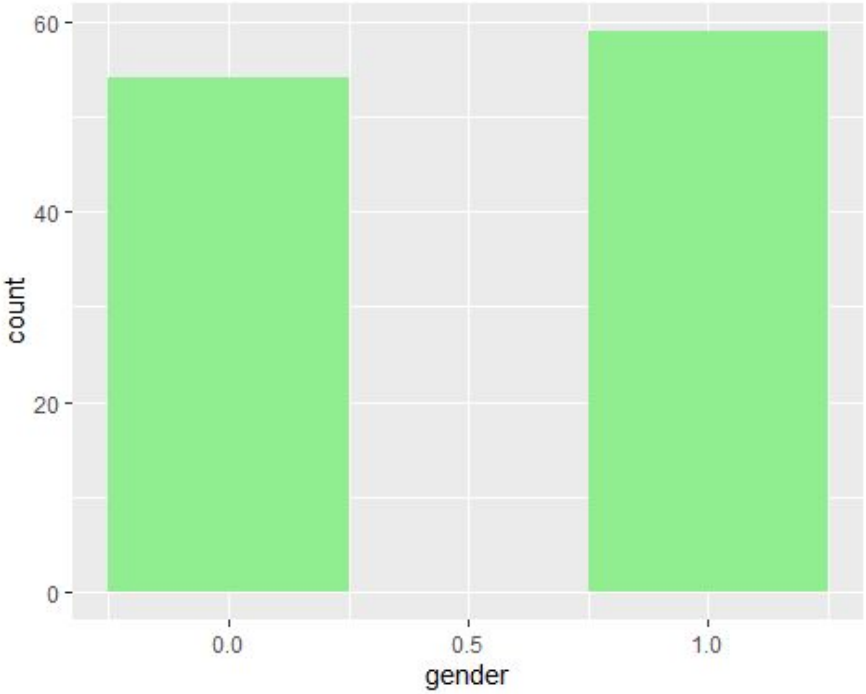
STAT 410
Starbucks Customer Survey

Michael Brandt, Thys Jacobi, Eunjin Park, Abigail Pearce

Data Introduction

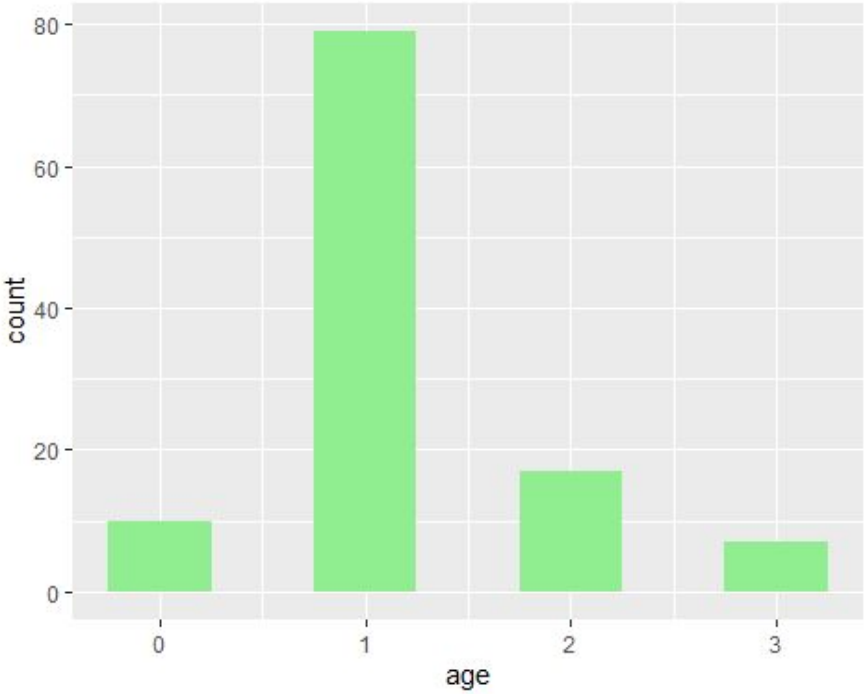
- “Starbucks Customer Survey” from Kaggle.com
    - The data was collected by randomly giving to customers after their visit and was asked multiple questions regarding themselves and about the service at the store.

Gender



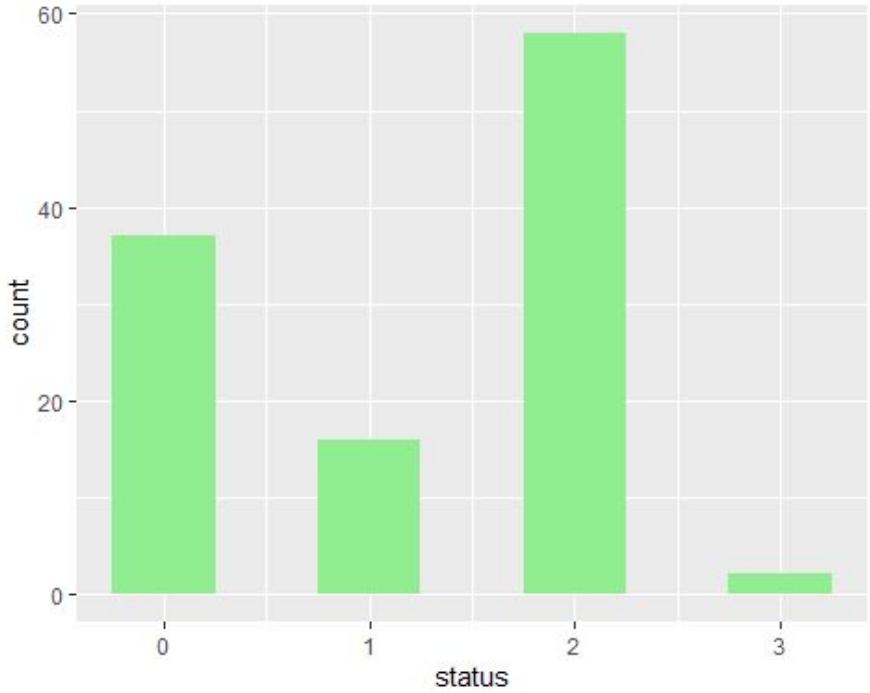
Factor	Gender	Count
0	Male	54
1	Female	59

