In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
```

In [2]:
```
data = pd.read_csv('acs_ny.csv')
```

In [3]:
```
data.head()
# x_col = ['Acres'[0], 'FamilyType'[2], 'NumUnits'[7],
#        'OwnRent'[10], 'YearBuilt'[11], 'FoodStamp'[14],
#        'HeatingFuel'[15], 'Language'[17]]
```

Out[3]:

	Acres	FamilyIncome	FamilyType	NumBedrooms	NumChildren	NumPeople	NumRo
0	10-Jan	150	Married	4	1	3	9
1	10-Jan	180	Female Head	3	2	4	6
2	10-Jan	280	Female Head	4	0	2	8
3	10-Jan	330	Female Head	2	1	2	4
4	10-Jan	330	Male Head	3	1	2	5

In [4]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 22745 entries, 0 to 22744
Data columns (total 18 columns):
Acres           22745 non-null object
FamilyIncome    22745 non-null int64
FamilyType      22745 non-null object
NumBedrooms     22745 non-null int64
NumChildren     22745 non-null int64
NumPeople       22745 non-null int64
NumRooms        22745 non-null int64
NumUnits        22745 non-null object
NumVehicles     22745 non-null int64
NumWorkers      22745 non-null int64
OwnRent         22745 non-null object
YearBuilt       22745 non-null object
HouseCosts      22745 non-null int64
ElectricBill    22745 non-null int64
FoodStamp       22745 non-null object
HeatingFuel     22745 non-null object
Insurance       22745 non-null int64
Language        22745 non-null object
dtypes: int64(10), object(8)
memory usage: 3.1+ MB
```

In [5]:
```
data.describe()
```

Out[5]:

	FamilyIncome	NumBedrooms	NumChildren	NumPeople	NumRooms	NumV
count	2.274500e+04	22745.000000	22745.000000	22745.000000	22745.000000	22745.
mean	1.102814e+05	3.385315	0.901165	3.390459	7.174764	2.1126
std	1.004539e+05	1.092477	1.159535	1.407659	2.345623	0.9691
min	5.000000e+01	0.000000	0.000000	2.000000	1.000000	0.0000
25%	5.254000e+04	3.000000	0.000000	2.000000	6.000000	2.0000
50%	8.700000e+04	3.000000	0.000000	3.000000	7.000000	2.0000
75%	1.338000e+05	4.000000	2.000000	4.000000	8.000000	3.0000
max	1.605000e+06	8.000000	12.000000	18.000000	21.000000	6.0000

The data is very clean with no missing values.

In [6]:
```
data.isnull().sum()
```

Out[6]:
```
Acres           0
FamilyIncome    0
FamilyType      0
NumBedrooms     0
NumChildren     0
NumPeople       0
NumRooms        0
NumUnits        0
NumVehicles     0
NumWorkers      0
OwnRent         0
YearBuilt       0
HouseCosts      0
ElectricBill    0
FoodStamp       0
HeatingFuel     0
Insurance       0
Language        0
dtype: int64
```

In [7]:
```
x = data.iloc[:, 2:]
y = data.iloc[:, 1:2]
```

We will make dummy variables for categorical variables: The categorical variables here are: Acres, FamilyType, NumUnits, OwnRent, YearBuilt, FoodStamp, HeatingFuel, Language

In [8]:
```
encoded_x = pd.get_dummies(x,drop_first = True)
```

In [9]:
```
from sklearn.model_selection import train_test_split
X_train, X_test, Y_train, Y_test = train_test_split(encoded_x, y, test_size =
0.2, random_state = 0)
```

In [10]:
```
from sklearn.linear_model import LinearRegression
regression = LinearRegression()
lm = regression.fit(X_train,Y_train)
```

In [11]:
```
y_pred = regression.predict(X_test)
```

In [12]:
```
print(lm.intercept_)
print(lm.coef_)
```

```
[-80800.16615241]
[[  2.48021866e+03   4.67784045e+03  -8.10461495e+03   4.73533561e+03
    6.71082735e+03   1.95796521e+04   2.70678110e+01   5.15552947e+01
    1.82433181e+01   1.02494604e+04   3.03508401e+04   1.13475607e+04
    9.60800367e+03   4.86757724e+04   8.16761059e+03   2.80985179e+03
    8.25876691e+03   6.44835907e+03   6.45777141e+03   1.31873527e+04
    1.68446923e+04   1.12023329e+04   1.93418763e+04   7.97085844e+03
    1.03771375e+04   1.45504044e+04   7.20327516e+03   4.56537299e+04
    6.35914343e+03  -1.21848069e+04   1.52077060e+04   1.62297445e+04
    7.94061885e+03   1.42126280e+04   1.17086936e+04   3.21869565e+03
    8.14541435e+02   6.85317121e+02   1.81225543e+03  -3.44355992e+03
   -1.23035662e+04]]
```

In [13]:
```
import sklearn
```

In [14]:
```
#we will take out the R^2 value
sklearn.metrics.r2_score(Y_test, y_pred)
```

Out[14]: 0.35256801039469621

In [15]:
```
import statsmodels.api as sm
```

```
C:\Program Files (x86)\Microsoft Visual Studio\Shared\Anaconda3_64\lib\site-p
ackages\statsmodels\compat\pandas.py:56: FutureWarning: The pandas.core.datet
ools module is deprecated and will be removed in a future version. Please use
the pandas.tseries module instead.
  from pandas.core import datetools
```

In [16]:
```
X = encoded_x
Y = data.iloc[:, 1:2]
## fit a OLS model with intercept on TV and Radio
X = sm.add_constant(X)
est = sm.OLS(Y, X).fit()

est.summary()
```

Out[16]:

