Assignment #5
Ji Jung

Introduction
The goal of this assignment is to design a model with some predictive explanatory value for the sale price of a home given some contextual information about many observed sales. Further variables, especially categorical will be considered. Analysis will be conducted using adjusted R-Squared, Mallow's Cp, Rsquared, Forward, Backward and Stepwise then compared for better modeling.

Results
PART A: Dummy Coding of Categorical Variables
Selecting a categorical variable

A categorical variable of HouseStyle is chosen as it contains eight categories.

1Story	One story
1.5Fin	One and one-half story: 2nd level finished
1.5Unf	One and one-half story: 2nd level unfinished
2Story	Two story
2.5Fin	Two and one-half story: 2nd level finished
2.5Unf	Two and one-half story: 2nd level unfinished
SFoyer	Split Foyer
SLvl	Split Level

The mean of HouseStyle is calculated in Table 1.

N	HouseStyle	SalePrice Mean
1108	1Story	174392.88
231	1.5Fin	140087.45
19	1.5Unf	109663.16
690	2Story	201178.07
5	2.5Fin	253000.00
21	2.5Unf	181900.00

63	SFoyer	142558.10
103	SLvl	167385.78

Table 1: Mean SalePrice given HouseStyle

The linear regression model is $\text{SalesPrice} = \beta_0 + \beta_1 \text{HouseStyle} + \epsilon$.

Dummy Code

The last category of SLvl is chosen to be the basis of interpretation. As such, the remainders are given dummy codes.

HouseStyle	hs_1	hs_2	hs_3	hs_4	hs_5	hs_6	hs_7
1Story	1	0	0	0	0	0	0
1.5Fin	0	1	0	0	0	0	0
1.5Unf	0	0	1	0	0	0	0
2Story	0	0	0	1	0	0	0
2.5Fin	0	0	0	0	1	0	0
2.5Unf	0	0	0	0	0	1	0
SFoyer	0	0	0	0	0	0	1
SLvl	0	0	0	0	0	0	0

Table 2: Modeling HouseStyle as an indicator Variable

The regression model is $\text{SalePrice} = \beta_0 + \beta_1 h_1 + \beta_2 h_2 + \beta_3 h_3 + \beta_4 h_4 + \beta_5 h_5 + \beta_6 h_6 + \beta_7 h_7 + \epsilon$. With parameter estimates, a model can be construct.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > $\lvert t \rvert$
Intercept	1	167386	6752.11484	24.79	<.0001
hs_1	1	7007.10056	7058.98098	0.99	0.3210
hs_2	1	-27298	8119.08454	-3.36	0.0008
hs_3	1	-57723	17110	-3.37	0.0008
hs_4	1	33792	7238.55483	4.67	<.0001

hs_5	1	85614	31381	2.73	0.0064
hs_6	1	14514	16407	0.88	0.3765
hs_7	1	-24828	10960	-2.27	0.0236

Source	
Root MSE	68526
R-Square	0.0811
Adj R-Square	0.0782
F Value	28.13

Table 3: Model Parameter Estimates for SalePrice = HouseStyle Indicator Variables

The configured model is $\text{SalesPrice} = 167{,}386 + 7{,}007.10056hs_1 - 2{,}298hs_2 - 57{,}723hs_3 + 37{,}792hs_4 + 85{,}614hs_5 + 14514hs_6 - 24{,}828hs_7$

Interpretations of this model is that if the HouseStyle is 1, or '1Story' then the model becomes:

SalePrice = 167,386 + 7,007.10056

If the HouseStyle is '1Story', then the SalePrice in this model is $174,387.10, Looking back at our mean table we see that '1Story' had a mean of $174,392.88, If the HouseStyle is 2, or '1.5Fin' then the model becomes:

SalePrice = 167,386 – 27,298

If the HouseStyle is '1.5Fin', then the SalePrice in this model is $140,088.00, Looking back at our mean table we see that '1.5Fin' had a mean of $140,087.45, If the HouseStyle is 3, or '1.5Unf' then the model becomes:

SalePrice = 167,386 – 57,723

If the HouseStyle is '1.5Unf', then the SalePrice in this model is $109,663.00, Looking back at our mean table we see that '1.5Unf' had a mean of $109,663.16, If the HouseStyle is 4, or '2Story' then the model becomes:

SalePrice = 167,386 + 33,792

If the HouseStyle is '2Story', then the SalePrice in this model is $201,178.00, Looking back at our mean table we see that '2Story' had a mean of $201,178.07, If the HouseStyle is 5, or '2.5Fin' then the model becomes:

