Final Project: An Analysis of Loan Data and Its Categorical Factors

Tejaswi Rachapudi
Group 2

Introduction

The amount of loan a borrower receives is generally dependent on many factors. Using data from the LendingClub (found here), we wish to determine which factors affect the loan amount. This dataset contains data for loans issued from 2007 to 2018. The original dataset has 2,260,668 observations with 145 variables, but we choose to subset the data by randomly sampling 925 observations per month, creating a subset of 55,500 observations. Of the 145 variables, we only choose to analyze 5 variables: 1 continuous and 4 categorical. The variables we use are as follows:

- **loan_amnt**: the amount of money requested by the borrower
- **grade:** the loan grade assigned by the LendingClub, calculated using a formula that takes into account credit score and other indicators of credit risk (details can be found here)
- **homeownership**: homeownership status provided by borrower during registration
- **term**: the number of payments on the loan
- **purpose**: the borrower's reason for loan request

Our goal is to determine the impact of categorical predictors on a continuous response. More specifically, we would like to determine how the loan amount differs within each type of grade, homeownership, term, and purpose.

We can do so by performing an analysis of variance (ANOVA), with loan_amnt as the continuous response, and grade, homeownership, term, and purpose as the categorical predictors. A simple tabulation of each predictor variable tells us the different levels of each predictor, and that we are working with unbalanced data. The tabulations are as follows:

	loan_amnt		
	Mean	Std	N
grade			
A	14049.85	8666.60	9718
B	13508.30	8774.06	15428
C	14309.14	9096.51	15997
D	15391.59	9248.36	8874
E	17042.05	9420.16	3889
F	19027.85	9381.52	1191
G	19430.46	8820.66	403

	loan_amnt		
	Mean	Std	N
term			
36 months	12543.54	8493.25	41707
60 months	20593.54	8010.04	13793

	loan_amnt		
	Mean	Std	N
homeownership			
ANY	15032.29	10814.69	24
MORTGAGE	15995.19	9389.82	26860
OWN	13972.97	8874.33	6606
RENT	12944.26	8416.41	22010

	loan_amnt		
	Mean	Std	N
purpose			
car	9546.67	7160.78	600
credit_card	14618.05	8617.53	11986
debt_consolidation	15370.85	8878.09	31655
home_improvement	14640.48	9747.62	3759
house	15951.71	10667.63	425
major_purchase	13624.19	10607.08	1261
medical	8960.71	7447.60	794
moving	8143.33	6607.06	405
other	10554.04	8914.10	3623
renewable_energy	9378.47	7292.11	36
small_business	17151.53	11150.93	523
vacation	6344.21	5844.71	432
wedding	3875.00	.	1

Methods

Since we are working with categorical variables and unbalanced data, we must use **proc glm** for n-way analyses instead of **proc anova**, which is used for balanced data. First, we start with one-way analyses with each predictor as the main effect and determine what this tells us about the significance of the models. Then, we look at all combinations of two-way analyses and possible significant models. We then create a model with all four predictors and determine the significance of that model. Lastly, we use least square means as an estimate to give equal weight to each observation. Using the Tukey-Kramer multiple comparison, we can determine significant groups for each predictor. From plots of the least square means, we see what conclusions we can make.

Results

The one-way analyses can be found in detail in Appendix a. Since all the predictors are significant in their individual models, we proceed further with all combinations of two-way analyses, which can be found in Appendix b. Since we have that all predictors are significant once again, we create a model with all four predictors and perform an analysis of variance. We test a few models with the predictors in different order, and see that the model with all 4 predictors is significant. Below, we see that the model with more than just error has a p-value of <0.0001, which means the model is significant.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	22	868192053855	39463275175	592.20	<.0001
Error	55477	3.6968704E12	66637893.2		
Corrected Total	55499	4.5650625E12			

The R_2 value indicates that 19% of the variation in loan_amnt is accounted for by grade, homeownership, purpose, and term.

