Test Set Preprocessing

In this milestone, you will go beyond the descriptive statistics you completed in the last milestone. This milestone is really about diving deeper to analyze your data, beyond descriptive stats. Maybe you need to analyze qualitative data or textual data to get a full picture.

Field	Description
amount_tsh	Total static head (amount water available to waterpoint)
date_recorded	The date the row was entered
funder	Who funded the well
gps_height	Altitude of the well
installer	Organization that installed the well
longitude	GPS coordinate
latitude	GPS coordinate
wpt_name	Name of the waterpoint if there is one
num_private	
basin	Geographic water basin
subvillage	Geographic location
region	Geographic location
region_code	Geographic location (coded)
district_code	Geographic location (coded)
lga	Geographic location
ward	Geographic location
population	Population around the well
public_meeting	True/False
recorded_by	Group entering this row of data
scheme_management	Who operates the waterpoint
scheme_name	Who operates the waterpoint
permit	If the waterpoint is permitted
construction_year	Year the waterpoint was constructed
extraction_type	The kind of extraction the waterpoint uses
extraction_type_group	The kind of extraction the waterpoint uses
extraction_type_class	The kind of extraction the waterpoint uses
management	How the waterpoint is managed
management_group	How the waterpoint is managed
payment	What the water costs
payment_type	What the water costs
water_quality	The quality of the water
quality_group	The quality of the water
quantity	The quantity of water
quantity_group	The quantity of water
source	The source of the water
source_type	The source of the water
source_class	The source of the water
waterpoint_type	The kind of waterpoint
waterpoint_type_group	The kind of waterpoint
functional	the waterpoint is operational and there are no repairs needed
non functional	the waterpoint is not operational

Import Libraries

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
import datetime

%matplotlib inline
sns.set_style('dark')
sns.set(font_scale=1.2)

import warnings
warnings.filterwarnings('ignore')

import feature_engine.missing_data_imputers as mdi
from feature_engine.outlier_removers import Winsorizer

from sklearn.preprocessing import LabelEncoder, StandardScaler, MinMaxScaler, OneHotEncoder

pd.set_option('display.max_columns',None)
#pd.set_option('display.max_rows',None)
pd.set_option('display.width', 1000)

np.random.seed(0)
np.set_printoptions(suppress=True)
```

In [2]:
```
df = pd.read_csv("test.csv",parse_dates=['date_recorded'])
```

In [3]:
```
df
```

Out[3]:

	id	amount_tsh	date_recorded	funder	gps_height	installer	longitude	latitude	wpt_name	num_private	basin
0	50785	0.0	2013-02-04	Dmdd	1996	DMDD	35.290799	-4.059696	Dinamu Secondary School	0	Interna
1	51630	0.0	2013-02-04	Government Of Tanzania	1569	DWE	36.656709	-3.309214	Kimnyak	0	Pangan
2	17168	0.0	2013-02-01	NaN	1567	NaN	34.767863	-5.004344	Puma Secondary	0	Interna
3	45559	0.0	2013-01-22	Finn Water	267	FINN WATER	38.058046	-9.418672	Kwa Mzee Pange	0	Ruvuma Southern Coas
4	49871	500.0	2013-03-27	Bruder	1260	BRUDER	35.006123	-10.950412	Kwa Mzee Turuka	0	Ruvuma Southern Coas
...	...	...	...	...	...	...	...	...	...	...	...
14845	39307	0.0	2011-02-24	Danida	34	Da	38.852669	-6.582841	Kwambwezi	0	Wami Ruvu
14846	18990	1000.0	2011-03-21	Hiap	0	HIAP	37.451633	-5.350428	Bonde La Mkondoa	0	Pangan
14847	28749	0.0	2013-03-04	NaN	1476	NaN	34.739804	-4.585587	Bwawani	0	Interna
14848	33492	0.0	2013-02-18	Germany	998	DWE	35.432732	-10.584159	Kwa John	0	Lake Nyasa
14849	68707	0.0	2013-02-13	Government Of Tanzania	481	Government	34.765054	-11.226012	Kwa Mzee Chagala	0	Lake Nyasa

14850 rows × 40 columns

Exploratory Data Analysis

In [4]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14850 entries, 0 to 14849
Data columns (total 40 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   id                     14850 non-null  int64
 1   amount_tsh             14850 non-null  float64
 2   date_recorded          14850 non-null  datetime64[ns]
 3   funder                 13981 non-null  object
 4   gps_height             14850 non-null  int64
 5   installer              13973 non-null  object
 6   longitude              14850 non-null  float64
 7   latitude               14850 non-null  float64
 8   wpt_name               14850 non-null  object
 9   num_private            14850 non-null  int64
 10  basin                  14850 non-null  object
 11  subvillage             14751 non-null  object
 12  region                 14850 non-null  object
 13  region_code            14850 non-null  int64
 14  district_code          14850 non-null  int64
 15  lga                    14850 non-null  object
 16  ward                   14850 non-null  object
 17  population             14850 non-null  int64
 18  public_meeting         14029 non-null  object
 19  recorded_by            14850 non-null  object
 20  scheme_management      13881 non-null  object
 21  scheme_name            7758 non-null   object
 22  permit                 14113 non-null  object
 23  construction_year      14850 non-null  int64
 24  extraction_type        14850 non-null  object
 25  extraction_type_group  14850 non-null  object
 26  extraction_type_class  14850 non-null  object
 27  management             14850 non-null  object
 28  management_group       14850 non-null  object
 29  payment                14850 non-null  object
 30  payment_type           14850 non-null  object
 31  water_quality          14850 non-null  object
 32  quality_group          14850 non-null  object
 33  quantity               14850 non-null  object
 34  quantity_group         14850 non-null  object
 35  source                 14850 non-null  object
 36  source_type            14850 non-null  object
 37  source_class           14850 non-null  object
 38  waterpoint_type        14850 non-null  object
 39  waterpoint_type_group  14850 non-null  object
dtypes: datetime64[ns](1), float64(3), int64(7), object(29)
memory usage: 4.5+ MB
```

