

always_allow_html: yes

title: “Final Assignment” output: word_document: default html_notebook: default pdf_document: default
html_document: default — # how does marital status effect hotels choices Importing data from original
database consumer class variable is created from the variables srchadultscount and srchchildrencount. Consumers
with srchadultcount of 1 and 2 would be transformed to “single person” and “couples” respectively. Consumers
with a srchchildrencount greater than 1 would be called “parents”. All other consumers are called “others”

loading and saving

```
load(file = "../Data/mydata.RData")  
# save(mydata, file = "../Data/mydata.RData")  
mydata <- data  
remove(data)
```

trim database

```
# mydata <- subset(mydata, select = -c(prop_brand_bool) )  
# mydata <- subset(mydata, select = -c(position) )  
# mydata <- subset(mydata, select = -c(srch_saturday_night_bool) )  
# mydata <- subset(mydata, select = -c(random_bool) )  
# mydata <- subset(mydata, select = -c(price_usd_normalized) )  
# mydata <- subset(mydata, select = -c(Pclass) )  
# mydata <- subset(mydata, select = -c(score) )
```

subsetting

```
mydata <- mydata[which(mydata$click_bool == 1),]  
mydata <- mydata[which(mydata$prop_review_score != 0),]  
mydata <- mydata[which(mydata$srch_id < 4000),]  
# mydata <- head(mydata, 1000)  
length((mydata$srch_id))
```

```
## [1] 2611
```

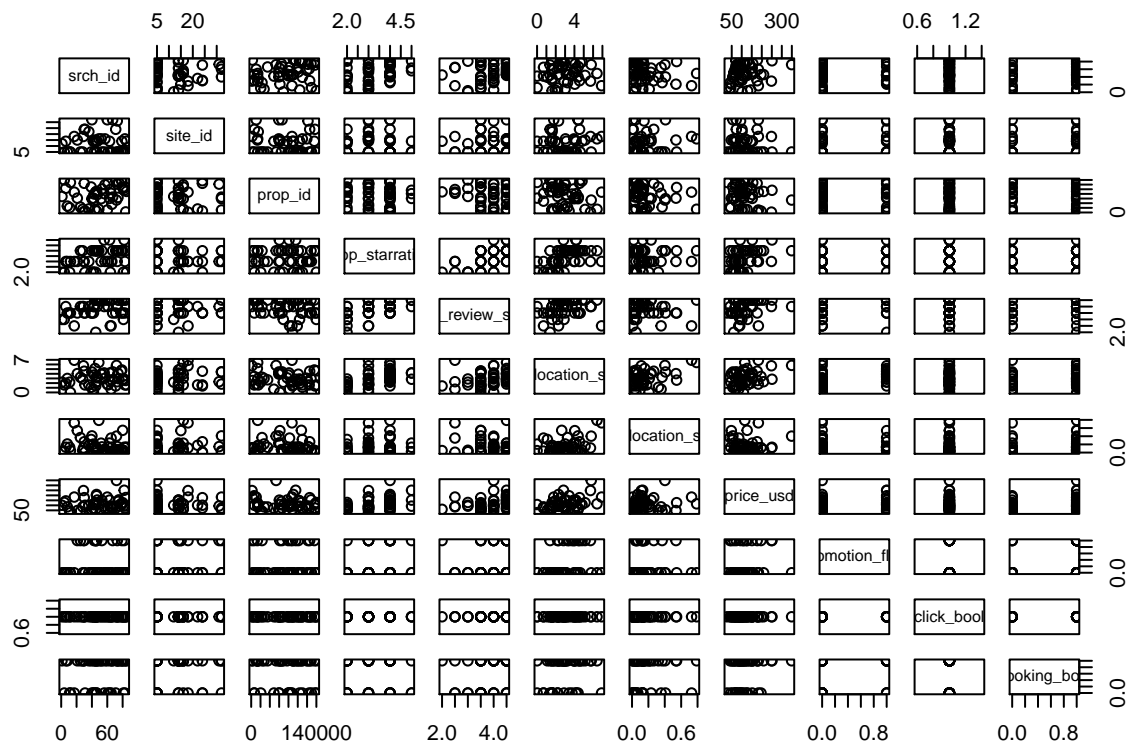
```
length((unique(mydata$srch_id)))
```

```
## [1] 2360
```

how many unique search ids

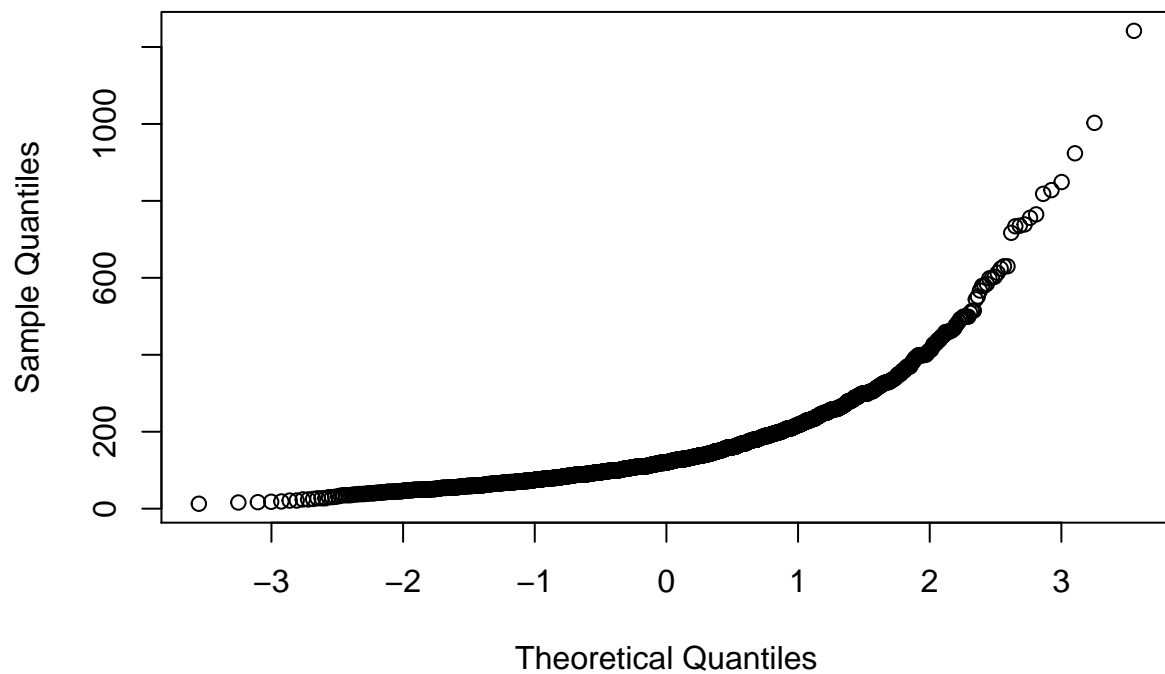
pairs plot

```
df <- subset(head(mydata, 50), select = -c(consumer))  
pairs(df)
```



```
qqnorm(mydata$price_usd)
```