OLS Regression Results

Dep. Variable:	FamilyIncome	**R-squared:**	0.344
Model:	OLS	**Adj. R-squared:**	0.343
Method:	Least Squares	**F-statistic:**	291.0
Date:	Sat, 14 Apr 2018	**Prob (F-statistic):**	0.00
Time:	10:17:50	**Log-Likelihood:**	-2.8944e+05
No. Observations:	22745	**AIC:**	5.790e+05
Df Residuals:	22703	**BIC:**	5.793e+05
Df Model:	41		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-7.314e+04	2.26e+04	-3.237	0.001	-1.17e+05	-2.89e+04
NumBedrooms	2610.5627	673.575	3.876	0.000	1290.309	3930.817
NumChildren	4137.2606	703.547	5.881	0.000	2758.261	5516.260
NumPeople	-7555.3019	650.785	-11.610	0.000	-8830.884	-6279.720
NumRooms	4643.8267	302.838	15.334	0.000	4050.244	5237.410
NumVehicles	6938.4598	678.174	10.231	0.000	5609.192	8267.727
NumWorkers	1.901e+04	784.745	24.228	0.000	1.75e+04	2.06e+04
HouseCosts	27.0085	0.592	45.643	0.000	25.849	28.168
ElectricBill	46.7250	5.670	8.241	0.000	35.612	57.838
Insurance	18.6187	0.680	27.384	0.000	17.286	19.951
FamilyType_Male Head	8882.8225	2804.646	3.167	0.002	3385.524	1.44e+04
FamilyType_Married	3.015e+04	1661.750	18.143	0.000	2.69e+04	3.34e+04
NumUnits_Single attached	1.093e+04	3669.308	2.979	0.003	3740.098	1.81e+04
NumUnits_Single detached	9592.0328	3278.715	2.926	0.003	3165.527	1.6e+04
OwnRent_Outright	5.398e+04	6790.797	7.948	0.000	4.07e+04	6.73e+04
OwnRent_Rented	7495.5002	2011.641	3.726	0.000	3552.545	1.14e+04
YearBuilt_1940-1949	-3670.2014	2.12e+04	-0.173	0.862	-4.52e+04	3.78e+04
YearBuilt_1950-1959	618.4751	2.11e+04	0.029	0.977	-4.08e+04	4.2e+04
YearBuilt_1960-1969	-124.0640	2.11e+04	-0.006	0.995	-4.16e+04	4.13e+04
YearBuilt_1970-1979	-306.6532	2.11e+04	-0.015	0.988	-4.18e+04	4.11e+04
YearBuilt_1980-1989	6833.9315	2.12e+04	0.323	0.747	-3.46e+04	4.83e+04

YearBuilt_1990-1999	8202.1665	2.12e+04	0.388	0.698	-3.33e+04	4.97e+04
YearBuilt_2000-2004	6246.4785	2.12e+04	0.294	0.769	-3.54e+04	4.79e+04
YearBuilt_2005	8598.6384	2.18e+04	0.395	0.693	-3.41e+04	5.13e+04
YearBuilt_2006	7281.1955	2.19e+04	0.332	0.740	-3.57e+04	5.03e+04
YearBuilt_2007	8208.4163	2.21e+04	0.372	0.710	-3.5e+04	5.14e+04
YearBuilt_2008	6459.8614	2.25e+04	0.288	0.774	-3.76e+04	5.05e+04
YearBuilt_2009	-1837.0954	2.29e+04	-0.080	0.936	-4.67e+04	4.3e+04
YearBuilt_2010	3.312e+04	2.41e+04	1.373	0.170	-1.41e+04	8.04e+04
YearBuilt_Before 1939	-551.3734	2.11e+04	-0.026	0.979	-4.19e+04	4.08e+04
FoodStamp_Yes	-1.185e+04	2266.129	-5.228	0.000	-1.63e+04	-7406.278
HeatingFuel_Electricity	1.016e+04	7037.505	1.444	0.149	-3632.872	2.4e+04
HeatingFuel_Gas	1.34e+04	6583.622	2.036	0.042	496.946	2.63e+04
HeatingFuel_None	3925.2730	1.82e+04	0.215	0.830	-3.18e+04	3.97e+04
HeatingFuel_Oil	9834.1494	6625.616	1.484	0.138	-3152.511	2.28e+04
HeatingFuel_Other	8165.9121	8826.806	0.925	0.355	-9135.231	2.55e+04
HeatingFuel_Solar	2076.5958	3.73e+04	0.056	0.956	-7.1e+04	7.52e+04
HeatingFuel_Wood	-2028.0935	6933.553	-0.293	0.770	-1.56e+04	1.16e+04
Language_English	3761.2304	3028.383	1.242	0.214	-2174.608	9697.069
Language_Other	817.7825	5642.628	0.145	0.885	-1.02e+04	1.19e+04
Language_Other European	-2152.2006	3422.223	-0.629	0.529	-8859.993	4555.592
Language_Spanish	-9620.3926	3490.135	-2.756	0.006	-1.65e+04	-2779.489

Omnibus:	16166.760	**Durbin-Watson:**	0.565
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	427635.583
Skew:	3.107	**Prob(JB):**	0.00
Kurtosis:	23.313	**Cond. No.**	3.35e+05

We will start doing backward elimination with p-value. In Model 3, we are going to use backward elimination to delete p-values less than 0.05.

I am going to do group elimination meaning, 0.9 p-value variables will be deleted first, and so on.