SalePrice = 167,386 + 85,614

If the HouseStyle is '2.5Fin', then the SalePrice in this model is $253,000.00, Looking back at our mean table we see that '2.5Fin' had a mean of $253,000.00, If the HouseStyle is 6, or '2.5Unf' then the model becomes:

SalePrice = 167,386 + 14,514

If the HouseStyle is '2.5Unf', then the SalePrice in this model is $181,900.00, Looking back at our mean table we see that '2.5Unf' had a mean of $181,900.00, If the HouseStyle is 7, or 'SFoyer' then the model becomes:

SalePrice = 167,386 – 24,828

If the HouseStyle is 'SFoyer', then the SalePrice in this model is $142,558.00, Looking back at our mean table we see that 'SFoyer' had a mean of $142,558.10,If the HouseStyle is 8 or 'SLvl' then the model becomes:

SalePrice = 167,386

That is to say, if the HouseStyle is 'SLvl', then the SalePrice in this model is $167,386, Looking back at our mean table we see that 'SLvl' had a mean of $167,385.78.

The Hypothesis Tests

$$H_0 : \beta_{1\ldots7} = 0 \; versus \; H_1 : \beta_{1\ldots7} = 0$$

Variables hs_5 and hs_6 do not yield statistically significant results while hs_2 , hs_3 , hs_4 , hs_5 and hs_7 show statistically significant results. It could be possible to fathom a method to write out the hypothesis testing in a more separate-but-joint fashion, but cannot find a reference to explain. This model may be unreliable to work with and interpret.

New Variable, New Dummy Codes

GarageType is chosen for it holds seven categories.

N	GarageType	SalePrice Mean
13	2Types	169446.15
1342	Attchd	198923.58
21	Basment	154383.33
135	BuiltIn	227725.25
8	CarPort	103943.75
617	Detchd	134038.98

104	NA	106865.14

Table 4: Mean SalePrice given GarageType

PART B: Automated Variable Selection Procedures

The Analysis

From previous assignments, strong continuous variables chosen along with discussed categorical variables.

Continuous Variable	Correlation to SalePrice	Prob > $\|r\|$ under H_0: ρ=0	Number of Observations
GrLivArea	0.70678	<0.0001	2930
GarageArea	0.64040	<0.0001	2929
TotalBsmtSF	0.63228	<0.0001	2929
FirstFlrSF	0.62168	<0.0001	2930
MasVnrArea	0.50828	<0.0001	2907
BsmtFinSF1	0.43291	<0.0001	2929
BsmtUnfSF	0.18286	<0.0001	2929

Table 5: Top seven continuous variable correlation to SalePrice

After running through analysis, a summary table can be constructed.

Model	Cont.	Ind.	Root MSE	C_p	R-Square	Adj. R-Square	F Value	AIC	BIC
Adj. R-Square	4	8	32905.57	8.9007	0.7892	0.7881	-	46382.3907	46384.5916
Mallow's C_p	4	7	32909.82	8.4711	0.7891	0.7880	-	46381.9726	46384.1409
AIC	4	7	32909.82	8.4711	0.7891	0.7880	-	46381.9726	46384.1409
Forward	5	7	32910.48	11.5634	0.78935	0.78802	592.60	46385.04	46387.29
Backward	4	7	32909.82	8.47108	0.78907	0.78803	753.98	46381.97	46384.14
Stepwise	4	7	32909.82	8.47108	0.78907	0.78803	753.98	46381.97	46384.14

Table 6: Automatic Variable Selection Model Comparison

Mallow's C_p and AIC would result in models of greatly reduced parameters. However, the results show that these models all performed well by incorporating almost all the continuous variables in them.

For Mallow's C_p and AIC, same results are calculated to minimize AIC. Mallow's C_p method found the models with the lowest AIC thus the same model is used for the AIC selection criteria.

Aside from model complexity, there is concern that OverallQual is a subjective categorical measurement, as opposed to HouseStyle which is an observable categorical measurement. This likely means that the model would be more useful in inference than for prediction. There are in-sample observations of OverallQual, but with out-of- samples, there is no systematic way of observing and characterizing Overallqual. It is not as obviously measurable as an explicit feature of the premises such as HouseStyle and GarageType.

