```
import numpy as np
import pandas as pd
from subprocess import check_output

central_df = pd.read_csv("ca-central-1.csv")

central_df.head()
```

```
   2017-05-06 17:29:01    c4.large  Linux/UNIX  ca-central-1a  0.0139
0  2017-05-06 17:29:01  m4.4xlarge     Windows  ca-central-1b  0.8328
1  2017-05-06 17:29:00  m4.4xlarge  Linux/UNIX  ca-central-1b  0.1051
2  2017-05-06 17:29:00  m4.2xlarge     Windows  ca-central-1b  0.4152
3  2017-05-06 17:29:00  m4.2xlarge  Linux/UNIX  ca-central-1b  0.0532
4  2017-05-06 17:28:49  m4.4xlarge  Linux/UNIX  ca-central-1b  0.1060
```

```
central_df.columns = ['datetime','os','instance_type','region','price']

central_df.head()
```

```
              datetime          os instance_type  region   price
0  2017-05-06 17:29:01  m4.4xlarge       Windows       1  0.8328
1  2017-05-06 17:29:00  m4.4xlarge    Linux/UNIX       1  0.1051
2  2017-05-06 17:29:00  m4.2xlarge       Windows       1  0.4152
3  2017-05-06 17:29:00  m4.2xlarge    Linux/UNIX       1  0.0532
4  2017-05-06 17:28:49  m4.4xlarge    Linux/UNIX       1  0.1060
```

```
#east_df.head()
central_df.dropna(inplace=True)

from sklearn.preprocessing import LabelEncoder
labelencoder = LabelEncoder()
central_df['price'] = labelencoder.fit_transform(central_df['price'])
central_df.head()
```

```
              datetime          os instance_type  region  price
0  2017-05-06 17:29:01  m4.4xlarge       Windows       1   5971
1  2017-05-06 17:29:00  m4.4xlarge    Linux/UNIX       1   1450
2  2017-05-06 17:29:00  m4.2xlarge       Windows       1   4451
3  2017-05-06 17:29:00  m4.2xlarge    Linux/UNIX       1    428
4  2017-05-06 17:28:49  m4.4xlarge    Linux/UNIX       1   1464
```

```
X1 = central_df.drop(['price','datetime'],axis=1)
central_df2  = pd.get_dummies(X1)
X1 = central_df2.values
y1 = central_df['price'].values

central_df.head()
```

```
              datetime          os instance_type  region  price
0  2017-05-06 17:29:01  m4.4xlarge       Windows       1   5971
1  2017-05-06 17:29:00  m4.4xlarge    Linux/UNIX       1   1450
2  2017-05-06 17:29:00  m4.2xlarge       Windows       1   4451
```

```
3  2017-05-06 17:29:00  m4.2xlarge    Linux/UNIX       1    428
4  2017-05-06 17:28:49  m4.4xlarge    Linux/UNIX       1   1464
```

```
central_df['instance_type'].value_counts()
```

```
instance_type
Linux/UNIX    549422
Windows       332103
Name: count, dtype: int64
```

```
central_df['os'].value_counts()
```

```
os
m4.large       188057
c4.large       144251
m4.2xlarge     130601
c4.xlarge       90271
m4.4xlarge      85187
m4.xlarge       68481
c4.2xlarge      26618
c4.8xlarge      23293
c4.4xlarge      17380
r4.large        12308
d2.xlarge       11657
m4.10xlarge      9471
m4.16xlarge      8644
r4.4xlarge       7765
r4.2xlarge       7141
i3.8xlarge       6783
r4.8xlarge       5986
d2.2xlarge       5053
r4.16xlarge      4325
x1.16xlarge      4108
i3.4xlarge       3824
d2.8xlarge       3075
i3.large         3007
i3.xlarge        2942
i3.2xlarge       2502
r4.xlarge        2463
d2.4xlarge       2293
i3.16xlarge      2101
x1.32xlarge      1938
Name: count, dtype: int64
```

```
central_df['price'].value_counts()
```

```
price
18      19420
19      18569
17      18135
1415    17701
```

```
20      17603
        ...
2476        1
7562        1
6929        1
5674        1
6279        1
Name: count, Length: 9122, dtype: int64
```

```
central_df2.head()
```

```
   region  os_c4.2xlarge  os_c4.4xlarge  os_c4.8xlarge  os_c4.large  \
0       1          False          False          False        False
1       1          False          False          False        False
2       1          False          False          False        False
3       1          False          False          False        False
4       1          False          False          False        False

   os_c4.xlarge  os_d2.2xlarge  os_d2.4xlarge  os_d2.8xlarge  os_d2.xlarge  \
0         False          False          False          False         False
1         False          False          False          False         False
2         False          False          False          False         False
3         False          False          False          False         False
4         False          False          False          False         False

   ...  os_r4.16xlarge  os_r4.2xlarge  os_r4.4xlarge  os_r4.8xlarge  \
0  ...           False          False          False          False
1  ...           False          False          False          False
2  ...           False          False          False          False
3  ...           False          False          False          False
4  ...           False          False          False          False

   os_r4.large  os_r4.xlarge  os_x1.16xlarge  os_x1.32xlarge  \
0        False         False           False           False
1        False         False           False           False
2        False         False           False           False
3        False         False           False           False
4        False         False           False           False

   instance_type_Linux/UNIX  instance_type_Windows
0                     False                   True
1                      True                  False
2                     False                   True
3                      True                  False
4                      True                  False

[5 rows x 32 columns]
```

```
from sklearn.model_selection import train_test_split
```

```
X1_train, X1_test, y1_train, y1_test = train_test_split(X1, y1,
test_size=0.2,random_state=42 )

from sklearn.linear_model import SGDRegressor
clf1 = SGDRegressor()
clf1.fit(X1_train,y1_train)

y1_rbf = clf1.predict(X1_test)
```

Random Forests

```
from sklearn.ensemble import RandomForestRegressor
clf = RandomForestRegressor(max_depth=2, random_state=0)

clf.fit(X1_train,y1_train)

Pred = clf.predict(X1_test)

print(Pred)
```

```
[3358.96139489  667.13528485 3358.96139489 ...  667.13528485  667.13528485
 3358.96139489]
```