D211_data_cleaning

April 21, 2024

1 D211 Data Cleaning

1.1 Darian Gurrola

1.2 External CSV Data Cleaning

[3]:
```python
#Import pandas to create a dataframe and import the external
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
```

[4]:
```python
# Import the initial csv file and assign to df_computer
df_initial = pd.read_csv('ACSST1Y2022.S2801-Data.csv', header=1)
```

[5]:
```python
print(df_initial.head())
```

```
     Geography Geographic Area Name  Estimate!!Total!!Total households  \
0  0400000US01              Alabama                            2016448
1  0400000US02               Alaska                             274574
2  0400000US04              Arizona                            2850377
3  0400000US05             Arkansas                            1216207
4  0400000US06           California                           13550586

   Margin of Error!!Total!!Total households  \
0                                     11475
1                                      3261
2                                     11519
3                                      8435
4                                     19485

   Estimate!!Total!!Total households!!TYPES OF COMPUTER!!Has one or more types
of computing devices:  \
0                                            1896102
1                                             267088
2                                            2747303
3                                            1143531
4                                           13141047
```

```
   Margin of Error!!Total!!Total households!!TYPES OF COMPUTER!!Has one or more
types of computing devices:  \
0                                              12178
1                                               3710
2                                              12297
3                                              10193
4                                              22219

   Estimate!!Total!!Total households!!TYPES OF COMPUTER!!Has one or more types
of computing devices:!!Desktop or laptop  \
0                                            1436344
1                                             226777
2                                            2378615
3                                             869727
4                                           11424134

   Margin of Error!!Total!!Total households!!TYPES OF COMPUTER!!Has one or more
types of computing devices:!!Desktop or laptop  \
0                                              12636
1                                               4304
2                                              15340
3                                              11023
4                                              28123

   Estimate!!Total!!Total households!!TYPES OF COMPUTER!!Has one or more types
of computing devices:!!Desktop or laptop!!Desktop or laptop with no other type
of computing device  \
0                                              44882
1                                               6538
2                                              73384
3                                              30002
4                                             268049

   Margin of Error!!Total!!Total households!!TYPES OF COMPUTER!!Has one or more
types of computing devices:!!Desktop or laptop!!Desktop or laptop with no other
type of computing device  \
0                                               3476
1                                               1334
2                                               4007
3                                               3178
4                                               7495

   …  \
0  …
1  …
2  …
3  …
```

```
4  …

   Margin of Error!!Percent!!Total households!!HOUSEHOLD INCOME IN THE PAST 12
MONTHS (IN 2022 INFLATION-ADJUSTED DOLLARS)!!$20,000 to $74,999:!!Without an
Internet subscription  \
0                                                0.6
1                                                1.5
2                                                0.6
3                                                0.8
4                                                0.2

   Estimate!!Percent!!Total households!!HOUSEHOLD INCOME IN THE PAST 12 MONTHS
(IN 2022 INFLATION-ADJUSTED DOLLARS)!!$75,000 or more:  \
0                                                (X)
1                                                (X)
2                                                (X)
3                                                (X)
4                                                (X)

   Margin of Error!!Percent!!Total households!!HOUSEHOLD INCOME IN THE PAST 12
MONTHS (IN 2022 INFLATION-ADJUSTED DOLLARS)!!$75,000 or more:  \
0                                                (X)
1                                                (X)
2                                                (X)
3                                                (X)
4                                                (X)

   Estimate!!Percent!!Total households!!HOUSEHOLD INCOME IN THE PAST 12 MONTHS
(IN 2022 INFLATION-ADJUSTED DOLLARS)!!$75,000 or more:!!With dial-up Internet
subscription alone  \
0                                                0.2
1                                                0.0
2                                                0.0
3                                                0.1
4                                                0.1

   Margin of Error!!Percent!!Total households!!HOUSEHOLD INCOME IN THE PAST 12
MONTHS (IN 2022 INFLATION-ADJUSTED DOLLARS)!!$75,000 or more:!!With dial-up
Internet subscription alone  \
0                                                0.1
1                                                0.1
2                                                0.1
3                                                0.1
4                                                0.1

   Estimate!!Percent!!Total households!!HOUSEHOLD INCOME IN THE PAST 12 MONTHS
(IN 2022 INFLATION-ADJUSTED DOLLARS)!!$75,000 or more:!!With a broadband
Internet subscription  \
```

```
0                                               95.4
1                                               96.3
2                                               96.3
3                                               95.6
4                                               97.3

   Margin of Error!!Percent!!Total households!!HOUSEHOLD INCOME IN THE PAST 12
MONTHS (IN 2022 INFLATION-ADJUSTED DOLLARS)!!$75,000 or more:!!With a broadband
Internet subscription  \
0                                                0.4
1                                                0.5
2                                                0.3
3                                                0.6
4                                                0.1

   Estimate!!Percent!!Total households!!HOUSEHOLD INCOME IN THE PAST 12 MONTHS
(IN 2022 INFLATION-ADJUSTED DOLLARS)!!$75,000 or more:!!Without an Internet
subscription  \
0                                                4.5
1                                                3.6
2                                                3.7
3                                                4.3
4                                                2.6

   Margin of Error!!Percent!!Total households!!HOUSEHOLD INCOME IN THE PAST 12
MONTHS (IN 2022 INFLATION-ADJUSTED DOLLARS)!!$75,000 or more:!!Without an
Internet subscription  \
0                                                0.4
1                                                0.5
2                                                0.3
3                                                0.6
4                                                0.1

   Unnamed: 126
0           NaN
1           NaN
2           NaN
3           NaN
4           NaN

[5 rows x 127 columns]
```

[6]:
```
#Select the columns that will be used in the analysis and assign to df_computer
```

```
df_computer = df_initial[["Geographic Area Name", "Estimate!!Total!!Total␣
↪households", "Estimate!!Total!!Total households!!TYPE OF INTERNET␣
↪SUBSCRIPTIONS!!With an Internet subscription:!!Broadband of any type!!
↪Cellular data plan", "Estimate!!Total!!Total households!!TYPE OF INTERNET␣
↪SUBSCRIPTIONS!!With an Internet subscription:!!Broadband of any type!!
↪Broadband such as cable, fiber optic or DSL", "Estimate!!Percent!!Total␣
↪households!!TYPES OF COMPUTER!!Has one or more types of computing devices:!!
↪Tablet or other portable wireless computer", "Estimate!!Percent!!Total␣
↪households!!TYPE OF INTERNET SUBSCRIPTIONS!!With an Internet subscription:!!
↪Broadband of any type!!Broadband such as cable, fiber optic or DSL",␣
↪"Estimate!!Percent!!Total households!!HOUSEHOLD INCOME IN THE PAST 12 MONTHS␣
↪(IN 2022 INFLATION-ADJUSTED DOLLARS)!!Less than $20,000:!!With a broadband␣
↪Internet subscription", "Estimate!!Percent!!Total households!!HOUSEHOLD␣
↪INCOME IN THE PAST 12 MONTHS (IN 2022 INFLATION-ADJUSTED DOLLARS)!!$20,000␣
↪to $74,999:!!With a broadband Internet subscription", "Estimate!!Percent!!
↪Total households!!HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2022␣
↪INFLATION-ADJUSTED DOLLARS)!!$75,000 or more:!!With a broadband Internet␣
↪subscription"]]
```

[7]:
```
print(df_computer.head())
```

```
  Geographic Area Name  Estimate!!Total!!Total households  \
0              Alabama                            2016448
1               Alaska                             274574
2              Arizona                            2850377
3             Arkansas                            1216207
4           California                           13550586

   Estimate!!Total!!Total households!!TYPE OF INTERNET SUBSCRIPTIONS!!With an
Internet subscription:!!Broadband of any type!!Cellular data plan  \
0                                            1639736
1                                             239052
2                                            2395194
3                                             988480
4                                           12035392

   Estimate!!Total!!Total households!!TYPE OF INTERNET SUBSCRIPTIONS!!With an
Internet subscription:!!Broadband of any type!!Broadband such as cable, fiber
optic or DSL  \
0                                            1377556
1                                             187858
2                                            2169440
3                                             815531
4                                           10729302

   Estimate!!Percent!!Total households!!TYPES OF COMPUTER!!Has one or more types
of computing devices:!!Tablet or other portable wireless computer  \
```

```
0                                                57.5
1                                                63.9
2                                                64.6
3                                                55.3
4                                                67.4

   Estimate!!Percent!!Total households!!TYPE OF INTERNET SUBSCRIPTIONS!!With an
Internet subscription:!!Broadband of any type!!Broadband such as cable, fiber
optic or DSL  \
0                                                68.3
1                                                68.4
2                                                76.1
3                                                67.1
4                                                79.2

   Estimate!!Percent!!Total households!!HOUSEHOLD INCOME IN THE PAST 12 MONTHS
(IN 2022 INFLATION-ADJUSTED DOLLARS)!!Less than $20,000:!!With a broadband
Internet subscription  \
0                                                72.2
1                                                75.7
2                                                75.2
3                                                69.7
4                                                79.7

   Estimate!!Percent!!Total households!!HOUSEHOLD INCOME IN THE PAST 12 MONTHS
(IN 2022 INFLATION-ADJUSTED DOLLARS)!!$20,000 to $74,999:!!With a broadband
Internet subscription  \
0                                                86.1
1                                                87.8
2                                                89.2
3                                                86.5
4                                                90.8

   Estimate!!Percent!!Total households!!HOUSEHOLD INCOME IN THE PAST 12 MONTHS
(IN 2022 INFLATION-ADJUSTED DOLLARS)!!$75,000 or more:!!With a broadband
Internet subscription
0                                                95.4
1                                                96.3
2                                                96.3
3                                                95.6
4                                                97.3
```

1.3 Check for missing values

[9]:
```python
#detect missing values in each variable of df_churn
df_computer.isnull().sum()
```

[9]:
```
Geographic Area Name
0
Estimate!!Total!!Total households
0
Estimate!!Total!!Total households!!TYPE OF INTERNET SUBSCRIPTIONS!!With an
Internet subscription:!!Broadband of any type!!Cellular data plan
0
Estimate!!Total!!Total households!!TYPE OF INTERNET SUBSCRIPTIONS!!With an
Internet subscription:!!Broadband of any type!!Broadband such as cable, fiber
optic or DSL            0
Estimate!!Percent!!Total households!!TYPES OF COMPUTER!!Has one or more types of
computing devices:!!Tablet or other portable wireless computer
0
Estimate!!Percent!!Total households!!TYPE OF INTERNET SUBSCRIPTIONS!!With an
Internet subscription:!!Broadband of any type!!Broadband such as cable, fiber
optic or DSL          0
Estimate!!Percent!!Total households!!HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN
2022 INFLATION-ADJUSTED DOLLARS)!!Less than $20,000:!!With a broadband Internet
subscription     0
Estimate!!Percent!!Total households!!HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN
2022 INFLATION-ADJUSTED DOLLARS)!!$20,000 to $74,999:!!With a broadband Internet
subscription    0
Estimate!!Percent!!Total households!!HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN
2022 INFLATION-ADJUSTED DOLLARS)!!$75,000 or more:!!With a broadband Internet
subscription       0
dtype: int64
```

1.4 Rename columns

[12]:
```
```

```
df_computer = df_computer.rename(columns = {"Geographic Area Name":"state",␣
↪"Estimate!!Total!!Total households":"total households", "Estimate!!Total!!
↪Total households!!TYPE OF INTERNET SUBSCRIPTIONS!!With an Internet␣
↪subscription:!!Broadband of any type!!Cellular data plan":"cellular data␣
↪plan (total)", "Estimate!!Total!!Total households!!TYPE OF INTERNET␣
↪SUBSCRIPTIONS!!With an Internet subscription:!!Broadband of any type!!
↪Broadband such as cable, fiber optic or DSL":"broadband such as cable, fiber␣
↪optic, or dsl (total)", "Estimate!!Percent!!Total households!!TYPES OF␣
↪COMPUTER!!Has one or more types of computing devices:!!Tablet or other␣
↪portable wireless computer":"has one or more tablet (percent)", "Estimate!!
↪Percent!!Total households!!TYPE OF INTERNET SUBSCRIPTIONS!!With an Internet␣
↪subscription:!!Broadband of any type!!Broadband such as cable, fiber optic␣
↪or DSL":"broadband such as cable, fiber optic, or dsl (percent)", "Estimate!!
↪Percent!!Total households!!HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2022␣
↪INFLATION-ADJUSTED DOLLARS)!!Less than $20,000:!!With a broadband Internet␣
↪subscription":"less than $20,000 with broadband internet (percent)",␣
↪"Estimate!!Percent!!Total households!!HOUSEHOLD INCOME IN THE PAST 12 MONTHS␣
↪(IN 2022 INFLATION-ADJUSTED DOLLARS)!!$20,000 to $74,999:!!With a broadband␣
↪Internet subscription":"$20,000 to $74,999 with broadband internet␣
↪(percent)", "Estimate!!Percent!!Total households!!HOUSEHOLD INCOME IN THE␣
↪PAST 12 MONTHS (IN 2022 INFLATION-ADJUSTED DOLLARS)!!$75,000 or more:!!With␣
↪a broadband Internet subscription":"$75,000 or more with broadband internet␣
↪(percent)"})
```

[13]:
```
print(df_computer.head())
```

```
        state  total households  cellular data plan (total)  \
0     Alabama           2016448                     1639736
1      Alaska            274574                      239052
2     Arizona           2850377                     2395194
3    Arkansas           1216207                      988480
4  California          13550586                    12035392

   broadband such as cable, fiber optic, or dsl (total)  \
0                                            1377556
1                                             187858
2                                            2169440
3                                             815531
4                                           10729302

   has one or more tablet (percent)  \
0                              57.5
1                              63.9
2                              64.6
3                              55.3
4                              67.4
```

```
   broadband such as cable, fiber optic, or dsl (percent)  \
0                                                 68.3
1                                                 68.4
2                                                 76.1
3                                                 67.1
4                                                 79.2

   less than $20,000 with broadband internet (percent)  \
0                                                 72.2
1                                                 75.7
2                                                 75.2
3                                                 69.7
4                                                 79.7

   $20,000 to $74,999 with broadband internet (percent)  \
0                                                 86.1
1                                                 87.8
2                                                 89.2
3                                                 86.5
4                                                 90.8

   $75,000 or more with broadband internet (percent)
0                                                 95.4
1                                                 96.3
2                                                 96.3
3                                                 95.6
4                                                 97.3
```

[14]:

```
#Rename values in State column

#Find unique values of variable
print(df_computer["state"].unique())
```

```
['Alabama' 'Alaska' 'Arizona' 'Arkansas' 'California' 'Colorado'
 'Connecticut' 'Delaware' 'District of Columbia' 'Florida' 'Georgia'
 'Hawaii' 'Idaho' 'Illinois' 'Indiana' 'Iowa' 'Kansas' 'Kentucky'
 'Louisiana' 'Maine' 'Maryland' 'Massachusetts' 'Michigan' 'Minnesota'
 'Mississippi' 'Missouri' 'Montana' 'Nebraska' 'Nevada' 'New Hampshire'
 'New Jersey' 'New Mexico' 'New York' 'North Carolina' 'North Dakota'
 'Ohio' 'Oklahoma' 'Oregon' 'Pennsylvania' 'Rhode Island' 'South Carolina'
 'South Dakota' 'Tennessee' 'Texas' 'Utah' 'Vermont' 'Virginia'
 'Washington' 'West Virginia' 'Wisconsin' 'Wyoming' 'Puerto Rico']
```

[15]:

```
#Create dictionary to store numeric values for variable
dict_state = {"state":
                    {"Alabama":"AL",
                     "Alaska":"AK",
```

```
                    "Arizona":"AZ",
                    "Arkansas":"AR",
                    "California":"CA",
                    "Colorado":"CO",
                    "Connecticut":"CT",
                    "Delaware":"DE",
                    "District of Columbia":"DC",
                    "Florida":"FL",
                    "Georgia":"GA",
                    "Hawaii":"HI",
                    "Idaho":"ID",
                    "Illinois":"IL",
                    "Indiana":"IN",
                    "Iowa":"IA",
                    "Kansas":"KS",
                    "Kentucky":"KY",
                    "Louisiana":"LA",
                    "Maine":"ME",
                    "Maryland":"MD",
                    "Massachusetts":"MA",
                    "Michigan":"MI",
                    "Minnesota":"MN",
                    "Mississippi":"MS",
                    "Missouri":"MO",
                    "Montana":"MT",
                    "Nebraska":"NE",
                    "Nevada":"NV",
                    "New Hampshire":"NH",
                    "New Jersey":"NJ",
                    "New Mexico":"NM",
                    "New York":"NY",
                    "North Carolina":"NC",
                    "North Dakota":"ND",
                    "Ohio":"OH",
                    "Oklahoma":"OK",
                    "Oregon":"OR",
                    "Pennsylvania":"PA",
                    "Rhode Island":"RI",
                    "South Carolina":"SC",
                    "South Dakota":"SD",
                    "Tennessee":"TN",
                    "Texas":"TX",
                    "Utah":"UT",
                    "Vermont":"VT",
                    "Virginia":"VA",
                    "Washington":"WA",
                    "West Virginia":"WV",
```

```
                      "Wisconsin":"WI",
                      "Wyoming":"WY",
                      "Puerto Rico":"PR"
                     }
                 }

#Replace categorical values with numeric values from dictionary
df_computer.replace(dict_state, inplace=True)

#Confirm categorical values have been replaced
print(df_computer["state"].unique())
```

```
['AL' 'AK' 'AZ' 'AR' 'CA' 'CO' 'CT' 'DE' 'DC' 'FL' 'GA' 'HI' 'ID' 'IL'
 'IN' 'IA' 'KS' 'KY' 'LA' 'ME' 'MD' 'MA' 'MI' 'MN' 'MS' 'MO' 'MT' 'NE'
 'NV' 'NH' 'NJ' 'NM' 'NY' 'NC' 'ND' 'OH' 'OK' 'OR' 'PA' 'RI' 'SC' 'SD'
 'TN' 'TX' 'UT' 'VT' 'VA' 'WA' 'WV' 'WI' 'WY' 'PR']
```

[16]:
```
df_computer.to_csv('computer_clean.csv', index=False)
```

1.5 Location Table Cleaning

[18]:
```
# Import the location csv file and assign to df_location
df_location = pd.read_csv('location.csv')
```

[19]:
```
print(len(df_location))
```

```
8583
```

[20]:
```
# Count rows in location table
print(df_location.head())
```

```
   location_id    zip           city state    county
0         5599  62419        Calhoun    IL  Richland
1         2737  32266  Neptune Beach    FL     Duval
2         1297  16424     Linesville    PA  Crawford
3         5181  58428         Dawson    ND    Kidder
4           30    952    Sabana Seca    PR  Toa Baja
```

[21]:
```
#Detect duplicate rows in df_location
print(df_location.duplicated().value_counts())
```

```
False    8583
Name: count, dtype: int64
```

[22]:
```
#detect missing values in each column of df_location
df_location.isnull().sum()
```

[22]:
```
location_id    0
zip            0
city           0
state          0
county         0
dtype: int64
```

1.6 Job Table Cleaning

[24]:
```python
# Import the job csv file and assign to df_job
df_job = pd.read_csv('job.csv')
```

[25]:
```python
print(df_job.head())
```

```
   job_id                          job_title
0       1                 Academic librarian
1       2              Accommodation manager
2       3              Accountant- chartered
3       4    Accountant- chartered certified
4       5  Accountant- chartered management
```

[92]:
```python
print(len(df_job))
```

```
639
```

[94]:
```python
#Detect duplicate rows in df_churn
print(df_job.duplicated().value_counts())
```

```
False    639
Name: count, dtype: int64
```

[96]:
```python
#detect missing values in each column of df_location
df_job.isnull().sum()
```

[96]:
```
job_id       0
job_title    0
dtype: int64
```

1.7 Customer Table Cleaning

[29]:
```python
# Import the customer csv file and assign to df_job
df_customer = pd.read_csv('customer.csv')
```

[30]:
```python
print(df_customer.head())
```

```
  customer_id       lat        lng  population  children  age    income  \
0     K409198  56.25100 -133.37571          38         0   68  28561.99
```

```
1     S120509  44.32893  -84.24080       10446          1    27  21704.77
2     K191035  45.35589 -123.24657        3735          4    50   9609.57
3      D90850  32.96687 -117.24798       13863          1    48  18925.23
4     K662701  29.38012  -95.80673       11352          0    83  40074.19

     marital churn  gender  …  email  contacts  yearly_equip_faiure  techie  \
0    Widowed    No    Male  …     10         0                    1      No
1    Married   Yes  Female  …     12         0                    1     Yes
2    Widowed    No  Female  …      9         0                    1     Yes
3    Married    No    Male  …     15         2                    0     Yes
4  Separated   Yes    Male  …     16         2                    1      No

  port_modem  tablet  job_id payment_id contract_id location_id
0        Yes     Yes     229          2           2        5599
1         No     Yes     468          1           1        2737
2        Yes      No      96          2           3        1297
3         No      No     552          4           3        5181
4        Yes      No     371          4           1          30

[5 rows x 24 columns]
```

[98]:
```python
print(len(df_customer))
```

```
10000
```

[100]:
```python
#Detect duplicate rows in df_customer
print(df_customer.duplicated().value_counts())
```

```
False    10000
Name: count, dtype: int64
```

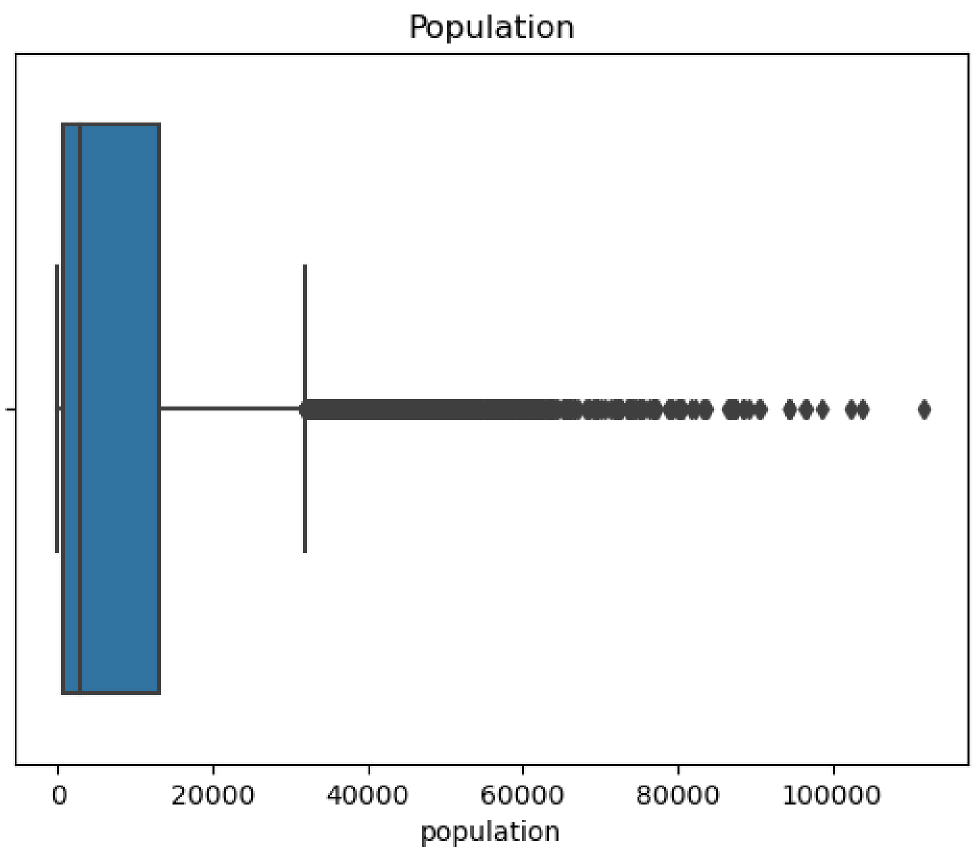
[32]:
```python
#detect missing values in each column of df_location
df_customer.isnull().sum()
```

[32]:
```
customer_id          0
lat                  0
lng                  0
population           0
children             0
age                  0
income               0
marital              0
churn                0
gender               0
tenure               0
monthly_charge       0
bandwidth_gp_year    0
outage_sec_week      0
```

```
email                    0
contacts                 0
yearly_equip_faiure      0
techie                   0
port_modem               0
tablet                   0
job_id                   0
payment_id               0
contract_id              0
location_id              0
dtype: int64
```

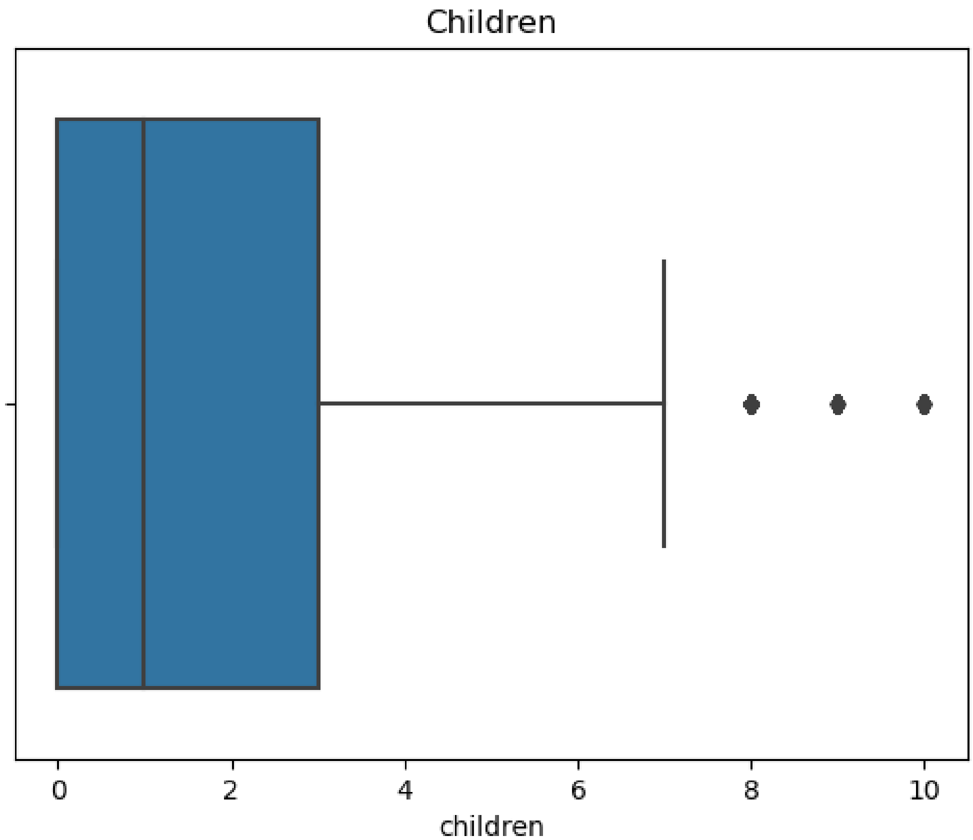
[33]:

```
#Detect outliers in Population variable
boxplot = sns.boxplot(x="population", data = df_customer).
↪set_title("Population")
```

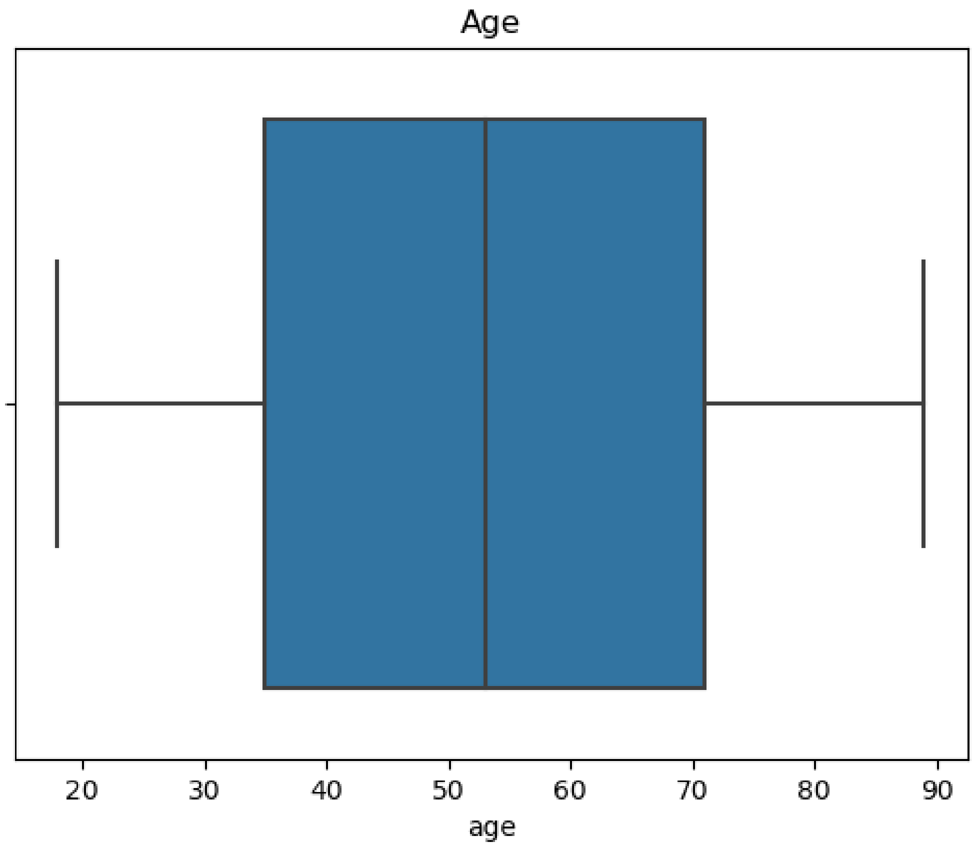


[34]:

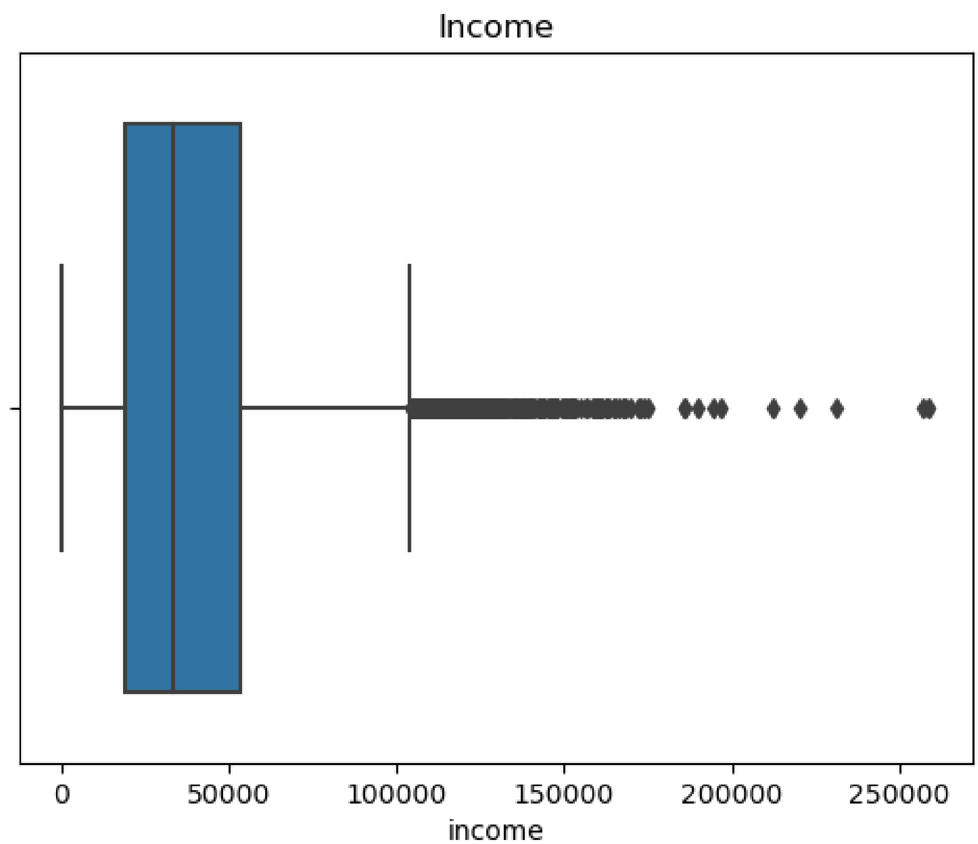
```
#Detect outliers in children variable
boxplot = sns.boxplot(x="children", data = df_customer).set_title("Children")
```