Age



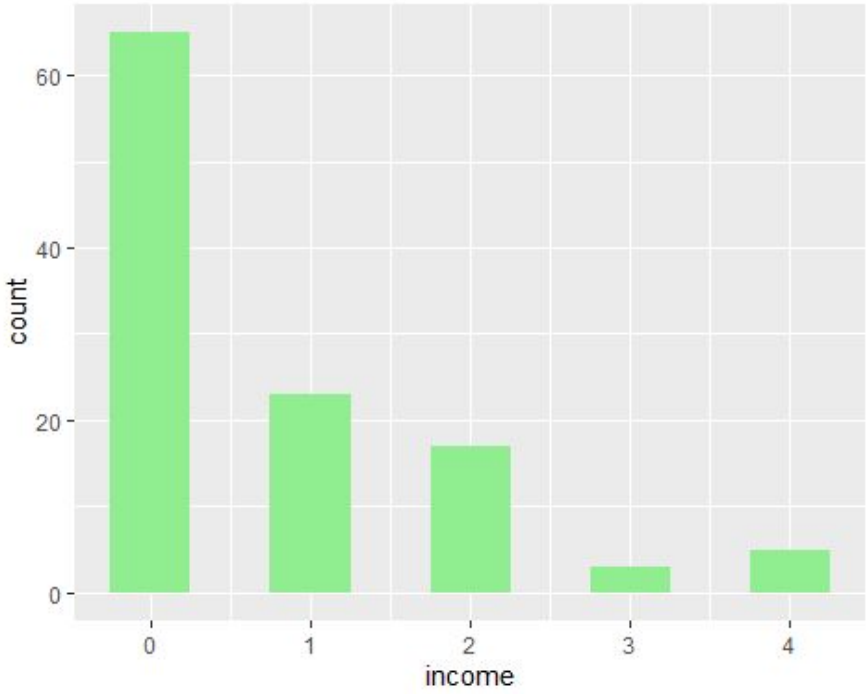
Factor	Age	Count
0	>20	10
1	20-29	79
2	30-39	17
3	<40	7

Status



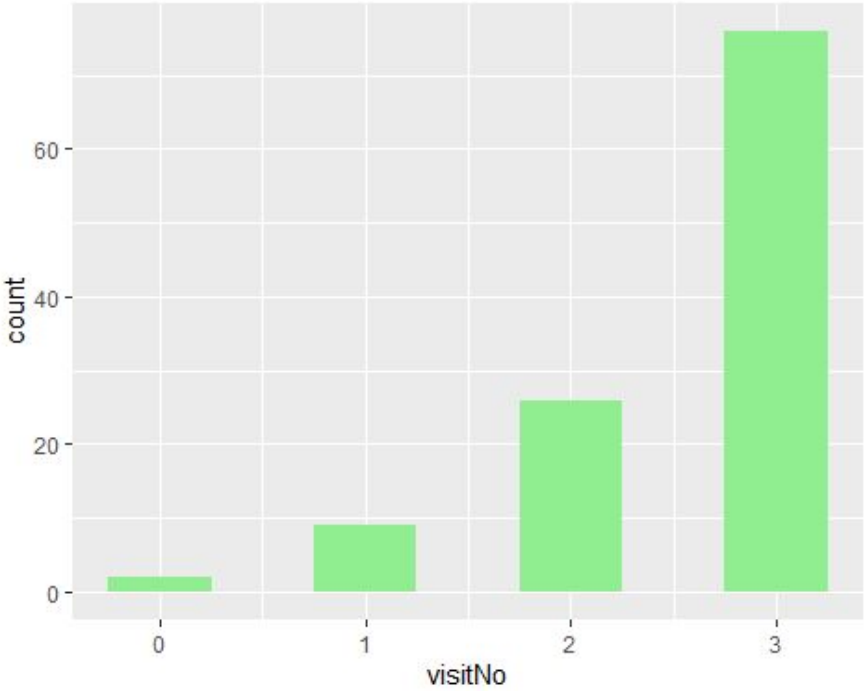
Factor	Status	Count
0	Student	37
1	Self-Employed	16
2	Employed	58
3	Housewife	2

Income



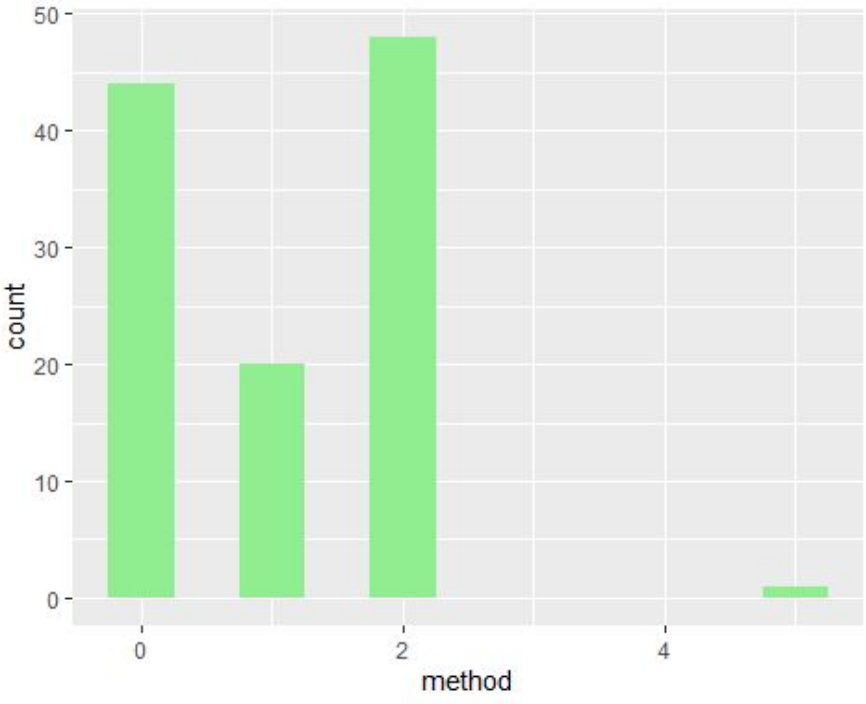
Factor	Income	Count
0	>$25,000	65
1	$25,000-$50,000	23
2	$50,001-$100,000	17
3	$100,001-$150,000	3
4	<$150,001	5

Visit Frequency



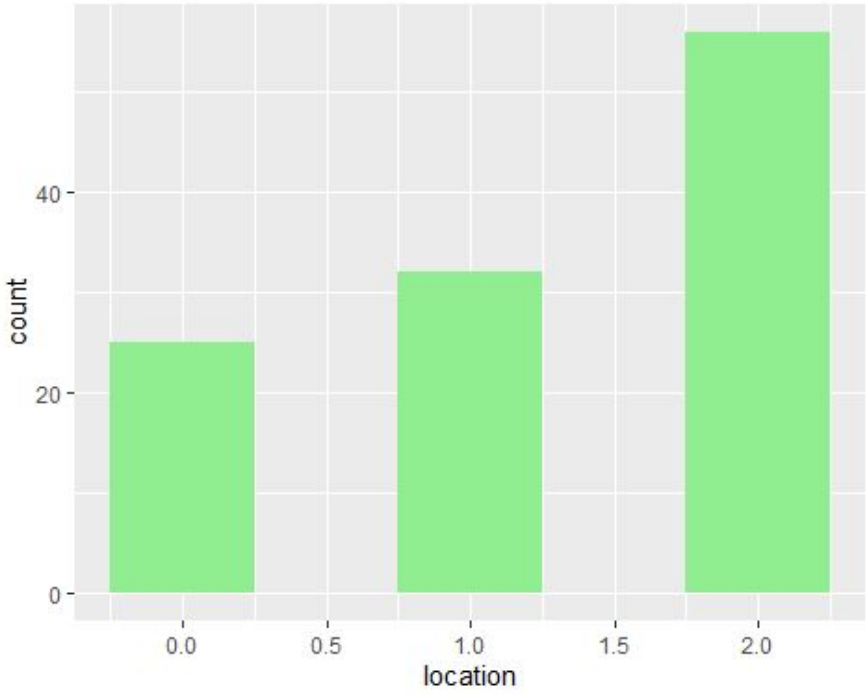
Factor	VisitNo	Count
0	Never	2
1	Monthly	9
2	Weekly	26
3	Daily	76

Method



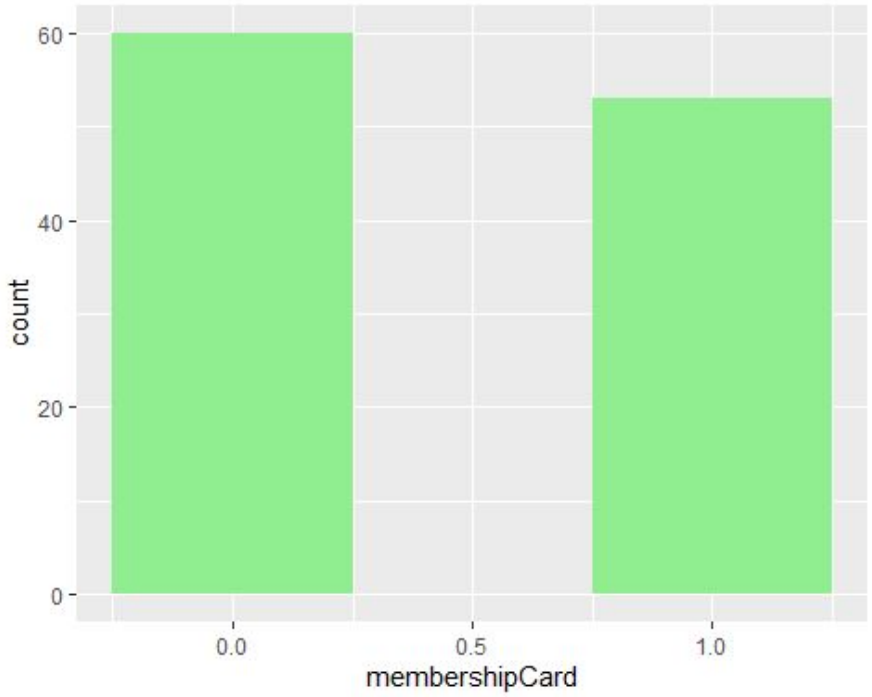
Factor	Method	Count
0	Dine-In	44
1	Drive-Thru	20
2	Take Away	48
3	Never	0
5	Others	1

Location



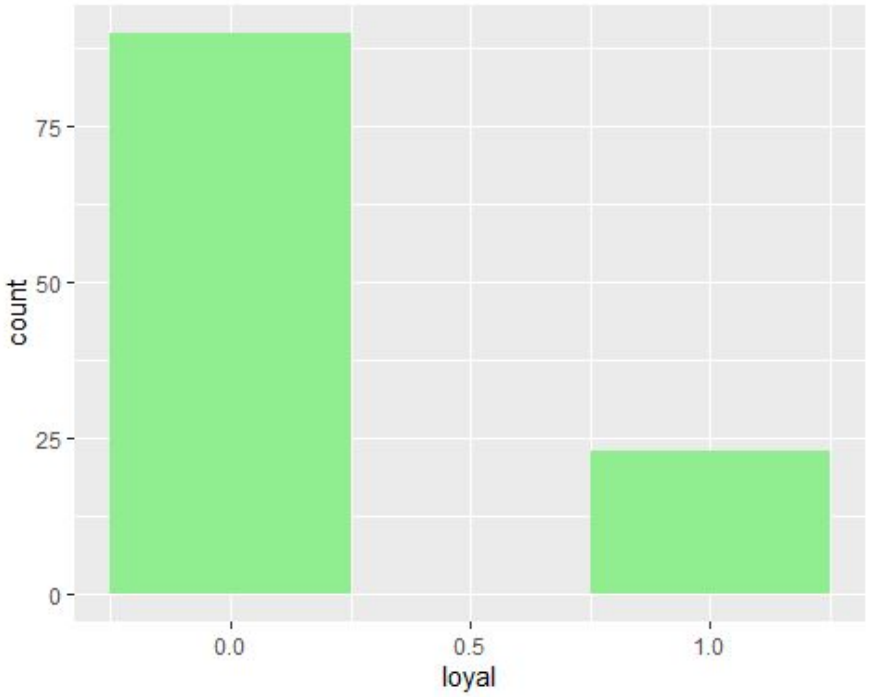
Factor	Location	Count
0	Within 1km	25
1	1km - 3km	32
2	Over 3km	56

Membership Card



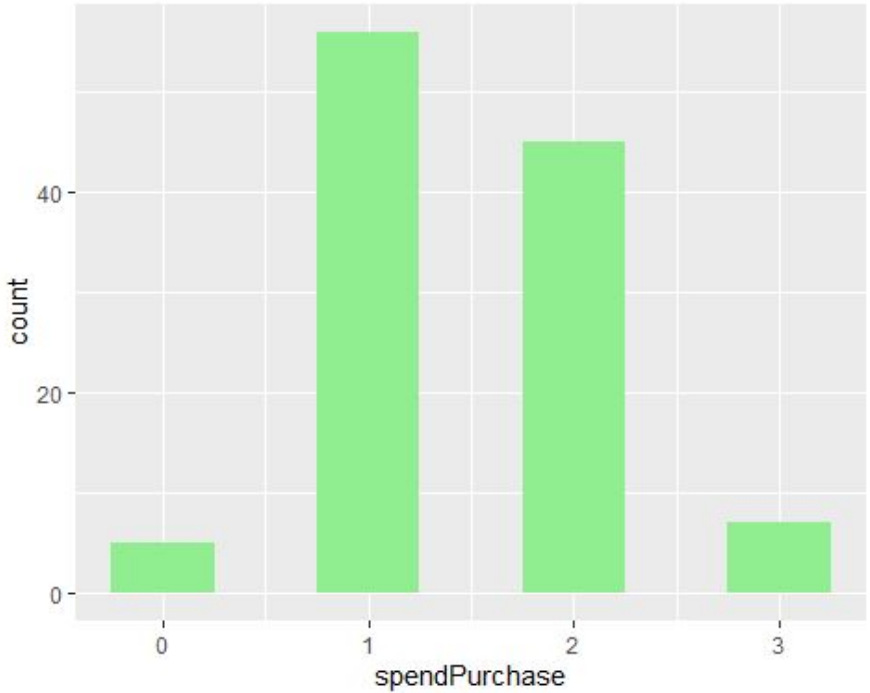
Factor	Membership_Card	Count
0	Yes	60
1	No	53

Loyal



Factor	Loyal	Count
0	Yes	90
1	No	23

Money Spent



Factor	Money_Spent	Count
0	0	5
1	>$20	56
2	$21-$40	45
3	<$41	7

Research Goals

Questions:

1) Which variables out of age, gender, status, and income have the most significant effect on visiting Starbucks at least weekly? What is the relationship that each of these variables has with the number of visits? Check assumptions of the model. Are any violated?
2) Based on the given variables, which ones are the strongest predictors in predicting whether a customer is loyal to Starbucks?
3) What variables have the greatest impact on purchase price?

Methods Used

Question 1:

- Created a binary response variable “highvisit” (going to Starbucks weekly or daily)
- GLM with predictors age, gender, status, and income, made each predictor into a factor
- Summary() function to see coefficients and p-values
- Challenges:
    - Necessary to create a binary response variable since data is categorical variables represented as numeric values
    - Since coefficients/p-values are relative to reference category, effect seen is factors compared to reference values (not just effect of the predictor alone)

Findings

- Nearly all prediction factors (except for “income1,” “income2,” and “income4”) of visiting Starbucks frequently had positive coefficients when the GLM was fit, meaning that they increase the probability of one going to Starbucks at least weekly
- Factors with noticeably larger coefficients: “status3” (housewife) and “income3” (income of 100,000 - 150,000 dollars per year), these traits may have a stronger influence on probability of going to Starbucks often
- At 0.05 level of significance, “income2” (income of 50,000 - 100,000 dollars per year) showed greatest evidence of effect being non-zero

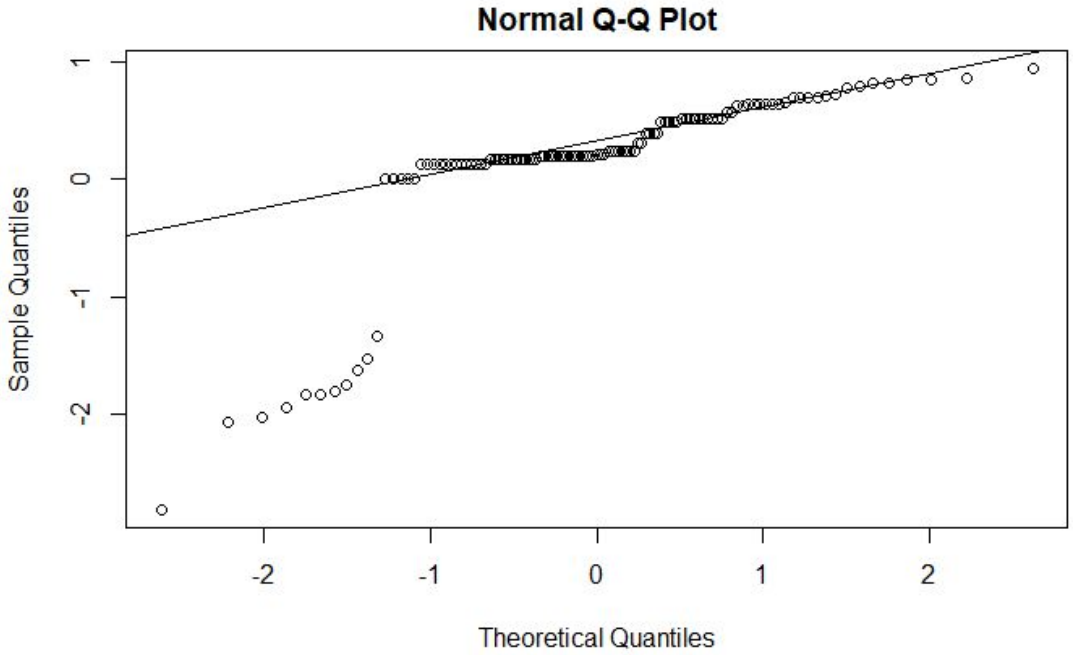
Methods Used

Question 1 cont.:

- Through qqnorm(), confirmed the hypothesis test.

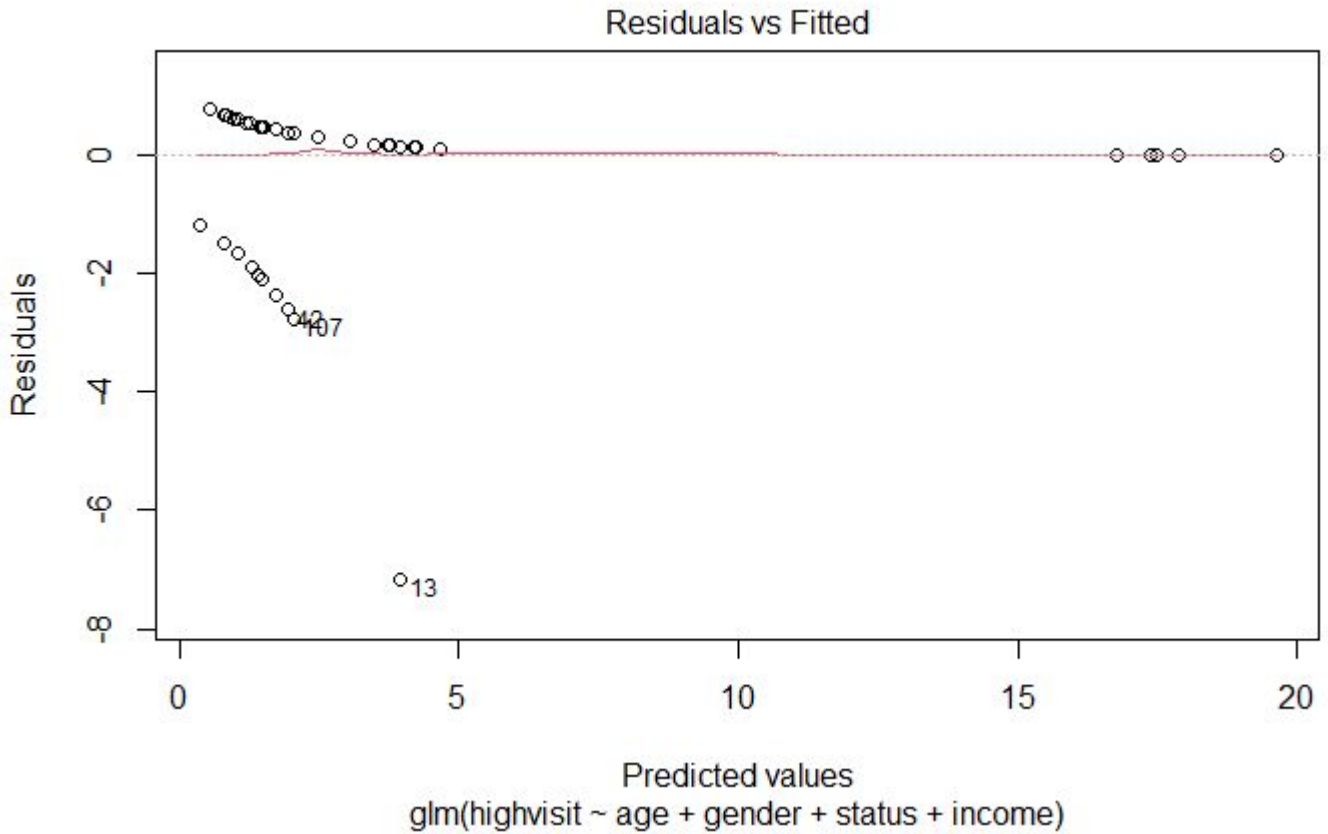
  -Relation between residuals and predicted values(glm model), QQplot between std residuals and theoretical quantiles, relation between residuals and leverage and qqplot between sample quantiles and theoretical quantiles

Normality of residuals

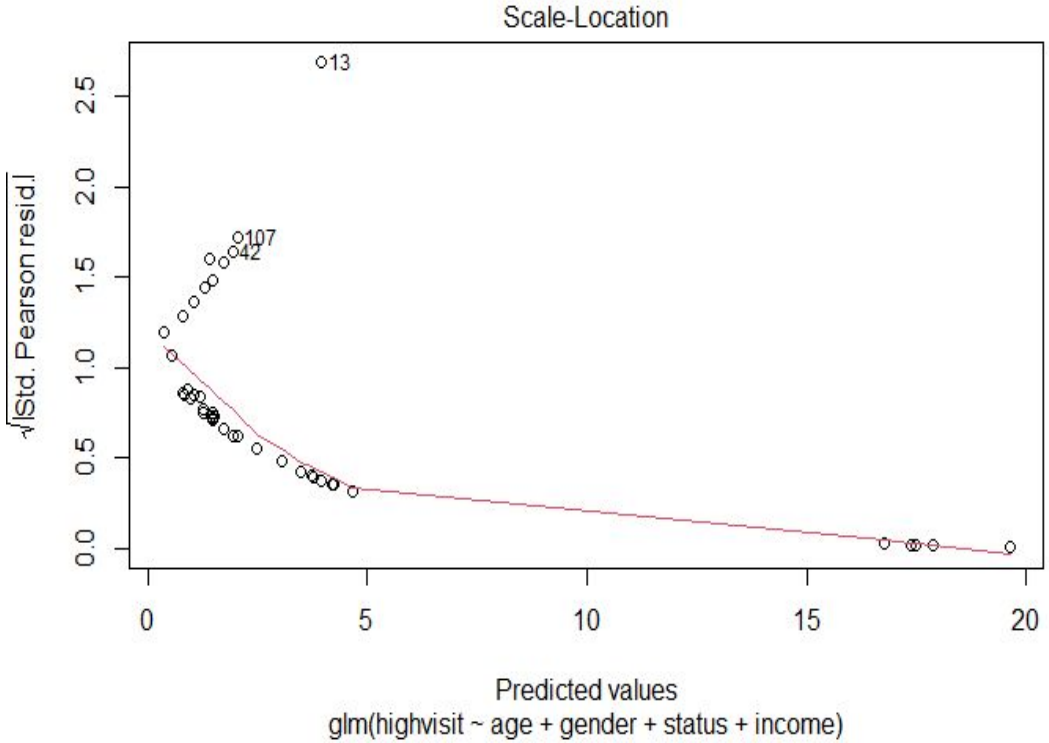


-The residuals follow the straight dashed line, then the assumption is fulfilled.