In [23]:
```
# From above statsmodel summary, we can find many variables with more than 0.9
 p-value. We will delete them all
encoded_x = encoded_x.drop(['YearBuilt_1950-1959', 'YearBuilt_1960-1969','Year
Built_1970-1979','YearBuilt_2009','YearBuilt_Before 1939','HeatingFuel_Solar'
], axis = 1)
```

In [25]:
```
X = encoded_x
Y = data.iloc[:, 1:2]
## fit a OLS model with intercept on TV and Radio
X = sm.add_constant(X)
est = sm.OLS(Y, X).fit()

est.summary()
```

Out[25]:

OLS Regression Results

Dep. Variable:	FamilyIncome	**R-squared:**	0.344
Model:	OLS	**Adj. R-squared:**	0.343
Method:	Least Squares	**F-statistic:**	340.9
Date:	Sat, 14 Apr 2018	**Prob (F-statistic):**	0.00
Time:	10:27:55	**Log-Likelihood:**	-2.8944e+05
No. Observations:	22745	**AIC:**	5.789e+05
Df Residuals:	22709	**BIC:**	5.792e+05
Df Model:	35		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-7.326e+04	8161.056	-8.977	0.000	-8.93e+04	-5.73e+04
NumBedrooms	2606.7109	673.204	3.872	0.000	1287.185	3926.236
NumChildren	4136.2667	703.158	5.882	0.000	2758.028	5514.505
NumPeople	-7553.1992	650.147	-11.618	0.000	-8827.531	-6278.867
NumRooms	4625.2048	301.270	15.352	0.000	4034.696	5215.714
NumVehicles	6954.1210	675.758	10.291	0.000	5629.589	8278.653
NumWorkers	1.901e+04	784.523	24.234	0.000	1.75e+04	2.05e+04
HouseCosts	27.0376	0.589	45.900	0.000	25.883	28.192
ElectricBill	46.8772	5.659	8.284	0.000	35.786	57.968
Insurance	18.6076	0.679	27.406	0.000	17.277	19.938
FamilyType_Male Head	8856.8573	2803.918	3.159	0.002	3360.985	1.44e+04
FamilyType_Married	3.013e+04	1661.143	18.136	0.000	2.69e+04	3.34e+04
NumUnits_Single attached	1.088e+04	3657.311	2.974	0.003	3707.539	1.8e+04
NumUnits_Single detached	9622.7903	3265.141	2.947	0.003	3222.891	1.6e+04
OwnRent_Outright	5.401e+04	6784.673	7.961	0.000	4.07e+04	6.73e+04
OwnRent_Rented	7472.3563	2010.797	3.716	0.000	3531.056	1.14e+04
YearBuilt_1940-1949	-3532.4866	2047.196	-1.726	0.084	-7545.132	480.158
YearBuilt_1980-1989	6984.9199	1983.495	3.522	0.000	3097.134	1.09e+04
YearBuilt_1990-1999	8352.5958	1977.022	4.225	0.000	4477.498	1.22e+04
YearBuilt_2000-2004	6395.1138	2649.218	2.414	0.016	1202.465	1.16e+04
YearBuilt_2005	8744.1218	5529.436	1.581	0.114	-2093.950	1.96e+04

YearBuilt_2006	7419.7959	6104.340	1.215	0.224	-4545.128	1.94e+04
YearBuilt_2007	8346.4661	6588.079	1.267	0.205	-4566.621	2.13e+04
YearBuilt_2008	6602.6127	7805.566	0.846	0.398	-8696.830	2.19e+04
YearBuilt_2010	3.334e+04	1.17e+04	2.856	0.004	1.05e+04	5.62e+04
FoodStamp_Yes	-1.189e+04	2264.031	-5.253	0.000	-1.63e+04	-7455.847
HeatingFuel_Electricity	1.021e+04	6938.464	1.472	0.141	-3389.883	2.38e+04
HeatingFuel_Gas	1.348e+04	6476.140	2.081	0.037	782.037	2.62e+04
HeatingFuel_None	4105.9152	1.82e+04	0.226	0.821	-3.15e+04	3.97e+04
HeatingFuel_Oil	9939.9156	6518.844	1.525	0.127	-2837.465	2.27e+04
HeatingFuel_Other	8185.2725	8750.160	0.935	0.350	-8965.641	2.53e+04
HeatingFuel_Wood	-2044.2819	6837.102	-0.299	0.765	-1.54e+04	1.14e+04
Language_English	3713.4246	3027.183	1.227	0.220	-2220.062	9646.911
Language_Other	778.7616	5641.416	0.138	0.890	-1.03e+04	1.18e+04
Language_Other European	-2172.6561	3421.497	-0.635	0.525	-8879.024	4533.712
Language_Spanish	-9663.2243	3488.635	-2.770	0.006	-1.65e+04	-2825.261

Omnibus:	16164.354	**Durbin-Watson:**	0.565
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	427463.092
Skew:	3.107	**Prob(JB):**	0.00
Kurtosis:	23.309	**Cond. No.**	8.22e+04

In [26]:
```
#Now we delete all with p= 0.7,
encoded_x = encoded_x.drop(['HeatingFuel_None','Language_Other'],axis = 1)
```

In [27]:
```
X = encoded_x
Y = data.iloc[:, 1:2]
## fit a OLS model with intercept on TV and Radio
X = sm.add_constant(X)
est = sm.OLS(Y, X).fit()

est.summary()
```

Out[27]:

OLS Regression Results

Dep. Variable:	FamilyIncome	**R-squared:**	0.344
Model:	OLS	**Adj. R-squared:**	0.343
Method:	Least Squares	**F-statistic:**	361.6
Date:	Sat, 14 Apr 2018	**Prob (F-statistic):**	0.00
Time:	10:30:07	**Log-Likelihood:**	-2.8944e+05
No. Observations:	22745	**AIC:**	5.789e+05
Df Residuals:	22711	**BIC:**	5.792e+05
Df Model:	33		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-7.255e+04	7704.337	-9.417	0.000	-8.76e+04	-5.74e+04
NumBedrooms	2606.8907	673.169	3.873	0.000	1287.433	3926.348
NumChildren	4139.0083	702.168	5.895	0.000	2762.712	5515.305
NumPeople	-7554.5929	650.034	-11.622	0.000	-8828.704	-6280.482
NumRooms	4626.0234	301.199	15.359	0.000	4035.652	5216.395
NumVehicles	6951.4066	675.612	10.289	0.000	5627.161	8275.652
NumWorkers	1.901e+04	784.459	24.237	0.000	1.75e+04	2.06e+04
HouseCosts	27.0374	0.589	45.938	0.000	25.884	28.191
ElectricBill	46.8964	5.656	8.291	0.000	35.810	57.983
Insurance	18.6120	0.679	27.427	0.000	17.282	19.942
FamilyType_Male Head	8854.2050	2803.781	3.158	0.002	3358.602	1.43e+04
FamilyType_Married	3.013e+04	1661.046	18.138	0.000	2.69e+04	3.34e+04
NumUnits_Single attached	1.088e+04	3657.140	2.975	0.003	3710.094	1.8e+04
NumUnits_Single detached	9621.6224	3264.986	2.947	0.003	3222.026	1.6e+04
OwnRent_Outright	5.402e+04	6784.166	7.963	0.000	4.07e+04	6.73e+04
OwnRent_Rented	7487.9838	2009.845	3.726	0.000	3548.550	1.14e+04
YearBuilt_1940-1949	-3525.0424	2046.899	-1.722	0.085	-7537.104	487.019
YearBuilt_1980-1989	6985.6097	1983.388	3.522	0.000	3098.033	1.09e+04
YearBuilt_1990-1999	8352.8195	1976.926	4.225	0.000	4477.910	1.22e+04
YearBuilt_2000-2004	6385.9146	2648.847	2.411	0.016	1193.992	1.16e+04
YearBuilt_2005	8753.1130	5529.034	1.583	0.113	-2084.173	1.96e+04

YearBuilt_2006	7405.0249	6103.088	1.213	0.225	-4557.445	1.94e+04
YearBuilt_2007	8360.3503	6586.594	1.269	0.204	-4549.826	2.13e+04
YearBuilt_2008	6607.4278	7805.169	0.847	0.397	-8691.238	2.19e+04
YearBuilt_2010	3.331e+04	1.17e+04	2.853	0.004	1.04e+04	5.62e+04
FoodStamp_Yes	-1.19e+04	2263.900	-5.255	0.000	-1.63e+04	-7459.354
HeatingFuel_Electricity	9691.5641	6549.695	1.480	0.139	-3146.287	2.25e+04
HeatingFuel_Gas	1.296e+04	6058.444	2.138	0.032	1080.887	2.48e+04
HeatingFuel_Oil	9420.5550	6104.457	1.543	0.123	-2544.599	2.14e+04
HeatingFuel_Other	7675.4579	8448.321	0.909	0.364	-8883.830	2.42e+04
HeatingFuel_Wood	-2557.7608	6449.836	-0.397	0.692	-1.52e+04	1.01e+04
Language_English	3512.6161	2637.675	1.332	0.183	-1657.407	8682.639
Language_Other European	-2374.4536	3082.942	-0.770	0.441	-8417.230	3668.323
Language_Spanish	-9856.7578	3157.619	-3.122	0.002	-1.6e+04	-3667.609

Omnibus:	16163.977	**Durbin-Watson:**	0.565
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	427429.819
Skew:	3.107	**Prob(JB):**	0.00
Kurtosis:	23.308	**Cond. No.**	6.21e+04

In [28]:

```
#Now we delete all with p= 0.6,
encoded_x = encoded_x.drop(['HeatingFuel_Wood'],axis = 1)
X = encoded_x
Y = data.iloc[:, 1:2]
## fit a OLS model with intercept on TV and Radio
X = sm.add_constant(X)
est = sm.OLS(Y, X).fit()

est.summary()
```

Out[28]:

OLS Regression Results

Dep. Variable:	FamilyIncome	**R-squared:**	0.344
Model:	OLS	**Adj. R-squared:**	0.344
Method:	Least Squares	**F-statistic:**	372.9
Date:	Sat, 14 Apr 2018	**Prob (F-statistic):**	0.00
Time:	10:31:51	**Log-Likelihood:**	-2.8944e+05
No. Observations:	22745	**AIC:**	5.789e+05
Df Residuals:	22712	**BIC:**	5.792e+05
Df Model:	32		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-7.475e+04	5335.131	-14.011	0.000	-8.52e+04	-6.43e+04
NumBedrooms	2603.0819	673.088	3.867	0.000	1283.783	3922.380
NumChildren	4135.6763	702.104	5.890	0.000	2759.504	5511.849
NumPeople	-7554.4911	650.022	-11.622	0.000	-8828.578	-6280.404
NumRooms	4627.3890	301.174	15.364	0.000	4037.067	5217.711
NumVehicles	6948.4491	675.558	10.285	0.000	5624.309	8272.590
NumWorkers	1.902e+04	784.412	24.242	0.000	1.75e+04	2.06e+04
HouseCosts	27.0381	0.589	45.940	0.000	25.884	28.192
ElectricBill	46.9166	5.656	8.296	0.000	35.831	58.002
Insurance	18.6116	0.679	27.427	0.000	17.282	19.942
FamilyType_Male Head	8848.7493	2803.695	3.156	0.002	3353.314	1.43e+04
FamilyType_Married	3.012e+04	1660.963	18.135	0.000	2.69e+04	3.34e+04
NumUnits_Single attached	1.088e+04	3657.072	2.975	0.003	3710.487	1.8e+04
NumUnits_Single detached	9620.9557	3264.925	2.947	0.003	3221.479	1.6e+04
OwnRent_Outright	5.402e+04	6784.030	7.963	0.000	4.07e+04	6.73e+04
OwnRent_Rented	7487.9379	2009.807	3.726	0.000	3548.578	1.14e+04
YearBuilt_1940-1949	-3525.3908	2046.861	-1.722	0.085	-7537.378	486.596
YearBuilt_1980-1989	6980.8201	1983.314	3.520	0.000	3093.388	1.09e+04
YearBuilt_1990-1999	8349.5363	1976.872	4.224	0.000	4474.732	1.22e+04
YearBuilt_2000-2004	6390.9546	2648.768	2.413	0.016	1199.188	1.16e+04
YearBuilt_2005	8768.0285	5528.804	1.586	0.113	-2068.806	1.96e+04

YearBuilt_2006	7376.9577	6102.564	1.209	0.227	-4584.486	1.93e+04
YearBuilt_2007	8353.7176	6586.451	1.268	0.205	-4556.177	2.13e+04
YearBuilt_2008	6583.8951	7804.799	0.844	0.399	-8714.045	2.19e+04
YearBuilt_2010	3.339e+04	1.17e+04	2.861	0.004	1.05e+04	5.63e+04
FoodStamp_Yes	-1.19e+04	2263.857	-5.255	0.000	-1.63e+04	-7458.775
HeatingFuel_Electricity	1.191e+04	3424.351	3.477	0.001	5193.679	1.86e+04
HeatingFuel_Gas	1.517e+04	2352.210	6.449	0.000	1.06e+04	1.98e+04
HeatingFuel_Oil	1.164e+04	2464.007	4.722	0.000	6805.754	1.65e+04
HeatingFuel_Other	9892.2332	6334.409	1.562	0.118	-2523.643	2.23e+04
Language_English	3510.2503	2637.619	1.331	0.183	-1659.663	8680.164
Language_Other European	-2377.8463	3082.873	-0.771	0.441	-8420.488	3664.795
Language_Spanish	-9853.8978	3157.552	-3.121	0.002	-1.6e+04	-3664.880

Omnibus:	16163.427	**Durbin-Watson:**	0.565
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	427394.290
Skew:	3.107	**Prob(JB):**	0.00
Kurtosis:	23.307	**Cond. No.**	4.78e+04

In [29]:

```
#Now we delete all with p= 0.4,
encoded_x = encoded_x.drop(['Language_Other European'],axis = 1)
X = encoded_x
Y = data.iloc[:, 1:2]
## fit a OLS model with intercept on TV and Radio
X = sm.add_constant(X)
est = sm.OLS(Y, X).fit()

est.summary()
```

Out[29]:

OLS Regression Results

Dep. Variable:	FamilyIncome	**R-squared:**	0.344
Model:	OLS	**Adj. R-squared:**	0.344
Method:	Least Squares	**F-statistic:**	384.9
Date:	Sat, 14 Apr 2018	**Prob (F-statistic):**	0.00
Time:	10:32:46	**Log-Likelihood:**	-2.8944e+05
No. Observations:	22745	**AIC:**	5.789e+05
Df Residuals:	22713	**BIC:**	5.792e+05
Df Model:	31		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-7.636e+04	4913.669	-15.539	0.000	-8.6e+04	-6.67e+04
NumBedrooms	2608.0706	673.051	3.875	0.000	1288.845	3927.297
NumChildren	4126.7104	702.002	5.878	0.000	2750.739	5502.682
NumPeople	-7544.8972	649.897	-11.609	0.000	-8818.740	-6271.055
NumRooms	4622.4075	301.102	15.352	0.000	4032.227	5212.588
NumVehicles	6932.1422	675.221	10.266	0.000	5608.662	8255.622
NumWorkers	1.902e+04	784.353	24.252	0.000	1.75e+04	2.06e+04
HouseCosts	27.0611	0.588	46.039	0.000	25.909	28.213
ElectricBill	46.7473	5.651	8.272	0.000	35.670	57.824
Insurance	18.6036	0.678	27.419	0.000	17.274	19.933
FamilyType_Male Head	8877.0397	2803.431	3.166	0.002	3382.124	1.44e+04
FamilyType_Married	3.016e+04	1660.300	18.164	0.000	2.69e+04	3.34e+04
NumUnits_Single attached	1.096e+04	3655.443	2.999	0.003	3797.034	1.81e+04
NumUnits_Single detached	9611.8740	3264.875	2.944	0.003	3212.496	1.6e+04
OwnRent_Outright	5.4e+04	6783.942	7.961	0.000	4.07e+04	6.73e+04
OwnRent_Rented	7513.8696	2009.508	3.739	0.000	3575.096	1.15e+04
YearBuilt_1940-1949	-3528.0050	2046.840	-1.724	0.085	-7539.950	483.941
YearBuilt_1980-1989	6962.8626	1983.160	3.511	0.000	3075.733	1.08e+04
YearBuilt_1990-1999	8335.6235	1976.772	4.217	0.000	4461.015	1.22e+04
YearBuilt_2000-2004	6374.1850	2648.655	2.407	0.016	1182.640	1.16e+04
YearBuilt_2005	8740.1313	5528.636	1.581	0.114	-2096.374	1.96e+04

YearBuilt_2006	7431.6102	6102.098	1.218	0.223	-4528.920	1.94e+04
YearBuilt_2007	8340.9728	6586.371	1.266	0.205	-4568.766	2.13e+04
YearBuilt_2008	6560.6776	7804.671	0.841	0.401	-8737.012	2.19e+04
YearBuilt_2010	3.348e+04	1.17e+04	2.868	0.004	1.06e+04	5.64e+04
FoodStamp_Yes	-1.188e+04	2263.764	-5.249	0.000	-1.63e+04	-7444.946
HeatingFuel_Electricity	1.195e+04	3423.789	3.491	0.000	5241.324	1.87e+04
HeatingFuel_Gas	1.519e+04	2351.988	6.460	0.000	1.06e+04	1.98e+04
HeatingFuel_Oil	1.164e+04	2463.985	4.722	0.000	6805.835	1.65e+04
HeatingFuel_Other	9933.8723	6334.123	1.568	0.117	-2481.442	2.23e+04
Language_English	5092.4852	1657.978	3.072	0.002	1842.735	8342.235
Language_Spanish	-8304.8503	2436.516	-3.408	0.001	-1.31e+04	-3529.112