Overall all models, aside from the forward selection, excluded the FirstFlrSF variable. The Forward selection logically had the highest C_p value, likely due in part to the inclusion of this model. FirstFlrSF variable is excluded as the there is an overall performance difference when considering the Adj. R-Square, and C_p is no different than Backward and Stepwise methods. In addition, Backward and Stepwise have resulted in models that are equivalent.

Indicator Variable Inclusion

SalePrice = −14106 + 70.86187 × GrLivArea + 79.56383 × GarageArea+52.14204 × TotalBsmtSF + 43.57188 × MasVnrArea−62056 × gt_1 − 151 × gt_2 − 4300 × gt_3 + 9457 × gt_4−15012 × gt_5 − 14633 × gt_6 + 0 × gt_7−2063 × hs_1 − 15081 × hs_2 + 3112 × hs_3 − 2752 × hs_4 −30202 × hs_5 − 21620 × hs_6 + 9301 × hs_7 + 0 × hs_8

Source	
Root MSE	32934
R-Square	0.7893
Adj. R-Square	0.7877
F Value	487.32

Table 7: Model Performance

It should be noted that both gt_4 and hs_8 were set to zero because they make the overall use of each set of indicator variables a linear combination. SAS automatically marks them as zero. Secondarily, every single indicator variable is found to be bias by SAS and only two are statistically significant (gt_1 and hs_2). All continuous variables included in this model are found to be statistically significant.

Interpretation of this model is difficult, it must be considered what a single unit increase means for each continuous variable, as well as the combinations for each indicator variable. The model is less interpretable for our business owners, and for the complexity increase there is no justifiable explanatory performance.

The model complexity of interpretation, as well as the relative performance (based on the Adj. R-Square and F Value criteria) are considered, considering other forays into this data set, to be poor. Hence, the indicator parameters is excluded.

SalePrice =
−27086+70.65635×GrLivArea+79.08423×GarageArea+57.38357×TotalBsmtSF+49.92249×MasVnrArea

Source	
Root MSE	34449
R-Square	0.7682
Adj. R-Square	0.7677
F Value	1842.16

Table 8: Model Performance

Dependent variables are all statistically significant and that the F Value increases significantly compared to the last model.

PART C: Validation Framework

Model Comparison

Model	Cont.	Ind.	Root MSE	C_p	R-Square	Adj. R-Square	F Value	AIC	BIC
Adj. R-Square	4	8	32905.57	8.9007	0.7892	0.7881	-	46382.3907	46384.5916
Mallow's C_p	4	7	32909.82	8.4711	0.7891	0.7880	-	46381.9726	46384.1409
AIC	4	7	32909.82	8.4711	0.7891	0.7880	-	46381.9726	46384.1409
Forward	5	7	32910.48	11.5634	0.78935	0.78802	592.60	46385.04	46387.29
Backward	4	7	32909.82	8.47108	0.78907	0.78803	753.98	46381.97	46384.14
Stepwise	4	7	32909.82	8.47108	0.78907	0.78803	753.98	46381.97	46384.14

Table 9 : Automatic Variable Selection Model Comparison

Model	Cont.	Ind.	Root MSE	C_p	R-Square	Adj. R-Square	F Value	AIC	BIC
Adj. R-Square	4	8	33521.67	11.5843	0.77270	0.7706	-	31795.7656	31798.1320
Mallow's C_p	4	7	32909.82	9.6048	0.7724	0.7706	-	31793.8094	31796.0918
AIC	4	7	32909.82	9.6048	0.7724	0.7706	-	31793.8094	31796.0918
Forward	5	7	32910.48	11.5634	0.78935	0.78802	592.60	46385.04	46387.29

Backward	4	7	33533.15	9.61115	0.77209	0.77044	465.97	31793.84	31796.06
Stepwise	4	7	33533.15	9.61115	0.77209	0.77044	465.97	31796.06	31793.84

Table 10: Automatic Variable Selection Model Comparison (Training Data)

Conclusions

There were significant challenges with this data set when using categorical variables. While these challenges are not unique to this data set, it seems that categorical variables are quite unwieldy when it comes to model complexity and interpretation.

Flexibility in modeling is key, a lot if idiosyncrasies in a data set can be teased out through re-examination and implementation of different models. Using a categorical variable that is a Lickert scale likely means observations are subjective in nature. It might be wise to first consider those as variables.

It may be a better method for utilizing categorical variables be to perform an extensive EDA with the continuous variables and then there is a confident model, adding indicator variables and check performance, complexity and interpret-ability.