SAI assignment 5

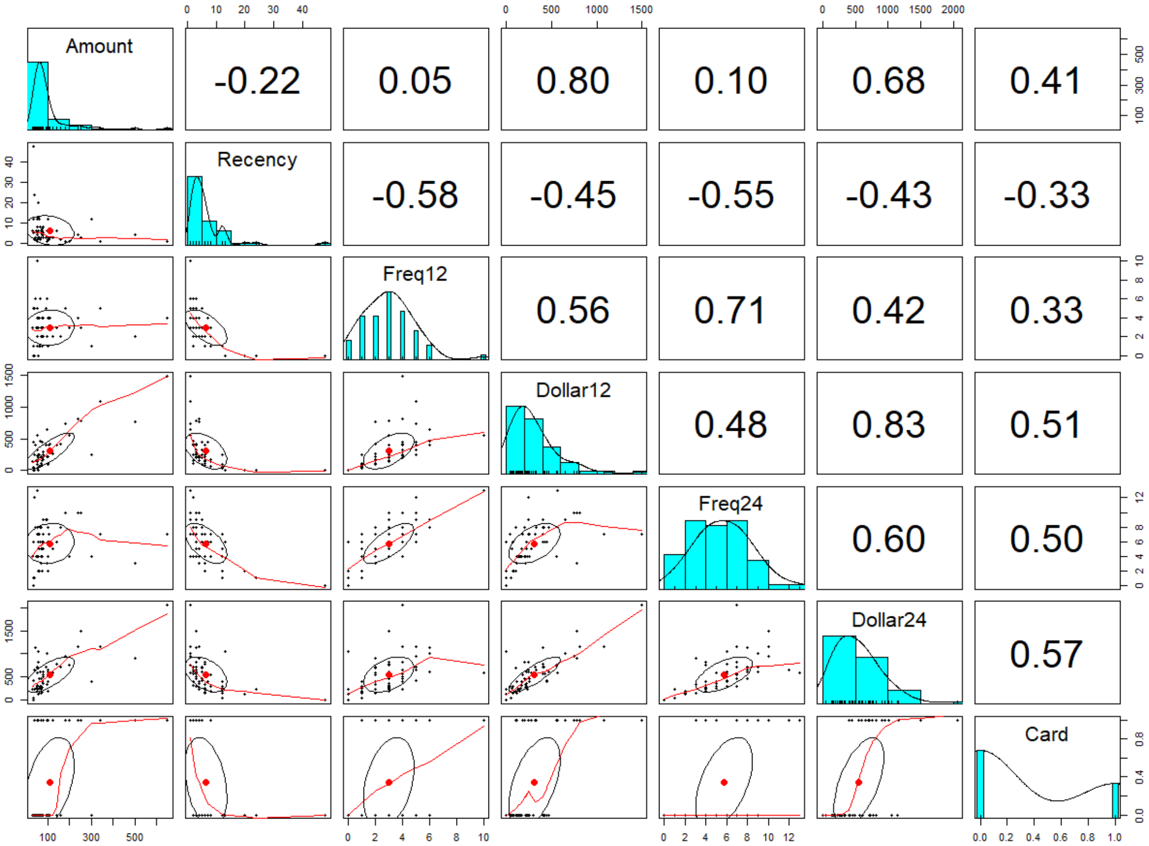
Caolan McDonagh – 21249929

Questions of interest

For this assignment we are investigating the effectiveness of regression models to make predictions. In this case, predicting the dollar amount spent by a customer on their most recent purchase, this is our target variable. Our explanatory variables consist of dollar amounts over the course of 12 and 24 months, the purchase frequency for 12 and 24 months and if the customer has a company credit card. We all of the above we can proceed with the below questions of interest.

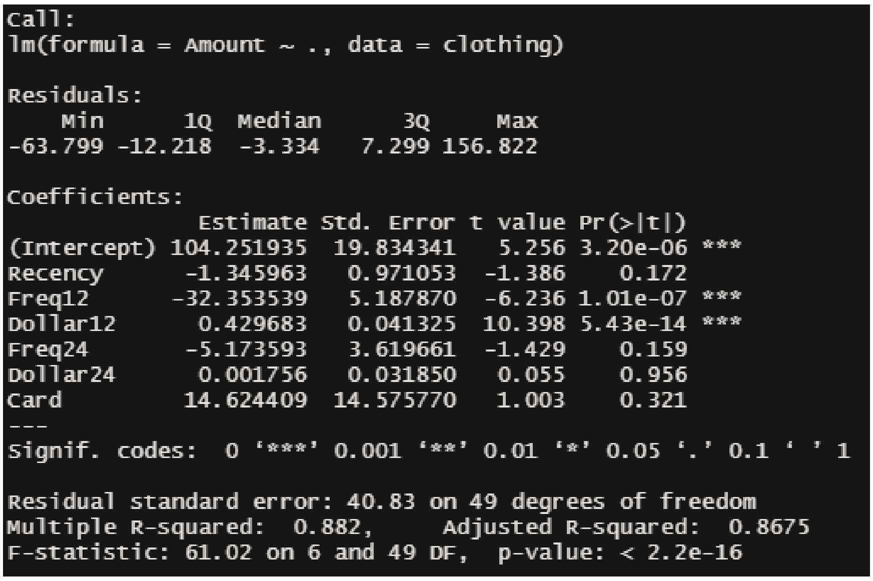
Question 1

The initial question of interest around this regression model is subset select.



The above is a pair.panels output of the clothing dataset as loaded. This gives us some good information revolving around the correlation between features via a histogram, Pearson correlation coefficient and bivariate scatter plot. Using Pearson's r, which ranges between -1 to 1 for a measure of linear correlation we can get a good idea of strong relationships to Amount, such as Dollar12.

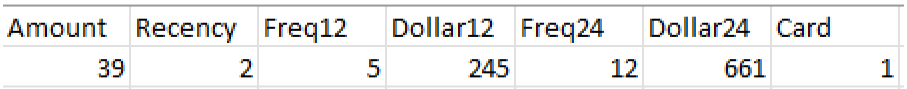
Before getting too bogged down in the numbers, it is generally best practice to look at the features logically first, and what we are looking for. Amount is the latest amount spent by a customer. To get a good idea of predicting their next purchase we'd like to look at existing expenditure data, this being the dollar and frequency amounts. For example, Card is an indicator if the customer has a private-label credit card with said retailer, but this doesn't really help with amount prediction, rather it just lets us know this customer is prone to shopping with this retailer. Recency shows when the last purchase was, but not helpful for an amount. Maybe we can use this in conjunction with another feature later? To break down the effectiveness of the features, we can produce a model using all features and interpret results.



```
Call:
lm(formula = Amount ~ ., data = clothing)

Residuals:
    Min      1Q  Median      3Q     Max 
-63.799 -12.218  -3.334   7.299 156.822 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 104.251935  19.834341   5.256 3.20e-06 ***
Recency      -1.345963   0.971053  -1.386    0.172    
Freq12      -32.353539   5.187870  -6.236 1.01e-07 ***
Dollar12      0.429683   0.041325  10.398 5.43e-14 ***
Freq24       -5.173593   3.619661  -1.429    0.159    
Dollar24      0.001756   0.031850   0.055    0.956    
Card         14.624409  14.575770   1.003    0.321    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40.83 on 49 degrees of freedom
Multiple R-squared:  0.882,     Adjusted R-squared:  0.8675 
F-statistic: 61.02 on 6 and 49 DF,  p-value: < 2.2e-16
```

Above is the summary of our maximum multiple linear regression model using all features present in the dataset. We first look to the F-statistic and associated p-value. We can see a break down by feature here, this reveals of our 6 features, only Freq12 and Dollar12 are statistically significant.

To properly understand coefficients, we need to understand what the model is trying to do, in the case of multiple linear regression it is trying to build the formula **y=mx+b** (b= intercept, m= slope), we build this via estimates for b and m by making use of ordinary least squares. The (intercept) row in the coefficients table above is where this line meets the y axis, using this we can begin to build the estimates. The baseline is $104, and dollars are added/subtracted based on the features to provide our Amount prediction, using the same formula. E.g [Using Freq12 estimate]:



Amount	Recency	Freq12	Dollar12	Freq24	Dollar24	Card
39	2	5	245	12	661	1

y=$-**32.35** (**2**) + $**104.2**

y=$**39.55**

55 cents off the actual amount isn't bad, but this was rather lucky all features considered. Using just this feature would work for some, but not most. Therefore, we look to *multiple* linear regression

The standard of error shows us how much uncertainty is present with our coefficient, this is rather high looking at this. Again, we can take note and use this when comparing our optimised subset.