Normal Q-Q Plot



```
shapiro.test(mydata$price_usd)
```

```
##
## Shapiro-Wilk normality test
```

```

##
## data: mydata$price_usd
## W = 0.76229, p-value < 2.2e-16
not normally distributed

library(outliers)
chisq.out.test(mydata$price_usd, variance = var(mydata$price_usd),opposite = TRUE)

##
## chi-squared test for outlier
##
## data: mydata$price_usd
## X-squared = 1.8554, p-value = 0.1732
## alternative hypothesis: lowest value 12.8 is an outlier
chisq.out.test(mydata$price_usd, variance = var(mydata$price_usd),opposite = FALSE)

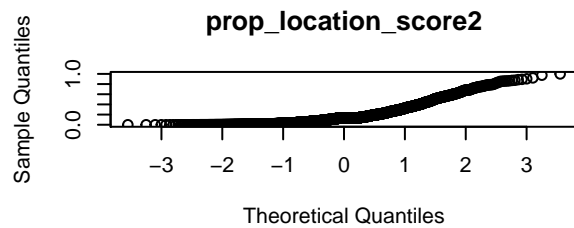
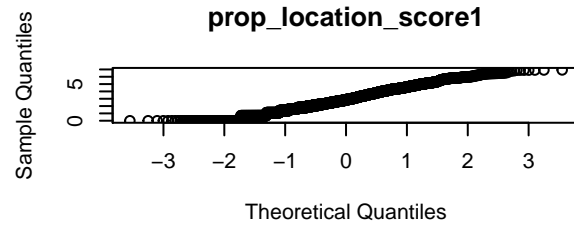
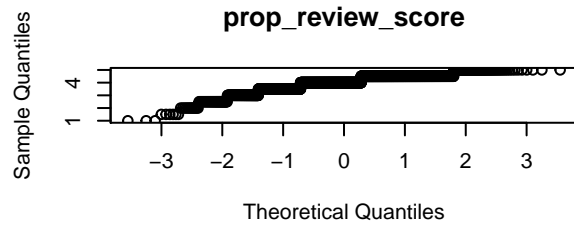
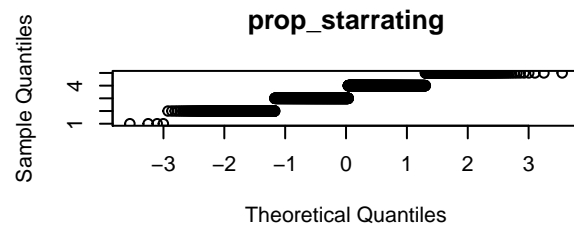
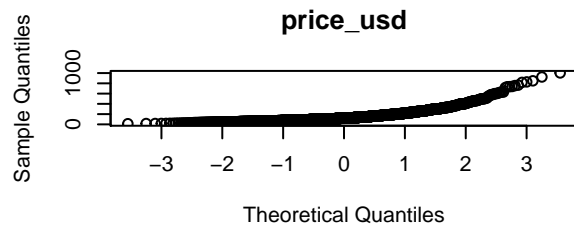
##
## chi-squared test for outlier
##
## data: mydata$price_usd
## X-squared = 121.28, p-value < 2.2e-16
## alternative hypothesis: highest value 1242 is an outlier
removing outliers
mydata <- mydata[which(mydata$price_usd < 1242 ),]
chisq.out.test(mydata$price_usd, variance = var(mydata$price_usd),opposite = TRUE)

##
## chi-squared test for outlier
##
## data: mydata$price_usd
## X-squared = 1.9331, p-value = 0.1644
## alternative hypothesis: lowest value 12.8 is an outlier
chisq.out.test(mydata$price_usd, variance = var(mydata$price_usd),opposite = FALSE)

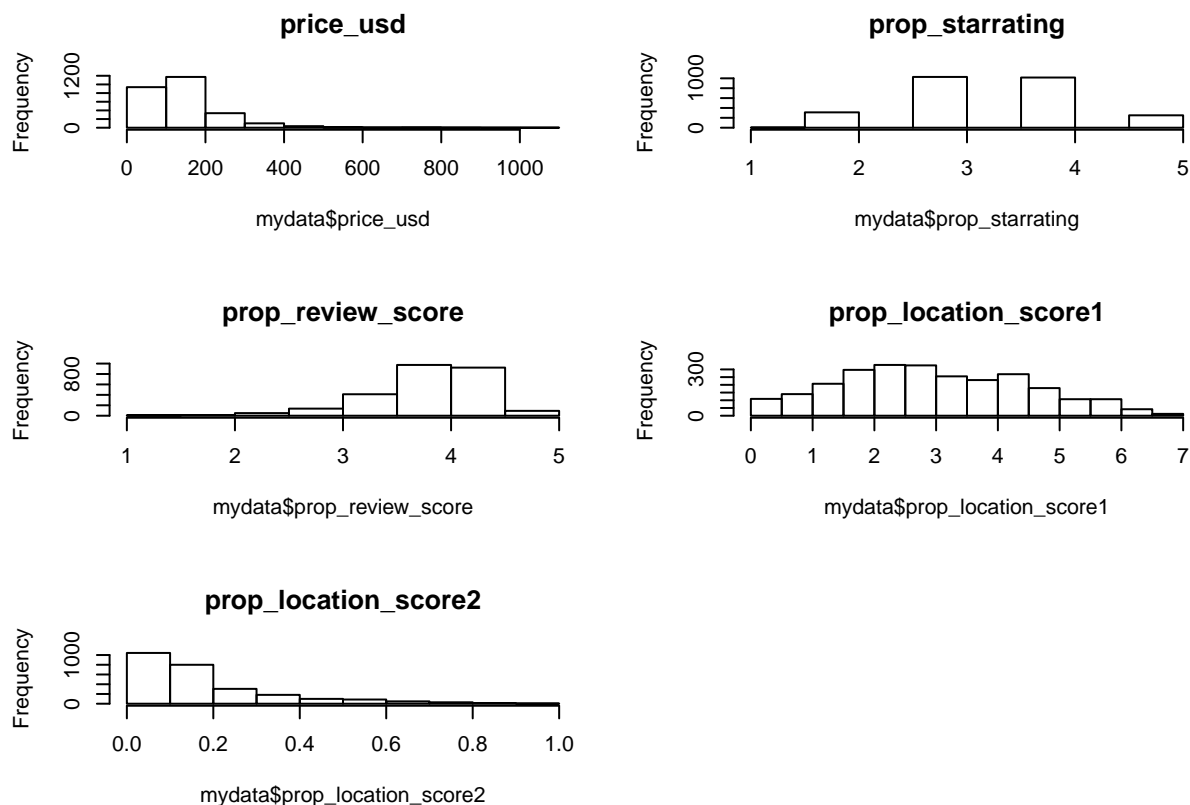
##
## chi-squared test for outlier
##
## data: mydata$price_usd
## X-squared = 77.699, p-value < 2.2e-16
## alternative hypothesis: highest value 1002.82 is an outlier
# library(plotly)
# library(ggplot2)
# plot_ly(x = mydata$price_usd, type = "histogram")
# ggplotly(p)
# ggplotly(p)

par(mfrow=c(3,2));
qqnorm(mydata$price_usd,main="price_usd")
qqnorm(mydata$prop_starrating,main="prop_starrating")
qqnorm(mydata$prop_review_score,main="prop_review_score")
qqnorm(mydata$prop_location_score1,main="prop_location_score1")
qqnorm(mydata$prop_location_score2,main="prop_location_score2")

```



```
par(mfrow=c(3,2));
hist(mydata$price_usd,main="price_usd")
hist(mydata$prop_starrating,main="prop_starrating")
hist(mydata$prop_review_score,main="prop_review_score")
hist(mydata$prop_location_score1,main="prop_location_score1")
hist(mydata$prop_location_score2,main="prop_location_score2")
```



```
df <- subset(mydata, select = -c(consumer,click_bool,srch_id,site_id,prop_id))
round(cor(df),3)
```

```
##               prop_starrating prop_review_score
## prop_starrating             1.000             0.424
## prop_review_score           0.424             1.000
## prop_location_score1        0.284             0.095
## prop_location_score2        0.018             0.004
## price_usd                   0.502             0.329
## promotion_flag              0.144            -0.019
## booking_bool               -0.057             0.027
##               prop_location_score1 prop_location_score2 price_usd
## prop_starrating             0.284             0.018     0.502
## prop_review_score           0.095             0.004     0.329
## prop_location_score1        1.000             0.279     0.269
## prop_location_score2        0.279             1.000     0.057
## price_usd                   0.269             0.057     1.000
## promotion_flag              0.152             0.005    -0.060
## booking_bool               -0.070             0.069    -0.172
##               promotion_flag booking_bool
## prop_starrating             0.144     -0.057
## prop_review_score          -0.019      0.027
## prop_location_score1        0.152     -0.070
## prop_location_score2        0.005      0.069
## price_usd                  -0.060     -0.172
## promotion_flag              1.000      0.047
## booking_bool                0.047      1.000
```

```
prop_starrating prop_review_score prop_location_score1 prop_location_score2 price_usd p
```

```
prop_starrating 1.00 0.31 0.30 0.05 0.47 0.19 0.00 prop_review_score 0.31 1.00 0.15 0.04 0.22 0.06 0.12
prop_location_score1 0.30 0.15 1.00 0.29 0.27 0.16 0.00 prop_location_score2 0.05 0.04 0.29 1.00 0.02
0.03 0.11 price_usd 0.47 0.22 0.27 0.02 1.00 -0.01 -0.11 promotion_flag 0.19 0.06 0.16 0.03 -0.01 1.00 0.11
booking_bool 0.00 0.12 0.00 0.11 -0.11 0.11 1.00
```

with 0.47 prop_starrating has the highest colinearity with price_usd followed by 0.31 for prop_starrating and prop_review_score

randomized block design

```
xtabs(price_usd ~ prop_starrating + consumer ,data=mydata)
```

```
##               consumer
## prop_starrating couple   other Parents   single
##               1  283.27    0.00    0.00    71.00
##               2 14137.50 3322.21 2472.32 6026.93
##               3 70900.22 14139.67 7777.78 26433.55
##               4 97350.56 21189.41 12108.00 39897.67
##               5 41575.44 8771.48 5074.91 13921.75
```

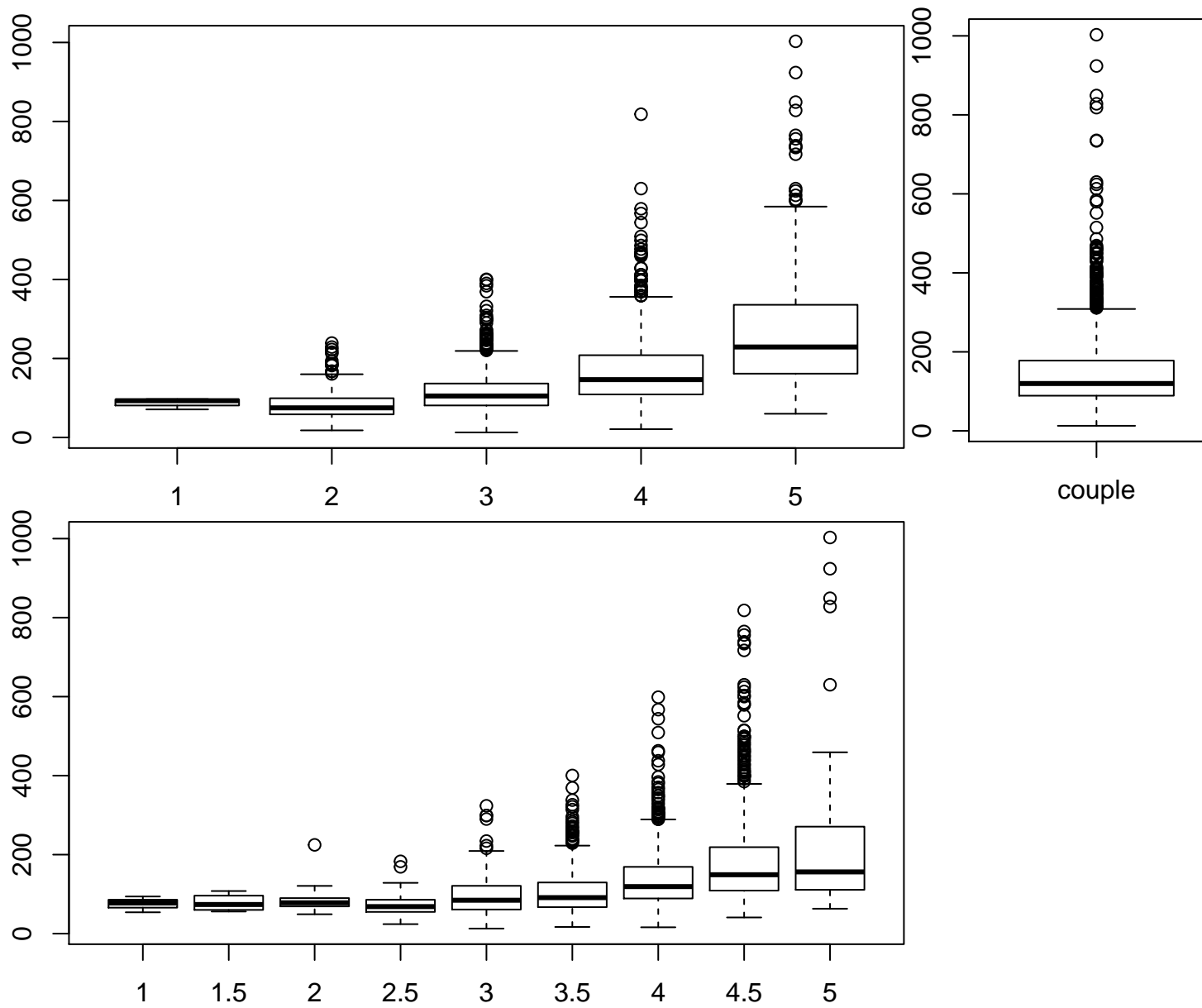
```
xtabs(price_usd ~ prop_starrating + prop_review_score ,data=mydata)
```

```
##               prop_review_score
## prop_starrating      1      1.5      2      2.5      3      3.5
##               1    0.00    0.00    0.00   71.00    0.00  187.94
##               2   94.00   130.00  602.96 2557.00 4836.28 7592.06
##               3   54.00   241.20  258.23  842.83 6289.98 20924.98
##               4   77.46   96.14  302.71  253.01 2007.35 14835.05
##               5    0.00    0.00    0.00    0.00  290.75 1369.09
##               prop_review_score
## prop_starrating      4      4.5      5
##               1    0.00    0.00   95.33
##               2  7519.50 2366.98  260.18
##               3 48109.98 38167.51 4362.51
##               4 67381.06 79444.55 6148.31
##               5 10287.25 47696.34 9700.15
```

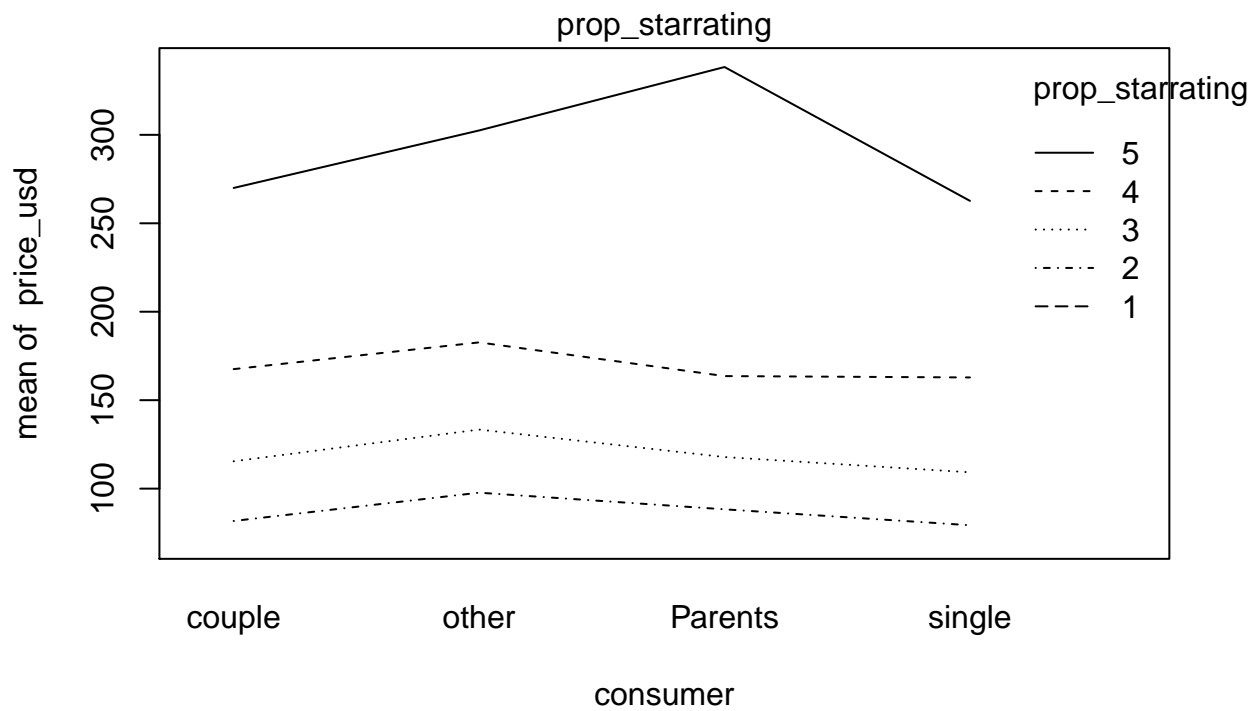
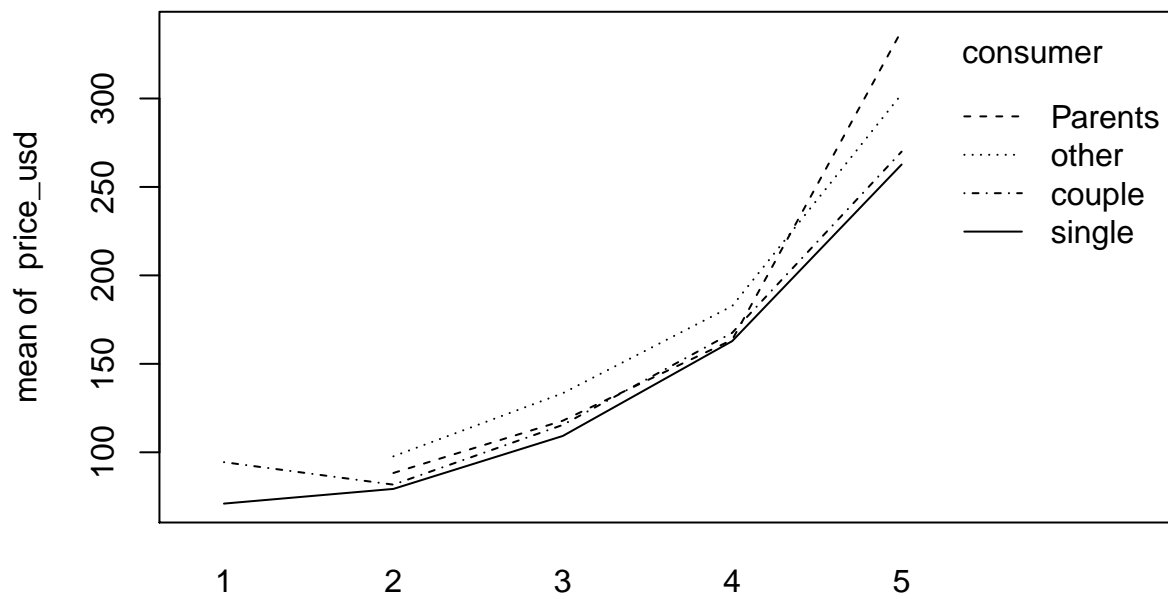
```
attach(mydata)
```

