# DescTC - python package

December 16, 2020

## 1 DescTC

https://github.com/marianealves/DescTC

#### 1.0.1 Installation

pip install DescTC

https://pypi.org/project/DescTC/

### 1.1 To save you time! <(ºvº)>**

The DescTC python package provides the distribution and valuable information about each variable of your dataset helping you to decide which data cleansing method should be used without having to type lots of commands one at a time.

Methods provided:

**DescTC.table( )**

```
Offers you the following information of each quantitative/qualitative variable:

- Type
- Quantity of zero numbers
- Quantity of NaN's
- % of NaN's
- Quantity of uniques values
- Quantity of outliers
- Min value / Lowest category
- Mean
- Median
- Mode
- Max value / Highest category
```

### DescTC.chart( )

```
Condense large amounts of information of each variable into easy-to-understand formats
that clearly and effectively communicate important points:

- Plot the distribution of each variable
- Box plot of each quantitative variables
- Plot the correlation between quantitative variables
```

### DescTC.printfullTable( )

```
- Useful to see the entire outcome independently on which environment you are executing the pa
```

*Please be aware that your data must be coverted to a pandas DataFrame with column names.*

*Use the help( ) function to display the documentation of the specified module.*

See below the package outcome using a pandas DataFrames example.

### Installing required packages

[1]:
```
!pip install pandas
!pip install numpy
!pip install matplotlib
!pip install seaborn
```

### Installing DescTC package

[2]:
```
!pip install DescTC
```

### Importing methods

[3]:
```
from DescTC import *
```

### Importing data

[4]:
```
import pandas as pd

df = pd.read_csv("census.csv")
```

### Creating new instance of DescTC( )

[5]:
```
test = DescTC(df)
```

### Printing head/tail of the DataFrame

[6]:
```
test.df
```

[6]:
```
       age          workclass  final-weight    education  education-num  \
0       39          State-gov         77516    Bachelors             13
1       50   Self-emp-not-inc         83311    Bachelors             13
2       38            Private        215646      HS-grad              9
3       53            Private        234721         11th              7
4       28            Private        338409    Bachelors             13
…     …                   …             …            …              …
32556   27            Private        257302   Assoc-acdm             12
32557   40            Private        154374      HS-grad              9
32558   58            Private        151910      HS-grad              9
32559   22            Private        201490      HS-grad              9
32560   52       Self-emp-inc        287927      HS-grad              9

            marital-status          occupation    relationship    race  \
0            Never-married        Adm-clerical   Not-in-family   White
1       Married-civ-spouse     Exec-managerial         Husband   White
2                 Divorced   Handlers-cleaners   Not-in-family   White
3       Married-civ-spouse   Handlers-cleaners         Husband   Black
4       Married-civ-spouse      Prof-specialty            Wife   Black
…                      …                 …             …      …
32556   Married-civ-spouse        Tech-support            Wife   White
32557   Married-civ-spouse   Machine-op-inspct         Husband   White
32558              Widowed        Adm-clerical       Unmarried   White
32559        Never-married        Adm-clerical       Own-child   White
32560   Married-civ-spouse     Exec-managerial            Wife   White

           sex  capital-gain  capital-loos  hour-per-week  native-country  \
0         Male          2174             0             40   United-States
1         Male             0             0             13   United-States
2         Male             0             0             40   United-States
3         Male             0             0             40   United-States
4       Female             0             0             40            Cuba
…         …            …             …             …              …
32556   Female             0             0             38   United-States
32557     Male             0             0             40   United-States
32558   Female             0             0             40   United-States
32559     Male             0             0             20   United-States
32560   Female         15024             0             40   United-States

      income
0      <=50K
1      <=50K
2      <=50K
3      <=50K
```

```
4       <=50K
…       …
32556   <=50K
32557    >50K
32558   <=50K
32559   <=50K
32560    >50K

[32561 rows x 15 columns]
```

**Accessing method: DescTC.table( )**

[7]:
```
test.table()
```

[7]:

| | Type | Quant.Zeros | Quant.NaNs | %NaNs | Quant.Uniques | \ |
|---|---|---|---|---|---|---|
| age | int64 | 0 | 0 | 0.00 | 73 | |
| workclass | object | 0 | 1836 | 5.64 | 9 | |
| final-weight | int64 | 0 | 0 | 0.00 | 21648 | |
| education | object | 0 | 0 | 0.00 | 16 | |
| education-num | int64 | 0 | 0 | 0.00 | 16 | |
| marital-status | object | 0 | 0 | 0.00 | 7 | |
| occupation | object | 0 | 1843 | 5.66 | 15 | |
| relationship | object | 0 | 0 | 0.00 | 6 | |
| race | object | 0 | 0 | 0.00 | 5 | |
| sex | object | 0 | 0 | 0.00 | 2 | |
| capital-gain | int64 | 29849 | 0 | 0.00 | 119 | |
| capital-loos | int64 | 31042 | 0 | 0.00 | 92 | |
| hour-per-week | int64 | 0 | 0 | 0.00 | 94 | |
| native-country | object | 0 | 583 | 1.79 | 42 | |
| income | object | 0 | 0 | 0.00 | 2 | |

| | Quant.Outliers | Min/Lowest | Mean | Median | \ |
|---|---|---|---|---|---|
| age | [121] | 17 | 38.5816 | 37 | |
| workclass | [0] | Never-worked | NaN | NaN | |
| final-weight | [347] | 12285 | 189778 | 178356 | |
| education | [0] | Preschool | NaN | NaN | |
| education-num | [219] | 1 | 10.0807 | 10 | |
| marital-status | [0] | Married-AF-spouse | NaN | NaN | |
| occupation | [0] | Armed-Forces | NaN | NaN | |
| relationship | [0] | Other-relative | NaN | NaN | |
| race | [0] | Other | NaN | NaN | |
| sex | [0] | Female | NaN | NaN | |
| capital-gain | [215] | 0 | 1077.65 | 0 | |
| capital-loos | [1470] | 0 | 87.3038 | 0 | |
| hour-per-week | [440] | 1 | 40.4375 | 40 | |
| native-country | [0] | Holand-Netherlands | NaN | NaN | |
| income | [0] | >50K | NaN | NaN | |

| | Mode | Max/Highest |
|---|---|---|
| age | 36 | 90 |
| workclass | Private | Private |
| final-weight | 123011 | 1484705 |
| education | HS-grad | HS-grad |
| education-num | 9 | 16 |
| marital-status | Married-civ-spouse | Married-civ-spouse |
| occupation | Prof-specialty | Prof-specialty |
| relationship | Husband | Husband |
| race | White | White |
| sex | Male | Male |
| capital-gain | 0 | 99999 |
| capital-loos | 0 | 4356 |
| hour-per-week | 40 | 99 |
| native-country | United-States | United-States |
| income | <=50K | <=50K |

**Other alternative for the table method:**

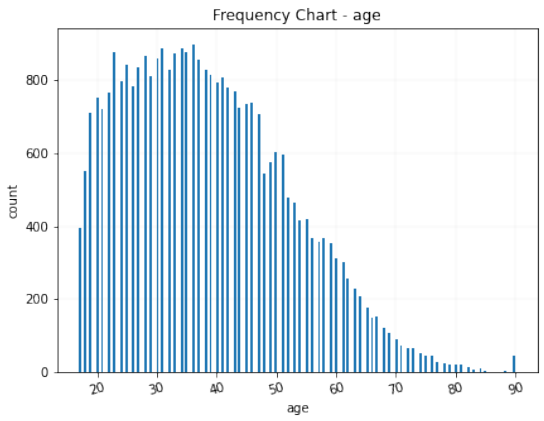
- The printfullTable method is useful to see the entire outcome independently on which environment you are executing the package.

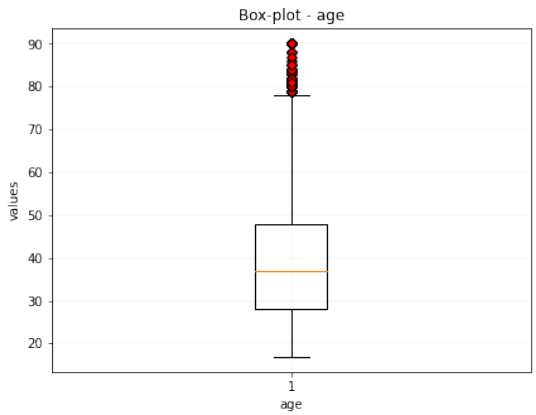
test.printfullTable( )

**Accessing method: DescTC.chart( )**

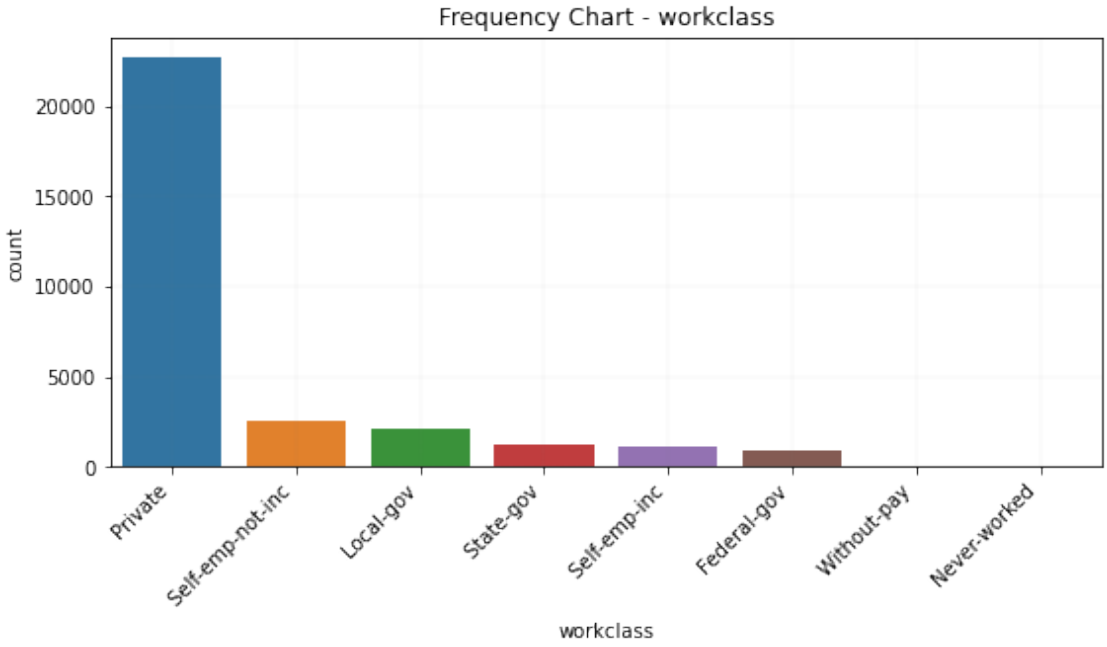
[8]:
```
test.chart()
```

Variable: age

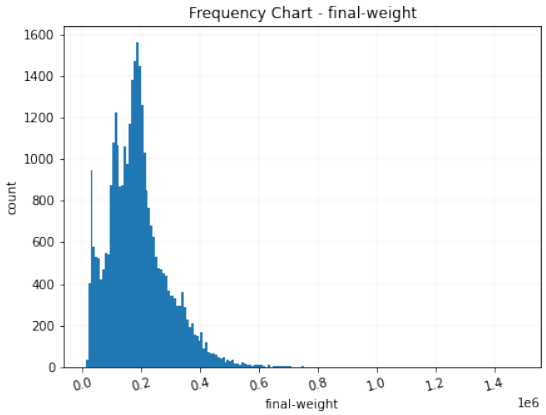


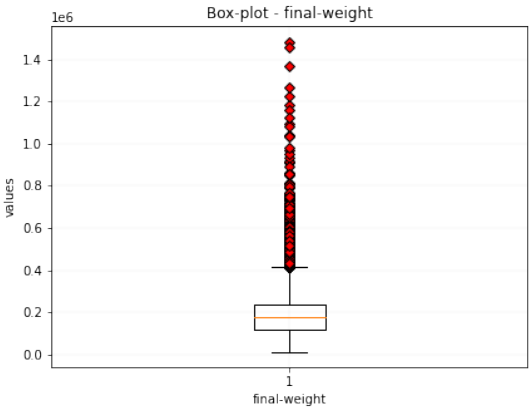




Variable: workclass

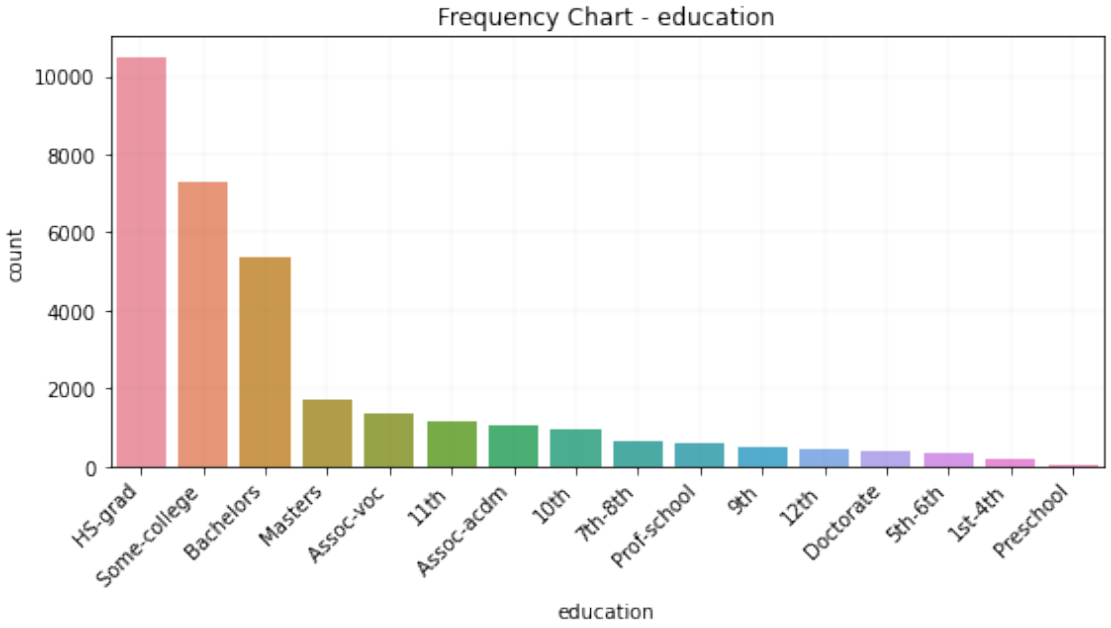



Variable: final-weight

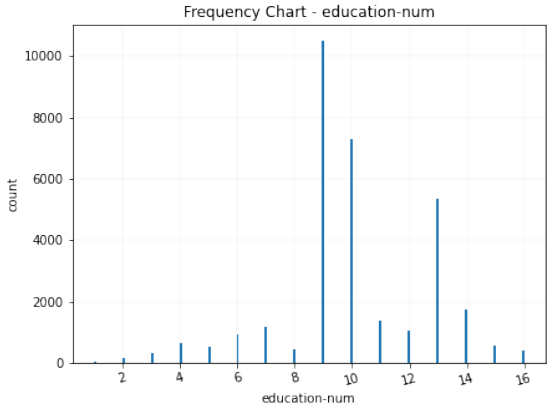


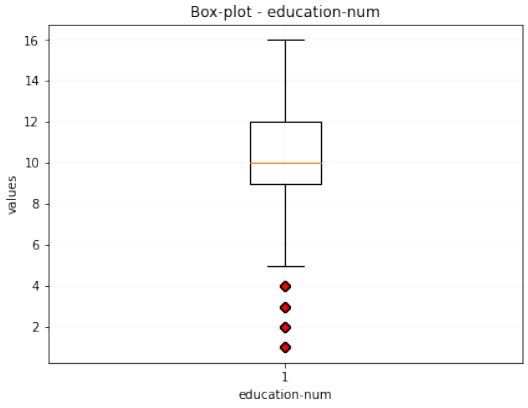




Variable: education

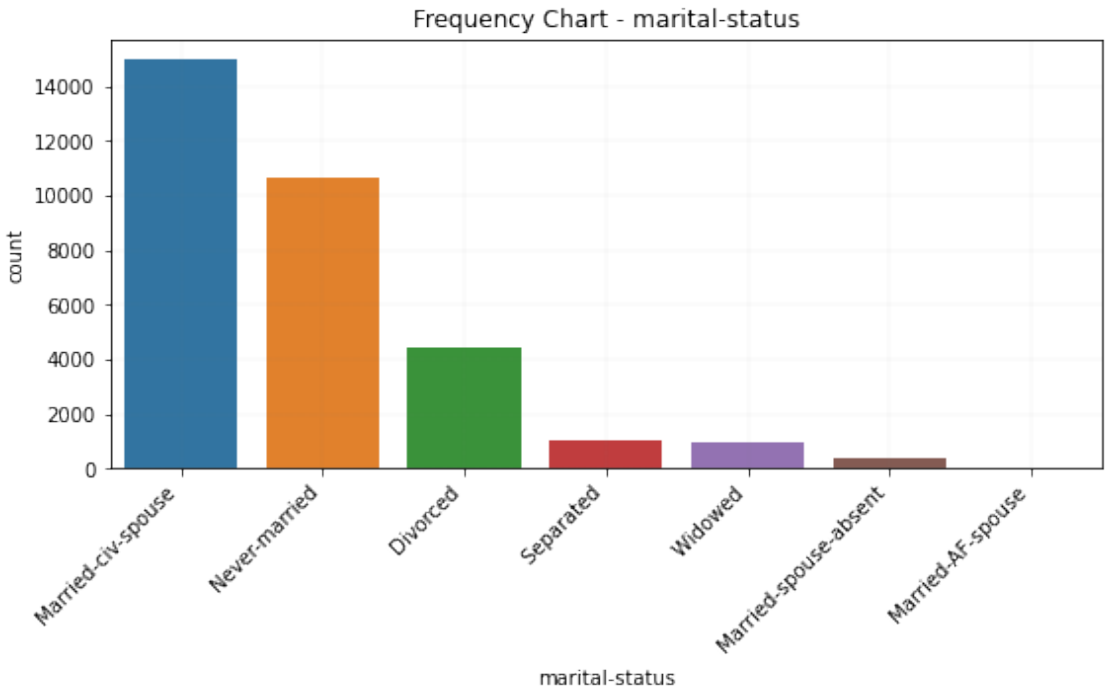



Variable: education-num

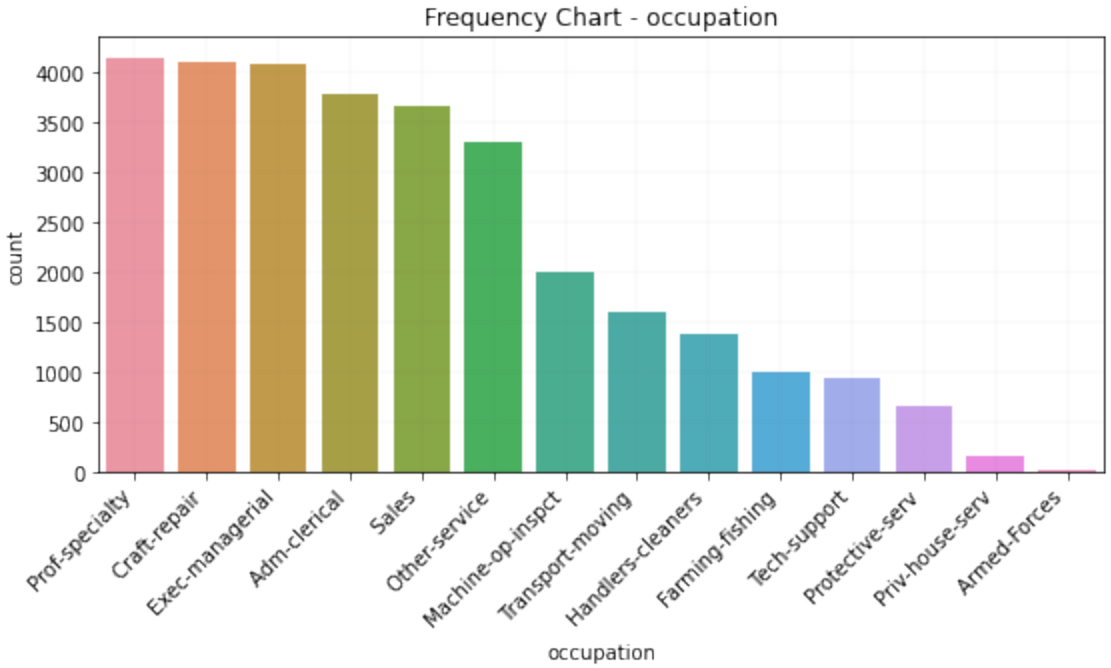


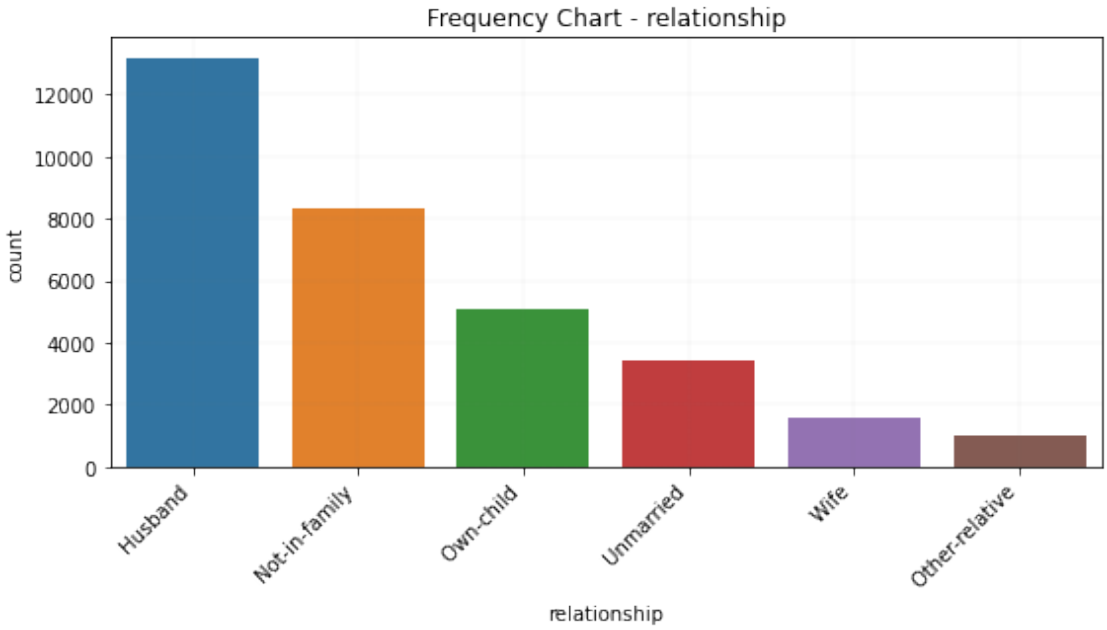

Variable: marital-status



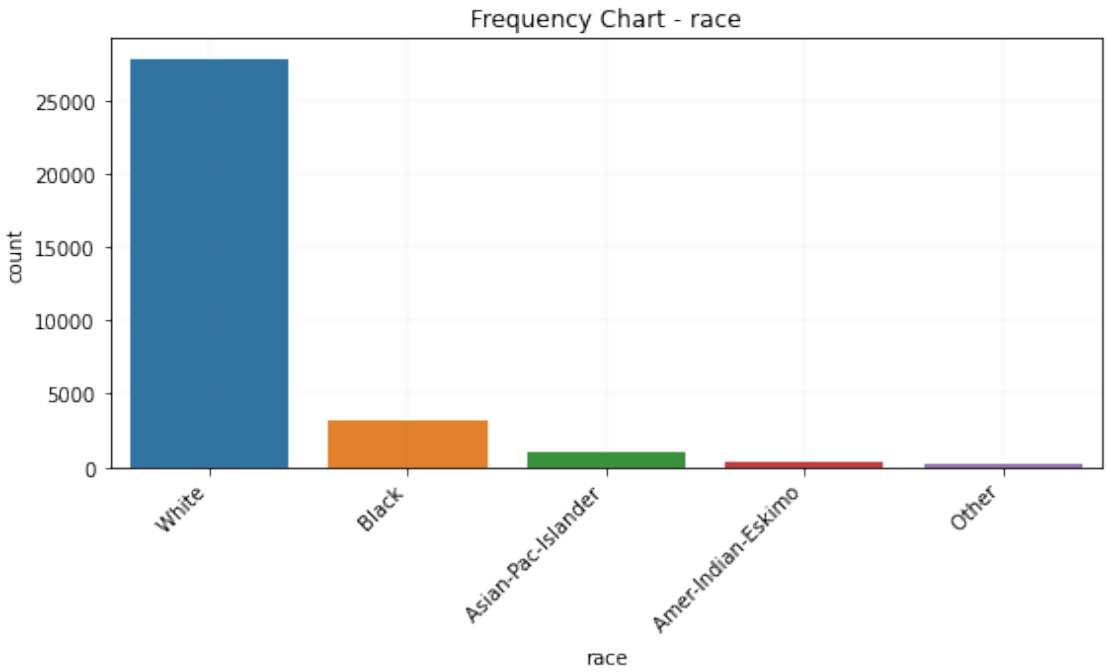

Variable: occupation



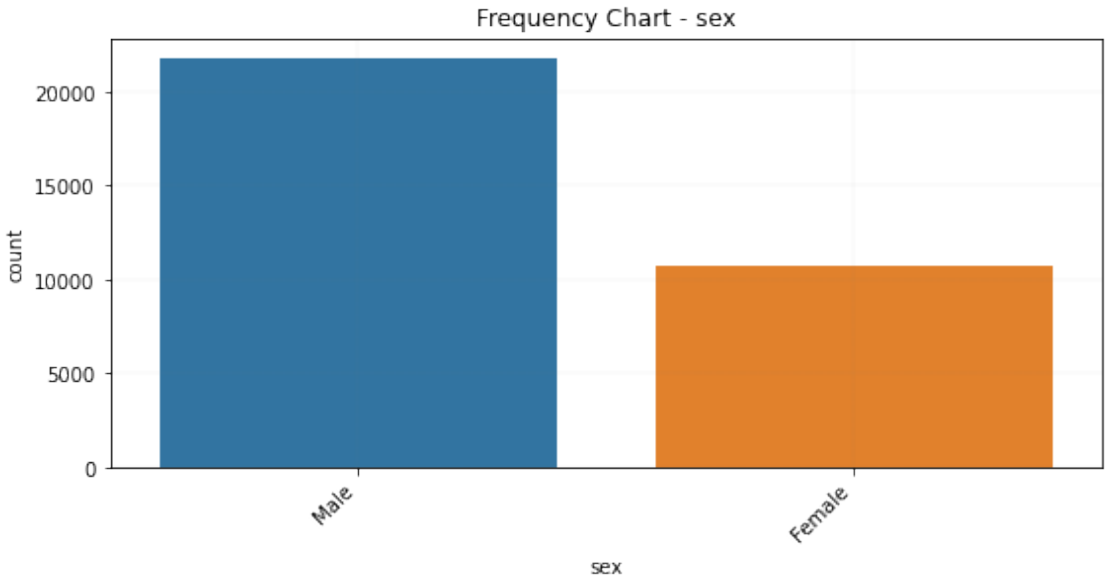
Variable: relationship



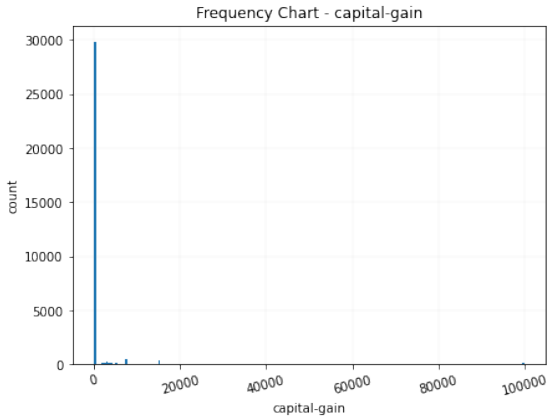

Variable: race

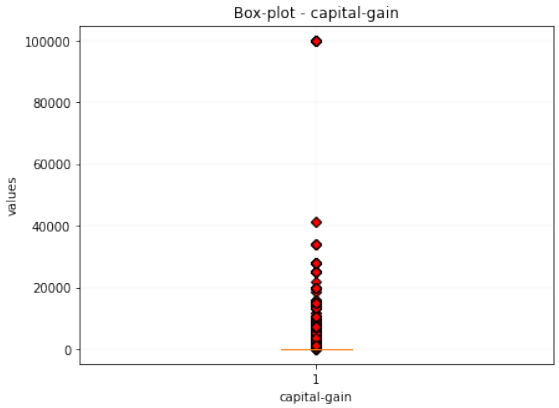



Variable: sex

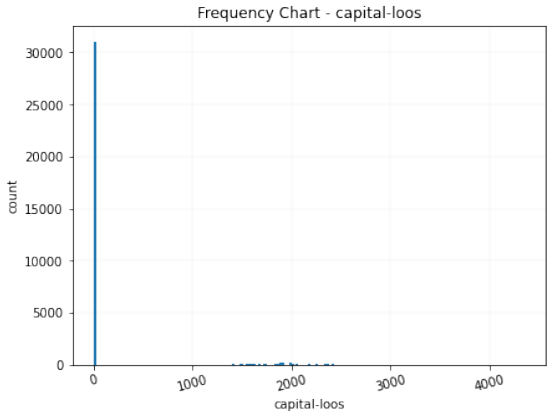


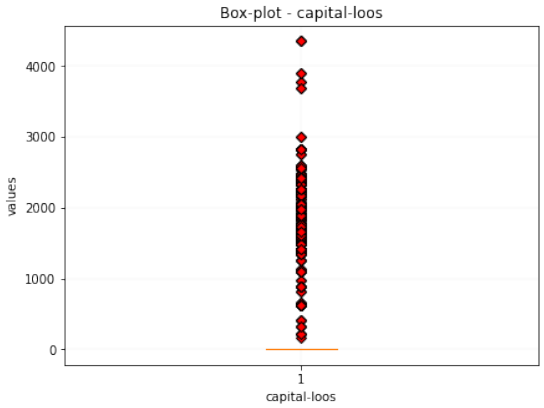
Variable: capital-gain



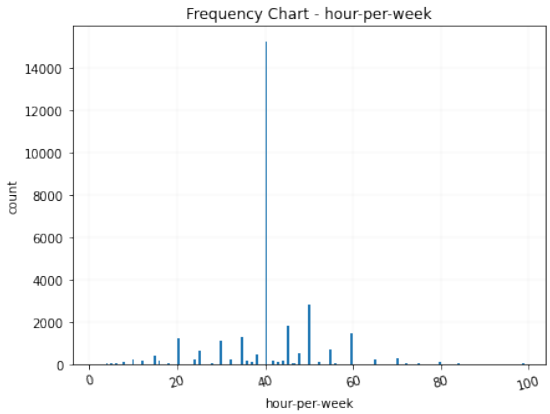



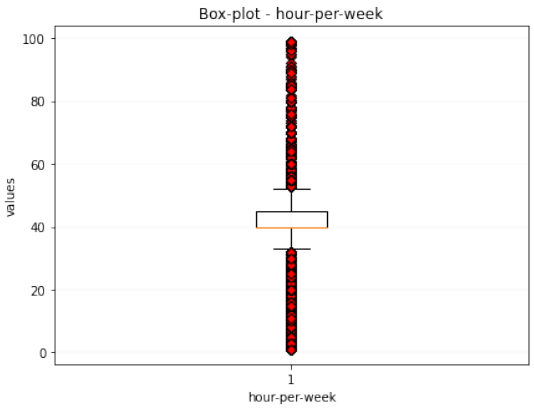

Variable: capital-loos



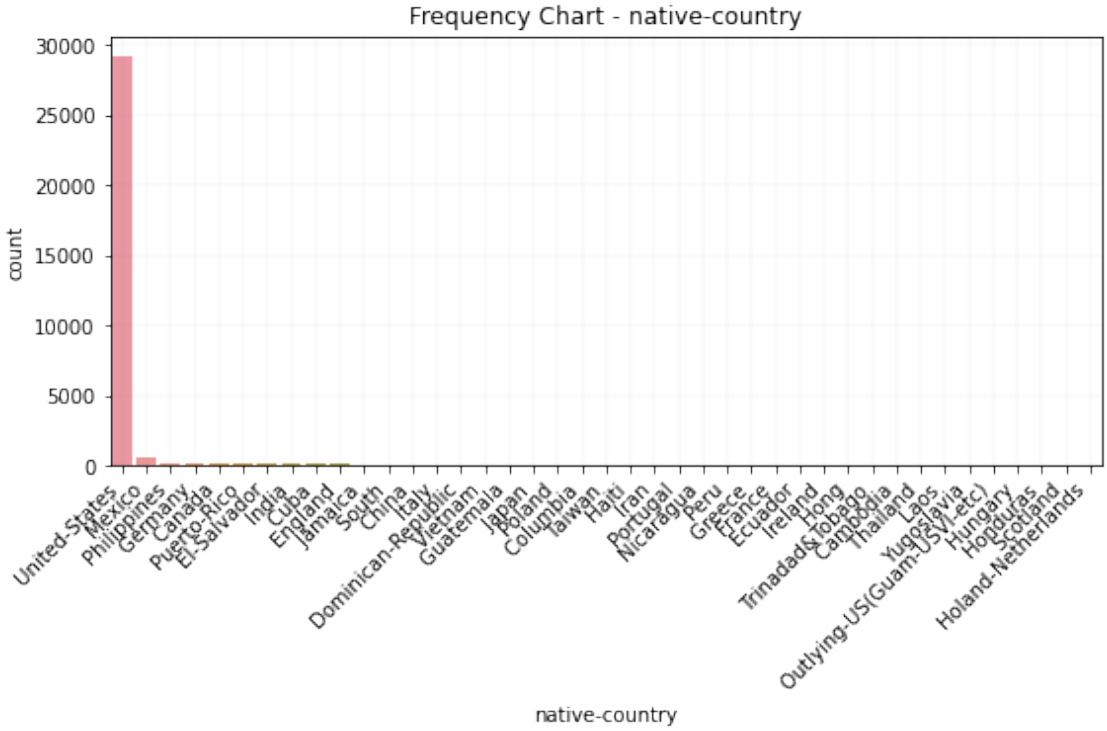




Variable: hour-per-week

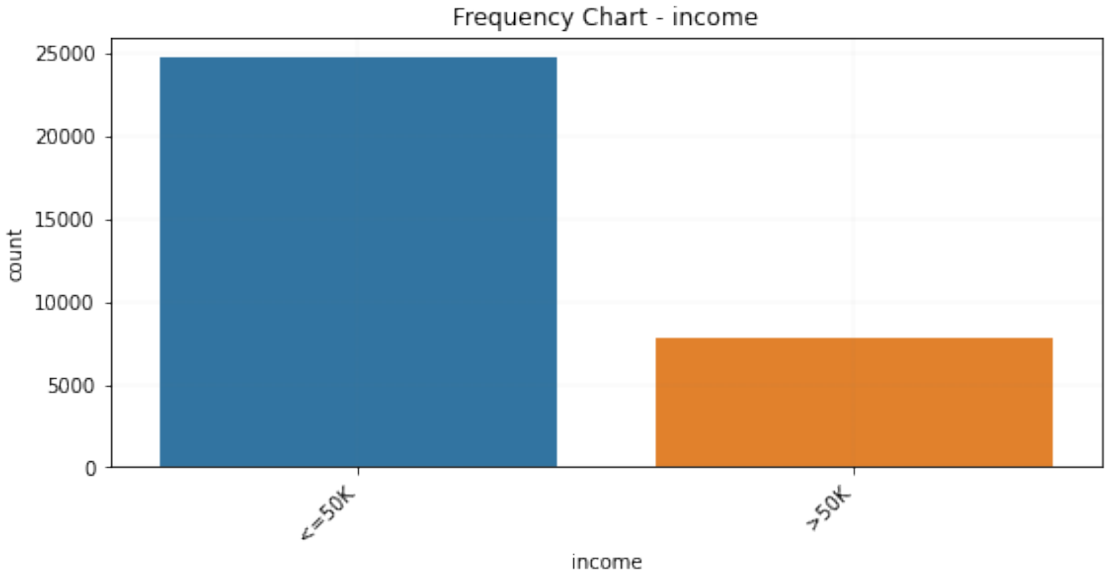


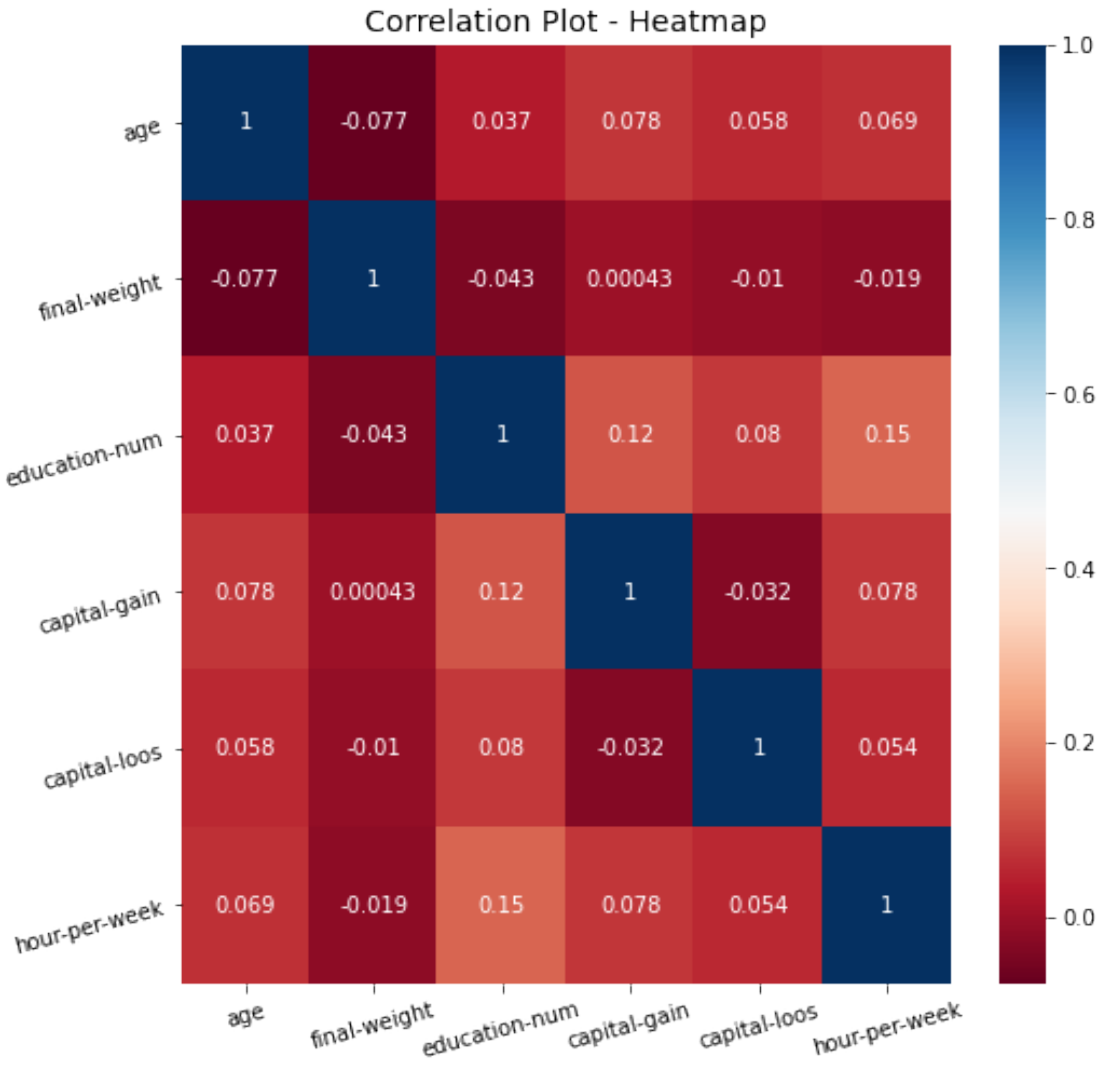




Variable: native-country




Variable: income


Frequency Chart - income
25000
20000
15000
10000
5000
0
count
<=50K
>50K
income




**Note:** The object data type can actually contain multiple different types. For instance, the column could include integers, floats, and strings which collectively are labeled as an object. Therefore, you may not get the box plot plotted from an object dtype variable.