We can also see the multiple R-squared and adjusted R-squared values, we can take note of the adjusted value for future comparisons after subset selection, the same for the F-statistic. This shows the relationship between predicator and response, when higher if helps us justify rejecting the null hypothesis, although this is over the whole model, rather than features.

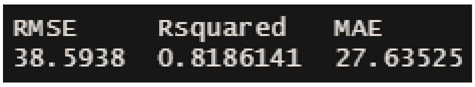
To build out best subset we can look at a few different factors, one of which is multicollinearity. This can cause overfitting in our model via variable changes being mirrored in one another, producing potentially unstable results and variance when the data or model are altered.

Using VIF (variance inflation factor) we can check for a numerical measure of multicollinearity:

Recency	Freq12	Dollar12	Freq24	Dollar24	Card
1.656238	3.082543	4.540948	3.256856	4.808116	1.599442

Generally, a VIF lower than 5 is considered acceptable, all the above variables are below 5 which is a good indicator at a lack of multicollinearity.

Checking for overfitting, I used repeated k-fold cross validation. 10 folds, 3 repeats:



```
RMSE      Rsquared   MAE
38.5938   0.8186141  27.63525
```

Comparing this back to the original model doesn't show much of a change when looking at the R-Squared values (Within 0.065). To further prevent overfitting and produce a more accurate model, we need to look to subset selection and removing features deemed worse when compared to other features. Initially I used stepwise regression in the forms of:

-**all possible**

-**step forward**

-**step both**

-**step backward**

Making use of the resulting **A**kaike **i**nformation **c**riterion and Sawa's **B**ayesian **i**nformation **c**riterion as measures of performance for the optimal subset size.

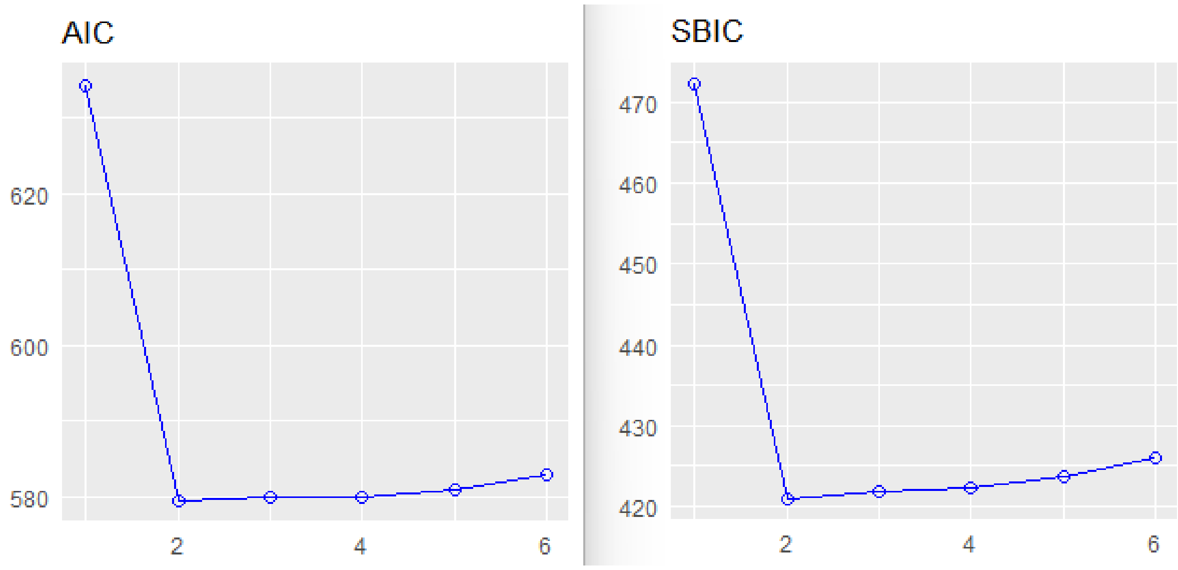
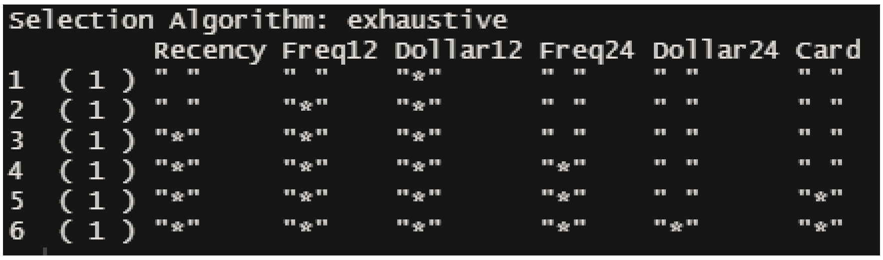


