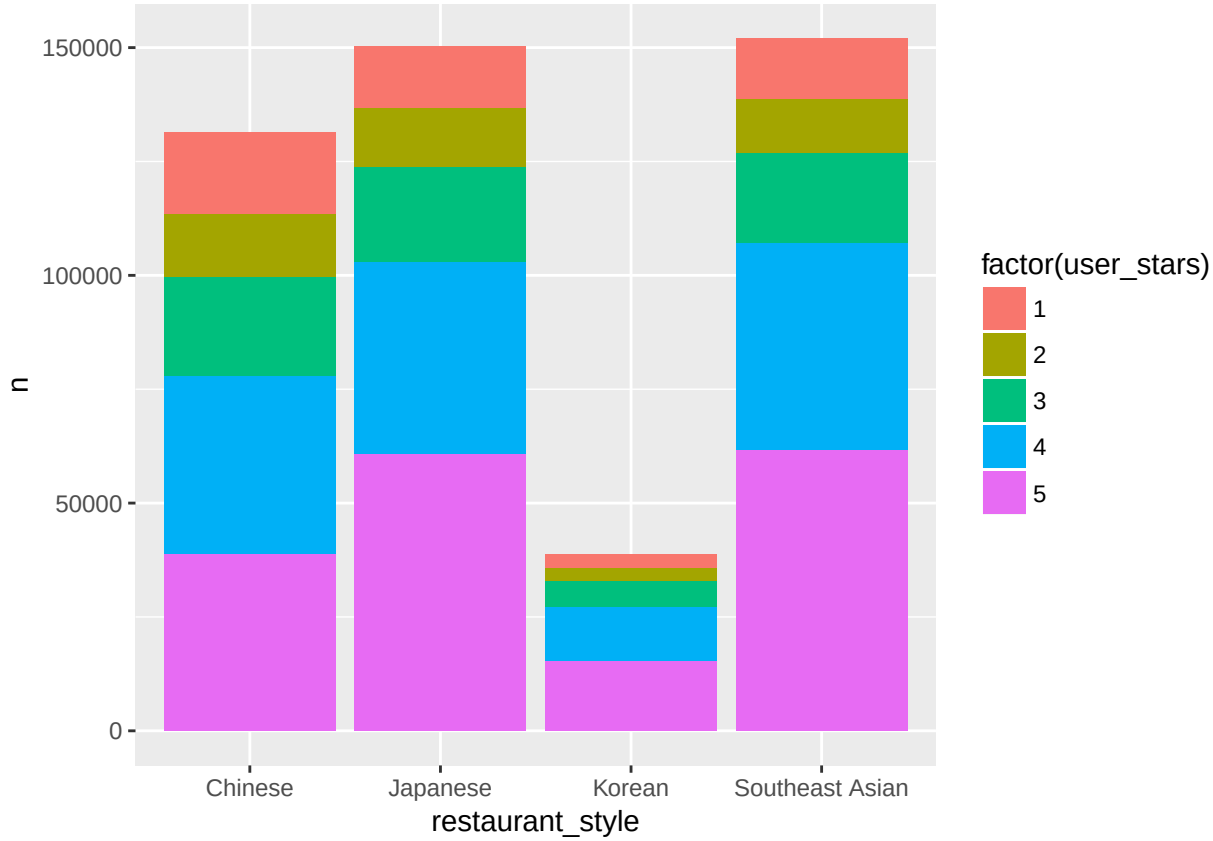
Midterm Project Yelp Data Challenge

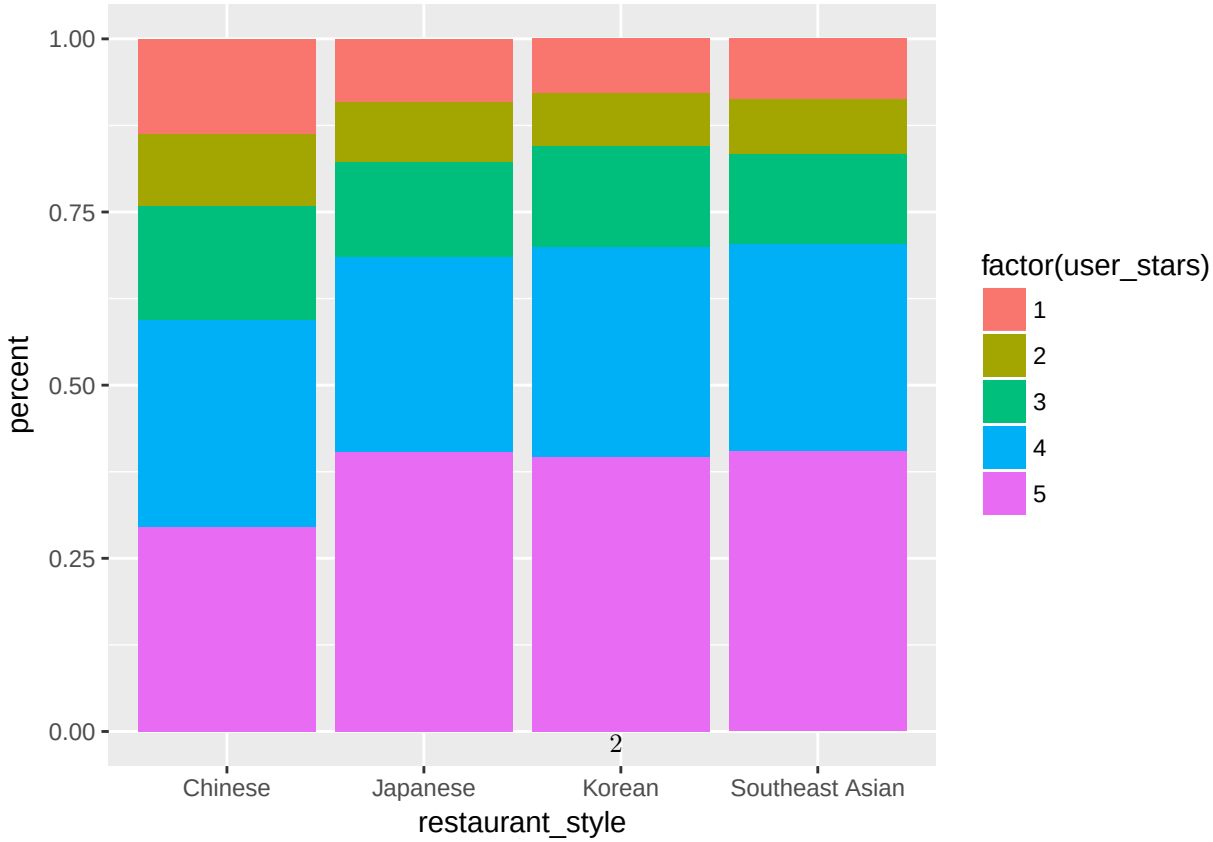
Xiang Zhao

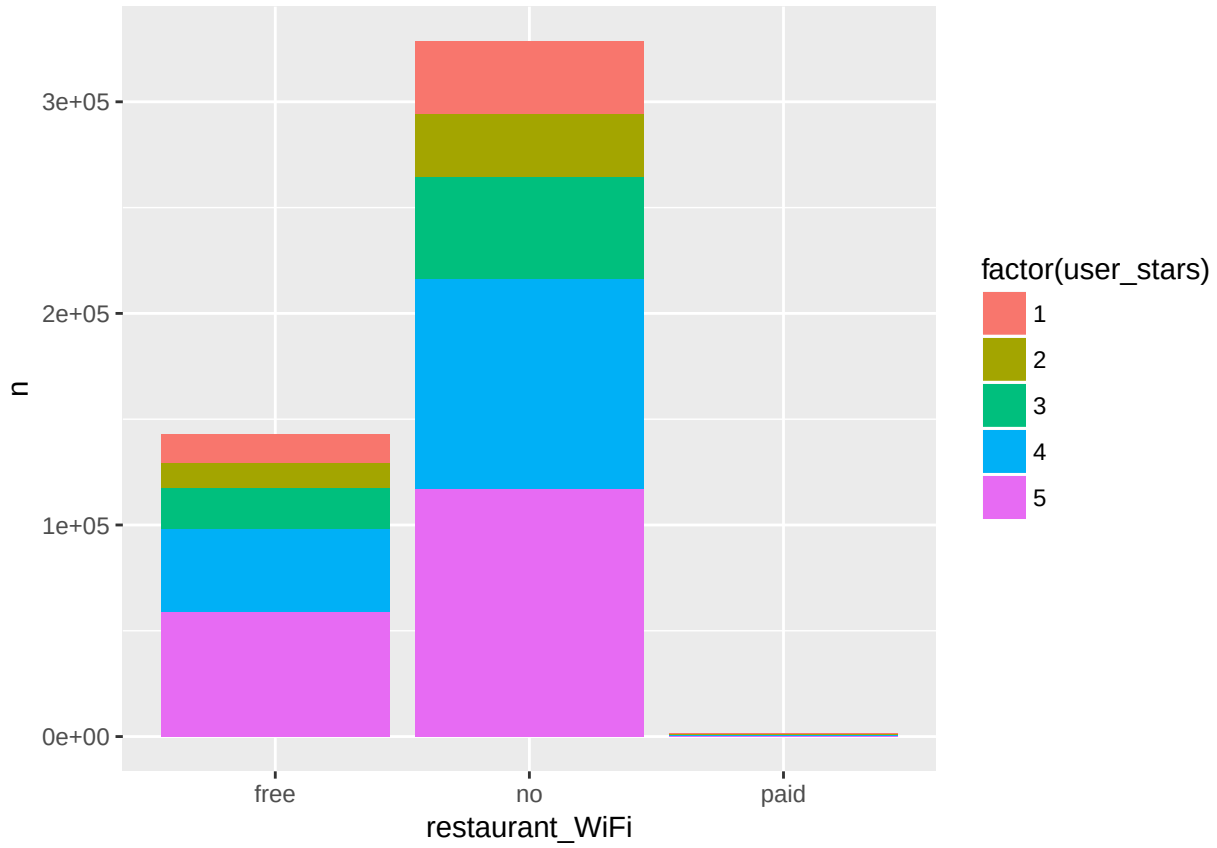
12/2/2017

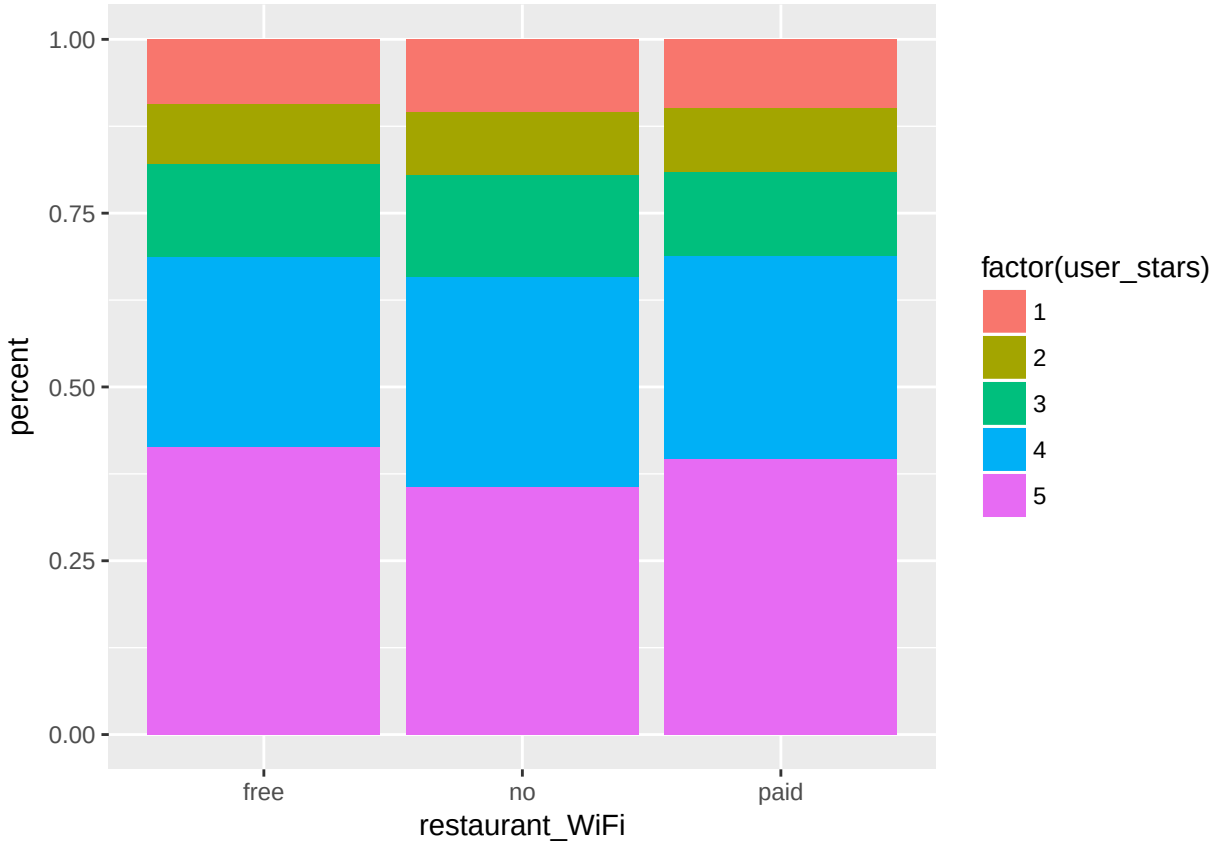
1. Introduction

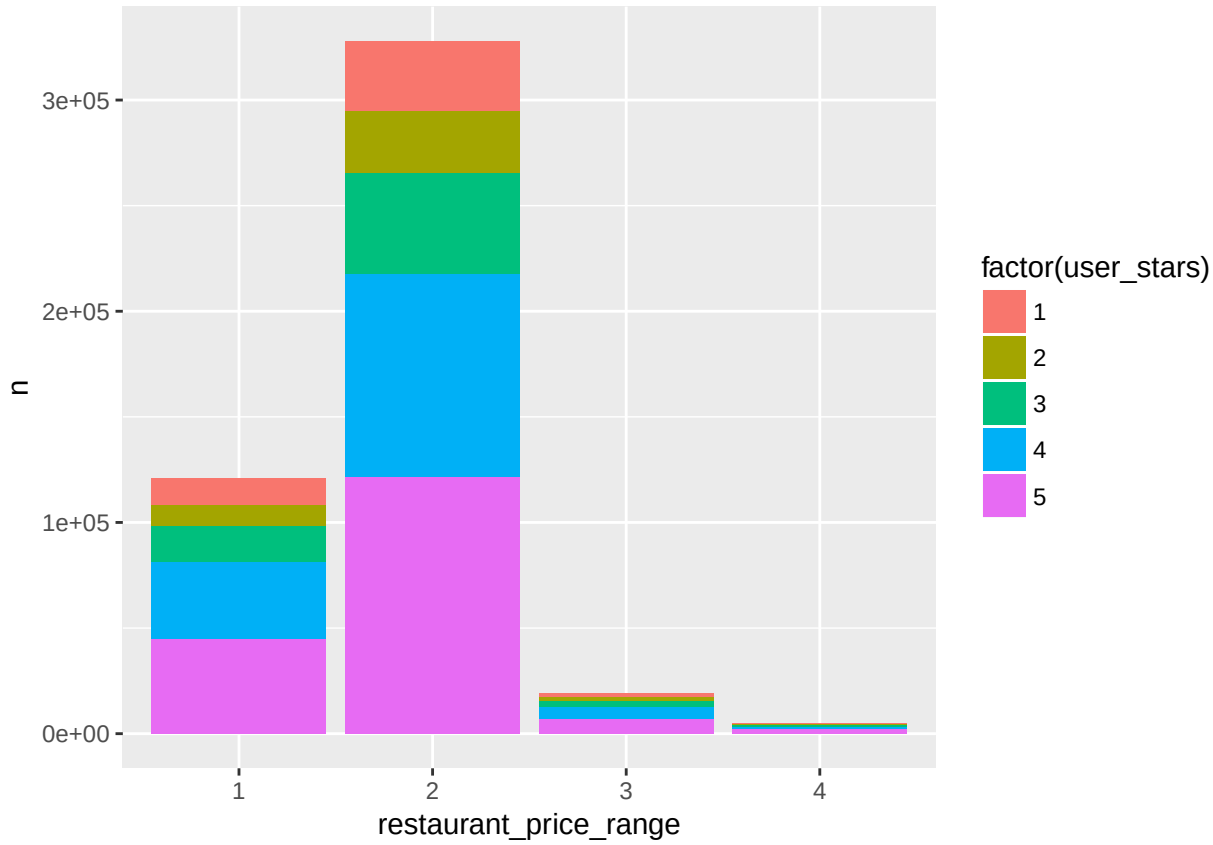
Ylep is a website and also a app collecting the information of business holders and users

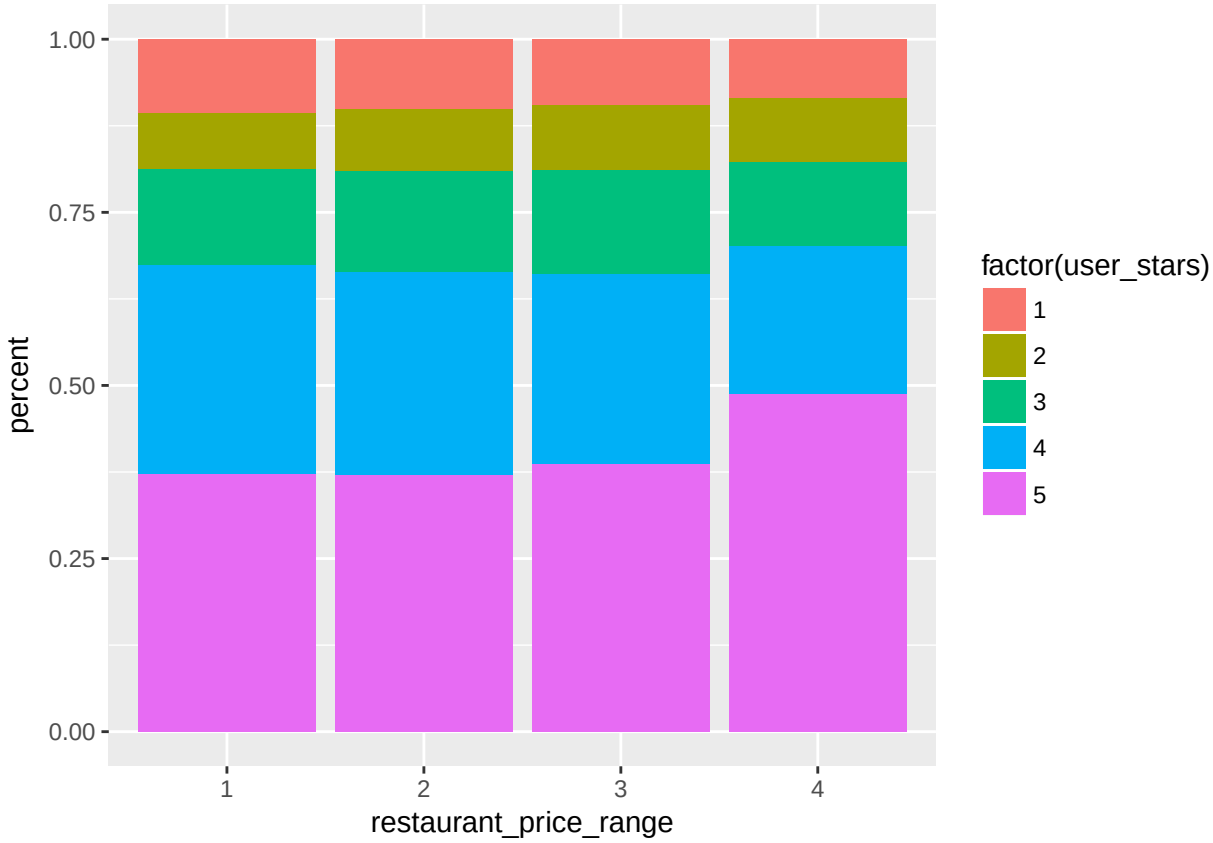


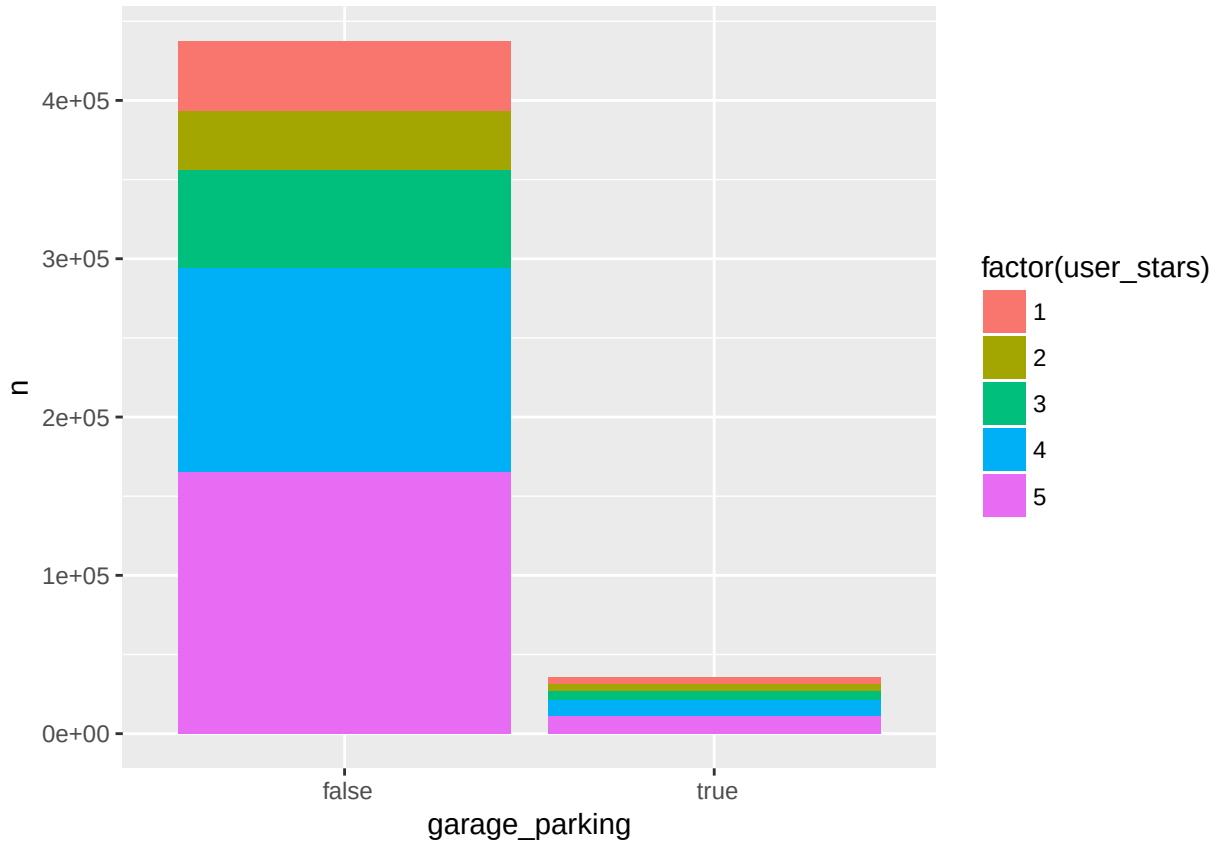


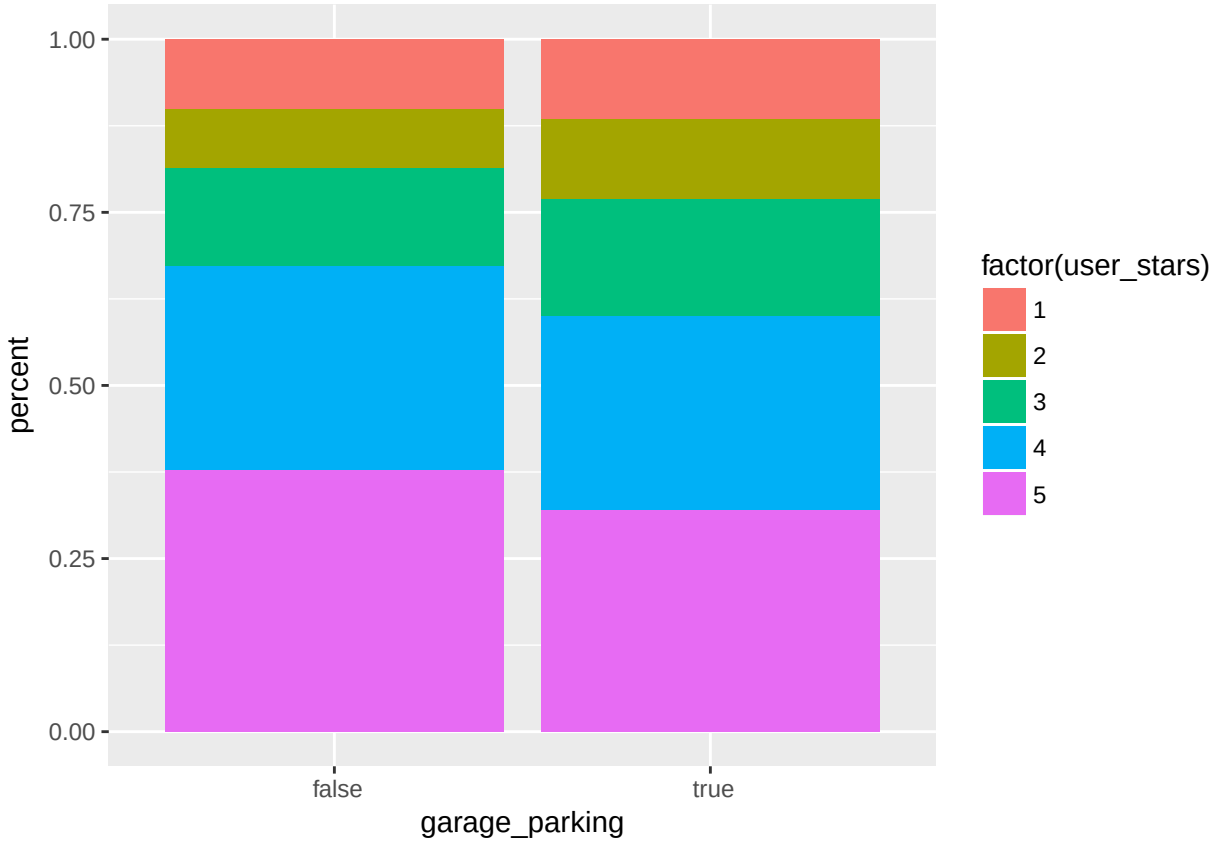

3e+05
2e+05
1e+05
0e+00
n
free
no
paid
restaurant_WiFi
factor(user_stars)
1
2
3
4
5

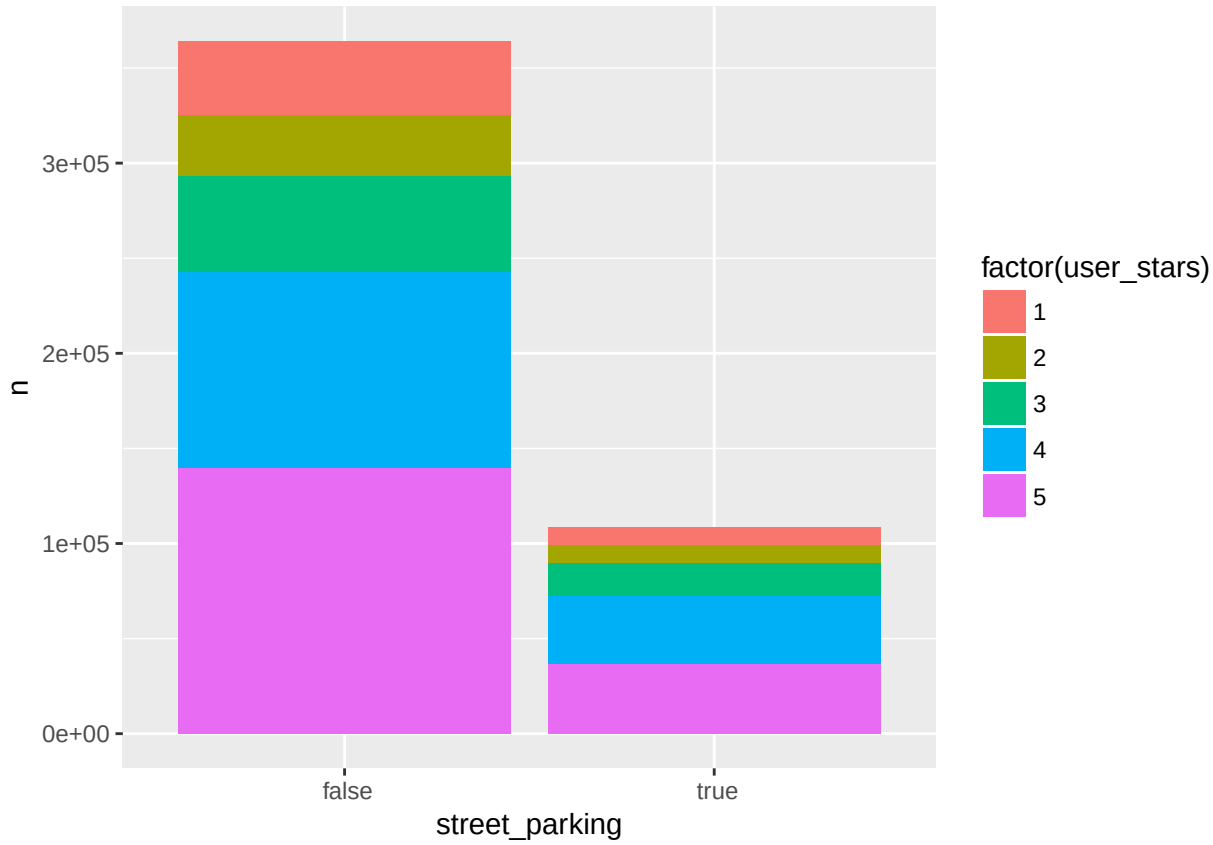

1.00
0.75
0.50
0.25
0.00
percent
free
no
paid
restaurant_WiFi
factor(user_stars)
1
2
3
4
5

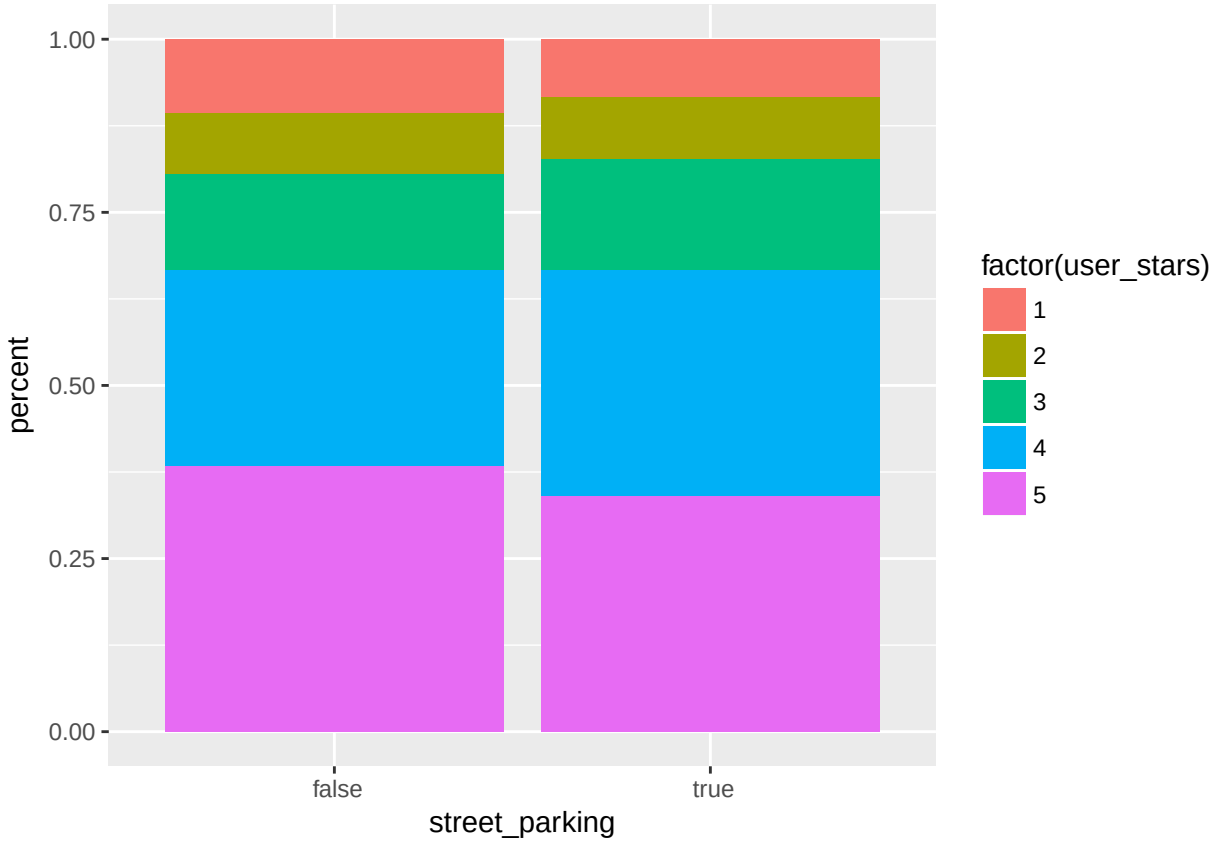

3e+05
2e+05
1e+05
0e+00
n
1
2
3
4
restaurant_price_range
factor(user_stars)
1
2
3
4
5

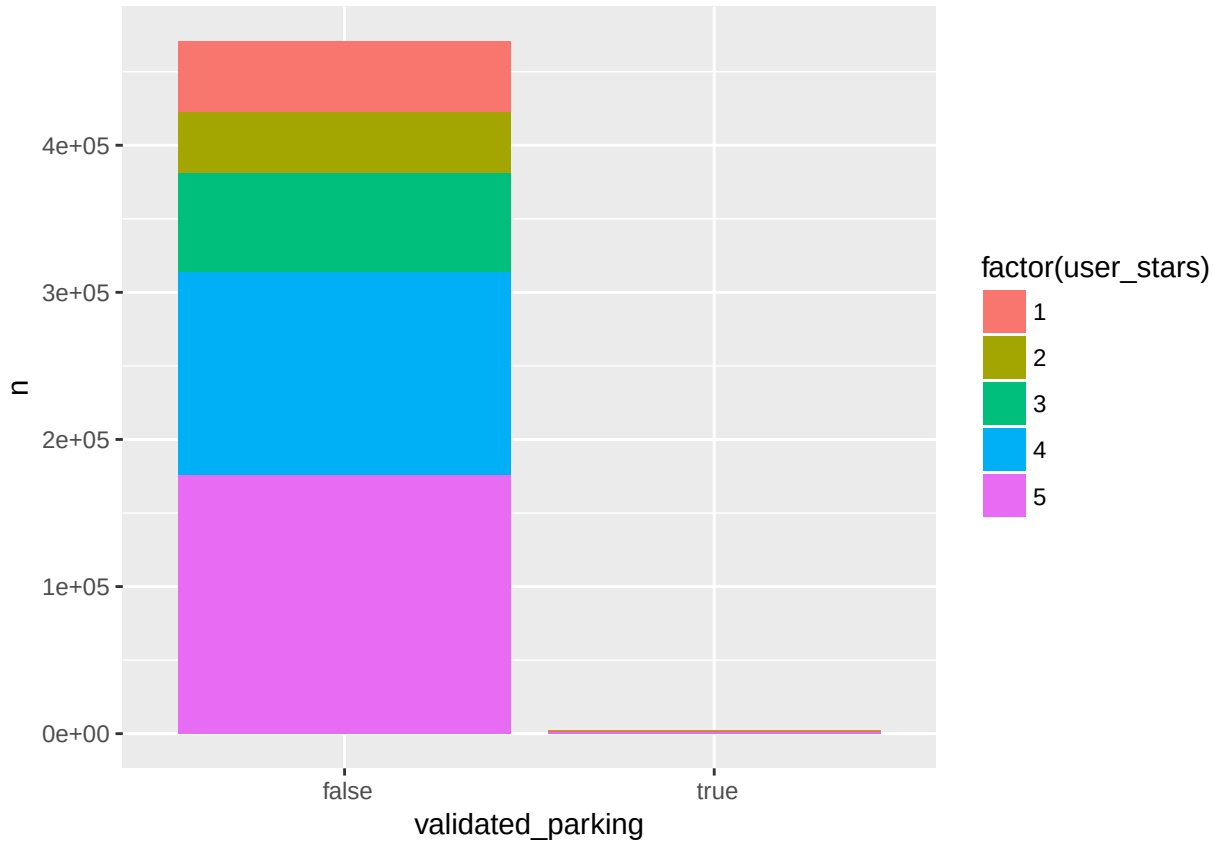

1.00
0.75
0.50
0.25
0.00
percent
1
2
3
4
restaurant_price_range
factor(user_stars)
1
2
3
4
5

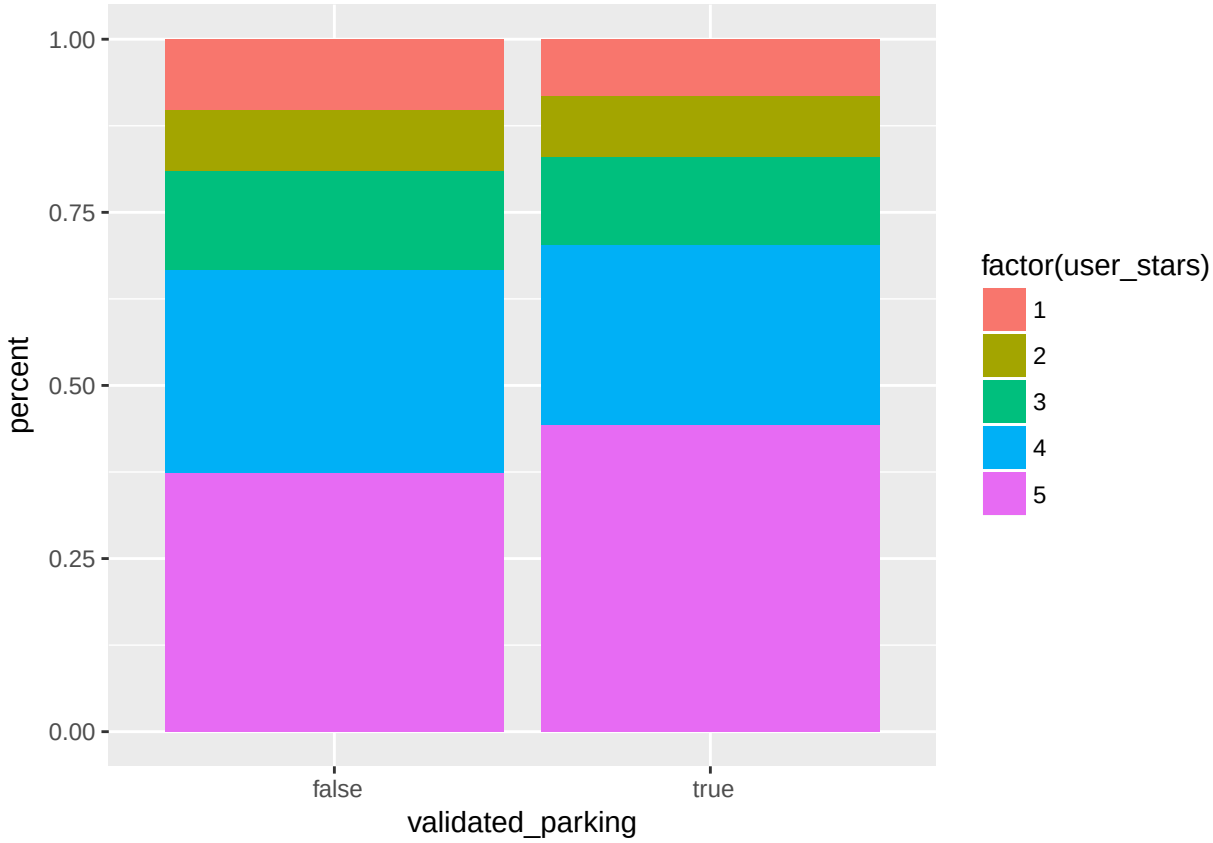

4e+05
3e+05
2e+05
1e+05
0e+00
n
false
true
garage_parking
factor(user_stars)
1
2
3
4
5

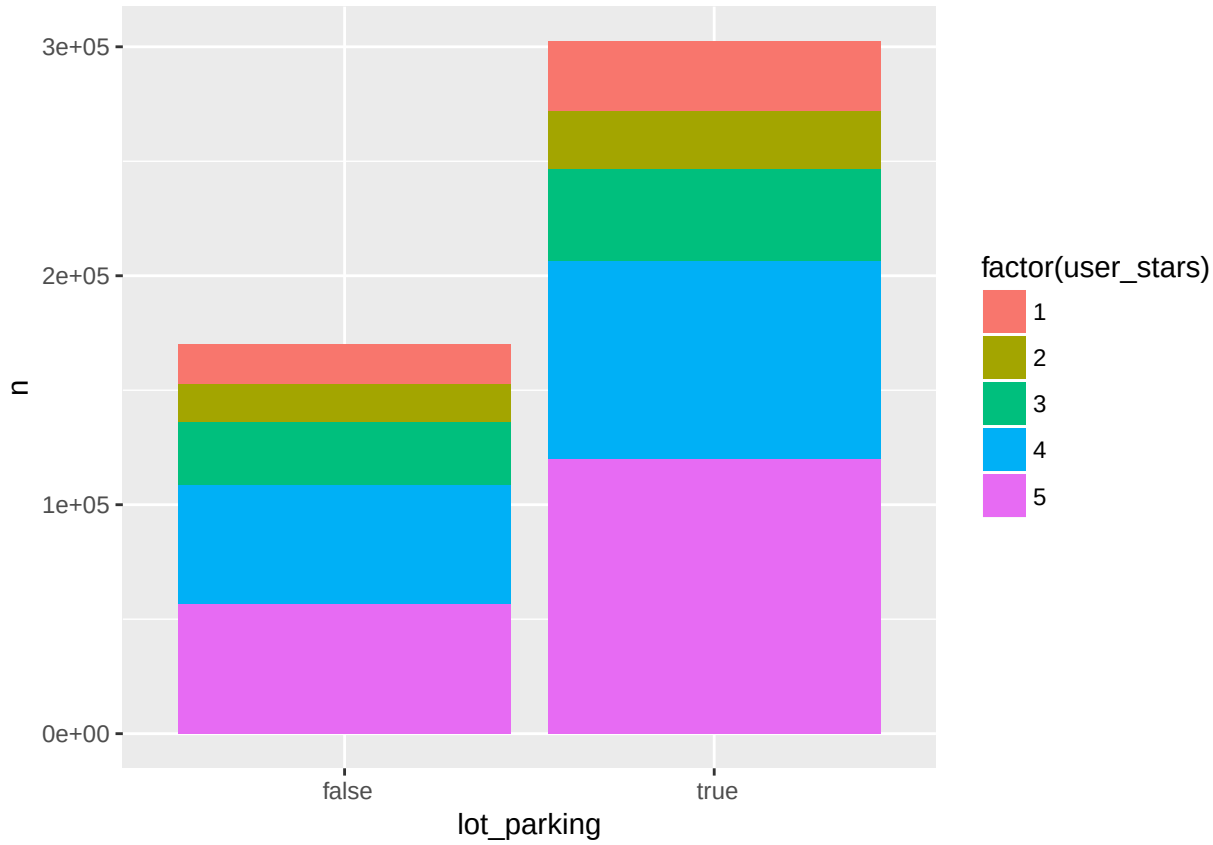

1.00
0.75
0.50
0.25
0.00
percent
false
true
garage_parking
factor(user_stars)
1
2
3
4
5

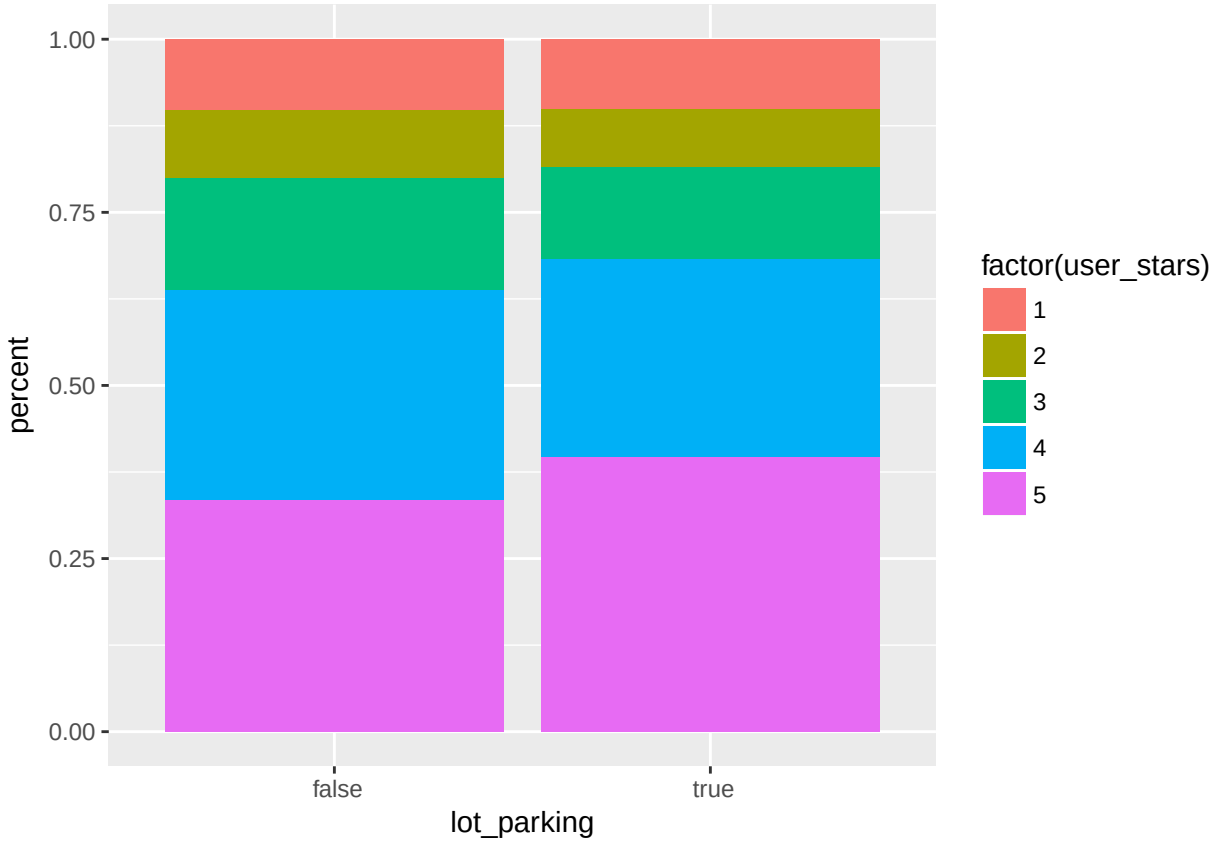

3e+05
2e+05
1e+05
0e+00
n
false
true
street_parking
factor(user_stars)
1
2
3
4
5

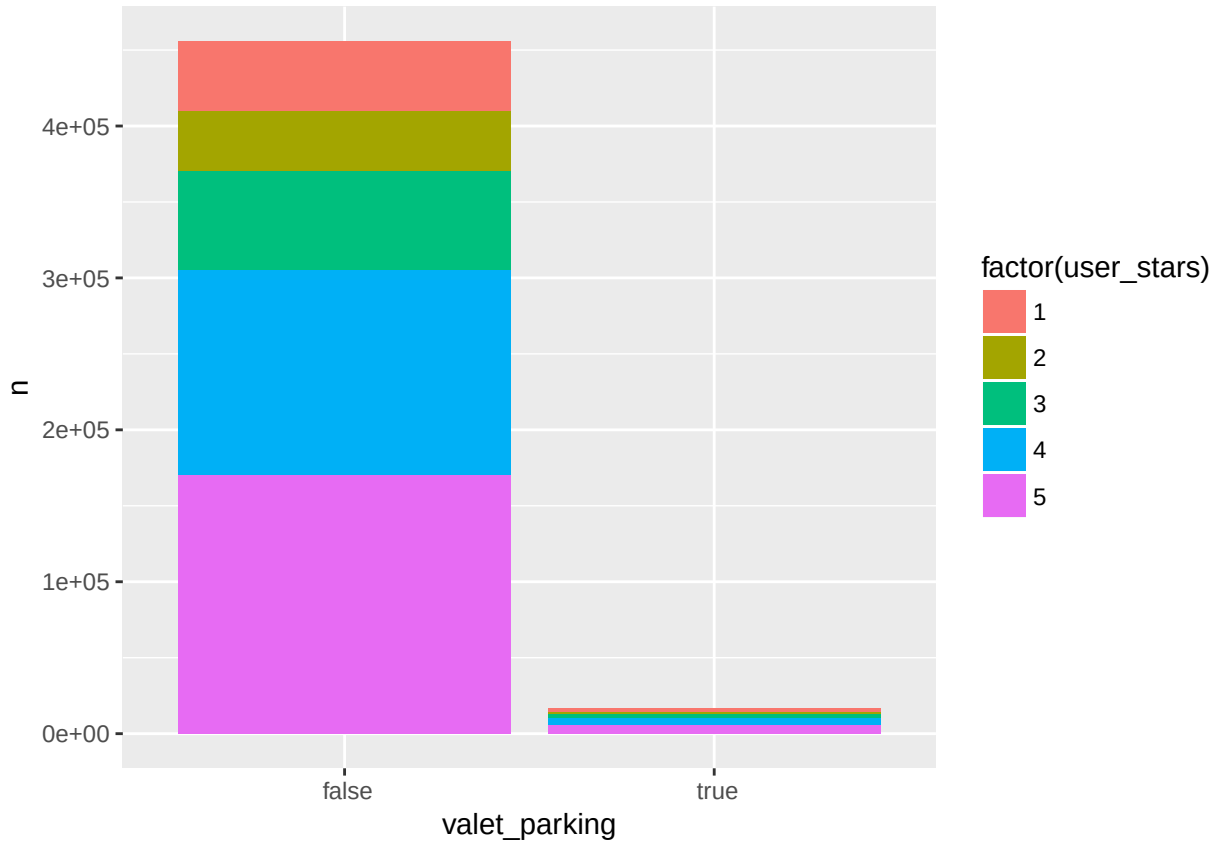

1.00
0.75
0.50
0.25
0.00
percent
false
true
street_parking
factor(user_stars)
1
2
3
4
5

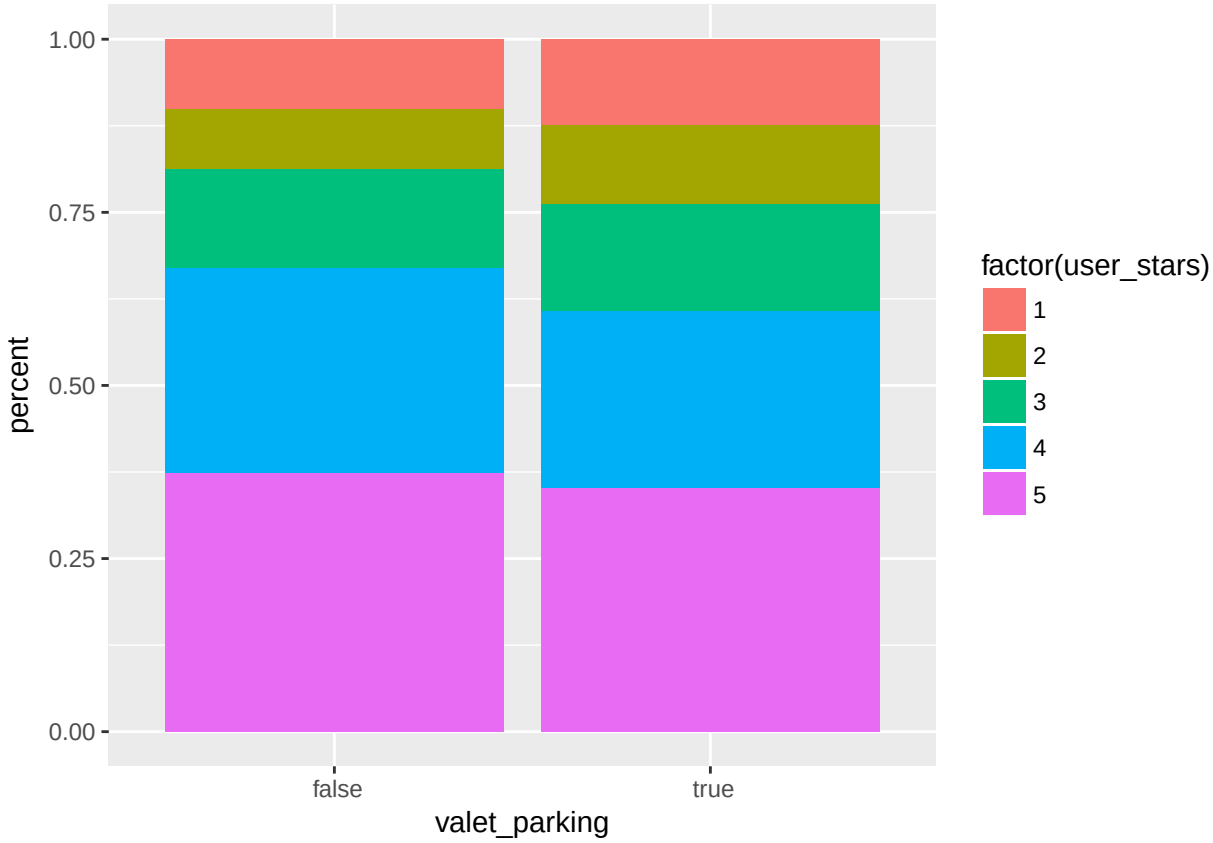

4e+05
3e+05
2e+05
1e+05
0e+00
n
false
true
validated_parking
factor(user_stars)
1
2
3
4
5

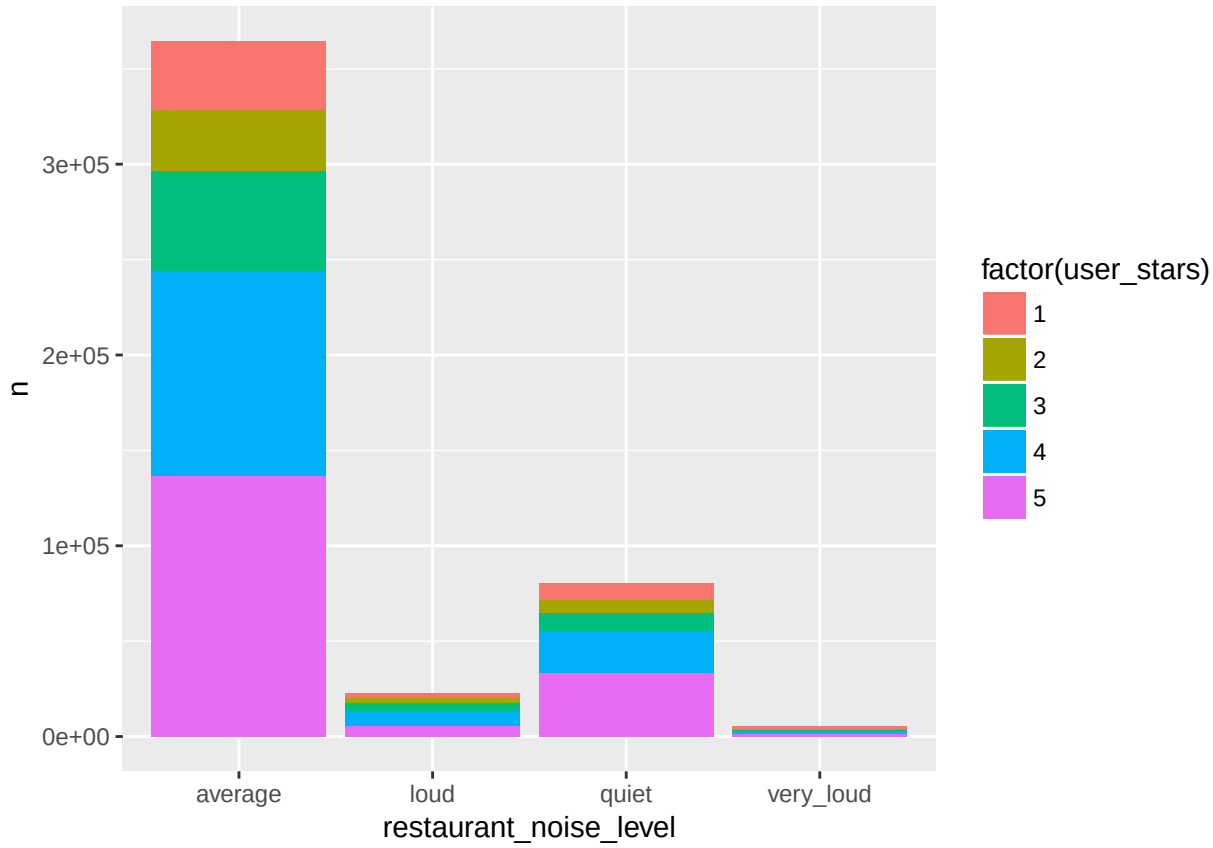

1.00
0.75
0.50
0.25
0.00
percent
false
true
validated_parking
factor(user_stars)
1
2
3
4
5

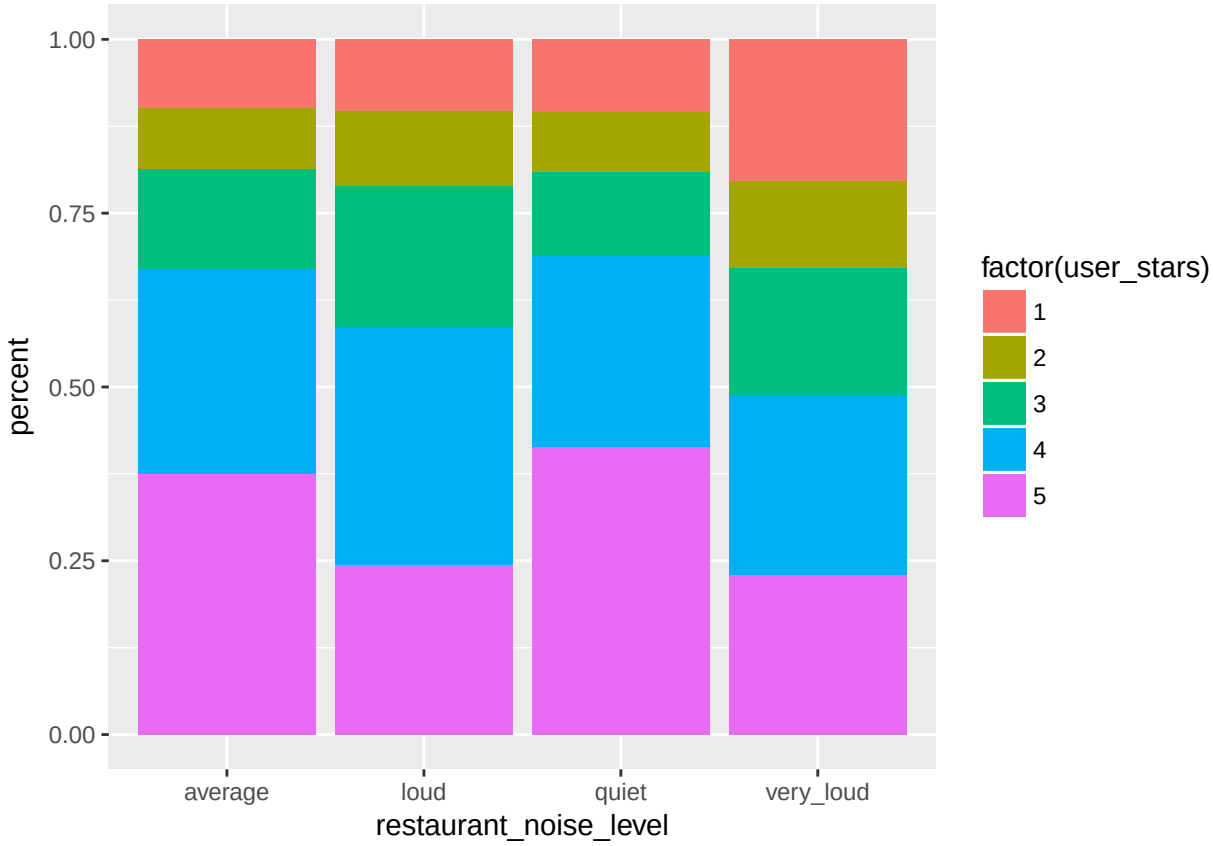

3e+05
2e+05
1e+05
0e+00
n
false
true
lot_parking
factor(user_stars)
1
2
3
4
5

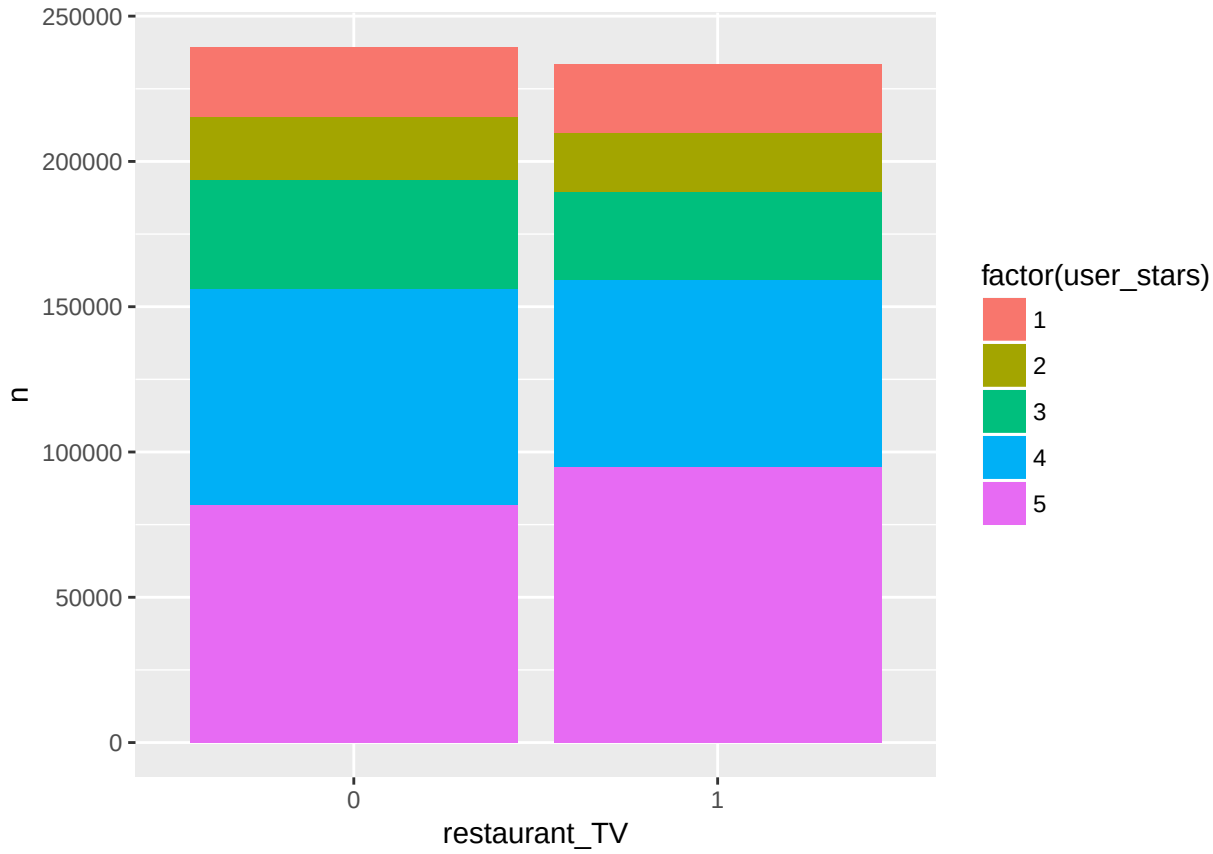

1.00
0.75
0.50
0.25
0.00
percent
false
true
lot_parking
factor(user_stars)
1
2
3
4
5

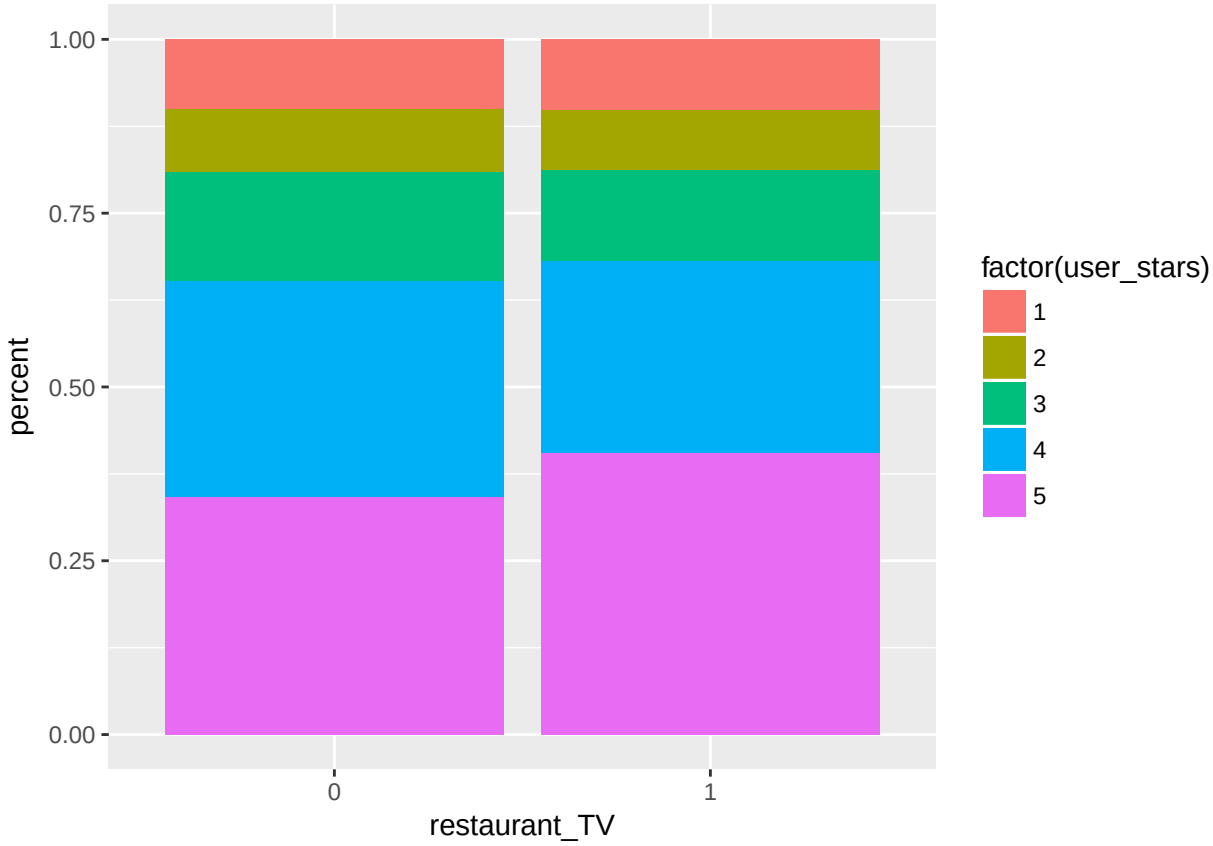

4e+05
3e+05
2e+05
1e+05
0e+00
n
false
true
valet_parking
factor(user_stars)
1
2
3
4
5

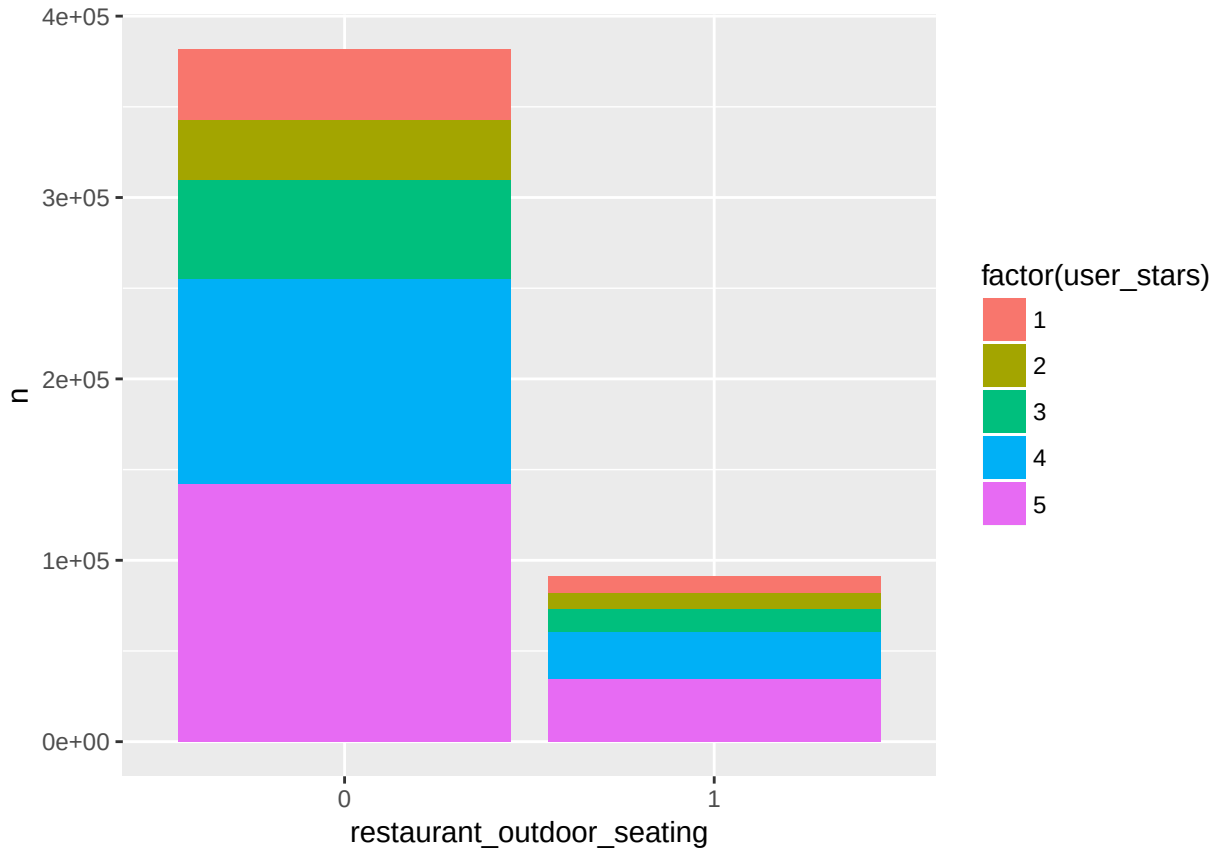

1.00
0.75
0.50
0.25
0.00
percent
false
true
valet_parking
factor(user_stars)
1
2
3
4
5

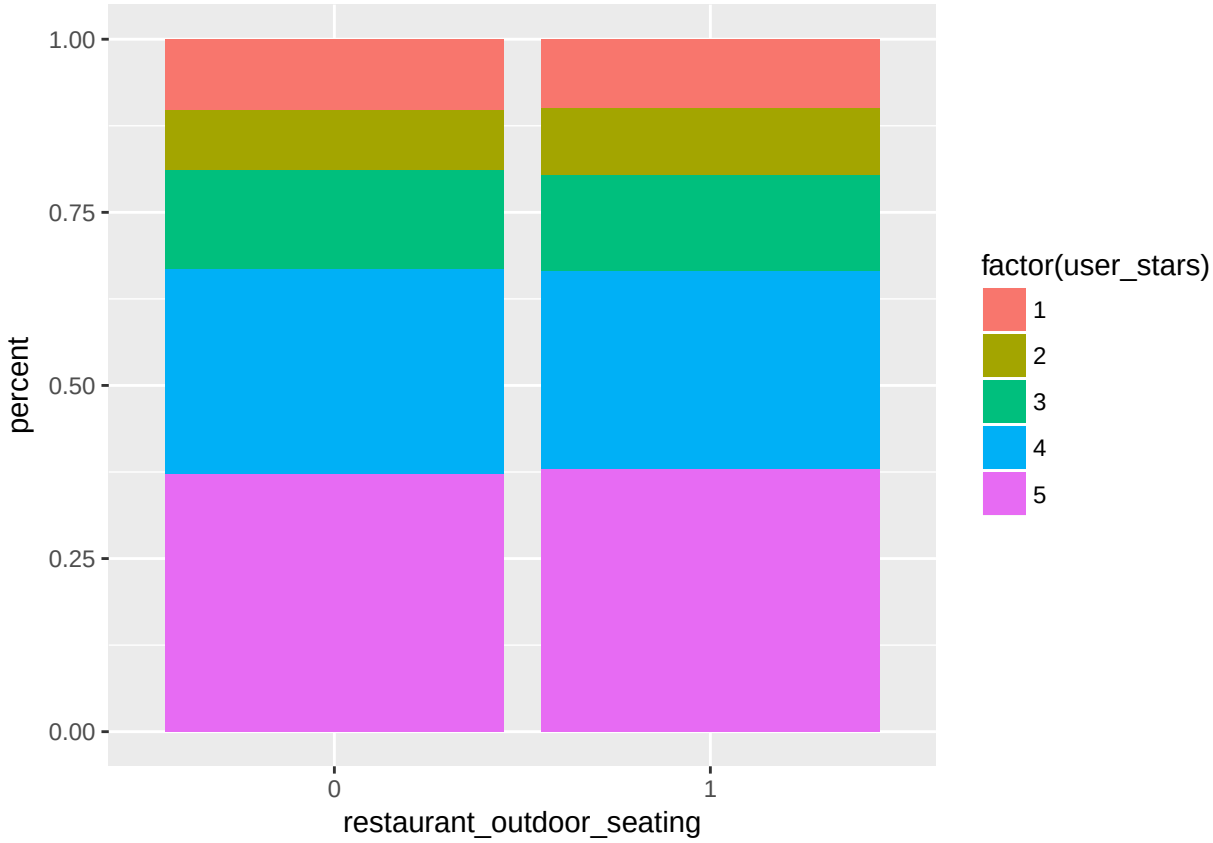

3e+05
2e+05
1e+05
0e+00
n
average
loud
quiet
very_loud
restaurant_noise_level
factor(user_stars)
1
2
3
4
5

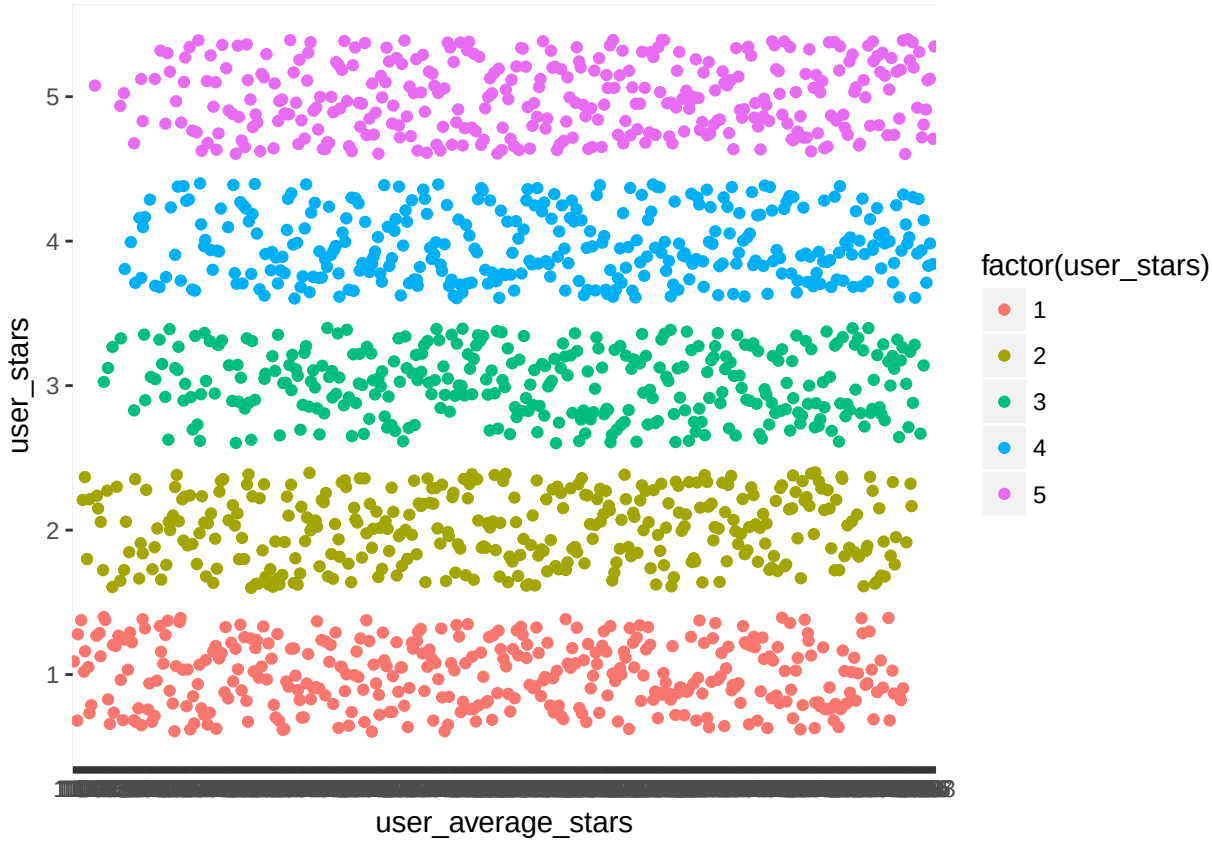

1.00
0.75
0.50
0.25
0.00
percent
average
loud
quiet
very_loud
restaurant_noise_level
factor(user_stars)
1
2
3
4
5

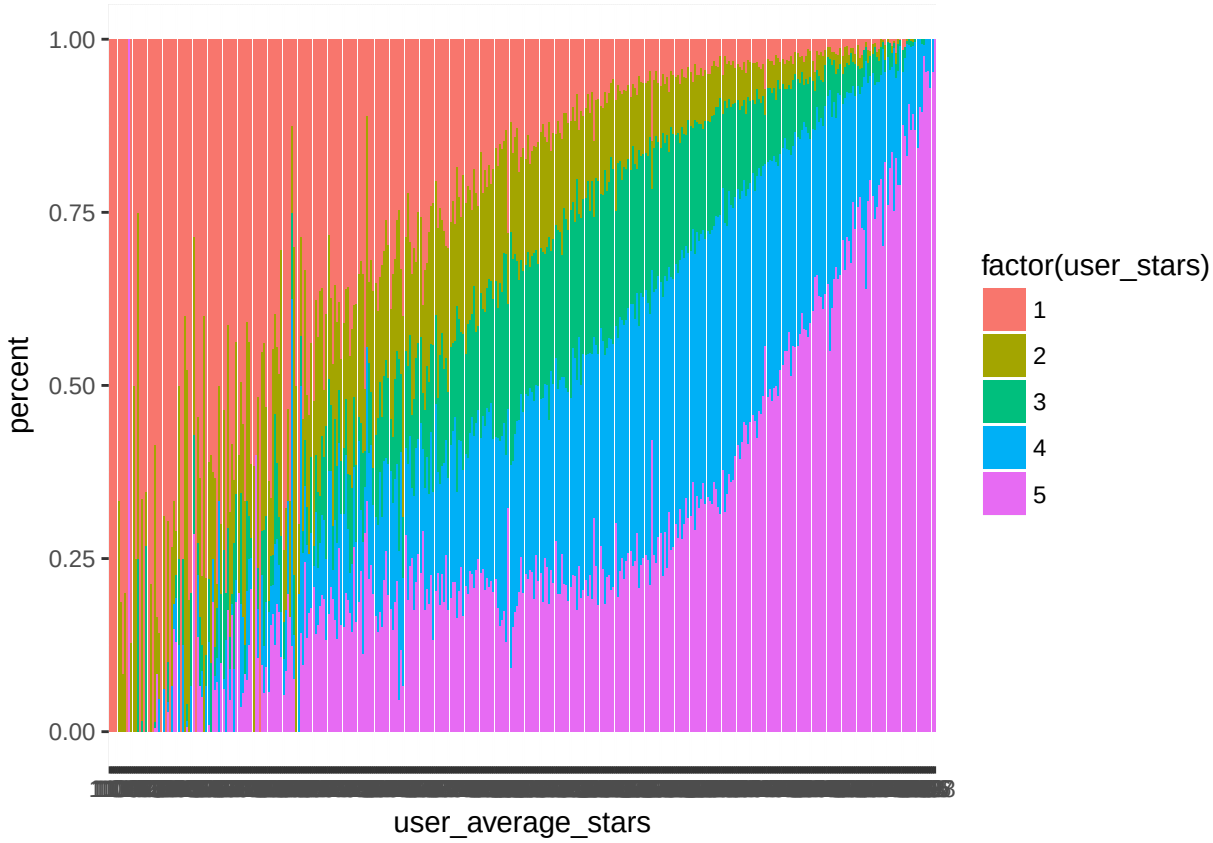

250000
200000
150000
100000
50000
0
n
0
1
restaurant_TV
factor(user_stars)
1
2
3
4
5


1.00
0.75
0.50
0.25
0.00
percent
0
1
restaurant_TV
factor(user_stars)
1
2
3
4
5


4e+05
3e+05
2e+05
1e+05
0e+00
n
0
1
restaurant_outdoor_seating
factor(user_stars)
1
2
3
4
5

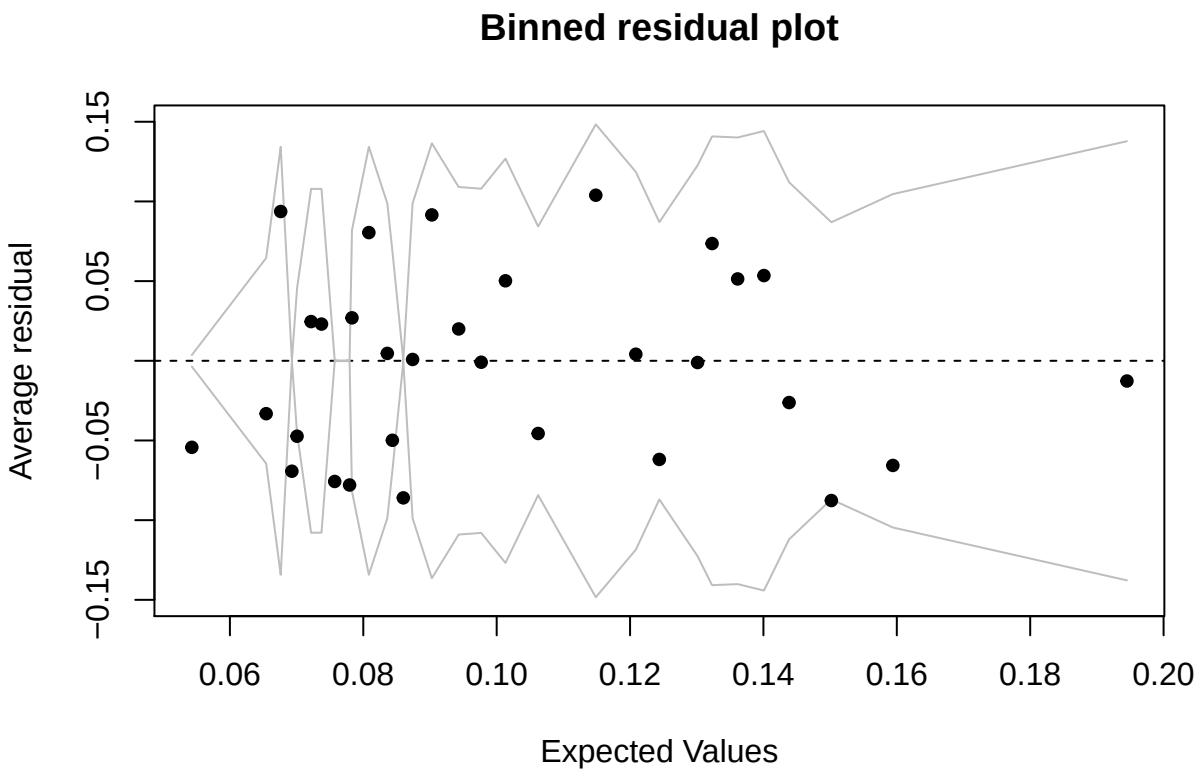

1.00
0.75
0.50
0.25
0.00
percent
0
1
restaurant_outdoor_seating
factor(user_stars)
1
2
3
4
5

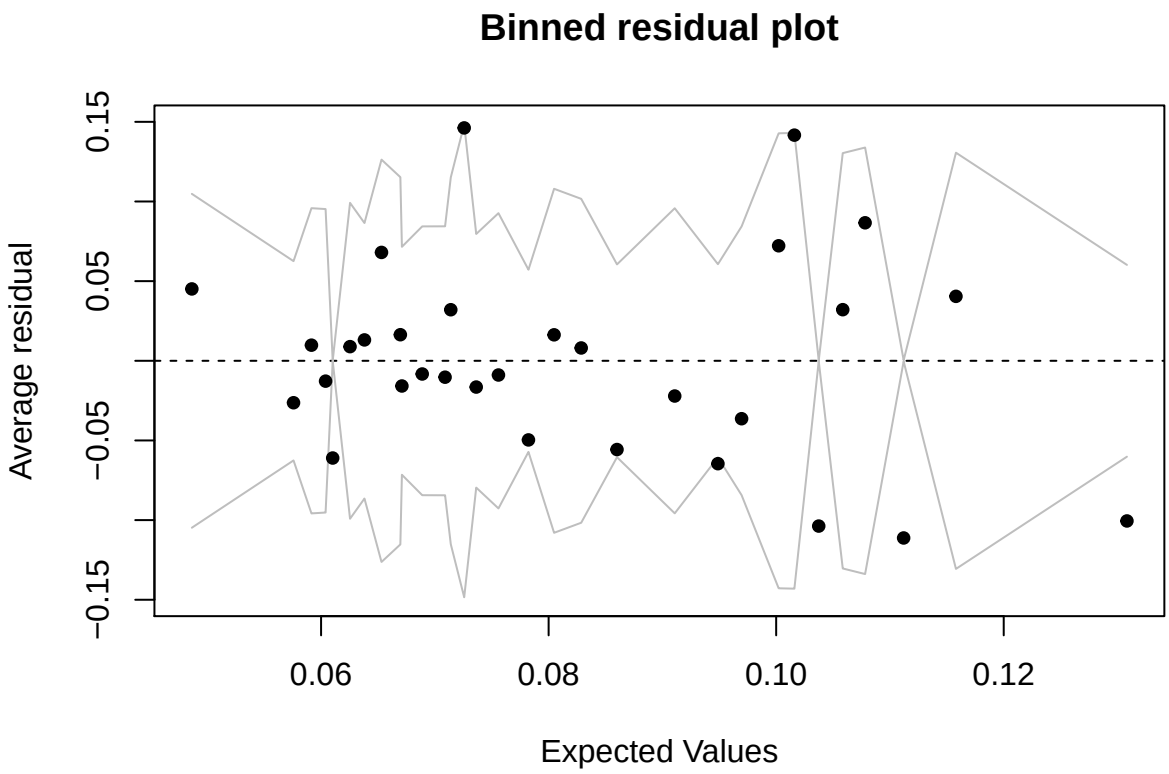

user_stars
5
4
3
2
1
user_average_stars
factor(user_stars)
1
2
3
4
5

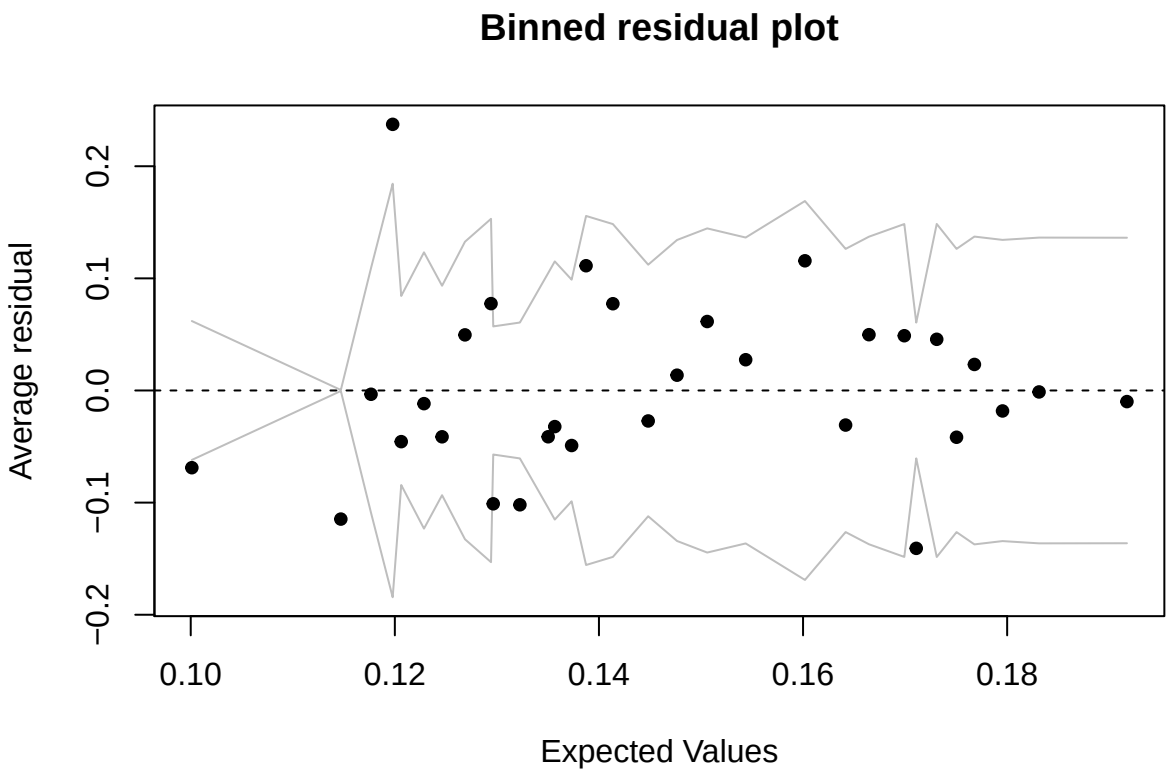


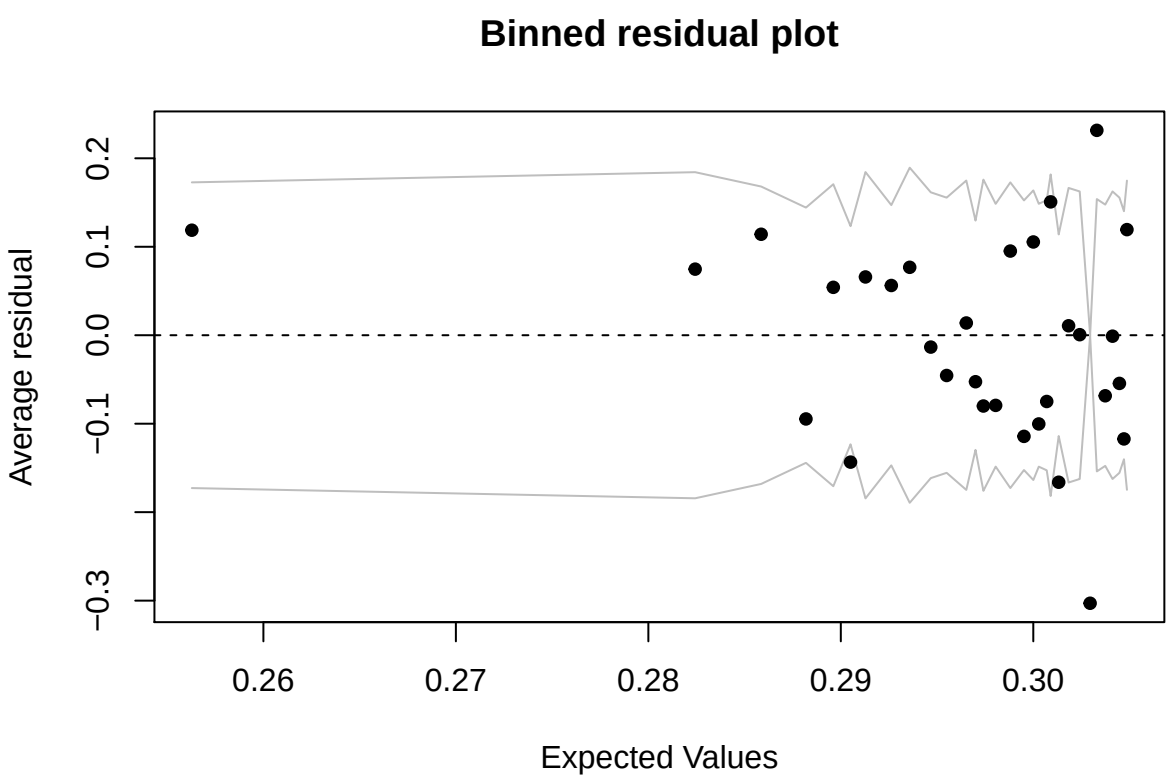

```
##
## Call:
## vglm(formula = ordered(user_stars) ~ restaurant_style + restaurant_WiFi +
##     restaurant_price_range + garage_parking + street_parking +
##     validated_parking + lot_parking + valet_parking + restaurant_noise_level +
##     restaurant_TV + restaurant_outdoor_seating, family = cumulative(parallel = T),
##     data = mysample)
##
##
## Pearson residuals:
##                     Min      1Q  Median      3Q   Max
## logit(P[Y<=1]) -1.972 -0.2224 -0.1632 -0.1326 3.534
## logit(P[Y<=2]) -1.294 -0.3237 -0.2170 -0.1706 4.724
## logit(P[Y<=3]) -1.373 -0.8520 -0.3104  0.4836 2.496
## logit(P[Y<=4]) -2.195 -1.0929  0.2863  1.0035 1.614
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept):1                    -1.69361    0.25479  -6.647 2.99e-11 ***
## (Intercept):2                    -0.99921    0.24710  -4.044 5.26e-05 ***
## (Intercept):3                    -0.19752    0.24414  -0.809  0.41849
## (Intercept):4                     1.06213    0.24651   4.309 1.64e-05 ***
## restaurant_styleJapanese         -0.47227    0.15330  -3.081  0.00207 **
## restaurant_styleKorean           -0.25571    0.23709  -1.079  0.28079
## restaurant_styleSoutheast Asian -0.70715    0.15655  -4.517 6.27e-06 ***
## restaurant_WiFino                -0.07427    0.13234  -0.561  0.57464
```

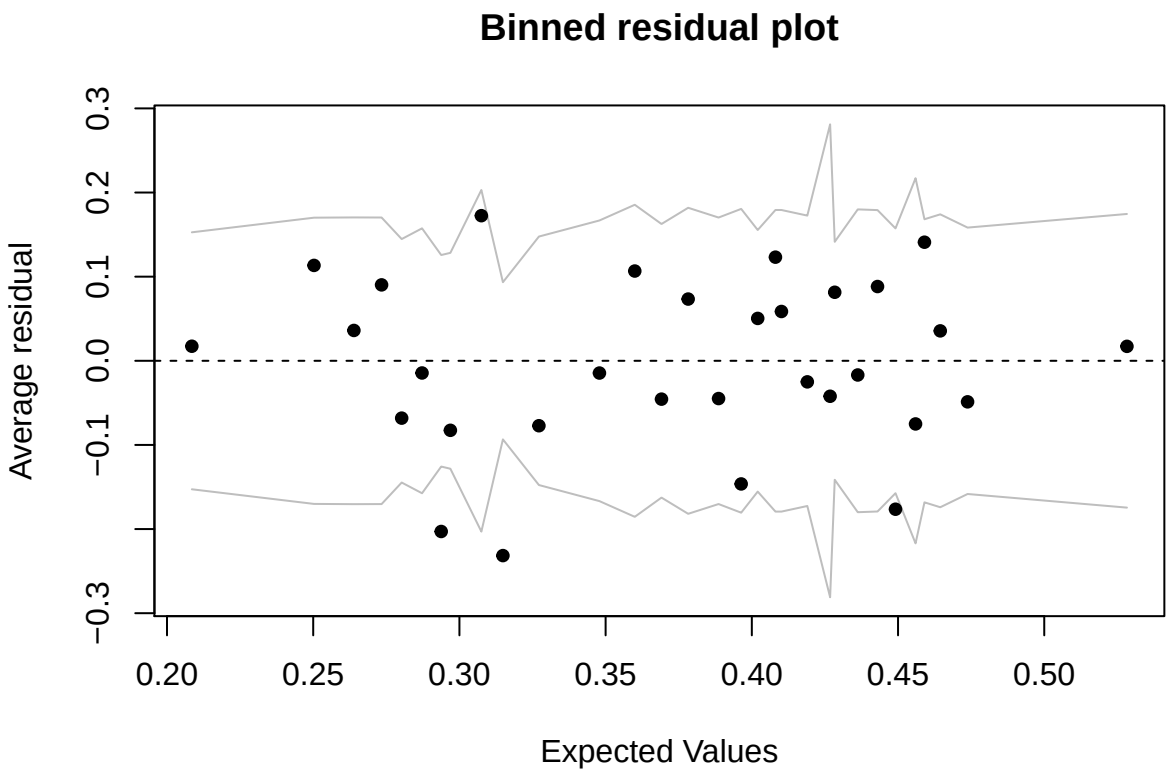
```
## restaurant_WiFipaid             -0.39088    0.95185  -0.411  0.68133
## restaurant_price_range2         -0.02863    0.13747  -0.208  0.83504
## restaurant_price_range3         -0.05292    0.35767  -0.148  0.88238
## restaurant_price_range4         -0.39202    0.75967  -0.516  0.60583
## garage_parking true              0.49583    0.28504   1.740  0.08194 .
## street_parking true              0.04260    0.18586   0.229  0.81871
## validated_parking true          -0.28863    0.72222  -0.400  0.68942
## lot_parking true                -0.08607    0.17520  -0.491  0.62321
## valet_parking true              -0.74023    0.39238  -1.887  0.05922 .
## restaurant_noise_levelloud       0.19020    0.26395   0.721  0.47115
## restaurant_noise_levelquiet      0.04670    0.16610   0.281  0.77859
## restaurant_noise_levelvery_loud  0.40010    0.54623   0.732  0.46387
## restaurant_TV1                  -0.11197    0.12475  -0.898  0.36943
## restaurant_outdoor_seating1      0.08134    0.15630   0.520  0.60278
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  4
##
## Names of linear predictors:
## logit(P[Y<=1]), logit(P[Y<=2]), logit(P[Y<=3]), logit(P[Y<=4])
##
## Residual deviance: 2859.178 on 3978 degrees of freedom
##
## Log-likelihood: -1429.589 on 3978 degrees of freedom
##
## Number of iterations: 6
##
## No Hauck-Donner effect found in any of the estimates
##
## Exponentiated coefficients:
##        restaurant_styleJapanese          restaurant_styleKorean
##                       0.6235872                       0.7743663
## restaurant_styleSoutheast Asian               restaurant_WiFino
##                       0.4930482                       0.9284180
##             restaurant_WiFipaid         restaurant_price_range2
##                       0.6764594                       0.9717792
##         restaurant_price_range3         restaurant_price_range4
##                       0.9484568                       0.6756908
##             garage_parking true             street_parking true
##                       1.6418670                       1.0435201
##          validated_parking true                lot_parking true
##                       0.7492867                       0.9175258
##              valet_parking true      restaurant_noise_levelloud
##                       0.4770058                       1.2094970
##     restaurant_noise_levelquiet restaurant_noise_levelvery_loud
##                       1.0478092                       1.4919794
##                   restaurant_TV1     restaurant_outdoor_seating1
##                       0.8940749                       1.0847404
```


Binned residual plot
Average residual
0.15
0.05
−0.05
−0.15
0.06
0.08
0.10
0.12
0.14
0.16
0.18
0.20
Expected Values


Binned residual plot
Average residual
0.15
0.05
−0.05
−0.15
0.06
0.08
0.10
0.12
Expected Values


Binned residual plot
Average residual
0.2
0.1
0.0
−0.1
−0.2
0.10
0.12
0.14
0.16
0.18
Expected Values

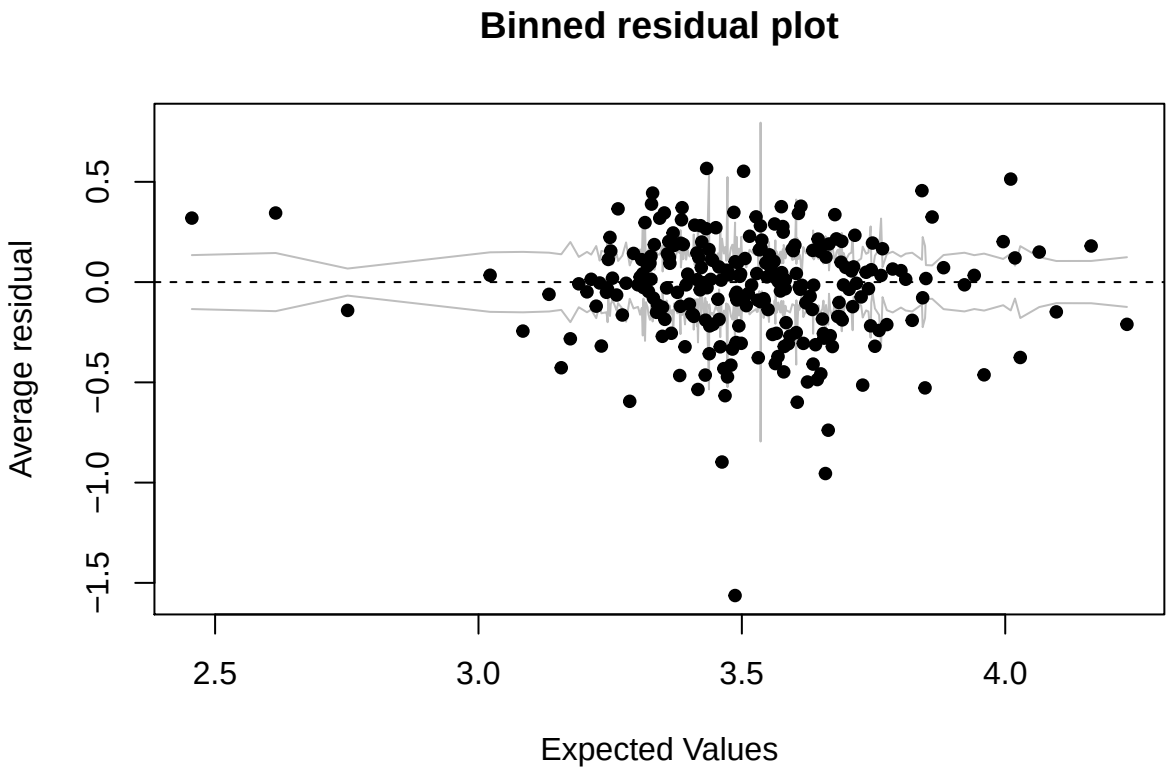

Binned residual plot
Average residual
0.2
0.1
0.0
−0.1
−0.3
0.26
0.27
0.28
0.29
0.30
Expected Values




```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## user_stars ~ restaurant_WiFi + restaurant_price_range + garage_parking +
##     street_parking + validated_parking + lot_parking + valet_parking +
##     restaurant_noise_level + restaurant_TV + restaurant_outdoor_seating +
##     (1 | restaurant_style) + (1 | restaurant_city:restaurant_state)
##    Data: mysample
##
## REML criterion at convergence: 3367.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.4308 -0.5698  0.2095  0.8388  1.4771
##
## Random effects:
##  Groups                           Name        Variance Std.Dev.
##  restaurant_city:restaurant_state (Intercept) 0.02125  0.1458
##  restaurant_style                 (Intercept) 0.04018  0.2004
##  Residual                                     1.65790  1.2876
## Number of obs: 1000, groups:
## restaurant_city:restaurant_state, 69; restaurant_style, 4
##
## Fixed effects:
##                               Estimate Std. Error t value
## (Intercept)                    3.57737    0.20070  17.825
## restaurant_WiFino              0.03174    0.09412   0.337
```

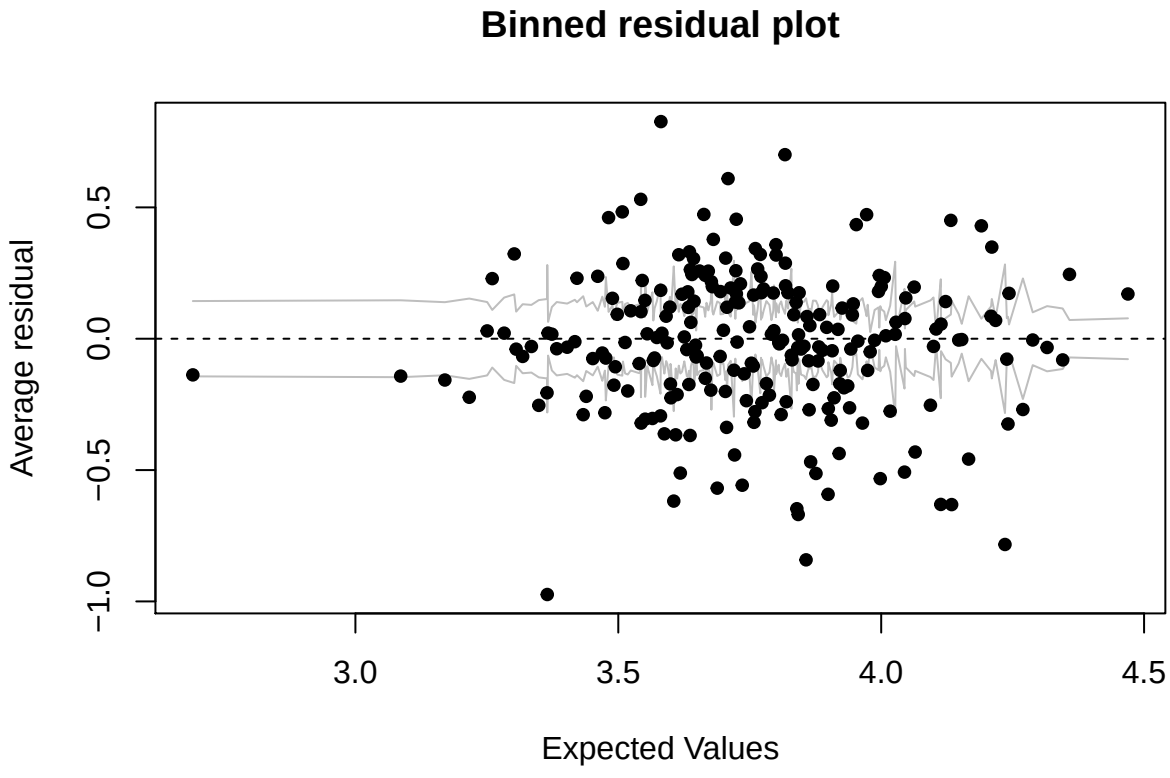
```
## restaurant_WiFipaid              0.55574    0.66437   0.836
## restaurant_price_range2         -0.01078    0.09760  -0.110
## restaurant_price_range3          0.17102    0.25480   0.671
## restaurant_price_range4          0.19852    0.52584   0.378
## garage_parking true             -0.48505    0.20850  -2.326
## street_parking true              0.11125    0.13729   0.810
## validated_parking true           0.22961    0.51047   0.450
## lot_parking true                 0.04649    0.12675   0.367
## valet_parking true               0.53025    0.27120   1.955
## restaurant_noise_levelloud      -0.04387    0.19186  -0.229
## restaurant_noise_levelquiet     -0.03405    0.11839  -0.288
## restaurant_noise_levelvery_loud -0.32985    0.39273  -0.840
## restaurant_TV1                   0.05675    0.08912   0.637
## restaurant_outdoor_seating1      0.02914    0.11375   0.256
##
## Correlation matrix not shown by default, as p = 16 > 12.
## Use print(x, correlation=TRUE)  or
##   vcov(x)     if you need it
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## user_stars ~ restaurant_WiFi + restaurant_price_range + garage_parking +
##     street_parking + validated_parking + lot_parking + valet_parking +
##     restaurant_noise_level + restaurant_TV + restaurant_outdoor_seating +
##     (1 | restaurant_city:restaurant_state)
##    Data: chinese
##
## REML criterion at convergence: 452290.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.4646 -0.5538  0.3092  0.9359  1.9417
##
## Random effects:
##  Groups                           Name        Variance Std.Dev.
##  restaurant_city:restaurant_state (Intercept) 0.1033   0.3214
##  Residual                                     1.8244   1.3507
## Number of obs: 131394, groups:  restaurant_city:restaurant_state, 190
##
## Fixed effects:
##                                 Estimate Std. Error t value
## (Intercept)                     3.460481   0.031447  110.04
## restaurant_WiFino              -0.074245   0.009681   -7.67
## restaurant_WiFipaid             0.391093   0.124722    3.14
## restaurant_price_range2        -0.076295   0.008556   -8.92
## restaurant_price_range3         0.241182   0.023064   10.46
## restaurant_price_range4         0.292546   0.056898    5.14
## garage_parking true             0.062425   0.016080    3.88
## street_parking true             0.236769   0.014132   16.75
## validated_parking true          0.318027   0.077457    4.11
## lot_parking true                0.169701   0.011287   15.03
## valet_parking true             -0.096321   0.029909   -3.22
## restaurant_noise_levelloud     -0.103661   0.016073   -6.45
## restaurant_noise_levelquiet     0.070449   0.010912    6.46
```

```
## restaurant_noise_levelvery_loud -0.838387   0.035536  -23.59
## restaurant_TV1                   0.081026   0.008341    9.71
## restaurant_outdoor_seating1     -0.116363   0.013031   -8.93
##
## Correlation matrix not shown by default, as p = 16 > 12.
## Use print(x, correlation=TRUE)  or
##   vcov(x)     if you need it
```



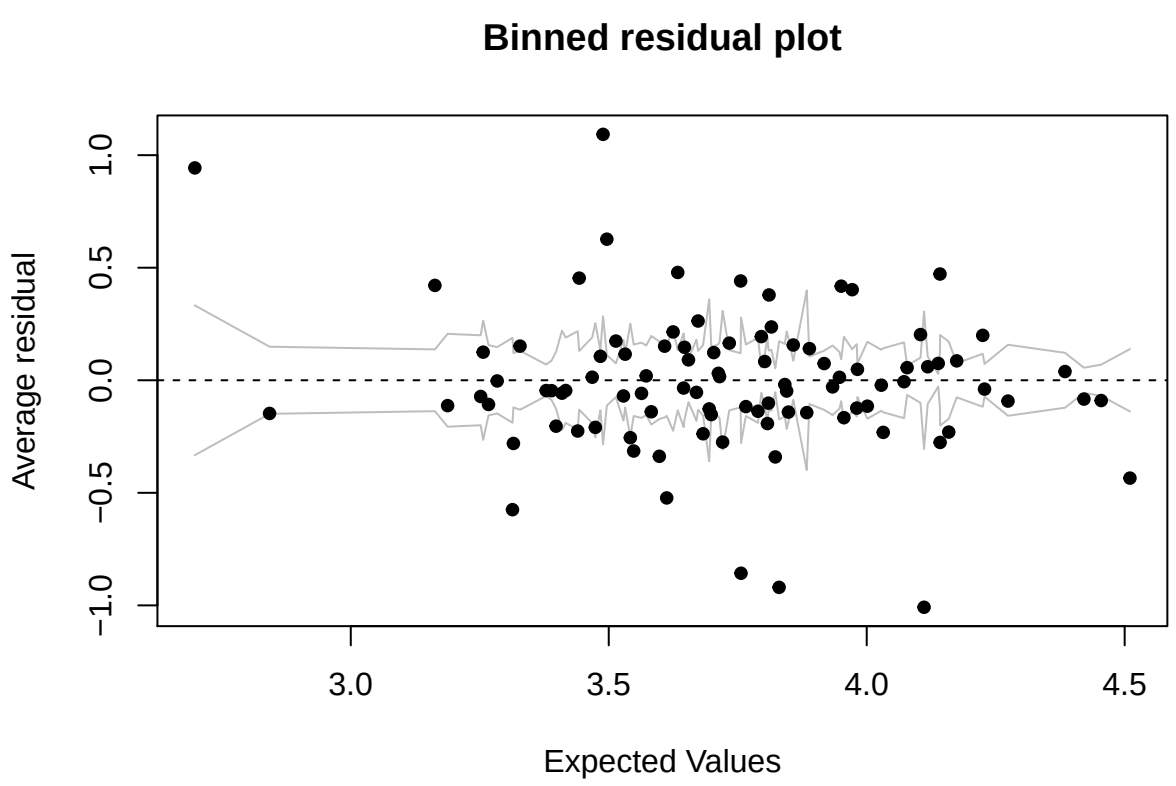

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## user_stars ~ restaurant_WiFi + restaurant_price_range + garage_parking +
##     street_parking + validated_parking + lot_parking + valet_parking +
##     restaurant_noise_level + restaurant_TV + restaurant_outdoor_seating +
##     (1 | restaurant_city:restaurant_state)
##    Data: japanese
##
## REML criterion at convergence: 498840.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.7861 -0.5712  0.2626  0.8048  2.0299
##
## Random effects:
##  Groups                           Name        Variance Std.Dev.
##  restaurant_city:restaurant_state (Intercept) 0.1465   0.3828
##  Residual                                     1.6089   1.2684
```

```
## Number of obs: 150429, groups:  restaurant_city:restaurant_state, 111
##
## Fixed effects:
##                                 Estimate Std. Error t value
## (Intercept)                     3.637550   0.041999   86.61
## restaurant_WiFino              -0.125252   0.007699  -16.27
## restaurant_WiFipaid            -0.325194   0.070338   -4.62
## restaurant_price_range2        -0.017203   0.010670   -1.61
## restaurant_price_range3         0.196647   0.017109   11.49
## restaurant_price_range4         0.124945   0.023910    5.23
## garage_parking true            -0.129792   0.013559   -9.57
## street_parking true             0.136821   0.012295   11.13
## validated_parking true          0.137396   0.042405    3.24
## lot_parking true                0.032307   0.010401    3.11
## valet_parking true             -0.238354   0.017193  -13.86
## restaurant_noise_levelloud     -0.081484   0.018166   -4.49
## restaurant_noise_levelquiet     0.182440   0.011740   15.54
## restaurant_noise_levelvery_loud 0.169690   0.028072    6.04
## restaurant_TV1                 -0.048817   0.008459   -5.77
## restaurant_outdoor_seating1    -0.057049   0.009343   -6.11
##
## Correlation matrix not shown by default, as p = 16 > 12.
## Use print(x, correlation=TRUE)  or
##   vcov(x)     if you need it
```




```
## Linear mixed model fit by REML ['lmerMod']
```

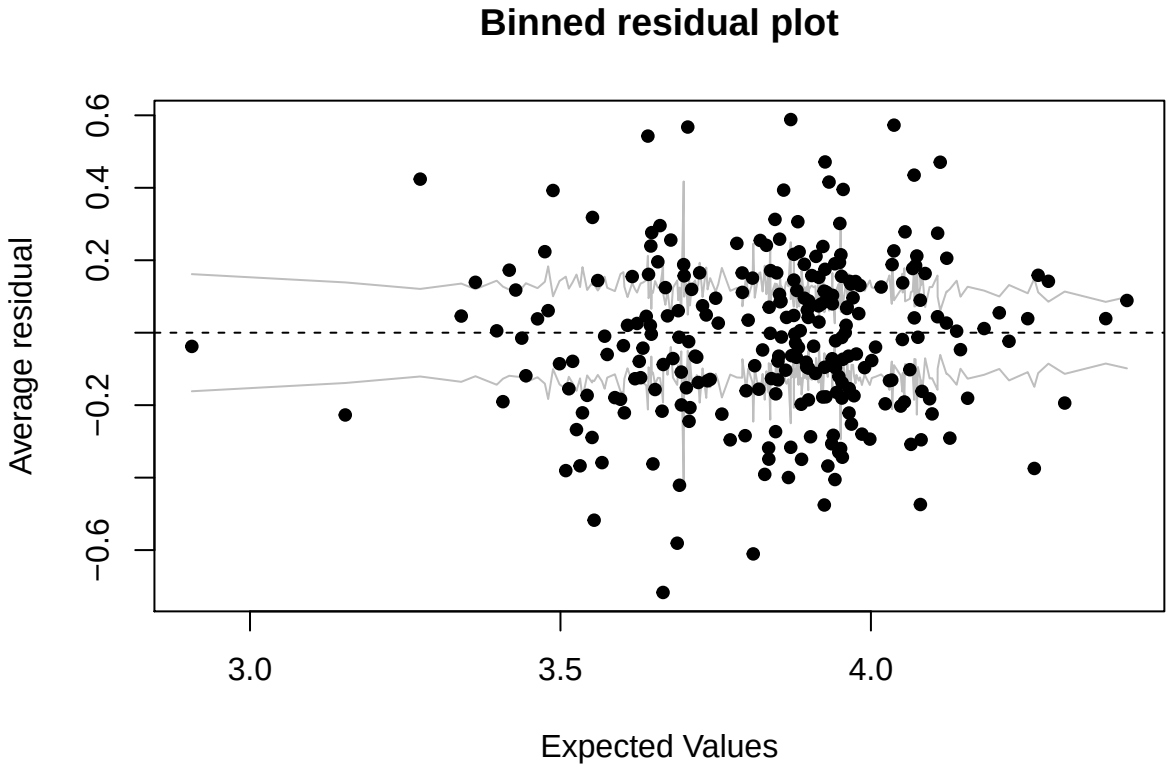
```
## Formula:
## user_stars ~ restaurant_WiFi + restaurant_price_range + garage_parking +
##     street_parking + validated_parking + lot_parking + valet_parking +
##     restaurant_noise_level + restaurant_TV + restaurant_outdoor_seating +
##     (1 | restaurant_city:restaurant_state)
##    Data: korean
##
## REML criterion at convergence: 124142.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0001 -0.4853  0.2527  0.7220  1.9299
##
## Random effects:
##  Groups                           Name        Variance Std.Dev.
##  restaurant_city:restaurant_state (Intercept) 0.1203   0.3469
##  Residual                                     1.4232   1.1930
## Number of obs: 38850, groups:  restaurant_city:restaurant_state, 41
##
## Fixed effects:
##                                 Estimate Std. Error t value
## (Intercept)                     4.363375   0.065494   66.62
## restaurant_WiFino              -0.315933   0.015895  -19.88
## restaurant_WiFipaid             0.017840   0.086046    0.21
## restaurant_price_range2        -0.281608   0.018035  -15.61
## restaurant_price_range3        -0.723019   0.091844   -7.87
## garage_parking true            -0.690390   0.086335   -8.00
## street_parking true            -0.142824   0.027089   -5.27
## validated_parking true         -0.330166   0.103130   -3.20
## lot_parking true               -0.003439   0.023524   -0.15
## valet_parking true              0.574760   0.075823    7.58
## restaurant_noise_levelloud     -0.134862   0.026442   -5.10
## restaurant_noise_levelquiet     0.022934   0.026623    0.86
## restaurant_noise_levelvery_loud -0.523394   0.168415   -3.11
## restaurant_TV1                 -0.124490   0.016837   -7.39
## restaurant_outdoor_seating1     0.034246   0.023349    1.47
##
## Correlation matrix not shown by default, as p = 15 > 12.
## Use print(x, correlation=TRUE)  or
##   vcov(x)     if you need it
```




```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## user_stars ~ restaurant_WiFi + restaurant_price_range + garage_parking +
##     street_parking + validated_parking + lot_parking + valet_parking +
##     restaurant_noise_level + restaurant_TV + restaurant_outdoor_seating +
##     (1 | restaurant_city:restaurant_state)
##    Data: se_asian
##
## REML criterion at convergence: 501276.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.7109 -0.5992  0.1649  0.8425  1.8204
##
## Random effects:
##  Groups                             Name        Variance Std.Dev.
##  restaurant_city:restaurant_state (Intercept) 0.09686  0.3112
##  Residual                                     1.57755  1.2560
## Number of obs: 152068, groups:  restaurant_city:restaurant_state, 140
##
## Fixed effects:
##                                Estimate Std. Error t value
## (Intercept)                    3.780289   0.032831  115.14
## restaurant_WiFino             -0.000177   0.007656   -0.02
## restaurant_WiFipaid           -0.084038   0.052226   -1.61
## restaurant_price_range2        0.016849   0.007487    2.25
```

```
## restaurant_price_range3          0.278611   0.053409    5.22
## restaurant_price_range4          0.579079   0.563010    1.03
## garage_parking true             -0.280772   0.020810  -13.49
## street_parking true              0.155595   0.011610   13.40
## validated_parking true           0.221032   0.063286    3.49
## lot_parking true                 0.106394   0.010640   10.00
## valet_parking true              -0.205557   0.030688   -6.70
## restaurant_noise_levelloud       0.014001   0.020605    0.68
## restaurant_noise_levelquiet      0.011927   0.008027    1.49
## restaurant_noise_levelvery_loud -0.739897   0.072382  -10.22
## restaurant_TV1                  -0.025635   0.007364   -3.48
## restaurant_outdoor_seating1     -0.003726   0.008817   -0.42

##
## Correlation matrix not shown by default, as p = 16 > 12.
## Use print(x, correlation=TRUE)  or
##   vcov(x)     if you need it
```



	Chinese Restaurant	Japanese Restaurant	Southeast Asian Restaurant	Korean Res
Intercept	3.4604814	3.6375503	3.7802886	4.
restaurant_WiFino	-0.0742453	-0.1252521	-0.0001770	-0.
restaurant_WiFipaid	0.3910925	-0.3251938	-0.0840379	0.
restaurant_price_range2	-0.0762948	-0.0172033	0.0168487	-0.
restaurant_price_range3	0.2411818	0.1966469	0.2786113	-0.
restaurant_price_range4	0.2925457	0.1249455	0.5790789	
garage_parking true	0.0624246	-0.1297917	-0.2807717	-0.

	Chinese Restaurant	Japanese Restaurant	Southeast Asian Restaurant	Korean Res
street_parking true	0.2367694	0.1368207	0.1555946	-0.
validated_parking true	0.3180269	0.1373965	0.2210321	-0.
lot_parking true	0.1697010	0.0323071	0.1063941	-0.
valet_parking true	-0.0963215	-0.2383541	-0.2055574	0.
restaurant_noise_levelloud	-0.1036612	-0.0814839	0.0140009	-0.
restaurant_noise_levelquiet	0.0704491	0.1824398	0.0119266	0.
restaurant_noise_levelvery_loud	-0.8383866	0.1696895	-0.7398966	-0.
restaurant_TV1	0.0810256	-0.0488169	-0.0256351	-0.
restaurant_outdoor_seating1	-0.1163634	-0.0570491	-0.0037261	0.

	city:state
Chinese	0.10330
Japanese	0.14650
Korean	0.12030
Southeast Asian	0.09686