Multilinear Regression on Auto-Mpg Dataset

Introduction

A linear regression model with stepwise regression (backwards elimination of t-tests with a significance level of 0.05) was applied to the Auto-Mpg Data dataset to find the best model to explain the data and predict the Mpg. What was interesting to discover was the large role the attributes Origin and Year played in predicting.

The dataset

The Auto-Mpg Data dataset is composed of 398 instances and 9 attributes and "concerns city-cycle fuel consumption in miles per gallon, to be predicted in terms of 3 multivalued discrete and 5 continuous attributes" (Quinlan, 1993).

1. Mpg: continuous
2. Cylinders: multi-valued discrete
3. Displacement: continuous
4. Horsepower: continuous
5. Weight: continuous
6. Acceleration: continuous
7. Model year: multi-valued discrete
8. Origin: multi-valued discrete
9. Car name: string (unique for each instance)

Exploring and Cleaning the Data

- Column titles were added.
- 6 NaNs found in the 'Horsepower' were replaced by the mean of the column.
- Horsepower was changed from a string to a float
- No duplicate rows were discovered

Chart 1

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 398 entries, 0 to 397
Data columns (total 9 columns):
Mpg             398 non-null float64
Cylinders       398 non-null int64
Displacement    398 non-null float64
Horsepower      398 non-null float64
Weight          398 non-null int64
Acceleration    398 non-null float64
Year            398 non-null int64
Origin          398 non-null int64
Car Name        398 non-null object
dtypes: float64(4), int64(4), object(1)
memory usage: 28.1+ KB
```

The data was checked for outliers. Running bar plots, histograms, and simple statistics showed nothing else out of the ordinary. Visualizing the data showed correlations among the variables Mpg, Weight, Displacement, and Horsepower (charts 3 – 5). It also became apparent that there is a relationship between Mpg and variables Year and Origin (charts 5 & 6).

Chart 2

	Mpg	Cylinders	Displacement	Horsepower	Weight	Acceleration	Year	Origin
count	398.000000	398.000000	398.000000	398.000000	398.000000	398.000000	398.000000	398.000000
mean	23.514573	5.454774	193.425879	104.469388	2970.424623	15.568090	76.010050	1.572864
std	7.815984	1.701004	104.269838	38.199187	846.841774	2.757689	3.697627	0.802055
min	9.000000	3.000000	68.000000	46.000000	1613.000000	8.000000	70.000000	1.000000
25%	17.500000	4.000000	104.250000	76.000000	2223.750000	13.825000	73.000000	1.000000
50%	23.000000	4.000000	148.500000	95.000000	2803.500000	15.500000	76.000000	1.000000
75%	29.000000	8.000000	262.000000	125.000000	3608.000000	17.175000	79.000000	2.000000
max	46.600000	8.000000	455.000000	230.000000	5140.000000	24.800000	82.000000	3.000000

Chart 3

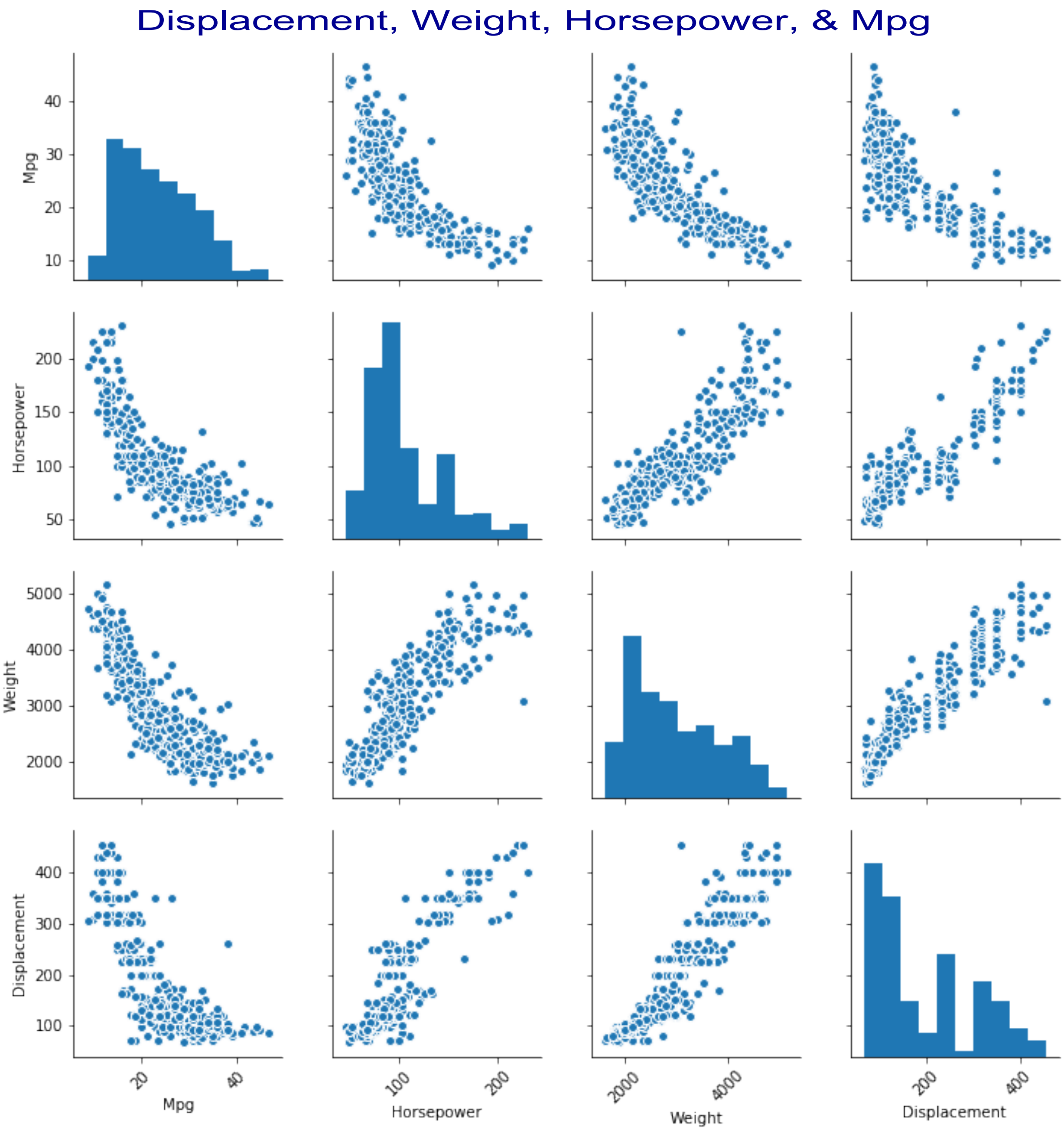

Displacement, Weight, Horsepower, & Mpg
Mpg
Horsepower
Weight
Displacement
10
20
30
40
50
100
150
200
2000
3000
4000
5000
300
400
Mpg
Horsepower
Weight
Displacement

Chart 4

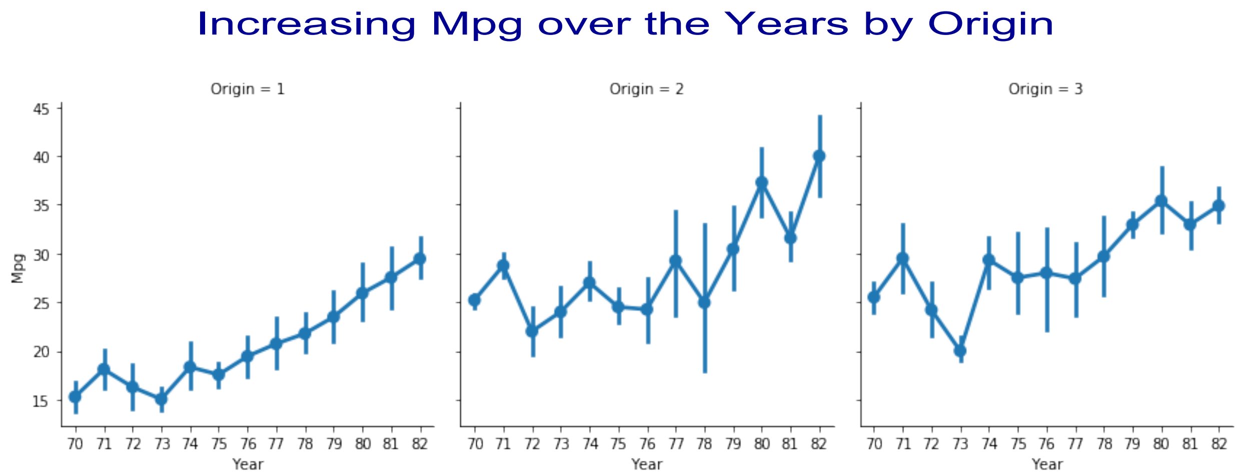


Chart 5

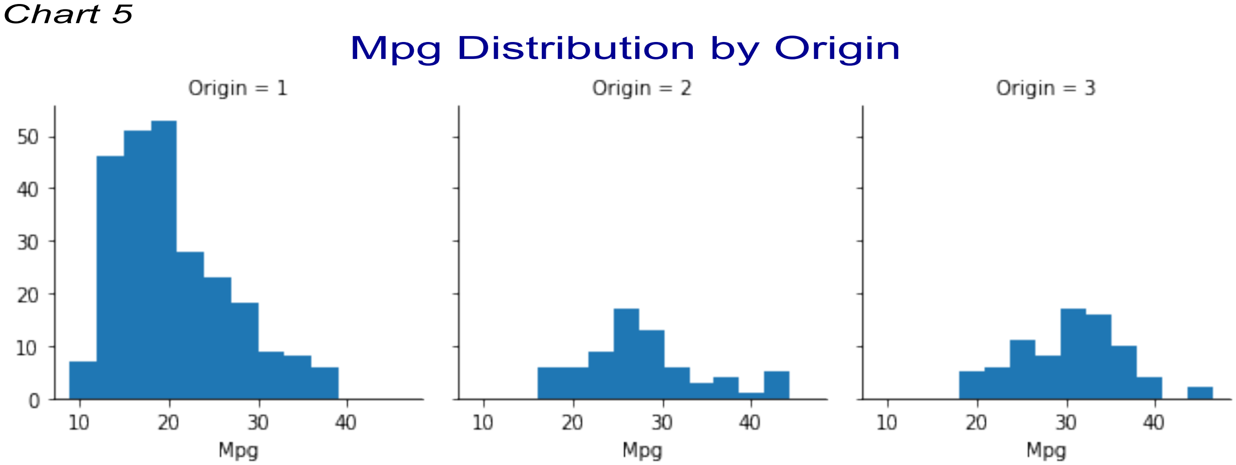
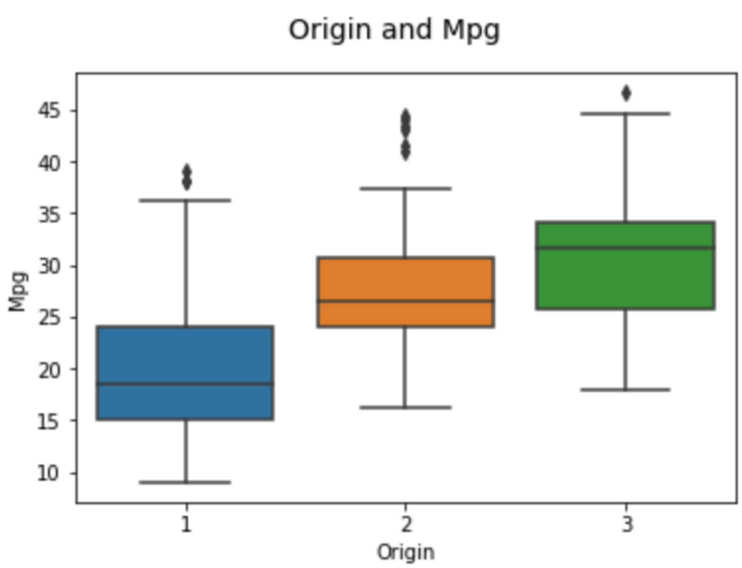


Chart 6



Preprocessing

Mpg was set as the dependent variable. The eight independent variables were changed into 19 through OneHotEncoder(code in appendix):

- CarName was dropped as each value was unique(e.g. 'Ford Mustang')
  - 8-1=7
- Origin(1,2,3) split into 3 dummy variables.
  - 7 -1 + 3 = 9 variables.
- Origin(1) removed to avoid the dummy variable trap.
  - 9-1 = 8 variables
- Year(1970-1982) split into 13 dummies.
  - (8-1 +13=20)
- Year(1) removed to avoid the dummy variable trap
  - (20-1=19)
- Leaving 19 variables to test.

Number	***Attribute***	***Dropped Attributes***
0	Constant	
		D. Year 70
1	D. (dummy)Year 71	
2	D. Year 72	
3	D. Year 73	
4	D. Year 74	
5	D. Year 75	
6	D. Year 76	
7	D. Year 77	
8	D. Year 78	
9	D. Year 79	
10	D. Year 80	
11	D. Year 81	
12	D. Year 82	
		D. Origin 1
13	D. Origin 2	
14	D. Origin 3	
15	Cylinders	
16	Displacement	
17	Horsepower	
18	Weight	
19	Acceleration	
		Year
		Origin
		CarName

Running the Model

A linear regression model using backwards elimination of t-tests with a significance level of 0.05 was used. After a number of iterations the best model was found to consist of 13 variables:

Chart 6

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.852
Model:	OLS	**Adj. R-squared:**	0.847
Method:	Least Squares	**F-statistic:**	169.9
Date:	Sun, 29 Jul 2018	**Prob (F-statistic):**	3.15e-150
Time:	18:46:08	**Log-Likelihood:**	-1002.5
No. Observations:	398	**AIC:**	2033.
Df Residuals:	384	**BIC:**	2089.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	38.0894	0.980	38.871	0.000	36.163	40.016
x1	1.2691	0.654	1.941	0.053	-0.016	2.555
x2	1.3988	0.593	2.360	0.019	0.234	2.564
x3	2.9333	0.637	4.608	0.000	1.682	4.185
x4	2.7504	0.578	4.762	0.000	1.615	3.886
x5	5.0264	0.632	7.950	0.000	3.783	6.269
x6	9.1470	0.657	13.916	0.000	7.855	10.439
x7	6.6328	0.652	10.174	0.000	5.351	7.915
x8	8.2003	0.643	12.749	0.000	6.936	9.465
x9	2.6086	0.522	4.998	0.000	1.582	3.635

The Adjusted R-squared of 0.847 means that most of the variance of the output variable is explained by the model.

Chart 7

Positional Number in the Model	Attribute	Number from OLS Regression Results	Coefficient
0:	Constant/bias	0	38.0878
4:	Year 74	1	1.267
6:	Year 76	2	1.3989
7:	Year 77	3	2.9336
8:	Year 78	4	2.7507
9:	Year 79	5	5.0265
10:	Year 80	6	9.1434
11:	Year 81	7	6.6309
12:	Year 82	8	8.1992
13:	Origin 2	9	2.6098
14:	Origin 3	10	2.5146
16:	Displacement	11	0.0147
17:	Horsepower	12	-0.0325
18:	Weight	13	-0.0060

Conclusion/Recommendation

The conclusion is that the optimal team of independent variables that can predict the Mpg with the highest statistical significance and the highest impact is:

38.0879 + Year74(1.267) + Year76(1.3989) + Year77(2.9336) + Year78(2.7507) + Year79(5.0265) + Year80(9.1434) + Year81(6.6309) + Year82(8.1992) + Origin2(2.6098) + Origin3(2.5146) + Displacement(0.0147) + Horsepower(-0.0325) + Weight(-0.0060)

This left a model in which Origin and Year played a bigger factor than the physical attributes (Acceleration, Weight, Displacement, Horsepower) and that, due to the number of dummy variables, looks complex.

The cities or countries designated by the Origins variables are not revealed. But if they are different countries it could be that different countries have different demands. The U.S. for a long period of time produced heavy and fuel inefficient automobiles.

The role the year plays in the model may seem puzzling. However, the 70s were the time of the fuel crises. “In response to the oil price shocks of the early 1970s, Congress passed the nation’s first Corporate Average Fuel Economy (CAFE) standards in 1975. The law called for a doubling of passenger-vehicle efficiency—to 27.5 miles (Lubitsch, 2011).

References

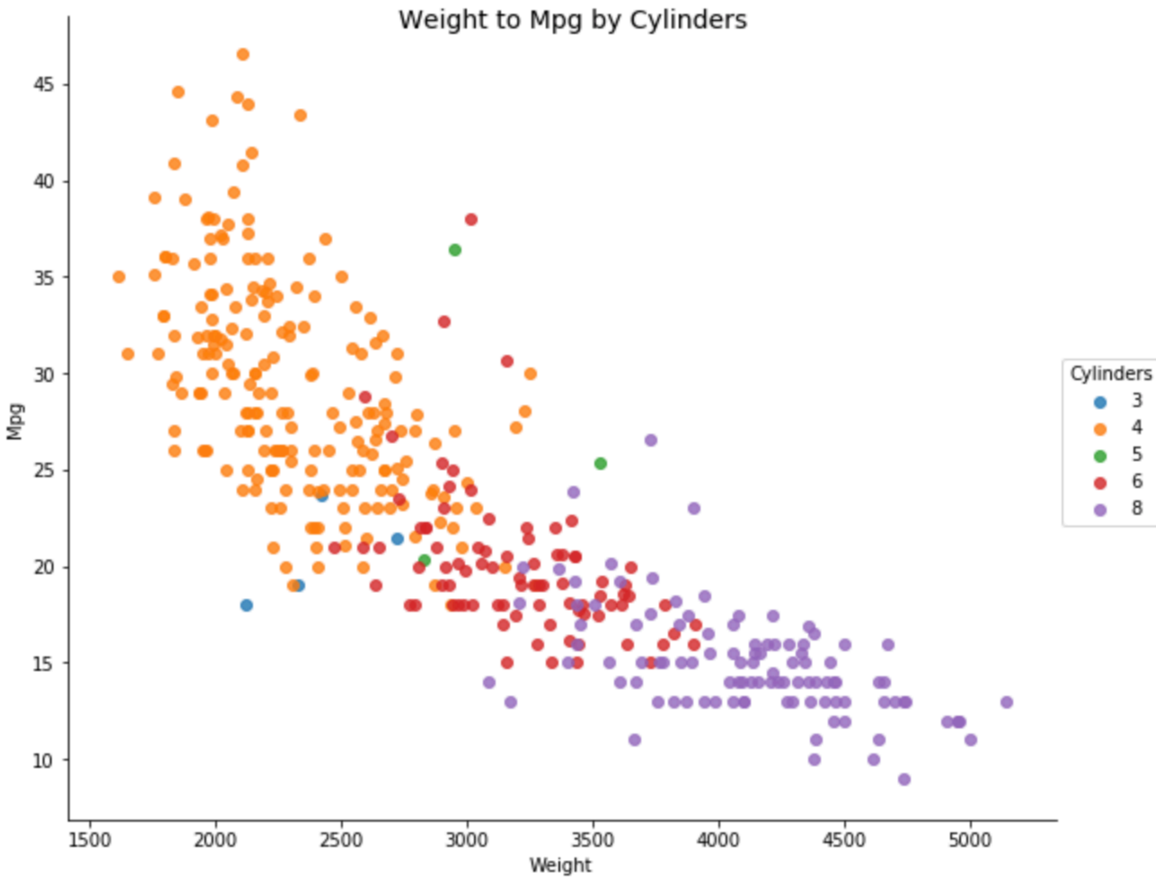
Jessica Frohman Lubetsky. *Clean Energy*, , 0AD, www.pewtrusts.org/~/media/assets/2011/04/history-of-fuel-economy-clean-energy-factsheet.pdf.

“UCI Machine Learning Repository: Auto MPG Data Set.” *UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set*, archive.ics.uci.edu/ml/datasets/Auto MPG.

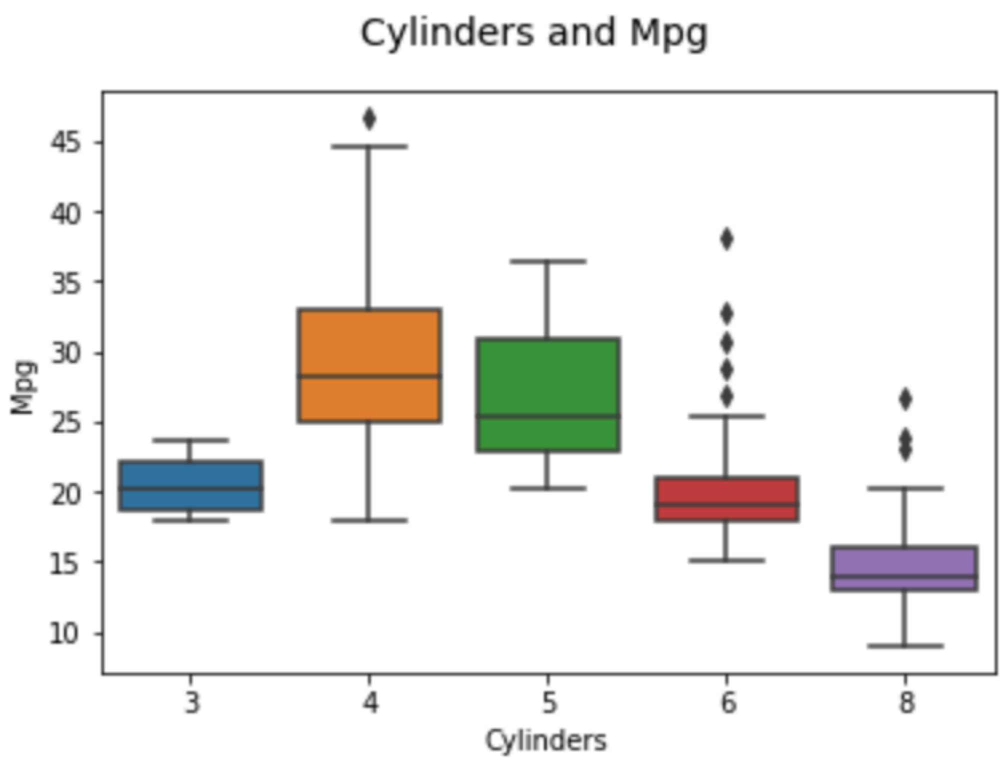
Appendix

More data exploration.

Graph 7



Graph 8



Code

OneHotEncoder – Convert Origin and Year to Dummy Variables

```
# Converting Origin into 3 Dummy Variables
from sklearn.preprocessing import LabelEncoder, OneHotEncoder
labelencoder = LabelEncoder()
X[:, 6] = labelencoder.fit_transform(X[:, 6])
onehotencoder = OneHotEncoder(categorical_features = [6])
X = onehotencoder.fit_transform(X).toarray()
```

```
# Removing One Dummy from the 3 Origin Dummies to Avoid the Dummy Variable Trap
X = X[:, 1:]
```

```
# Converting Year Attribute into 13 Dummy Variables
from sklearn.preprocessing import LabelEncoder, OneHotEncoder
labelencoder = LabelEncoder()
X[:, 7] = labelencoder.fit_transform(X[:, 7])
onehotencoder = OneHotEncoder(categorical_features = [7])
X = onehotencoder.fit_transform(X).toarray()
```

```
# Removing One Dummy from the 13 Year Dummies to Avoid the Dummy Variable Trap
X = X[:, 1:]
```