In [5]:
```
df.describe()
```

Out[5]:

	id	amount_tsh	gps_height	longitude	latitude	num_private	region_code	district_code	population
count	14850.000000	14850.000000	14850.000000	14850.000000	1.485000e+04	14850.000000	14850.000000	14850.000000	14850.000000
mean	37161.972929	322.826983	655.147609	34.061605	-5.684724e+00	0.415084	15.139057	5.626397	184.114209
std	21359.364833	2510.968644	691.261185	6.593034	2.940803e+00	8.167910	17.191329	9.673842	469.499332
min	10.000000	0.000000	-57.000000	0.000000	-1.156459e+01	0.000000	1.000000	0.000000	0.000000
25%	18727.000000	0.000000	0.000000	33.069455	-8.443970e+00	0.000000	5.000000	2.000000	0.000000
50%	37361.500000	0.000000	344.000000	34.901215	-5.049750e+00	0.000000	12.000000	3.000000	20.000000
75%	55799.750000	25.000000	1308.000000	37.196594	-3.320594e+00	0.000000	17.000000	5.000000	220.000000
max	74249.000000	200000.000000	2777.000000	40.325016	-2.000000e-08	669.000000	99.000000	80.000000	11469.000000

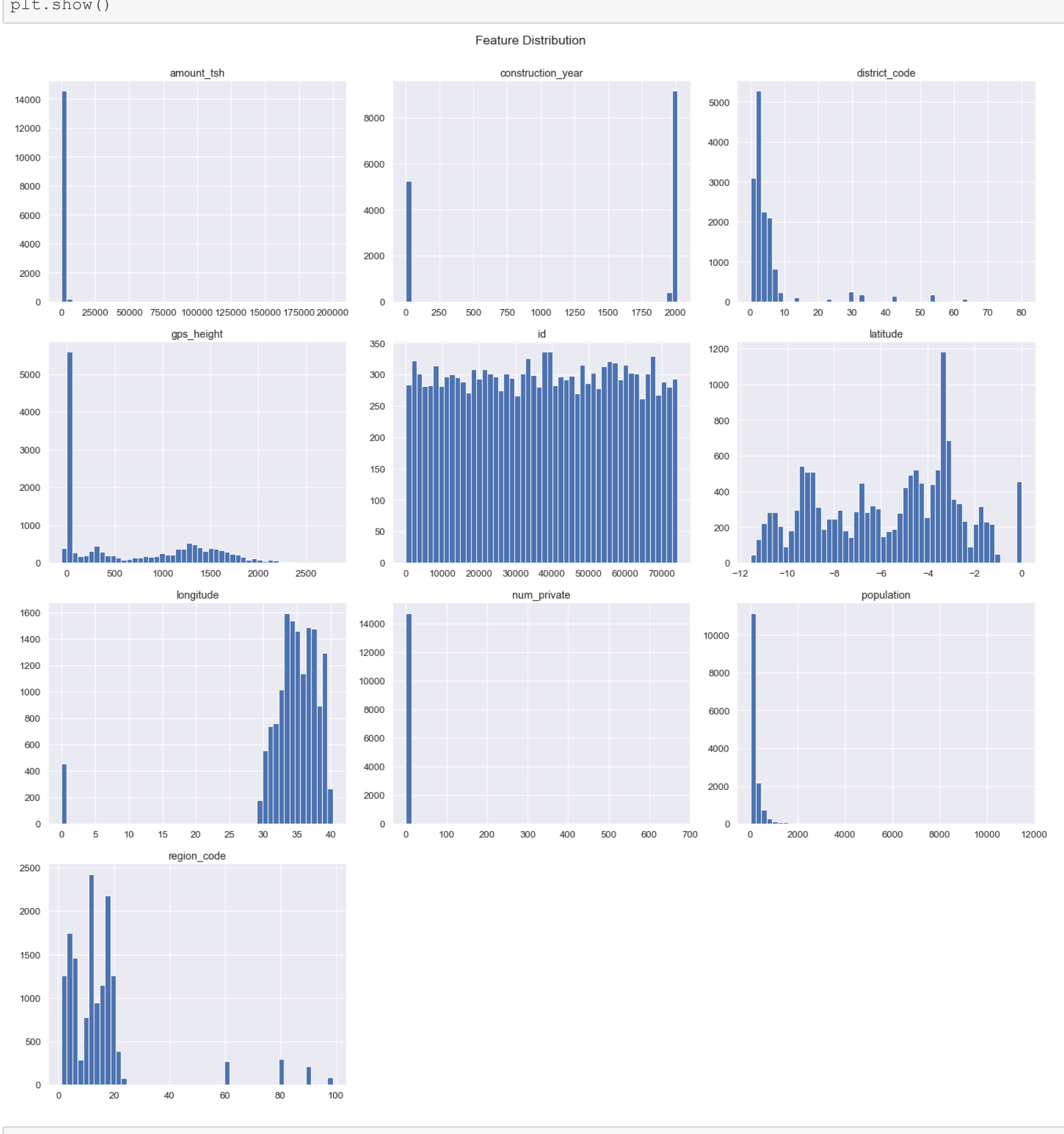
In [6]:
```
df.columns
```

Out[6]:
```
Index(['id', 'amount_tsh', 'date_recorded', 'funder', 'gps_height', 'installer', 'longitude', 'latitude', 'wpt_name', 'num_private', 'basin', 'subvillage', 'region', 'region_code', 'district_code', 'lga', 'ward', 'population', 'public_meeting', 'recorded_by', 'scheme_management', 'scheme_name', 'permit', 'construction_year', 'extraction_type', 'extraction_type_group', 'extraction_type_class', 'management', 'management_group', 'payment', 'payment_type', 'water_quality', 'quality_group', 'quantity', 'quantity_group', 'source', 'source_type', 'source_class', 'waterpoint_type', 'waterpoint_type_group'], dtype='object')
```

