By **Shubhendra Kumar**

**Email: kshubhendra8860@gmail.com**

# List down all the columns with missing values.

Output:

['STATE',

'TENURE\_IN\_MONTHS',

'CLOSESTSTOREDISTANCE',

'AGE',

'INCOME',

'LENGTH OF RESIDENCE',

'NUMBER OF PERSONS IN LIVING UNIT',

'NUMBER OF ADULTS IN LIVING UNIT',

'MOSAIC',

'CAPE: AGE: POP: MEDIAN AGE',

'CAPE: AGE: POP: % 0-17',

'CAPE: AGE: POP: % 18-99+',

'CAPE: AGE: POP: % 65-99+',

'CAPE: ETHNIC: POP: % WHITE ONLY',

'CAPE: ETHNIC: POP: % BLACK ONLY',

'CAPE: ETHNIC: POP: % ASIAN ONLY',

'CAPE: ETHNIC: POP: % HISPANIC',

'CAPE: DENSITY: PERSONS PER HH FOR POP IN HH',

'CAPE: HHSIZE: HH: AVERAGE HOUSEHOLD SIZE',

'CAPE: TYP: HH: % MARRIED COUPLE FAMILY',

'CAPE: CHILD: HH: % WITH PERSONS LT18',

'CAPE: CHILD: HH: % MARR COUPLE FAMW- PERSONS LT18',

'CAPE: CHILD: HH: % MARR COUPLE FAMW-O PERSONS LT18',

'CAPE: LANG: HH: % SPANISH SPEAKING',

'CAPE: EDUC: POP25+: MEDIAN EDUCATION ATTAINED',

'CAPE: HOMVAL: OOHU: MEDIAN HOME VALUE',

'CAPE: BUILT: HU: MEDIAN HOUSING UNIT AGE',

'CAPE: TENANCY: OCCHU: % OWNER OCCUPIED',

'CAPE: TENANCY: OCCHU: % RENTER OCCUPIED',

'CAPE: EDUC: ISPSA',

'CAPE: EDUC: ISPSA DECILE',

'CAPE: INC: FAMILY INC STATE DECILE',

'CAPE: INC: HH: MEDIAN FAMILY HOUSEHOLD INCOME']

# Categorize the columns based upon their data type and print, for ex: print all the numeric variables and other data types as well

## Columns with int64 data type:

Int64 Columns: ['CUSTOMERID', 'LCPCOUNT', 'FEMALE', 'HS\_DIPLOMA', 'SOME\_COLLEGE', 'BACH\_GRAD\_DEG', 'LT\_HS\_DIPLOMA', 'MARRIED', 'MNGMNT\_OFFICEADMIN', 'TECH\_PROF', 'SALES\_JOB', 'BLUE\_COLLAR', 'FARMER', 'RETIRED', 'SFDU', 'MFDU', 'HOMEOWNER', 'MAIL\_RESP\_MULTI', 'MAIL\_RESP\_SINGLE', 'METRO', 'URBAN', 'MOR BANK: UPSCALE MERCHANDISE BUYER', 'MOR BANK: MALE MERCHANDISE BUYER', 'MOR BANK: FEMALE MERCHANDISE BUYER', 'MOR BANK: CRAFTS-HOBBY MERCHANDISE BUYER', 'MOR BANK: GARDENING-FARMING BUYER', 'MOR BANK: BOOK BUYER', 'MOR BANK: COLLECT-SPECIAL FOODS BUYER', 'MOR BANK: GIFTS AND GADGETS BUYER', 'MOR BANK: GENERAL MERCHANDISE BUYER', 'MOR BANK: FAMILY AND GENERAL MAGAZINE', 'MOR BANK: FEMALE ORIENTED MAGAZINE', 'MOR BANK: MALE SPORTS MAGAZINE', 'MOR BANK: RELIGIOUS MAGAZINE', 'MOR BANK: GARDENING-FARMING MAGAZINE', 'MOR BANK: CULINARY INTERESTS MAGAZINE', 'MOR BANK: HEALTH AND FITNESS MAGAZINE', 'MOR BANK: DO-IT-YOURSELFERS', 'MOR BANK: NEWS AND FINANCIAL', 'MOR BANK: PHOTOGRAPHY', 'MOR BANK: OPPORTUNITY SEEKERS AND CE', 'MOR BANK: RELIGIOUS CONTRIBUTOR', 'MOR BANK: POLITICAL CONTRIBUTOR', 'MOR BANK: HEALTH AND INSTITUTION CONTRIBUTOR', 'MOR BANK: GENERAL CONTRIBUTOR', 'MOR BANK: MISCELLANEOUS', 'MOR BANK: ODDS AND ENDS', 'MOR BANK: DEDUPED CATEGORY HIT COUNT', 'MOR BANK: NON-DEDUPED CATEGORY HIT COUNT', 'MORTGAGE-HOME PURCHASE: HOME PURCHASE PRICE', 'FREQUENCY', 'QUANTITY', 'FREQUENCY\_2Y', 'QUANTITY\_2Y', '9-Repairs & Appraisals Spend', 'MDAYQTY\_L1Y', 'MDAYQTY\_L2Y', 'MDAYQTY\_L3Y', 'MDAYQTY\_L4Y', 'MDAYSHOPPER\_L1Y', 'MDAYSHOPPER\_L2Y', 'MDAYSHOPPER\_L3Y', 'MDAYSHOPPER\_L4Y']

## Columns with int64 data type:

Columns with float64: ['TENURE\_IN\_MONTHS', 'CLOSESTSTOREDISTANCE', 'AGE', 'INCOME', 'LENGTH OF RESIDENCE', 'NUMBER OF PERSONS IN LIVING UNIT', 'NUMBER OF ADULTS IN LIVING UNIT', 'CAPE: AGE: POP: MEDIAN AGE', 'CAPE: AGE: POP: % 0-17', 'CAPE: AGE: POP: % 18-99+', 'CAPE: AGE: POP: % 65-99+', 'CAPE: ETHNIC: POP: % WHITE ONLY', 'CAPE: ETHNIC: POP: % BLACK ONLY', 'CAPE: ETHNIC: POP: % ASIAN ONLY', 'CAPE: ETHNIC: POP: % HISPANIC', 'CAPE: DENSITY: PERSONS PER HH FOR POP IN HH', 'CAPE: TYP: HH: % MARRIED COUPLE FAMILY', 'CAPE: CHILD: HH: % WITH PERSONS LT18', 'CAPE: CHILD: HH: % MARR COUPLE FAMW- PERSONS LT18', 'CAPE: CHILD: HH: % MARR COUPLE FAMW-O PERSONS LT18', 'CAPE: LANG: HH: % SPANISH SPEAKING', 'CAPE: EDUC: POP25+: MEDIAN EDUCATION ATTAINED', 'CAPE: HOMVAL: OOHU: MEDIAN HOME VALUE', 'CAPE: BUILT: HU: MEDIAN HOUSING UNIT AGE', 'CAPE: TENANCY: OCCHU: % OWNER OCCUPIED', 'CAPE: TENANCY: OCCHU: % RENTER OCCUPIED', 'CAPE: EDUC: ISPSA', 'CAPE: EDUC: ISPSA DECILE', 'CAPE: INC: FAMILY INC STATE DECILE', 'CAPE: INC: HH: MEDIAN FAMILY HOUSEHOLD INCOME', 'TOTALSALES', 'TOTALSALES\_2Y', '1-Engagement Spend', '2-Wedding Bands Spend', '3-Fashion Diamonds Spend', '4-Fashion Jewelry Spend', '5-Close Out Spend', '6-Promotional Items Spend', '8-Marketing Premium SKUs Spend', '10-Pre Owned Spend', '11-Watches Spend', '12-Misc Merchandise Spend', '15-Store Events Spend', '16-Single Stone Jewelry Spend', 'MDAYREV\_L1Y', 'MDAYREV\_L2Y', 'MDAYREV\_L3Y', 'MDAYREV\_L4Y']

## Columns with object data type:

columns with Object: ['STATE', 'PRIVATELABELTENDERFLAG', 'MOSAIC']

# List the columns with duplicates b) Print before and after

Columns with duplicate values: {'7-Cost Only Spend', 'CAPE: HHSIZE: HH: AVERAGE HOUSEHOLD SIZE', 'CHILDREN'}

## Remove them

Shape before removing: (10000, 117)

Shape after removing: (10000, 114)

## b) Print before

CUSTOMERID STATE LCPCOUNT PRIVATELABELTENDERFLAG TENURE\_IN\_MONTHS \

0 5001 TX 1 N -9.0

1 5002 OH 0 Y 9.0

2 5003 TX 0 N 12.0

3 5004 TN 0 N -1.0

4 5005 TX 0 N 16.0

CLOSESTSTOREDISTANCE FEMALE AGE HS\_DIPLOMA SOME\_COLLEGE ... \

0 NaN 0 NaN 0 0 ...

1 8.728943 0 NaN 0 0 ...

2 NaN 0 NaN 0 0 ...

3 NaN 0 NaN 0 0 ...

4 NaN 0 NaN 0 0 ...

MDAYREV\_L3Y MDAYREV\_L4Y MDAYQTY\_L1Y MDAYQTY\_L2Y MDAYQTY\_L3Y \

0 0.0 0.0 0 0 0

1 0.0 0.0 0 0 0

2 0.0 0.0 0 0 0

3 0.0 0.0 0 0 0

4 0.0 0.0 0 0 0

MDAYQTY\_L4Y MDAYSHOPPER\_L1Y MDAYSHOPPER\_L2Y MDAYSHOPPER\_L3Y \

0 0 0 0 0

1 0 0 0 0

2 0 0 0 0

3 0 0 0 0

4 0 0 0 0

MDAYSHOPPER\_L4Y

0 0

1 0

2 0

3 0

4 0

[5 rows x 117 columns]

## Print after

CUSTOMERID STATE LCPCOUNT PRIVATELABELTENDERFLAG TENURE\_IN\_MONTHS \

0 5001 TX 1 N -9.0

1 5002 OH 0 Y 9.0

2 5003 TX 0 N 12.0

3 5004 TN 0 N -1.0

4 5005 TX 0 N 16.0

CLOSESTSTOREDISTANCE FEMALE AGE HS\_DIPLOMA SOME\_COLLEGE ... \

0 NaN 0 NaN 0 0 ...

1 8.728943 0 NaN 0 0 ...

2 NaN 0 NaN 0 0 ...

3 NaN 0 NaN 0 0 ...

4 NaN 0 NaN 0 0 ...

MDAYREV\_L3Y MDAYREV\_L4Y MDAYQTY\_L1Y MDAYQTY\_L2Y MDAYQTY\_L3Y \

0 0.0 0.0 0 0 0

1 0.0 0.0 0 0 0

2 0.0 0.0 0 0 0

3 0.0 0.0 0 0 0

4 0.0 0.0 0 0 0

MDAYQTY\_L4Y MDAYSHOPPER\_L1Y MDAYSHOPPER\_L2Y MDAYSHOPPER\_L3Y \

0 0 0 0 0

1 0 0 0 0

2 0 0 0 0

3 0 0 0 0

4 0 0 0 0

MDAYSHOPPER\_L4Y

0 0

1 0

2 0

3 0

4 0

[5 rows x 114 columns]

# List the constant columns

Constant column: ['MARRIED']

## Remove them

Orginal df Shape: (10000, 114)

Cleaned df Shape: (10000, 113)

## Print before

CUSTOMERID STATE LCPCOUNT PRIVATELABELTENDERFLAG TENURE\_IN\_MONTHS \

0 5001 TX 1 N -9.0

1 5002 OH 0 Y 9.0

2 5003 TX 0 N 12.0

3 5004 TN 0 N -1.0

4 5005 TX 0 N 16.0

CLOSESTSTOREDISTANCE FEMALE AGE HS\_DIPLOMA SOME\_COLLEGE ... \

0 NaN 0 NaN 0 0 ...

1 8.728943 0 NaN 0 0 ...

2 NaN 0 NaN 0 0 ...

3 NaN 0 NaN 0 0 ...

4 NaN 0 NaN 0 0 ...

MDAYREV\_L3Y MDAYREV\_L4Y MDAYQTY\_L1Y MDAYQTY\_L2Y MDAYQTY\_L3Y \

0 0.0 0.0 0 0 0

1 0.0 0.0 0 0 0

2 0.0 0.0 0 0 0

3 0.0 0.0 0 0 0

4 0.0 0.0 0 0 0

MDAYQTY\_L4Y MDAYSHOPPER\_L1Y MDAYSHOPPER\_L2Y MDAYSHOPPER\_L3Y \

0 0 0 0 0

1 0 0 0 0

2 0 0 0 0

3 0 0 0 0

4 0 0 0 0

MDAYSHOPPER\_L4Y

0 0

1 0

2 0

3 0

4 0

[5 rows x 114 columns]

## Print after

CUSTOMERID STATE LCPCOUNT PRIVATELABELTENDERFLAG TENURE\_IN\_MONTHS \

0 5001 TX 1 N -9.0

1 5002 OH 0 Y 9.0

2 5003 TX 0 N 12.0

3 5004 TN 0 N -1.0

4 5005 TX 0 N 16.0

CLOSESTSTOREDISTANCE FEMALE AGE HS\_DIPLOMA SOME\_COLLEGE ... \

0 NaN 0 NaN 0 0 ...

1 8.728943 0 NaN 0 0 ...

2 NaN 0 NaN 0 0 ...

3 NaN 0 NaN 0 0 ...

4 NaN 0 NaN 0 0 ...

MDAYREV\_L3Y MDAYREV\_L4Y MDAYQTY\_L1Y MDAYQTY\_L2Y MDAYQTY\_L3Y \

0 0.0 0.0 0 0 0

1 0.0 0.0 0 0 0

2 0.0 0.0 0 0 0

3 0.0 0.0 0 0 0

4 0.0 0.0 0 0 0

MDAYQTY\_L4Y MDAYSHOPPER\_L1Y MDAYSHOPPER\_L2Y MDAYSHOPPER\_L3Y \

0 0 0 0 0

1 0 0 0 0

2 0 0 0 0

3 0 0 0 0

4 0 0 0 0

MDAYSHOPPER\_L4Y

0 0

1 0

2 0

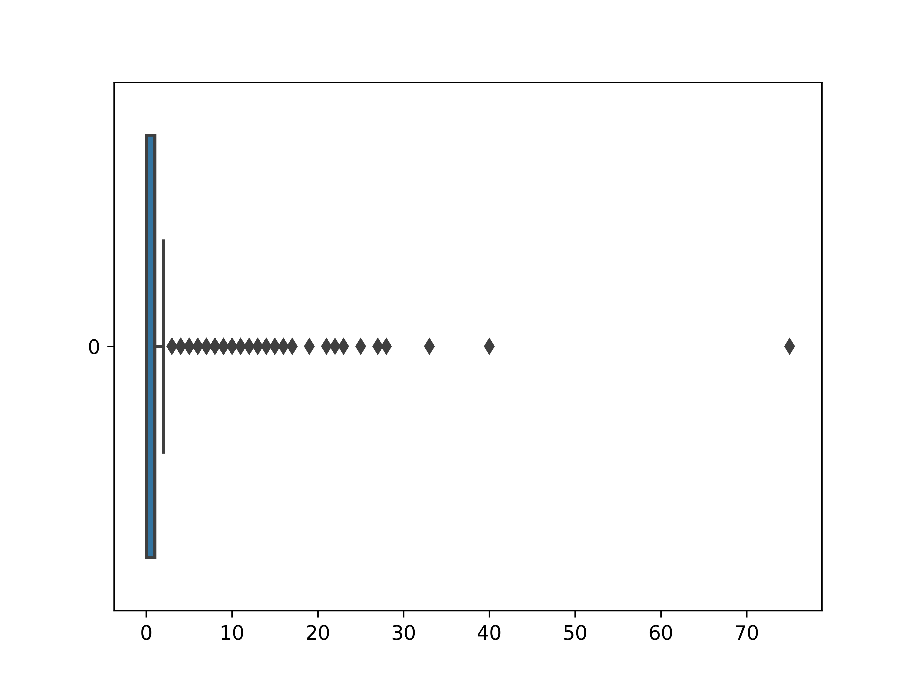
3 0

4 0

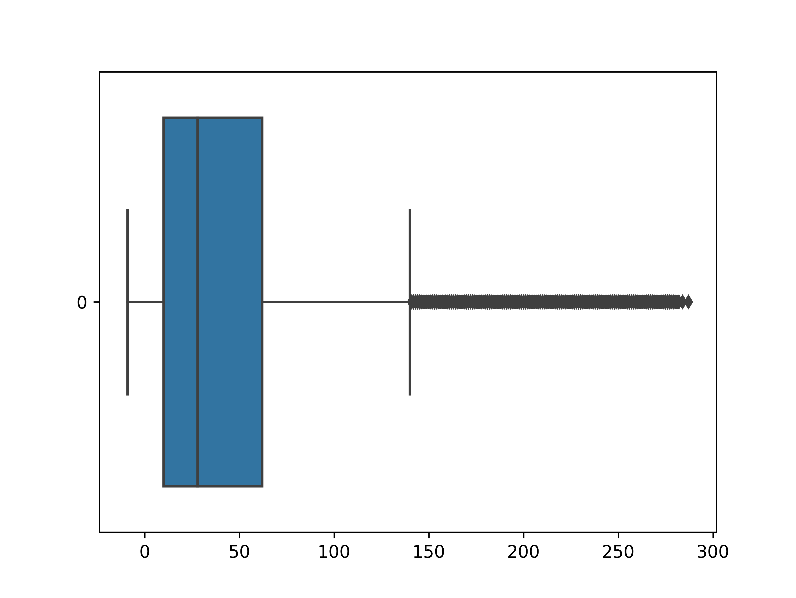
[5 rows x 113 columns]

# Create box plot to visualise the outliers of all the numeric columns

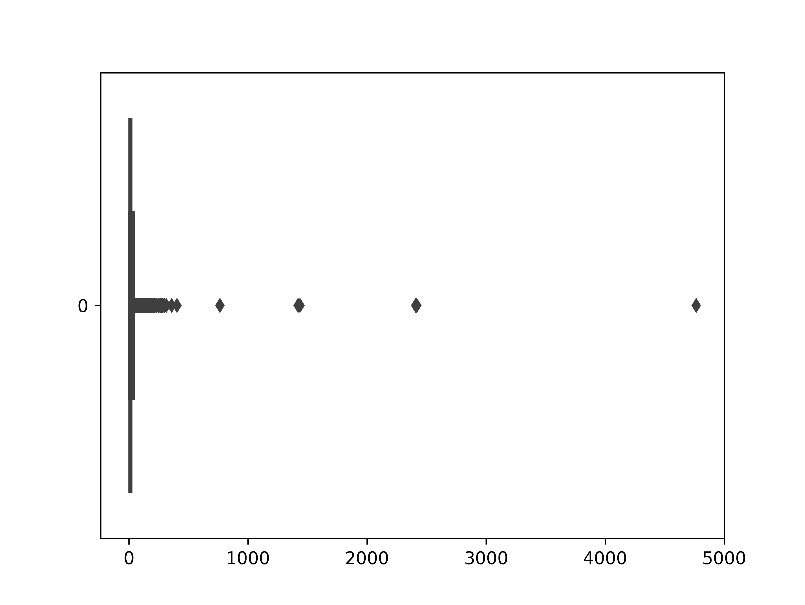
LCPCOUNT:



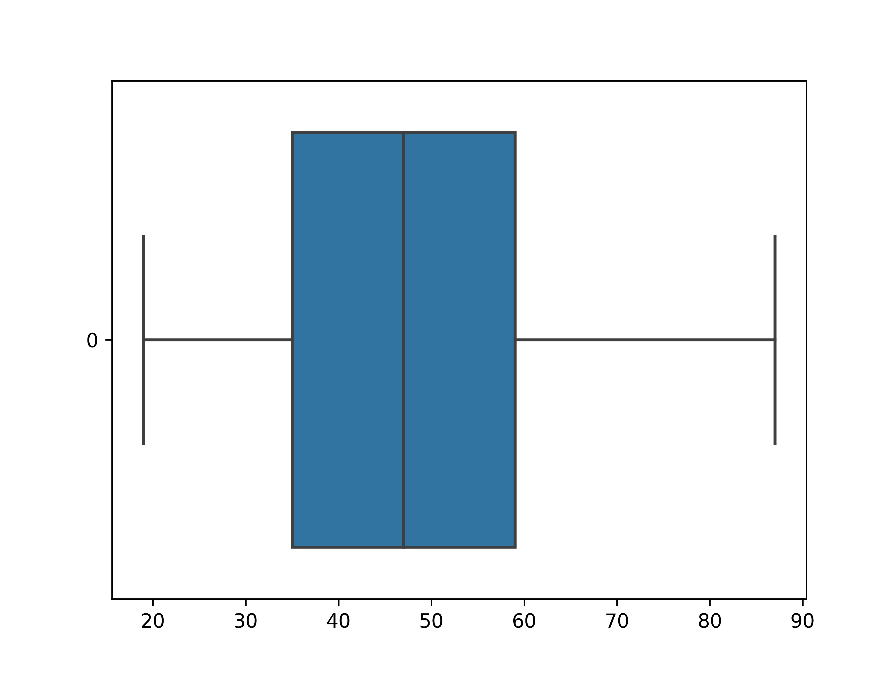
TENURE\_IN\_MONTHS:



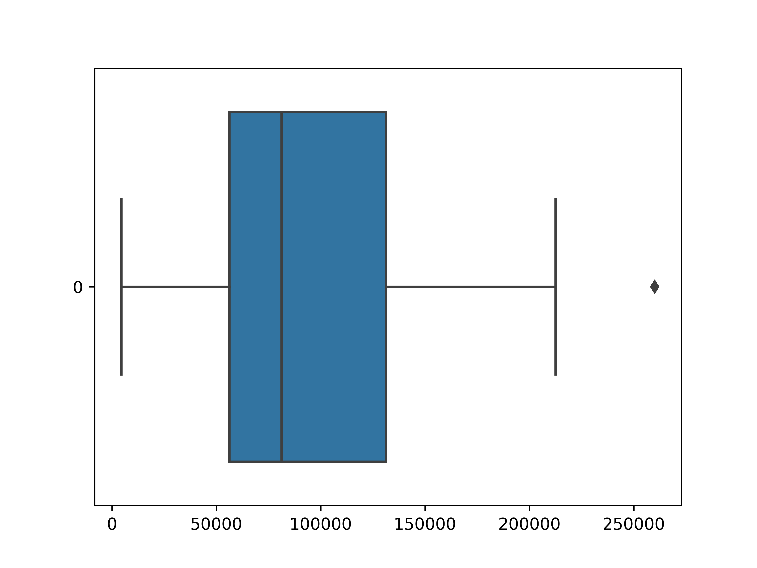
CLOSESTSTOREDISTANCE:



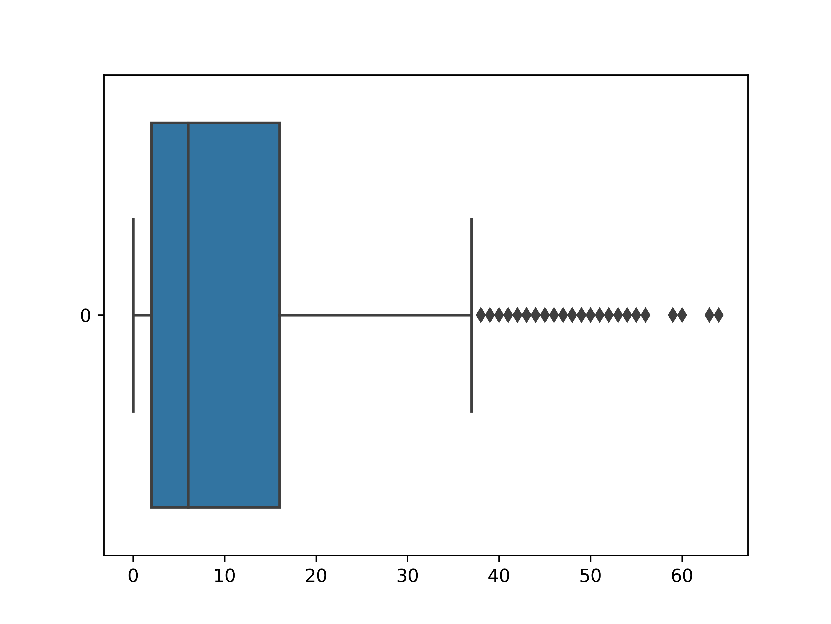
AGE:



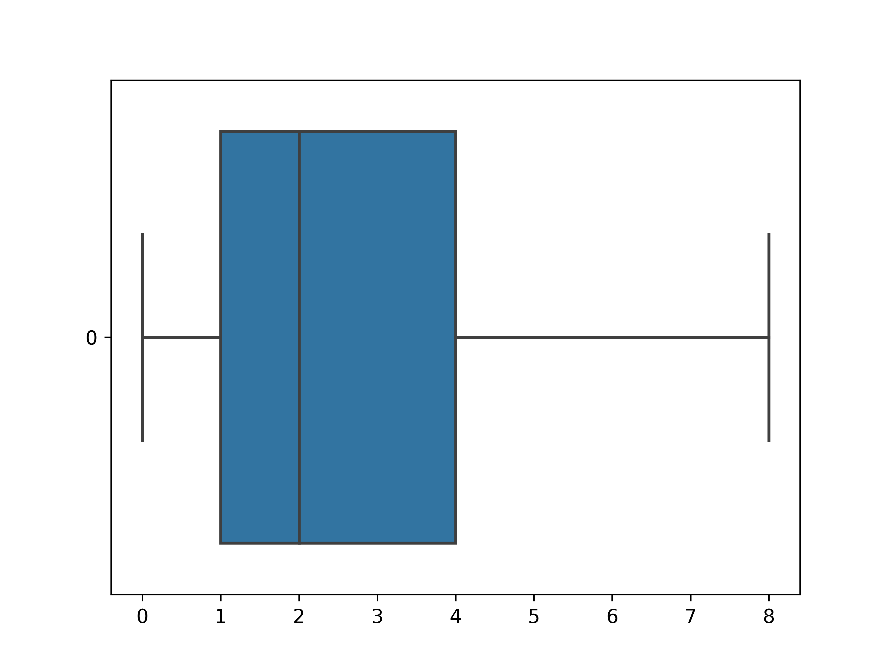
INCOME:



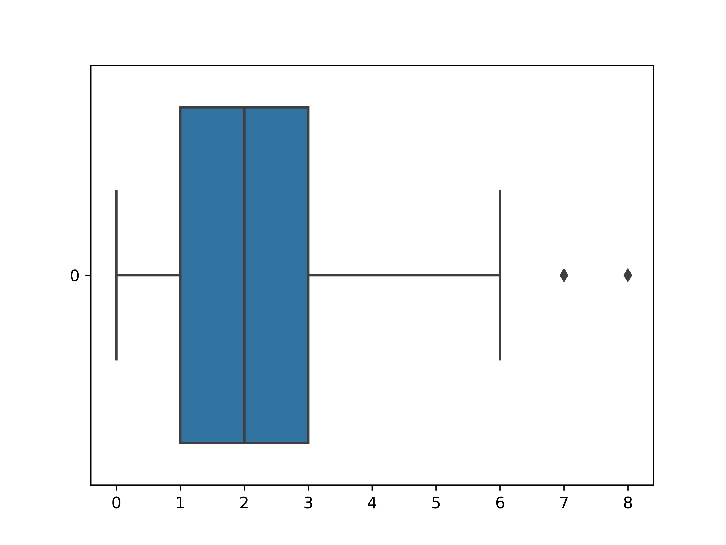
LENGTH OF RESIDENCE:



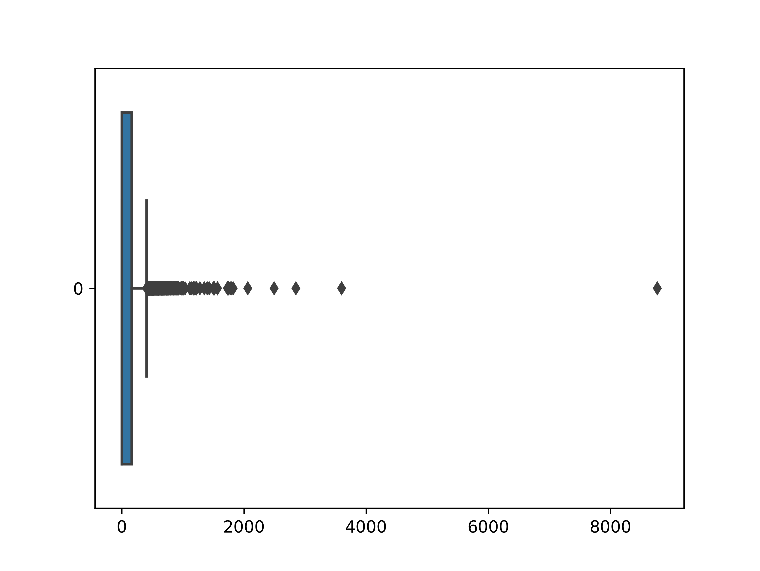
NUMBER OF PERSONS IN LIVING UNIT:



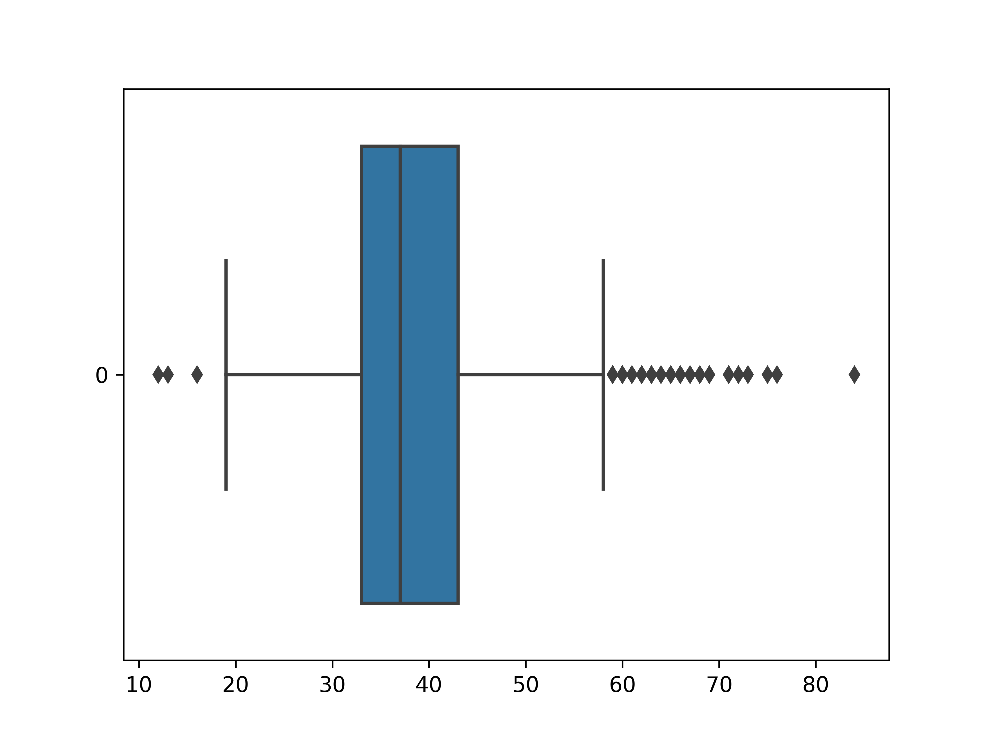
NUMBER OF ADULTS IN LIVING UNIT:



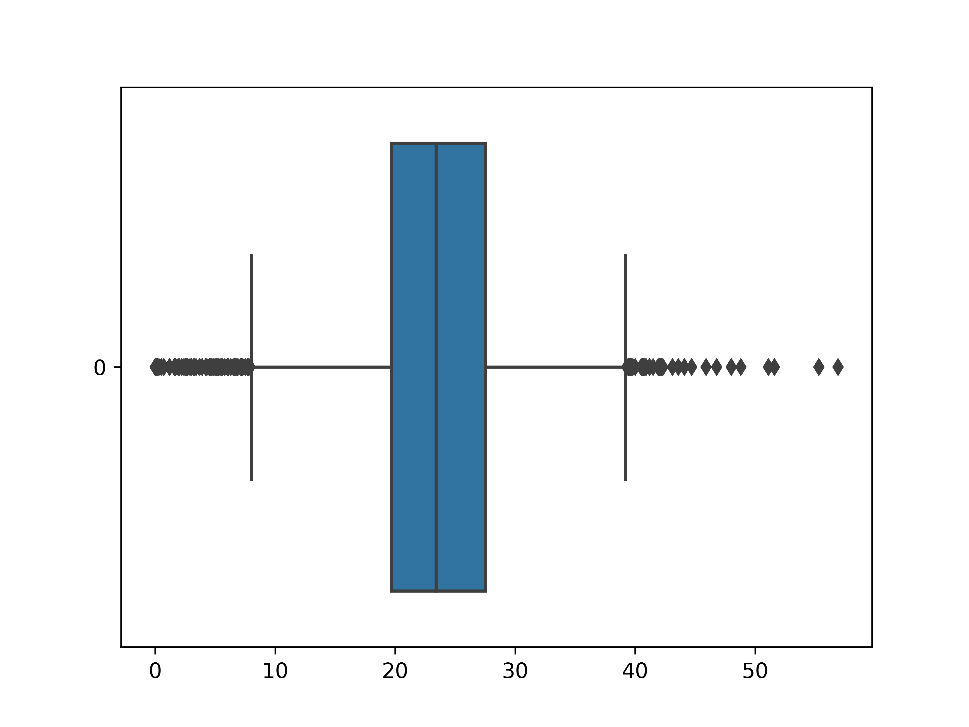
MORTGAGE-HOME PURCHASE: HOME PURCHASE PRICE:

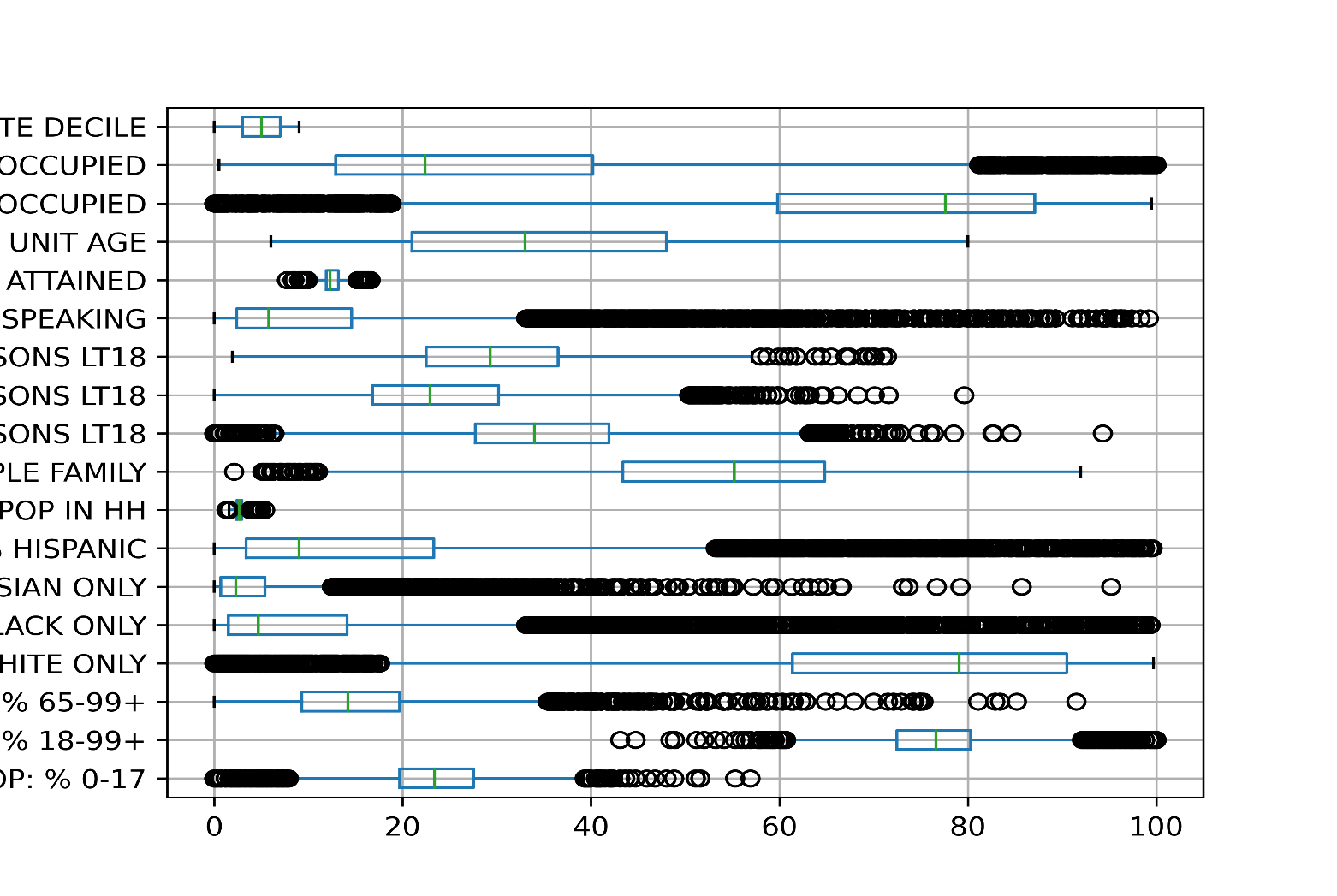


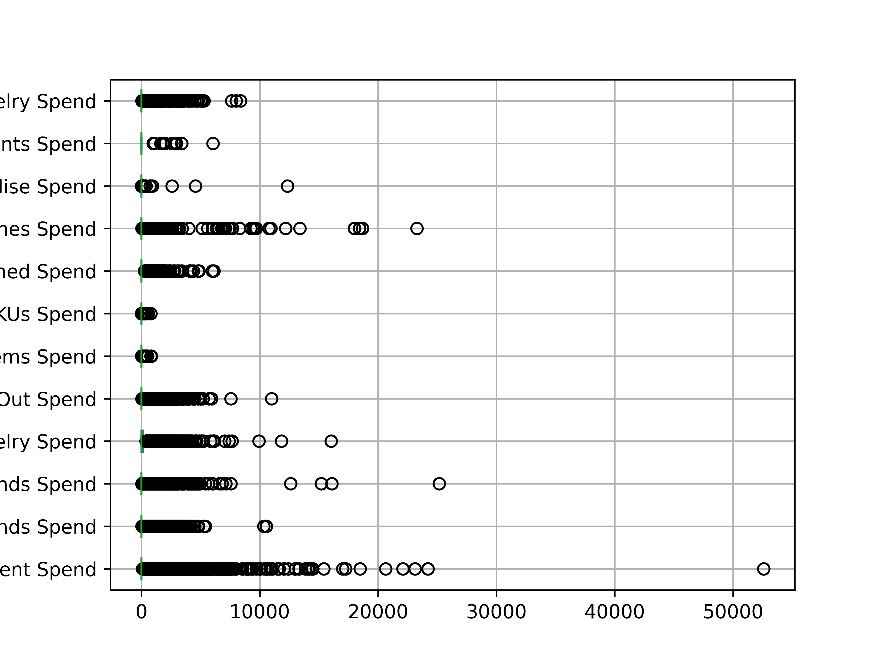
CAPE: AGE: POP: MEDIAN AGE:

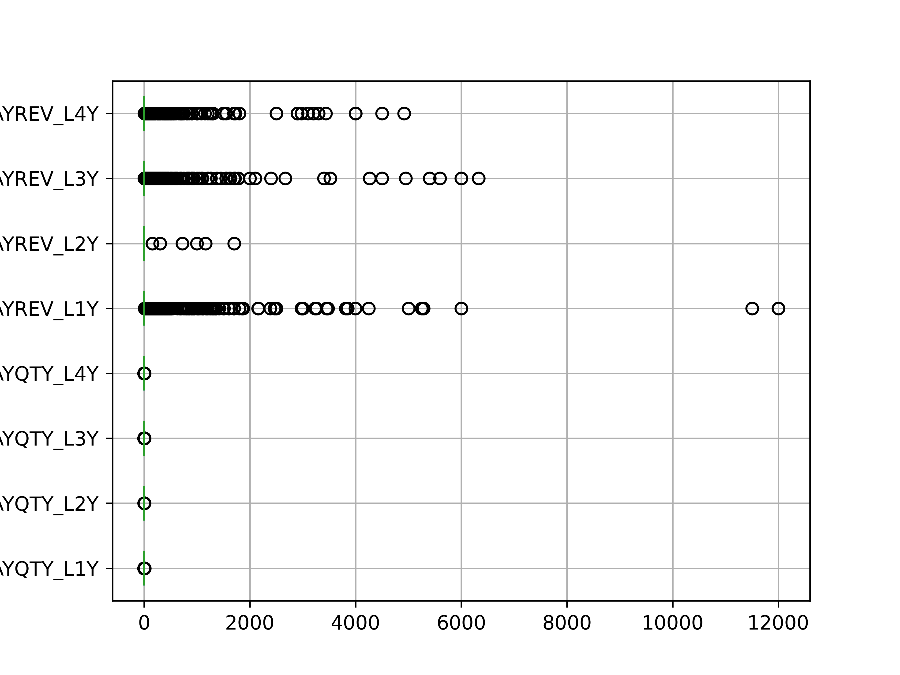


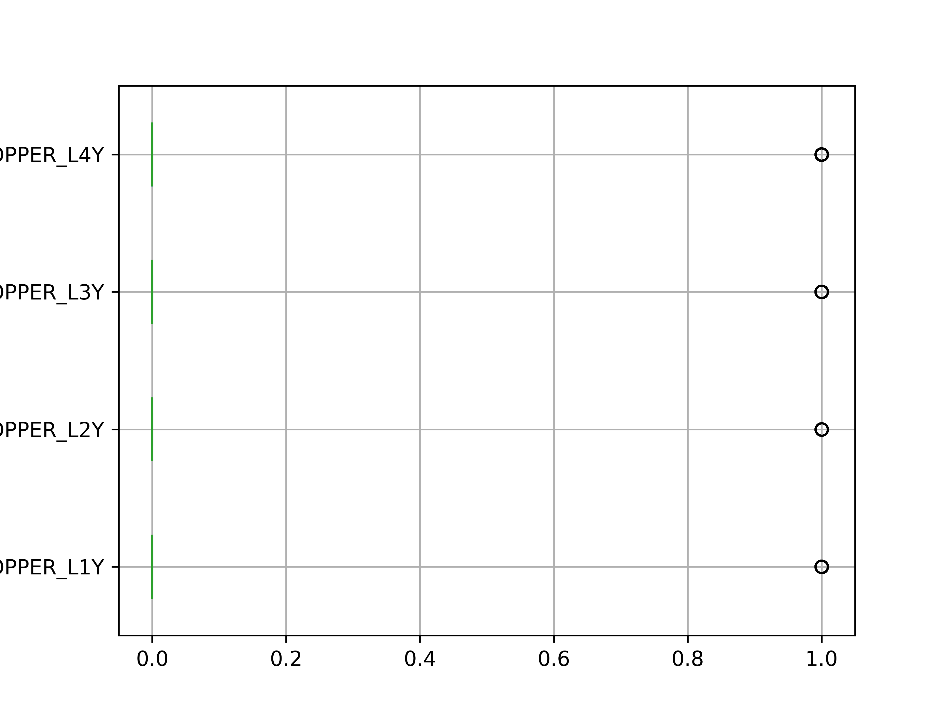
CAPE: AGE: POP: % 0-17:

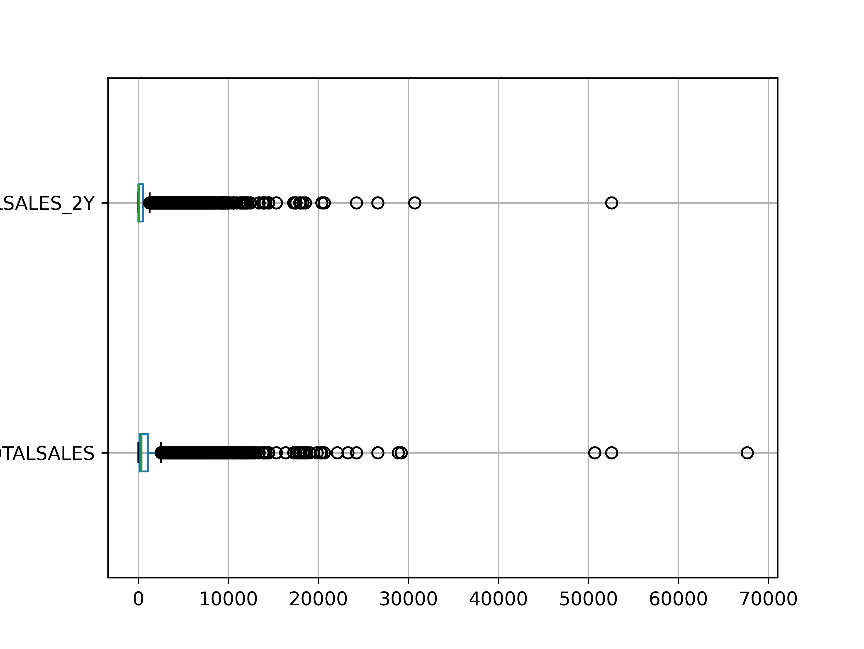


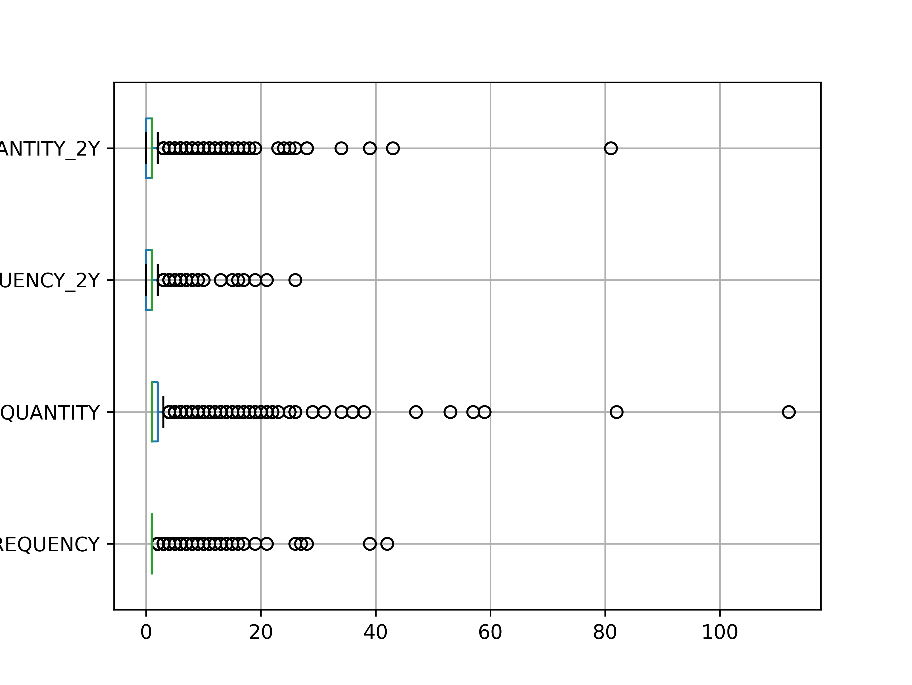


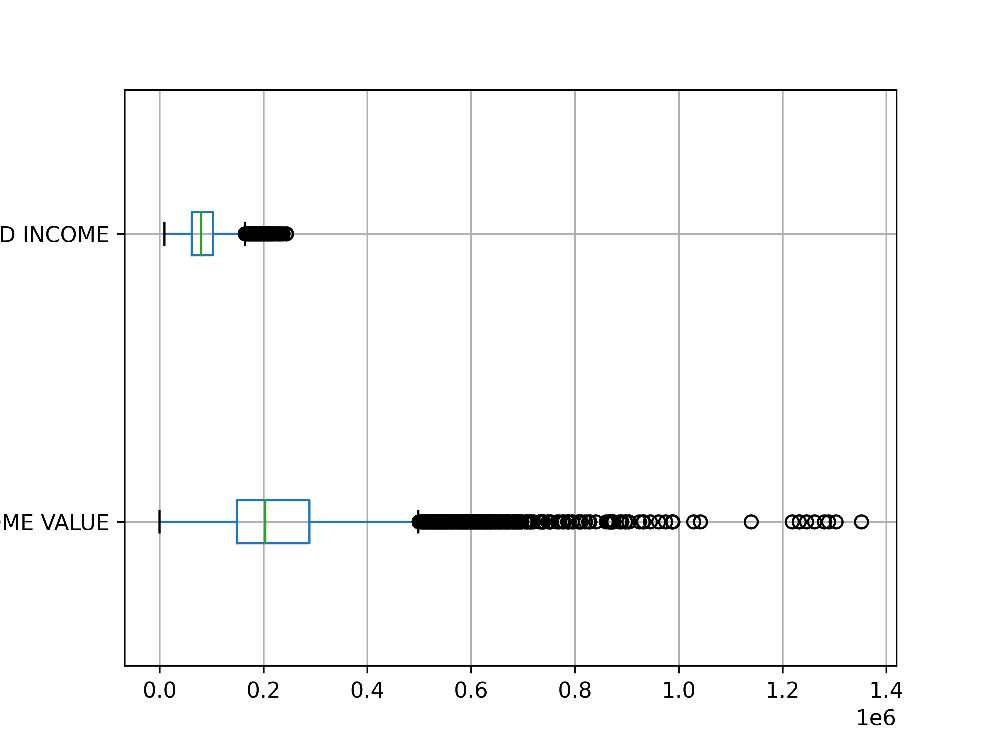












# Create charts for any 6 columns and show their distribution

