Northeastern University
College of Professional Studies

ALY 6020 – PAREDICTIVE ANALYTICS

Prof. Justin Grosz

Module 2 – Building the car of the future

03/05/2023

Pooja Kairamkonda

INTRODUCTION

Car manufacturing involves complex process, beginning with the design and construction of its fundamental parts and ending with the creation of its enormous framework. Few of the elements like MPG, industry, materials, horsepower, build quality will decide the performance & efficiency of the car. In this report, major important attributes of a car dataset such as displacement, acceleration, cylinders, horsepower, manufacturing year were used to perform the analysis. The results were analyzed to identify the trends & patterns of individual attributes including acceleration, displacement, cylinders & weight will that effectively impact the miles per gallon (MPG) of the car. The aim of this project is to predict the MPG based on the important attributes of a vehicle. A linear regression model was used to predict the outcome and determine the significance of attributes that will contribute to build a proper car. In order to optimize the results and select the best significant features that help in to achieve higher MPG over others, feature selection technique was used.

DATA CLEANING

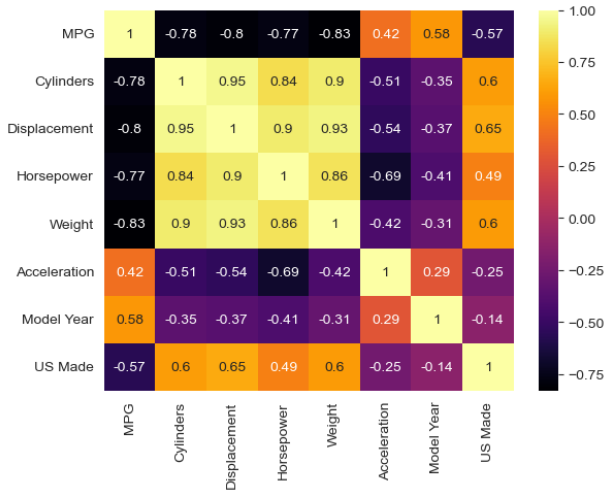
The dataset ‘car.csv’ includes 398 records and 8 variables. At a first glance, the dataset was in a good shape as the datatype of each variable was accurate except ‘Horsepower’ which was object. After fixing the datatype to numeric, the column was found to have 6 null values.
Here we have performed numerical analysis to check the data distribution of the variable and observe the outliers. With this initial analysis, the data was found to be right skewed and there are outliers present in the data. The null values were replaced with median since the mean was not the best representation of central tendency. This median value preserves the

distribution of the data and does not introduce bias to the analysis. Furthermore, no duplicate records were found in the dataset. The final dataset was used for exploratory data analysis in the next part.

EXPLORATORY DATA ANALYSIS

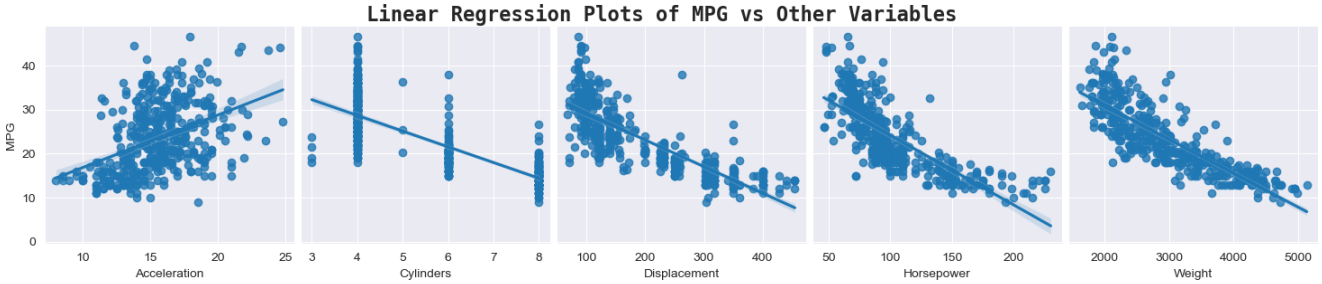
In this section, exploration of variables that may drive the MPG of the vehicle were analyzed using the a variety of graphs & plots that could possibly seem to impact the MPG factor.

A subset of numeric columns that can effectively have an impact on the car performance , was used to observe the outliers in the dataset. For the variables 'Cylinders' and 'Weight' the outliers were not noticed. Furthermore, a histogram of all these variables was generated to observe that data distribution, among which ‘Acceleration’ was found to have normal distribution. These graphs are included in the Appendix 1. In order to explore the relationship between the pair of numeric variables & target variable ‘MPG’, a correlation plot was displayed.



Upon observing the correlation matrix, we can clearly say that, the attributes that were initially considered to possibly have an impact, such as cylinders, displacement, horsepower and wight are negatively correlated with ‘MPG’. A vehicles mileage will decrease as these

numbers increase, according to a significant negative connection between weight, horsepower, displacement, and cylinders. To understand the significance between these variables and the target variable, regression plots were used that clearly indicates a positive correlation between Acceleration & MPG. On the other hand, a strong negative correlation was observed with 'Weight' indicating heavier vehicles will have lower mileage.



The positive correlation between MPG & Model year lead us to observe the yearly advancement in the car models as shown in Appendix 1. Here we have appended the attribute and added '19' for better representation of year value as '1970' instead of '70'. However, the analysis performed in the later part will help us finalize the attributes to focus on in order to higher MPG.

ANALYSIS

Now that we have completed initial analysis, we will conduct the further analysis of by developing models based on the initial findings & deductions made from the EDA. In the section, we are focusing on building a linear regression model with all the independent variables and one dependent variable 'MPG' which is our target variable, to make predictions on how these variables will have an impact and recommend variables to stakeholders that contribute to higher MPG over other.

The dataset was split into 80%-20% train-test split. A linear regression model was fit on the dataset and the OLS summary of regression analysis was obtained to observe & interpret the results.

Here, it was observed that, initially the R-squared value of 82.0% indicates variation in MPG due to independent variables including Acceleration, Cylinders, Weight, Displacement, Model Year, & US made. The statistical parameters in the summary help improve the results by using optimization techniques. From the coefficient column we can clearly observe that, the variables Cylinders, Model year, US made have high impact on MPG than displacement, horsepower, weight and acceleration. In the p-value column, the variables with high p-value (>0.05) were focused since they indicate how the measurement of those variables have no effect on the target variable. The variables with lesser p-value are more significant. The AIC value of 1687, need to be narrowed down during optimization for better prediction.

After observing the initial analysis results, we have chosen backward feature selection technique to eliminate variables and lower the AIC value in order to get a better model.

The variables ‘displacement’, ‘weight’, ‘model year’, and ‘US made’ were provided as part of the findings produced via backward selection, all of which had p-values < 0.05. The variables ‘acceleration’, ‘horsepower’, ‘cylinders’ were eliminated due to high p-value as a result of multicollinearity. From the below table we can observe that, the AIC values initially started decreasing using backward selection & then increased once the model was fit with all the variables having p-value < 0.05.

Linear regression model	R-squared	AIC
Model 1	82.00%	1687
Model 2	81.90%	1685
Model 3	81.90%	1684

Model 4	81.70%	1685

Based on the p-value and AIC, for each model the independent variables were eliminated as follows:

Model 2: Acceleration (p-value: 0.508)

Model 3: Acceleration, Cylinder (p-value: 0.329)

Model 4: Acceleration, Cylinder, Horsepower (p-value: 0.116)

Even though the 'Horsepower' was included in Model 3, it lead us with better AIC value and based on the summary of this model we have analyzed our results.

CONCLUSION & RECOMMENDATIONS

Based on the analysis result the model provided us the variables 'displacement', 'Model year', 'Weight', 'US made'. After carefully analyzing the results, we can say that few of the such as 'Model year' & 'US made' variables from this set do not really help in calculating MPG of a vehicle and doesn't really contribute to the fact that how efficiency of a vehicle can be increased. For any vehicle, the weight will decided by the number of cylinders inside an engine, and other parts of the vehicle. A larger engine will include more number of cylinders leading to a more displacement and horsepower. As the model has already provided 'Weight' which can be targeted in order to produce lighter weight parts and a compact engine since heavy cars will consume more fuel resulting in low MPG. However, lightweight cars can cause greater damage and human injuries, thus we must be careful to take safety precautions when designing them and not compromising structural integrity. Also, considering the displacement factor, a lower

displacement can be achieved by using smaller engine with less torque but more fuel efficiency. The model summary results are represented in Appendix 1.

Moreover, after observing the AIC results, we can conclude that, model 3 with least AIC can better predict the results. On the other it includes an additional variable 'Horsepower' with p-value > 0.05. As mentioned, a lightweight vehicle requires less energy to accelerate and maintain a constant speed. A lighter vehicle can accelerate and change directions more readily because it has a higher power-to-weight ratio. With the use of advanced technologies such as direct injection, turbocharging, and variable valve timing to improve efficiency and reduce emissions. Having said that, it is possible to maintain a vehicle at low horsepower to improve the MPG. We can advise the business stakeholders to consider weight and displacement when designing automobiles that will contribute to high MPG because these parameters are related and can deliver better performance if utilized accurately.

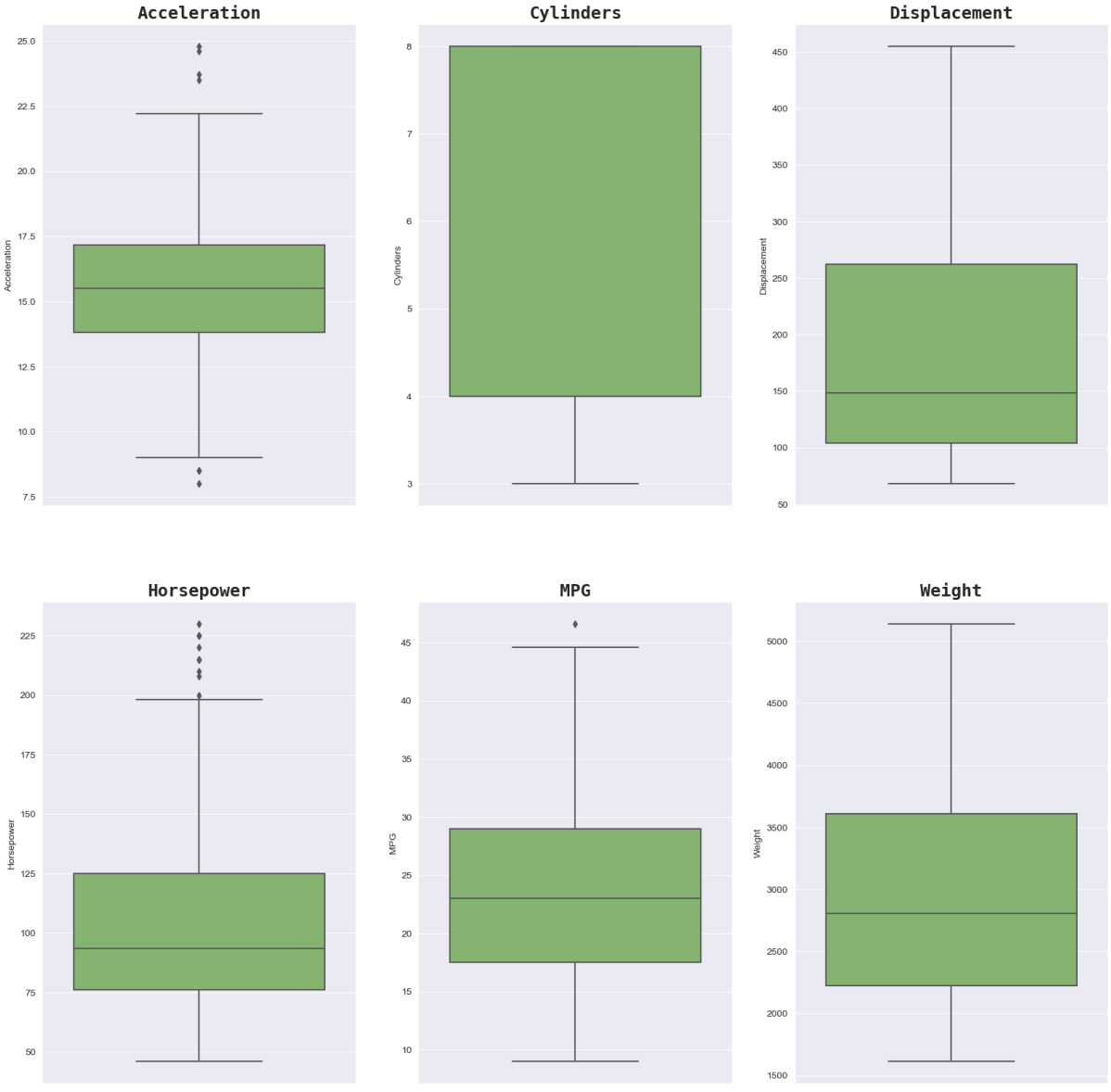
REFERENCES

1. Md Sohel Mahmood (Apr 22, 2022). Simple Explanation of Statsmodel Linear Regression Model Summary. Retrieved from - https://towardsdatascience.com/simple-explanation-of-statsmodel-linear-regression-model-summary-35961919868b
2. Tim McAleer (Dec 5, 2020). Interpreting Linear Regression Through statsmodels .summary() . Retrieved from - https://medium.com/swlh/interpreting-linear-regression-through-statsmodels-summary-4796d359035a#
3. Sagar Rawale. (Aug 1, 2018). Feature Selection Methods in Machine Learning. Retrieved from - https://medium.com/@sagar.rawale3/feature-selection-methods-in-

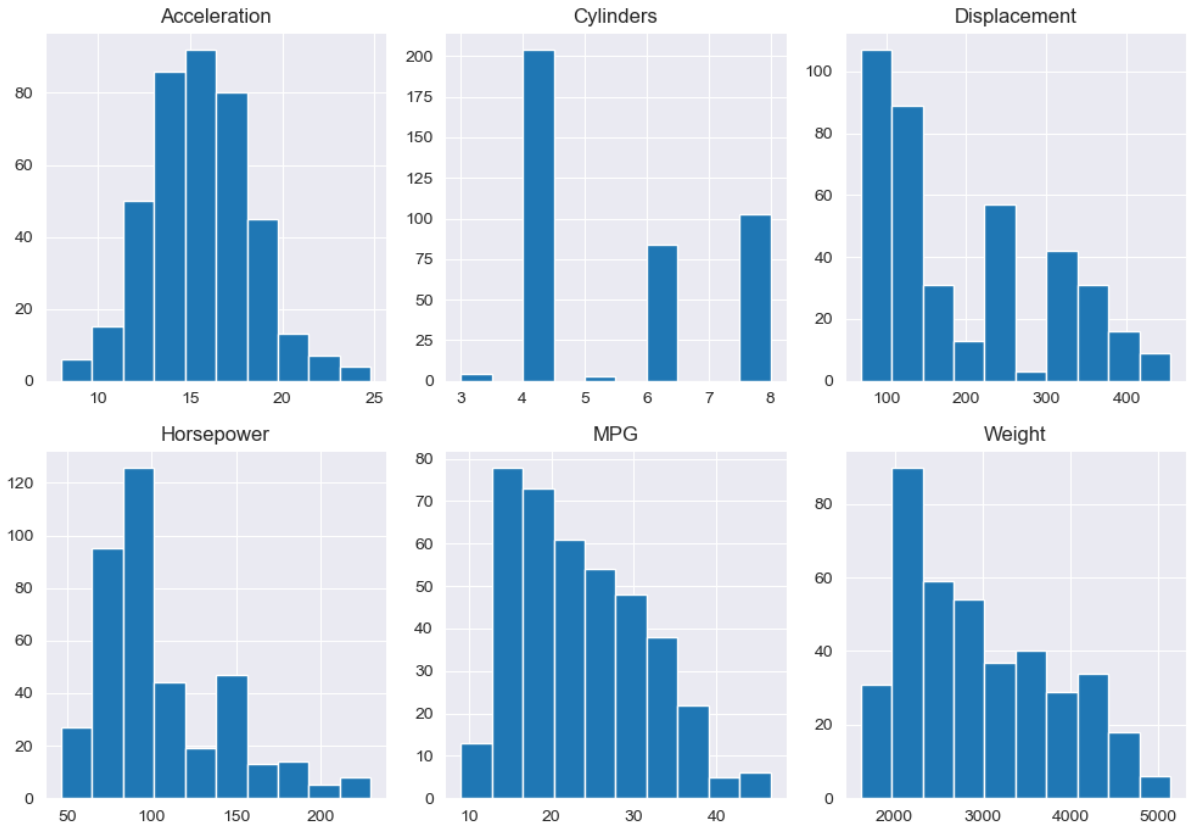
machine-learning-eaeef12019cc

APPENDIX 1

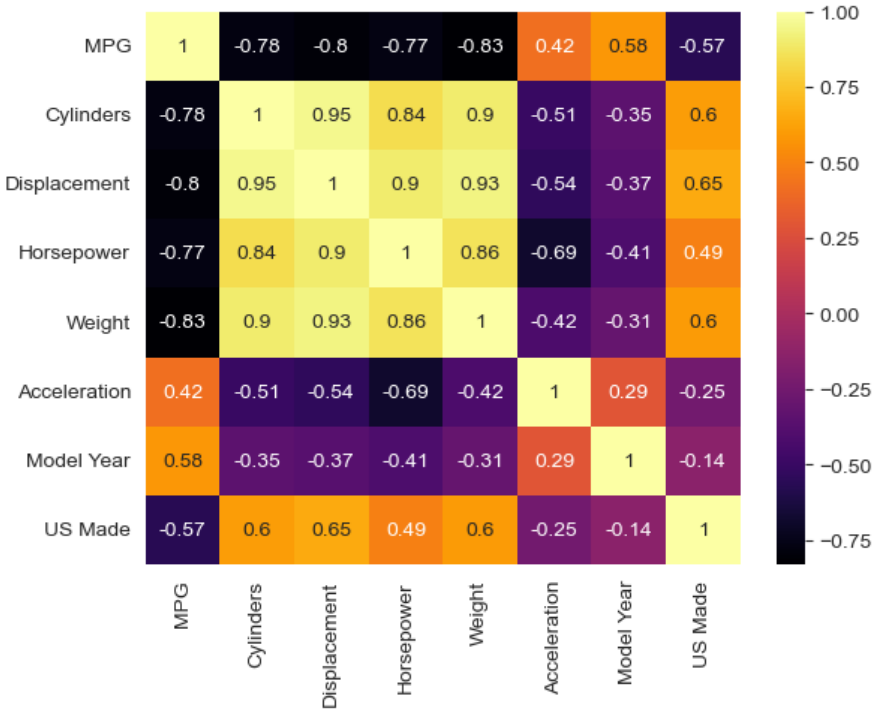
1. Detecting outliers of numerical variables



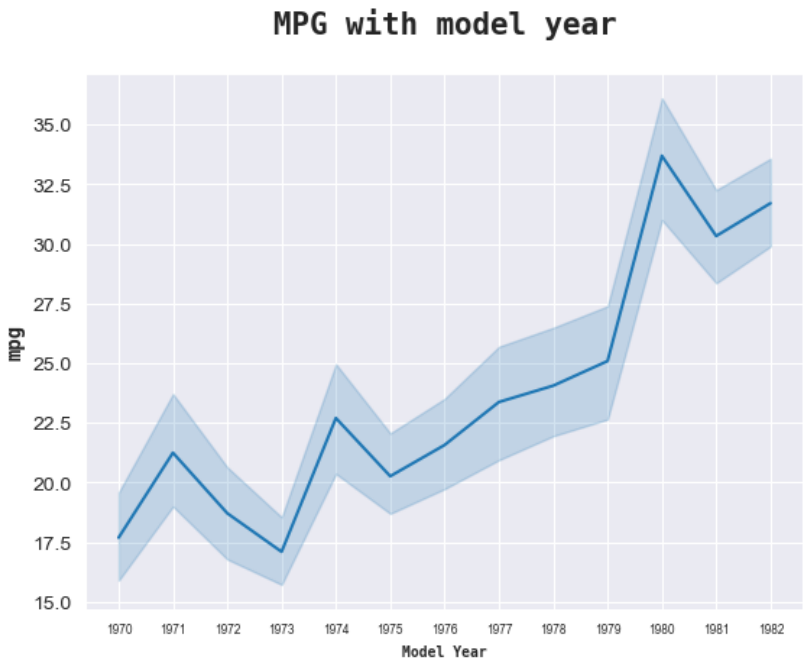
2. Data distribution of numerical variables



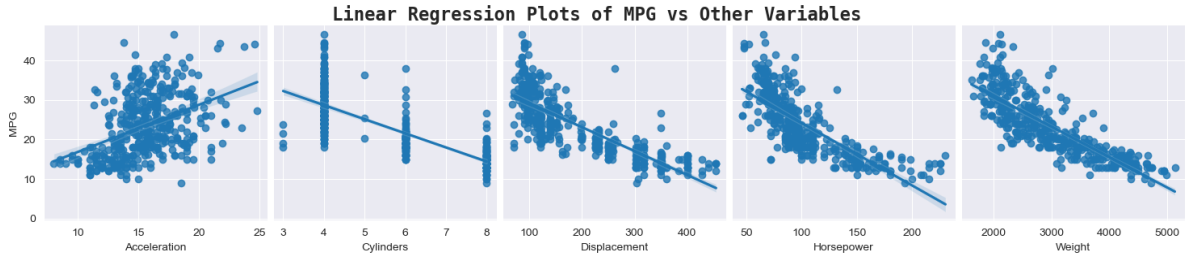
3. Heatmap of correlation



4. MPG with model year



5. Linear regression plots of MPG and other variables



6. OLS summary: Model 1

OLS Regression Results

Dep. Variable:	MPG	R-squared:	0.820
Model:	OLS	Adj. R-squared:	0.815
Method:	Least Squares	F-statistic:	201.1
Date:	Sun, 05 Mar 2023	Prob (F-statistic):	3.12e-111
Time:	11:50:21	Log-Likelihood:	-835.28
No. Observations:	318	AIC:	1687.
Df Residuals:	310	BIC:	1717.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-17.8466	5.214	-3.423	0.001	-28.106	-7.587
Cylinders	-0.3360	0.363	-0.927	0.355	-1.049	0.377
Displacement	0.0225	0.009	2.596	0.010	0.005	0.040
Horsepower	-0.0137	0.015	-0.891	0.374	-0.044	0.017
Weight	-0.0070	0.001	-9.704	0.000	-0.008	-0.006
Acceleration	0.0737	0.111	0.663	0.508	-0.145	0.293
Model Year	0.8114	0.057	14.182	0.000	0.699	0.924
US Made	-2.6104	0.543	-4.809	0.000	-3.678	-1.542

Omnibus:	12.968	Durbin-Watson:	1.896
Prob(Omnibus):	0.002	Jarque-Bera (JB):	18.009
Skew:	0.318	Prob(JB):	0.000123
Kurtosis:	3.977	Cond. No.	8.45e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.45e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

7. OLS summary: Model 3 (final model)

OLS Regression Results

Dep. Variable:	MPG	R-squared:	0.819
Model:	OLS	Adj. R-squared:	0.816
Method:	Least Squares	F-statistic:	281.8
Date:	Sun, 05 Mar 2023	Prob (F-statistic):	2.21e-113
Time:	13:46:38	Log-Likelihood:	-835.99
No. Observations:	318	AIC:	1684.
Df Residuals:	312	BIC:	1707.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-17.1067	4.527	-3.779	0.000	-26.014	-8.199
Displacement	0.0167	0.007	2.551	0.011	0.004	0.030
Horsepower	-0.0187	0.012	-1.575	0.116	-0.042	0.005
Weight	-0.0069	0.001	-10.634	0.000	-0.008	-0.006
Model Year	0.8084	0.057	14.206	0.000	0.696	0.920
US Made	-2.5876	0.541	-4.787	0.000	-3.651	-1.524

Omnibus:	13.935	Durbin-Watson:	1.892
Prob(Omnibus):	0.001	Jarque-Bera (JB):	19.824
Skew:	0.331	Prob(JB):	4.96e-05
Kurtosis:	4.028	Cond. No.	7.34e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.34e+04. This might indicate that there are
strong multicollinearity or other numerical problems.