Omnibus:	16161.923	**Durbin-Watson:**	0.565
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	427251.159
Skew:	3.106	**Prob(JB):**	0.00
Kurtosis:	23.303	**Cond. No.**	4.78e+04

In [30]:
```
#Now we delete all with p= 0.2,
encoded_x = encoded_x.drop(['YearBuilt_2008', 'YearBuilt_2006', 'YearBuilt_200
7'],axis = 1)
X = encoded_x
Y = data.iloc[:, 1:2]
## fit a OLS model with intercept on TV and Radio
X = sm.add_constant(X)
est = sm.OLS(Y, X).fit()

est.summary()
```

Out[30]:

OLS Regression Results

Dep. Variable:	FamilyIncome	**R-squared:**	0.344
Model:	OLS	**Adj. R-squared:**	0.344
Method:	Least Squares	**F-statistic:**	426.0
Date:	Sat, 14 Apr 2018	**Prob (F-statistic):**	0.00
Time:	10:34:07	**Log-Likelihood:**	-2.8944e+05
No. Observations:	22745	**AIC:**	5.789e+05
Df Residuals:	22716	**BIC:**	5.792e+05
Df Model:	28		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-7.58e+04	4905.035	-15.453	0.000	-8.54e+04	-6.62e+04
NumBedrooms	2619.0779	673.035	3.891	0.000	1299.883	3938.273
NumChildren	4152.7295	701.848	5.917	0.000	2777.059	5528.400
NumPeople	-7552.0026	649.896	-11.620	0.000	-8825.844	-6278.162
NumRooms	4618.5698	301.076	15.340	0.000	4028.439	5208.700
NumVehicles	6946.8977	675.166	10.289	0.000	5623.526	8270.270
NumWorkers	1.9e+04	784.234	24.222	0.000	1.75e+04	2.05e+04
HouseCosts	27.1382	0.586	46.278	0.000	25.989	28.288
ElectricBill	46.5228	5.650	8.234	0.000	35.448	57.597
Insurance	18.5609	0.678	27.371	0.000	17.232	19.890
FamilyType_Male Head	8884.4803	2803.415	3.169	0.002	3389.595	1.44e+04
FamilyType_Married	3.021e+04	1660.084	18.199	0.000	2.7e+04	3.35e+04
NumUnits_Single attached	1.067e+04	3652.187	2.922	0.003	3513.374	1.78e+04
NumUnits_Single detached	9289.0741	3260.502	2.849	0.004	2898.267	1.57e+04
OwnRent_Outright	5.43e+04	6781.573	8.007	0.000	4.1e+04	6.76e+04
OwnRent_Rented	7485.8947	2009.456	3.725	0.000	3547.224	1.14e+04
YearBuilt_1940-1949	-3734.4653	2044.029	-1.827	0.068	-7740.901	271.971
YearBuilt_1980-1989	6706.7591	1978.666	3.390	0.001	2828.439	1.06e+04
YearBuilt_1990-1999	8065.0437	1971.738	4.090	0.000	4200.301	1.19e+04
YearBuilt_2000-2004	6093.2458	2644.632	2.304	0.021	909.587	1.13e+04
YearBuilt_2005	8430.1621	5526.358	1.525	0.127	-2401.877	1.93e+04

YearBuilt_2010	3.315e+04	1.17e+04	2.841	0.005	1.03e+04	5.6e+04
FoodStamp_Yes	-1.195e+04	2263.535	-5.278	0.000	-1.64e+04	-7510.726
HeatingFuel_Electricity	1.2e+04	3423.380	3.504	0.000	5286.204	1.87e+04
HeatingFuel_Gas	1.514e+04	2351.840	6.438	0.000	1.05e+04	1.98e+04
HeatingFuel_Oil	1.148e+04	2462.733	4.663	0.000	6656.216	1.63e+04
HeatingFuel_Other	9950.1717	6333.120	1.571	0.116	-2463.177	2.24e+04
Language_English	5076.7954	1657.859	3.062	0.002	1827.279	8326.312
Language_Spanish	-8355.2078	2436.400	-3.429	0.001	-1.31e+04	-3579.698

Omnibus:	16163.763	**Durbin-Watson:**	0.564
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	427407.942
Skew:	3.107	**Prob(JB):**	0.00
Kurtosis:	23.307	**Cond. No.**	4.78e+04

In [31]:
```
#Now we delete all with p>0.1,
encoded_x = encoded_x.drop(['YearBuilt_2005', 'HeatingFuel_Other'],axis = 1)
X = encoded_x
Y = data.iloc[:, 1:2]
## fit a OLS model with intercept on TV and Radio
X = sm.add_constant(X)
est = sm.OLS(Y, X).fit()

est.summary()
```

Out[31]:

OLS Regression Results

Dep. Variable:	FamilyIncome	**R-squared:**	0.344
Model:	OLS	**Adj. R-squared:**	0.343
Method:	Least Squares	**F-statistic:**	458.5
Date:	Sat, 14 Apr 2018	**Prob (F-statistic):**	0.00
Time:	10:35:37	**Log-Likelihood:**	-2.8944e+05
No. Observations:	22745	**AIC:**	5.789e+05
Df Residuals:	22718	**BIC:**	5.792e+05
Df Model:	26		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-7.413e+04	4829.883	-15.348	0.000	-8.36e+04	-6.47e+04
NumBedrooms	2616.5505	673.046	3.888	0.000	1297.335	3935.766
NumChildren	4180.9262	701.545	5.960	0.000	2805.850	5556.003
NumPeople	-7568.3387	649.828	-11.647	0.000	-8842.047	-6294.631
NumRooms	4624.6413	301.076	15.360	0.000	4034.512	5214.771
NumVehicles	6954.3109	675.194	10.300	0.000	5630.884	8277.738
NumWorkers	1.899e+04	784.268	24.213	0.000	1.75e+04	2.05e+04
HouseCosts	27.1921	0.586	46.420	0.000	26.044	28.340
ElectricBill	46.4833	5.650	8.227	0.000	35.409	57.558
Insurance	18.5484	0.678	27.354	0.000	17.219	19.877
FamilyType_Male Head	8866.7660	2803.552	3.163	0.002	3371.612	1.44e+04
FamilyType_Married	3.022e+04	1659.841	18.209	0.000	2.7e+04	3.35e+04
NumUnits_Single attached	1.035e+04	3648.388	2.836	0.005	3194.979	1.75e+04
NumUnits_Single detached	8935.8820	3255.409	2.745	0.006	2555.058	1.53e+04
OwnRent_Outright	5.466e+04	6777.767	8.064	0.000	4.14e+04	6.79e+04
OwnRent_Rented	7564.0075	2009.057	3.765	0.000	3626.119	1.15e+04
YearBuilt_1940-1949	-3827.5128	2042.920	-1.874	0.061	-7831.776	176.751
YearBuilt_1980-1989	6604.3559	1976.412	3.342	0.001	2730.453	1.05e+04
YearBuilt_1990-1999	7928.7821	1968.991	4.027	0.000	4069.426	1.18e+04
YearBuilt_2000-2004	5926.5775	2642.537	2.243	0.025	747.023	1.11e+04
YearBuilt_2010	3.288e+04	1.17e+04	2.817	0.005	1e+04	5.58e+04

FoodStamp_Yes	-1.199e+04	2263.605	-5.295	0.000	-1.64e+04	-7549.406
HeatingFuel_Electricity	1.068e+04	3334.181	3.202	0.001	4140.022	1.72e+04
HeatingFuel_Gas	1.388e+04	2223.622	6.242	0.000	9522.484	1.82e+04
HeatingFuel_Oil	1.018e+04	2339.380	4.351	0.000	5593.872	1.48e+04
Language_English	5051.2895	1657.865	3.047	0.002	1801.761	8300.818
Language_Spanish	-8387.2569	2436.343	-3.443	0.001	-1.32e+04	-3611.858

Omnibus:	16159.192	**Durbin-Watson:**	0.564
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	426755.216
Skew:	3.106	**Prob(JB):**	0.00
Kurtosis:	23.291	**Cond. No.**	4.78e+04

In [33]:

```
#Now we delete all with p>0.05,
encoded_x = encoded_x.drop(['YearBuilt_1940-1949'],axis = 1)
X = encoded_x
Y = data.iloc[:, 1:2]
## fit a OLS model with intercept on TV and Radio
X = sm.add_constant(X)
est = sm.OLS(Y, X).fit()

est.summary()
```

Out[33]:

OLS Regression Results

Dep. Variable:	FamilyIncome	**R-squared:**	0.344
Model:	OLS	**Adj. R-squared:**	0.343
Method:	Least Squares	**F-statistic:**	476.7
Date:	Sat, 14 Apr 2018	**Prob (F-statistic):**	0.00
Time:	10:36:54	**Log-Likelihood:**	-2.8944e+05
No. Observations:	22745	**AIC:**	5.789e+05
Df Residuals:	22719	**BIC:**	5.791e+05
Df Model:	25		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-7.452e+04	4825.774	-15.441	0.000	-8.4e+04	-6.51e+04
NumBedrooms	2636.3236	673.000	3.917	0.000	1317.197	3955.450
NumChildren	4224.0566	701.206	6.024	0.000	2849.645	5598.469
NumPeople	-7600.0411	649.644	-11.699	0.000	-8873.387	-6326.695
NumRooms	4638.0626	301.007	15.408	0.000	4048.068	5228.057
NumVehicles	6987.1404	675.004	10.351	0.000	5664.086	8310.194
NumWorkers	1.895e+04	784.086	24.174	0.000	1.74e+04	2.05e+04
HouseCosts	27.1882	0.586	46.411	0.000	26.040	28.336
ElectricBill	46.5914	5.650	8.246	0.000	35.517	57.666
Insurance	18.5476	0.678	27.351	0.000	17.218	19.877
FamilyType_Male Head	8925.6972	2803.530	3.184	0.001	3430.586	1.44e+04
FamilyType_Married	3.029e+04	1659.546	18.252	0.000	2.7e+04	3.35e+04
NumUnits_Single attached	1.015e+04	3647.062	2.783	0.005	2999.787	1.73e+04
NumUnits_Single detached	8793.9796	3254.707	2.702	0.007	2414.530	1.52e+04
OwnRent_Outright	5.479e+04	6777.747	8.084	0.000	4.15e+04	6.81e+04
OwnRent_Rented	7573.1879	2009.162	3.769	0.000	3635.094	1.15e+04
YearBuilt_1980-1989	6963.5534	1967.200	3.540	0.000	3107.707	1.08e+04
YearBuilt_1990-1999	8285.0467	1959.895	4.227	0.000	4443.518	1.21e+04
YearBuilt_2000-2004	6281.6839	2635.877	2.383	0.017	1115.185	1.14e+04
YearBuilt_2010	3.324e+04	1.17e+04	2.848	0.004	1.04e+04	5.61e+04
FoodStamp_Yes	-1.2e+04	2263.708	-5.303	0.000	-1.64e+04	-7567.808

HeatingFuel_Electricity	1.067e+04	3334.365	3.201	0.001	4138.780	1.72e+04
HeatingFuel_Gas	1.38e+04	2223.291	6.206	0.000	9438.981	1.82e+04
HeatingFuel_Oil	1.012e+04	2339.290	4.326	0.000	5534.049	1.47e+04
Language_English	5103.1042	1657.726	3.078	0.002	1853.849	8352.360
Language_Spanish	-8346.1558	2436.379	-3.426	0.001	-1.31e+04	-3570.687

Omnibus:	16165.026	**Durbin-Watson:**	0.564
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	427155.870
Skew:	3.107	**Prob(JB):**	0.00
Kurtosis:	23.300	**Cond. No.**	4.78e+04