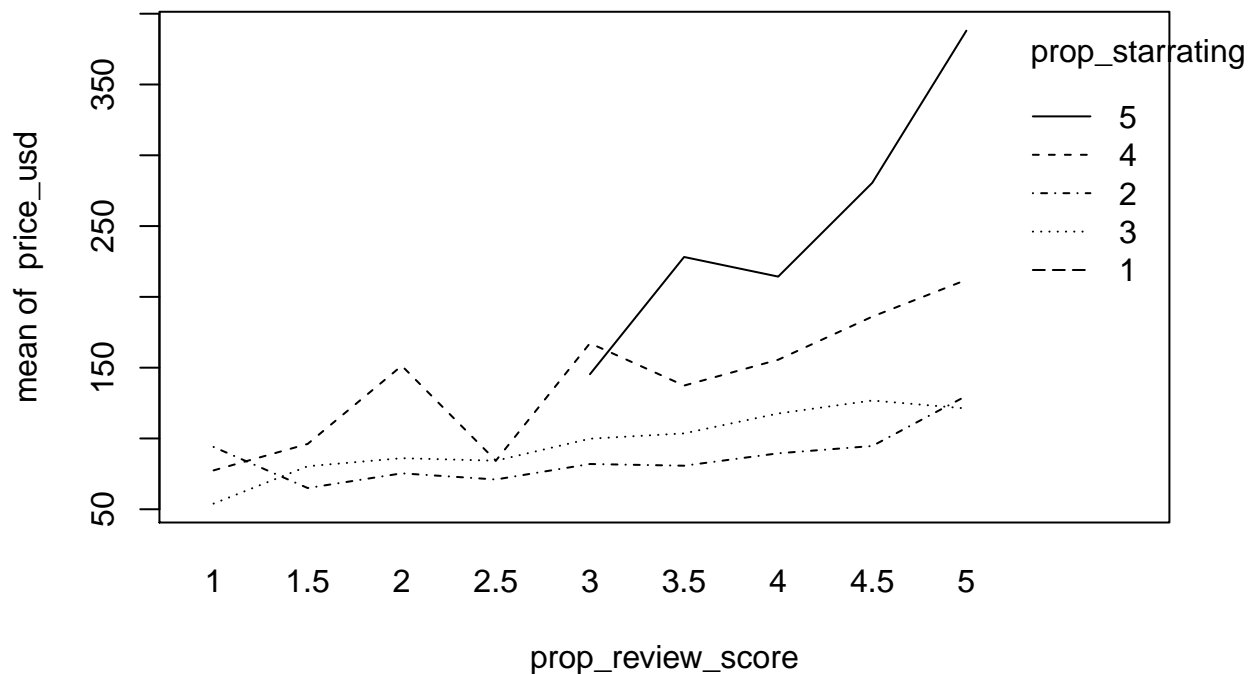
```
# par(mfrow=c(3,3))
```

```
boxplot(price_usd~prop_starrating); boxplot(price_usd~consumer) ; boxplot(price_usd~prop_review_score)
```



```
interaction.plot(prop_starrating,consumer,price_usd); interaction.plot(consumer,prop_starrating,price_usd)
```





```
mydata$prop_starrating=factor(mydata$prop_starrating)
mydata$prop_review_score=factor(mydata$prop_review_score)

aovpen=lm(price_usd~prop_starrating+prop_review_score,data=mydata)

anova(aovpen)
```

```
## Analysis of Variance Table
##
## Response: price_usd
##              Df    Sum Sq Mean Sq F value    Pr(>F)
## prop_starrating  4  6891830 1722958 263.075 < 2.2e-16 ***
## prop_review_score  8   654198   81775  12.486 < 2.2e-16 ***
## Residuals      2597 17008542    6549
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aovpen)
```

```
##
## Call:
## lm(formula = price_usd ~ prop_starrating + prop_review_score,
##     data = mydata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -218.34  -44.54  -14.29   27.88   689.93
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    41.344     61.947   0.667   0.5046
## prop_starrating2    4.234     40.845   0.104   0.9174
## prop_starrating3   24.803     40.731   0.609   0.5426
```

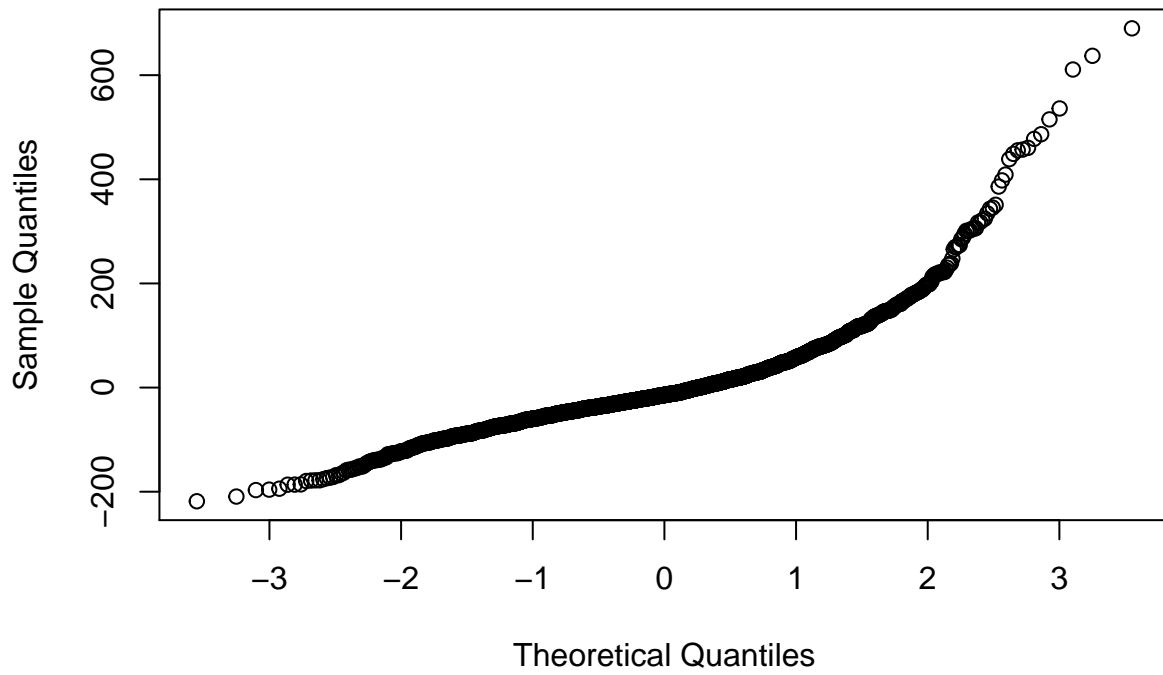
```

## prop_starrating4      72.390      40.767      1.776      0.0759 .
## prop_starrating5     169.608      41.022      4.135 3.67e-05 ***
## prop_review_score1.5   10.668      57.228      0.186      0.8521
## prop_review_score2     28.720      51.858      0.554      0.5797
## prop_review_score2.5   20.780      48.161      0.431      0.6662
## prop_review_score3     35.156      47.245      0.744      0.4569
## prop_review_score3.5   33.086      46.899      0.705      0.4806
## prop_review_score4     44.191      46.815      0.944      0.3453
## prop_review_score4.5   67.388      46.840      1.439      0.1504
## prop_review_score5    101.942      47.524      2.145      0.0320 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 80.93 on 2597 degrees of freedom
## Multiple R-squared:  0.3073, Adjusted R-squared:  0.3041
## F-statistic: 96.02 on 12 and 2597 DF,  p-value: < 2.2e-16
drop1(aovpen,test="Chisq")

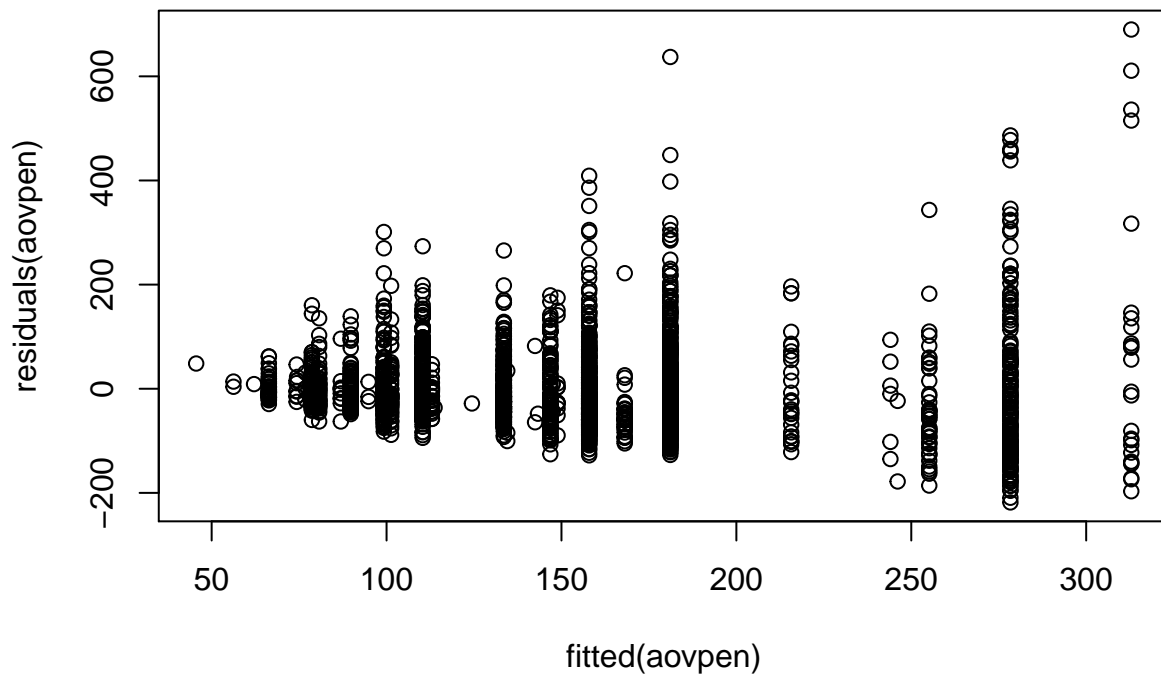
## Single term deletions
##
## Model:
## price_usd ~ prop_starrating + prop_review_score
##              Df Sum of Sq      RSS   AIC  Pr(>Chi)
## <none>                        17008542 22947
## prop_starrating    4   4564659 21573201 23560 < 2.2e-16 ***
## prop_review_score   8    654198 17662741 23030 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
qqnorm(residuals(aovpen))

```

Normal Q-Q Plot



```
plot(fitted(aovpen),residuals(aovpen))
```



prop_starrating is more significant than consumer type and prop_review_score but also both significant

Logistic Regression

An experiment with: an outcome Y that is 0 or 1 (“binary dependent variable”); one or more numerical explanatory variables X_1, \dots, X_p . one or more factor explanatory variables. (“independent variable”). The purpose is to explain Y by a function of X.

```
# tot=xtabs(~prop_review_score+price_usd,data=mydata);
# hist(mydata$price_usd,main="price_usd");
# round(xtabs(booking_bool~prop_review_score+price_usd,data=mydata)/tot,2)
#
# totage=xtabs(~prop_review_score,data=mydata)
# barplot(xtabs(booking_bool~prop_review_score,data=mydata)/totage)
#
# mydata$prop_review_score2 <- mydata$prop_review_score^2
# myglm=glm(booking_bool~prop_review_score+prop_review_score2+price_usd,data=mydata,family=binomial)
# summary(myglm)
```

```
myglm=glm(booking_bool~
           prop_starrating
           +prop_review_score
           +prop_location_score1
           +prop_location_score2
           +price_usd
           +promotion_flag
           +consumer
           ,data=mydata,family=binomial)
summary(myglm)
```

```
##
## Call:
## glm(formula = booking_bool ~ prop_starrating + prop_review_score +
##      prop_location_score1 + prop_location_score2 + price_usd +
##      promotion_flag + consumer, family = binomial, data = mydata)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.9877  -1.2780   0.7871   0.9419   2.2879
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.3931195   1.6996472  -1.408 0.159128
## prop_starrating2    2.0123613   1.1719039   1.717 0.085948 .
## prop_starrating3    1.8186350   1.1688833   1.556 0.119738
## prop_starrating4    1.8553442   1.1701929   1.586 0.112852
## prop_starrating5    2.0271625   1.1799129   1.718 0.085785 .
## prop_review_score1.5 -0.7377834   1.6520493  -0.447 0.655173
## prop_review_score2    0.9377906   1.3552031   0.692 0.488942
## prop_review_score2.5  1.2259385   1.2684957   0.966 0.333819
## prop_review_score3    0.7297695   1.2435368   0.587 0.557305
## prop_review_score3.5  1.3953167   1.2354380   1.129 0.258725
## prop_review_score4    1.7001720   1.2332443   1.379 0.168013
## prop_review_score4.5  1.7227357   1.2340611   1.396 0.162718
## prop_review_score5    1.5845126   1.2522512   1.265 0.205753
## prop_location_score1 -0.0897714   0.0307003  -2.924 0.003454 **
## prop_location_score2  1.3253697   0.2705760   4.898 9.67e-07 ***
```

```
## price_usd          -0.0043143  0.0005711  -7.554 4.23e-14 ***
## promotion_flag     0.2115738  0.0962113   2.199 0.027874 *
## consumerother      0.0894607  0.1374709   0.651 0.515201
## consumerParents   -0.0270289  0.1655778  -0.163 0.870329
## consumersingle     0.3819815  0.1050033   3.638 0.000275 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##    Null deviance: 3442.2  on 2609  degrees of freedom
## Residual deviance: 3278.5  on 2590  degrees of freedom
## AIC: 3318.5
##
## Number of Fisher Scoring iterations: 4
drop1(myglm,test="Chisq")

## Single term deletions
##
## Model:
## booking_bool ~ prop_starrating + prop_review_score + prop_location_score1 +
##   prop_location_score2 + price_usd + promotion_flag + consumer
##              Df Deviance    AIC    LRT Pr(>Chi)
## <none>                3278.5 3318.5
## prop_starrating      4   3284.7 3316.7  6.151 0.188168
## prop_review_score     8   3315.3 3339.3 36.792 1.257e-05 ***
## prop_location_score1  1   3287.1 3325.1  8.592 0.003377 **
## prop_location_score2  1   3303.8 3341.8 25.332 4.826e-07 ***
## price_usd            1   3341.6 3379.6 63.067 1.998e-15 ***
## promotion_flag       1   3283.4 3321.4  4.875 0.027250 *
## consumer             3   3292.7 3326.7 14.193 0.002654 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
myglm=glm(booking_bool~
  prop_review_score
  +prop_location_score1
  +prop_location_score2
  +price_usd
  +promotion_flag
  +consumer
  ,data=mydata,family=binomial)
summary(myglm)

##
## Call:
## glm(formula = booking_bool ~ prop_review_score + prop_location_score1 +
##   prop_location_score2 + price_usd + promotion_flag + consumer,
##   family = binomial, data = mydata)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.9782  -1.2881   0.7901   0.9427   2.2252
##
```

```
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.5089489   1.2359990  -0.412 0.680507
## prop_review_score1.5 -0.7238342   1.6527538  -0.438 0.661418
## prop_review_score2    0.9868408   1.3573111   0.727 0.467192
## prop_review_score2.5  1.2568074   1.2699084   0.990 0.322329
## prop_review_score3    0.7441285   1.2467621   0.597 0.550608
## prop_review_score3.5  1.3632721   1.2391255   1.100 0.271250
## prop_review_score4    1.6587396   1.2369201   1.341 0.179913
## prop_review_score4.5  1.6905400   1.2376793   1.366 0.171972
## prop_review_score5    1.5343434   1.2556226   1.222 0.221716
## prop_location_score1 -0.0920050   0.0302119  -3.045 0.002324 **
## prop_location_score2  1.2830149   0.2686645   4.776 1.79e-06 ***
## price_usd         -0.0041031   0.0005139  -7.985 1.41e-15 ***
## promotion_flag      0.2190510   0.0944278   2.320 0.020353 *
## consumerother       0.0929271   0.1373123   0.677 0.498560
## consumerParents    -0.0220893   0.1652853  -0.134 0.893684
## consumersingle      0.3807132   0.1048163   3.632 0.000281 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##    Null deviance: 3442.2  on 2609  degrees of freedom
## Residual deviance: 3284.7  on 2594  degrees of freedom
## AIC: 3316.7
##
## Number of Fisher Scoring iterations: 4
```

```
drop1(myglm,test="Chisq")
```

```
## Single term deletions
##
## Model:
## booking_bool ~ prop_review_score + prop_location_score1 + prop_location_score2 +
##   price_usd + promotion_flag + consumer
##               Df Deviance    AIC    LRT Pr(>Chi)
## <none>                3284.7 3316.7
## prop_review_score      8  3322.5 3338.5 37.891 7.884e-06 ***
## prop_location_score1   1  3294.0 3324.0  9.316 0.002271 **
## prop_location_score2   1  3308.7 3338.7 24.016 9.555e-07 ***
## price_usd              1  3355.6 3385.6 70.994 < 2.2e-16 ***
## promotion_flag         1  3290.1 3320.1  5.431 0.019781 *
## consumer               3  3298.8 3324.8 14.097 0.002776 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mydata$consumer=factor(mydata$consumer)
mydata$prop_review_score=factor(mydata$prop_review_score)
myglm=glm(booking_bool~
  prop_review_score
+prop_location_score2
+price_usd
+promotion_flag
+consumer
```

```
,data=mydata,family=binomial)
summary(myglm)
```

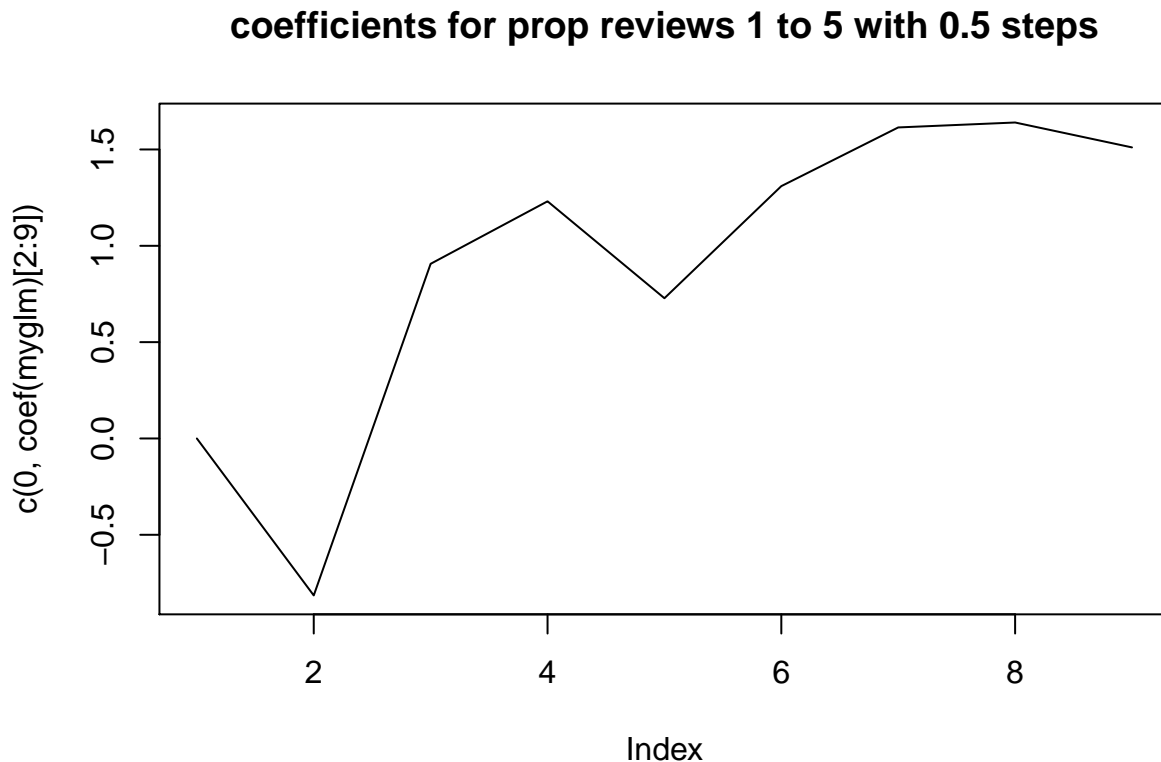
```
##
## Call:
## glm(formula = booking_bool ~ prop_review_score + prop_location_score2 +
##      price_usd + promotion_flag + consumer, family = binomial,
##      data = mydata)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.9566  -1.3009   0.8048   0.9368   2.2769
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.625446   1.231149  -0.508 0.611440
## prop_review_score1.5 -0.814697   1.648238  -0.494 0.621106
## prop_review_score2    0.906527   1.352144   0.670 0.502579
## prop_review_score2.5  1.230928   1.265531   0.973 0.330724
## prop_review_score3    0.728006   1.242430   0.586 0.557907
## prop_review_score3.5  1.309675   1.234693   1.061 0.288813
## prop_review_score4    1.614361   1.232521   1.310 0.190262
## prop_review_score4.5  1.639964   1.233256   1.330 0.183589
## prop_review_score5    1.510766   1.251291   1.207 0.227291
## prop_location_score2  1.050305   0.254956   4.120 3.8e-05 ***
## price_usd        -0.004498   0.000503  -8.943 < 2e-16 ***
## promotion_flag     0.167049   0.092677   1.802 0.071469 .
## consumerother      0.076274   0.136896   0.557 0.577409
## consumerParents   -0.004631   0.164779  -0.028 0.977580
## consumersingle     0.382782   0.104743   3.654 0.000258 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 3442.2  on 2609  degrees of freedom
## Residual deviance: 3294.0  on 2595  degrees of freedom
## AIC: 3324
##
## Number of Fisher Scoring iterations: 4
```

```
drop1(myglm,test="Chisq")
```

```
## Single term deletions
##
## Model:
## booking_bool ~ prop_review_score + prop_location_score2 + price_usd +
##      promotion_flag + consumer
##              Df Deviance    AIC    LRT Pr(>Chi)
## <none>                3294.0 3324.0
## prop_review_score      8   3330.9 3344.9 36.934 1.184e-05 ***
## prop_location_score2   1   3311.7 3339.7 17.682 2.611e-05 ***
## price_usd              1   3384.7 3412.7 90.688 < 2.2e-16 ***
## promotion_flag         1   3297.2 3325.2  3.271 0.070501 .
```

```
## consumer          3   3308.1 3332.1 14.138  0.002723 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(c(0,coef(myglm)[2:9]),type="l",main = "coefficients for prop reviews 1 to 5 with 0.5 steps" )
```



```
drop1(myglm,test="Chisq")
```

```
## Single term deletions
##
## Model:
## booking_bool ~ prop_review_score + prop_location_score2 + price_usd +
##   promotion_flag + consumer
##
```

	Df	Deviance	AIC	LRT	Pr(>Chi)	
<none>		3294.0	3324.0			
prop_review_score	8	3330.9	3344.9	36.934	1.184e-05	***
prop_location_score2	1	3311.7	3339.7	17.682	2.611e-05	***
price_usd	1	3384.7	3412.7	90.688	< 2.2e-16	***
promotion_flag	1	3297.2	3325.2	3.271	0.070501	.
consumer	3	3308.1	3332.1	14.138	0.002723	**

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```