R-Square	Coeff Var	Root MSE	loan_amnt Mean
0.190182	56.12708	8163.204	14544.15

The Type I and Type II SS for the model with grade, homeownership, purpose, and term have p-values of <0.0001, which means we reject the null that the coefficients of these predictors are 0, indicating the model with these predictors is significant. The output for the other models with the order of predictors changed is not shown, but all the models have similar output and are significant.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
grade	6	84015835623	14002639271	210.13	<.0001
homeownership	3	122920667836	40973555945	614.87	<.0001
purpose	12	164437079098	13703089925	205.64	<.0001
term	1	496818471298	496818471298	7455.49	<.0001

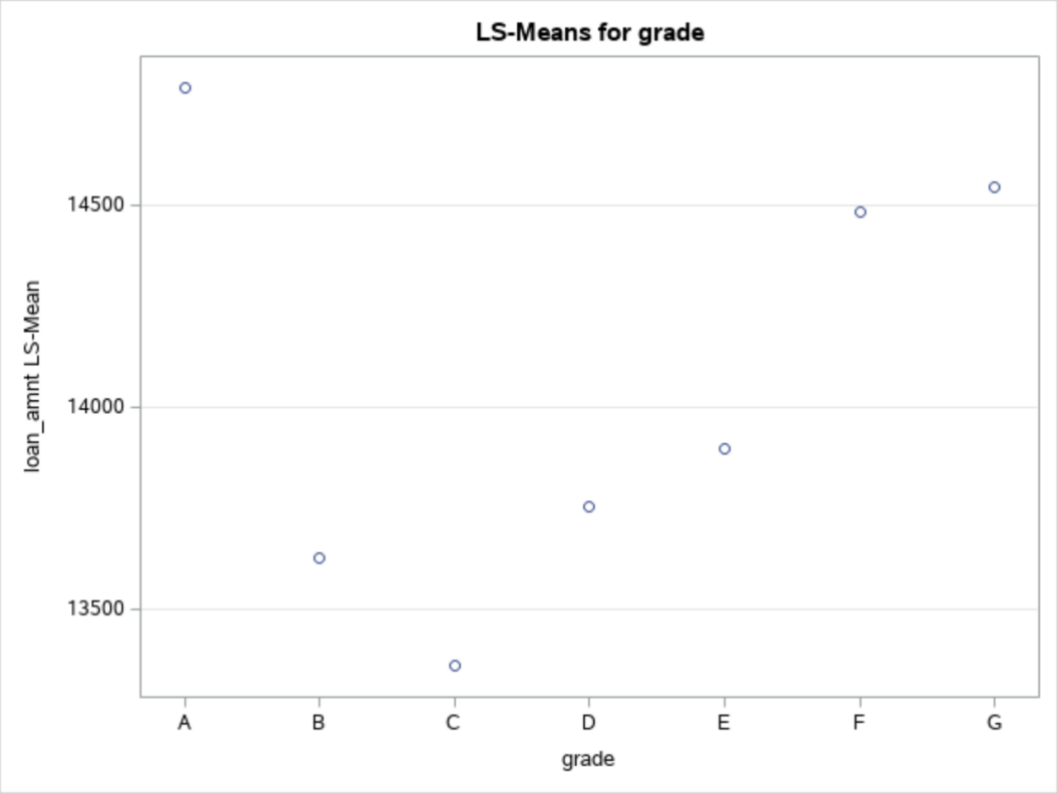
Source	DF	Type III SS	Mean Square	F Value	Pr > F
grade	6	13532045799	2255340966.5	33.84	<.0001
homeownership	3	60252973935	20084324645	301.39	<.0001
purpose	12	110647772206	9220647683.9	138.37	<.0001
term	1	496818471298	496818471298	7455.49	<.0001

Now that we know a model with all four predictors is significant, we can do further analysis to see how the loan_amnt changes within each predictor. We start with grade. The most significant groups are those that do not contain 0 in the 95% confidence intervals. Grade A has the largest difference in means with all other grades, and most of the groups containing grade A are significant, since the intervals do not contain 0. Only one group comparison with grade B is significant, while most comparisons with grade C are significant. No group comparisons with grade D, E, F, and G are significant.

grade	loan_amnt LSMEAN	LSMEAN Number
A	14791.1924	1
B	13626.5677	2
C	13359.7401	3
D	13755.3586	4
E	13898.8101	5
F	14482.3908	6
G	14546.6809	7

Least Squares Means for Effect grade				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	1164.624690	850.455414	1478.793966
1	3	1431.452240	1110.802374	1752.102107
1	4	1035.833837	664.874895	1406.792779
1	5	892.382241	406.992890	1377.771592
1	6	308.801557	-458.231430	1075.834545
1	7	244.511515	-999.298104	1488.321134
2	3	266.827550	-8.714436	542.369536
2	4	-128.790853	-458.505248	200.923542
2	5	-272.242449	-722.972125	178.487226
2	6	-855.823133	-1599.854352	-111.791914
2	7	-920.113175	-2149.445734	309.219383
3	4	-395.618403	-715.741167	-75.495640
3	5	-539.070000	-977.474060	-100.665939
3	6	-1122.650683	-1856.782286	-388.519080
3	7	-1186.940726	-2409.718592	35.837141
4	5	-143.451596	-609.529393	322.626200
4	6	-727.032280	-1476.323363	22.258804
4	7	-791.322322	-2022.796584	440.151939
5	6	-583.580683	-1382.129884	214.968517
5	7	-647.870726	-1909.226123	613.484671
6	7	-64.290043	-1451.685587	1323.105501