In [35]:
```
#Looks like nothing changed, so we are going to delete the dummy variables whi
ch had one or two variables deleted during backward elimination.
#Year_built, heating_fuel, language,
encoded_x = encoded_x.drop(['YearBuilt_1980-1989', 'YearBuilt_1990-1999', 'Yea
rBuilt_2000-2004', 'YearBuilt_2010', 'HeatingFuel_Electricity','HeatingFuel_Ga
s','HeatingFuel_Oil','Language_English','Language_Spanish'],axis = 1)
X = encoded_x
Y = data.iloc[:, 1:2]
## fit a OLS model with intercept on TV and Radio
X = sm.add_constant(X)
est = sm.OLS(Y, X).fit()

est.summary()
```

Out[35]:

OLS Regression Results

Dep. Variable:	FamilyIncome	**R-squared:**	0.341
Model:	OLS	**Adj. R-squared:**	0.340
Method:	Least Squares	**F-statistic:**	733.8
Date:	Sat, 14 Apr 2018	**Prob (F-statistic):**	0.00
Time:	10:42:18	**Log-Likelihood:**	-2.8950e+05
No. Observations:	22745	**AIC:**	5.790e+05
Df Residuals:	22728	**BIC:**	5.792e+05
Df Model:	16		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-5.313e+04	3949.502	-13.452	0.000	-6.09e+04	-4.54e+04
NumBedrooms	2578.8053	674.023	3.826	0.000	1257.675	3899.936
NumChildren	4744.3057	697.199	6.805	0.000	3377.749	6110.863
NumPeople	-8421.7969	640.641	-13.146	0.000	-9677.497	-7166.097
NumRooms	4819.3631	300.597	16.033	0.000	4230.173	5408.553
NumVehicles	7291.4816	666.525	10.940	0.000	5985.048	8597.916
NumWorkers	1.896e+04	785.495	24.133	0.000	1.74e+04	2.05e+04
HouseCosts	27.1698	0.575	47.249	0.000	26.043	28.297
ElectricBill	46.7511	5.624	8.312	0.000	35.727	57.776
Insurance	18.3705	0.678	27.110	0.000	17.042	19.699
FamilyType_Male Head	8427.4340	2809.411	3.000	0.003	2920.796	1.39e+04
FamilyType_Married	3.028e+04	1659.834	18.241	0.000	2.7e+04	3.35e+04
NumUnits_Single attached	6911.5196	3569.842	1.936	0.053	-85.615	1.39e+04
NumUnits_Single detached	4790.7069	3171.850	1.510	0.131	-1426.337	1.1e+04
OwnRent_Outright	5.983e+04	6755.382	8.856	0.000	4.66e+04	7.31e+04
OwnRent_Rented	6224.2296	2000.751	3.111	0.002	2302.621	1.01e+04
FoodStamp_Yes	-1.202e+04	2268.823	-5.297	0.000	-1.65e+04	-7571.857

Omnibus:	16137.176	**Durbin-Watson:**	0.559
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	424616.932
Skew:	3.101	**Prob(JB):**	0.00
Kurtosis:	23.238	**Cond. No.**	2.77e+04

In [36]:
```
#Now we delete all with p>0.05,
encoded_x = encoded_x.drop(['NumUnits_Single attached','NumUnits_Single detach
ed'],axis = 1)
X = encoded_x
Y = data.iloc[:, 1:2]
## fit a OLS model with intercept on TV and Radio
X = sm.add_constant(X)
est = sm.OLS(Y, X).fit()

est.summary()
```

Out[36]:

OLS Regression Results

Dep. Variable:	FamilyIncome	**R-squared:**	0.341
Model:	OLS	**Adj. R-squared:**	0.340
Method:	Least Squares	**F-statistic:**	838.3
Date:	Sat, 14 Apr 2018	**Prob (F-statistic):**	0.00
Time:	10:43:24	**Log-Likelihood:**	-2.8950e+05
No. Observations:	22745	**AIC:**	5.790e+05
Df Residuals:	22730	**BIC:**	5.792e+05
Df Model:	14		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-4.825e+04	2695.165	-17.904	0.000	-5.35e+04	-4.3e+04
NumBedrooms	2521.7605	671.611	3.755	0.000	1205.356	3838.165
NumChildren	4643.8966	693.021	6.701	0.000	3285.528	6002.265
NumPeople	-8329.5494	636.134	-13.094	0.000	-9576.415	-7082.683
NumRooms	4828.7868	299.570	16.119	0.000	4241.609	5415.964
NumVehicles	7117.4488	653.861	10.885	0.000	5835.837	8399.061
NumWorkers	1.902e+04	784.937	24.225	0.000	1.75e+04	2.06e+04
HouseCosts	27.3112	0.570	47.881	0.000	26.193	28.429
ElectricBill	46.8050	5.624	8.322	0.000	35.781	57.829
Insurance	18.3862	0.678	27.135	0.000	17.058	19.714
FamilyType_Male Head	8288.0088	2808.607	2.951	0.003	2782.946	1.38e+04
FamilyType_Married	3.019e+04	1659.301	18.195	0.000	2.69e+04	3.34e+04
OwnRent_Outright	6.167e+04	6608.745	9.331	0.000	4.87e+04	7.46e+04
OwnRent_Rented	6257.0759	1985.692	3.151	0.002	2364.984	1.01e+04
FoodStamp_Yes	-1.227e+04	2264.513	-5.419	0.000	-1.67e+04	-7832.763

Omnibus:	16128.441	**Durbin-Watson:**	0.560
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	423938.818
Skew:	3.099	**Prob(JB):**	0.00
Kurtosis:	23.222	**Cond. No.**	2.70e+04

As we see, we could not improve our score, meaning we have to clean the data and use assumptions of Multiple Regression to take out variables before modelling.