Figure 1 Best stepwise AIC and SBIC

All the mentioned stepwise methods produced very similar results, opting for 2 to 4 features being the optimal subset size. With this the process of elimination can begin. The regsubsets() method from the leaps library produces a best subset selection using residual sum of squares, as seen below:



```
Selection Algorithm: exhaustive
         Recency Freq12 Dollar12 Freq24 Dollar24 Card
1  ( 1 ) " "     " "    "*"      " "    " "      " " 
2  ( 1 ) " "     "*"    "*"      " "    " "      " " 
3  ( 1 ) "*"     "*"    "*"      " "    " "      " " 
4  ( 1 ) "*"     "*"    "*"      "*"    " "      " " 
5  ( 1 ) "*"     "*"    "*"      "*"    " "      "*" 
6  ( 1 ) "*"     "*"    "*"      "*"    "*"      "*" 
```

Here Recency, Freq12 and Dollar12 are shown present in the most common in a given model. Taking a step back and looking at this logically, this makes sense. Dollar12 and Freq12 show a good historical purchasing history for use in the prediction, where recency can be used as an indicator for frequency shoppers/more likely to spend.

Using these outright wouldn't be optimal, instead we can look to create more useful variables.

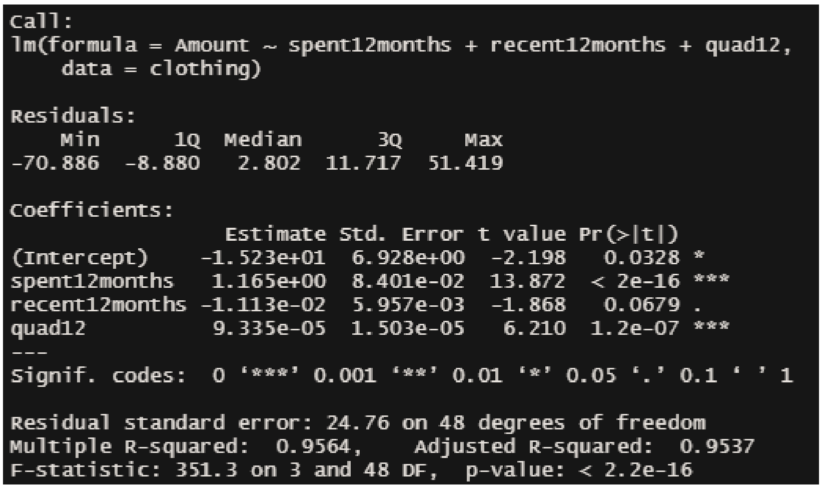
With dollar12 and freq12 we can create an explanatory variable (Dollar12 / Freq12) where dollar12 is in response to freq12. This will give us an average purchase amount in those 12 months, which is more relevant in predicting the most recent amount.

Dollar12 was our highest correlating feature and was present in the most models when using residual sum of squares. It has been shown to be a feature we should keep in our subset, but we can make better use of it by squaring it. By doing so we can more accurately see the effect Dollar12 has on Amount with this new quadratic variable. This will help with any nonlinear relationships.

Recency in conjunction with Dollar12 creates an interaction term based on the dollar12 spent and their last visit. This is useful to the regression as dollars spent by customers is related to when the customer last made a purchase. More recent purchases can indicate a more frequent shopper, especially when looking back at their dollar spent history.