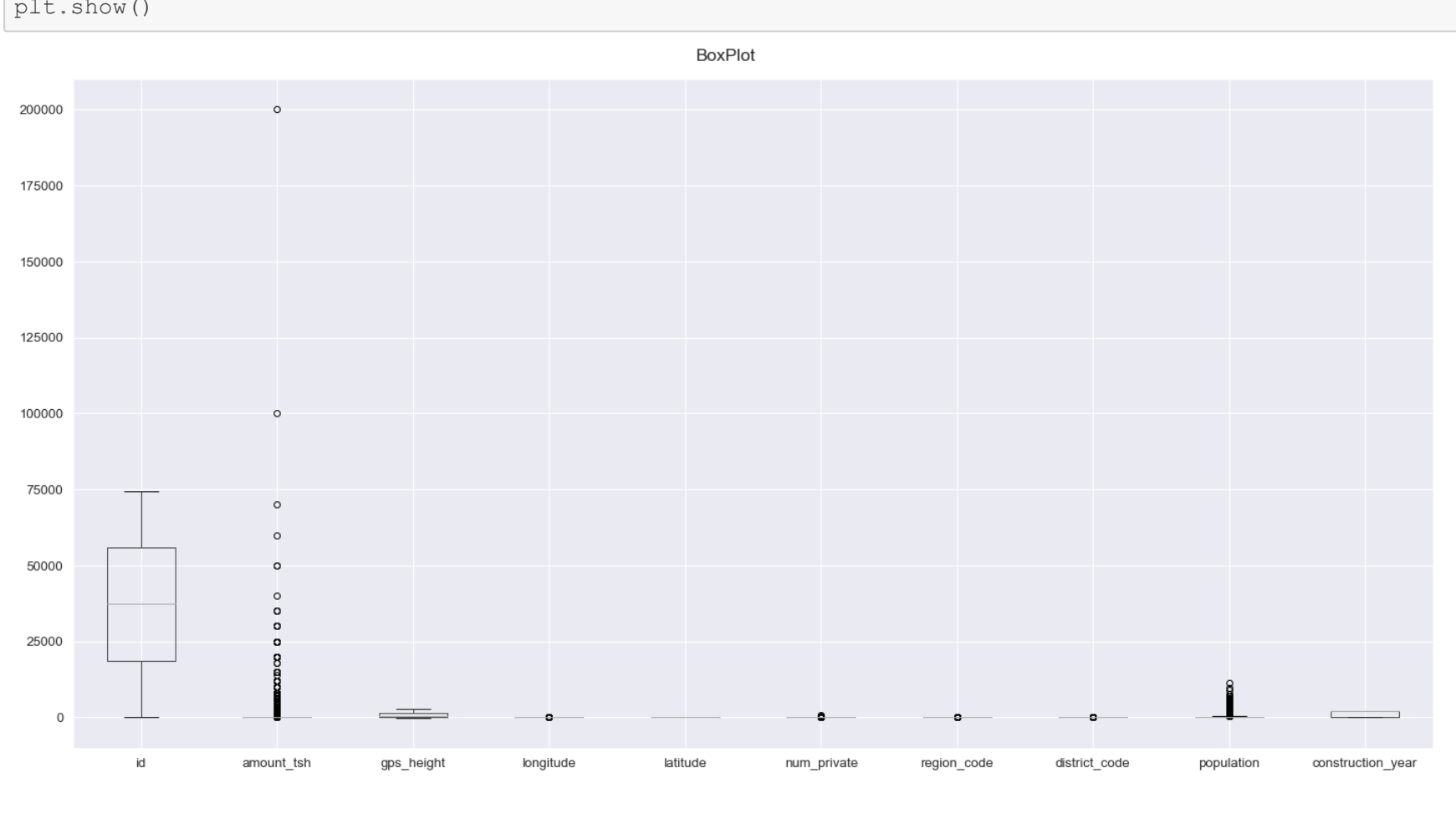
Data Visualization

Univariate Data Exploration

In [7]:
```
df.hist(bins=50, figsize=(20,20))
plt.suptitle('Feature Distribution', x=0.5, y=1.02, ha='center', fontsize='large')
plt.tight_layout()
plt.show()
```



In [8]:
```
df.boxplot(figsize=(20,10))
plt.suptitle('BoxPlot', x=0.5, y=1.02, ha='center', fontsize='large')
plt.tight_layout()
plt.show()
```



Drop similarly and irrelevant columns

In [9]:
```
df.columns
```

Out[9]:
```
Index(['id', 'amount_tsh', 'date_recorded', 'funder', 'gps_height', 'installer', 'longitude', 'latitude', 'wpt_name', 'num_private', 'basin', 'subvillage', 'region', 'region_code', 'district_code', 'lga', 'ward', 'population', 'public_meeting', 'recorded_by', 'scheme_management', 'scheme_name', 'permit', 'construction_year', 'extraction_type', 'extraction_type_group', 'extraction_type_class', 'management', 'management_group', 'payment', 'payment_type', 'water_quality', 'quality_group', 'quantity', 'quantity_group', 'source', 'source_type', 'source_class', 'waterpoint_type', 'waterpoint_type_group'], dtype='object')
```

In [10]:
```
df.drop(['id','date_recorded', 'funder', 'gps_height', 'installer', 'longitude', 'latitude',
        'wpt_name', 'num_private','subvillage', 'region', 'region_code', 'district_code', 'lga', 'ward'
,
        'recorded_by','scheme_name','extraction_type', 'extraction_type_group','management',
        'payment','water_quality','quantity','source', 'source_type','waterpoint_type'],axis=1,inplace=
True)
```

In [11]:
```
df
```

Out[11]:

	amount_tsh	basin	population	public_meeting	scheme_management	permit	construction_year	extraction_type_class
0	0.0	Internal	321	True	Parastatal	True	2012	other
1	0.0	Pangani	300	True	VWC	True	2000	gravity
2	0.0	Internal	500	True	VWC	NaN	2010	other
3	0.0	Ruvuma / Southern Coast	250	NaN	VWC	True	1987	other
4	500.0	Ruvuma / Southern Coast	60	NaN	Water Board	True	2000	gravity
...	...	...	...	...	...	...	...	...
14845	0.0	Wami / Ruvu	20	True	VWC	True	1988	motorpump
14846	1000.0	Pangani	2960	True	VWC	False	1994	handpump
14847	0.0	Internal	200	True	VWC	NaN	2010	gravity
14848	0.0	Lake Nyasa	150	True	VWC	True	2009	gravity
14849	0.0	Lake Nyasa	40	True	VWC	True	2008	gravity

14850 rows × 14 columns

In [12]:
```
#Save to csv
#df.to_csv("test1.csv",index=False)
```

In [13]:
```
df["construction_year"] = df["construction_year"].replace(to_replace=0, value=np.nan)
```

In [14]:
```
df
```

Out[14]:

	amount_tsh	basin	population	public_meeting	scheme_management	permit	construction_year	extraction_type_class
0	0.0	Internal	321	True	Parastatal	True	2012.0	other
1	0.0	Pangani	300	True	VWC	True	2000.0	gravity
2	0.0	Internal	500	True	VWC	NaN	2010.0	other
3	0.0	Ruvuma / Southern Coast	250	NaN	VWC	True	1987.0	other
4	500.0	Ruvuma / Southern Coast	60	NaN	Water Board	True	2000.0	gravity
...	...	...	...	...	...	...	...	...
14845	0.0	Wami / Ruvu	20	True	VWC	True	1988.0	motorpump
14846	1000.0	Pangani	2960	True	VWC	False	1994.0	handpump
14847	0.0	Internal	200	True	VWC	NaN	2010.0	gravity
14848	0.0	Lake Nyasa	150	True	VWC	True	2009.0	gravity
14849	0.0	Lake Nyasa	40	True	VWC	True	2008.0	gravity

14850 rows × 14 columns

In [15]:
```
df['age'] = 2020 - df["construction_year"]
```

In [16]:
```
df
```

Out[16]:

	amount_tsh	basin	population	public_meeting	scheme_management	permit	construction_year	extraction_type_class
0	0.0	Internal	321	True	Parastatal	True	2012.0	other
1	0.0	Pangani	300	True	VWC	True	2000.0	gravity
2	0.0	Internal	500	True	VWC	NaN	2010.0	other
3	0.0	Ruvuma / Southern Coast	250	NaN	VWC	True	1987.0	other
4	500.0	Ruvuma / Southern Coast	60	NaN	Water Board	True	2000.0	gravity
...	...	...	...	...	...	...	...	...
14845	0.0	Wami / Ruvu	20	True	VWC	True	1988.0	motorpump
14846	1000.0	Pangani	2960	True	VWC	False	1994.0	handpump
14847	0.0	Internal	200	True	VWC	NaN	2010.0	gravity
14848	0.0	Lake Nyasa	150	True	VWC	True	2009.0	gravity
14849	0.0	Lake Nyasa	40	True	VWC	True	2008.0	gravity

14850 rows × 15 columns

In [17]:
```
df.drop(['construction_year'],axis=1,inplace=True)
```

In [18]:
```
df
```

Out[18]:

	amount_tsh	basin	population	public_meeting	scheme_management	permit	extraction_type_class	management_group
0	0.0	Internal	321	True	Parastatal	True	other	parastatal
1	0.0	Pangani	300	True	VWC	True	gravity	user-group
2	0.0	Internal	500	True	VWC	NaN	other	user-group
3	0.0	Ruvuma / Southern Coast	250	NaN	VWC	True	other	user-group
4	500.0	Ruvuma / Southern Coast	60	NaN	Water Board	True	gravity	user-group
...	...	...	...	...	...	...	...	...
14845	0.0	Wami / Ruvu	20	True	VWC	True	motorpump	user-group
14846	1000.0	Pangani	2960	True	VWC	False	handpump	user-group
14847	0.0	Internal	200	True	VWC	NaN	gravity	user-group
14848	0.0	Lake Nyasa	150	True	VWC	True	gravity	user-group
14849	0.0	Lake Nyasa	40	True	VWC	True	gravity	user-group

14850 rows × 14 columns

In [19]:
```
#Save to csv
#df.to_csv("test2.csv",index=False)
```

In [20]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14850 entries, 0 to 14849
Data columns (total 14 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   amount_tsh             14850 non-null  float64
 1   basin                  14850 non-null  object
 2   population             14850 non-null  int64
 3   public_meeting         14029 non-null  object
 4   scheme_management      13881 non-null  object
 5   permit                 14113 non-null  object
 6   extraction_type_class  14850 non-null  object
 7   management_group       14850 non-null  object
 8   payment_type           14850 non-null  object
 9   quality_group          14850 non-null  object
 10  quantity_group         14850 non-null  object
 11  source_class           14850 non-null  object
 12  waterpoint_type_group  14850 non-null  object
 13  age                    9590 non-null   float64
dtypes: float64(2), int64(1), object(11)
memory usage: 1.6+ MB
```

Treat Missing Values

In [21]:
```
df.isnull().sum()
```

Out[21]:
```
amount_tsh                  0
basin                       0
population                  0
public_meeting            821
scheme_management         969
permit                    737
extraction_type_class       0
management_group            0
payment_type                0
quality_group               0
quantity_group              0
source_class                0
waterpoint_type_group       0
age                      5260
dtype: int64
```

In [22]:
```
df.dropna(inplace=True)
```

In [23]:
```
df.isnull().sum()
```

Out[23]:
```
amount_tsh               0
basin                    0
population               0
public_meeting           0
scheme_management        0
permit                   0
extraction_type_class    0
management_group         0
payment_type             0
quality_group            0
quantity_group           0
source_class             0
waterpoint_type_group    0
age                      0
dtype: int64
```

In [24]:
```
df
```

Out[24]:

	amount_tsh	basin	population	public_meeting	scheme_management	permit	extraction_type_class	management_group
0	0.0	Internal	321	True	Parastatal	True	other	parastatal
1	0.0	Pangani	300	True	VWC	True	gravity	user-group
5	0.0	Pangani	200	True	VWC	True	gravity	user-group
6	0.0	Rufiji	600	True	VWC	True	handpump	user-group
7	0.0	Ruvuma / Southern Coast	1	True	Water Board	True	submersible	user-group
...	...	...	...	...	...	...	...	...
14843	0.0	Lake Tanganyika	1400	True	Water authority	False	other	user-group
14845	0.0	Wami / Ruvu	20	True	VWC	True	motorpump	user-group
14846	1000.0	Pangani	2960	True	VWC	False	handpump	user-group
14848	0.0	Lake Nyasa	150	True	VWC	True	gravity	user-group
14849	0.0	Lake Nyasa	40	True	VWC	True	gravity	user-group

8073 rows × 14 columns

In [25]:
```
df.reset_index(drop=True,inplace=True)
```

In [26]:
```
df
```

Out[26]:

	amount_tsh	basin	population	public_meeting	scheme_management	permit	extraction_type_class	management_group
0	0.0	Internal	321	True	Parastatal	True	other	parastatal
1	0.0	Pangani	300	True	VWC	True	gravity	user-group
2	0.0	Pangani	200	True	VWC	True	gravity	user-group
3	0.0	Rufiji	600	True	VWC	True	handpump	user-group
4	0.0	Ruvuma / Southern Coast	1	True	Water Board	True	submersible	user-group
...	...	...	...	...	...	...	...	...
