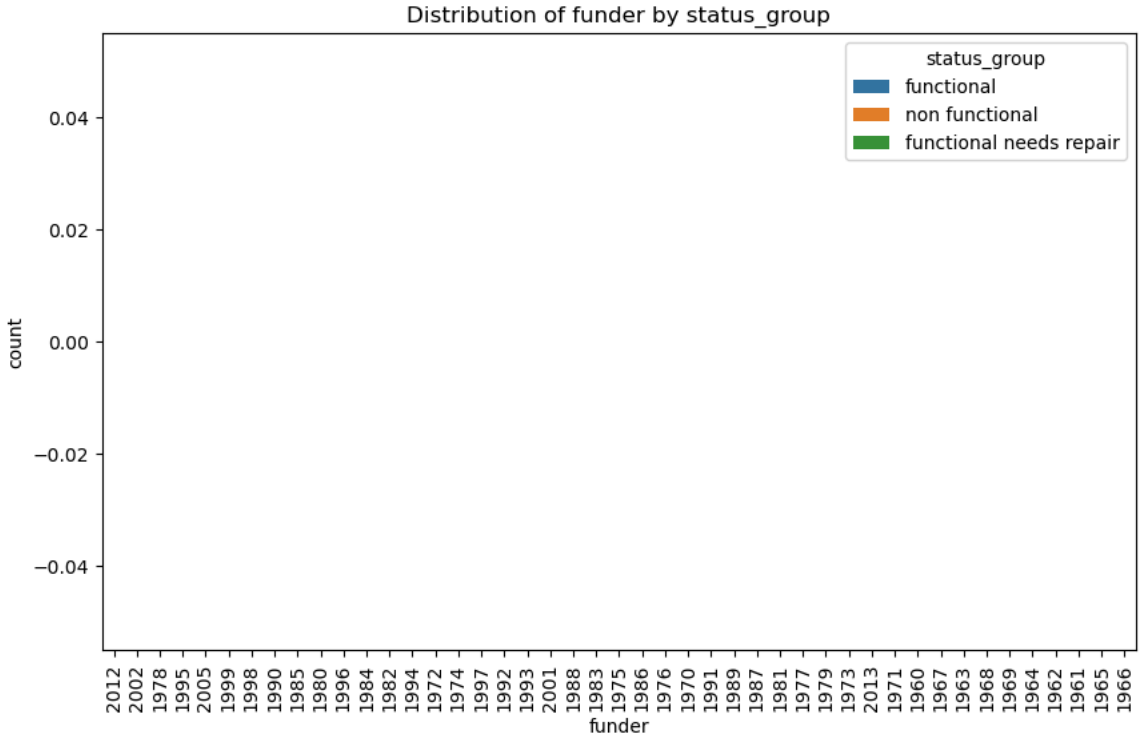

Distribution of district_code by status_group
district_code
80
70
60
50
40
30
20
10
0
functional
non functional
functional needs repair
status_group

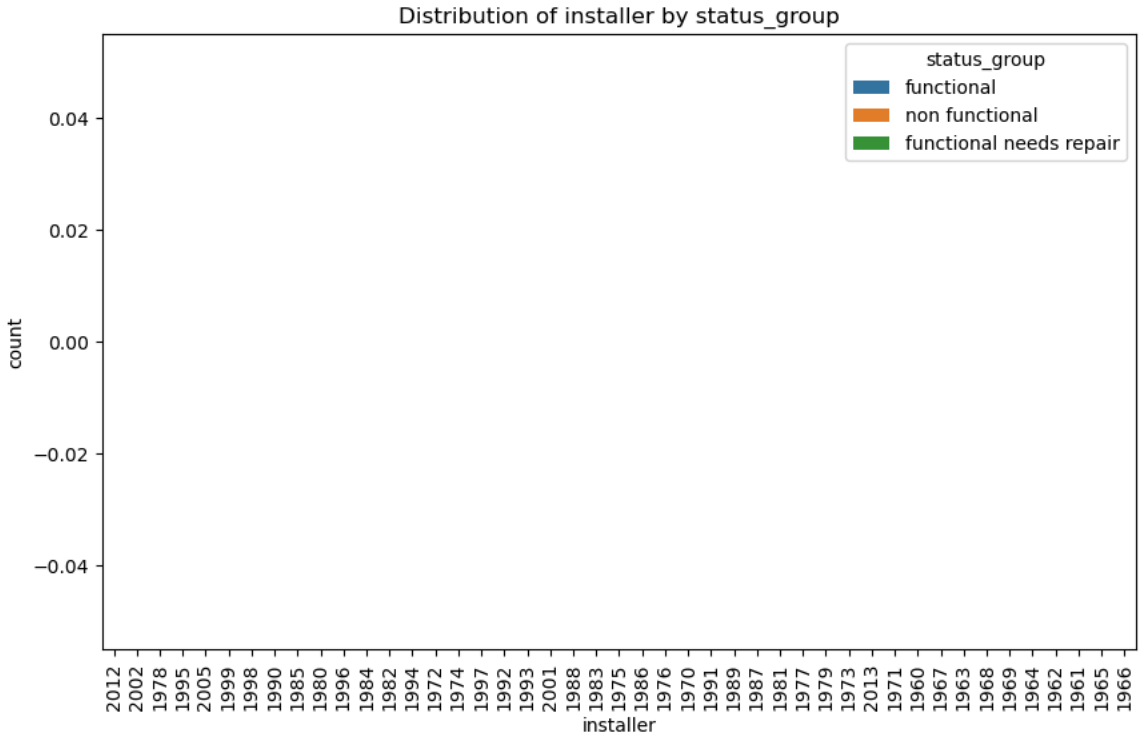

Distribution of population by status_group
population
30000
25000
20000
15000
10000
5000
0
functional
non functional
functional needs repair
status_group


Distribution of construction_year by status_group
construction_year
2000
1750
1500
1250
1000
750
500
250
0
functional
non functional
functional needs repair
status_group

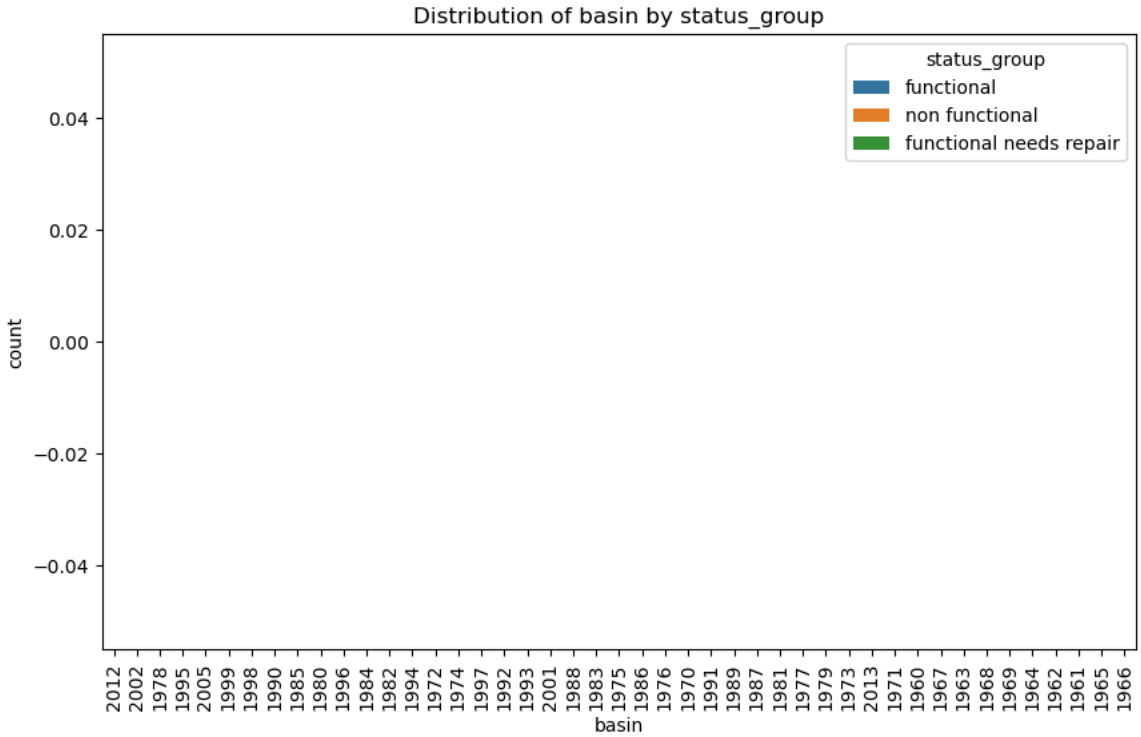
In [19]:
```
order = imputed_data[col].value_counts().index[10:]
for col in imputed_data.select_dtypes(include=['object']).columns:
    plt.figure(figsize=(10, 6))
    sns.countplot(x=col, hue='status_group', data=imputed_data, order=orde
r)
    plt.xticks(rotation=90)
    plt.title(f'Distribution of {col} by status_group')
    plt.show()
```

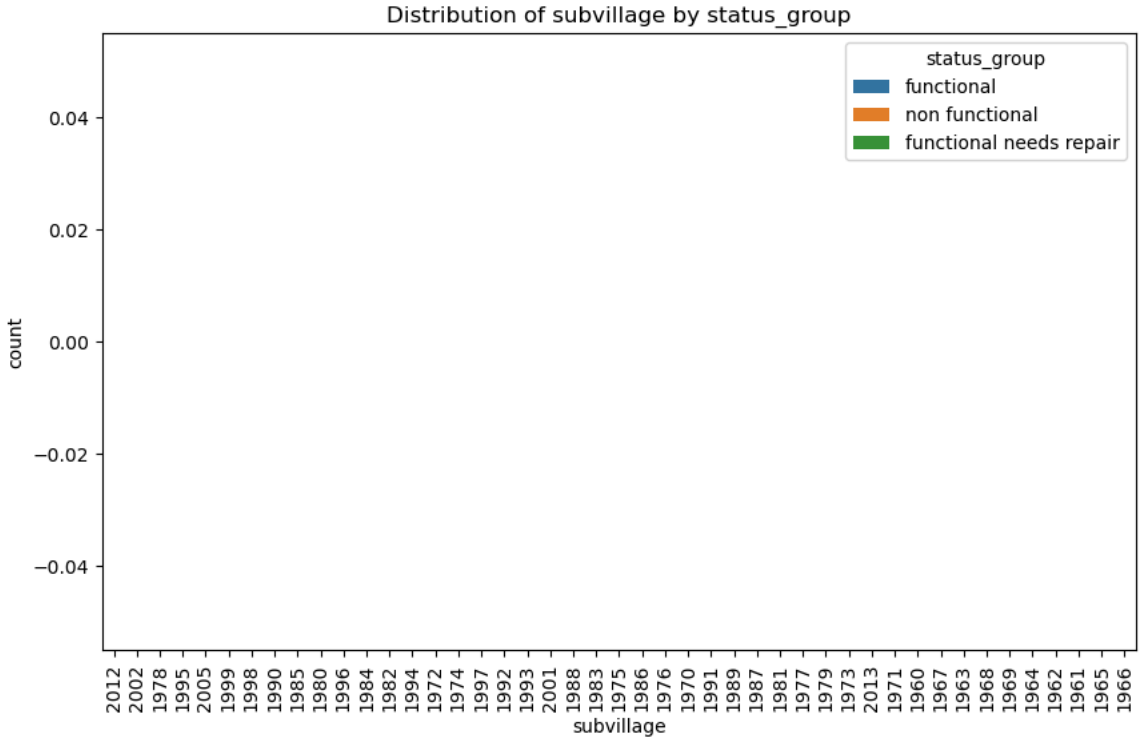

Distribution of date_recorded by status_group
status_group
functional
non functional
functional needs repair
count
0.04
0.02
0.00
−0.02
−0.04
2012 2002 1978 1995 2005 1999 1998 1990 1985 1980 1996 1984 1982 1994 1972 1974 1997 1992 1993 2001 1988 1983 1975 1986 1976 1970 1991 1989 1987 1981 1977 1979 1973 2013 1971 1960 1967 1963 1968 1969 1964 1962 1961 1965 1966
date_recorded

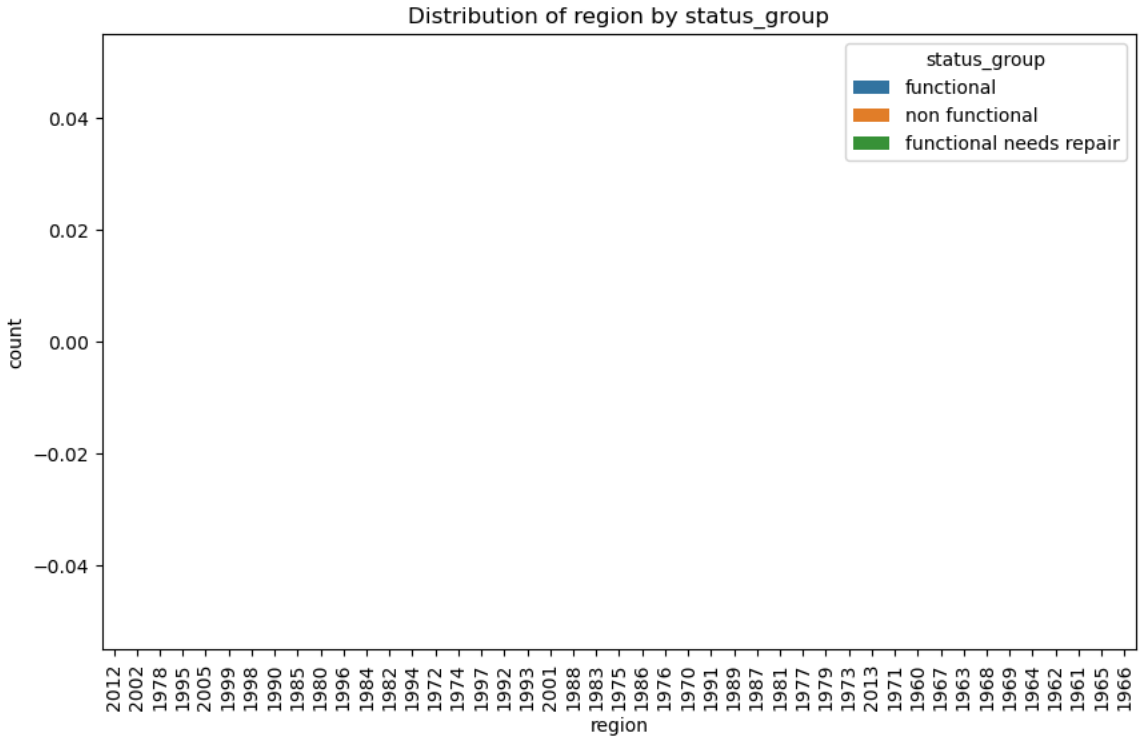

Distribution of funder by status_group
status_group
functional
non functional
functional needs repair
count
0.04
0.02
0.00
−0.02
−0.04
2012 2002 1978 1995 2005 1999 1998 1990 1985 1980 1996 1984 1982 1994 1972 1974 1997 1992 1993 2001 1988 1983 1975 1986 1976 1970 1991 1989 1987 1981 1977 1979 1973 2013 1971 1960 1967 1963 1968 1969 1964 1962 1961 1965 1966
funder

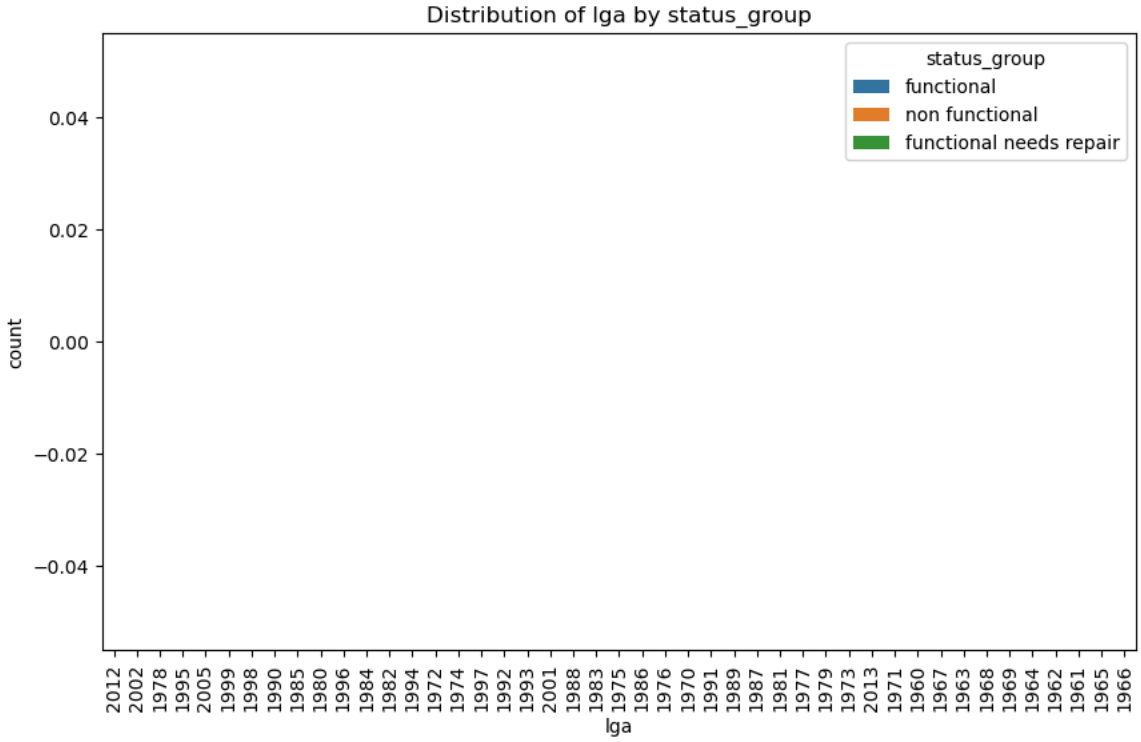

Distribution of installer by status_group
status_group
functional
non functional
functional needs repair
count
0.04
0.02
0.00
−0.02
−0.04
2012 2002 1978 1995 2005 1999 1998 1990 1985 1980 1996 1984 1982 1994 1972 1974 1997 1992 1993 2001 1988 1983 1975 1986 1976 1970 1991 1989 1987 1981 1977 1979 1973 2013 1971 1960 1967 1963 1968 1969 1964 1962 1961 1965 1966
installer

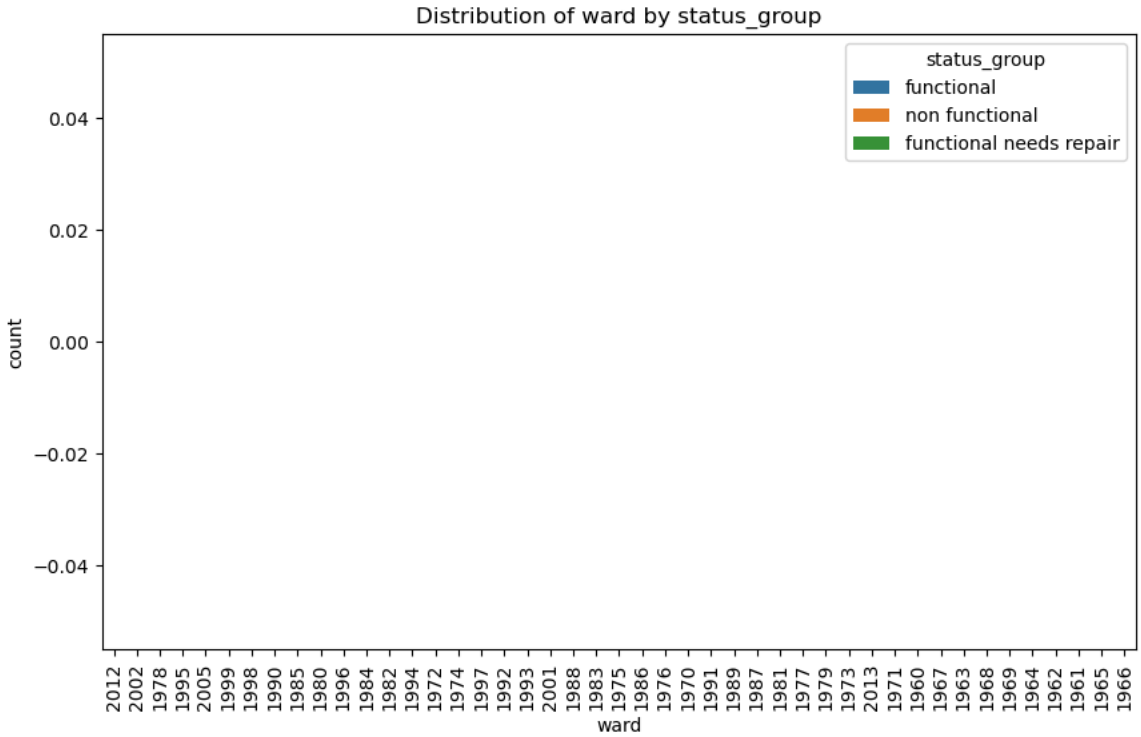

Distribution of wpt_name by status_group
status_group
functional
non functional
functional needs repair
count
0.04
0.02
0.00
−0.02
−0.04
2012 2002 1978 1995 2005 1999 1998 1990 1985 1980 1996 1984 1982 1994 1972 1974 1997 1992 1993 2001 1988 1983 1975 1986 1976 1970 1991 1989 1987 1981 1977 1979 1973 2013 1971 1960 1967 1963 1968 1969 1964 1962 1961 1965 1966
wpt_name

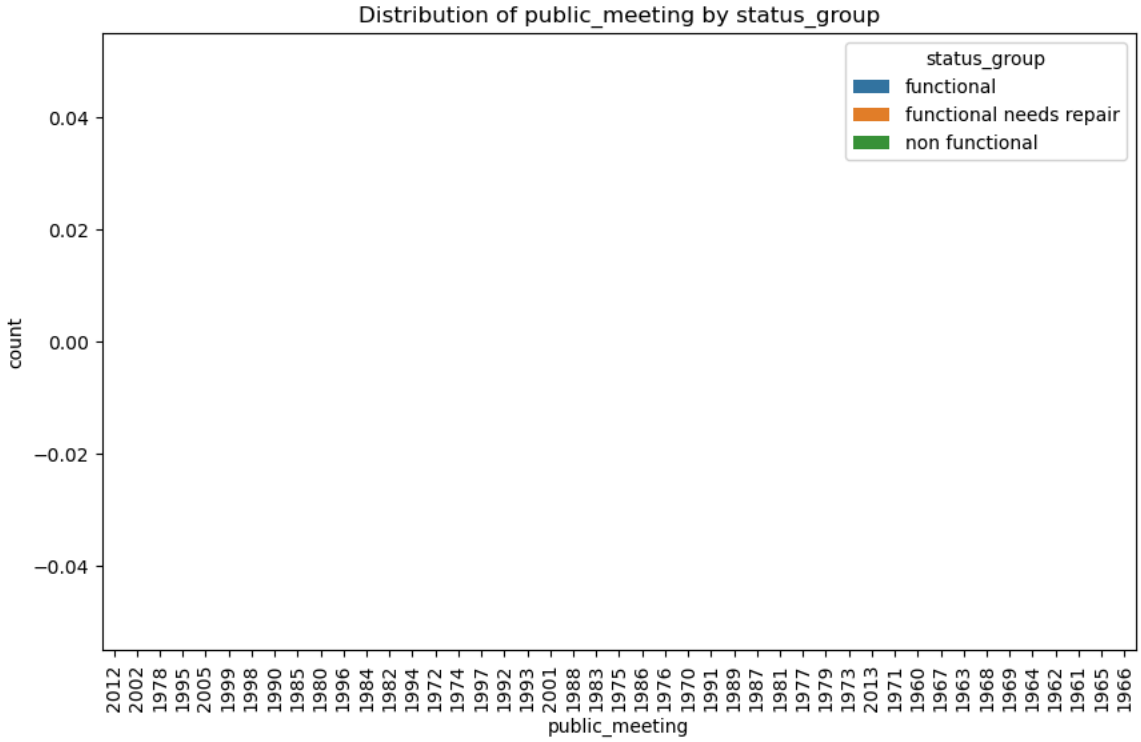

Distribution of basin by status_group
status_group
functional
non functional
functional needs repair
count
0.04
0.02
0.00
−0.02
−0.04
2012 2002 1978 1995 2005 1999 1998 1990 1985 1980 1996 1984 1982 1994 1972 1974 1997 1992 1993 2001 1988 1983 1975 1986 1976 1970 1991 1989 1987 1981 1977 1979 1973 2013 1971 1960 1967 1963 1968 1969 1964 1962 1961 1965 1966
basin


Distribution of subvillage by status_group
status_group
functional
non functional
functional needs repair
count
0.04
0.02
0.00
−0.02
−0.04
2012 2002 1978 1995 2005 1999 1998 1990 1985 1980 1996 1984 1982 1994 1972 1974 1997 1992 1993 2001 1988 1983 1975 1986 1976 1970 1991 1989 1987 1981 1977 1979 1973 2013 1971 1960 1967 1963 1968 1969 1964 1962 1961 1965 1966
subvillage

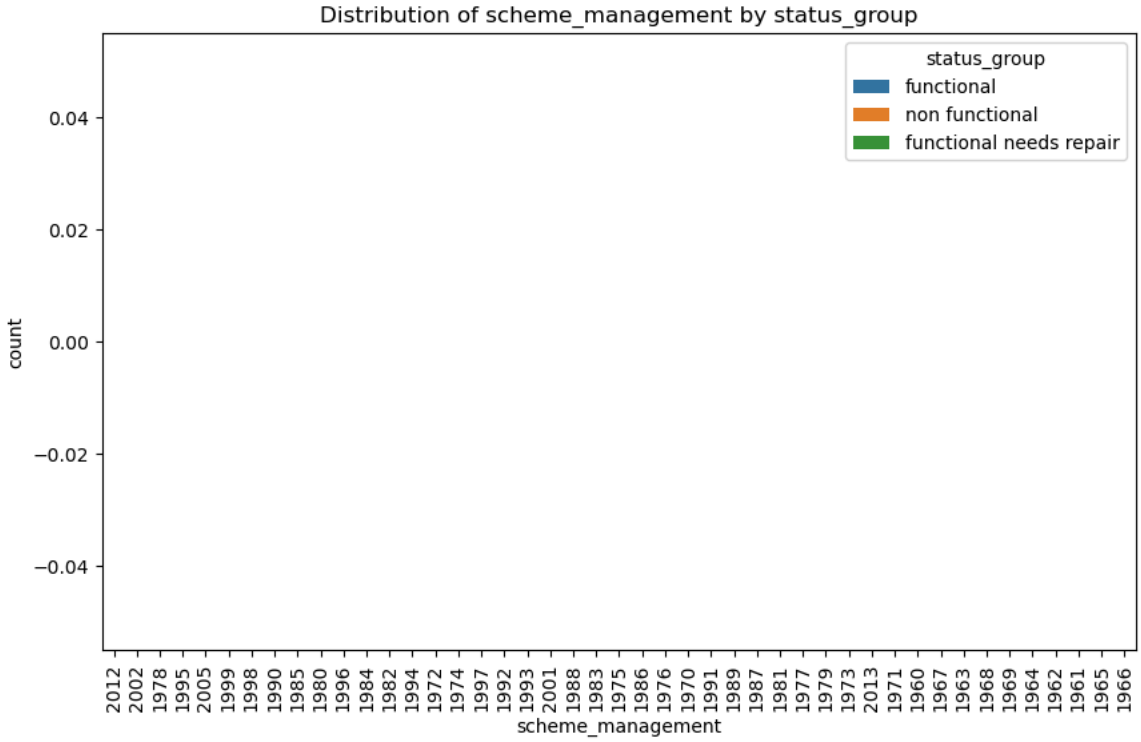

Distribution of region by status_group
status_group
functional
non functional
functional needs repair
count
0.04
0.02
0.00
−0.02
−0.04
2012 2002 1978 1995 2005 1999 1998 1990 1985 1980 1996 1984 1982 1994 1972 1974 1997 1992 1993 2001 1988 1983 1975 1986 1976 1970 1991 1989 1987 1981 1977 1979 1973 2013 1971 1960 1967 1963 1968 1969 1964 1962 1961 1965 1966
region

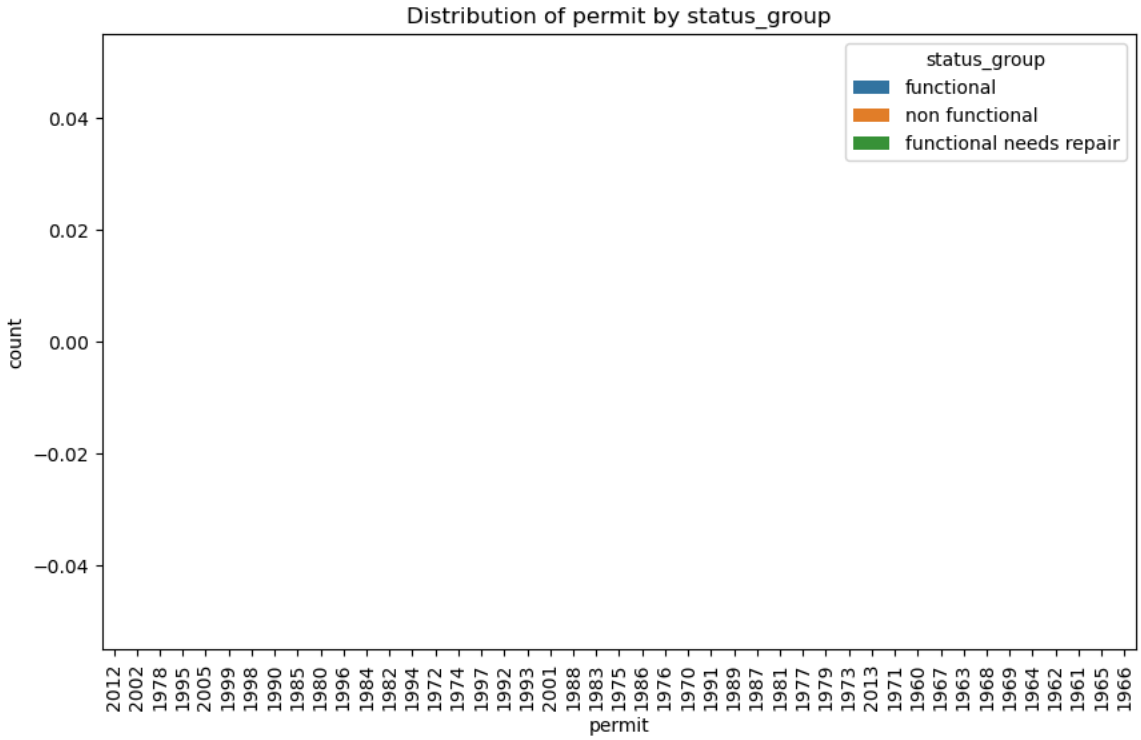

Distribution of lga by status_group
status_group
functional
non functional
functional needs repair
count
0.04
0.02
0.00
−0.02
−0.04
2012 2002 1978 1995 2005 1999 1998 1990 1985 1980 1996 1984 1982 1994 1972 1974 1997 1992 1993 2001 1988 1983 1975 1986 1976 1970 1991 1989 1987 1981 1977 1979 1973 2013 1971 1960 1967 1963 1968 1969 1964 1962 1961 1965 1966
lga

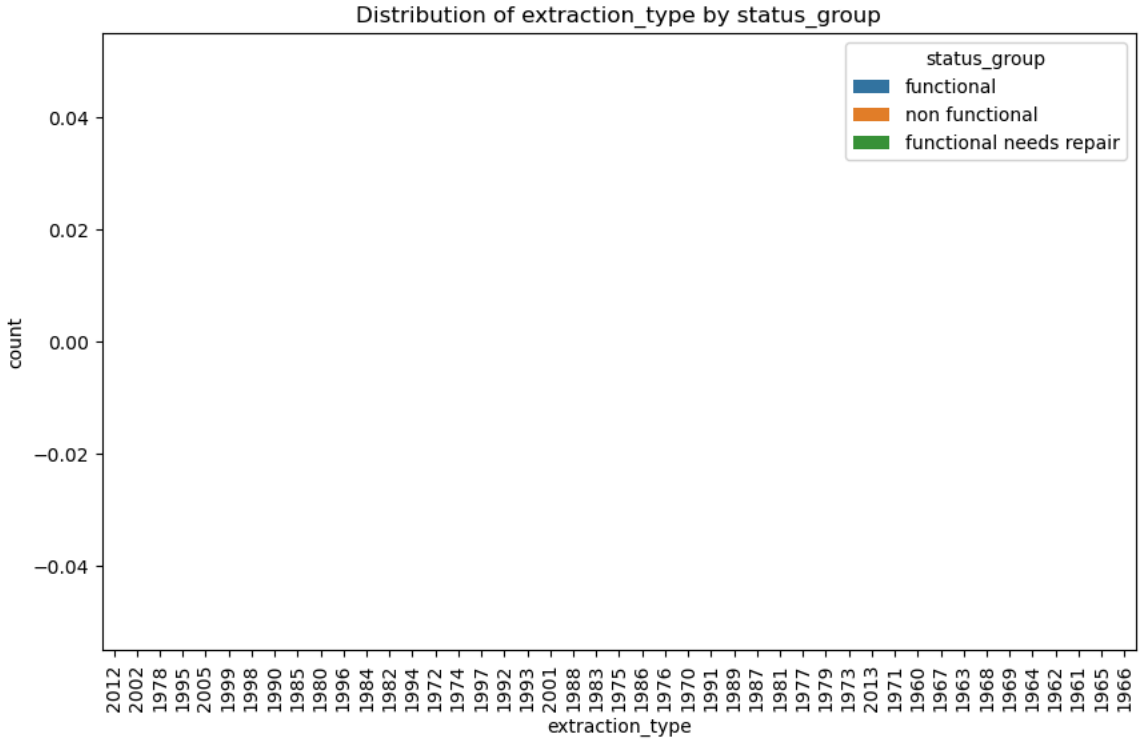

Distribution of ward by status_group
status_group
functional
non functional
functional needs repair
count
0.04
0.02
0.00
−0.02
−0.04
2012 2002 1978 1995 2005 1999 1998 1990 1985 1980 1996 1984 1982 1994 1972 1974 1997 1992 1993 2001 1988 1983 1975 1986 1976 1970 1991 1989 1987 1981 1977 1979 1973 2013 1971 1960 1967 1963 1968 1969 1964 1962 1961 1965 1966
ward

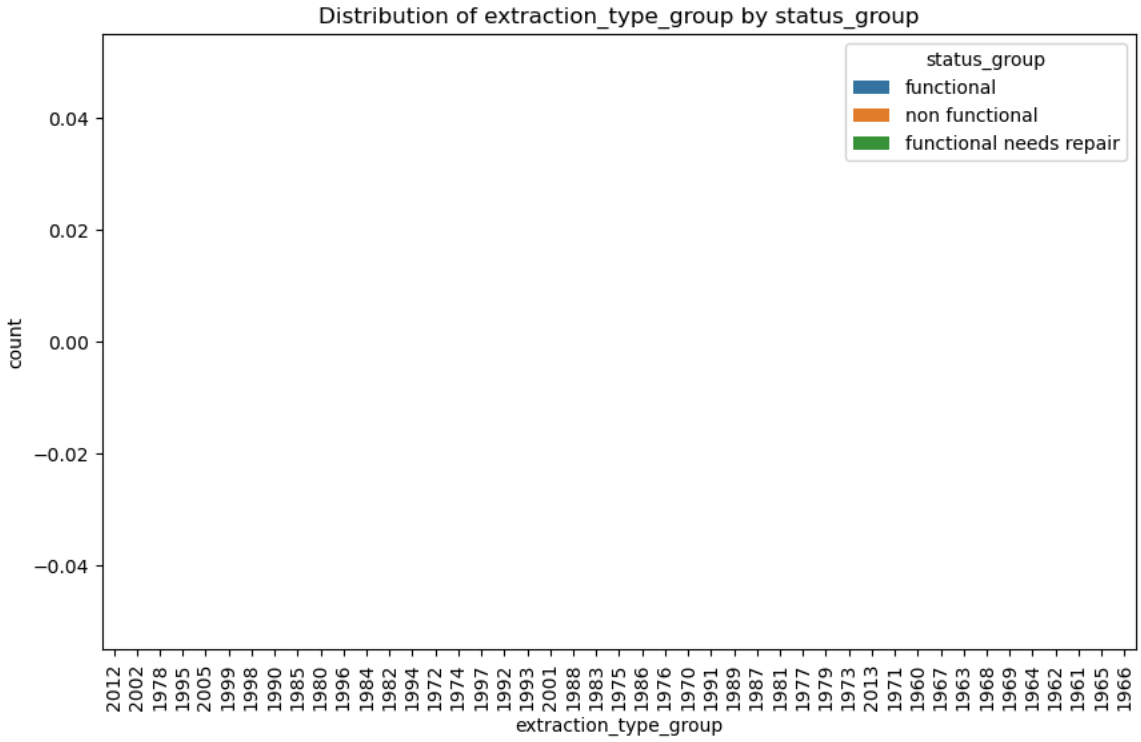

Distribution of public_meeting by status_group
status_group
functional
functional needs repair
non functional
count
0.04
0.02
0.00
−0.02
−0.04
2012 2002 1978 1995 2005 1999 1998 1990 1985 1980 1996 1984 1982 1994 1972 1974 1997 1992 1993 2001 1988 1983 1975 1986 1976 1970 1991 1989 1987 1981 1977 1979 1973 2013 1971 1960 1967 1963 1968 1969 1964 1962 1961 1965 1966
public_meeting

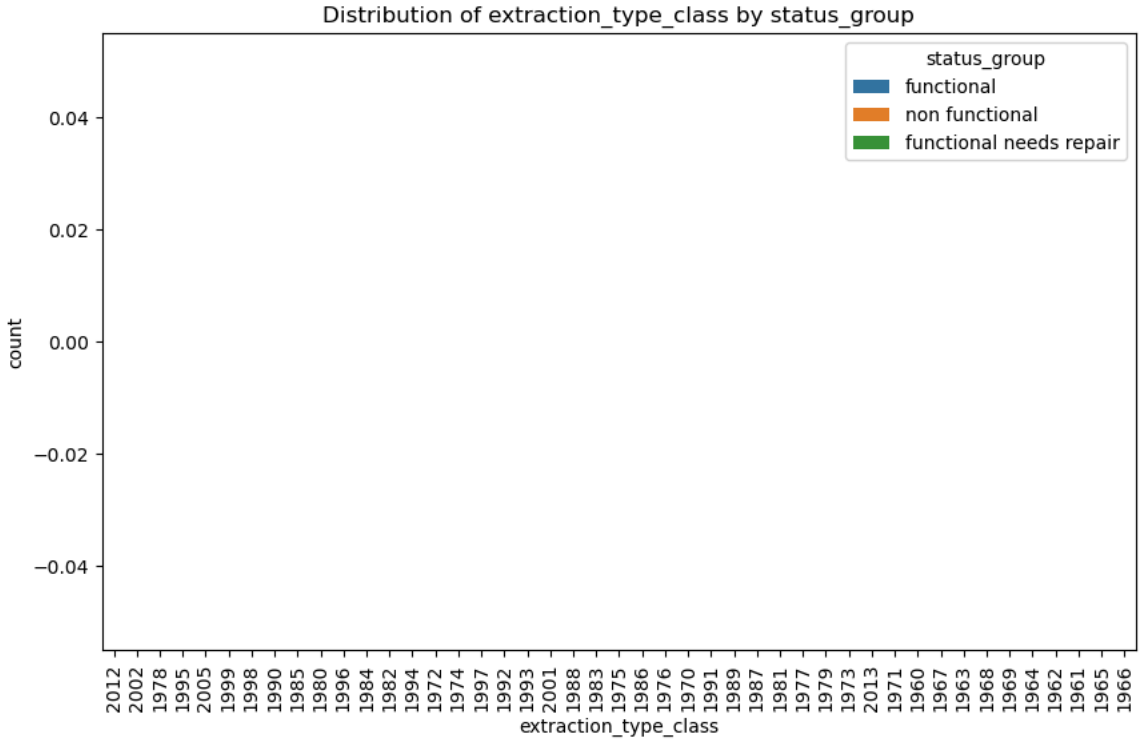

Distribution of recorded_by by status_group
status_group
functional
non functional
functional needs repair
count
0.04
0.02
0.00
−0.02
−0.04
2012 2002 1978 1995 2005 1999 1998 1990 1985 1980 1996 1984 1982 1994 1972 1974 1997 1992 1993 2001 1988 1983 1975 1986 1976 1970 1991 1989 1987 1981 1977 1979 1973 2013 1971 1960 1967 1963 1968 1969 1964 1962 1961 1965 1966
recorded_by

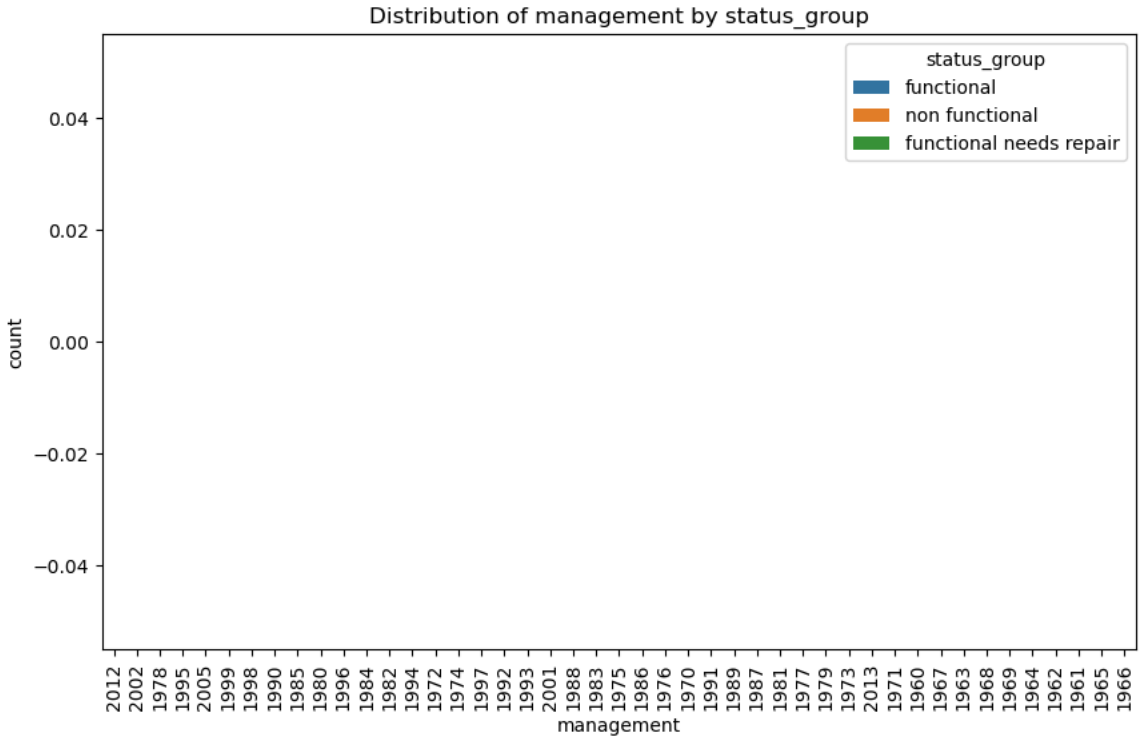

Distribution of scheme_management by status_group
status_group
functional
non functional
functional needs repair
count
0.04
0.02
0.00
−0.02
−0.04
2012 2002 1978 1995 2005 1999 1998 1990 1985 1980 1996 1984 1982 1994 1972 1974 1997 1992 1993 2001 1988 1983 1975 1986 1976 1970 1991 1989 1987 1981 1977 1979 1973 2013 1971 1960 1967 1963 1968 1969 1964 1962 1961 1965 1966
scheme_management

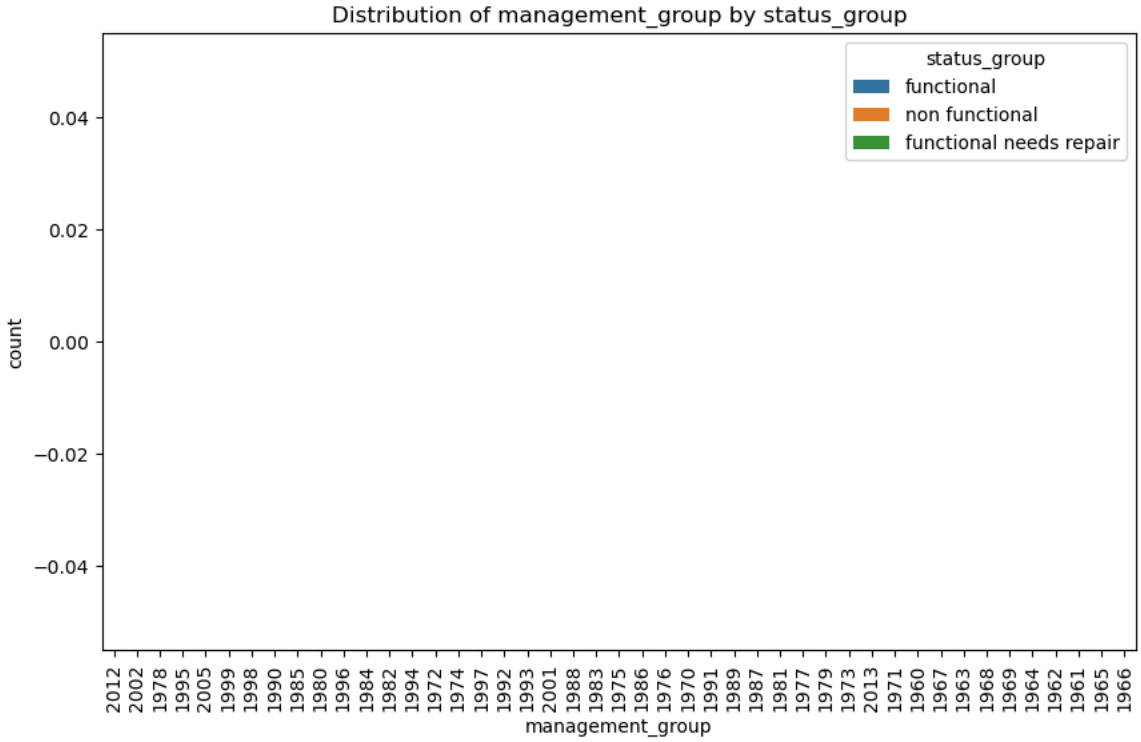

Distribution of permit by status_group
status_group
functional
non functional
functional needs repair
count
0.04
0.02
0.00
−0.02
−0.04
2012 2002 1978 1995 2005 1999 1998 1990 1985 1980 1996 1984 1982 1994 1972 1974 1997 1992 1993 2001 1988 1983 1975 1986 1976 1970 1991 1989 1987 1981 1977 1979 1973 2013 1971 1960 1967 1963 1968 1969 1964 1962 1961 1965 1966
permit

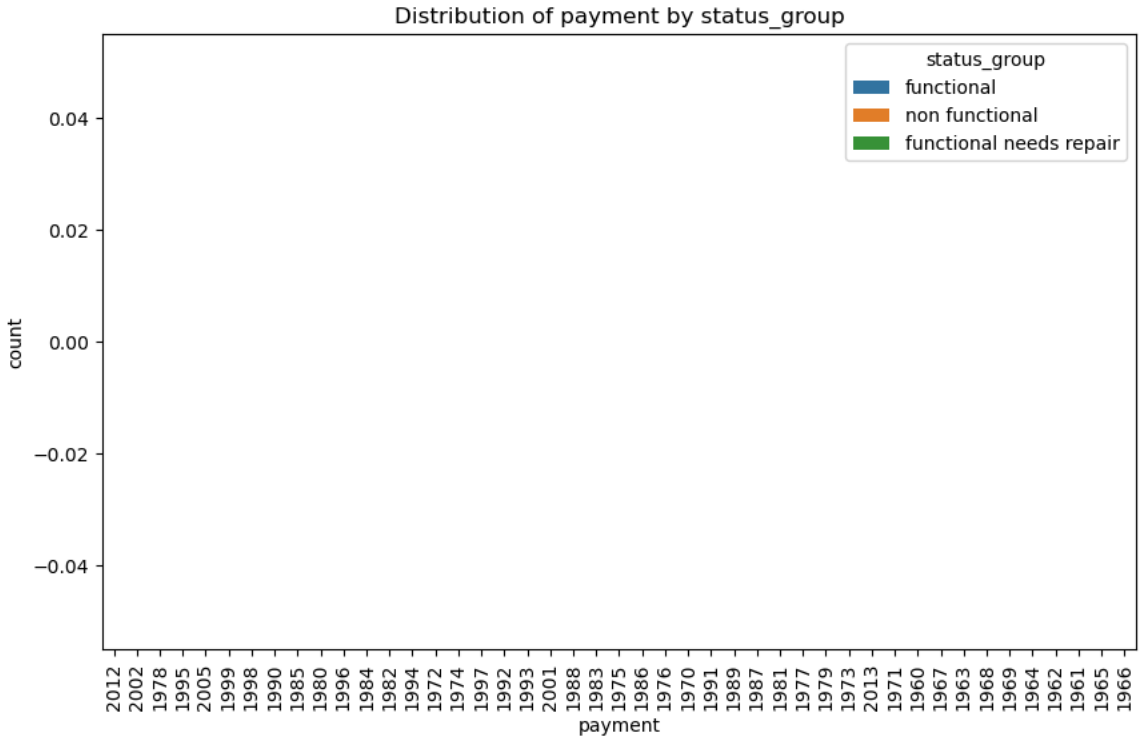

Distribution of extraction_type by status_group
status_group
functional
non functional
functional needs repair
count
0.04
0.02
0.00
−0.02
−0.04
2012 2002 1978 1995 2005 1999 1998 1990 1985 1980 1996 1984 1982 1994 1972 1974 1997 1992 1993 2001 1988 1983 1975 1986 1976 1970 1991 1989 1987 1981 1977 1979 1973 2013 1971 1960 1967 1963 1968 1969 1964 1962 1961 1965 1966
extraction_type

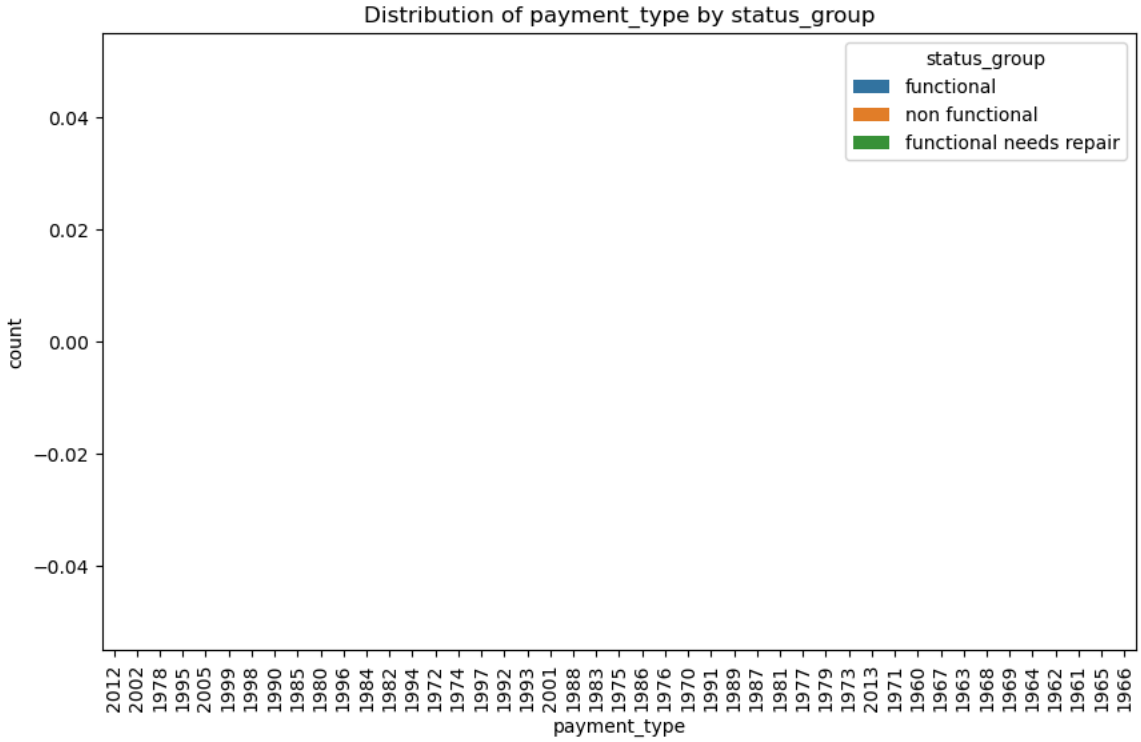

Distribution of extraction_type_group by status_group
status_group
functional
non functional
functional needs repair
count
0.04
0.02
0.00
−0.02
−0.04
2012 2002 1978 1995 2005 1999 1998 1990 1985 1980 1996 1984 1982 1994 1972 1974 1997 1992 1993 2001 1988 1983 1975 1986 1976 1970 1991 1989 1987 1981 1977 1979 1973 2013 1971 1960 1967 1963 1968 1969 1964 1962 1961 1965 1966
extraction_type_group

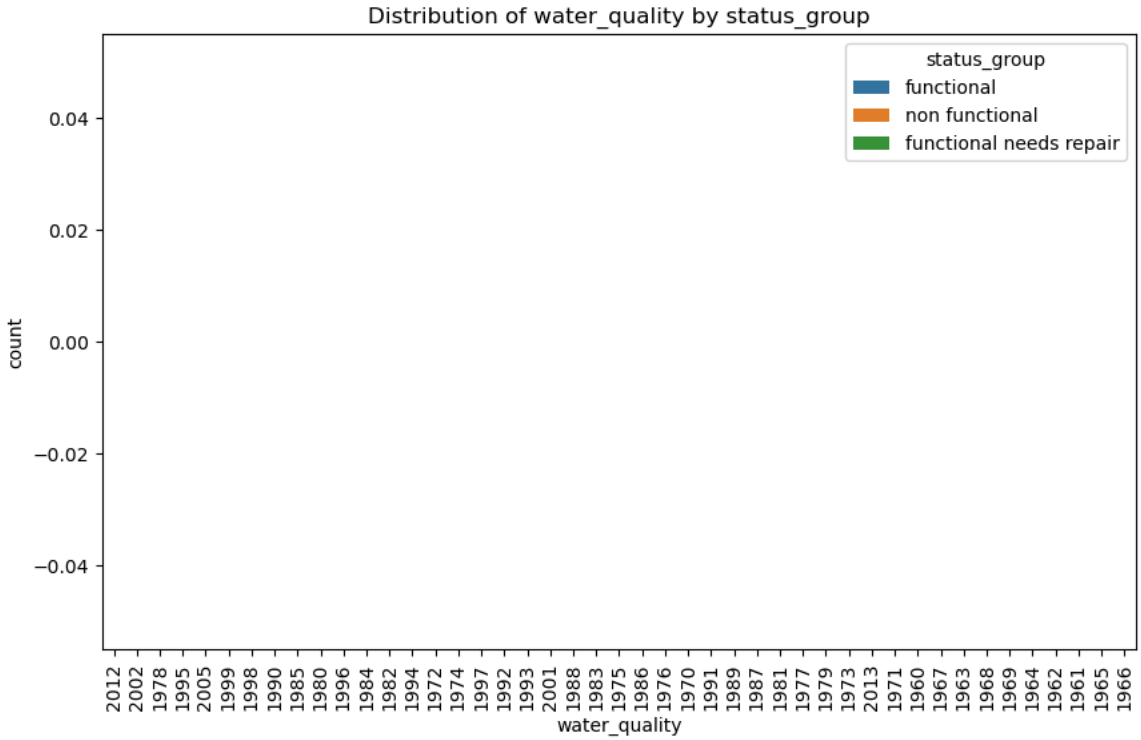

Distribution of extraction_type_class by status_group
status_group
functional
non functional
functional needs repair
count
0.04
0.02
0.00
−0.02
−0.04
2012 2002 1978 1995 2005 1999 1998 1990 1985 1980 1996 1984 1982 1994 1972 1974 1997 1992 1993 2001 1988 1983 1975 1986 1976 1970 1991 1989 1987 1981 1977 1979 1973 2013 1971 1960 1967 1963 1968 1969 1964 1962 1961 1965 1966
extraction_type_class

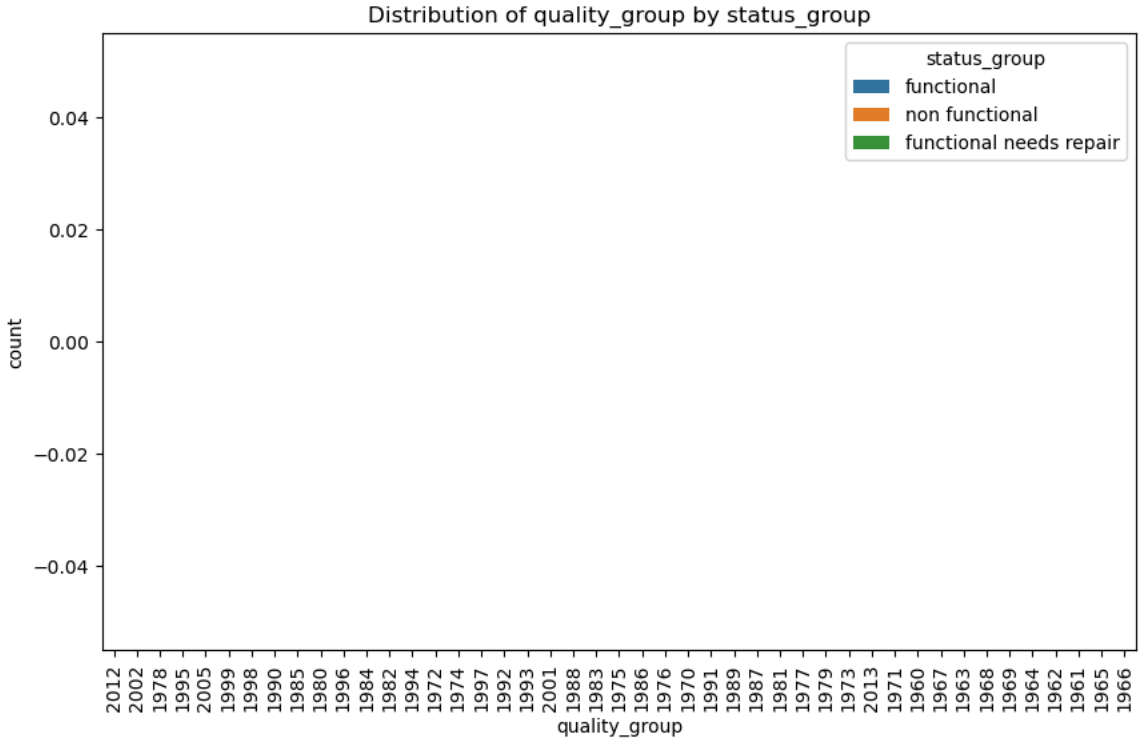

Distribution of management by status_group
status_group
functional
non functional
functional needs repair
count
0.04
0.02
0.00
−0.02
−0.04
2012 2002 1978 1995 2005 1999 1998 1990 1985 1980 1996 1984 1982 1994 1972 1974 1997 1992 1993 2001 1988 1983 1975 1986 1976 1970 1991 1989 1987 1981 1977 1979 1973 2013 1971 1960 1967 1963 1968 1969 1964 1962 1961 1965 1966
management

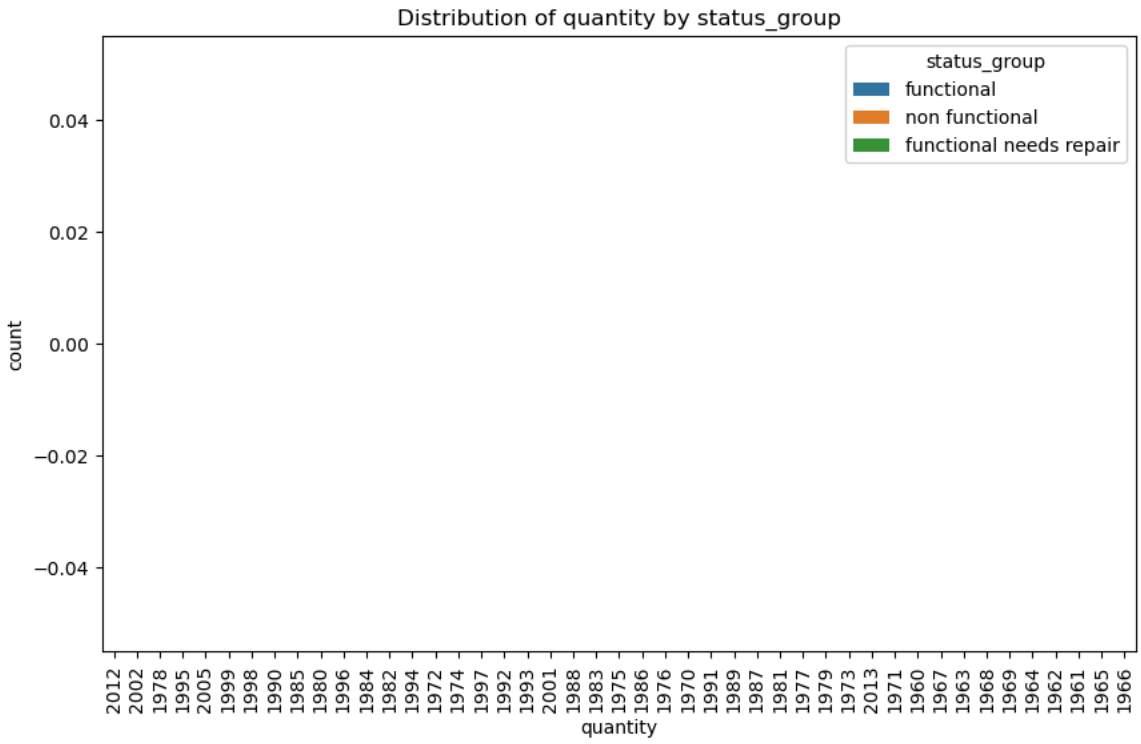

Distribution of management_group by status_group
status_group
functional
non functional
functional needs repair
count
0.04
0.02
0.00
−0.02
−0.04
2012 2002 1978 1995 2005 1999 1998 1990 1985 1980 1996 1984 1982 1994 1972 1974 1997 1992 1993 2001 1988 1983 1975 1986 1976 1970 1991 1989 1987 1981 1977 1979 1973 2013 1971 1960 1967 1963 1968 1969 1964 1962 1961 1965 1966
management_group

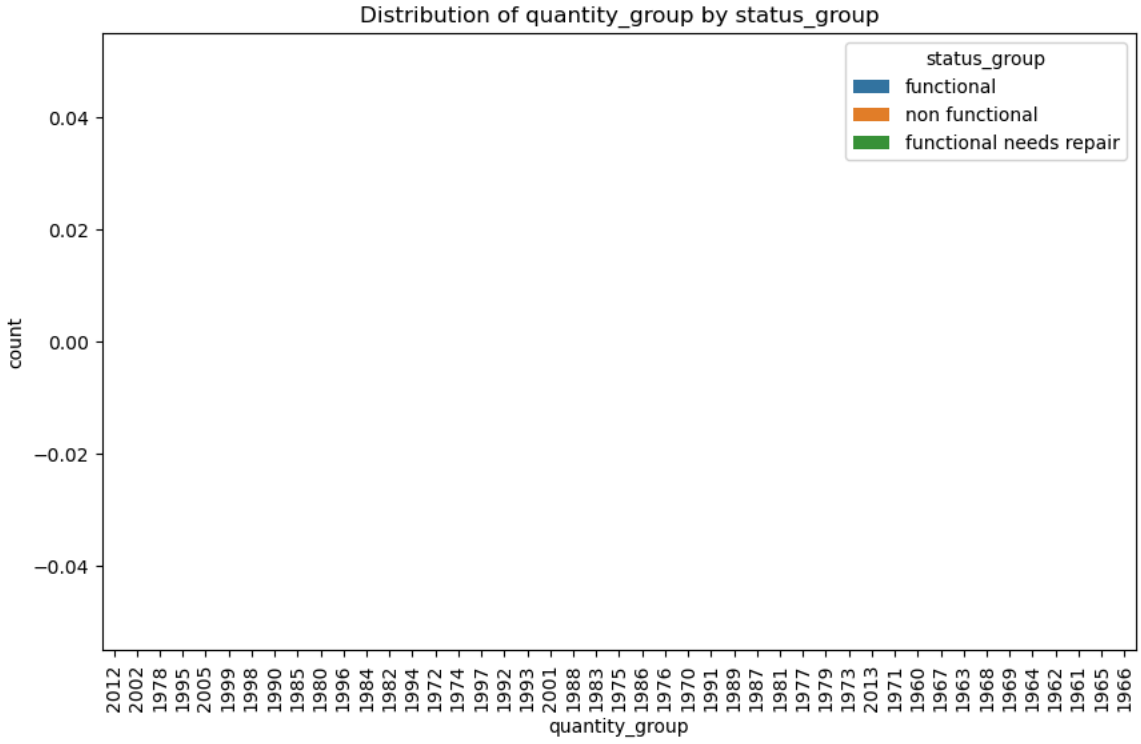

Distribution of payment by status_group
status_group
functional
non functional
functional needs repair
count
0.04
0.02
0.00
−0.02
−0.04
2012 2002 1978 1995 2005 1999 1998 1990 1985 1980 1996 1984 1982 1994 1972 1974 1997 1992 1993 2001 1988 1983 1975 1986 1976 1970 1991 1989 1987 1981 1977 1979 1973 2013 1971 1960 1967 1963 1968 1969 1964 1962 1961 1965 1966
payment

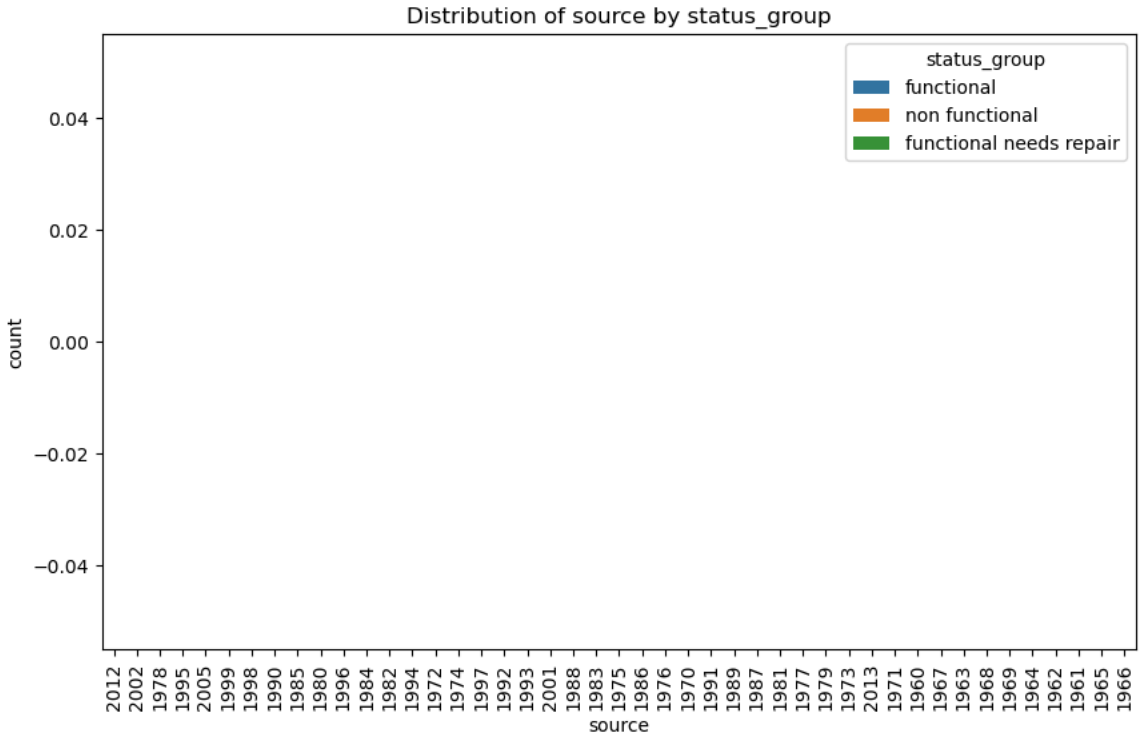

Distribution of payment_type by status_group
status_group
functional
non functional
functional needs repair
count
0.04
0.02
0.00
−0.02
−0.04
2012 2002 1978 1995 2005 1999 1998 1990 1985 1980 1996 1984 1982 1994 1972 1974 1997 1992 1993 2001 1988 1983 1975 1986 1976 1970 1991 1989 1987 1981 1977 1979 1973 2013 1971 1960 1967 1963 1968 1969 1964 1962 1961 1965 1966
payment_type

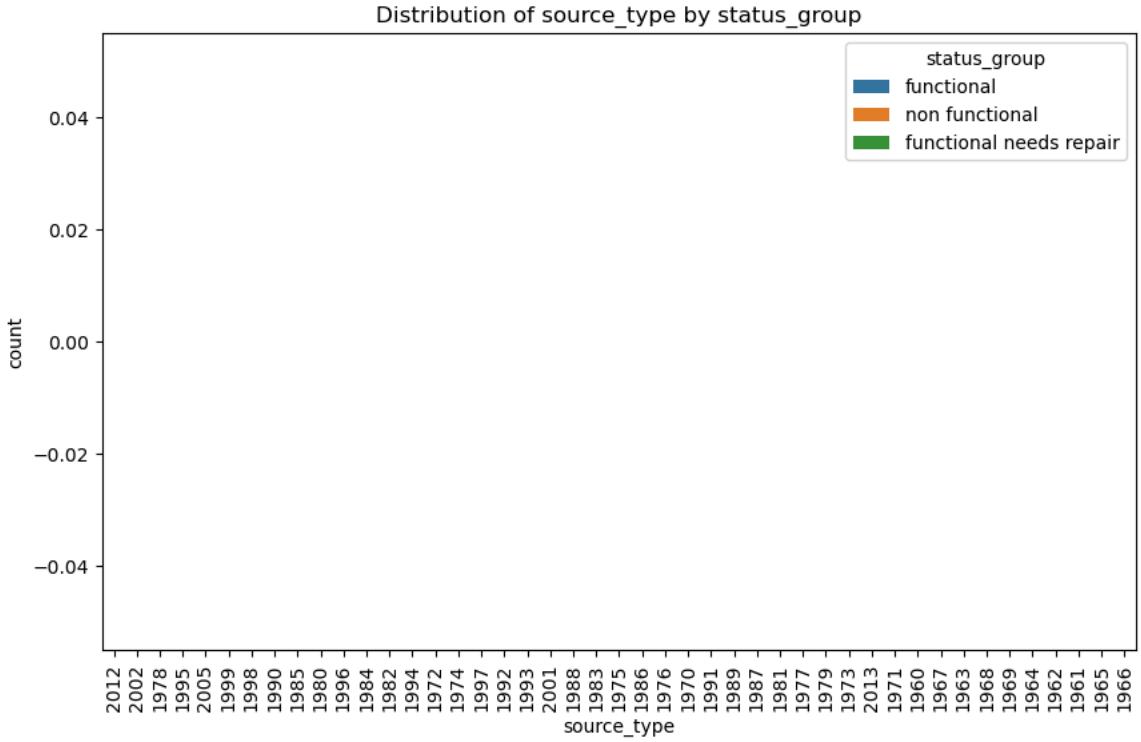

Distribution of water_quality by status_group
status_group
functional
non functional
functional needs repair
count
0.04
0.02
0.00
−0.02
−0.04
2012 2002 1978 1995 2005 1999 1998 1990 1985 1980 1996 1984 1982 1994 1972 1974 1997 1992 1993 2001 1988 1983 1975 1986 1976 1970 1991 1989 1987 1981 1977 1979 1973 2013 1971 1960 1967 1963 1968 1969 1964 1962 1961 1965 1966
water_quality


Distribution of quality_group by status_group
status_group
functional
non functional
functional needs repair
count
0.04
0.02
0.00
−0.02
−0.04
2012 2002 1978 1995 2005 1999 1998 1990 1985 1980 1996 1984 1982 1994 1972 1974 1997 1992 1993 2001 1988 1983 1975 1986 1976 1970 1991 1989 1987 1981 1977 1979 1973 2013 1971 1960 1967 1963 1968 1969 1964 1962 1961 1965 1966
quality_group

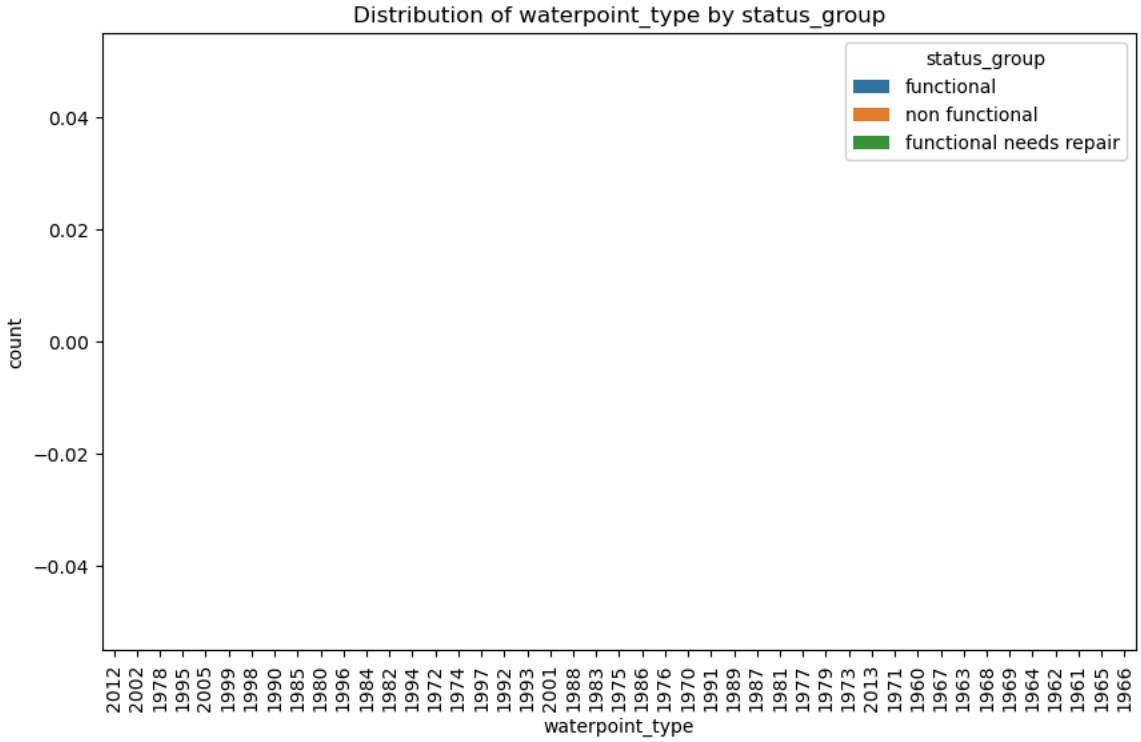

Distribution of quantity by status_group
status_group
functional
non functional
functional needs repair
count
0.04
0.02
0.00
−0.02
−0.04
2012 2002 1978 1995 2005 1999 1998 1990 1985 1980 1996 1984 1982 1994 1972 1974 1997 1992 1993 2001 1988 1983 1975 1986 1976 1970 1991 1989 1987 1981 1977 1979 1973 2013 1971 1960 1967 1963 1968 1969 1964 1962 1961 1965 1966
quantity

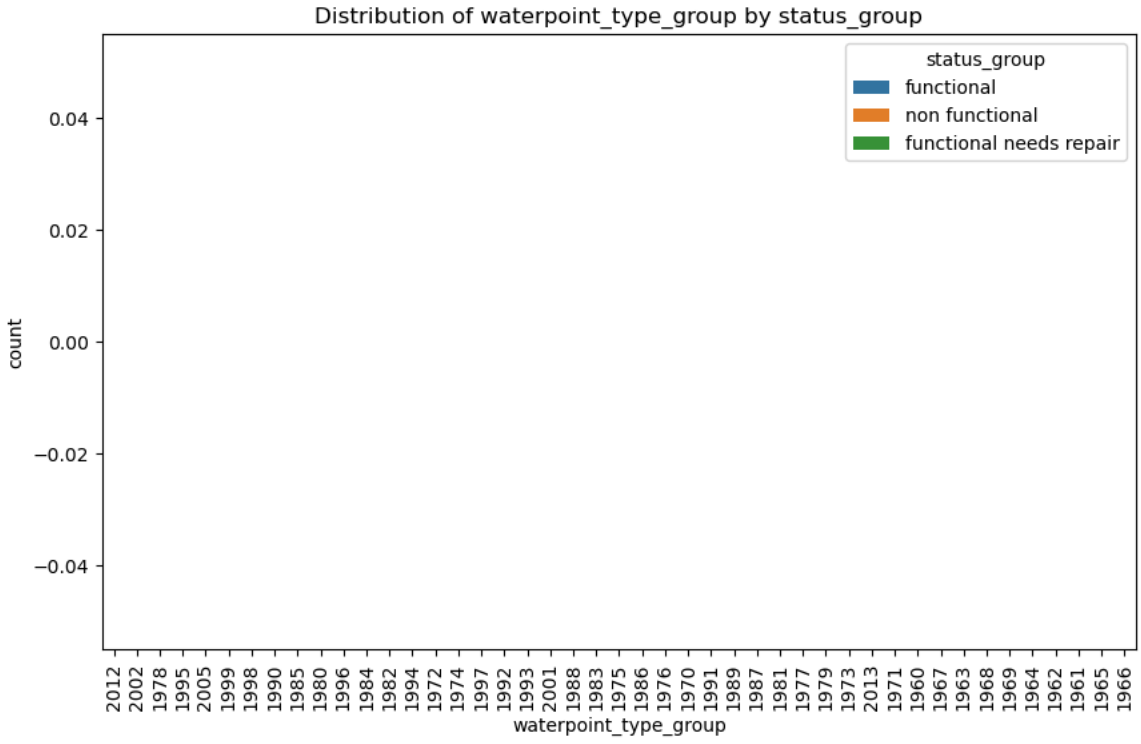

Distribution of quantity_group by status_group
status_group
functional
non functional
functional needs repair
count
0.04
0.02
0.00
−0.02
−0.04
2012 2002 1978 1995 2005 1999 1998 1990 1985 1980 1996 1984 1982 1994 1972 1974 1997 1992 1993 2001 1988 1983 1975 1986 1976 1970 1991 1989 1987 1981 1977 1979 1973 2013 1971 1960 1967 1963 1968 1969 1964 1962 1961 1965 1966
quantity_group

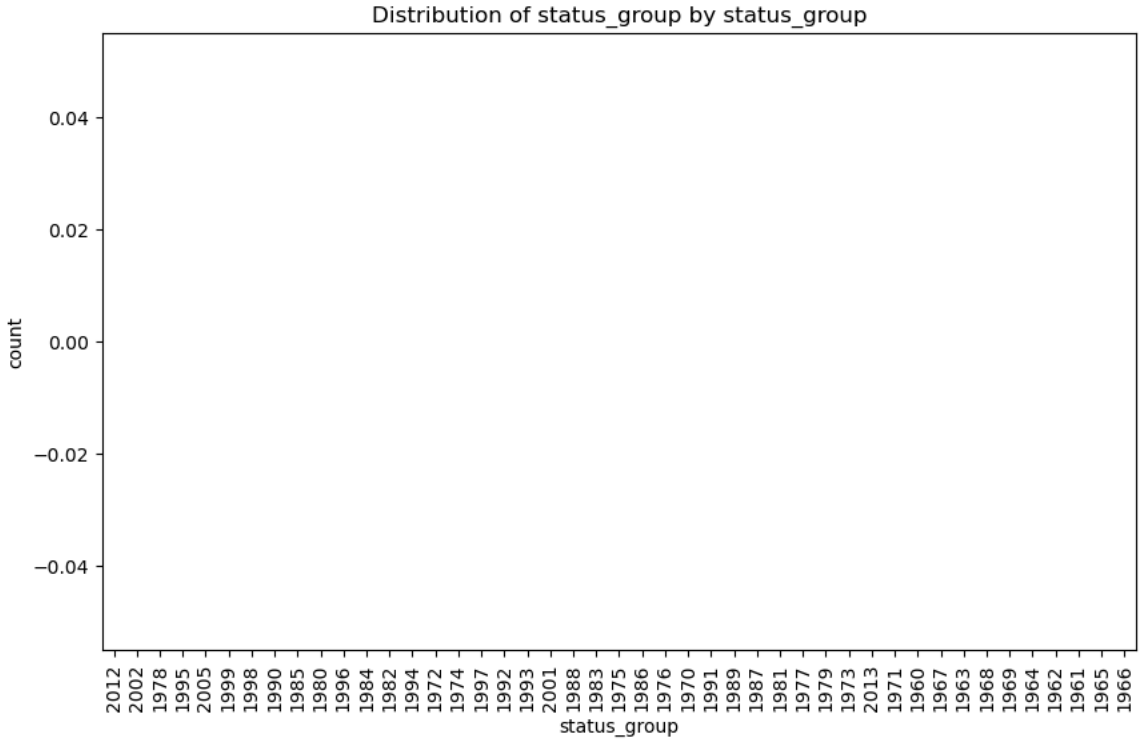

Distribution of source by status_group
status_group
functional
non functional
functional needs repair
count
0.04
0.02
0.00
−0.02
−0.04
2012 2002 1978 1995 2005 1999 1998 1990 1985 1980 1996 1984 1982 1994 1972 1974 1997 1992 1993 2001 1988 1983 1975 1986 1976 1970 1991 1989 1987 1981 1977 1979 1973 2013 1971 1960 1967 1963 1968 1969 1964 1962 1961 1965 1966
source


Distribution of source_type by status_group
status_group
functional
non functional
functional needs repair
count
0.04
0.02
0.00
−0.02
−0.04
2012 2002 1978 1995 2005 1999 1998 1990 1985 1980 1996 1984 1982 1994 1972 1974 1997 1992 1993 2001 1988 1983 1975 1986 1976 1970 1991 1989 1987 1981 1977 1979 1973 2013 1971 1960 1967 1963 1968 1969 1964 1962 1961 1965 1966
source_type


Distribution of source_class by status_group
status_group
functional
non functional
functional needs repair
count
0.04
0.02
0.00
−0.02
−0.04
2012 2002 1978 1995 2005 1999 1998 1990 1985 1980 1996 1984 1982 1994 1972 1974 1997 1992 1993 2001 1988 1983 1975 1986 1976 1970 1991 1989 1987 1981 1977 1979 1973 2013 1971 1960 1967 1963 1968 1969 1964 1962 1961 1965 1966
source_class


Distribution of waterpoint_type by status_group
status_group
functional
non functional
functional needs repair
count
0.04
0.02
0.00
−0.02
−0.04
2012 2002 1978 1995 2005 1999 1998 1990 1985 1980 1996 1984 1982 1994 1972 1974 1997 1992 1993 2001 1988 1983 1975 1986 1976 1970 1991 1989 1987 1981 1977 1979 1973 2013 1971 1960 1967 1963 1968 1969 1964 1962 1961 1965 1966
waterpoint_type


Distribution of waterpoint_type_group by status_group
status_group
functional
non functional
functional needs repair
count
0.04
0.02
0.00
−0.02
−0.04
2012 2002 1978 1995 2005 1999 1998 1990 1985 1980 1996 1984 1982 1994 1972 1974 1997 1992 1993 2001 1988 1983 1975 1986 1976 1970 1991 1989 1987 1981 1977 1979 1973 2013 1971 1960 1967 1963 1968 1969 1964 1962 1961 1965 1966
waterpoint_type_group

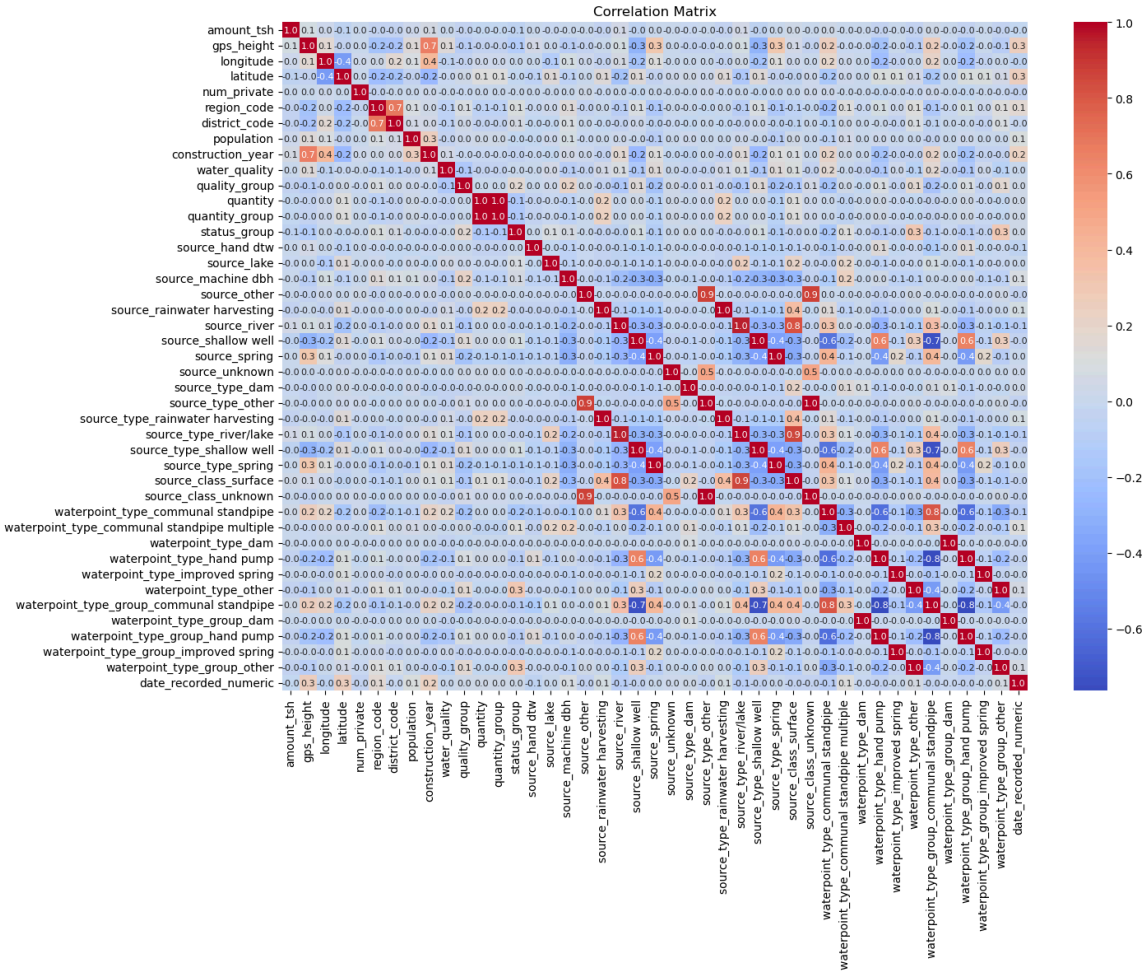

Distribution of status_group by status_group
count
0.04
0.02
0.00
−0.02
−0.04
2012 2002 1978 1995 2005 1999 1998 1990 1985 1980 1996 1984 1982 1994 1972 1974 1997 1992 1993 2001 1988 1983 1975 1986 1976 1970 1991 1989 1987 1981 1977 1979 1973 2013 1971 1960 1967 1963 1968 1969 1964 1962 1961 1965 1966
status_group

In [20]:
```
from sklearn.preprocessing import LabelEncoder

# Label encode ordinal columns
label_encode_cols = ['quality_group', 'quantity_group', 'status_group', 'wa
ter_quality', 'quantity']
le_dict = {}
for col in label_encode_cols:
    le = LabelEncoder()
    imputed_data[col] = le.fit_transform(imputed_data[col].astype(str))
    le_dict[col] = le

# One-hot encode nominal columns using pd.get_dummies
onehot_encode_cols = ['source', 'source_type', 'source_class', 'waterpoint_
type', 'waterpoint_type_group']
t_data = pd.get_dummies(imputed_data, columns=onehot_encode_cols, drop_firs
t=True,  dtype=int)
print(t_data.head())
```

```
      id  amount_tsh date_recorded        funder  gps_height     installer
\
0  69572      6000.0    14/03/2011         Roman        1390         Roman
1   8776         0.0    06/03/2013       Grumeti        1399       GRUMETI
2  34310        25.0    25/02/2013  Lottery Club         686  World vision
3  67743         0.0    28/01/2013        Unicef         263        UNICEF
4  19728         0.0    13/07/2011   Action In A           0       Artisan

   longitude   latitude              wpt_name  num_private  ...  \
0  34.938093  -9.856322                  none            0  ...
1  34.698766  -2.147466              Zahanati            0  ...
2  37.460664  -3.821329           Kwa Mahundi            0  ...
3  38.486161 -11.155298  Zahanati Ya Nanyumbu            0  ...
4  31.130847  -1.825359               Shuleni            0  ...

  waterpoint_type_communal standpipe multiple waterpoint_type_dam  \
0                                          0                   0
1                                          0                   0
2                                          1                   0
3                                          1                   0
4                                          0                   0

  waterpoint_type_hand pump  waterpoint_type_improved spring  \
0                         0                                0
1                         0                                0
2                         0                                0
3                         0                                0
4                         0                                0

   waterpoint_type_other waterpoint_type_group_communal standpipe  \
0                      0                                        1
1                      0                                        1
2                      0                                        1
3                      0                                        1
4                      0                                        1

  waterpoint_type_group_dam  waterpoint_type_group_hand pump  \
0                         0                                0
1                         0                                0
2                         0                                0
3                         0                                0
4                         0                                0

  waterpoint_type_group_improved spring waterpoint_type_group_other
0                                     0                           0
1                                     0                           0
2                                     0                           0
3                                     0                           0
4                                     0                           0

[5 rows x 63 columns]
```

Checking for the relationship between the features for feature selection

In [21]:
```
t_data_no_id = t_data.drop('id', axis=1)

t_data_no_id['date_recorded'] = pd.to_datetime(t_data_no_id['date_recorde
d'], format='%d/%m/%Y')
# Converting datetime to a numeric timestamp (in seconds)
t_data_no_id['date_recorded_numeric'] = t_data_no_id['date_recorded'].astyp
e('int64') // 10**9

# Now compute the correlation matrix using the new numeric date column
numeric_data = t_data_no_id.select_dtypes(include=['number'])
corr = numeric_data.corr()
plt.figure(figsize=(15, 12))
sns.heatmap(corr, annot=True, fmt=".1f", cmap='coolwarm',annot_kws={"size":
8})
plt.title("Correlation Matrix")
plt.tight_layout()
plt.show()
```



There seems to be very little colinearity between the features and our target variable, meaning most linear regression models might not suffice.

Feature Selection

To avoid multicollinearity, we shall drop one of the columns in quantity/quantity_group, district code and region code

In [25]:
```
t_data_no_id.describe
```

Out[25]:
```
<bound method NDFrame.describe of        amount_tsh date_recorded
funder  gps_height  \
0          6000.0    2011-03-14                  Roman        1390
1             0.0    2013-03-06                Grumeti        1399
2            25.0    2013-02-25            Lottery Club         686
3             0.0    2013-01-28                 Unicef         263
4             0.0    2011-07-13             Action In A           0
...           ...           ...                     ...         ...
59395        10.0    2013-05-03         Germany Republi        1210
59396      4700.0    2011-05-07             Cefa-njombe        1212
59397         0.0    2011-04-11  Government Of Tanzania           0
59398         0.0    2011-03-08                  Malec           0
59399         0.0    2011-03-23              World Bank         191

          installer  longitude   latitude              wpt_name  num_priva
te  \
0             Roman  34.938093  -9.856322                  none
0
1           GRUMETI  34.698766  -2.147466              Zahanati
0
2      World vision  37.460664  -3.821329           Kwa Mahundi
0
3            UNICEF  38.486161 -11.155298  Zahanati Ya Nanyumbu
0
4           Artisan  31.130847  -1.825359               Shuleni
0
...             ...        ...        ...                   ...
...
59395           CES  37.169807  -3.253847   Area Three Namba 27
0
59396          Cefa  35.249991  -9.070629     Kwa Yahona Kuvala
0
59397           DWE  34.017087  -8.750434               Mashine
0
59398          Musa  35.861315  -6.378573                Mshoro
0
59399         World  38.104048  -6.747464       Kwa Mzee Lugawa
0

                        basin  ... waterpoint_type_dam  \
0                  Lake Nyasa  ...                   0
1                Lake Victoria  ...                   0
2                      Pangani  ...                   0
3      Ruvuma / Southern Coast  ...                   0
4                Lake Victoria  ...                   0
...                        ...  ...                 ...
59395                  Pangani  ...                   0
59396                   Rufiji  ...                   0
59397                   Rufiji  ...                   0
59398                   Rufiji  ...                   0
59399              Wami / Ruvu  ...                   0

      waterpoint_type_hand pump  waterpoint_type_improved spring  \
0                             0                                 0
1                             0                                 0
2                             0                                 0
3                             0                                 0
4                             0                                 0
...                         ...                               ...
59395                         0                                 0
59396                         0                                 0
```

```
59397                         1                                0
59398                         1                                0
59399                         1                                0

       waterpoint_type_other waterpoint_type_group_communal standpipe  \
0                          0                                        1
1                          0                                        1
2                          0                                        1
3                          0                                        1
4                          0                                        1
...                      ...                                      ...
59395                      0                                        1
59396                      0                                        1
59397                      0                                        0
59398                      0                                        0
59399                      0                                        0

      waterpoint_type_group_dam  waterpoint_type_group_hand pump  \
0                             0                                 0
1                             0                                 0
2                             0                                 0
3                             0                                 0
4                             0                                 0
...                         ...                               ...
59395                         0                                 0
59396                         0                                 0
59397                         0                                 1
59398                         0                                 1
59399                         0                                 1

      waterpoint_type_group_improved spring waterpoint_type_group_other  \
0                                         0                           0
1                                         0                           0
2                                         0                           0
3                                         0                           0
4                                         0                           0
...                                     ...                         ...
59395                                     0                           0
59396                                     0                           0
59397                                     0                           0
59398                                     0                           0
59399                                     0                           0

      date_recorded_numeric
0                1300060800
1                1362528000
2                1361750400
3                1359331200
4                1310515200
...                     ...
59395            1367539200
59396            1304726400
59397            1302480000
59398            1299542400
59399            1300838400

[59400 rows x 63 columns]>
```

Building the machine learning model

In [48]:
```
from sklearn.model_selection import train_test_split
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import confusion_matrix, classification_report
numeric_df = t_data_no_id.select_dtypes(include=[np.number])
X = numeric_df.drop(columns=['status_group'], errors='ignore')
y = t_data_no_id['status_group']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, r
andom_state=42)
pipeline = Pipeline([
    ('scaler', StandardScaler()),
    ('classifier', RandomForestClassifier(n_estimators=100, random_state=4
2))
])

pipeline.fit(X_train, y_train)
```

Out[48]:

```
▸ Pipeline i ? (https://scikit-learn.org/1.4/modules/generated/sklearn.pipeline.Pipel
    ▸ StandardScaler ? (https://scikit-learn.org/1.4/modules/generated/sklearn.preprocessing.Standar
    ▸ RandomForestClassifier ? (https://scikit-learn.org/1.4/modules/generated/sklearn.ensemble.RandomF
```

In [50]:
```
test_values.describe
```

Out[50]:
```
<bound method NDFrame.describe of           id  amount_tsh date_recorded
funder  gps_height  \
0      50785         0.0    04/02/2013                    Dmdd        1996
1      51630         0.0    04/02/2013  Government Of Tanzania        1569
2      17168         0.0    01/02/2013                     NaN        1567
3      45559         0.0    22/01/2013              Finn Water         267
4      49871       500.0    27/03/2013                  Bruder        1260
...      ...         ...           ...                     ...         ...
14845  39307         0.0    24/02/2011                  Danida          34
14846  18990      1000.0    21/03/2011                    Hiap           0
14847  28749         0.0    04/03/2013                     NaN        1476
14848  33492         0.0    18/02/2013                 Germany         998
14849  68707         0.0    13/02/2013  Government Of Tanzania         481

        installer  longitude   latitude                wpt_name  num_priv
ate  \
0            DMDD  35.290799  -4.059696  Dinamu Secondary School
0
1             DWE  36.656709  -3.309214                  Kimnyak
0
2             NaN  34.767863  -5.004344           Puma Secondary
0
3      FINN WATER  38.058046  -9.418672           Kwa Mzee Pange
0
4          BRUDER  35.006123 -10.950412          Kwa Mzee Turuka
0
...           ...        ...        ...                      ...
...
14845          Da  38.852669  -6.582841                Kwambwezi
0
14846        HIAP  37.451633  -5.350428         Bonde La Mkondoa
0
14847         NaN  34.739804  -4.585587                  Bwawani
0
14848         DWE  35.432732 -10.584159                 Kwa John
0
14849  Government  34.765054 -11.226012         Kwa Mzee Chagala
0

       ... payment_type water_quality quality_group      quantity  \
0      ...    never pay          soft          good      seasonal
1      ...    never pay          soft          good  insufficient
2      ...    never pay          soft          good  insufficient
3      ...      unknown          soft          good           dry
4      ...      monthly          soft          good        enough
...    ...          ...           ...           ...           ...
14845  ...    never pay          soft          good        enough
14846  ...     annually         salty         salty  insufficient
14847  ...    never pay          soft          good  insufficient
14848  ...    never pay          soft          good  insufficient
14849  ...    never pay          soft          good           dry

       quantity_group                source           source_type  \
0            seasonal  rainwater harvesting  rainwater harvesting
1        insufficient                spring                spring
2        insufficient  rainwater harvesting  rainwater harvesting
3                 dry          shallow well          shallow well
4              enough                spring                spring
...               ...                   ...                   ...
14845          enough                 river            river/lake
14846    insufficient          shallow well          shallow well
```

```
14847    insufficient                   dam                   dam
14848    insufficient                 river            river/lake
14849             dry                spring                spring

       source_class     waterpoint_type waterpoint_type_group
0           surface               other                 other
1       groundwater  communal standpipe    communal standpipe
2           surface               other                 other
3       groundwater               other                 other
4       groundwater  communal standpipe    communal standpipe
...             ...                 ...                   ...
14845       surface  communal standpipe    communal standpipe
14846   groundwater           hand pump             hand pump
14847       surface  communal standpipe    communal standpipe
14848       surface  communal standpipe    communal standpipe
14849   groundwater  communal standpipe    communal standpipe

[14850 rows x 40 columns]>
```

In [52]:
```
test_data = test_values.copy()
cat_imputer = SimpleImputer(strategy='most_frequent')
test_data[['permit', 'scheme_management', 'public_meeting',
              'subvillage', 'funder', 'installer', 'wpt_name']] = \
    cat_imputer.fit_transform(test_data[['permit', 'scheme_management', 'pu
blic_meeting',
                                                 'subvillage', 'funder', 'insta
ller', 'wpt_name']])
print(test_data.isna().sum())
print(test_data.head())
```

```
id                         0
amount_tsh                 0
date_recorded              0
funder                     0
gps_height                 0
installer                  0
longitude                  0
latitude                   0
wpt_name                   0
num_private                0
basin                      0
subvillage                 0
region                     0
region_code                0
district_code              0
lga                        0
ward                       0
population                 0
public_meeting             0
recorded_by                0
scheme_management          0
scheme_name             7242
permit                     0
construction_year          0
extraction_type            0
extraction_type_group      0
extraction_type_class      0
management                 0
management_group           0
payment                    0
payment_type               0
water_quality              0
quality_group              0
quantity                   0
quantity_group             0
source                     0
source_type                0
source_class               0
waterpoint_type            0
waterpoint_type_group      0
dtype: int64
      id  amount_tsh date_recorded                  funder  gps_height  \
0  50785         0.0    04/02/2013                    Dmdd        1996
1  51630         0.0    04/02/2013  Government Of Tanzania        1569
2  17168         0.0    01/02/2013  Government Of Tanzania        1567
3  45559         0.0    22/01/2013              Finn Water         267
4  49871       500.0    27/03/2013                  Bruder        1260

    installer  longitude   latitude                 wpt_name  num_private
\
0        DMDD  35.290799  -4.059696  Dinamu Secondary School            0
1         DWE  36.656709  -3.309214                  Kimnyak            0
2         DWE  34.767863  -5.004344           Puma Secondary            0
3  FINN WATER  38.058046  -9.418672           Kwa Mzee Pange            0
4      BRUDER  35.006123 -10.950412          Kwa Mzee Turuka            0

   ... payment_type water_quality quality_group      quantity  quantity_gr
oup  \
0  ...    never pay          soft          good      seasonal        seaso
nal
1  ...    never pay          soft          good  insufficient    insuffici
```

```
ent
2  ...    never pay          soft          good  insufficient    insuffici
ent
3  ...      unknown          soft          good           dry
dry
4  ...      monthly          soft          good        enough          eno
ugh

                 source           source_type  source_class  \
0  rainwater harvesting  rainwater harvesting       surface
1                spring                spring   groundwater
2  rainwater harvesting  rainwater harvesting       surface
3          shallow well          shallow well   groundwater
4                spring                spring   groundwater

      waterpoint_type waterpoint_type_group
0               other                 other
1  communal standpipe    communal standpipe
2               other                 other
3               other                 other
4  communal standpipe    communal standpipe

[5 rows x 40 columns]
```

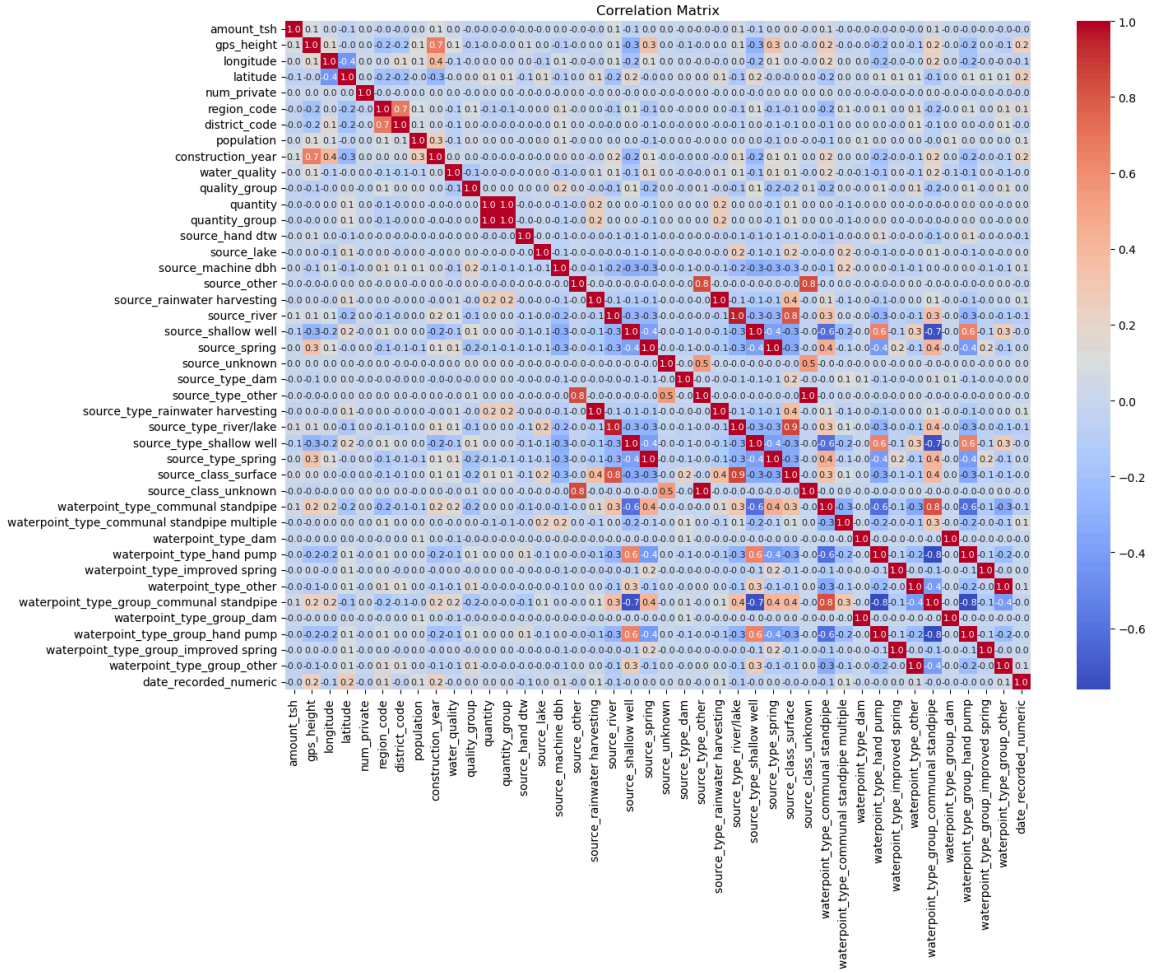
In [54]:
```
# Label encode ordinal columns
label_encode_cols = ['quality_group', 'quantity_group',  'water_quality',
'quantity']
le_dict = {}
for col in label_encode_cols:
    le = LabelEncoder()
    test_data[col] = le.fit_transform(test_data[col].astype(str))
    le_dict[col] = le

# One-hot encode nominal columns using pd.get_dummies
onehot_encode_cols = ['source', 'source_type', 'source_class', 'waterpoint_
type', 'waterpoint_type_group']
act_data = pd.get_dummies(test_data, columns=onehot_encode_cols, drop_first
=True,  dtype=int)
print(act_data.head())
```

```
      id  amount_tsh date_recorded                 funder  gps_height  \
0  50785         0.0    04/02/2013                   Dmdd        1996
1  51630         0.0    04/02/2013  Government Of Tanzania        1569
2  17168         0.0    01/02/2013  Government Of Tanzania        1567
3  45559         0.0    22/01/2013             Finn Water         267
4  49871       500.0    27/03/2013                 Bruder        1260

    installer  longitude   latitude                wpt_name  num_private
\
0        DMDD  35.290799  -4.059696  Dinamu Secondary School            0
1         DWE  36.656709  -3.309214                  Kimnyak            0
2         DWE  34.767863  -5.004344           Puma Secondary            0
3  FINN WATER  38.058046  -9.418672           Kwa Mzee Pange            0
4      BRUDER  35.006123 -10.950412          Kwa Mzee Turuka            0

   ... waterpoint_type_communal standpipe multiple waterpoint_type_dam  \
0  ...                                          0                   0
1  ...                                          0                   0
2  ...                                          0                   0
3  ...                                          0                   0
4  ...                                          0                   0

  waterpoint_type_hand pump  waterpoint_type_improved spring  \
0                         0                                0
1                         0                                0
2                         0                                0
3                         0                                0
4                         0                                0

   waterpoint_type_other waterpoint_type_group_communal standpipe  \
0                      1                                        0
1                      0                                        1
2                      1                                        0
3                      1                                        0
4                      0                                        1

  waterpoint_type_group_dam  waterpoint_type_group_hand pump  \
0                         0                                0
1                         0                                0
2                         0                                0
3                         0                                0
4                         0                                0

  waterpoint_type_group_improved spring waterpoint_type_group_other
0                                     0                           1
1                                     0                           0
2                                     0                           1
3                                     0                           1
4                                     0                           0

[5 rows x 63 columns]
```

In [82]:
```
# since most of the features are categorical in nature we shall encode as f
ollows
# Label encode ordinal columns
label_encode_cols = ['quality_group', 'quantity_group',  'water_quality',
'quantity']
le_dict = {}
for col in label_encode_cols:
    le = LabelEncoder()
    test_data[col] = le.fit_transform(test_data[col].astype(str))
    le_dict[col] = le

# One-hot encode nominal columns using pd.get_dummies
onehot_encode_cols = ['source', 'source_type', 'source_class', 'waterpoint_
type', 'waterpoint_type_group']
act_data = pd.get_dummies(test_data, columns=onehot_encode_cols, drop_first
=True,  dtype=int)
print(act_data.head())
```

```
      id  amount_tsh date_recorded                 funder  gps_height  \
0  50785         0.0    04/02/2013                   Dmdd        1996
1  51630         0.0    04/02/2013  Government Of Tanzania        1569
2  17168         0.0    01/02/2013  Government Of Tanzania        1567
3  45559         0.0    22/01/2013             Finn Water         267
4  49871       500.0    27/03/2013                 Bruder        1260

    installer  longitude   latitude                wpt_name  num_private
\
0        DMDD  35.290799  -4.059696  Dinamu Secondary School            0
1         DWE  36.656709  -3.309214                  Kimnyak            0
2         DWE  34.767863  -5.004344           Puma Secondary            0
3  FINN WATER  38.058046  -9.418672           Kwa Mzee Pange            0
4      BRUDER  35.006123 -10.950412          Kwa Mzee Turuka            0

   ... waterpoint_type_communal standpipe multiple waterpoint_type_dam  \
0  ...                                          0                   0
1  ...                                          0                   0
2  ...                                          0                   0
3  ...                                          0                   0
4  ...                                          0                   0

  waterpoint_type_hand pump  waterpoint_type_improved spring  \
0                         0                                0
1                         0                                0
2                         0                                0
3                         0                                0
4                         0                                0

   waterpoint_type_other waterpoint_type_group_communal standpipe  \
0                      1                                        0
1                      0                                        1
2                      1                                        0
3                      1                                        0
4                      0                                        1

  waterpoint_type_group_dam  waterpoint_type_group_hand pump  \
0                         0                                0
1                         0                                0
2                         0                                0
3                         0                                0
4                         0                                0

  waterpoint_type_group_improved spring waterpoint_type_group_other
0                                     0                           1
1                                     0                           0
2                                     0                           1
3                                     0                           1
4                                     0                           0

[5 rows x 63 columns]
```

In [58]:
```
act_data_no_id = act_data.drop('id', axis=1)

act_data_no_id['date_recorded'] = pd.to_datetime(act_data_no_id['date_recor
ded'], format='%d/%m/%Y')
# Converting datetime to a numeric timestamp (in seconds)
act_data_no_id['date_recorded_numeric'] = act_data_no_id['date_recorded'].a
stype('int64') // 10**9

# Now compute the correlation matrix using the new numeric date column
numeric_data = act_data_no_id.select_dtypes(include=['number'])
corr = numeric_data.corr()
plt.figure(figsize=(15, 12))
sns.heatmap(corr, annot=True, fmt=".1f", cmap='coolwarm',annot_kws={"size":
8})
plt.title("Correlation Matrix")
plt.tight_layout()
plt.show()
```



In [60]:
```
numeric_d = act_data_no_id.select_dtypes(include=[np.number])
X_test = numeric_d
X_test.head
```

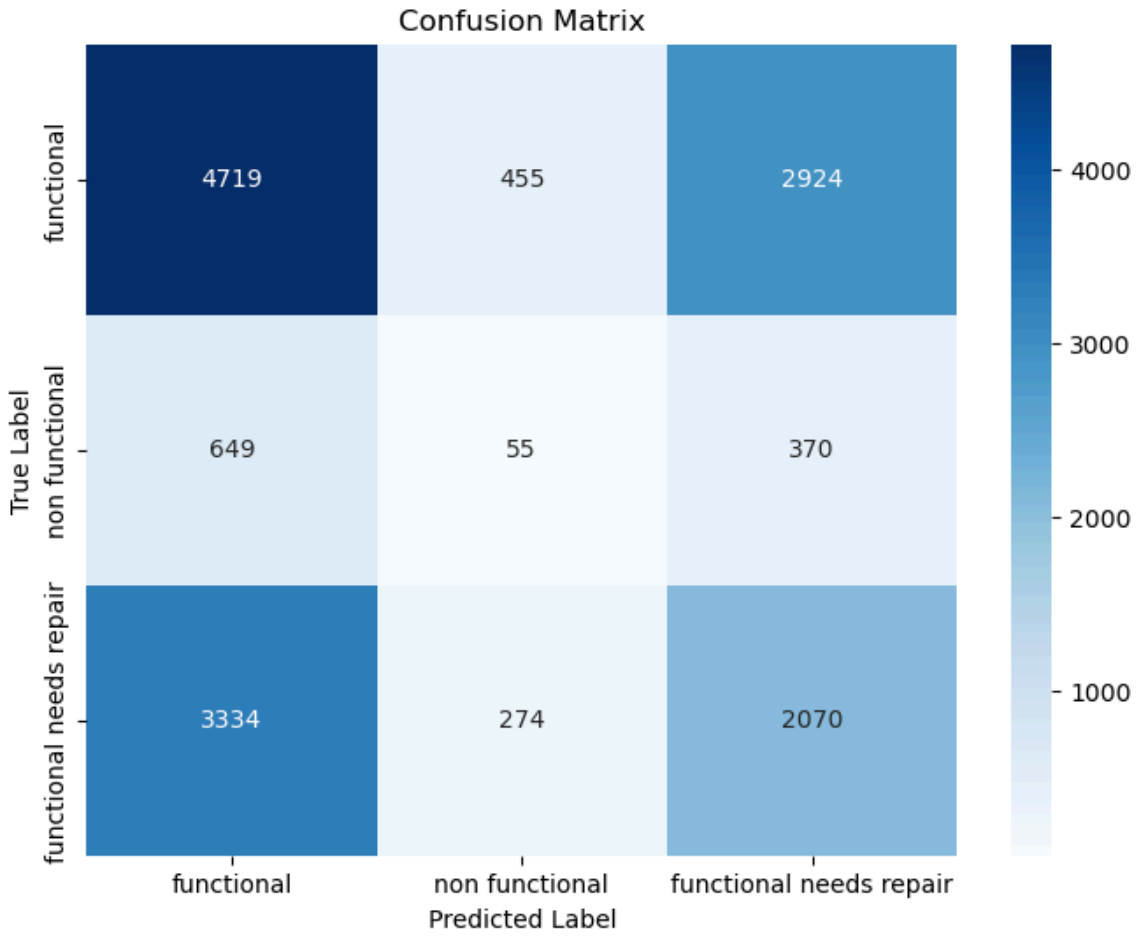
Out[60]:
```
<bound method NDFrame.head of        amount_tsh  gps_height  longitude   l
atitude  num_private  region_code  \
0             0.0        1996  35.290799  -4.059696            0
21
1             0.0        1569  36.656709  -3.309214            0
2
2             0.0        1567  34.767863  -5.004344            0
13
3             0.0         267  38.058046  -9.418672            0
80
4           500.0        1260  35.006123 -10.950412            0
10
...           ...         ...        ...        ...          ...
...
14845         0.0          34  38.852669  -6.582841            0
6
14846      1000.0           0  37.451633  -5.350428            0
4
14847         0.0        1476  34.739804  -4.585587            0
13
14848         0.0         998  35.432732 -10.584159            0
10
14849         0.0         481  34.765054 -11.226012            0
10

       district_code  population  construction_year  water_quality  ...  \
0                  3         321               2012              6  ...
1                  2         300               2000              6  ...
2                  2         500               2010              6  ...
3                 43         250               1987              6  ...
4                  3          60               2000              6  ...
...              ...         ...                ...            ...  ...
14845              1          20               1988              6  ...
14846              7        2960               1994              4  ...
14847              2         200               2010              6  ...
14848              2         150               2009              6  ...
14849              3          40               2008              6  ...

       waterpoint_type_dam  waterpoint_type_hand pump  \
0                        0                          0
1                        0                          0
2                        0                          0
3                        0                          0
4                        0                          0
...                    ...                        ...
14845                    0                          0
14846                    0                          1
14847                    0                          0
14848                    0                          0
14849                    0                          0

       waterpoint_type_improved spring  waterpoint_type_other  \
0                                    0                      1
1                                    0                      0
2                                    0                      1
3                                    0                      1
4                                    0                      0
...                                ...                    ...
14845                                0                      0
14846                                0                      0
14847                                0                      0
```

```
14848                                0                      0
14849                                0                      0

       waterpoint_type_group_communal standpipe  waterpoint_type_group_dam
\
0                                            0                          0
1                                            1                          0
2                                            0                          0
3                                            0                          0
4                                            1                          0
...                                        ...                        ...
14845                                        1                          0
14846                                        0                          0
14847                                        1                          0
14848                                        1                          0
14849                                        1                          0

       waterpoint_type_group_hand pump  waterpoint_type_group_improved spr
ing  \
0                                    0
0
1                                    0
0
2                                    0
0
3                                    0
0
4                                    0
0
...                                ...
...
14845                                0
0
14846                                1
0
14847                                0
0
14848                                0
0
14849                                0
0

       waterpoint_type_group_other  date_recorded_numeric
0                                1             1359936000
1                                0             1359936000
2                                1             1359676800
3                                1             1358812800
4                                0             1364342400
...                            ...                    ...
14845                            0             1298505600
14846                            0             1300665600
14847                            0             1362355200
14848                            0             1361145600
14849                            0             1360713600

[14850 rows x 42 columns]>
```

Evaluating the model

In [68]:
```
y_pred = pipeline.predict(X_test)
cm = confusion_matrix(y_test, y_pred)
print("Confusion Matrix:")
print(cm)
plt.figure(figsize=(8, 6))
class_names = le.classes_
target_names = ['functional', 'non functional', 'functional needs repair']

sns.heatmap(cm, annot=True, fmt="d", cmap="Blues",
            xticklabels=target_names, yticklabels=target_names)
plt.xlabel('Predicted Label')
plt.ylabel('True Label')
plt.title('Confusion Matrix')
plt.show()

print(classification_report(y_test, y_pred, target_names=target_names))
```

```
Confusion Matrix:
[[4719  455 2924]
 [ 649   55  370]
 [3334  274 2070]]
```




```
                    precision    recall  f1-score   support

        functional       0.54      0.58      0.56      8098
    non functional       0.07      0.05      0.06      1074
functional needs repair  0.39      0.36      0.37      5678

          accuracy                           0.46     14850
         macro avg       0.33      0.33      0.33     14850
      weighted avg       0.45      0.46      0.45     14850
```

Our model could predict 54% of the occurences in the functional class, and 39% of the functional needs repair class. It, however, could only predict 7% of the non-functional class.For the functional and functional-needs-repair classes f1-scores of 0.58 and 0.37 reflects a poor performance by the model. For the non-functional class ,moreover, a f1-score of 0.06 indicates that both precision and recall are low for this class. The actual number of occurences "support" for the non-functional class is very small compared to the other classes. The model generally had a poor score only being able to predict less than 10% of the non-functional class and barely 60% for the functional class which was the best performing class. A low number of samples is most likely the reason for the poor precision, recall and f1 score

In [ ]: