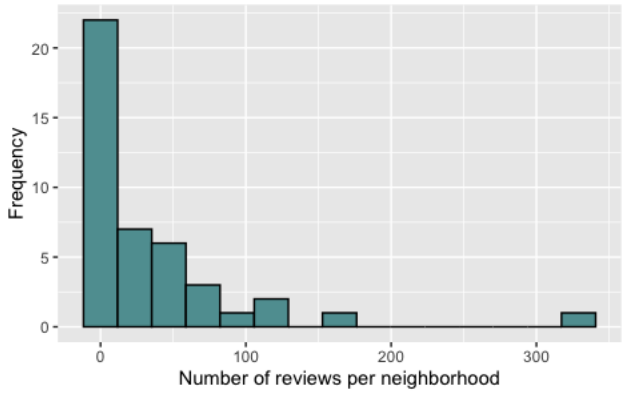
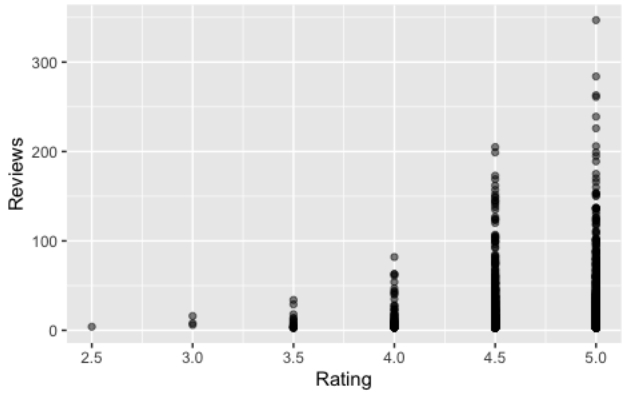
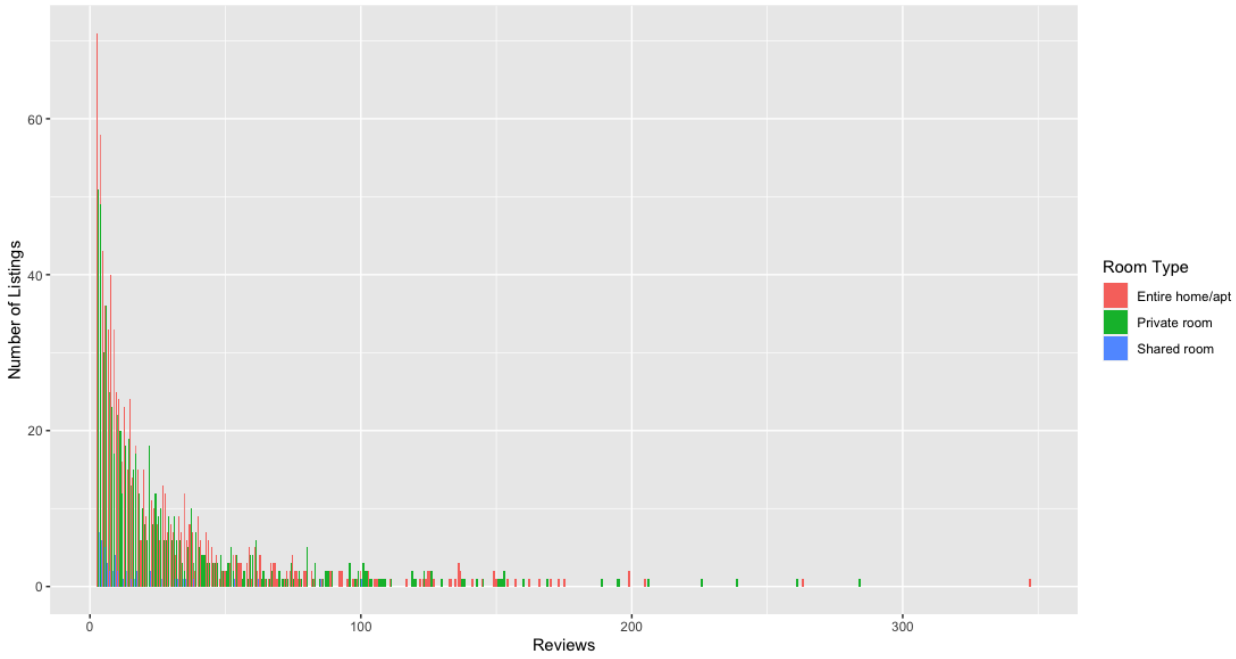
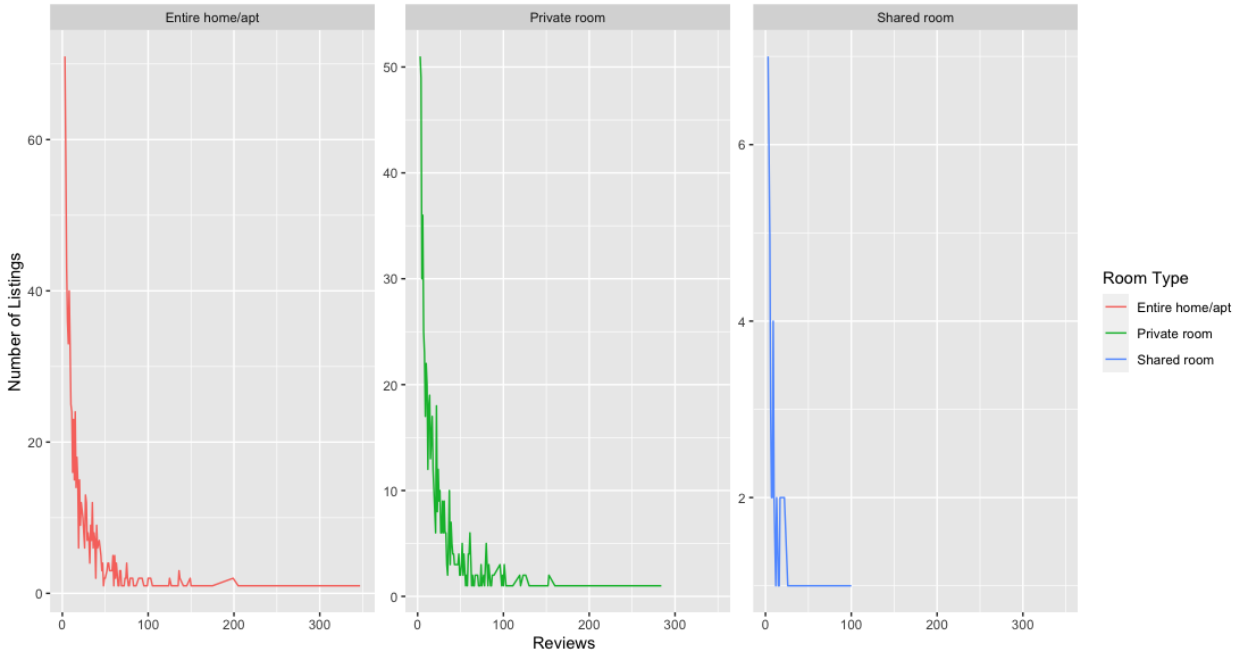
1a)

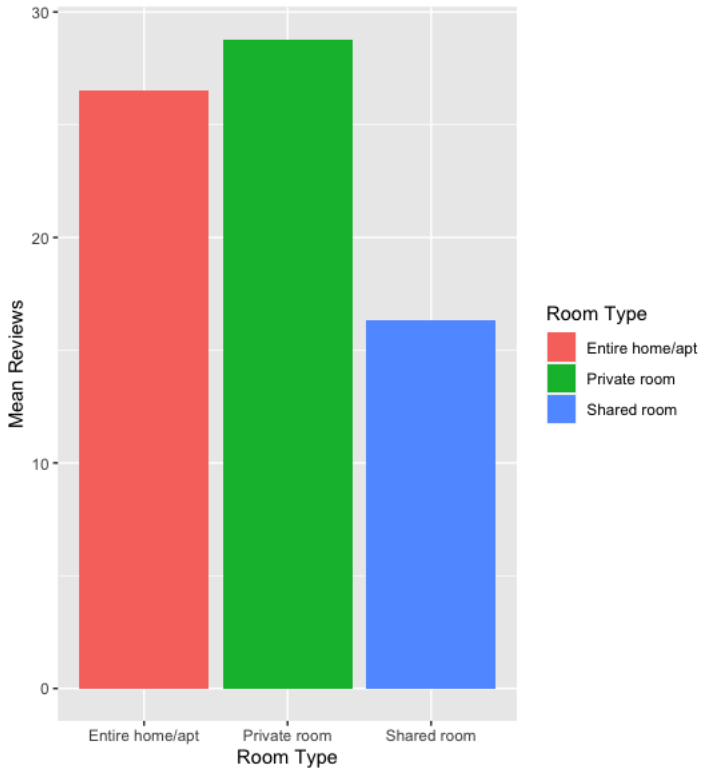


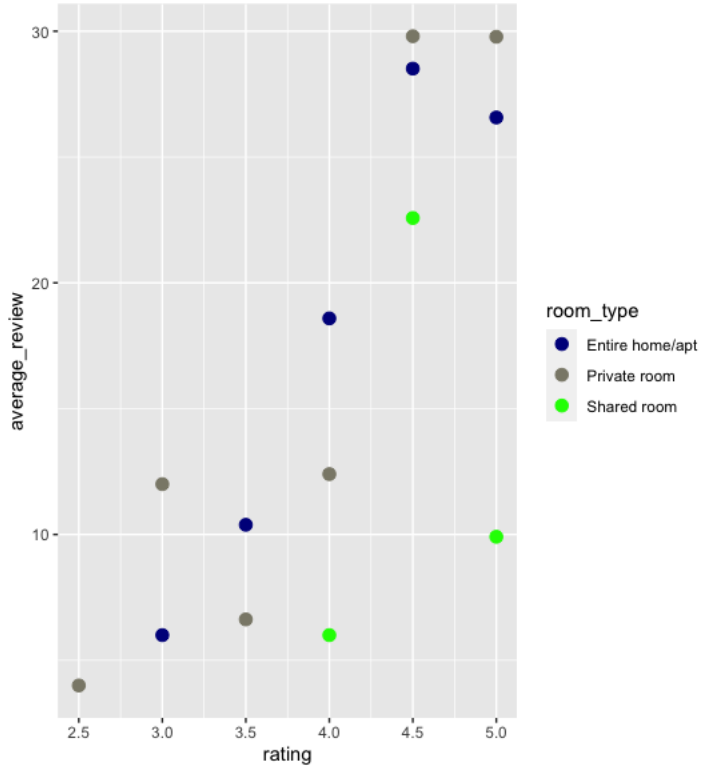


We can clearly see that as the rating increases the number of reviews increases


Number of Listings
60
40
20
0
0
100
200
300
Reviews
Room Type
Entire home/apt
Private room
Shared room


Entire home/apt
Private room
Shared room
Number of Listings
60
40
20
0
50
40
30
20
10
0
6
4
2
0
100
200
300
Reviews
Room Type
Entire home/apt
Private room
Shared room


30
20
10
0
Mean Reviews
Entire home/apt
Private room
Shared room
Room Type
Room Type
Entire home/apt
Private room
Shared room


30
20
10
average_review
2.5
3.0
3.5
4.0
4.5
5.0
rating
room_type
Entire home/apt
Private room
Shared room

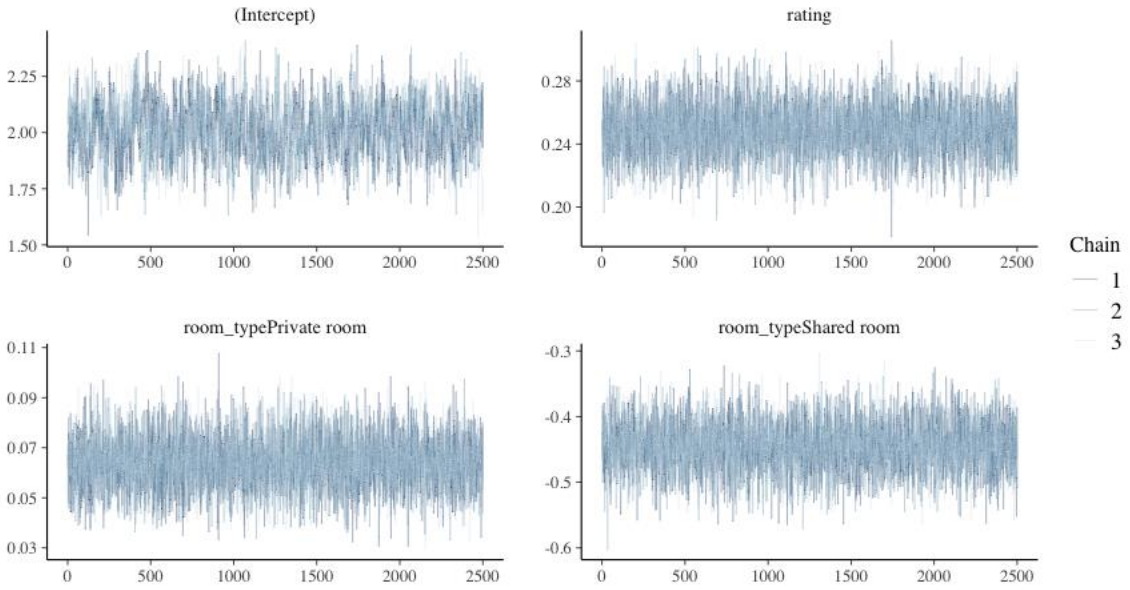
There is no clear trend visible btw the reviews and accommodation type. The number of data for Entire home/apt and Private Room is greater than Shared Room. As shown in previous plots, there is clear relationship btw number of reviews and ratings.

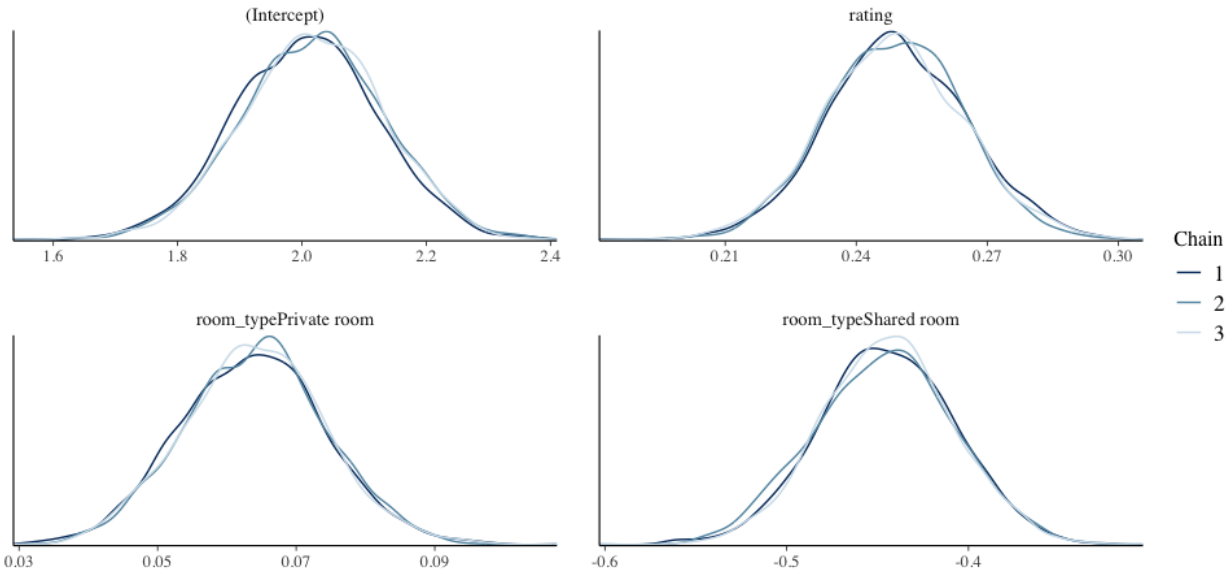
b) # Fit the model using STAN glmer

```
airbnb_model = stan_glmer(reviews ~ rating + room_type + (1|neighborhood),
                          data = AirBnB_Data, family = poisson(),
                          prior_intercept = normal(0, 2.5, autoscale = TRUE),
                          prior = normal(0, 2.5, autoscale = TRUE),
                          prior_covariance = decov(reg = 1, conc = 1, shape = 1, scale = 1),
                          chains = 3, iter = 1000*5)
```

The prior is setup as relatively broad for our intercept and regression coefficient indicating that we have weak belief. We used decovariance prior to specify prior covariance. It has four hyperparameters reg, conc, scale and shape. It controls scale shape and degree of shrinkage of the covariance matrix. The prior covariance is relatively diffused indicating weak prior belief on the variability of the random effects across neighborhood.
The autoscale = TRUE takes care of returning a better prior by looking at the data.

c)





term	estimate	std.error	conf.low	conf.high
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1 (Intercept)	2.02	0.115	1.83	2.20
2 rating	0.249	0.0160	0.223	0.276
3 room_typePrivate room	0.0641	0.0101	0.0470	0.0811
4 room_typeShared room	-0.445	0.0373	-0.508	-0.384

The expected number of reviews when the rating is zero and the room type is Entire home/apt is 2.015 and the 90% interval is (1.82, 2.204)

The expected increase in reviews when the rating increases by 1 is 0.2489 when the room type is Entire home/apt keeping everything else constant. The 90 % interval is (0.2263, 0.2755).

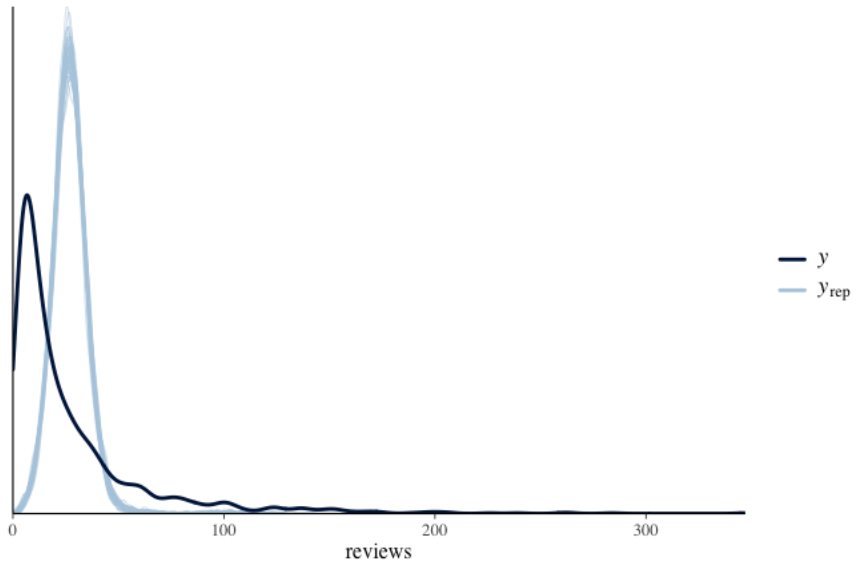
The expected increase in reviews when the rating is zero and room type is Private Room compared to Entire Home/apt is 0.06397 and the 90% interval is (0.0469, 0.081)

The expected decrease in reviews when the rating is zero and room type is Shared Room compared to Entire Home/apt is 0.4453 (since decrease) and the 90% interval is (-0.5076, -0.384)

Sigma[neighborhood:(Intercept),(Intercept)]
Min. :0.1321
1st Qu.:0.2829
Median :0.3321
Mean :0.3438
3rd Qu.:0.3927
Max. :0.9147

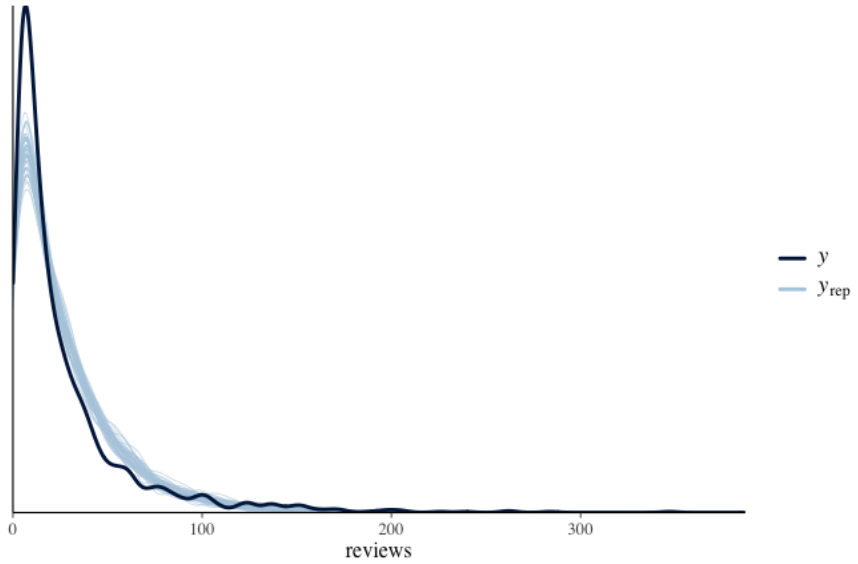
The expected standard deviation of the random effects for the intercepts across neighborhoods is 0.3438 keeping everything else constant. The value represent heterogeneity across various neighborhood after accounting for rating and room type. The value is fairly moderate so we expect some heterogeneity in number of reviews based on neighbor. The 90% interval is (0.2259, 0.501)

d)



The model does not seem to approximate the true distribution of the data.

e)



The assumption that the reviews is negative binomial is able to approximate the true distribution of data very accurately.

The poisson distribution assumes that the mean and variance is same. However, that might not be the case always for count data. A negative binomial accounts for over dispersion. So, when we have count data with over dispersion, we use negative binomial. Also, we see that across neighbor standard deviation is significant, so we can clearly say there is over dispersion. Moreover, the mean of reviews is 27.2 and variance is 1220.9. So, negative binomial is a good fit.

f)

term	estimate	std.error	conf.low	conf.high
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1 (Intercept)	1.99	0.394	1.34	2.68
2 rating	0.265	0.0830	0.122	0.402
3 room_typePrivate room	0.0682	0.0527	-0.0169	0.156
4 room_typeShared room	-0.471	0.152	-0.711	-0.209

The expected number of reviews when the rating is zero and the room type is Entire home/apt is 1.99 and the 90% interval is (1.34, 2.68)

The expected increase in reviews when the rating increases by 1 is 0.265 when the room type is Entire home/apt keeping everything else constant. The 90 % interval is (0.122, 0.402).

The expected increase in reviews when the rating is zero and room type is Private Room compared to Entire Home/apt is 0.0682 and the 90% interval is (-0.0169, 0.156). Since the 90 % interval contains zero it is not very significant.

The expected decrease in reviews when the rating is zero and room type is Shared Room compared to Entire Home/apt is 0.471 (since decrease) and the 90% interval is (-0.711, -0.209).

Sigma[neighborhood:(Intercept),(Intercept)]
Min. :6.100e-07
1st Qu.:1.464e-02
Median :2.780e-02
Mean :3.405e-02
3rd Qu.:4.571e-02
Max. :2.777e-01

The expected standard deviation of the random effects for the intercepts across neighborhoods is 0.03405 keeping everything else constant. The value represent heterogeneity across various neighborhood after accounting for rating and room type. The value is low, so we expect homogenity in number of reviews based on neighbor. The 90% interval is (0.0032, 0.875).

g)

Min. : 0.000
1st Qu.: 2.000
Median : 5.000
Mean : 8.627
3rd Qu.: 11.000
Max. :149.000

If the rating is zero, the expected number of reviews for a new listing of Private Room in Avondale will be 8.627. The 90% interval is (0,28).