8068	0.0	Lake Tanganyika	1400	True	Water authority	False	other	user-group
8069	0.0	Wami / Ruvu	20	True	VWC	True	motorpump	user-group
8070	1000.0	Pangani	2960	True	VWC	False	handpump	user-group
8071	0.0	Lake Nyasa	150	True	VWC	True	gravity	user-group
8072	0.0	Lake Nyasa	40	True	VWC	True	gravity	user-group

8073 rows × 14 columns

In [27]:
```
#Save to csv
#df.to_csv("test3.csv",index=False)
```

Correlation

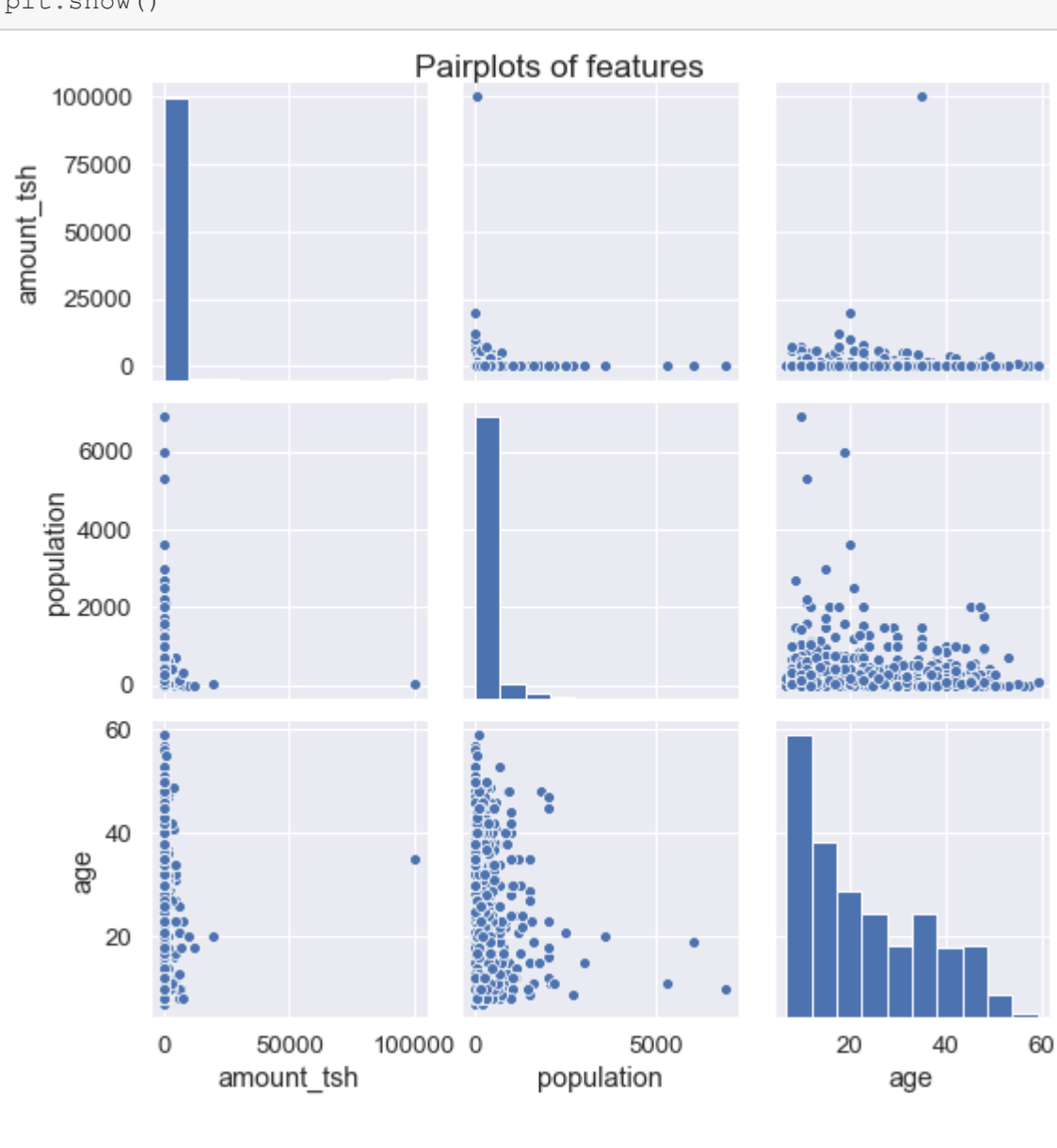
In [28]:
```
df.corr()
```

Out[28]:

	amount_tsh	population	age
amount_tsh	1.000000	0.007590	-0.012931
population	0.007590	1.000000	-0.030211
age	-0.012931	-0.030211	1.000000

In [29]:
```
plt.figure(figsize=(16,5))
sns.heatmap(df.corr(),cmap="coolwarm",annot=True,fmt='.2f',linewidths=2)
plt.show()
```

In [30]:
```
sns.pairplot(df[['amount_tsh','population','age']].sample(1000))
plt.suptitle('Pairplots of features', x=0.5, y=1.02, ha='center', fontsize='large')
plt.show()
```



Treat Duplicate Values

In [31]:
```
df.duplicated(keep='first').sum()
```

Out[31]:
```
1498
```

In [32]:
```
df[df.duplicated(keep=False)] #Check duplicate values
```

Out[32]:

	amount_tsh	basin	population	public_meeting	scheme_management	permit	extraction_type_class	management_group
2	0.0	Pangani	200	True	VWC	True	gravity	user-group
4	0.0	Ruvuma / Southern Coast	1	True	Water Board	True	submersible	user-group
6	0.0	Pangani	1	True	Water Board	True	gravity	user-group
7	5.0	Pangani	75	True	Water Board	True	gravity	user-group
8	1000.0	Rufiji	1	True	VWC	True	gravity	user-group
...	...	...	...	...	...	...	...	...
8055	5.0	Pangani	45	True	Water Board	True	gravity	user-group
8061	1000.0	Lake Nyasa	0	True	VWC	False	gravity	user-group
8062	0.0	Rufiji	40	True	VWC	False	handpump	user-group
8064	3000.0	Lake Nyasa	0	True	VWC	False	gravity	user-group
8067	600.0	Lake Tanganyika	230	True	WUG	True	gravity	user-group

2197 rows × 14 columns

Treat Outliers

In [33]:
```
df.columns
```

Out[33]:
```
Index(['amount_tsh', 'basin', 'population', 'public_meeting', 'scheme_management', 'permit', 'extraction_type_class', 'management_group', 'payment_type', 'quality_group', 'quantity_group', 'source_class', 'waterpoint_type_group', 'age'], dtype='object')
```

In [34]:
```
df.describe()
```

Out[34]:

	amount_tsh	population	age
count	8073.000000	8073.000000	8073.000000
mean	544.618296	265.841571	22.966060
std	3293.672667	554.746029	12.426148
min	0.000000	0.000000	7.000000
25%	0.000000	20.000000	12.000000
50%	0.000000	125.000000	19.000000
75%	250.000000	300.000000	33.000000
max	200000.000000	11469.000000	60.000000

In [35]:
```
windsorizer = Winsorizer(distribution='skewed',tail='right',fold=3.0, variables=['amount_tsh','populati
on'])
```

In [36]:
```
windsorizer.fit(df)
```

Out[36]:
```
Winsorizer(distribution='skewed', fold=3.0,
           variables=['amount_tsh', 'population'])
```

In [37]:
```
df2 = windsorizer.transform(df)
```

In [38]:
```
df2
```

Out[38]:

	amount_tsh	basin	population	public_meeting	scheme_management	permit	extraction_type_class	management_group
0	0.0	Internal	321.0	True	Parastatal	True	other	parastatal
1	0.0	Pangani	300.0	True	VWC	True	gravity	user-group
2	0.0	Pangani	200.0	True	VWC	True	gravity	user-group
3	0.0	Rufiji	600.0	True	VWC	True	handpump	user-group
4	0.0	Ruvuma / Southern Coast	1.0	True	Water Board	True	submersible	user-group
...	...	...	...	...	...	...	...	...