After all of this we can create out final model:

finalModel <- lm(formula = Amount ~ spent12months + recent12months + quad12 , data = clothing)



```
Call:
lm(formula = Amount ~ spent12months + recent12months + quad12, 
    data = clothing)

Residuals:
    Min      1Q  Median      3Q     Max 
-70.886  -8.880   2.802  11.717  51.419 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -1.523e+01  6.928e+00  -2.198   0.0328 *  
spent12months   1.165e+00  8.401e-02  13.872  < 2e-16 ***
recent12months -1.113e-02  5.957e-03  -1.868   0.0679 .  
quad12          9.335e-05  1.503e-05   6.210  1.2e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.76 on 48 degrees of freedom
Multiple R-squared:  0.9564,	Adjusted R-squared:  0.9537 
F-statistic: 351.3 on 3 and 48 DF,  p-value: < 2.2e-16
```

Looking at the summary of the regression model, we can already see a substantial improvement in the Adjusted R-Squared score (0.9537 vs 0.8675) implying this subset is a better fit than using all features. The residual standard of error has almost halved on top of this, implying better predictions.

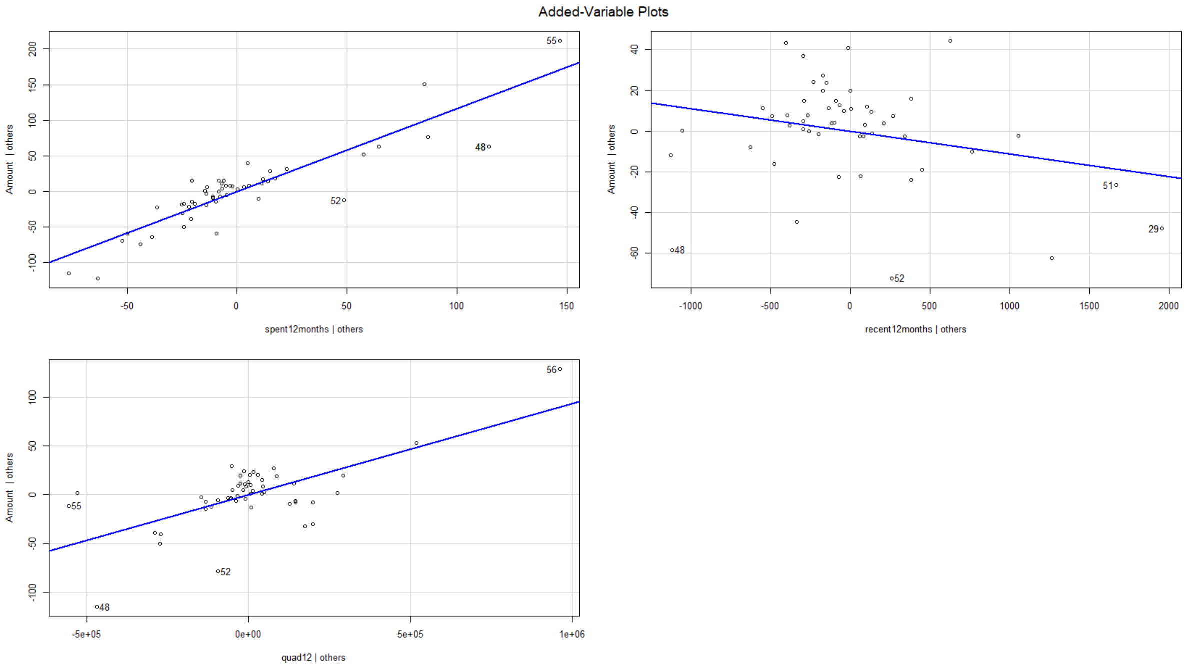


Figure 2 Added Variable plot

The above added variable plot shows a nice correlation between the features and Amount.

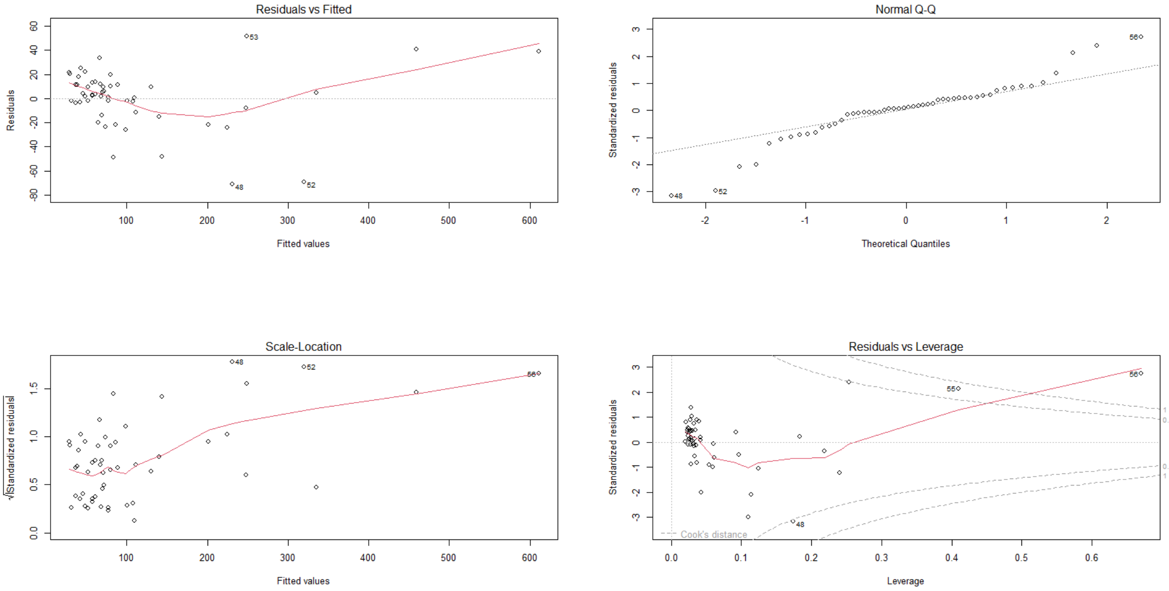


Figure 3 Regression Diagnostics.

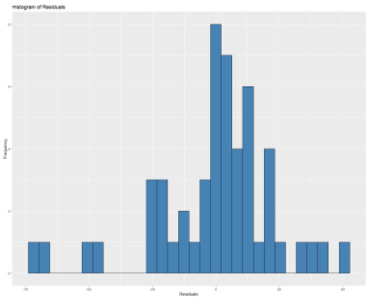
The above graphs come from our final regression model. The first Residuals vs Fitted plot looks okay, points are adhering to the identity line without signs of heteroscedasticity. No explicit pattern that may indicate anything wrong.

The normal Q-Q plot looks okay besides heavy tails, which would indicate some extreme values. When comparing this to the Q-Q plot of the original regression of all features, it is much smaller in the tails.

A coherent spread is kept in the Scale-Location plot, again in favour of no heteroscedasticity.

Finally, the residual vs leverage plot. There are a handful of influential points present here, one of which is rather far from the Cook's distance line, removing this may result in a batter coefficient and R squared value. This is value [56], which when looking at the data is the highest spender in the last 12- and 24-months period. This could be considered an outlier, but I won't remove it in this case as it is not astronomical when compared to other high spenders.

We can check the residuals of this regression model to insure we are normally distributed. We can see the rough bellshape present, confirming to an assumption of normality.



To note, when running a model with all features, including the newly created. The P values of quad12 and spent12months are the most statistically significant of all features by a large margin. Much closer to 0 than any other original features.

Question 2

Lasso (Least absolute shrinkage and selection operator) regression is highly regarded regularization method used in statistics, for avoiding overfitting. It is used to amplify existing regression methods to produce more accurate predictions via shrinkage. This is where we shrink our data values towards a centre (mean). Lasso in particular leans towards fewer features in the model and works to eliminate multicollinearity. In doing so, it essentially automates the subset selection procedure through eliminating the features caught in the lasso tightening. We can use this to verify the steps taken in Q1 by checking what features are eliminated (reduced to 0).

BootLasso simulation 95% confidence interval of all features, including newly created:

Upr/Lwr	Recency	Freq12	Dollar12	Freq24	Dollar24	Card	Quad12	Spent12Months	Recent12Months
Lower	-2.737	0	-0.166	-4.610	-0.025	-8.907	1.060	-0.009	0.0001
Upper	1.362	8.523	-0.112	0.930	0.245	15.131	1.372	0.008	0.0002

Multiple Linear Regression 95% confidence interval of all features:

Upr/Lwr	Recency	Freq12	Dollar12	Freq24	Dollar24	Card	Quad12	Spent12Months	Recent12Months
Lower	-4.771	-9.672	-3.967	-5.535	-4.940	-1.014	1.646	1.089	-1.302
Upper	2.147	15.991	-0.112	2.092	0.029	20.897	0.0003	1.555	0.015

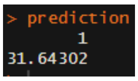
BootLasso estimated coefficients:

Recency	Freq12	Dollar12	Freq24	Dollar24	Card	Quad12	Spent12Months	Recent12Months
0	0	-0.082	-2.305	-0.00009	3.304	1.118	-0.0029	0.00015

From the above we can see the lasso pulling into its diamond shape. Recency, Freq12 were caught and shrunk to 0, and Dollar24 takes a large reduction. This is adhering to the subset selection attempts made in question 1, leaving me confident in my subset selection efforts for use in prediction in the next question.

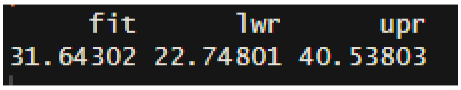
Question 3

Using the predict() method and our “finalModel” for the first customer, we end up with a prediction of **$31.64**.



```
> prediction
       1
31.64302
```

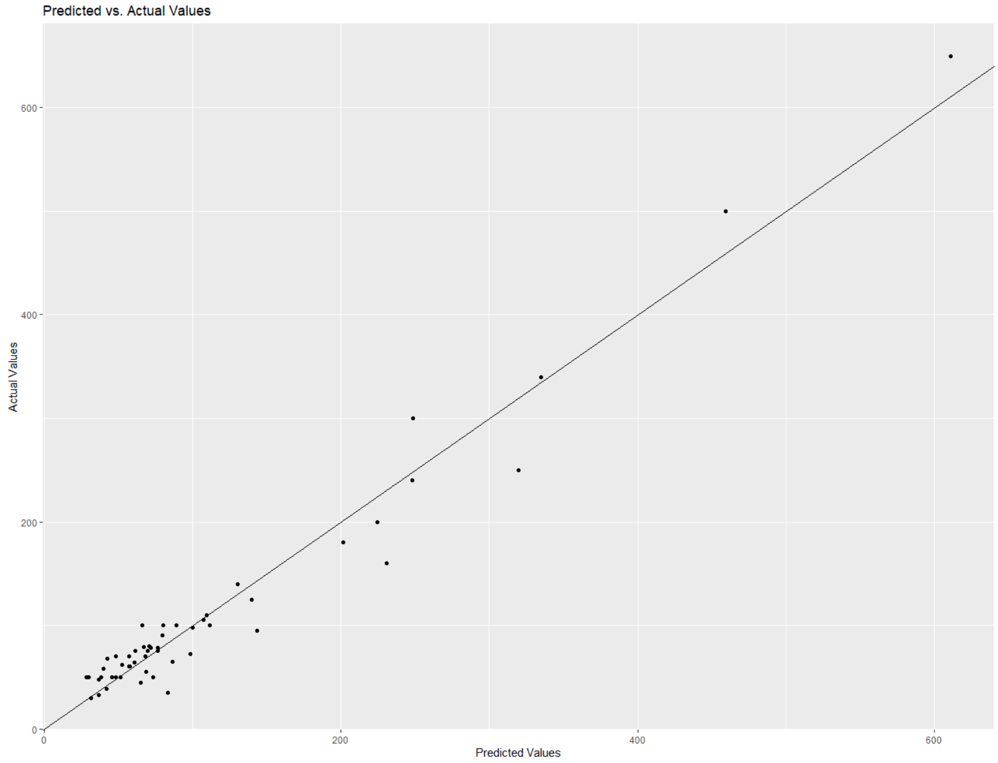
This is an overprediction of **$1.64** (~5%). Not bad! A pretty accurate prediction all things accounted for. Below is a interval estimation based on our final model:



```
     fit      lwr      upr
31.64302 22.74801 40.53803
```

A lower of $22.74 and upper of $40.53. This is, in my opinion, a “healthy” interval estimation for what our model is attempting to achieve. The upper and lower are not unreasonably inaccurate, and provide a good estimation given the use case of this model.

Below is some extra information on all predicted values vs the actual values contained in the dataset. Again, good linearity in my oppinion, indicating a reasonbly accurate regression model for predicting the amount spent by a customer on a transaction.


Predicted vs. Actual Values
Actual Values
Predicted Values
0
200
400
600
0
200
400
600