The graph below shows that grade A has the highest loan amounts. It seems that generally, as grade increases, loan_amnt also increases.

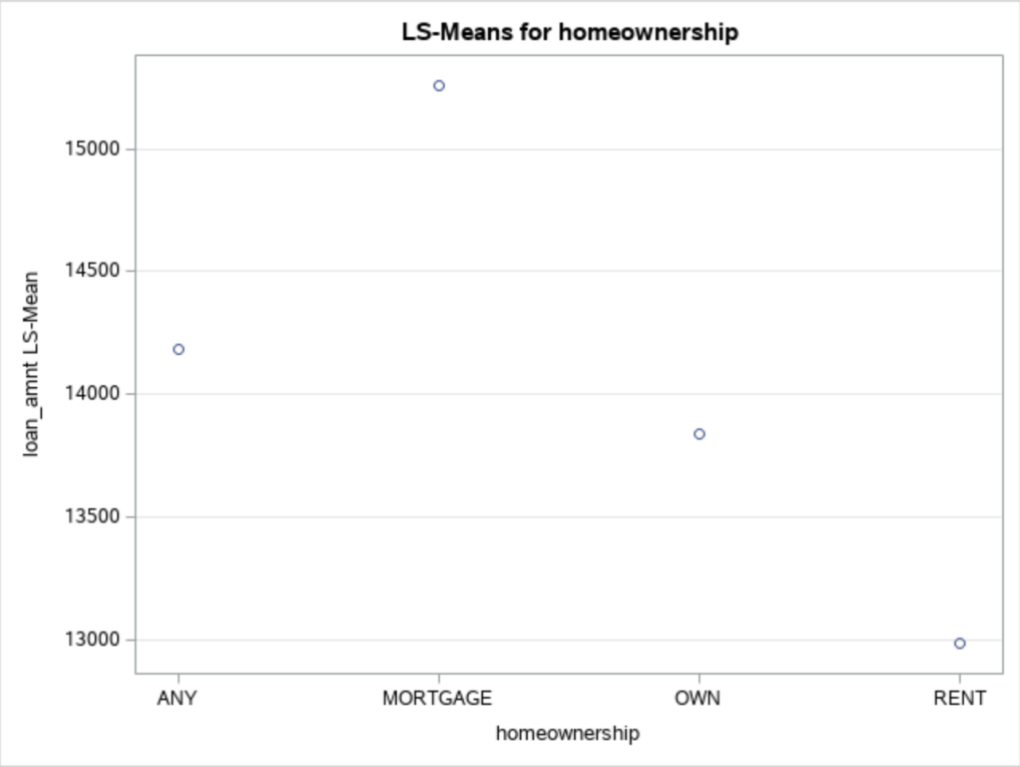


Next, we analyze the least square means for the predictor homeownership. The groups mortgage-own, mortgage-rent, and own-rent are all significant, while homeownership type "any" appears to be insignificant, as 0 is in the confidence interval for all groups with "any".

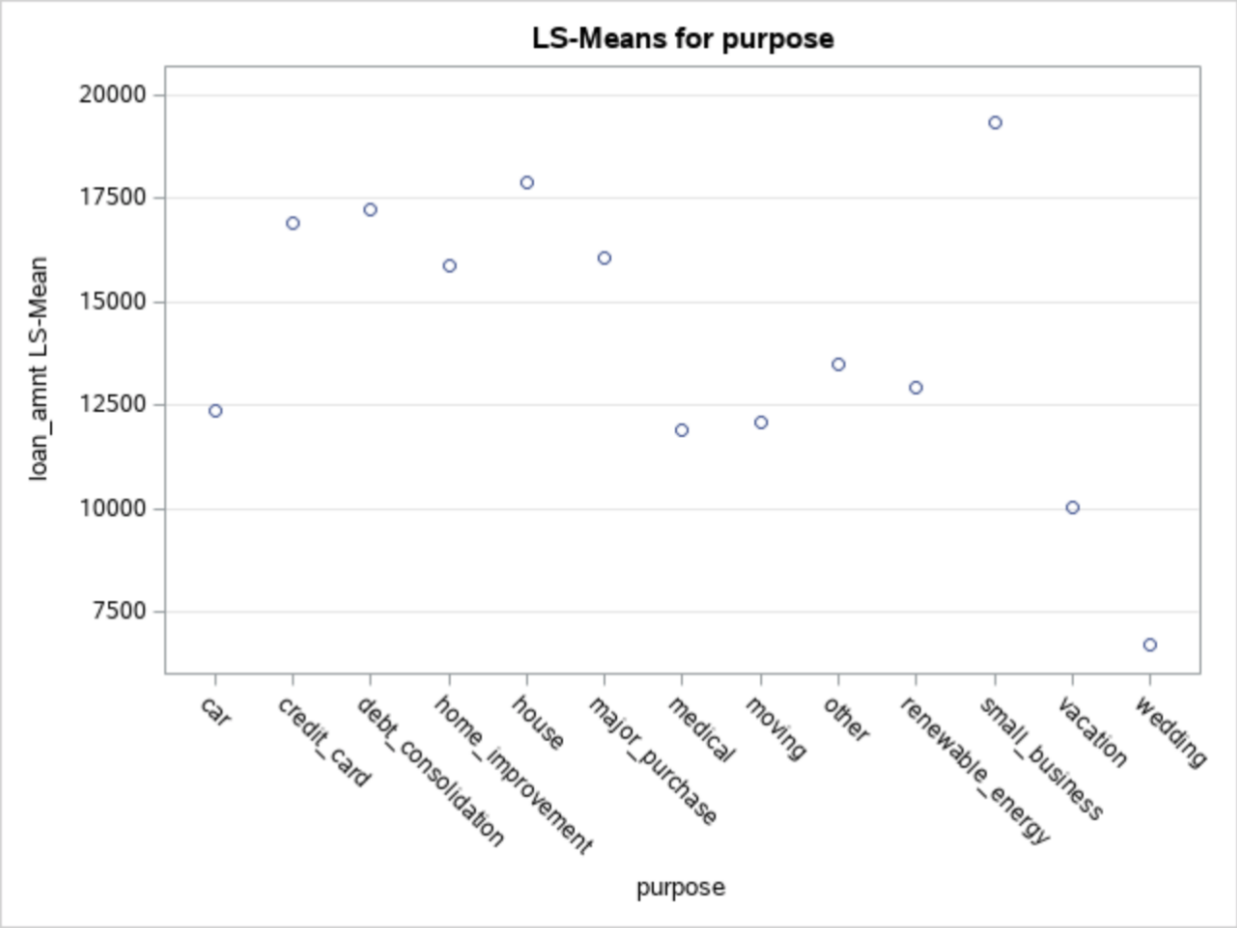
homeownership	loan_amnt LSMEAN	LSMEAN Number
ANY	14182.5403	1
MORTGAGE	15259.6448	2
OWN	13836.9200	3
RENT	12984.1753	4

Least Squares Means for Effect homeownership				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-1077.104540	-5360.267384	3206.058305
1	3	345.620286	-3943.526400	4634.766972
1	4	1198.364925	-3085.602271	5482.332121
2	3	1422.724826	1133.663720	1711.785932
2	4	2275.469464	2079.474439	2471.464489
3	4	852.744639	556.528607	1148.960670

The plot below shows that those with mortgages receive larger loans, while those who are renting receive smaller loans.



Next, we look at purpose. Each level of purpose has significant group comparisons. Although we will not be going into detail about which groups specifically, we can see from the plot below that borrowers receive larger loans for small businesses, houses, credit card, and debt consolidation. Borrowers receive smaller loans for weddings and vacations.

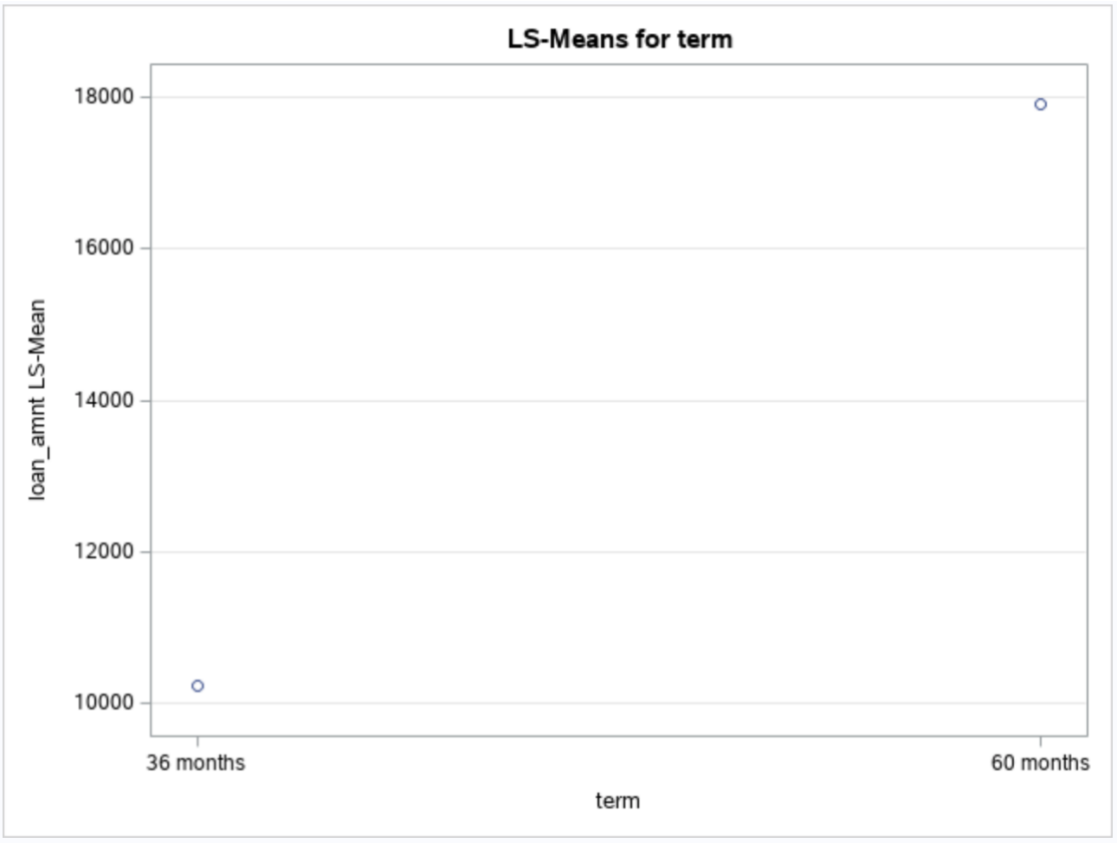


The analysis for term is very straightforward. There is a clear difference between taking out a loan for 36 months vs. 60 months. The Tukey-Kramer Adjustment for Multiple Comparisons test below shows that we must reject the null that the lsmeans for a 36-month term is equivalent to the lsmeans for a 60-month term. In other words, a very small p-value indicates that there is a significant difference between a 36-month term and a 60-month term. This observation is further supported by the confidence interval, which does not contain 0.

term	loan_amnt LSMEAN	H0:LSMean1=LSMean2
		Pr > |t|
36 months	10224.0620	<.0001
60 months	17907.5782	

Least Squares Means for Effect term				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-7683.516148	-7857.929366	-7509.102930

The plot below displays the clear distinction between the two levels of term. It is apparent that a loan taken out for 60 months is larger than a loan taken out for 36 months.



Conclusions

From the Results section, we have determined that grade, homeownership, purpose, and term are all significant predictors of loan_amnt. Our goal was to see how loan amounts differ within each loan grade, each homeownership type, each purpose, and each term. After performing these analyses, we reach many conclusions. It seems that having a loan grade of A results in greater loan amounts. This must mean that borrowers who have grade A have better credit scores, and thus are given larger loans in confidence that they will pay back their loans in good time. Also, taking out a loan for 60 months results in a larger loan amount than a loan taken out for only 36 months. Larger loans are taken out for those with small businesses, those with credit cards and debt, while smaller loans are taken out for weddings and vacations. It seems that those who have mortgages take out larger loans, while those who are renting take out smaller loans. Some improvements that could be made include analyzing with interaction terms to see if there are any possible interactions between these predictors. A surprising finding is that average loan amounts increase as loan grade increases. We would have expected the loan amounts to decrease as loan grade increases, as a loan grade of A is better than a loan grade of G. This analysis gives investors and others a good idea of how these different predictors affect how much money a borrower receives.

Appendix

a. Starting with grade as the predictor for loan_amnt, we see below that the p-value is <0.0001, which means the model with more than just error is significant. The F value tells us there is 173 times as much variation than expected, which also means the model is significant.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	84015835623	14002639271	173.41	<.0001
Error	55493	4.4810466E12	80749763.381		
Corrected Total	55499	4.5650625E12			

The R_2 value is 0.018, meaning only 1.8% of the variation in loan_amnt is explained by grade, which does not seem to be practically significant. However, the small p-value in the table below signifies that grade is a significant predictor of loan_amnt. We conclude that the beta for grade is not 0 in the model, and therefore, it is significant.

R-Square	Coeff Var	Root MSE	loan_amnt Mean
0.018404	61.78491	8986.087	14544.15

Source	DF	Anova SS	Mean Square	F Value	Pr > F
grade	6	84015835623	14002639271	173.41	<.0001

Next, we perform a one-way analysis with homeownership as the main effect. The p-value is extremely small again, so we conclude the model with more than just error is significant. The F value tells us there is 478 times as much variation than expected, which also means the model is significant.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	115052753929	38350917976	478.27	<.0001
Error	55496	4.4500097E12	80186134.154		
Corrected Total	55499	4.5650625E12			

The R_2 value is 0.025, indicating only 2.5% of the variation in loan_amnt is explained by homeownership, which is not practically significant, but the next table indicates homeownership is significant.

R-Square	Coeff Var	Root MSE	loan_amnt Mean
0.025203	61.56891	8954.671	14544.15

Source	DF	Anova SS	Mean Square	F Value	Pr > F
homeownership	3	115052753929	38350917976	478.27	<.0001

We perform the same analysis with purpose as the main effect and we see once again that the model with more than just error is significant. An F value of 180 indicates there is 180 times as much variation than expected.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	171333487591	14277790633	180.31	<.0001
Error	55487	4.393729E12	79184835.499		
Corrected Total	55499	4.5650625E12			

An R_2 value of 0.0375 tells us that only 3.75% of the variation in loan_amnt is accounted for by the predictor purpose. We see from the next table that purpose is a significant predictor, as it has an extremely small p-value.

R-Square	Coeff Var	Root MSE	loan_amnt Mean
0.037531	61.18329	8898.586	14544.15

Source	DF	Anova SS	Mean Square	F Value	Pr > F
purpose	12	171333487591	14277790633	180.31	<.0001

Lastly, we analyze the model with term as the only predictor. We notice below that the model with more than only error is significant again, as the p-value is <0.0001. There is 9,574 times as much variation than expected.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	671685781941	671685781941	9574.52	<.0001
Error	55498	3.8933767E12	70153459.097		
Corrected Total	55499	4.5650625E12			

The R_2 value tells us that 14.7% of the variance in loan_amnt is explained by term. We observe that term is a significant predictor as its p-value is extremely small. Therefore, the beta for term does not equal 0, and is in fact significant in the model.

R-Square	Coeff Var	Root MSE	loan_amnt Mean
0.147136	57.58858	8375.766	14544.15

Source	DF	Anova SS	Mean Square	F Value	Pr > F
term	1	671685781941	671685781941	9574.52	<.0001

b. First, we analyze the model with grade and homeownership. The first table tells us that the model with more than only error is significant, since it has a p-value of <0.0001. There is 293 times as much variation than expected.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	206936503460	22992944829	292.76	<.0001
Error	55490	4.358126E12	78538943.079		
Corrected Total	55499	4.5650625E12			

The R_2 value indicates only 4.5% of the variation in loan_amnt is accounted for by both grade and homeownership. We can take a look at the Type I and Type III sum of squares to determine the amount of variation explained by the model when a term is added to or dropped from the model, respectively. The Type I and Type II SS for the model with both grade and homeownership have p-values of <0.0001, which means we reject the null that the coefficients of grade and homeownership are 0, indicating the model with these predictors is significant.

R-Square	Coeff Var	Root MSE	loan_amnt Mean
0.045330	60.93325	8862.220	14544.15

Source	DF	Type I SS	Mean Square	F Value	Pr > F
grade	6	84015835623	14002639271	178.29	<.0001
homeownership	3	122920667836	40973555945	521.70	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
grade	6	91883749531	15313958255	194.99	<.0001
homeownership	3	122920667836	40973555945	521.70	<.0001

We do the same for the following models:

- Grade and purpose
- Grade and term
- Homeownership and purpose
- Homeownership and term
- Purpose and term

We arrive at the same conclusions as before (output is excluded for the sake of space). All models are significant.