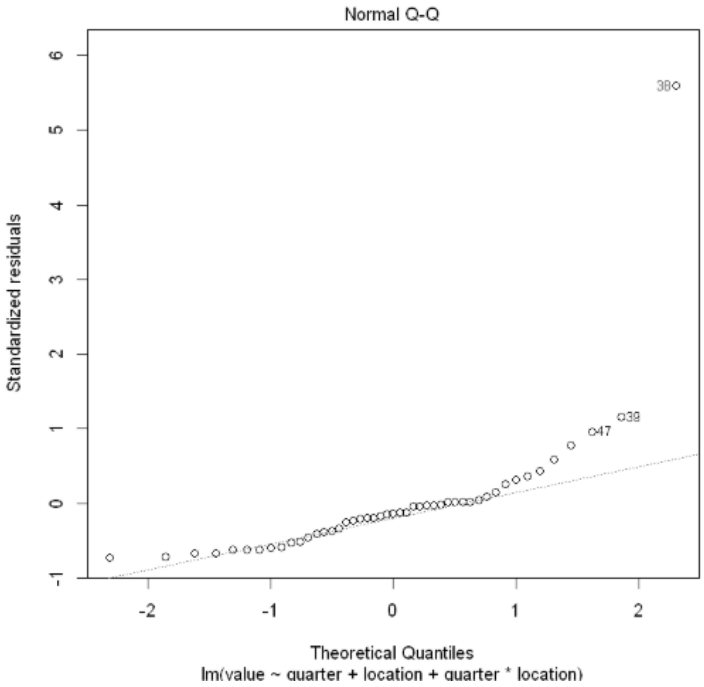
Hursh Desai

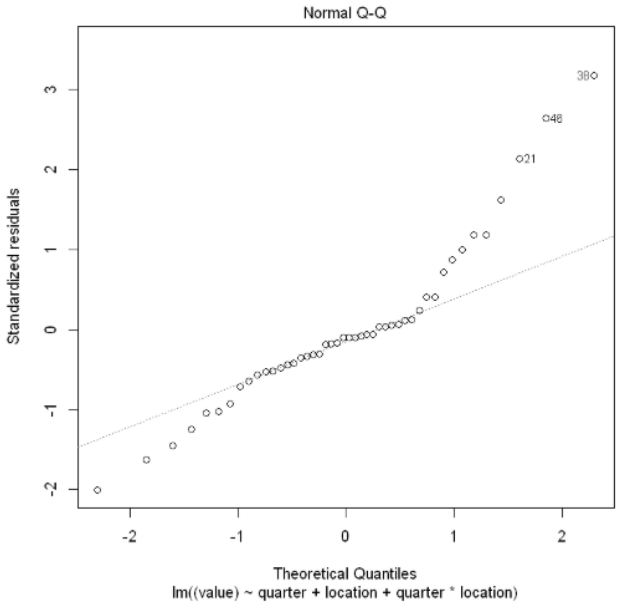
I was curious whether there was a relationship between the time of year or location of a company's headquarters and the amount of money raised in that company's IPO. I found data for all of the IPO's that took place domestically in 2018. There were 48 companies in total. The categories I chose to show location were picked to try and get as close to an even number of companies across all levels for both categorical variables. Any company that was headquartered in and around the bay area was grouped into the Bay Area. There were many companies from various places around China however making individual categories for each city clearly wouldn't have yielded enough categories so I grouped all those cities into simply China. Other consists of any of company headquartered outside of those two areas. That includes multiple IPO's for companies outside of the Bay Area but still in the US as well as some companies such as Spotify that at headquartered in Sweden. The time of the year is clearly segmented into the 4 quarters in a year each quarter being 3 months long. This is table for the observations. Surprisingly, there were a lot of companies in those other areas that IPO in the second quarter of the year.

	location		
quarter	Bay_Area	China	Other
1st Quarter	2	4	4
2nd Quarter	3	2	11
3rd Quarter	4	5	5
4th Quarter	3	4	1

