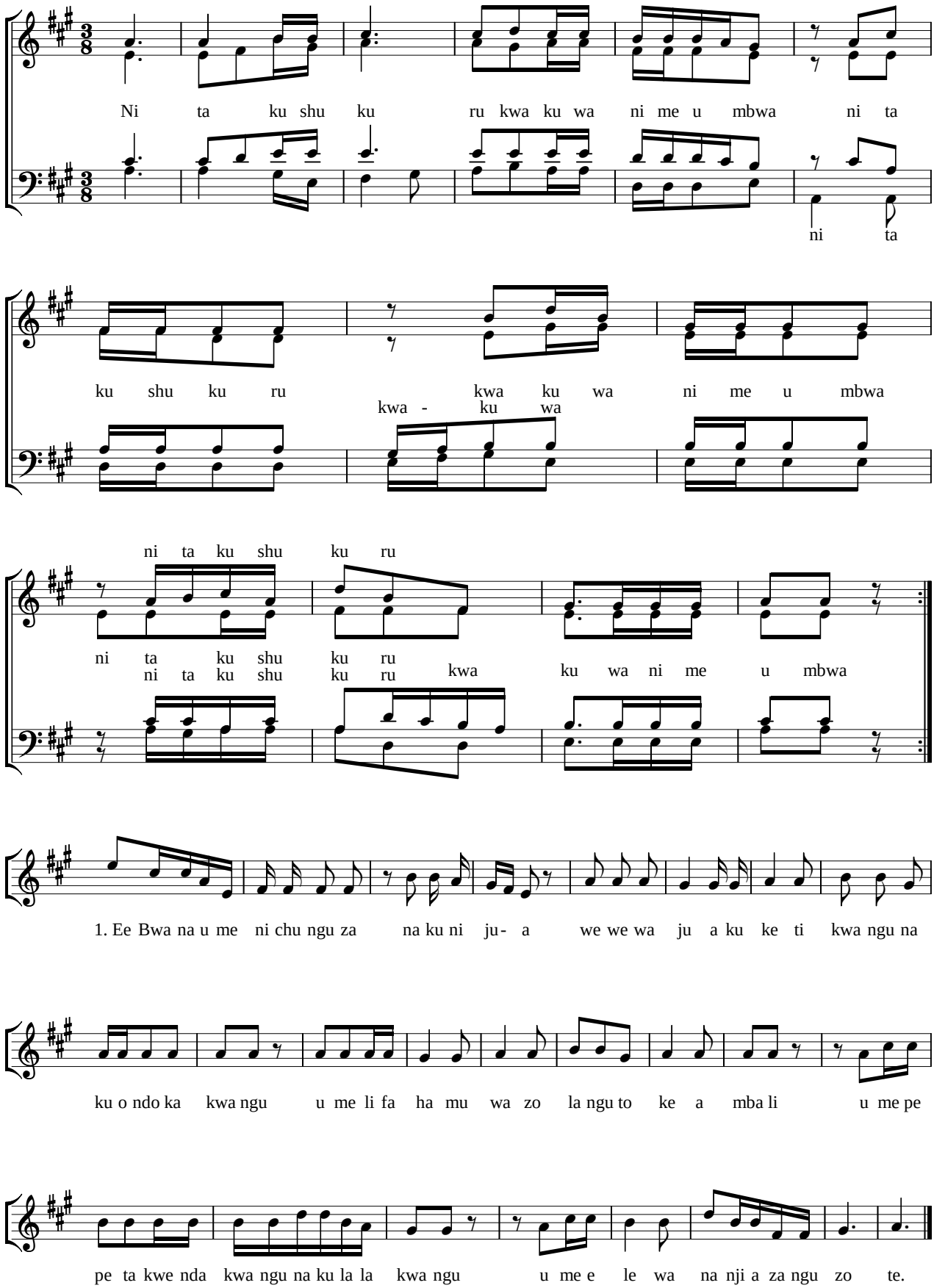
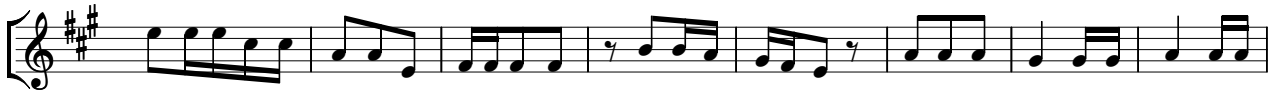
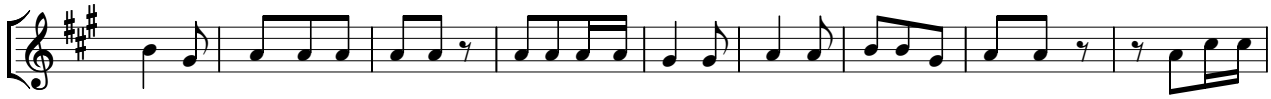
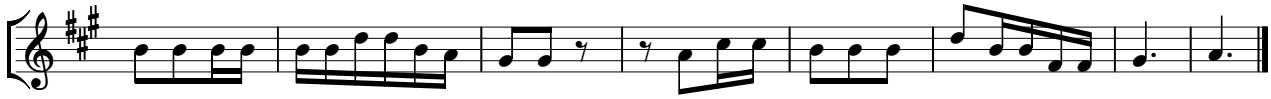
NITAKUSHUKURU KWA KUWA NIMEUMBWA

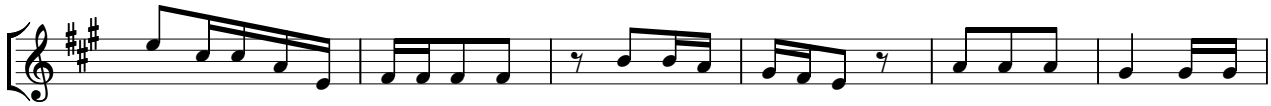
By Baraka Kabuje
Sange - Mbeya
19 June, 2018.

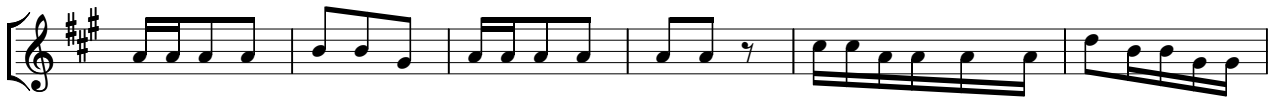


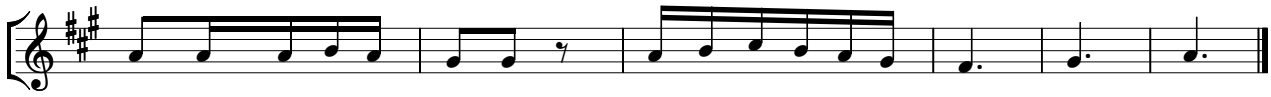

2.Ma a na we we ndi we u li ye u mba m ti ma wa- ngu u li ni u nga tu mbo ni tu


mbo ni mwa ma ma ya ngu ni ta ku shu ku ru kwa ku wa ni me u mbwa kwa ji nsi


ya a ja bu ya a ja bu ya ku ti sha ma te ndo ma te ndo ya ko ni ya a ja bu


3.Na fsi ya ngu ya ju a sa na ya ju a sa - na mi fu pa ya ngu mi


fu pa ya ngu ha i ku si ti ri ka kwa ko ni li po u mbwa kwa si ri ni li po


u mbwa kwa u - sta di pa nde za chi ni za n - chi.

Step 1. Import numpy and pandas library

http://pandas.pydata.org/ (http://pandas.pydata.org/)

In [278]:
```
# your code here
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
```

Step 2. Load tanzania house-hold data set

In [279]:
```
data = pd.read_csv('data/householdtz.csv')
```

The columns are organized as:

- HOUSEHOLD number :house_id
- Number of HOUSEHOLD: households_number
- Number of eligible women in HH: households_women
- Number of eligible men in HH: households_men
- Number of children 5 and under: households_children
- Region: region
- Type of place of residence: residence_type
- Place of residence: residence_place
- Source of drinking water: drink_water_source
- Time to get to water source: time_water_source
- Type of toilet facility: toilet_type
- Has electricity: electricity
- Has radio: radio
- Has television: tv
- Has refrigerator: refrigerator
- Main floor material: floor_material
- Main wall material: wall_material
- Main roof material: roof_material
- Rooms used for sleeping: sleeping_room
- Has a landline telephone: landline
- Share toilet with other households:share_toilet
- Type of cooking fuel: cooking_fuel_type
- Have bednet for sleeping: bednet
- Anything done to water to make safe to drink: treat_water
- Water usually treated by: boil: water_boil
- Water usually treated by: add bleach/chlorine: water_bleach
- Water usually treated by: strain through a cloth: water_strain
- Water usually treated by: use water filter:water_filter
- Water usually treated by: solar disinfection:water_solar
- Water usually treated by: let it stand and settle:water_settle
- Water usually treated by: other: water_other
- Number of households sharing toilet: house_share_toilet
- Wealth index: wealth_index
- Wealth index factor score (5 decimals): wealth_index_score
- Mainland/Zanzibar: location
- Who provides water at main source: water_provider
- Paraffin lamp:paraffin_lamp
- Iron:iron
- Main source of energy for lighting:lighting_source
- How far is the nearest market place (kilometers):distance_market
- How many meals per day:meals_day
- How far is the nearest health facility:distance_health

Step 3. Inspect the data-set and answer the following questions.

- How many rows does it contain? How many column?

Step 4 Statistically describe numerical features

- What is the minimum market walking distance?
- What is the average number of meals per day?
- What is maximum number of women in households?
- What is the average time a person take to reach to water source?
- What is maximum distance a person walk to reach health facility?
- What is the avarage number of households per family?

In [13]:
```
# your code here
data.shape

#From Output 13 we see that the data set has 401 rows and 43 columns
```

Out[13]:
```
(401, 43)
```

In [19]:
```
data['distance_market'].min()

#Output 19 above that: The minimum market walking distance is zero(0)
```

Out[19]:
```
0
```

In [20]:
```
data['meals_day'].mean()

#Output 20 shows that the average number of meals per day is 2.4164588528678306.
#But Now since the number meals is whole we can approximate it to be 2,
#that is the average number of meals per day is APPROXIMATELY 2
```

Out[20]:
```
2.4164588528678306
```

In [22]:
```
data['households_women'].max()

#From output 22 we see that the maximum number of women in households is 5
```

Out[22]:
```
5
```

In [24]:
```
data['time_water_source'].mean()

#in output 24 we get a result that the average time a person take to reach to water
```

Out[24]:
```
56.52618453865337
```

In [25]:
```
data['distance_health'].max()

#From output 25 we see that the maximum distance a person walk to reach health faci
```

Out[25]:
```
36
```

In [26]:
```
data['households_number'].mean()

 #Program Output shows that, the avarage number of households per family is 4.72069
 #but the number of households is whole, now we can approximate it to be 5
 #that is the average number of households per family is APPROXIMATELY 5
```

Out[26]:
```
4.720698254364089
```

Counting categorical values

Use: `.value_counts()`

In [27]:
```
# e.g
data.location.value_counts()
```

Out[27]:
```
Mainland rural    329
Mainland urban     68
Unguja              4
Name: location, dtype: int64
```

Question 1. What are first three regions with lowest average number of house-holds per family?

In [236]:
```
pd.crosstab(data.region, data.households_number)


#From the output 236 we see that the lowest average number of households per family
#Now from that Column we see ARUSHA, DODOMA and KAGERA being the first regions with
#lowest average number of households per family which is 1
```

Out[236]:

households_number	1	2	3	4	5	6	7	8	9	10	11	13	15
region													
Arusha	2	4	3	9	10	8	5	1	0	0	0	0	0
Dodoma	14	24	34	39	35	26	15	14	4	3	3	2	0
Iringa	0	0	2	1	1	0	0	1	0	0	0	0	0
Kagera	1	0	0	1	0	0	0	0	0	0	0	0	0
Kigoma	0	1	2	0	2	2	3	2	1	3	0	0	0
Kilimanjaro	2	2	3	6	6	6	2	0	0	0	0	0	0
Lindi	0	0	0	0	1	0	0	0	0	0	0	0	0
Mbeya	1	0	1	3	0	0	1	1	0	0	0	0	0
Morogoro	0	0	0	3	1	0	2	0	0	0	1	0	0
Mtwara	0	3	3	4	2	2	2	0	0	0	0	0	0
Mwanza	0	1	0	3	2	0	0	0	0	1	1	0	0
Pwani	2	0	0	0	0	0	0	1	0	0	0	0	0
Ruvuma	0	1	0	1	0	0	0	0	0	0	0	0	0
Singida	1	0	1	1	1	0	0	0	0	0	0	0	0
Tabora	0	0	0	0	1	0	0	0	0	0	0	0	1
Tanga	8	6	5	4	4	5	7	0	2	1	0	0	0
Zanzibar North	0	0	0	1	0	0	0	0	1	1	1	0	0

Question 2. How many people in Kigoma live in urban area ?

In [87]:
```
pd.crosstab(data.residence_type, data.region)

#From output 87 there are 5 people in kigoma who live in urban
```

Out[87]:

region	Arusha	Dodoma	Iringa	Kagera	Kigoma	Kilimanjaro	Lindi	Mbeya	Morogoro	Mtwara	Mw
residence_type											
Rural	28	174	5	2	11	18	1	7	7	16	
Urban	14	39	0	0	5	9	0	0	0	0	

Question 3. What types of toilet facility do people use. How many in each type?

In [89]:
```
# your code here
data.toilet_type.value_counts()

#The output is shown in output 89 below
```

Out[89]:
```
Pit latrine - without slab / open pit          266
No facility/bush/field                          72
Pit latrine - with slab                         32
Flush - to pit latrine                          18
Pit latrine - ventilated improved pit (VIP)      8
OTHER                                            2
Flush - to septic tank                           2
99                                               1
Name: toilet_type, dtype: int64
```

Question 4. What is the common source of drinking water for people living in Arusha?

In [90]:
```
# your code
data.drink_water_source.value_counts()


#From output 90 below, River/dam/lake/ponds/stream/canal/irirgation
#channel is the source of water with highest number which is 101.
#So we can say that, it is the most common source of drinking water for people livi
```

Out[90]:
```
River/dam/lake/ponds/stream/canal/irirgation channel    101
Public tap/standpipe                                     89
Open public well                                         75
Spring                                                   37
Neighbour tap                                            37
Protected public well                                    36
Piped to yard/plot                                        6
Piped into dwelling                                       5
Tanker truck                                              4
Neighbour open well                                       4
Neighbour borehole                                        3
Cart with small tank                                      2
Protected well in dwelling                                1
Bottled water                                             1
Name: drink_water_source, dtype: int64
```

Question 5. In which region people travel long distance to reach health facility?

In [91]:
```
# your code
pd.crosstab(data.region, data.distance_health)


#From output 91 below the highest distance is 36.
#A region which falls in that category is Tanga
```

Out[91]:

distance_health	0	1	2	3	4	5	6	7	8	9	10	11	12	13	15	16	19	24	36
region																			
Arusha	4	10	7	9	8	2	0	0	2	0	0	0	0	0	0	0	0	0	0
Dodoma	40	50	28	35	21	3	0	0	6	6	6	17	0	1	0	0	0	0	0
Iringa	1	3	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Kagera	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Kigoma	5	2	5	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Kilimanjaro	10	7	5	3	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Lindi	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Mbeya	3	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Morogoro	0	3	1	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0
Mtwara	1	0	1	2	2	0	0	0	0	0	2	0	2	0	6	0	0	0	0
Mwanza	4	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Pwani	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ruvuma	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
Singida	1	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Tabora	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
Tanga	0	2	13	1	6	0	1	1	9	0	1	1	3	0	0	1	0	1	2
Zanzibar North	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Question 6.Find number of people who do not treat their drinking water in each region

In [99]:
```
# your code
pd.crosstab(data.region, data.treat_water)


#From the output 99 below the column with No represent number of people who do not
```

Out[99]:

region	No	Yes
Arusha	29	13
Dodoma	168	45
Iringa	3	2
Kagera	1	1
Kigoma	8	8
Kilimanjaro	12	15
Lindi	1	0
Mbeya	5	2
Morogoro	4	3
Mtwara	15	1
Mwanza	4	4
Pwani	3	0
Ruvuma	1	1
Singida	3	1
Tabora	2	0
Tanga	24	18
Zanzibar North	3	1

Let us convert categorical values into numeric label

In [108]:
```
data.electricity.value_counts()
```

Out[108]:
```
No     377
Yes     22
9        2
Name: electricity, dtype: int64
```

In [111]:
```
data['electricity'] = data['electricity'].map({'No': 0,'Yes': 1})
```

To find percent of people with electricity in both urban and rural area we use:

In [112]:
```
data.groupby('residence_type')[['electricity']].mean()
```

Out[112]:

residence_type	electricity
Rural	NaN
Urban	NaN

This immediately gives us some insight: overall, one of every five family ($0.212 * 5 = 1$) in urban area has electricity, while it is only one in 40 families in rural area.

Question 7.Find distribution of family with refrigerator in urban and rural area.

In [102]:
```
# your code
pd.crosstab(data.residence_type, data.refrigerator)

#The distribution of family with refrigerator in urban and rural area is shown in
```

Out[102]:

residence_type	refrigerator No	Yes
Rural	333	0
Urban	65	3

Exploratory Data Analysis

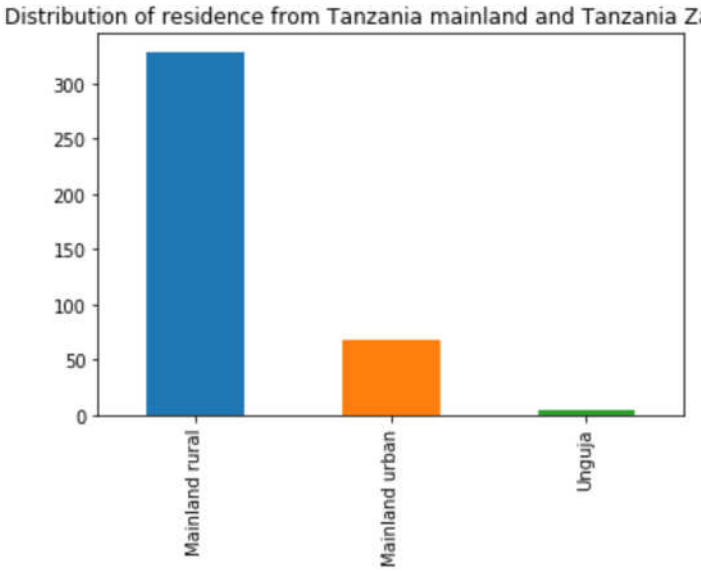
Data Exploration is the key to getting insights from data. Practitioners say a good data exploration strategy can solve even complicated problems in few hours.

Let look our data graphically

For example: Distribution of residence from Tanzania mainland and Tanzania Zanzibar

Note: use *.value_counts()* to get unique values of categorical column.

In [189]:
```
data.location.value_counts().plot(kind='bar')
plt.title('Distribution of residence from Tanzania mainland and Tanzania Zanzibar')
```



In [118]:
```
tz_rural = len(data[data.location=='Mainland rural'])
tz_urban = len(data[data.location=='Mainland urban'])
tz_ugunja = len(data[data.location=='Unguja'])
total = data.shape[0]
print("Number of people from Tanzania-rural: %d " %tz_rural)
print("Number of people from Tanzania-urban: %d " %tz_urban)
print("Number of people from Tanzania-Zanzibar: %d " %tz_ugunja)
```

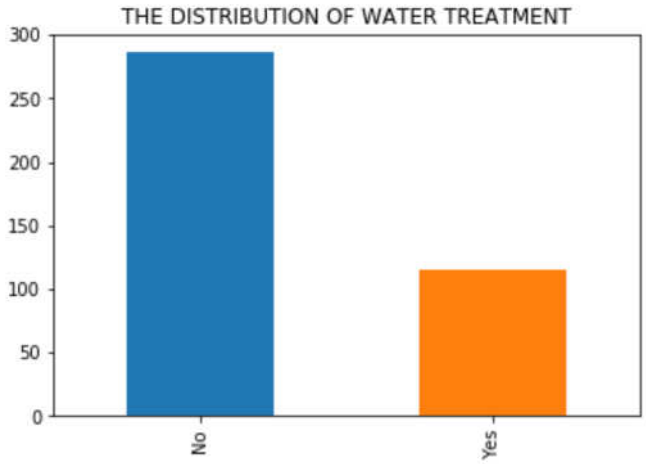
```
Number of people from Tanzania-rural: 329
Number of people from Tanzania-urban: 68
Number of people from Tanzania-Zanzibar: 4
```

Question 8 & 9. Plot the distribution of water treatment. What percent of people do not treat their drinking water?

In [143]:
```
# your code here
data.treat_water.value_counts().plot(kind='bar')
plt.title('THE DISTRIBUTION OF WATER TREATMENT')

#BELOW IS THE PLOT FOR THE DISTRIBUTION OF WATER TREATMENT
```

Out[143]:
```
Text(0.5,1,'THE DISTRIBUTION OF WATER TREATMENT')
```



In [139]:
```
data.treat_water.value_counts('percentage')*100

# The output 139 below shows the percent of people do not(NO) and do(YES) treat the
#Thereby we get that the percent of people do not treat their drinking water is 71
```

Out[139]:
```
No     71.321696
Yes    28.678304
Name: treat_water, dtype: float64
```

Let break down a plot by some categories using searborn FacetGrid.

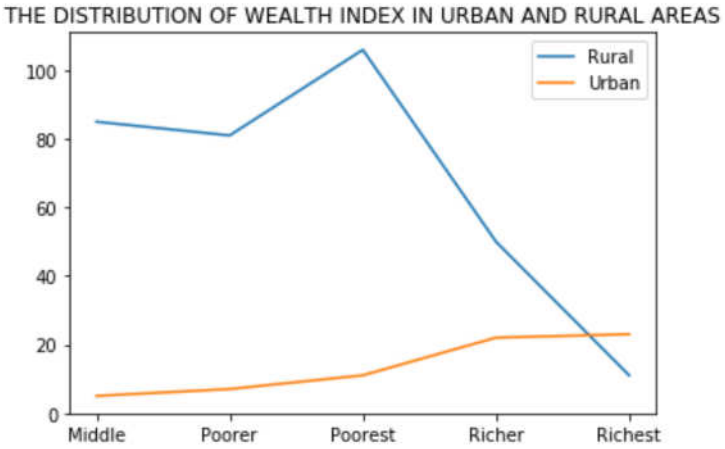
In [ ]:
```
g = sns.FacetGrid(data, col='residence_type')
g.map(sns.distplot, "household_number")
```

Question 10. Plot the distribution of wealth index in urban and rural area.

In [257]:
```
pd.crosstab(data.wealth_index, data.residence_type)
plt.plot(pd.crosstab(data.wealth_index, data.residence_type))
plt.legend(pd.crosstab(data.wealth_index, data.residence_type))
plt.title('THE DISTRIBUTION OF WEALTH INDEX IN URBAN AND RURAL AREAS')


#THE GRAPH BELOW SHOWS THE DISTRIBUTION OF WEALTH INDEX IN URBAN AND RURAL AREA
```

Out[257]: Text(0.5,1,'THE DISTRIBUTION OF WEALTH INDEX IN URBAN AND RURAL AREAS')



In [ ]: