Gurteg Singh
ECN 145

Report 2

1.) Logfile etc.

2.) Observations:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
vehtype	137393	5	CAR	81380	NaN	NaN	NaN	NaN	NaN	NaN	NaN
epatmpg	137393	NaN	NaN	NaN	26.6604	7.20752	6.4	22.3	25.8392	29.5	65.8
hybrid	136898	NaN	NaN	NaN	0.0375024	0.18999	0	0	0	0	1
domestic	136600	NaN	NaN	NaN	0.554539	0.497018	0	0	1	1	1
automaker	136600	3	Domestic	75750	NaN	NaN	NaN	NaN	NaN	NaN	NaN
vehage	133856	NaN	NaN	NaN	6.35381	4.73063	1	3	5	9	25
fueltype	137393	NaN	NaN	NaN	3.9731	0.282807	1	4	4	4	4
gscost	137393	NaN	NaN	NaN	3.06889	0.138542	2.86	2.94	3.03	3.2	3.78
rural	137393	NaN	NaN	NaN	0.295998	0.456492	0	0	0	1	1
children_under_16	137393	NaN	NaN	NaN	0.355688	0.797348	0	0	0	0	9
adults	137393	NaN	NaN	NaN	1.82955	0.739767	1	1	2	2	9
hh_income	126693	NaN	NaN	NaN	58.9579	30.4328	5	30	60	90	100
ba_grad	137059	NaN	NaN	NaN	0.465807	0.498831	0	0	0	1	1
hhstate	137393	51	TX	20464	NaN	NaN	NaN	NaN	NaN	NaN	NaN

After omitting missing values there were 122,967 counts.
Average vehicle age is about 6.33 years, and average epatmpg is about 26.69 miles per gallon (after dropping missing values).
There are about 3.88% hybrid vehicles and 55.30% domestic vehicles.

3.)

vehtype	
CAR	0.583425
LARGE SUV	0.051794
MINIVAN	0.085714
PICKUP	0.130124
SMALL SUV	0.148942

Of the market share, Cars, Large SUVs, Minivans, Pickups, and small SUVs make up about 58.34%, 5.18%, 8.57%, 13.01, and 14.89% respectively.

4.)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                epatmpg   R-squared:                       0.047
Model:                            OLS   Adj. R-squared:                  0.047
Method:                 Least Squares   F-statistic:                     1009.
Date:                Thu, 22 Oct 2020   Prob (F-statistic):               0.00
Time:                        23:11:23   Log-Likelihood:            -4.1501e+05
No. Observations:              122967   AIC:                         8.300e+05
Df Residuals:                  122960   BIC:                         8.301e+05
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept           18.3743      0.452     40.640      0.000      17.488      19.260
children_under_16   -0.6473      0.025    -25.657      0.000      -0.697      -0.598
vehage              -0.2626      0.005    -56.261      0.000      -0.272      -0.253
rural               -1.0185      0.045    -22.840      0.000      -1.106      -0.931
gscost               3.6641      0.147     24.966      0.000       3.376       3.952
hh_income           -0.0239      0.001    -29.373      0.000      -0.026      -0.022
ba_grad              1.4867      0.046     32.274      0.000       1.396       1.577
==============================================================================
Omnibus:                    54984.262   Durbin-Watson:                   1.995
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           394429.993
Skew:                           2.018   Prob(JB):                         0.00
Kurtosis:                      10.790   Cond. No.                     1.57e+03
==============================================================================
```

Here we can see that our model and explanatory variables are all statistically significant, and our R squared covers about 4% of the data variance. The coefficients of our data state that while holding all other variables constant, each additional increase in a particular variable will lead to it's unique coefficient value increase in epatmpg. For example, *holding all variables constant*, increasing a child under 16 count in a family will decrease that family's mpg by about 0.64 miles per gallon. Some of the variables are bicany and only take a value of 0 or 1. He we can also see that on average, an individual living in a rural residence has about a 1 mpg lower vehicle than that of an individual living in an urban residence. With binary variables, their counterparts are included into the intercept.

5.)

OLS Regression Results

Dep. Variable:	epatmpg	R-squared:	0.327
Model:	OLS	Adj. R-squared:	0.327
Method:	Least Squares	F-statistic:	5966.
Date:	Thu, 22 Oct 2020	Prob (F-statistic):	0.00
Time:	23:11:33	Log-Likelihood:	-3.9364e+05
No. Observations:	122967	AIC:	7.873e+05
Df Residuals:	122956	BIC:	7.874e+05
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	23.0639	0.381	60.573	0.000	22.318	23.810
C(vehtype)[T.LARGE SUV]	-10.4794	0.079	-132.613	0.000	-10.634	-10.324
C(vehtype)[T.MINIVAN]	-5.6009	0.063	-88.372	0.000	-5.725	-5.477
C(vehtype)[T.PICKUP]	-9.1892	0.053	-174.564	0.000	-9.292	-9.086
C(vehtype)[T.SMALL SUV]	-6.3962	0.050	-128.708	0.000	-6.494	-6.299
children_under_16	-0.0038	0.022	-0.173	0.862	-0.046	0.039
vehage	-0.2862	0.004	-72.804	0.000	-0.294	-0.279
rural	-0.2035	0.038	-5.396	0.000	-0.277	-0.130
gscost	2.8654	0.123	23.208	0.000	2.623	3.107
hh_income	-0.0084	0.001	-12.238	0.000	-0.010	-0.007
ba_grad	0.8091	0.039	20.783	0.000	0.733	0.885

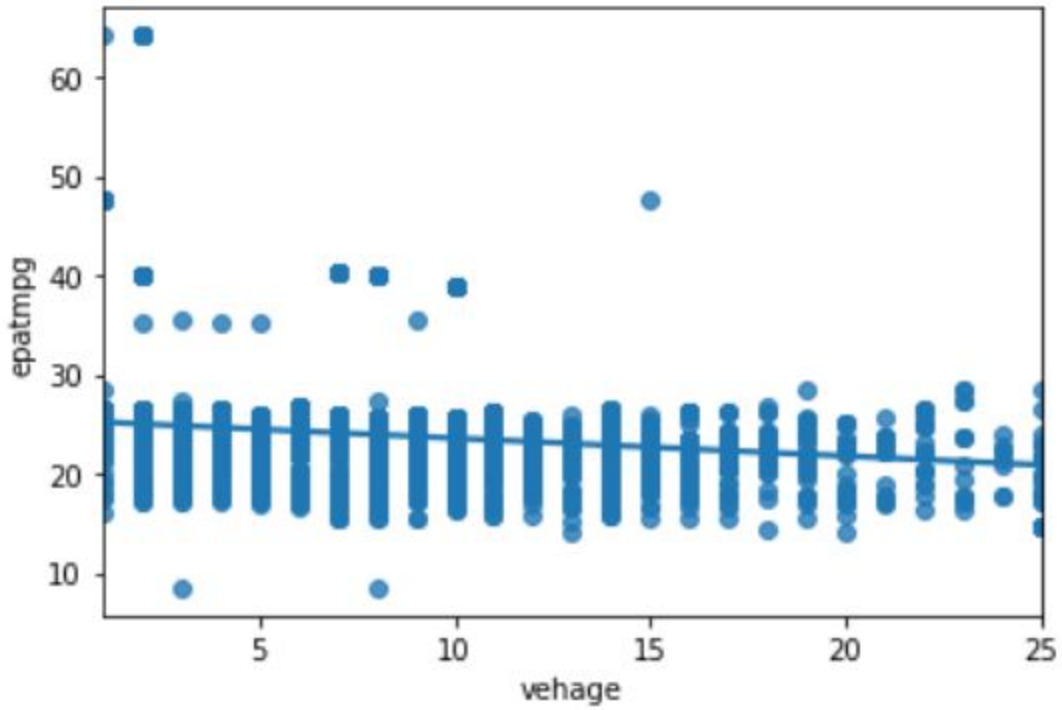
Omnibus:	69890.150	Durbin-Watson:	1.996
Prob(Omnibus):	0.000	Jarque-Bera (JB):	808532.230
Skew:	2.537	Prob(JB):	0.00
Kurtosis:	14.492	Cond. No.	1.58e+03

Here we can see that now, we can see how mpg is affected as a whole depending on the type of car an individual is driving. By including car as the base value for the mpg, we can see that every other type of vehicle is much less fuel efficient. For example, *holding all other variables constant,* a Large Suv is on average 5.6 miles per gallon less than that of a car. Since this is a dummy variable, we can set the other types of vehicle values as 0 as they are not relevant to the Large Suv.

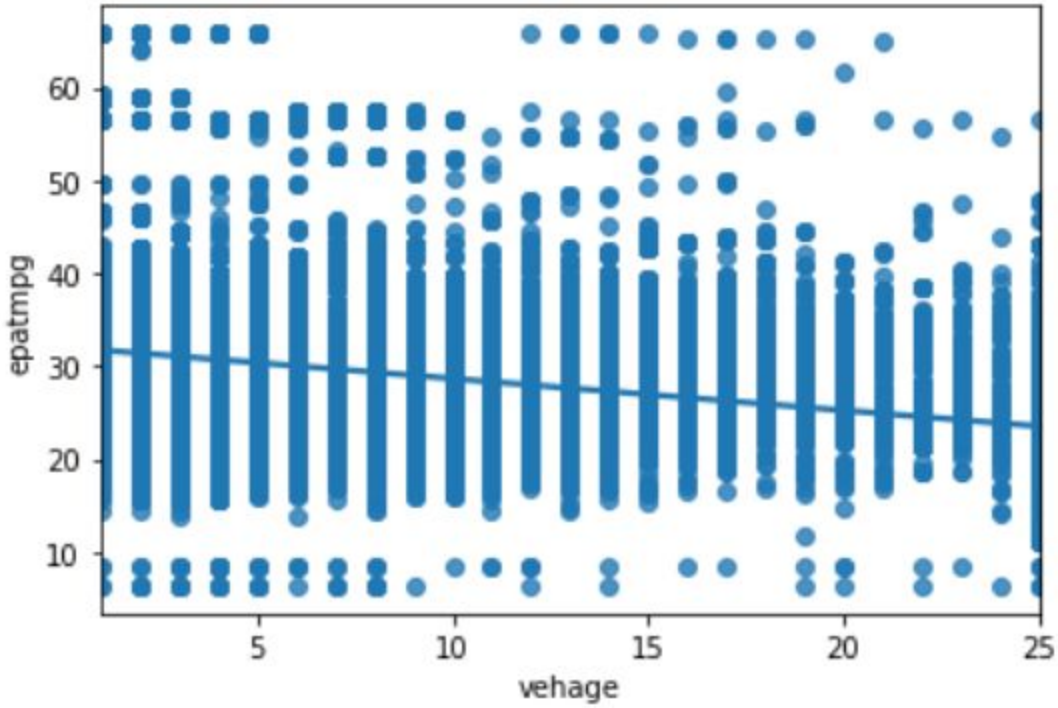
6.)
In doing so, we can now omit variable bias and see that in fact, it is not children_under_16 that effects mpg (as it is not statistically significant anymore). This may be because families with kids may tend to prefer driving minivans or other larger vehicles and that is instead, what plays the role of decreasing a vehicle's mpg.

7.)

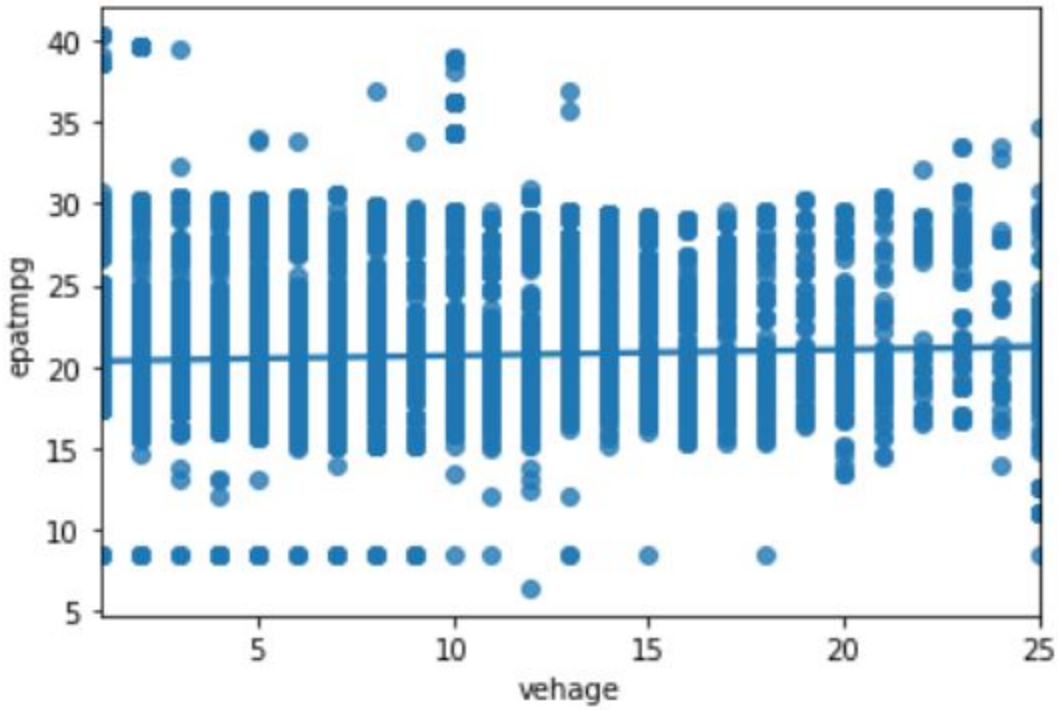
Minivan:


epatmpg
vehage
60
50
40
30
20
10
5
10
15
20
25

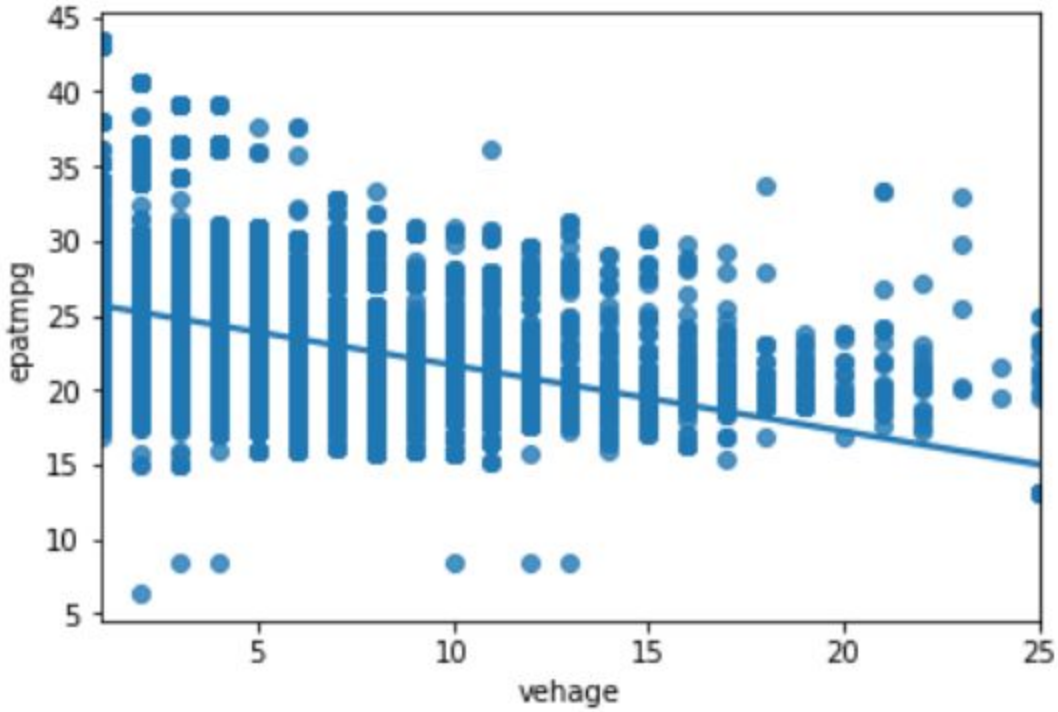
Cars:


epatmpg
vehage
60
50
40
30
20
10
5
10
15
20
25

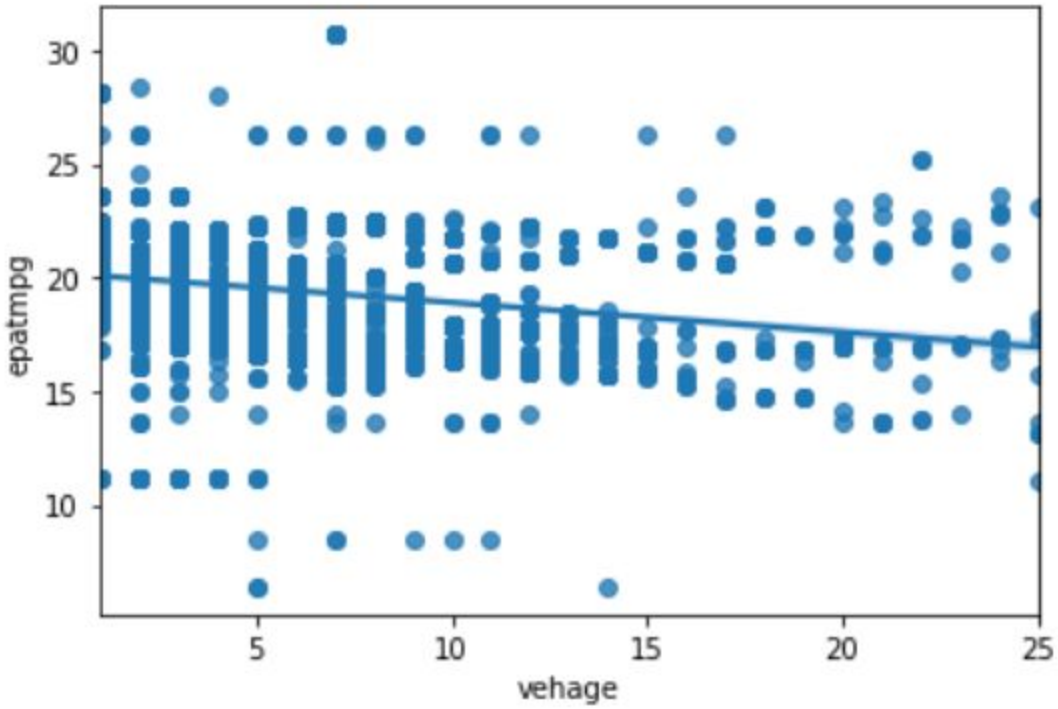
Pickup:



Small Suv:



Large Suvs:



While graphing each scatter plot we can see that the overall trend seems to be that *on average* as a vehicle gets older, the it's mileage per gallon tends to decrease as well. The only outliers are pickup trucks, that seem to have *on average* increasing mileage per gallon as they tend to age. This may be that their niche is more power based rather than consumption based, and so those who prefer pick-up trucks tend to have a different utility/ much different preference in comparison to those who prefer miles per gallon.

8.)

MNLogit Regression Results

Dep. Variable:	vehtype	No. Observations:	122967
Model:	MNLogit	Df Residuals:	122939
Method:	MLE	Df Model:	24
Date:	Thu, 22 Oct 2020	Pseudo R-squ.:	0.03914
Time:	21:10:43	Log-Likelihood:	-1.4501e+05
converged:	True	LL-Null:	-1.5091e+05
Covariance Type:	nonrobust	LLR p-value:	0.000

vehtype=LARGE SUV	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.6161	0.305	-2.023	0.043	-1.213	-0.019
vehage	0.0006	0.003	0.188	0.851	-0.006	0.007
gscost	-1.0453	0.099	-10.550	0.000	-1.239	-0.851
rural	0.3184	0.029	11.068	0.000	0.262	0.375
children_under_16	0.5894	0.013	45.876	0.000	0.564	0.615
hh_income	0.0187	0.001	32.881	0.000	0.018	0.020
ba_grad	-0.3750	0.030	-12.395	0.000	-0.434	-0.316

vehtype=MINIVAN	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2674	0.244	1.094	0.274	-0.212	0.746
vehage	0.0044	0.002	1.814	0.070	-0.000	0.009
gscost	-0.8501	0.080	-10.684	0.000	-1.006	-0.694
rural	0.1523	0.024	6.434	0.000	0.106	0.199
children_under_16	0.6952	0.010	66.335	0.000	0.675	0.716
hh_income	-0.0002	0.000	-0.364	0.716	-0.001	0.001
ba_grad	0.0248	0.025	1.008	0.314	-0.023	0.073

vehtype=PICKUP	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0596	0.202	-5.235	0.000	-1.456	-0.663
vehage	-0.0054	0.002	-2.651	0.008	-0.009	-0.001
gscost	-0.2655	0.066	-4.040	0.000	-0.394	-0.137
rural	0.6826	0.018	37.046	0.000	0.647	0.719
children_under_16	0.1999	0.012	16.682	0.000	0.176	0.223
hh_income	0.0071	0.000	19.659	0.000	0.006	0.008
ba_grad	-0.7186	0.021	-34.414	0.000	-0.760	-0.678

vehtype=SMALL SUV	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1668	0.186	-6.259	0.000	-1.532	-0.801
vehage	-0.0510	0.002	-23.242	0.000	-0.055	-0.047
gscost	-0.1600	0.060	-2.646	0.008	-0.278	-0.041
rural	0.2413	0.018	13.079	0.000	0.205	0.277
children_under_16	0.1793	0.011	16.054	0.000	0.157	0.201
hh_income	0.0079	0.000	23.278	0.000	0.007	0.009
ba_grad	-0.0406	0.019	-2.133	0.033	-0.078	-0.003

After regressing with marginal effects~

MNLogit Marginal Effects

Dep. Variable: vehtype
Method: dydx
At: overall

vehtype=CAR	dy/dx	std err	z	P>|z|	[0.025	0.975]
vehage	0.0045	0.000	13.785	0.000	0.004	0.005
gscost	0.1006	0.010	10.040	0.000	0.081	0.120
rural	-0.0863	0.003	-29.057	0.000	-0.092	-0.080
children_under_16	-0.0779	0.002	-43.922	0.000	-0.081	-0.074
hh_income	-0.0017	5.5e-05	-30.650	0.000	-0.002	-0.002
ba_grad	0.0653	0.003	20.755	0.000	0.059	0.071
vehtype=LARGE SUV	**dy/dx**	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
vehage	0.0005	0.000	2.968	0.003	0.000	0.001
gscost	-0.0421	0.005	-9.016	0.000	-0.051	-0.033
rural	0.0076	0.001	5.724	0.000	0.005	0.010
children_under_16	0.0214	0.001	37.170	0.000	0.020	0.023
hh_income	0.0008	2.75e-05	28.311	0.000	0.001	0.001
ba_grad	-0.0128	0.001	-9.068	0.000	-0.016	-0.010
vehtype=MINIVAN	**dy/dx**	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
vehage	0.0010	0.000	5.709	0.000	0.001	0.001
gscost	-0.0533	0.006	-9.102	0.000	-0.065	-0.042
rural	-0.0008	0.002	-0.455	0.649	-0.004	0.003
children_under_16	0.0446	0.001	61.120	0.000	0.043	0.046
hh_income	-0.0003	3.18e-05	-9.247	0.000	-0.000	-0.000
ba_grad	0.0123	0.002	6.880	0.000	0.009	0.016
vehtype=PICKUP	**dy/dx**	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
vehage	0.0003	0.000	1.514	0.130	-9.77e-05	0.001
gscost	-0.0098	0.007	-1.388	0.165	-0.024	0.004
rural	0.0670	0.002	34.306	0.000	0.063	0.071
children_under_16	0.0070	0.001	5.793	0.000	0.005	0.009
hh_income	0.0005	3.81e-05	13.184	0.000	0.000	0.001
ba_grad	-0.0764	0.002	-34.165	0.000	-0.081	-0.072
vehtype=SMALL SUV	**dy/dx**	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
vehage	-0.0063	0.000	-23.626	0.000	-0.007	-0.006
gscost	0.0045	0.007	0.620	0.535	-0.010	0.019
rural	0.0124	0.002	5.673	0.000	0.008	0.017
children_under_16	0.0049	0.001	3.883	0.000	0.002	0.007
hh_income	0.0007	4.06e-05	17.205	0.000	0.001	0.001
ba_grad	0.0116	0.002	5.110	0.000	0.007	0.016

Here we've done a logistic regression, which is useful for discrete probability modelling. Since usual LPM models tend to violate the [0,1] constraint probability models are confined to, we instead use the logit model to better limit the graphs with the utilization of natural log functions. A natural log graphs a function in the format: ((e^(f(x))/(1+(e^(f(x))). With this in place, a value will never be above 1, or below 0. However, due to the nature of this model, the coefficients of the results are in log units, and as such, must be derived (derivated) into percentage form. Hence, we then find the *marginal effects* of such variable coefficients in percentages, and use these to interpret our data. Now our model regresses what the probability is of a particular individual to choose a certain vehicle in regards to each variable. For example, for those living in rural residences, *holding all other variables constant*, the marginal effect - in terms of probability that the individual buys a car decreases by 8.4%. In contrast, that same individual has about a 6.7% chance increase in the probability for buying a pick-up truck. For locations with high gas prices, most of the population would probably buy cars since an additional increase in the price of gas (***HOLDING ALL OTHER VARIABLES CONSTANT - I'm tired of repeating this)*** leads to an about 10% increase (with diminishing returns) in the probability for buying a car. However, gas prices for pick-up trucks, and small suv's are no longer correlated (not statistically significant). Households with each additional child tend to have a 2.14% increase in the likelihood of buying a large suv, and a 4.46% increase in the chance they buy a minivan.

9.)
MFX:

MNLogit Marginal Effects

Dep. Variable: vehtype
Method: dydx
At: overall

vehtype=CAR	dy/dx	std err	z	P>|z|	[0.025	0.975]
vehage	0.0045	0.000	13.785	0.000	0.004	0.005
gscost	0.1006	0.010	10.040	0.000	0.081	0.120
rural	-0.0863	0.003	-29.057	0.000	-0.092	-0.080
children_under_16	-0.0779	0.002	-43.922	0.000	-0.081	-0.074
hh_income	-0.0017	5.5e-05	-30.650	0.000	-0.002	-0.002
ba_grad	0.0653	0.003	20.755	0.000	0.059	0.071

vehtype=LARGE SUV	dy/dx	std err	z	P>|z|	[0.025	0.975]
vehage	0.0005	0.000	2.968	0.003	0.000	0.001
gscost	-0.0421	0.005	-9.016	0.000	-0.051	-0.033
rural	0.0076	0.001	5.724	0.000	0.005	0.010
children_under_16	0.0214	0.001	37.170	0.000	0.020	0.023
hh_income	0.0008	2.75e-05	28.311	0.000	0.001	0.001
ba_grad	-0.0128	0.001	-9.068	0.000	-0.016	-0.010

vehtype=MINIVAN	dy/dx	std err	z	P>|z|	[0.025	0.975]
vehage	0.0010	0.000	5.709	0.000	0.001	0.001
gscost	-0.0533	0.006	-9.102	0.000	-0.065	-0.042
rural	-0.0008	0.002	-0.455	0.649	-0.004	0.003
children_under_16	0.0446	0.001	61.120	0.000	0.043	0.046
hh_income	-0.0003	3.18e-05	-9.247	0.000	-0.000	-0.000
ba_grad	0.0123	0.002	6.880	0.000	0.009	0.016

vehtype=PICKUP	dy/dx	std err	z	P>|z|	[0.025	0.975]
vehage	0.0003	0.000	1.514	0.130	-9.77e-05	0.001
gscost	-0.0098	0.007	-1.388	0.165	-0.024	0.004
rural	0.0670	0.002	34.306	0.000	0.063	0.071
children_under_16	0.0070	0.001	5.793	0.000	0.005	0.009
hh_income	0.0005	3.81e-05	13.184	0.000	0.000	0.001
ba_grad	-0.0764	0.002	-34.165	0.000	-0.081	-0.072

vehtype=SMALL SUV	dy/dx	std err	z	P>|z|	[0.025	0.975]
vehage	-0.0063	0.000	-23.626	0.000	-0.007	-0.006
gscost	0.0045	0.007	0.620	0.535	-0.010	0.019
rural	0.0124	0.002	5.673	0.000	0.008	0.017
children_under_16	0.0049	0.001	3.883	0.000	0.002	0.007
hh_income	0.0007	4.06e-05	17.205	0.000	0.001	0.001
ba_grad	0.0116	0.002	5.110	0.000	0.007	0.016

Here, we can see that with an additional increase in gasoline tax, market share of cars and small suvs increase, and the others tend to decrease.

10.)

Dep. Variable:	vehage	**R-squared:**	0.006
Model:	OLS	**Adj. R-squared:**	0.006
Method:	Least Squares	**F-statistic:**	393.8
Date:	Fri, 23 Oct 2020	**Prob (F-statistic):**	3.40e-171
Time:	04:22:19	**Log-Likelihood:**	-3.6466e+05
No. Observations:	122967	**AIC:**	7.293e+05
Df Residuals:	122964	**BIC:**	7.293e+05
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	6.5782	0.017	384.367	0.000	6.545	6.612
children_under_16	-0.4357	0.017	-26.390	0.000	-0.468	-0.403
rural	-0.2766	0.029	-9.455	0.000	-0.334	-0.219

Omnibus:	26464.455	**Durbin-Watson:**	1.989
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	52429.508
Skew:	1.303	**Prob(JB):**	0.00
Kurtosis:	4.854	**Cond. No.**	2.70

Hypothesis 0: Families with children under 16 tend to have no correlation with the age of cars that they would buy. BetaChildren_under_16 = 0 (no correlation)

Hypothesis Alternative: Families with children tend to buy newer models (Beta Children_under_16 is *not* = 0) (correlation)
Here I tested out to see if families with children would buy newer/safer cars for the children. As it turns out, as vehicles tend to age, less and less families seem to have a use for them. Each variable, as it turns out, tends to be statistically significant, so we can reject the null hypothesis, and say This is interesting as it shows that families tend to utilize newer models of vehicles for increasing amounts of children, as each additional child creates an about 0.43 year decrease in the vehicle age. Interestingly so, when compared to urban cities, it seems that rural families tend to decrease the use of older vehicles as well, by a factor of 0.27 *holding all other variables constant*.