8068	0.0	Lake Tanganyika	1140.0	True	Water authority	False	other	user-group
8069	0.0	Wami / Ruvu	20.0	True	VWC	True	motorpump	user-group
8070	1000.0	Pangani	1140.0	True	VWC	False	handpump	user-group
8071	0.0	Lake Nyasa	150.0	True	VWC	True	gravity	user-group
8072	0.0	Lake Nyasa	40.0	True	VWC	True	gravity	user-group

8073 rows × 14 columns

In [39]:
```
df2.describe()
```

Out[39]:

	amount_tsh	population	age
count	8073.000000	8073.000000	8073.000000
mean	196.662145	218.979066	22.966060
std	339.750561	272.294408	12.426148
min	0.000000	0.000000	7.000000
25%	0.000000	20.000000	12.000000
50%	0.000000	125.000000	19.000000
75%	250.000000	300.000000	33.000000
max	1000.000000	1140.000000	60.000000

In [40]:
```
windsorizer.left_tail_caps_
```

Out[40]:
```
{}
```

In [41]:
```
windsorizer.right_tail_caps_
```

Out[41]:
```
{'amount_tsh': 1000.0, 'population': 1140.0}
```

In [42]:
```
#Save to csv
#df2.to_csv("test4.csv",index=False)
```

In [43]:
```
df2.columns
```

Out[43]:
```
Index(['amount_tsh', 'basin', 'population', 'public_meeting', 'scheme_management', 'permit', 'extraction_type_class', 'management_group', 'payment_type', 'quality_group', 'quantity_group', 'source_class', 'waterpoint_type_group', 'age'], dtype='object')
```

In [46]:
```
#Rearrange columns
df2 = df2[['amount_tsh','age','population','basin','public_meeting', 'scheme_management', 'permit',
          'extraction_type_class', 'management_group', 'payment_type', 'quality_group', 'quantity_grou
p',
          'source_class', 'waterpoint_type_group']]
```

In [47]:
```
df2
```
Out[47]:

	amount_tsh	age	population	basin	public_meeting	scheme_management	permit	extraction_type_class	management_group
0	0.0	8.0	321.0	Internal	True	Parastatal	True	other	parastatal
1	0.0	20.0	300.0	Pangani	True	VWC	True	gravity	user-group
2	0.0	30.0	200.0	Pangani	True	VWC	True	gravity	user-group
3	0.0	13.0	600.0	Rufiji	True	VWC	True	handpump	user-group
4	0.0	38.0	1.0	Ruvuma / Southern Coast	True	Water Board	True	submersible	user-group
...	...	...	...	...	...	...	...	...	...
8068	0.0	25.0	1140.0	Lake Tanganyika	True	Water authority	False	other	user-group
8069	0.0	32.0	20.0	Wami / Ruvu	True	VWC	True	motorpump	user-group
8070	1000.0	26.0	1140.0	Pangani	True	VWC	False	handpump	user-group
8071	0.0	11.0	150.0	Lake Nyasa	True	VWC	True	gravity	user-group
8072	0.0	12.0	40.0	Lake Nyasa	True	VWC	True	gravity	user-group

8073 rows × 14 columns

Perform One-Hot Encoding

In [48]:
```
df2.info()
```
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8073 entries, 0 to 8072
Data columns (total 14 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   amount_tsh             8073 non-null   float64
 1   age                    8073 non-null   float64
 2   population             8073 non-null   float64
 3   basin                  8073 non-null   object
 4   public_meeting         8073 non-null   object
 5   scheme_management      8073 non-null   object
 6   permit                 8073 non-null   object
 7   extraction_type_class  8073 non-null   object
 8   management_group       8073 non-null   object
 9   payment_type           8073 non-null   object
 10  quality_group          8073 non-null   object
 11  quantity_group         8073 non-null   object
 12  source_class           8073 non-null   object
 13  waterpoint_type_group  8073 non-null   object
dtypes: float64(3), object(11)
memory usage: 883.1+ KB
```