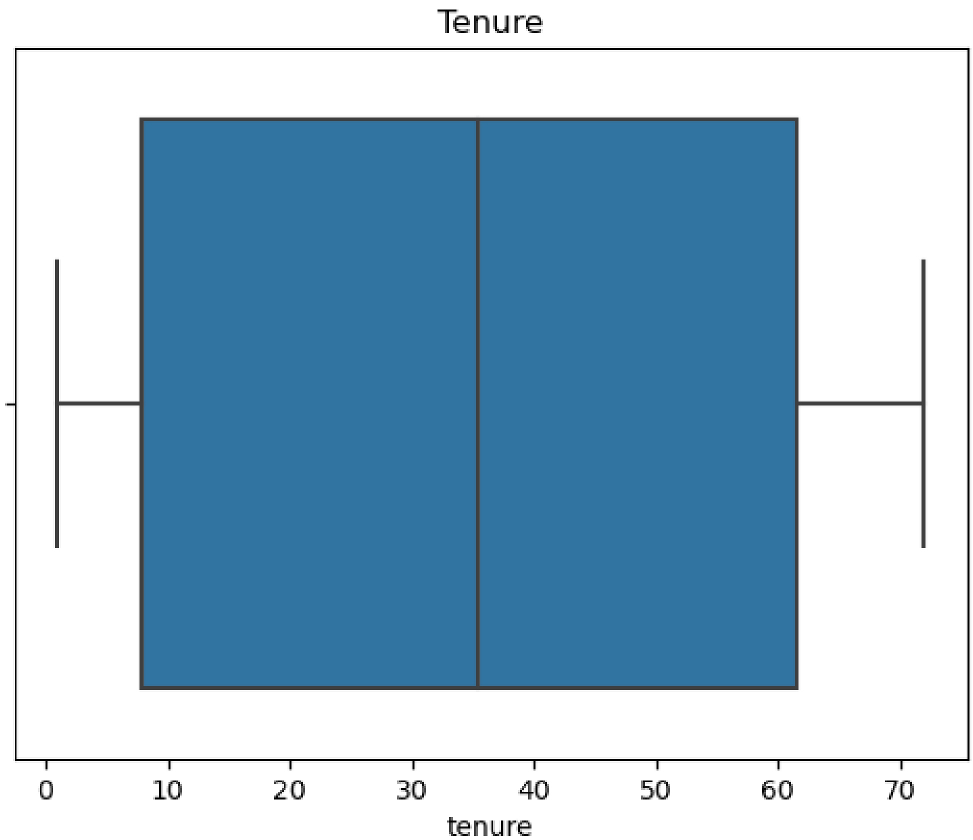
[35]:
```
#Detect outliers in age variable
boxplot = sns.boxplot(x="age", data = df_customer).set_title("Age")
```



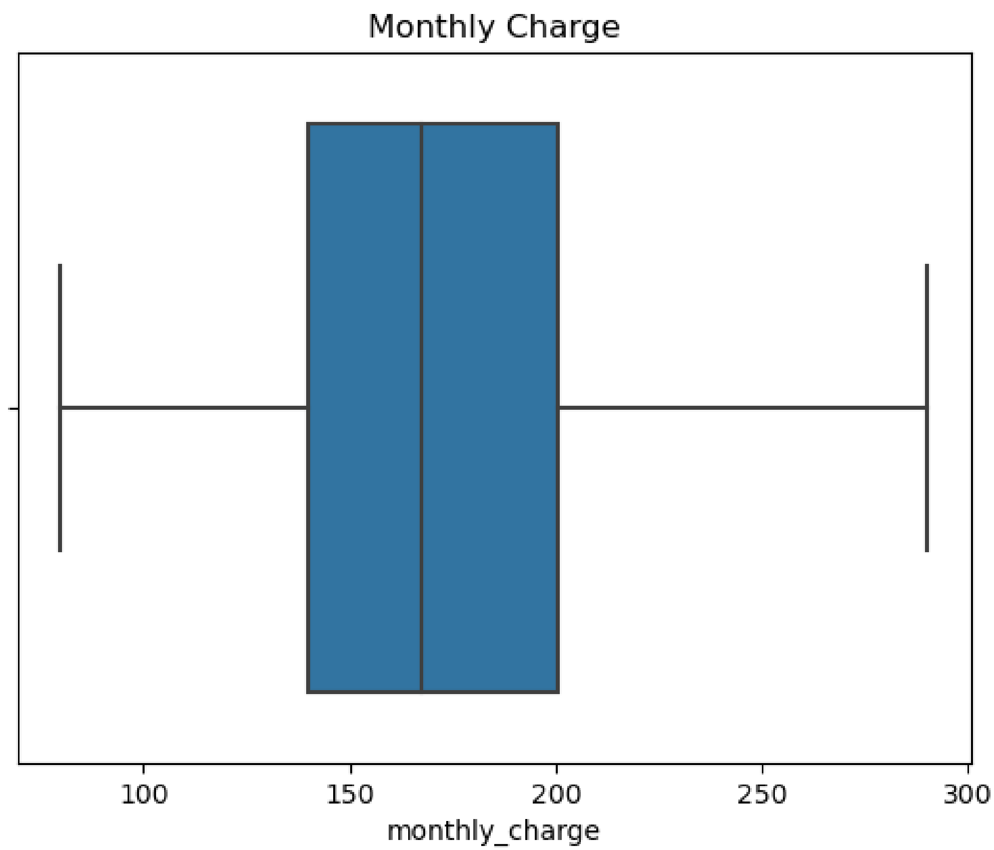
[36]:
```
#Detect outliers in income variable
boxplot = sns.boxplot(x="income", data = df_customer).set_title("Income")
```



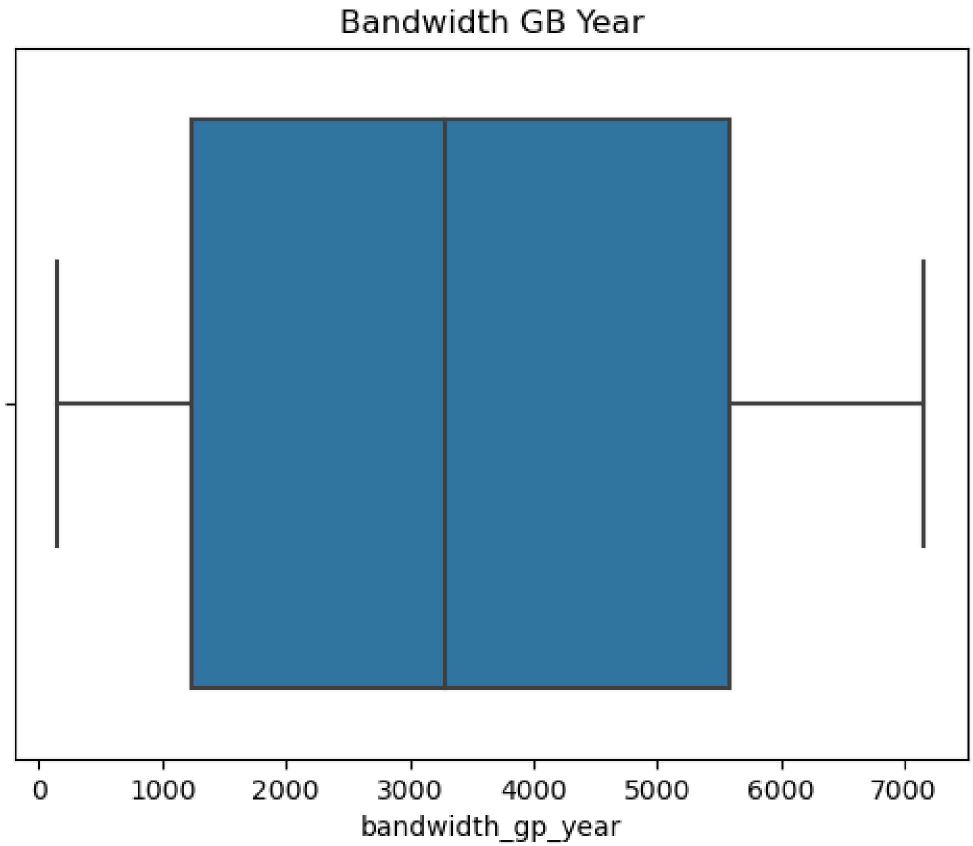
[37]:
```
#Detect outliers in tenure variable
boxplot = sns.boxplot(x="tenure", data = df_customer).set_title("Tenure")
```



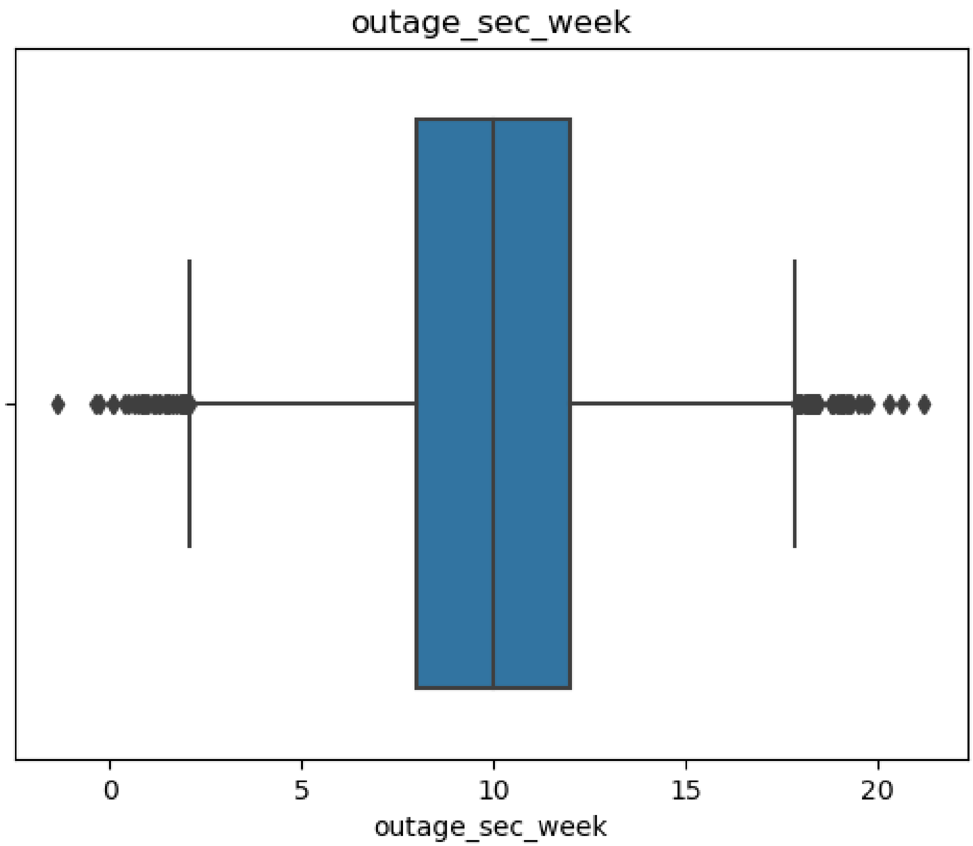
[38]:
```
#Detect outliers in monthly_charge variable
boxplot = sns.boxplot(x="monthly_charge", data = df_customer).
 ↪set_title("Monthly Charge")
```



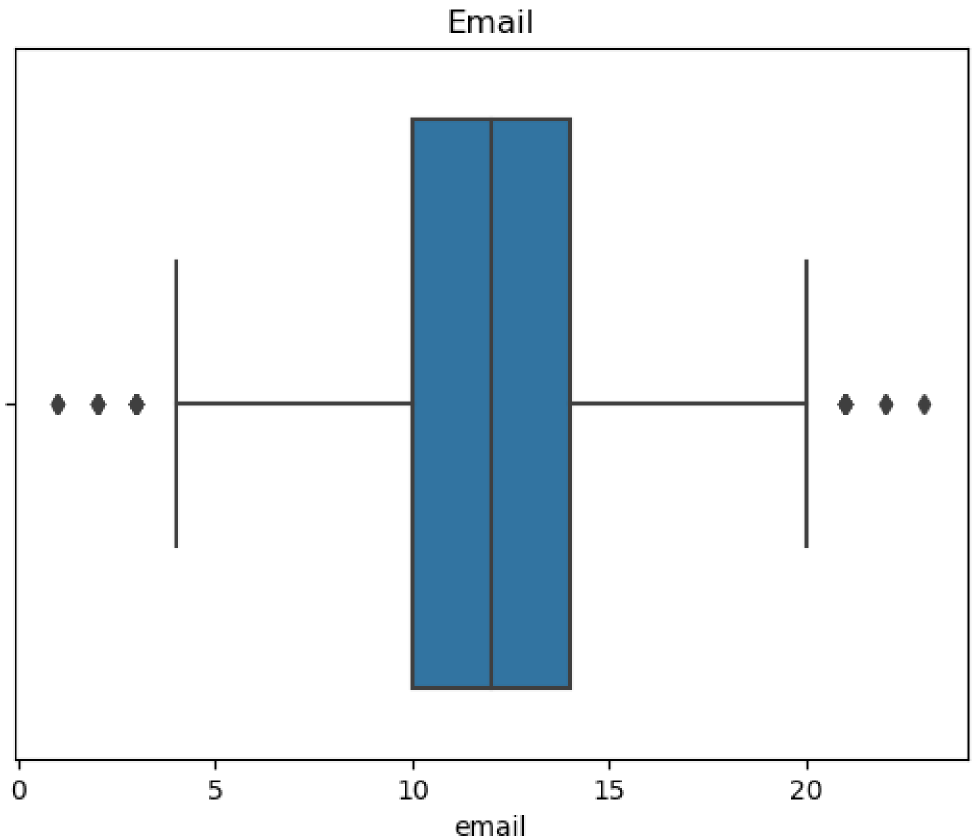
[39]:
```
#Detect outliers in bandwidth_gp_year variable
boxplot = sns.boxplot(x="bandwidth_gp_year", data = df_customer).
 ↪set_title("Bandwidth GB Year")
```



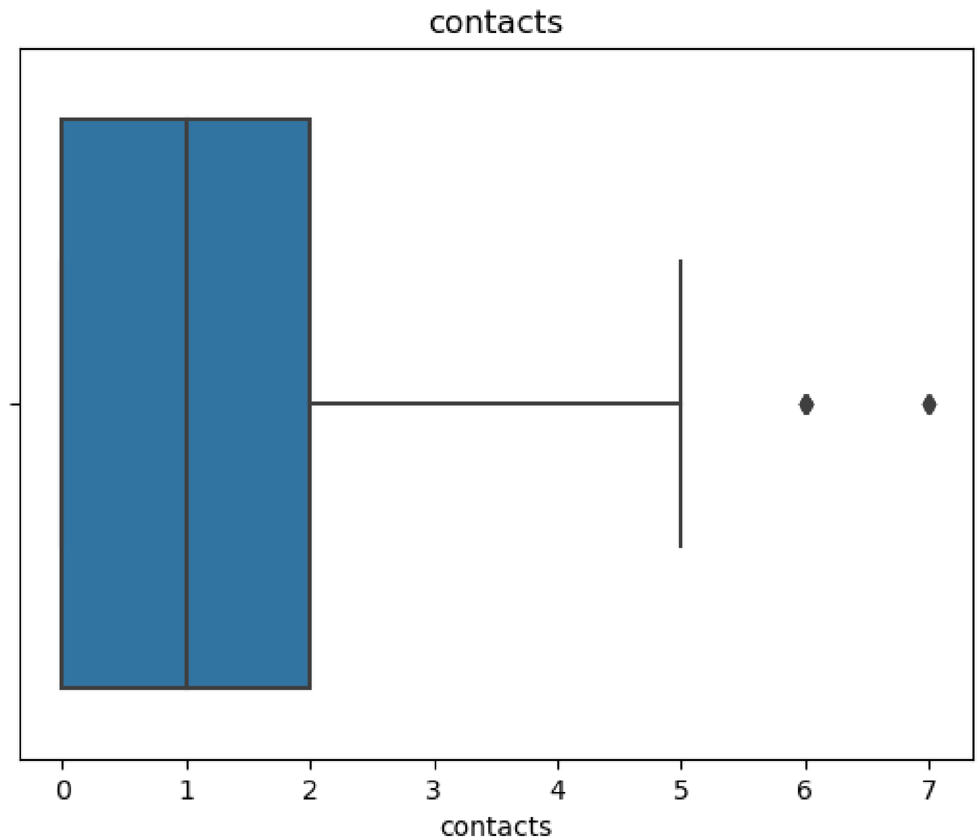
[40]:
```
#Detect outliers in outage_sec_week variable
boxplot = sns.boxplot(x="outage_sec_week", data = df_customer).
 ↪set_title("outage_sec_week")
```



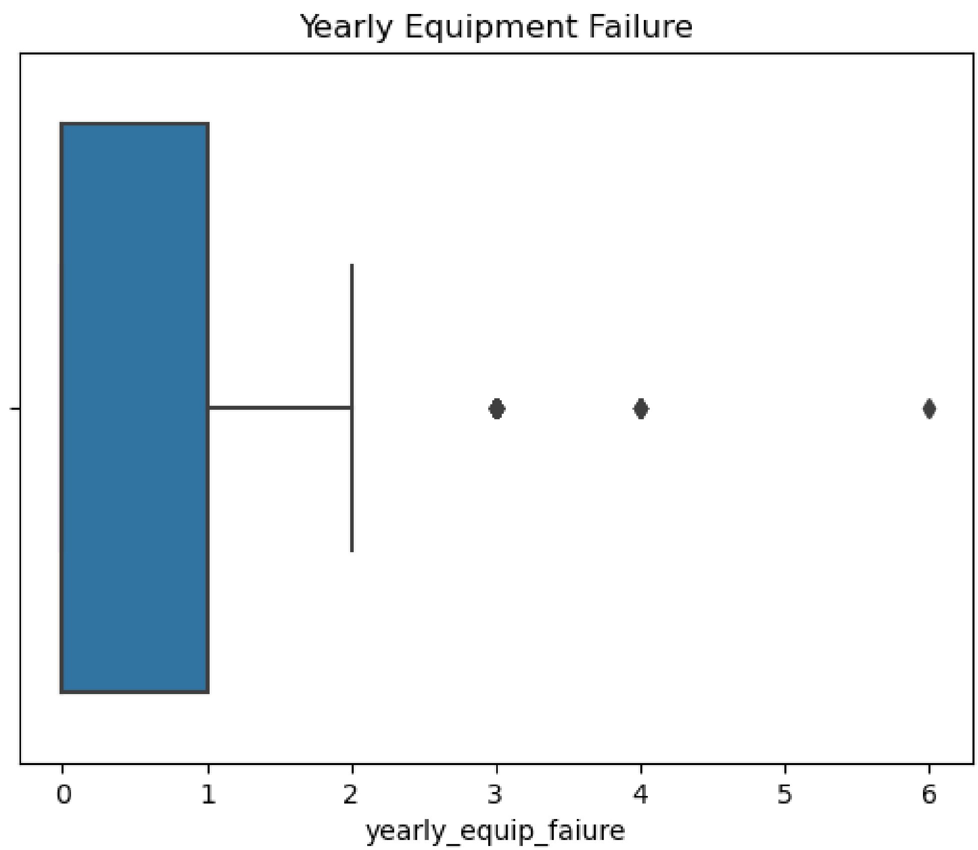
[41]:
```
#Detect outliers in email variable
boxplot = sns.boxplot(x="email", data = df_customer).set_title("Email")
```



[42]:
```
#Detect outliers in contacts variable
boxplot = sns.boxplot(x="contacts", data = df_customer).set_title("contacts")
```




```
[43]: #Detect outliers in yearly_equip_faiure variable
boxplot = sns.boxplot(x="yearly_equip_faiure", data = df_customer).
 ↪set_title("Yearly Equipment Failure")
```



[44]:
```
print(min(df_customer['bandwidth_gp_year']))
```

```
155.5067148
```

1.8 Payment Table Cleaning

[46]:
```
# Import the payment csv file and assign to df_payment
df_payment = pd.read_csv('payment.csv')
```

[47]:
```
print(df_payment.head())
```

```
   payment_id              payment_type
0           1  Bank Transfer Automatic
1           2    Credit Card Automatic
2           3         Electronic Check
3           4             Mailed Check
```

1.9 Contract Table Cleaning

[49]:
```
# Import the contract csv file and assign to df_contract
df_contract = pd.read_csv('contract.csv')
```

[50]:
```
print(df_contract.head())
```

```
   contract_id        duration
0            1  Month-to-month
1            2        One year
2            3        Two Year
```