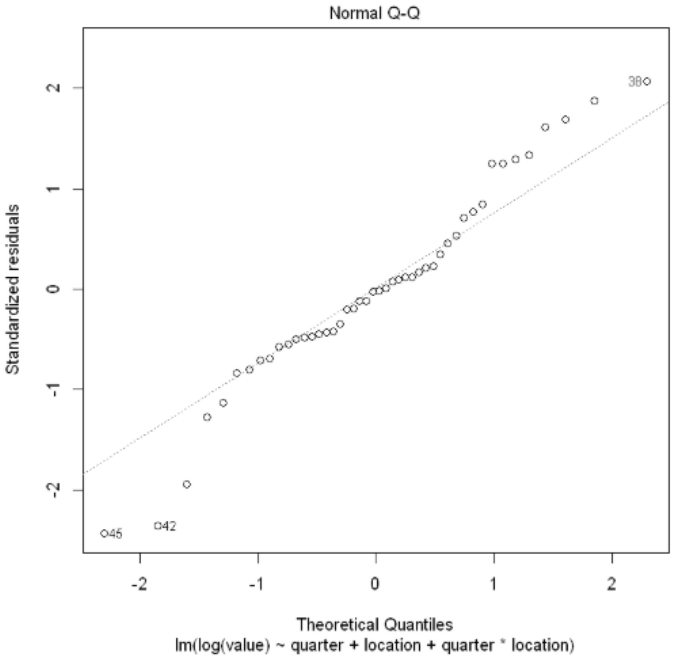
Other does seem to have

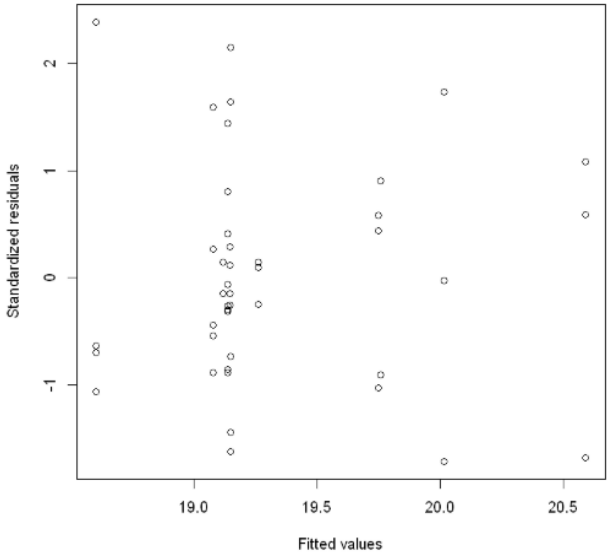


Here we can clearly see that data point 38 which is spotify is an outlier. This is most probably because it was put in the Other category for location because it is from Sweden however, its IPO was the most anticipated of the year and was raised an enormous amount of money. So I will remove this data point seeing as it is also from the 2nd quarter of other which has the most amount of data it shouldn't ruin the data set that much.



Clearly even after removing it there is a very long right tail which makes sense because the value of the company does have to do with money which usually causes the long right tail. A solution for this would be to take logs.





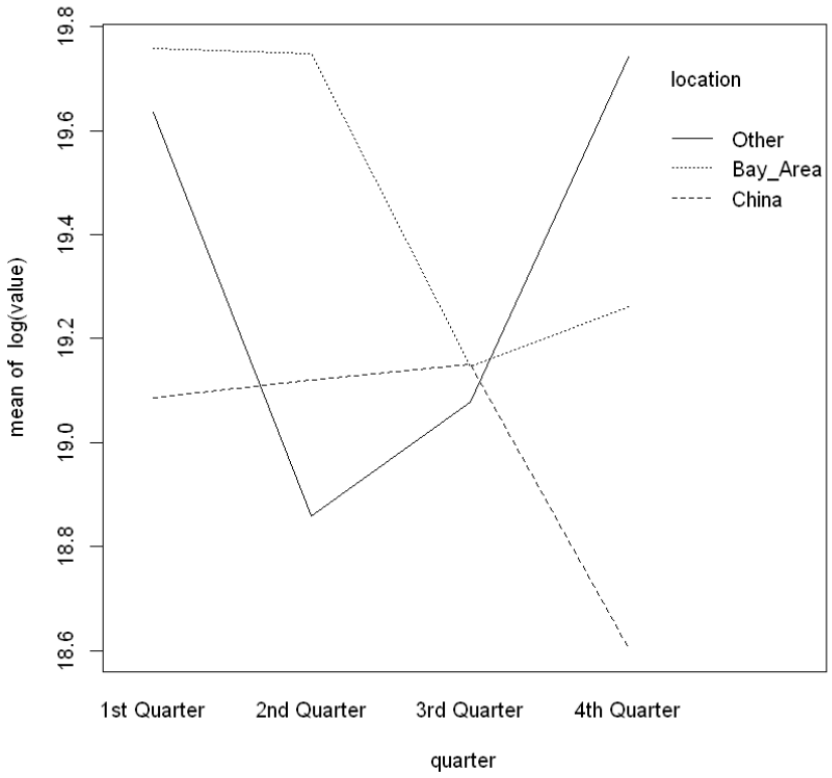
Taking logs of the money raised does help with the long right tail a but however it still looks like the data is not completely normal and there is still a long left tail for some reason.

I wanted to take care of the huge outlier that was Spotify first just in case it ruined the analysis of deciding whether the interaction effect between Location and Quarter made a significant difference if added into the model.

A anova: 4 × 4

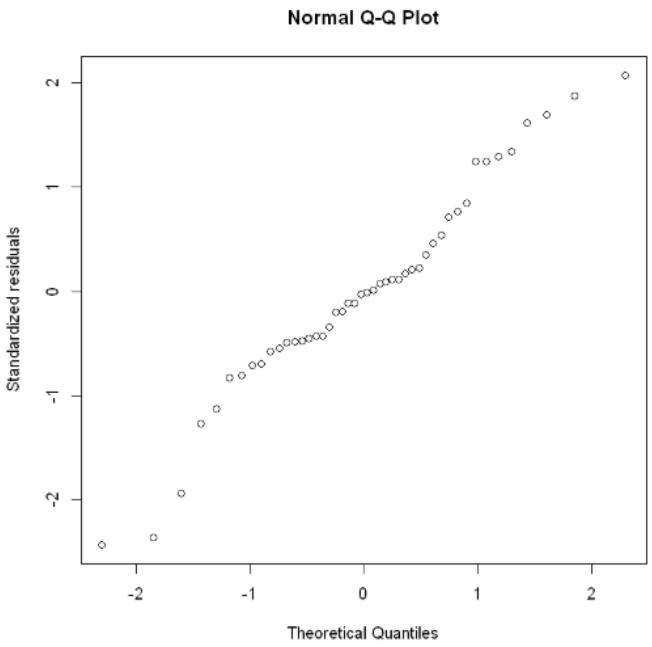
	Sum Sq	Df	F value	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>
quarter	1.379109	3	0.2472993	0.8626927
location	1.603535	2	0.4313144	0.6530609
quarter:location	2.481312	6	0.2224721	0.9668675
Residuals	65.061291	35	NA	NA

Looking at the F-tests we can see that the p-value is very high meaning that there is no reason to include the interaction effect since it is not adding anything to the model. But, it also looks like quarter and location also does not add anything to the model. It might be that there is no relationship between the quarter a company IPO's in and what location their headquarters are in with the logged values of the amount of money they raised.

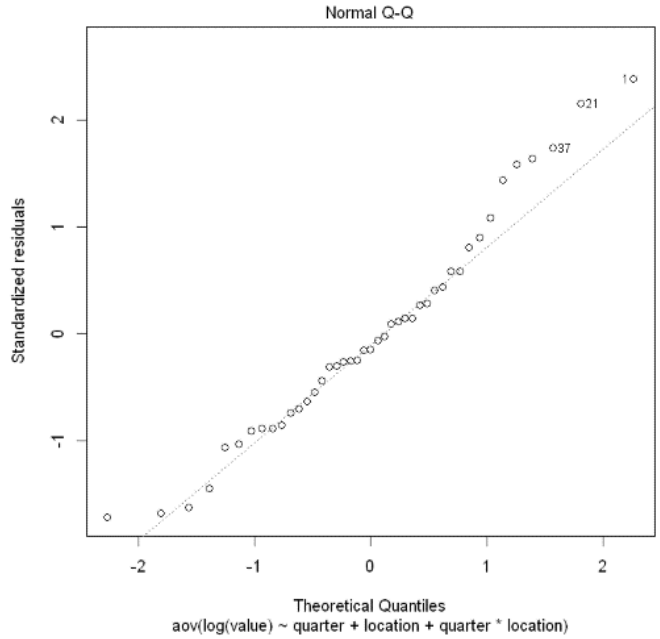


The interaction that are shown from this interaction plot might just be because of random noise and not because of an actual association between the quarter with location and IPO value.

We should however, first check for non-constant variance between our groups.



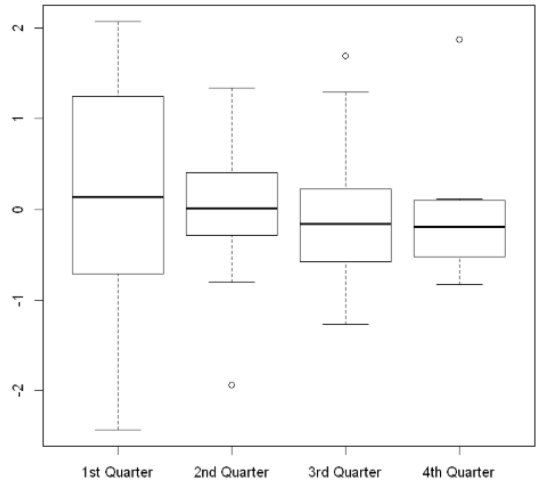
After we took the Spotify outlier out and logged the IPO value the Normal Q-Q plot still seems to have a slight long left and right tail. So we can take out the other three outliers that are a part of the long left tail: One Stop Systems (OSS), Senmiao Technology, and HyreCar. This gives us this normal Q-Q plot using standardized residuals:

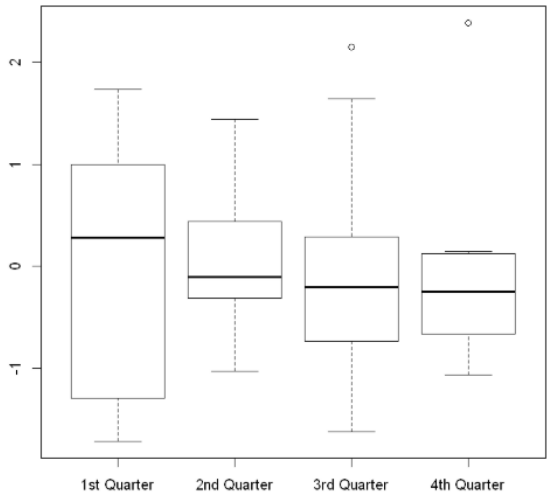


This seems to be taking care of a lot of the non-normality. To see if there is non-constant variance in our groups we can run a Levene's Test to see if the variances are equal across subgroups.

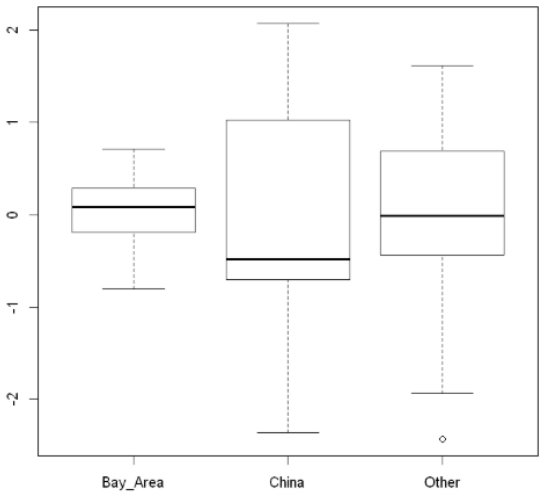
	Sum Sq	Df	F value	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>
quarter	1.170783	3	1.486781	0.236750891
location	3.044665	2	5.799647	0.007091339
quarter:location	3.214419	5	2.449202	0.054784826
Residuals	8.399586	32	NA	NA

This Levene's Test shows that because the p-value for quarter main effect is so low that there is some clear non constant variance for that subgroup. The p-value for the interaction effect is also quite low but that could just be because of the high variance in location that is being taken into account by it.





A side by side comparison of the box plots before and after taking those outliers out shows that the it has heavily equalized the amount of variance especially in the 1st quarter. Before, the plot of the standardized residuals against the quarters showed that the first quarter has much more vairance than the other quarters especially when compared to the 4th quarter.



Plotting the standardized residuals against the location also shows that China has a much higher variance than the other two location and the bay area has a very small variance when compared to the other two locations. This is probably where all of the non constant variance is coming from.

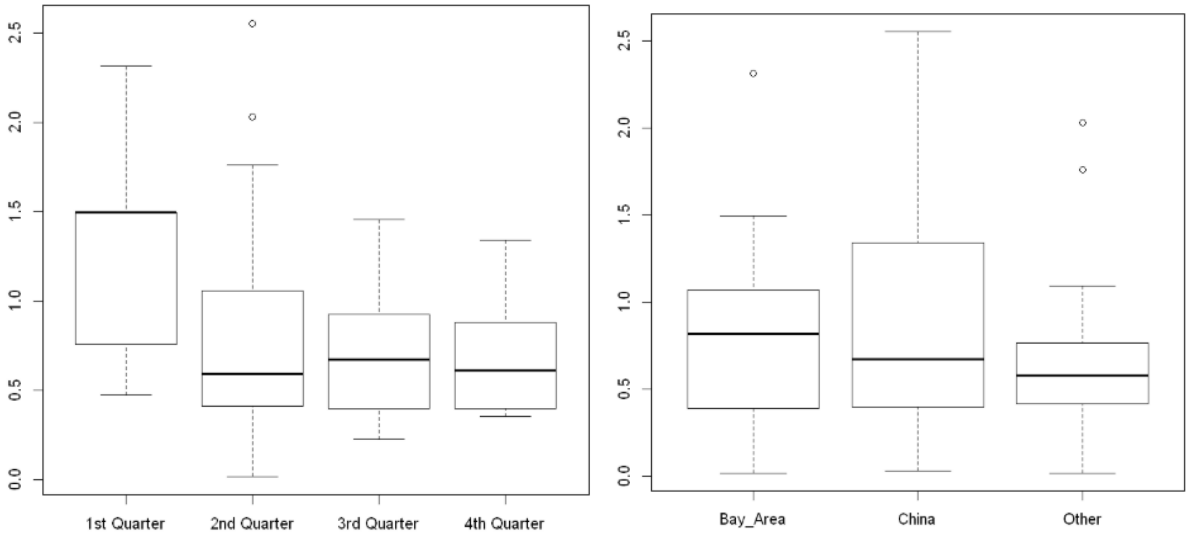
It seems as though doing a Weighted Least Squares makes sense in this case to take into account the differing amounts of variance between categories. Now we must go back to the start because these outliers may no longer be outliers once we take into account the differences in variance. However, I think it is safe to still not include Spotify since it is simply that much of and outlier that even taking it into account using WLS would still lead to an inefficient regression.

A anova: 4 × 4

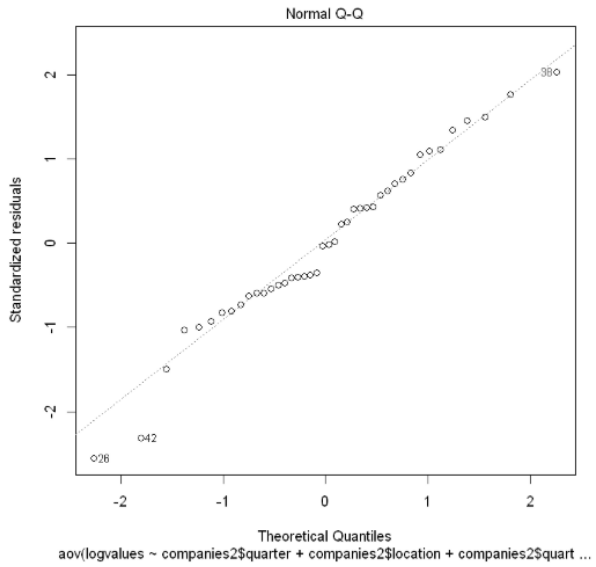
	Sum Sq	Df	F value	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>
companies2$quarter	0.3946943	3	0.4927058	0.6898612
companies2$location	0.3878369	2	0.7262182	0.4915363
companies2*quarter* location : *companies2*	0.3399295	4	0.3182563	0.8636367
Residuals	8.5448011	32	NA	NA

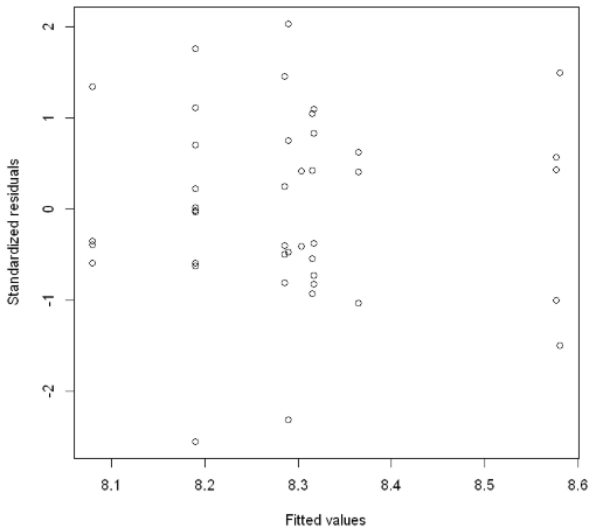
However, even after using WLS it seems like the interaction effect is still not needed.

Doing a Levene's Test shows that there is still non-constant variance which can also be seen in the boxplots of the subgroups against the absolute value of the standardized residuals so that we can compare levels.



These boxplots also show the same result. There still seems to be non-constant variance.





Looking at the Q-Q plot of the standardized residuals though and standardized residuals vs fitted values, shows us that there is still a little left tail caused by some outliers. Those two outliers were there even before we used WLS regression: Senmiao Technology and HyreCar. Even after taking those out it does not benefit the case for the interaction effect seeing as the p-value for the partial F- Test is still not close to significant. It appears there is no interaction effects between the quarter and location that are associated with the IPO value of a company.

Now it is down to the main effects. Do either or both show any evidence of a statistically significant relationship with the IPO value. Also, if we compare them are there any differences among the different subgroups.

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = logvalues ~ companies2$quarter + companies2$location, weight = wt)

$`companies2$quarter`
                                  diff        lwr       upr     p adj
2nd Quarter-1st Quarter -0.3407076822 -0.9609746 0.2795593 0.4582414
3rd Quarter-1st Quarter -0.3703346770 -0.9906016 0.2499323 0.3853182
4th Quarter-1st Quarter -0.3696055078 -1.0667254 0.3275144 0.4888010
3rd Quarter-2nd Quarter -0.0296269948 -0.4796159 0.4203620 0.9979628
4th Quarter-2nd Quarter -0.0288978256 -0.5800195 0.5222238 0.9989661
4th Quarter-3rd Quarter  0.0007291692 -0.5503925 0.5518508 1.0000000

$`companies2$location`
                      diff        lwr       upr     p adj
China-Bay_Area -0.04368448 -0.4686301 0.3812611 0.9656591
Other-Bay_Area -0.06278639 -0.4877320 0.3621592 0.9304412
Other-China    -0.01910190 -0.4273765 0.3891726 0.9927811
```

Looking at the Tukey comparison these different subgroups are not significantly different from each other. The p-values are very large and the confidence interval includes 0 for every comparison.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
quarter	3	3.34	1.1119	0.873	0.463
location	2	0.60	0.3015	0.237	0.790
Residuals	39	49.65	1.2731		

The partial F-tests also do not show any statistically significant relationship between either quarter or location on IPO value. Thus the means of these regression coefficients do not really mean anything if there is little association between any of these predictors and the response variable. However, as we can see even though there is normality there is still large amounts of non-constant variance that I have been unable to get rid of. One core assumption of linear regression is still not checked so any conclusions made from this data cannot be fully trusted. Still I can only conclude that this is little to no relationship between the quarter of the year or the location in which a company is headquartered in and the amount of money that is raised by that company in an IPO, at least for 2018.