In [49]:
```
df3 = pd.get_dummies(df2,drop_first=True)
```

In [50]:
```
df3
```
Out[50]:

	amount_tsh	age	population	basin_Lake Nyasa	basin_Lake Rukwa	basin_Lake Tanganyika	basin_Lake Victoria	basin_Pangani	basin_Rufiji	basin_Ruvuma / Southern Coast
0	0.0	8.0	321.0	0	0	0	0	0	0	0
1	0.0	20.0	300.0	0	0	0	0	1	0	0
2	0.0	30.0	200.0	0	0	0	0	1	0	0
3	0.0	13.0	600.0	0	0	0	0	0	1	0
4	0.0	38.0	1.0	0	0	0	0	0	0	1
...	...	...	...	...	...	...	...	...	...	...
8068	0.0	25.0	1140.0	0	0	1	0	0	0	0
8069	0.0	32.0	20.0	0	0	0	0	0	0	0
8070	1000.0	26.0	1140.0	0	0	0	0	1	0	0
8071	0.0	11.0	150.0	1	0	0	0	0	0	0
8072	0.0	12.0	40.0	1	0	0	0	0	0	0

8073 rows × 55 columns

In [51]:
```
df3.columns
```
Out[51]:
```
Index(['amount_tsh', 'age', 'population', 'basin_Lake Nyasa', 'basin_Lake Rukwa', 'basin_Lake Tanganyika', 'basin_Lake Victoria', 'basin_Pangani', 'basin_Rufiji', 'basin_Ruvuma / Southern Coast', 'basin_Wami / Ruvu', 'public_meeting_True', 'scheme_management_Other', 'scheme_management_Parastatal', 'scheme_management_Private operator', 'scheme_management_SWC', 'scheme_management_Trust', 'scheme_management_VWC', 'scheme_management_WUA', 'scheme_management_WUG', 'scheme_management_Water Board', 'scheme_management_Water authority', 'permit_True', 'extraction_type_class_handpump', 'extraction_type_class_motorpump', 'extraction_type_class_other', 'extraction_type_class_rope pump', 'extraction_type_class_submersible', 'extraction_type_class_wind-powered', 'management_group_other', 'management_group_parastatal', 'management_group_unknown', 'management_group_user-group', 'payment_type_monthly', 'payment_type_never pay', 'payment_type_on failure', 'payment_type_other',
       'payment_type_per bucket', 'payment_type_unknown', 'quality_group_fluoride', 'quality_group_good', 'quality_group_milky', 'quality_group_salty', 'quality_group_unknown', 'quantity_group_enough', 'quantity_group_insufficient', 'quantity_group_seasonal', 'quantity_group_unknown', 'source_class_surface', 'source_class_unknown', 'waterpoint_type_group_communal standpipe', 'waterpoint_type_group_dam', 'waterpoint_type_group_hand pump', 'waterpoint_type_group_improved spring', 'waterpoint_type_group_other'],
      dtype='object')
```

In [53]:
```
#Save to csv
#df3.to_csv("test5.csv",index=False)
```

In [ ]: