```
%pip install -U dataprep
```

```
Requirement already satisfied: cloudpickle>=1.1.1 in /usr/local/lib/python3.7/dist-pa
Requirement already satisfied: fsspec>=0.6.0 in /usr/local/lib/python3.7/dist-package
Requirement already satisfied: toolz>=0.8.2 in /usr/local/lib/python3.7/dist-packages
Requirement already satisfied: partd>=0.3.10 in /usr/local/lib/python3.7/dist-package
Requirement already satisfied: Werkzeug>=2.0 in /usr/local/lib/python3.7/dist-package
Requirement already satisfied: itsdangerous>=2.0 in /usr/local/lib/python3.7/dist-pac
Requirement already satisfied: importlib-metadata>=3.6.0 in /usr/local/lib/python3.7/
Requirement already satisfied: click>=8.0 in /usr/local/lib/python3.7/dist-packages (
Requirement already satisfied: Six in /usr/local/lib/python3.7/dist-packages (from fl
Requirement already satisfied: zipp>=0.5 in /usr/local/lib/python3.7/dist-packages (f
Requirement already satisfied: ipykernel>=4.5.1 in /usr/local/lib/python3.7/dist-pack
Requirement already satisfied: traitlets>=4.3.1 in /usr/local/lib/python3.7/dist-pack
Requirement already satisfied: nbformat>=4.2.0 in /usr/local/lib/python3.7/dist-packa
Requirement already satisfied: jupyterlab-widgets>=1.0.0 in /usr/local/lib/python3.7/

Requirement already satisfied: ipython-genutils~=0.2.0 in /usr/local/lib/python3.7/di
Requirement already satisfied: ipython>=4.0.0 in /usr/local/lib/python3.7/dist-packag
Requirement already satisfied: widgetsnbextension~=3.6.0 in /usr/local/lib/python3.7/
Requirement already satisfied: jupyter-client in /usr/local/lib/python3.7/dist-packag
Requirement already satisfied: pickleshare in /usr/local/lib/python3.7/dist-packages
Requirement already satisfied: prompt-toolkit<2.0.0,>=1.0.4 in /usr/local/lib/python3
Requirement already satisfied: simplegeneric>0.8 in /usr/local/lib/python3.7/dist-pac
Requirement already satisfied: pexpect in /usr/local/lib/python3.7/dist-packages (fro
Requirement already satisfied: pygments in /usr/local/lib/python3.7/dist-packages (fr
Requirement already satisfied: setuptools>=18.5 in /usr/local/lib/python3.7/dist-pack
Requirement already satisfied: decorator in /usr/local/lib/python3.7/dist-packages (f
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.7/dist-packa
Requirement already satisfied: ply in /usr/local/lib/python3.7/dist-packages (from js
Requirement already satisfied: jupyter-core in /usr/local/lib/python3.7/dist-packages
Requirement already satisfied: jsonschema!=2.5.0,>=2.4 in /usr/local/lib/python3.7/di
Requirement already satisfied: pyrsistent!=0.17.0,!=0.17.1,!=0.17.2,>=0.14.0 in /usr/
Requirement already satisfied: importlib-resources>=1.4.0 in /usr/local/lib/python3.7
Requirement already satisfied: joblib in /usr/local/lib/python3.7/dist-packages (from
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in /usr/local/lib/python3.7/d
Requirement already satisfied: pytz>=2017.3 in /usr/local/lib/python3.7/dist-packages
Requirement already satisfied: locket in /usr/local/lib/python3.7/dist-packages (from
Requirement already satisfied: wcwidth in /usr/local/lib/python3.7/dist-packages (fro
Requirement already satisfied: greenlet!=0.4.17 in /usr/local/lib/python3.7/dist-pack
Requirement already satisfied: asttokens<3.0.0,>=2.0.0 in /usr/local/lib/python3.7/di
Requirement already satisfied: pure_eval<1.0.0 in /usr/local/lib/python3.7/dist-packa
Requirement already satisfied: executing<0.9.0,>=0.8.3 in /usr/local/lib/python3.7/di
Requirement already satisfied: notebook>=4.4.1 in /usr/local/lib/python3.7/dist-packa
Requirement already satisfied: nbconvert in /usr/local/lib/python3.7/dist-packages (f
Requirement already satisfied: Send2Trash in /usr/local/lib/python3.7/dist-packages (
Requirement already satisfied: terminado>=0.8.1 in /usr/local/lib/python3.7/dist-pack
Requirement already satisfied: pyzmq>=13 in /usr/local/lib/python3.7/dist-packages (f
Requirement already satisfied: ptyprocess in /usr/local/lib/python3.7/dist-packages (
Requirement already satisfied: matplotlib in /usr/local/lib/python3.7/dist-packages (
Requirement already satisfied: idna>=2.0 in /usr/local/lib/python3.7/dist-packages (f
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.7/dist-pac
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.7/dist-packages
Requirement already satisfied: entrypoints>=0.2.2 in /usr/local/lib/python3.7/dist-pa
Requirement already satisfied: bleach in /usr/local/lib/python3.7/dist-packages (from
```

```
Requirement already satisfied: testpath in /usr/local/lib/python3.7/dist-packages (fr
Requirement already satisfied: mistune<2,>=0.8.1 in /usr/local/lib/python3.7/dist-pac
Requirement already satisfied: pandocfilters>=1.4.1 in /usr/local/lib/python3.7/dist-
Requirement already satisfied: defusedxml in /usr/local/lib/python3.7/dist-packages (
Requirement already satisfied: webencodings in /usr/local/lib/python3.7/dist-packages
```

```
# This Python 3 environment comes with many helpful analytics libraries installed
# It is defined by the kaggle/python Docker image: https://github.com/kaggle/docker-python
# For example, here's several helpful packages to load

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will list all files und

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# You can write up to 20GB to the current directory (/kaggle/working/) that gets preserved as
# You can also write temporary files to /kaggle/temp/, but they won't be saved outside of the


from dataprep.eda import create_report
from sklearn.model_selection import train_test_split
from sklearn.metrics import auc, classification_report, roc_auc_score
from sklearn.ensemble import RandomForestClassifier
import warnings
warnings.filterwarnings('ignore')
```

Reading Data

```
train=pd.read_csv('/content/beginner.csv')
```

```
train.head()
```

	YEAR	SERIAL	MONTH	HWTFINL	CPSID	ASECFLAG	HFLAG	ASECWTH	REGION	ST
0	2015	16566	3	NaN	0	1.0	NaN	1671.32	12	
1	2020	49554	6	1930.4505	20200504991400	NaN	NaN	NaN	33	

```
print("Shape of Train Dataset: "+str(train.shape))
```

```
Shape of Train Dataset: (54737, 29)
```

4	2018	31356	3	NaN	20170102597500	1.0	NaN	1719.16	22	

If we exclude the Column we plan to predict, we have 28 columns to use as features and we are going to predict the INCWAGE column

Let's check the type of these columns

```
train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 54737 entries, 0 to 54736
Data columns (total 29 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   YEAR       54737 non-null  int64  
 1   SERIAL     54737 non-null  int64  
 2   MONTH      54737 non-null  int64  
 3   HWTFINL    30160 non-null  float64
 4   CPSID      54737 non-null  int64  
 5   ASECFLAG   26695 non-null  float64
 6   HFLAG      2028 non-null   float64
 7   ASECWTH    24577 non-null  float64
 8   REGION     54737 non-null  int64  
 9   STATEFIP   54737 non-null  int64  
 10  NFAMS      54737 non-null  int64  
 11  PERNUM     54737 non-null  int64  
 12  WTFINL     30160 non-null  float64
 13  CPSIDP     54737 non-null  int64  
 14  ASECWT     24577 non-null  float64
 15  AGE        54737 non-null  int64  
 16  SEX        54737 non-null  int64  
 17  RACE       54737 non-null  int64  
 18  MARST      54737 non-null  int64  
 19  BPL        54737 non-null  int64  
 20  EMPSTAT    54737 non-null  int64  
 21  OCC        54737 non-null  int64  
 22  UHRSWORKT  54737 non-null  int64  
 23  WKSTAT     54737 non-null  int64  
 24  JOBCERT    30160 non-null  float64
 25  EDUC       54737 non-null  int64  
 26  EDDIPGED   30160 non-null  float64
 27  INCWAGE    24577 non-null  float64
```

```
 28  OINCWAGE   24577 non-null  float64
dtypes: float64(10), int64(19)
memory usage: 12.1 MB
```

We have 8 columns considered as Categorical features and the rest are numerical features

Let's check for missing values

```
train.isna().sum()
```

```
YEAR             0
SERIAL           0
MONTH            0
HWTFINL      24577
CPSID            0
ASECFLAG     28042
HFLAG        52709
ASECWTH      30160
REGION           0
STATEFIP         0
NFAMS            0
PERNUM           0
WTFINL       24577
CPSIDP           0
ASECWT       30160
AGE              0
SEX              0
RACE             0
MARST            0
BPL              0
EMPSTAT          0
OCC              0
UHRSWORKT        0
WKSTAT           0
JOBCERT      24577
EDUC             0
EDDIPGED     24577
INCWAGE      30160
OINCWAGE     30160
dtype: int64
```

To have quick EDA we will use a powerful library that provides insights on the dataset

```
report = create_report(train, title='My Report')
```

```
report
```

My Report Overview Variables ≡ Interactions Correlations

Missing Values

Overview

Dataset Statistics

Number of Variables	29
Number of Rows	54737
Missing Cells	299699
Missing Cells (%)	18.9%
Duplicate Rows	0
Duplicate Rows (%)	0.0%
Total Size in Memory	12.1 MB
Average Row Size in Memory	232.0 B
Variable Types	Numerical: 21 Categorical: 8

Dataset Insights

HWTFINL and WTFINL have similar distributions	Similar Distribution
CPSID and CPSIDP have similar distributions	Similar Distribution
ASECWTH and ASECWT have similar distributions	Similar Distribution
HWTFINL has 24577 (44.9%) missing values	Missing
ASECFLAG has 28042 (51.23%) missing values	Missing
HFLAG has 52709 (96.3%) missing values	Missing
ASECWTH has 30160 (55.1%) missing values	Missing
WTFINL has 24577 (44.9%) missing values	Missing
ASECWT has 30160 (55.1%) missing values	Missing
JOBCERT has 24577 (44.9%) missing values	Missing

1 2 3 4

Variables

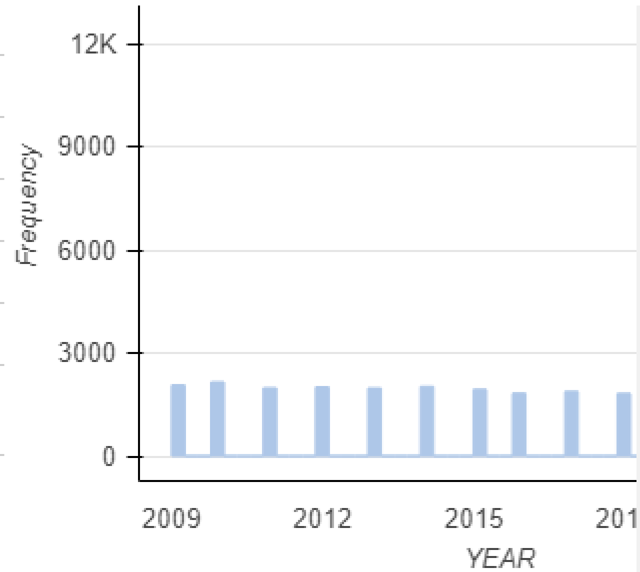
Sort by Feature order ☐ Reverse order

YEAR

Approximate

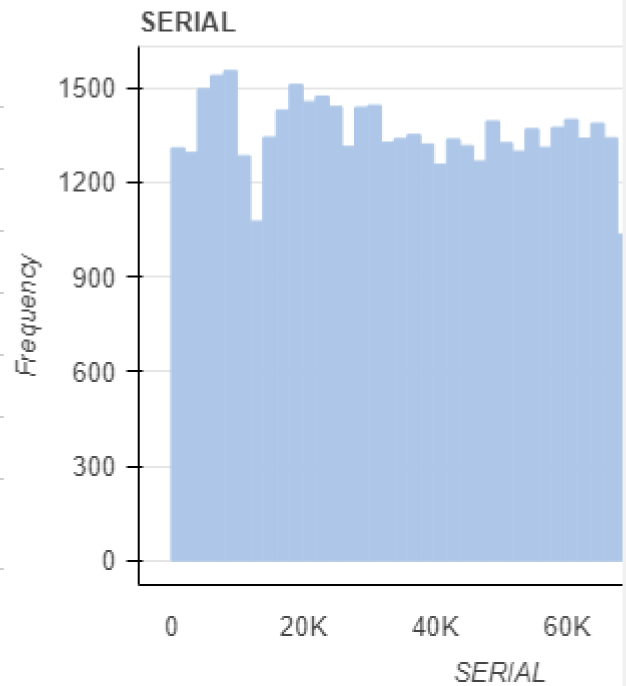
istinct ount	14
pproximate nique (%)	0.0%
issing	0
issing (%)	0.0%
nfinite	0
nfinite (%)	0.0%
emory Size	855.3 KB

Mean	2017.8782
Minimum	2009
Maximum	2022
Zeros	0
Zeros (%)	0.0%
Negatives	0
Negatives (%)	0.0%



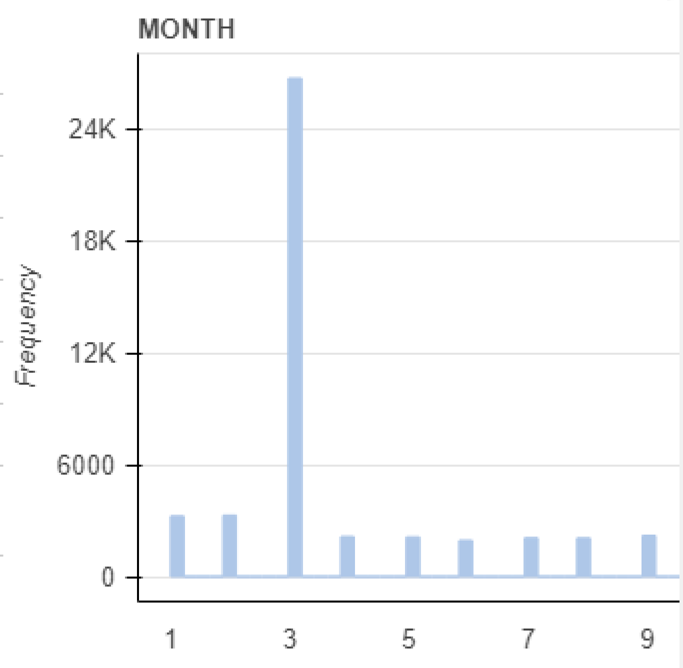
pproximate istinct ount	40584
pproximate nique (%)	74.1%
issing	0
issing (%)	0.0%
finite	0
finite (%)	0.0%
emory Size	855.3 KB

Mean	40716.2568
Minimum	5
Maximum	99417
Zeros	0
Zeros (%)	0.0%
Negatives	0
Negatives (%)	0.0%



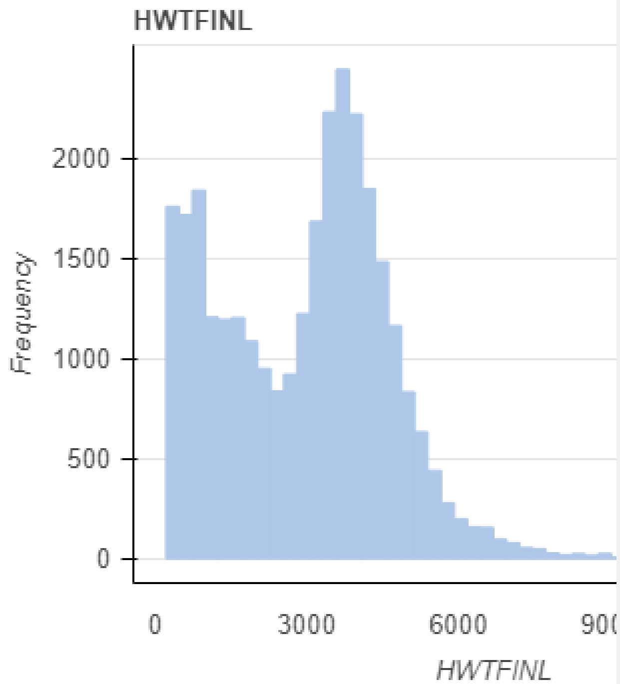
Approximate Distinct Count	12
Approximate Unique (%)	0.0%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Memory Size	855.3 KB

Mean	4.9545
Minimum	1
Maximum	12
Zeros	0
Zeros (%)	0.0%
Negatives	0
Negatives (%)	0.0%



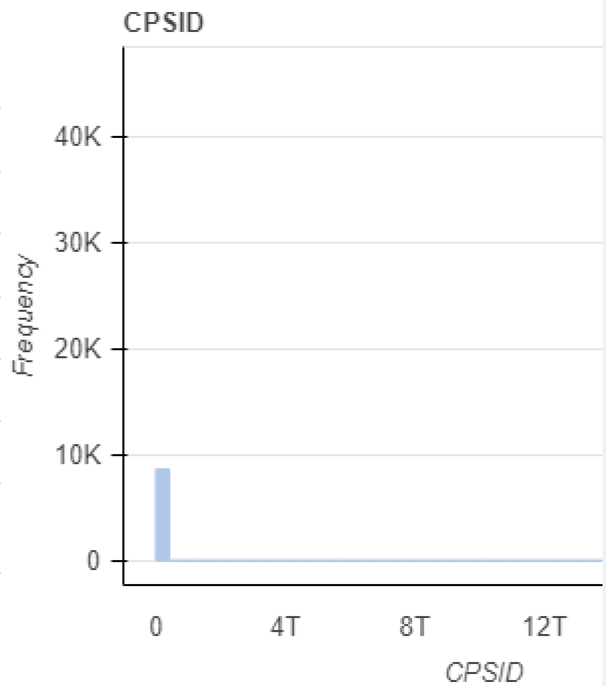
pproximate istinct ount	29375
pproximate nique (%)	97.4%
issing	24577
issing (%)	44.9%
finite	0
finite (%)	0.0%
emory Size	471.2 KB

Mean	2969.6539
Minimum	212.2263
Maximum	13173.7325
Zeros	0
Zeros (%)	0.0%
Negatives	0
Negatives (%)	0.0%



pproximate stinct punt	43204
pproximate nique (%)	78.9%
issing	0
issing (%)	0.0%
finite	0
finite (%)	0.0%
emory Size	855.3 KB

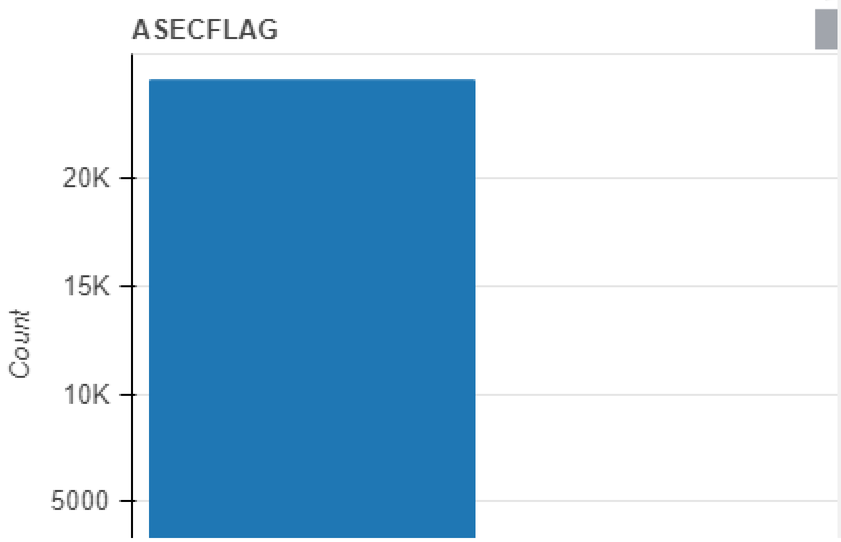
Mean	1.7002×10^{13}
Minimum	0
Maximum	2.022×10^{13}
Zeros	8617
Zeros (%)	15.7%
Negatives	0
Negatives (%)	0.0%

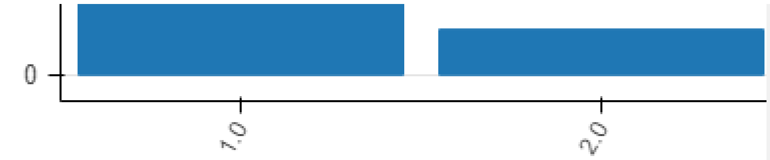


ASECFLAG
categorical

Show Details

Approximate Distinct Count	2
Approximate Unique (%)	0.0%
Missing	28042
Missing (%)	51.2%
Memory Size	1.7 MB

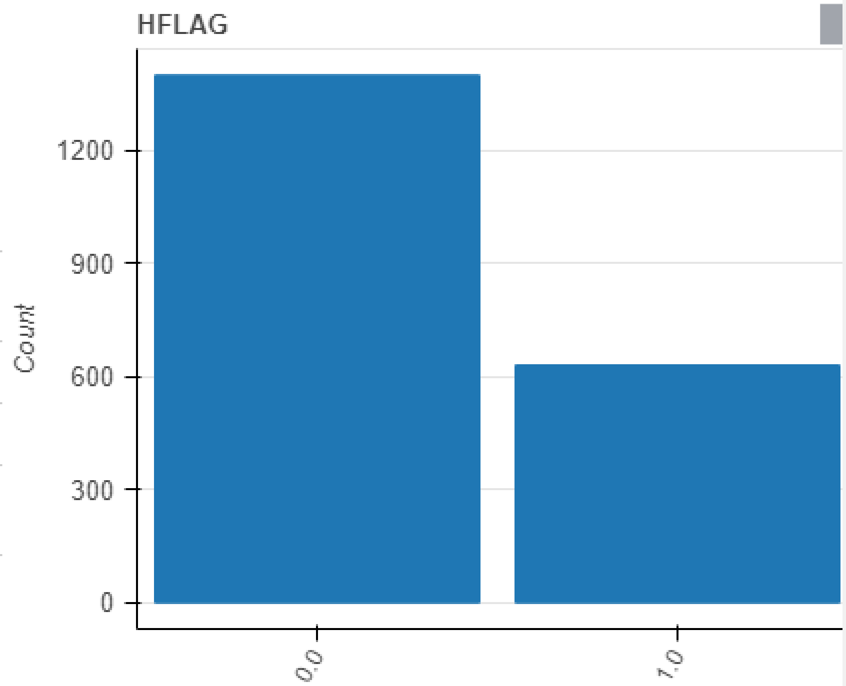




HFLAG
categorical

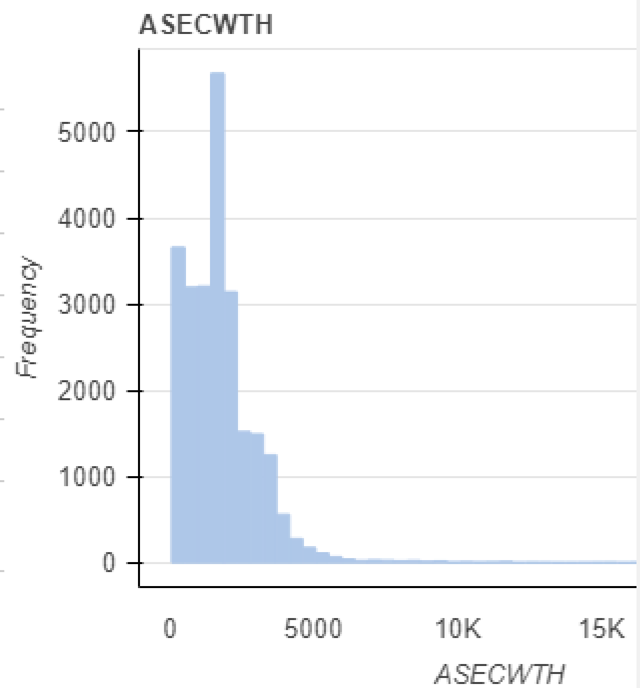
Show Details

Approximate Distinct Count	2
Approximate Unique (%)	0.1%
Missing	52709
Missing (%)	96.3%
Memory Size	134.7 KB



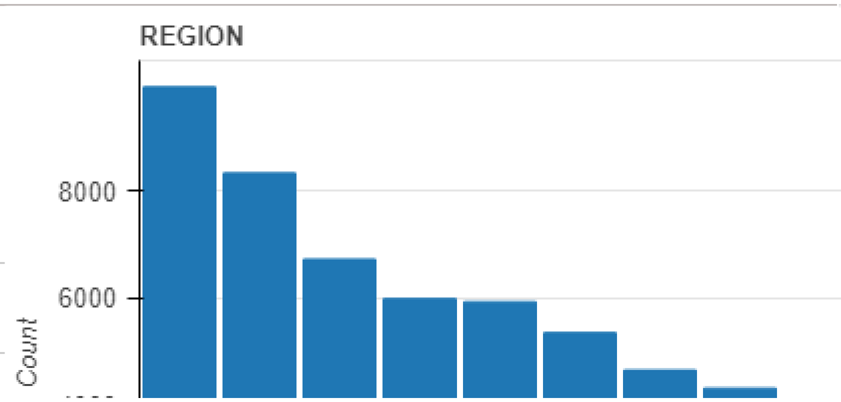
Approximate Distinct Count	22998
Approximate Unique (%)	93.6%
Missing	30160
Missing (%)	55.1%
Infinite	0
Infinite (%)	0.0%
Memory Size	384.0 KB

Mean	1763.6123
Minimum	52.51
Maximum	22869.35
Zeros	0
Zeros (%)	0.0%
Negatives	0
Negatives (%)	0.0%



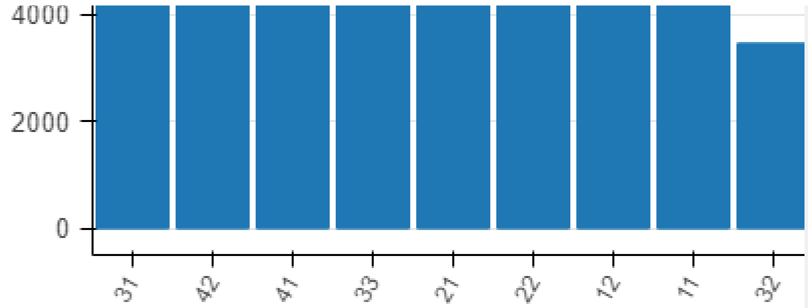
REGION
categorical

Approximate Distinct Count	9
Approximate Unique (%)	0.0%
Missing	0



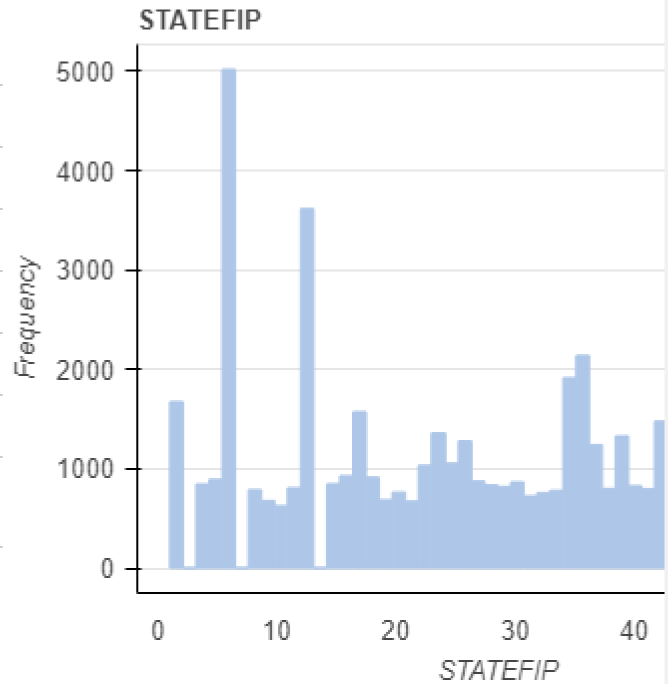
Show Details

Missing (%)	0.0%
Memory Size	3.5 MB



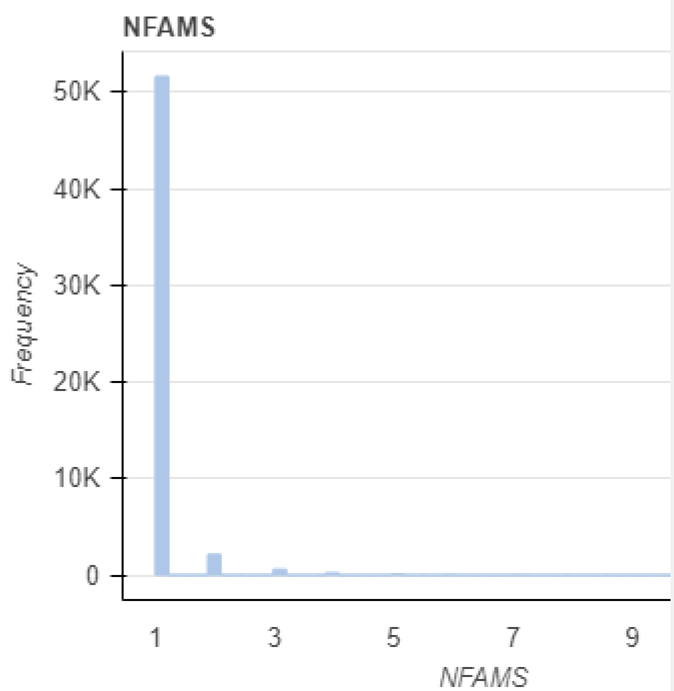
Approximate Distinct Count	51
Approximate Unique (%)	0.1%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Memory Size	855.3 KB

Mean	27.9278
Minimum	1
Maximum	56
Zeros	0
Zeros (%)	0.0%
Negatives	0
Negatives (%)	0.0%



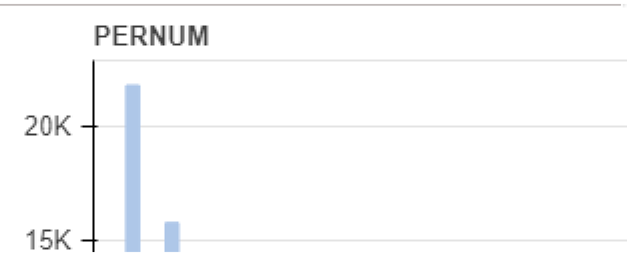
Approximate Distinct Count	11
Approximate Unique (%)	0.0%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Memory Size	855.3 KB

Mean	1.0848
Minimum	1
Maximum	12
Zeros	0
Zeros (%)	0.0%
Negatives	0
Negatives (%)	0.0%



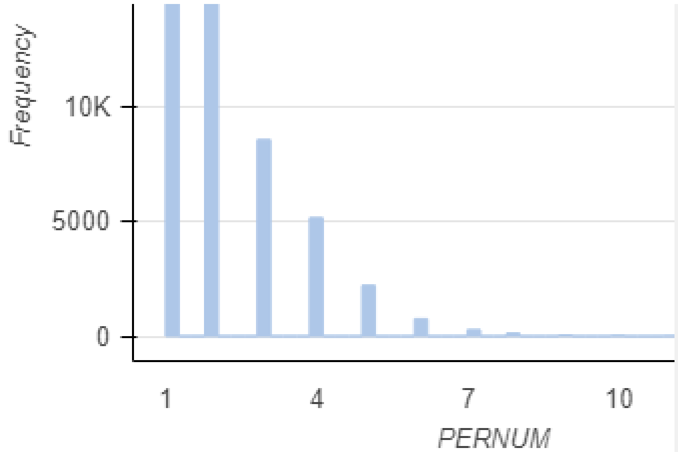
Approximate Distinct Count	14
Approximate	0.0%

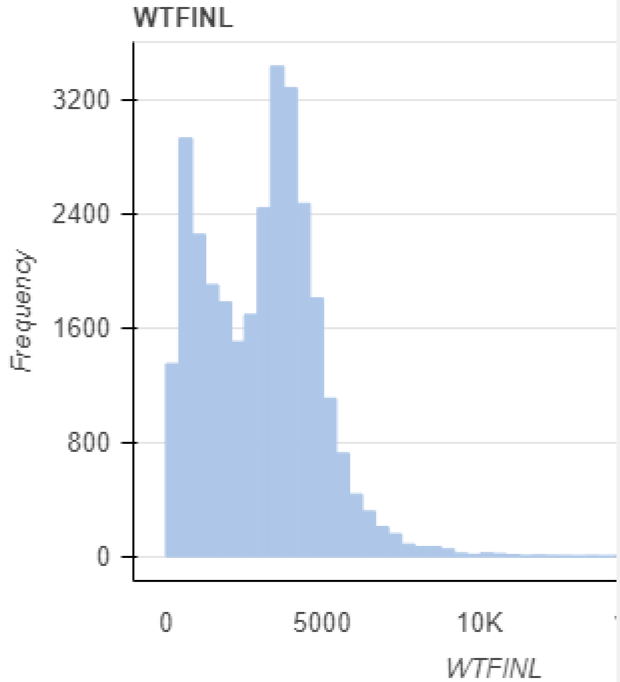
Mean	2.1785
Minimum	1
Maximum	14



Unique (%)	0.0%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Memory Size	855.3 KB

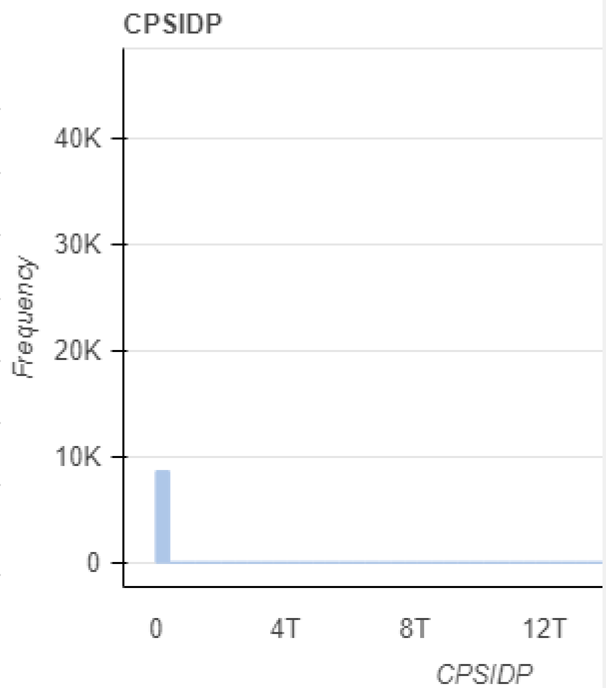
Maximum	14
Zeros	0
Zeros (%)	0.0%
Negatives	0
Negatives (%)	0.0%





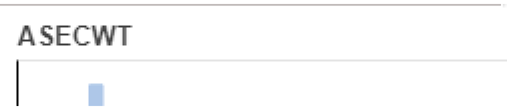
pproximate istinct ount	29870
pproximate nique (%)	99.0%
issing	24577
issing (%)	44.9%
finite	0
finite (%)	0.0%
emory Size	471.2 KB

Mean	3041.3939
Minimum	0
Maximum	20917.4515
Zeros	85
Zeros (%)	0.2%
Negatives	0
Negatives (%)	0.0%



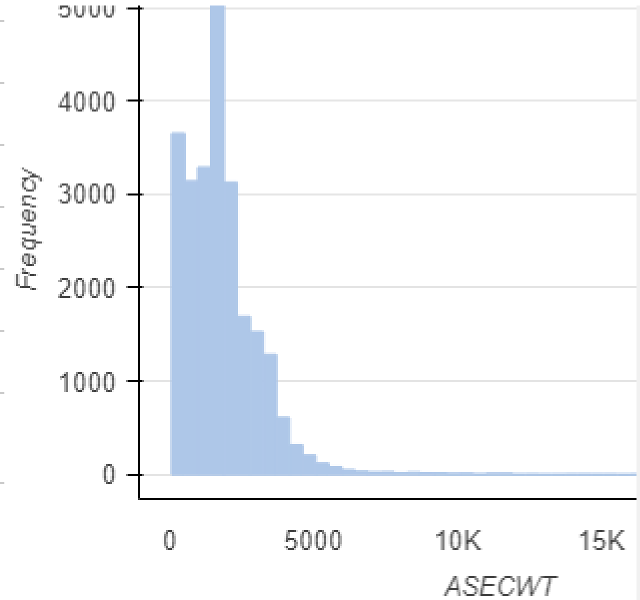
oproximate stinct ount	45306
oproximate nique (%)	82.8%
issing	0
issing (%)	0.0%
finite	0
finite (%)	0.0%
emory Size	855.3 KB

Mean	1.7002×10^{13}
Minimum	0
Maximum	2.022×10^{13}
Zeros	8617
Zeros (%)	15.7%
Negatives	0
Negatives (%)	0.0%



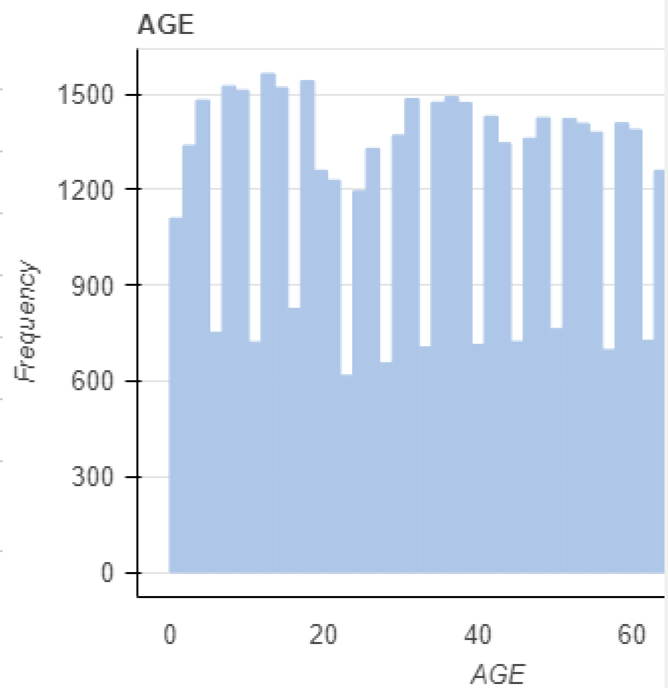
Approximate Distinct Count	23483
Approximate Unique (%)	95.5%
Missing	30160
Missing (%)	55.1%
Infinite	0
Infinite (%)	0.0%
Memory Size	384.0 KB

Mean	1802.5003
Minimum	54.58
Maximum	22869.35
Zeros	0
Zeros (%)	0.0%
Negatives	0
Negatives (%)	0.0%



Approximate Distinct Count	82
Approximate Unique (%)	0.1%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Memory Size	855.3 KB

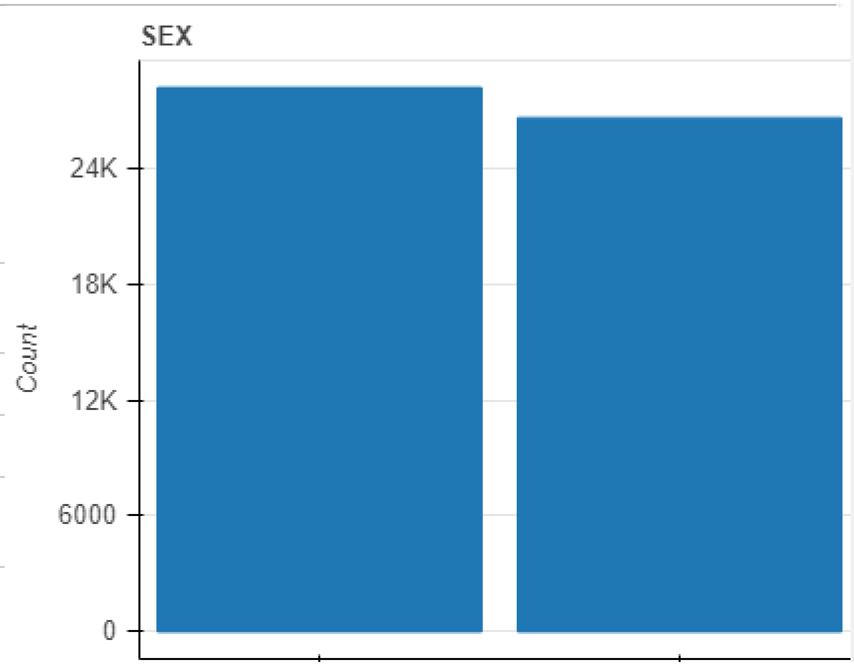
Mean	38.8724
Minimum	0
Maximum	85
Zeros	503
Zeros (%)	0.9%
Negatives	0
Negatives (%)	0.0%



SEX
categorical

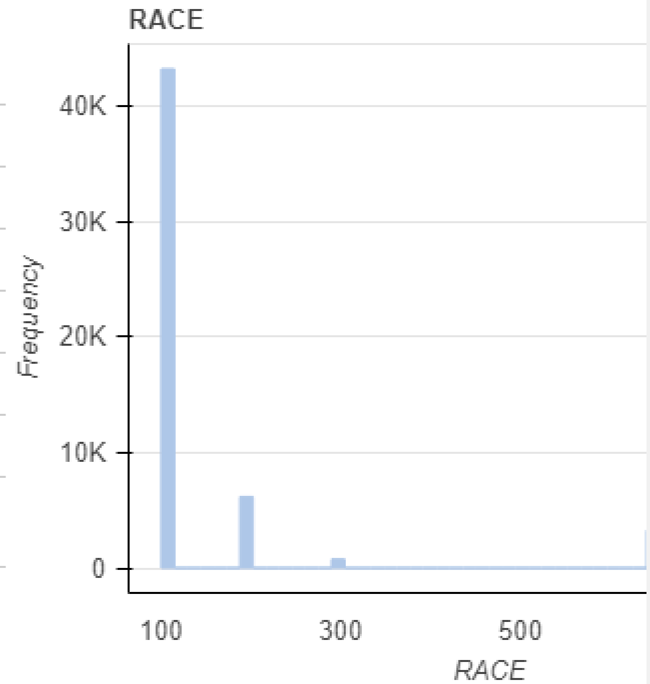
Show Details

Approximate Distinct Count	2
Approximate Unique (%)	0.0%
Missing	0
Missing (%)	0.0%
Memory Size	3.4 MB



Approximate Distinct Count	**24**
Approximate Unique (%)	**0.0%**
Missing	**0**
Missing (%)	**0.0%**
Infinite	**0**
Infinite (%)	**0.0%**
Memory Size	**855.3 KB**

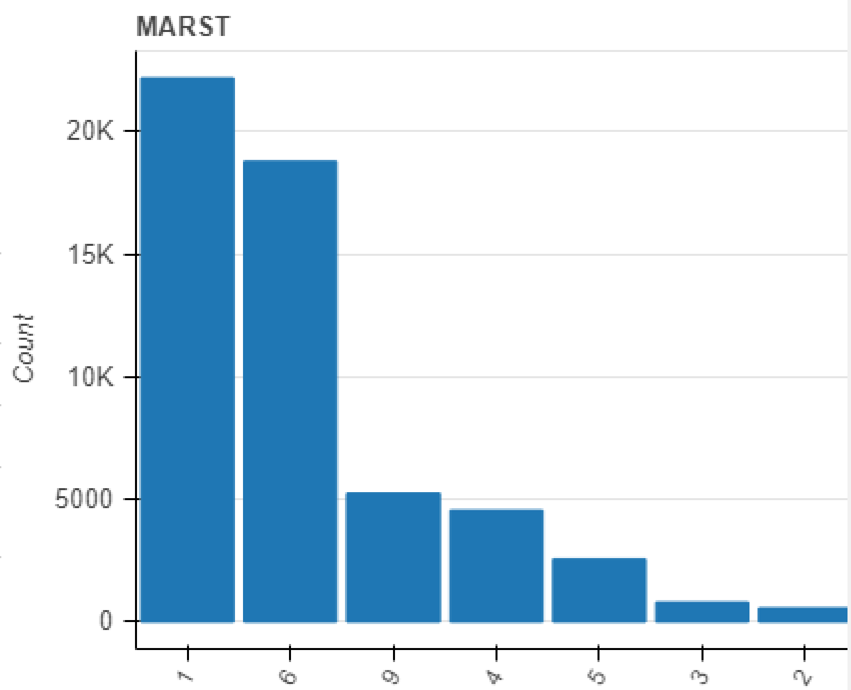
Mean	**163.7212**
Minimum	**100**
Maximum	**830**
Zeros	**0**
Zeros (%)	**0.0%**
Negatives	**0**
Negatives (%)	**0.0%**



MARST
categorical

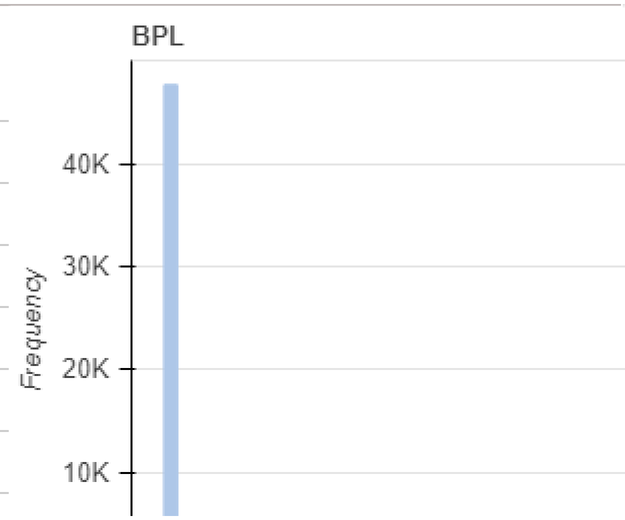
Show Details

Approximate Distinct Count	**7**
Approximate Unique (%)	**0.0%**
Missing	**0**
Missing (%)	**0.0%**
Memory Size	**3.4 MB**



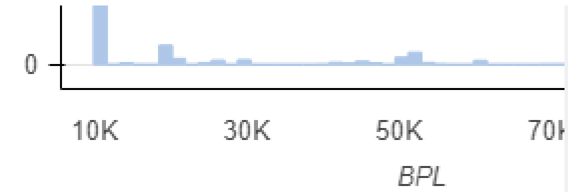
Approximate Distinct Count	**161**
Approximate Unique (%)	**0.3%**
Missing	**0**
Missing (%)	**0.0%**
Infinite	**0**

Mean	**13279.508**
Minimum	**9900**
Maximum	**96000**
Zeros	**0**
Zeros (%)	**0.0%**
Negatives	**0**



Infinite (%)	0.0%
Memory Size	855.3 KB

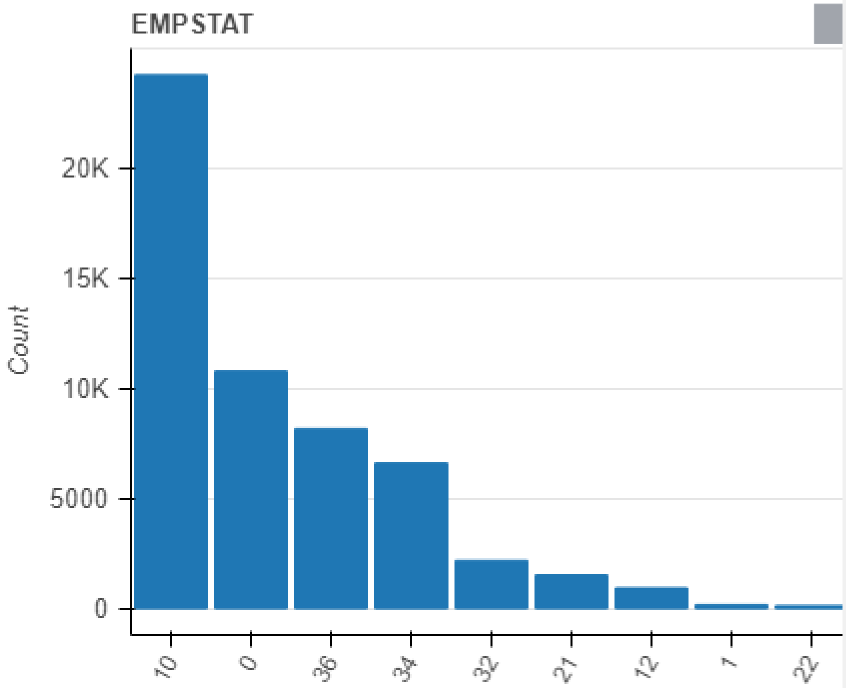
Negatives (%)	0.0%



EMPSTAT
categorical

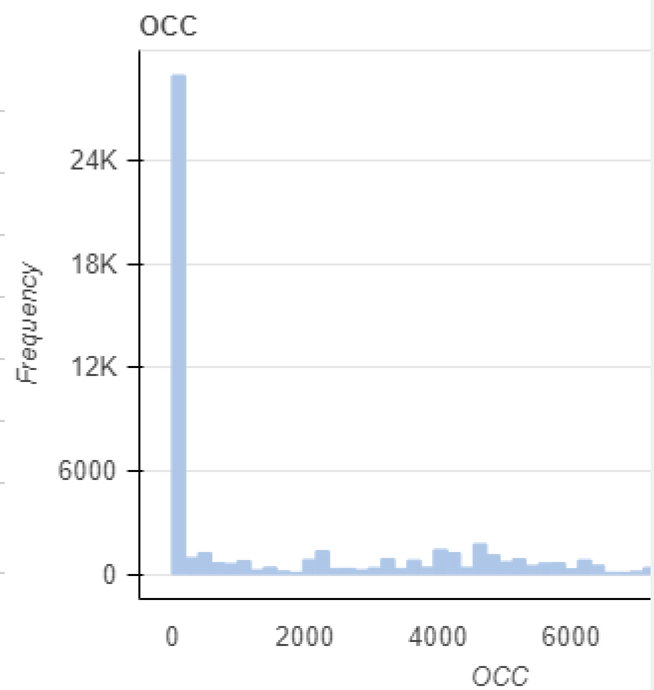
Show Details

Approximate Distinct Count	9
Approximate Unique (%)	0.0%
Missing	0
Missing (%)	0.0%
Memory Size	3.5 MB



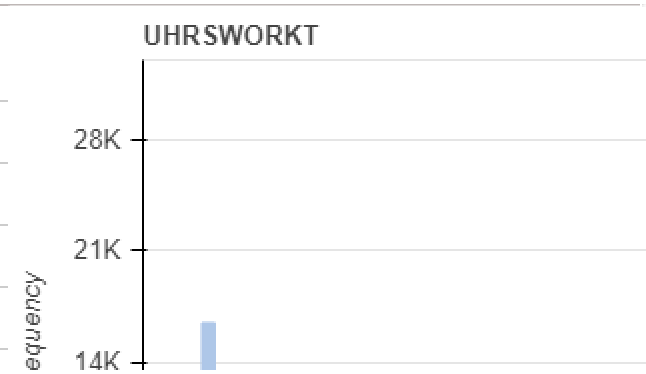
Approximate Distinct Count	654
Approximate Unique (%)	1.2%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Memory Size	855.3 KB

Mean	2021.451
Minimum	0
Maximum	9840
Zeros	27716
Zeros (%)	50.6%
Negatives	0
Negatives (%)	0.0%



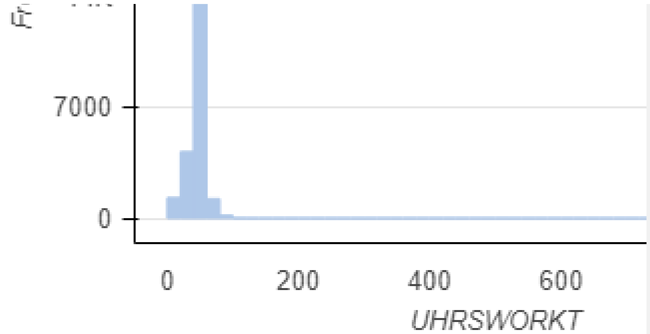
Approximate Distinct Count	100
Approximate Unique (%)	0.2%
Missing	0

Mean	590.1611
Minimum	0
Maximum	999
Zeros	18



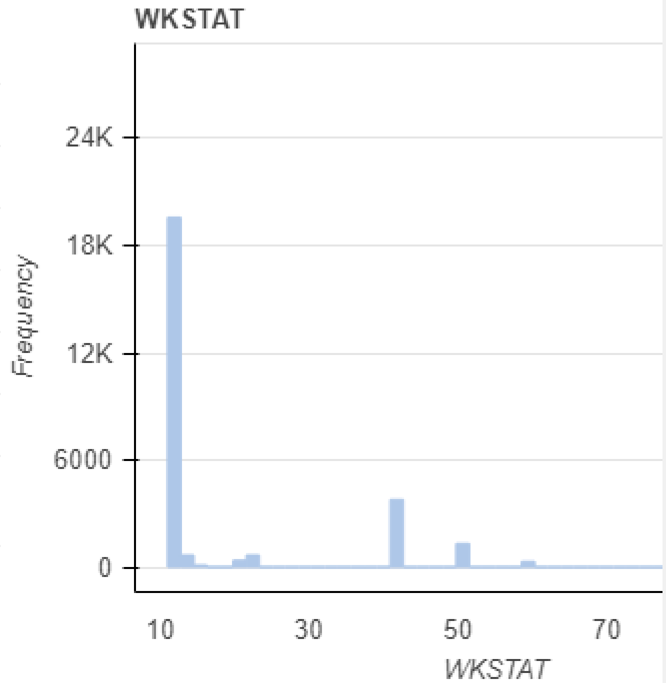
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Memory Size	855.3 KB

Zeros (%)	0.0%
Negatives	0
Negatives (%)	0.0%



Approximate Distinct Count	12
Approximate Unique (%)	0.0%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Memory Size	855.3 KB

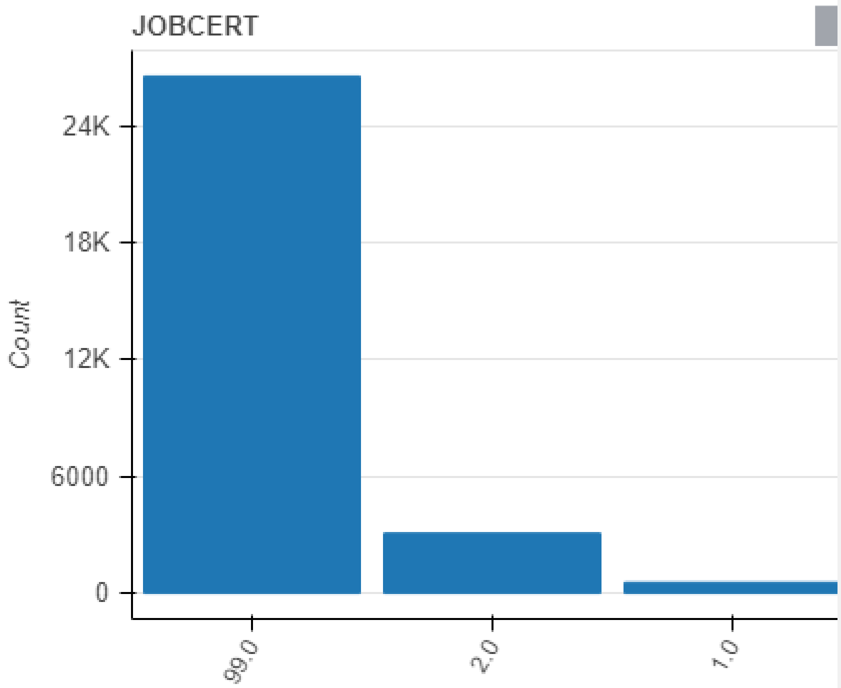
Mean	59.4543
Minimum	11
Maximum	99
Zeros	0
Zeros (%)	0.0%
Negatives	0
Negatives (%)	0.0%



JOBCERT
categorical

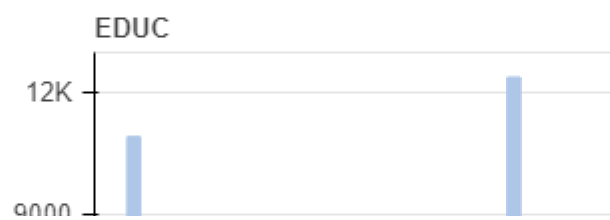
Show Details

Approximate Distinct Count	3
Approximate Unique (%)	0.0%
Missing	24577
Missing (%)	44.9%
Memory Size	2.0 MB



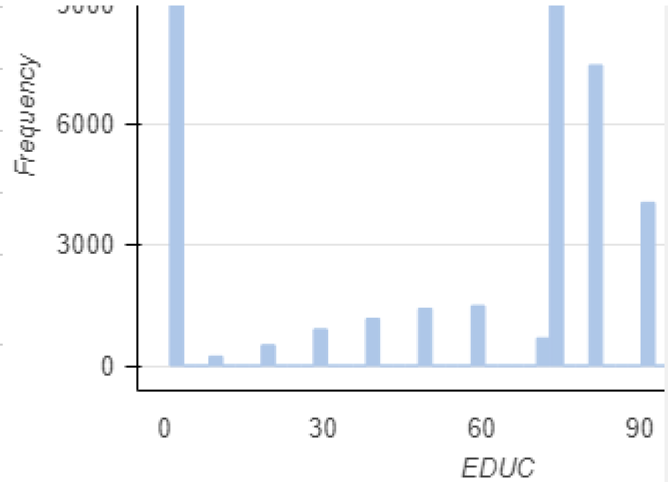
Approximate Distinct Count	17

Mean	68.4491
Minimum	1



Approximate Unique (%)	0.0%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Memory Size	855.3 KB

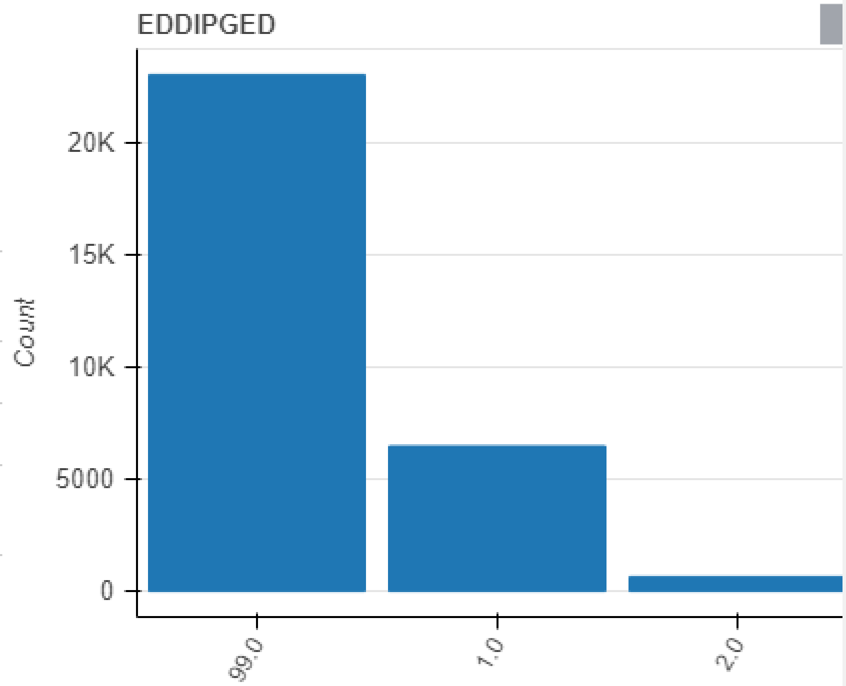
Maximum	125
Zeros	0
Zeros (%)	0.0%
Negatives	0
Negatives (%)	0.0%



EDDIPGED
categorical

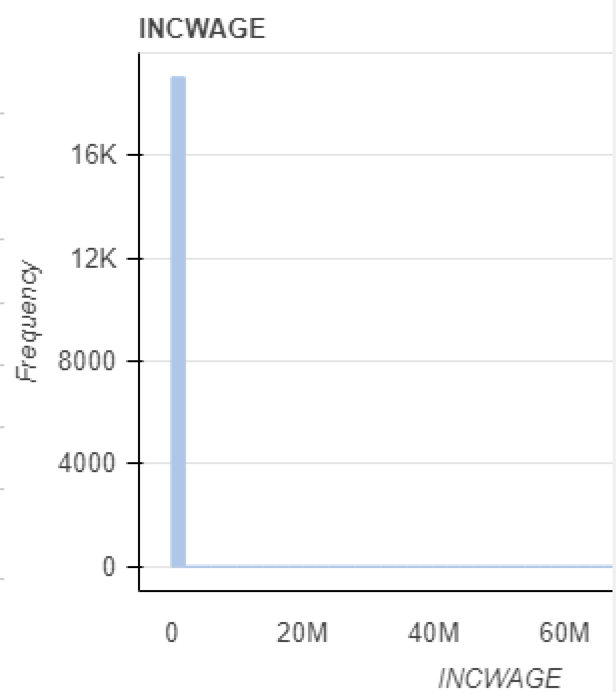
Show Details

Approximate Distinct Count	3
Approximate Unique (%)	0.0%
Missing	24577
Missing (%)	44.9%
Memory Size	2.0 MB



pproximate stinct unt	1450
pproximate ique (%)	5.9%
ssing	30160
ssing (%)	55.1%
inite	0
inite (%)	0.0%
mory Size	384.0 KB

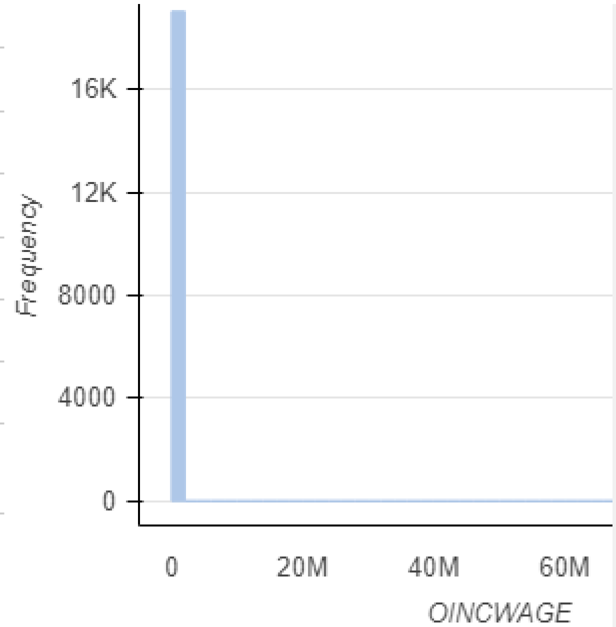
Mean	2.2576×10^{07}
Minimum	0
Maximum	1×10^{08}
Zeros	7496
Zeros (%)	13.7%
Negatives	0
Negatives (%)	0.0%



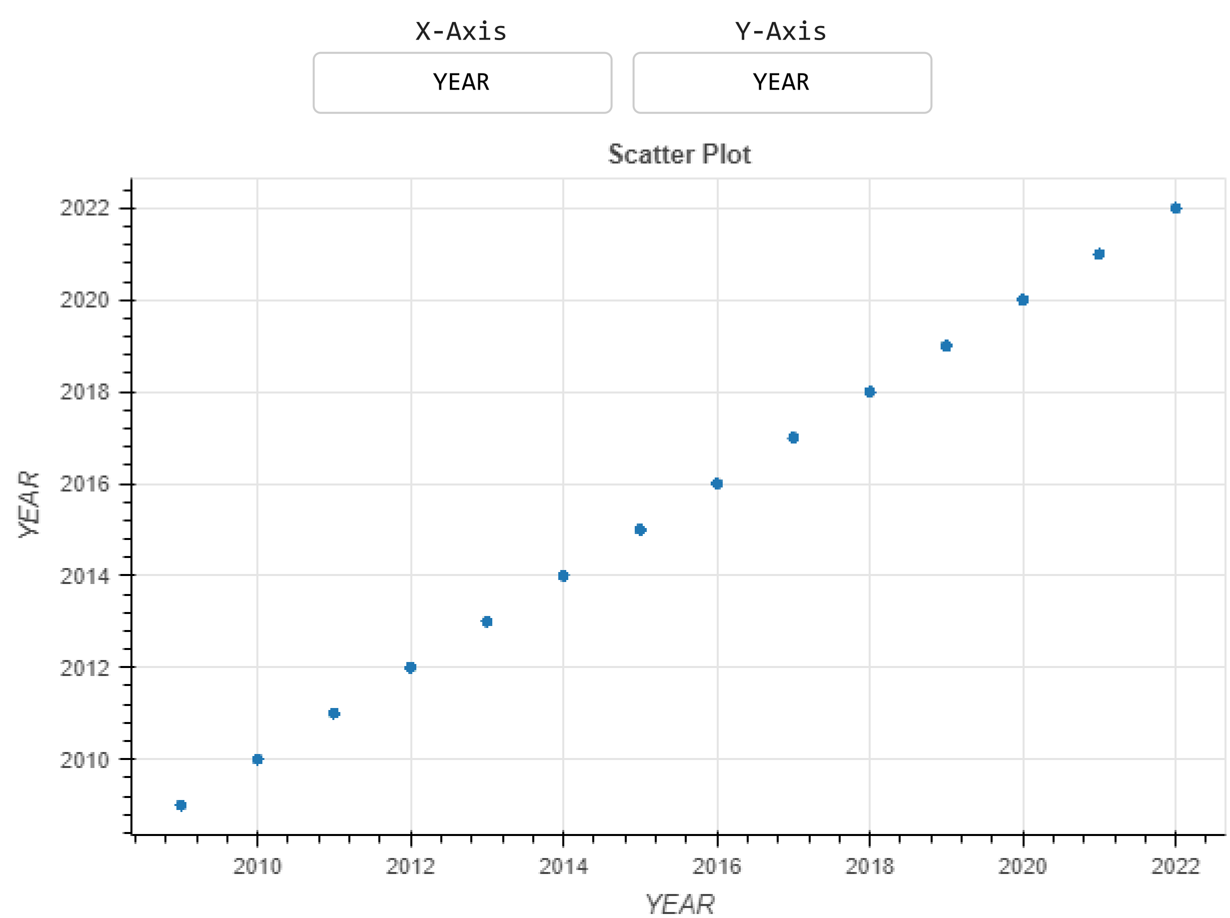
OINCWAGE

proximate stinct unt	278
proximate ique (%)	1.1%
ssing	30160
ssing (%)	55.1%
inite	0
inite (%)	0.0%
mory Size	384.0 KB

Mean	2.2554×10^{07}
Minimum	0
Maximum	1×10^{08}
Zeros	17838
Zeros (%)	32.6%
Negatives	0
Negatives (%)	0.0%



Interactions



Correlations

Cleaning the data

```
train.dropna()
```

YEAR	SERIAL	MONTH	HWTFINL	CPSID	ASECFLAG	HFLAG	ASECWTH	REGION	STATEFIP	...

0 rows × 29 columns

What about Categorical variables ? => Ordinal encoder

```
# Handling categorical attributes
from sklearn.preprocessing import OrdinalEncoder
ordinal_encoder = OrdinalEncoder()
train['EDDIPGED'] = ordinal_encoder.fit_transform(train[['EDDIPGED']])
train['JOBCERT'] = ordinal_encoder.fit_transform(train[['JOBCERT']])
train['EMPSTAT'] = ordinal_encoder.fit_transform(train[['EMPSTAT']])
train['MARST'] = ordinal_encoder.fit_transform(train[['MARST']])
train['SEX'] = ordinal_encoder.fit_transform(train[['SEX']])
train['REGION'] = ordinal_encoder.fit_transform(train[['REGION']])
train['HFLAG'] = ordinal_encoder.fit_transform(train[['HFLAG']])
train['ASECFLAG'] = ordinal_encoder.fit_transform(train[['ASECFLAG']])
```

```
df=train
```

```
#Dropping non correlated columns to INCWAGE
l=['YEAR', 'SERIAL','REGION', 'STATEFIP', 'NFAMS', 'SEX', 'RACE', 'BPL', 'HFLAG',]
df.drop(l, inplace=True, axis=1)
```

```
#Removing NAN values
```

```
for i in df.columns:
  vals = pd.to_numeric(df[i], errors='coerce')
  df[i] = vals.fillna(vals.mean())
```

```
df
```

	MONTH	HWTFINL	CPSID	ASECFLAG	ASECWTH	PERNUM	WTF
0	3	2969.653939	0	0.000000	1671.320000	1	3041.393
1	6	1930.450500	20200504991400	0.079341	1763.612314	3	1718.642
2	3	2969.653939	0	0.000000	433.890000	2	3041.393
3	3	2969.653939	20091201475800	0.000000	2996.700000	1	3041.393
4	3	2969.653939	20170102597500	0.000000	1719.160000	3	3041.393
...	...	...	...	...	...	...	
54732	10	766.129300	20191004211300	0.079341	1763.612314	1	766.129
54733	3	2969.653939	20100300901100	0.000000	1477.030000	1	3041.393
54734	11	873.029400	20201006551800	0.079341	1763.612314	3	1468.447
54735	12	4327.645400	20201005723900	0.079341	1763.612314	2	4464.994
54736	3	2969.653939	0	0.000000	570.480000	2	3041.393

54737 rows × 20 columns

```
df = df[df.get("INCWAGE") != 0]
```

```
vals = pd.to_numeric(df['INCWAGE'], errors='coerce')
vals
```

```
0        1.000000e+05
1        2.257584e+07
5        1.000000e+08
6        2.257584e+07
7        2.257584e+07
             ...
54732    2.257584e+07
54733    3.600000e+03
54734    2.257584e+07
54735    2.257584e+07
54736    1.000000e+08
Name: INCWAGE, Length: 47241, dtype: float64
```

The target variable is very unbalanced so you need to figure out how to fight overfitting in this problem

For now we will use only the numerical features

```
df.columns
```

```
Index(['MONTH', 'HWTFINL', 'CPSID', 'ASECFLAG', 'ASECWTH', 'PERNUM', 'WTFINL',
       'CPSIDP', 'ASECWT', 'AGE', 'MARST', 'EMPSTAT', 'OCC', 'UHRSWORKT',
       'WKSTAT', 'JOBCERT', 'EDUC', 'EDDIPGED', 'INCWAGE', 'OINCWAGE'],
      dtype='object')
```

```
main_cols=['MONTH', 'HWTFINL', 'CPSID', 'ASECFLAG', 'ASECWTH', 'PERNUM', 'WTFINL',
       'ASECWT', 'AGE', 'MARST', 'EMPSTAT', 'OCC', 'UHRSWORKT', 'WKSTAT',
       'JOBCERT', 'EDUC', 'EDDIPGED']
```

```
X=df[main_cols]
y=df.INCWAGE
```

```
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.33,random_state=42)
```

```
from sklearn.ensemble import RandomForestRegressor
clf=RandomForestRegressor(max_depth=100, random_state=101,
                          max_features=None, min_samples_leaf=15)
clf.fit(X_train,y_train)
```

```
RandomForestRegressor(max_depth=100, max_features=None, min_samples_leaf=15,
                      random_state=101)
```

```
y_pred=clf.predict(X_test)
print(y_pred)
```

```
[4.54636958e+04 2.08239721e+04 2.25758421e+07 ... 9.99999990e+07
 2.25758421e+07 9.99999990e+07]
```

```
print(clf.score(X_train, y_train))
print(clf.score(X_test, y_test))
```

```
0.9999993441476379
0.999999255366242
```

0s completed at 9:35 PM