Linearity

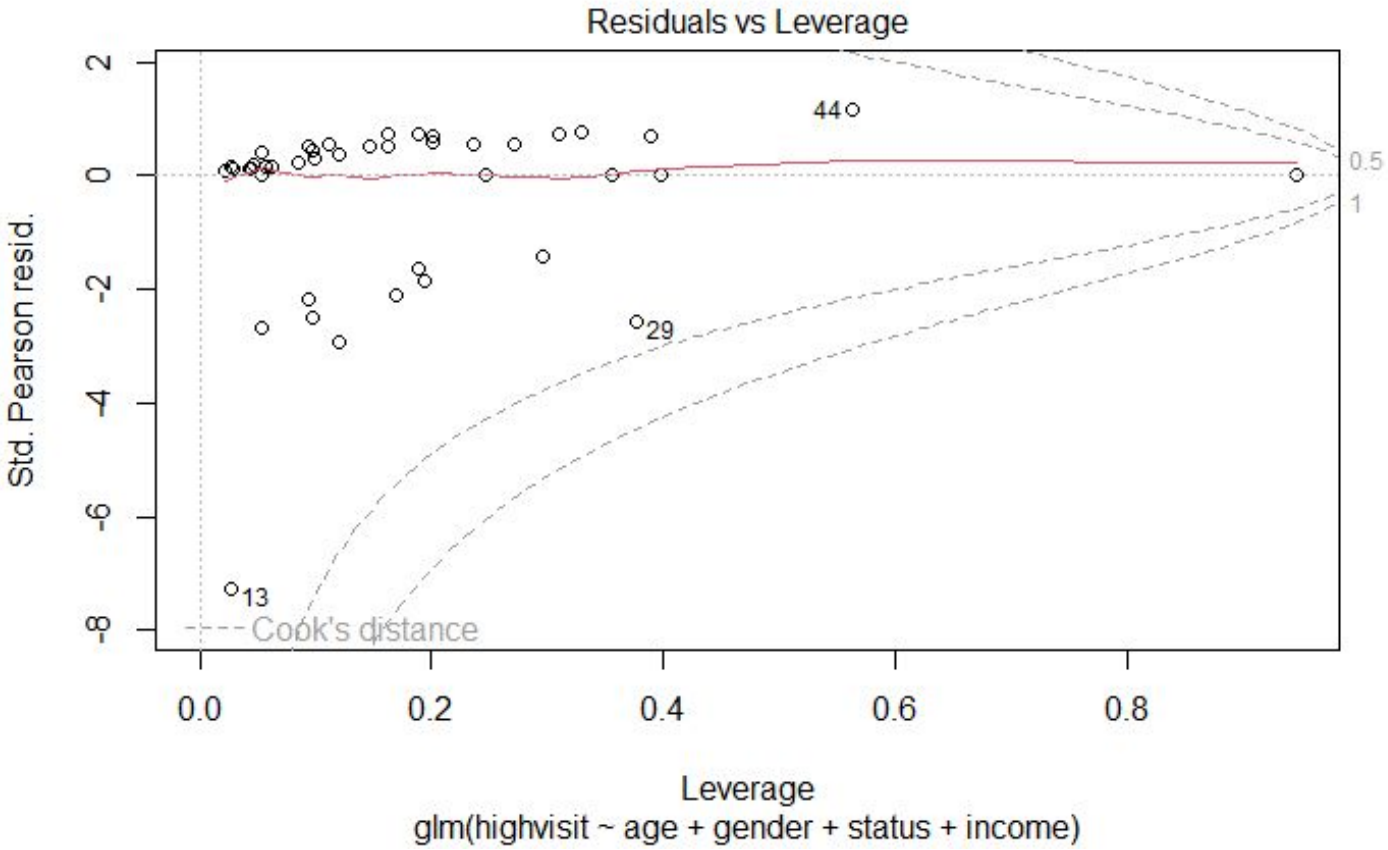

Residuals vs Fitted
Residuals
0
-2
-4
-6
-8
0
5
10
15
20
42
107
13
Predicted values
glm(highvisit ~ age + gender + status + income)

About the homoscedasticity of residuals



-The scale location plot suggests some non-linearity here,
but what we can also see is that the spread of magnitudes seems to be lowest in the fitted values close to 0, highest in the fitted values around 1.8, and medium around 0.9. This suggests heteroskedasticity.

Influenced values



Findings

- By qqplot the majority of the residuals show the linearity normality of residuals assumption is satisfied.
- By residuals and fitted diagram residuals are not spread equally, so it would be hard to say for having a linear relationship.
- By scale- location diagram, residuals are spread equally from horizontal line so we could confirm homoscedasticity of residuals.

Methods Used

Question 2:

- Created a logistic regression model to see which factors are the largest predictors in figuring out whether a customer is loyal to Starbucks.
- Used prediction models of all the predators to figure out which factors of each predictors are the most impactful.
- Using these findings create a model to show what is the most loyal customer to Starbucks.

Model

```
loyalty_fit <- glm(loyal ~ gender + age + status +
income + visitNo + method + location + membershipCard,
data = starbucks, family = binomial(link = "logit"))
```

Predict

Using the most common variable in each predictor(Female/20 to 29/Employed/Less than RM25,000/Take away/more than 3km/owning a membership card) we will go through the different variables and create prediction models and figure out which factors of each variable have the biggest impact.

Factors

```
## [1] "Gender"
## [1] 1 0
##           1           2
## 0.08438504 0.07637594
```

Male is more likely to be loyal.

```
## [1] "Age"
## [1] 1 2 3 0
##           1           2           3           4
## 0.08438504 0.06117710 0.04404488 0.11531564
```

Age 40+ is more likely to be loyal.

```
## [1] "Status"
## [1] 0 2 1 3
##           1           2           3           4
## 0.12776222 0.08438504 0.10409332 0.06812444
```

Housewife is more likely to be loyal.

```
## [1] "Income"
## [1] 0 2 1 3 4
##           1           2           3           4
5
## 0.08438504 0.17449477 0.12247986 0.24249513
0.32651369
```

Income >$25,000 is more likely to be loyal.

```
## [1] "Method"
## [1] 0 2 1 5
##           1           2           3           4
## 0.11783517 0.08438504 0.09987188 0.05016949
```

Other is more likely to be loyal.

```
## [1] "Location"
## [1] 0 1 2
##           1           2           3
## 0.06681392 0.07512887 0.08438504
```

Within 1km is more likely to be loyal.

```
## [1] "Visit"
## [1] 3 2 1 0
##             1             2             3
4
## 0.0843850363 0.0115969061 0.0014914727
0.0001901238
```

Daily visitors are more likely to be loyal

```
## [1] "Membership_Card"
## [1] 0 1
##           1           2
## 0.08438504 0.29732453
```

Having a Membership card is more likely to be loyal

Best Factors

Using the factors that affect loyalty we see that a man who is 40+ and a housewife, makes $25,000, gets Starbucks everyday, lives within 1km of Starbucks, gets his Starbucks by other methods, while having a Starbucks membership is the most likely to be loyal to Starbucks.

```
1 - BestPredict
```

```
##         1
## 0.999962
```

Logical Factors

Since the previous person does not exist we will switch the gender to female to match with being a housewife, and then change the method to the next lowest since there is only 1 data point that corresponds to method of “other” so it will become “take away” and we see they still almost have a 100% chance of being loyal

```
1 - LogicalPredict
```

```
##         1
## 0.9999262
```

Question 2 Conclusions

Strongest Impactors:

- We see that having a membership card is a strongest predictor when it comes to showing loyalty, adding about 20% to the chances of being loyal.
- On the personal level the strongest predictor of loyalty is the person's income and it shows that the more money people make the less likely they are to be loyal.

Surprises:

- I expected to see a larger impact that method had, they were evenly distributed and I expected Drive-Thru to be the strongest predictor of loyalty since the ease of access is what Starbucks is known for.
- Same thing with visitNo, I would have expected a larger gap between all of the factors since I associate loyalty with going daily or weekly, but there was still loyalty with the people who go monthly.

Summary

```
summary(loyalty_fit)
##
## Call:
## glm(formula = loyal ~ gender + age + status + income + visitNo +
##     method + location + membershipCard, family = binomial(link = "logit"),
##     data = starbucks)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.26891  -0.64957  -0.35941  -0.06946   2.67735
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -7.7475     3.1881  -2.430   0.0151 *
## gender           0.1084     0.6064   0.179   0.8581
## age             -0.3466     0.4811  -0.720   0.4712
## status          -0.2317     0.3362  -0.689   0.4908
## income           0.4151     0.3334   1.245   0.2131
## visitNo          2.0611     1.0404   1.981   0.0476 *
## method          -0.1856     0.2914  -0.637   0.5242
## location         0.1262     0.3796   0.333   0.7395
## membershipCard   1.5241     0.6443   2.365   0.0180 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 114.191  on 112  degrees of freedom
## Residual deviance:  89.632  on 104  degrees of freedom
## AIC: 107.63
##
## Number of Fisher Scoring iterations: 7
```

Looking at the summary we can see that the biggest impacts are visitNo, membershipCard. If we want to look at the person themselves we see income, age, and status have the biggest impact.

Q: What variables have an impact on purchase price?

Method

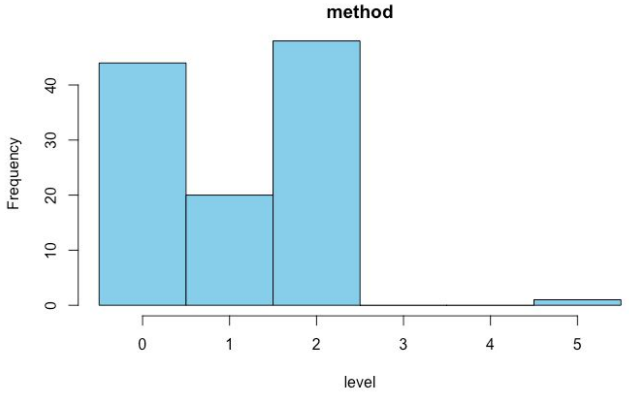
1. Convert all variables to factors due to the categorical nature of each variable.
2. Create a new binary response variable highPurchase from spendPurchase in order to perform logistic regression.
3. Use stepwise regression in both directions to find which variables are most impactful for determining is a purchase is high
4. Use this model to find what variable levels have the best chance at causing a high purchase.

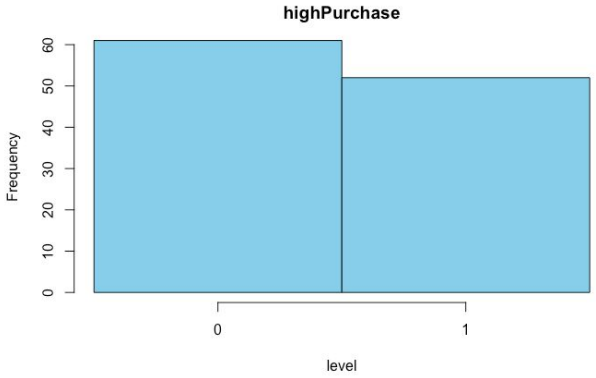
Model after stepwise regression

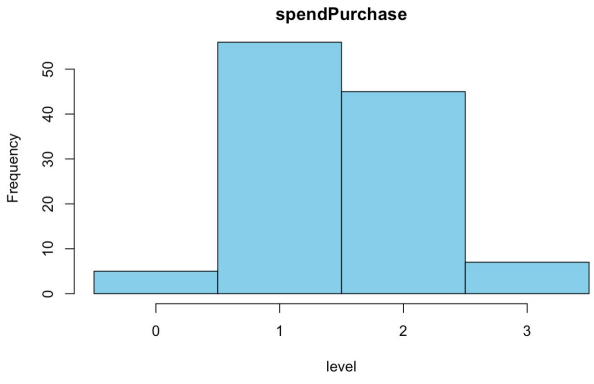
Logistic regression model with logit link.

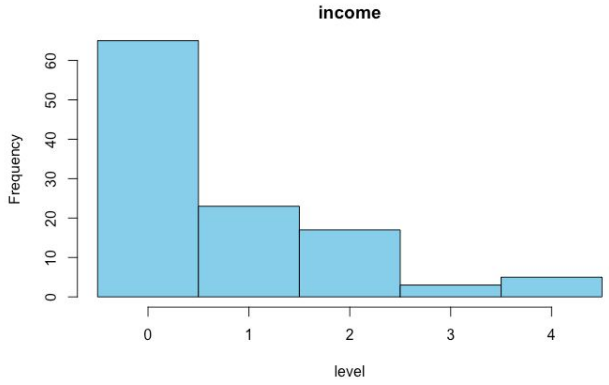
Formula: highPurchase ~ income + method

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.0398     0.3961  -2.625  0.00866 **
income1        1.1716     0.5137   2.281  0.02257 *
income2        0.9044     0.5871   1.540  0.12345
income3       17.5511  1384.1109   0.013  0.98988
income4        1.7339     1.2122   1.430  0.15260
method1        1.3839     0.6196   2.234  0.02551 *
method2        0.1680     0.4599   0.365  0.71493
method5       17.6059  2399.5448   0.007  0.99415
```

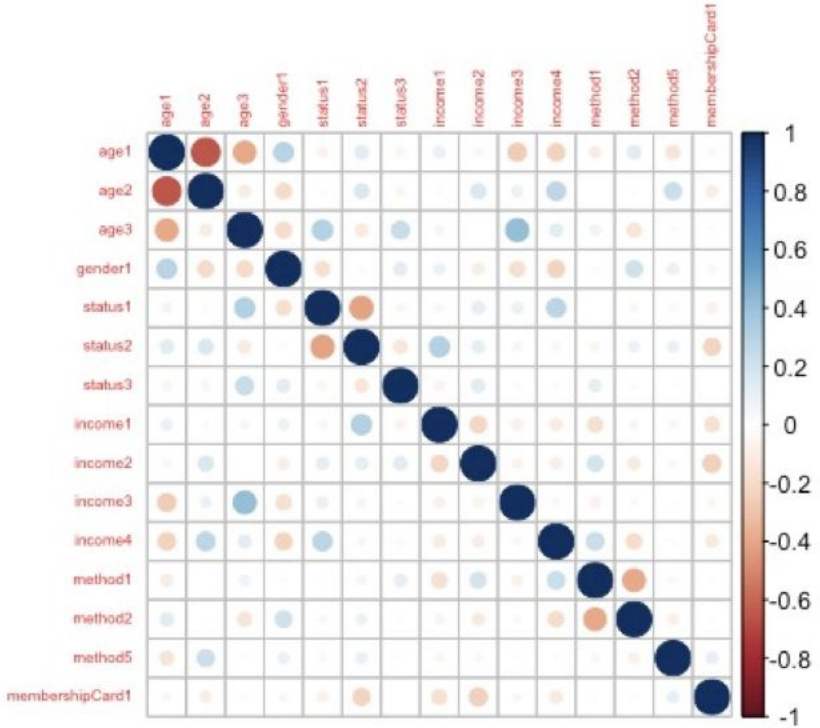


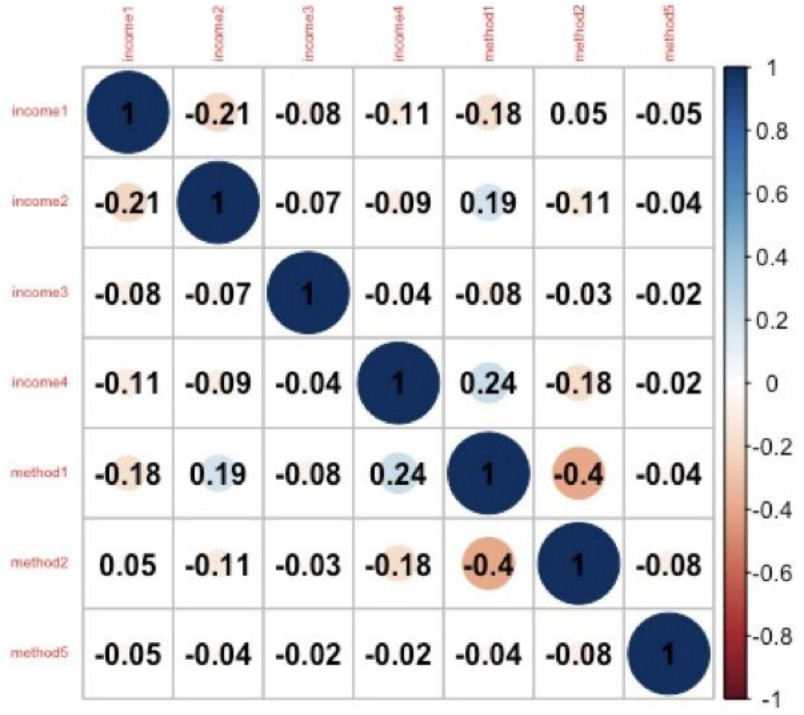




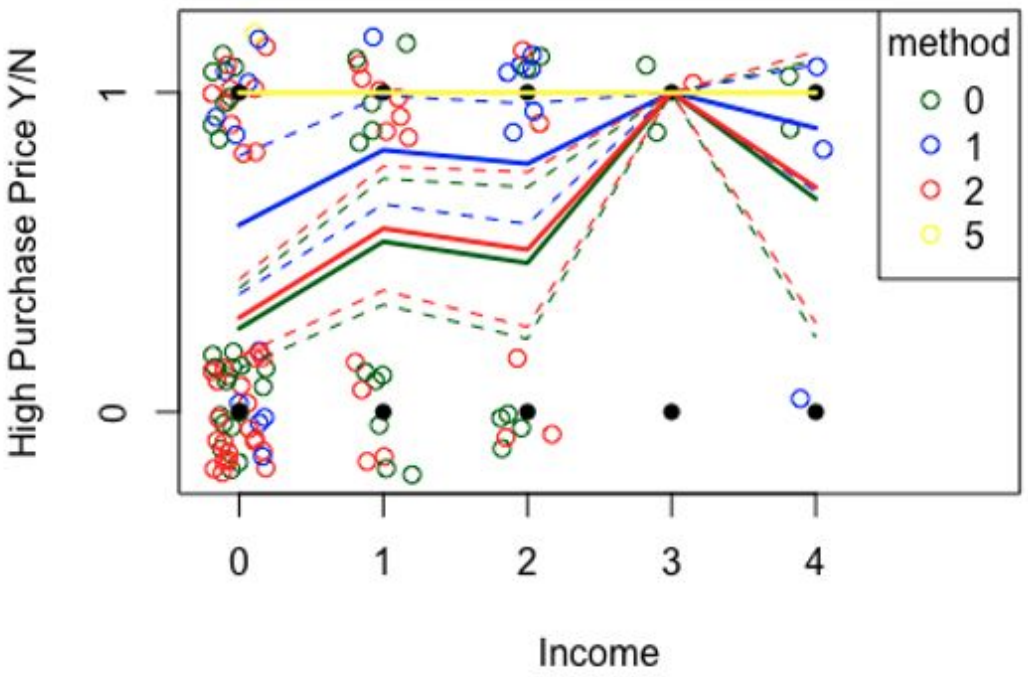


Correlation plots:





Model Plots



Notes

There is only 1 observation with method 5. Perhaps the method was misreported?

All observations where income=3 have high purchase price. This is what is causing the models to all converge at (3, 1)

Methods

0: Dine in

1: Drive-Thru

2: Take Away

5: Never Buy

Question 3 Conclusions

- We can see from the plot that higher income levels lead to a high purchase more often than lower income levels, which was expected.
- Also, people who order in a drive-thru are much more likely to spend a lot of money and people who order take out are slightly more likely to spend a lot of money compared to people who order dine in.

Surprises

- I was surprised to find that status was not a relevant factor in determining if a purchase is high or not. Same with having a membership card
- I thought that dine ins would have the highest chance of spending more money, but that was the opposite.

Challenges

- Converting between numeric type and factor type without losing factor values