AustinCityBnb

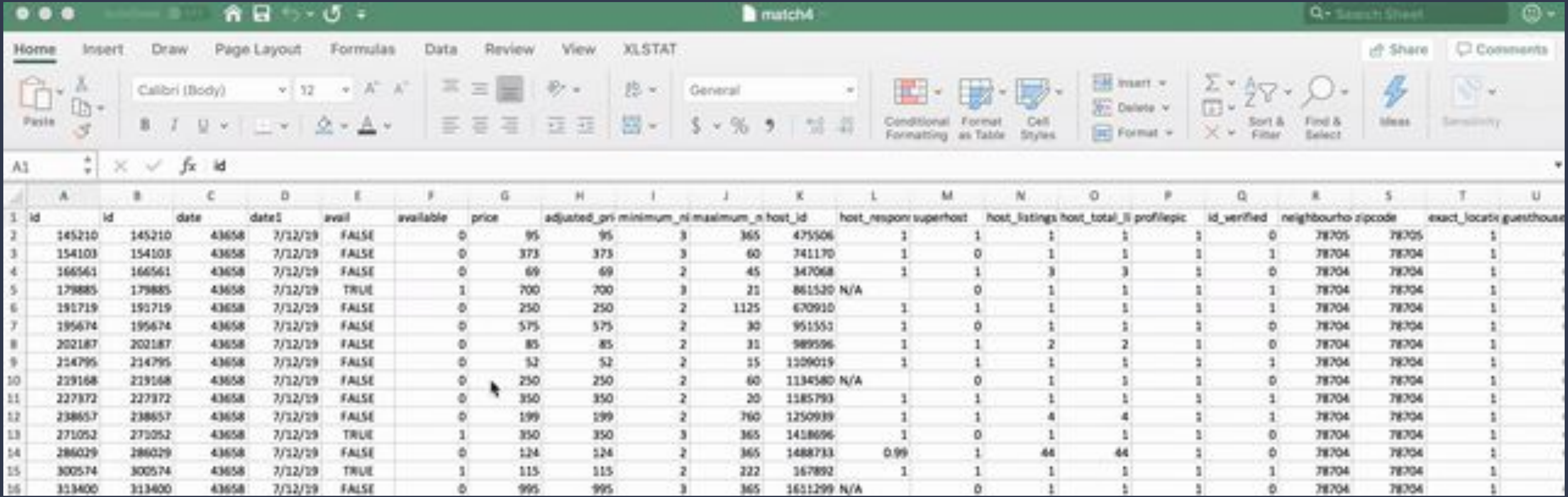
Valuing a local event from the perspective of Airbnb Hosts

Maya Rowen, Sijia Yu, Zoha Zahid

AUSTIN
CITY
LIMITS®

+

airbnb



- Data pulled on July 12, 2019
- Available listings from July 12, 2019 till July 11, 2020
- Attributes include:
    - Listing features (bedrooms, bathrooms, shared/private, etc)
    - Price
    - Location
    - Host information (SuperHost, responsive rate, number of listings etc)
    - Availability (till July 11, 2020)
- **260,976 rows of data!**

Hypotheses

Price bumps

AUSTIN CITY LIMITS®

High importance value of location

Changes in booking behavior

Lower availability of listings

Overview

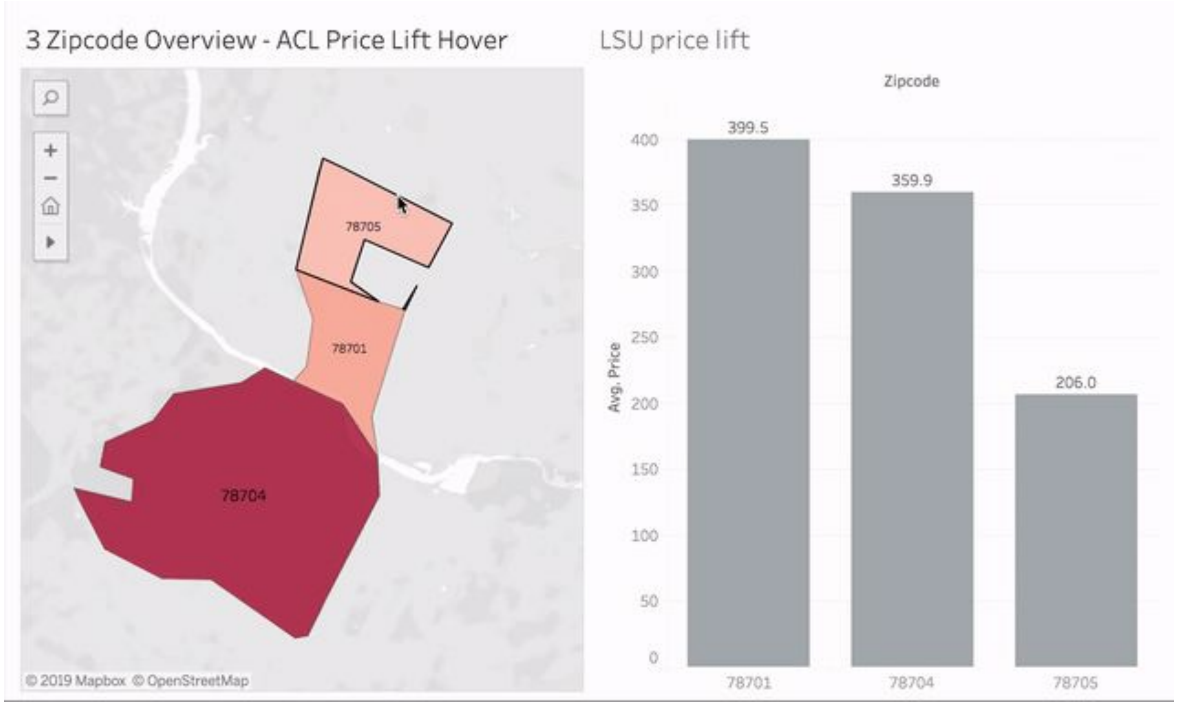
*Data pulled on the date of July 12, 2019

Overview

*Data pulled on the date of July 12, 2019

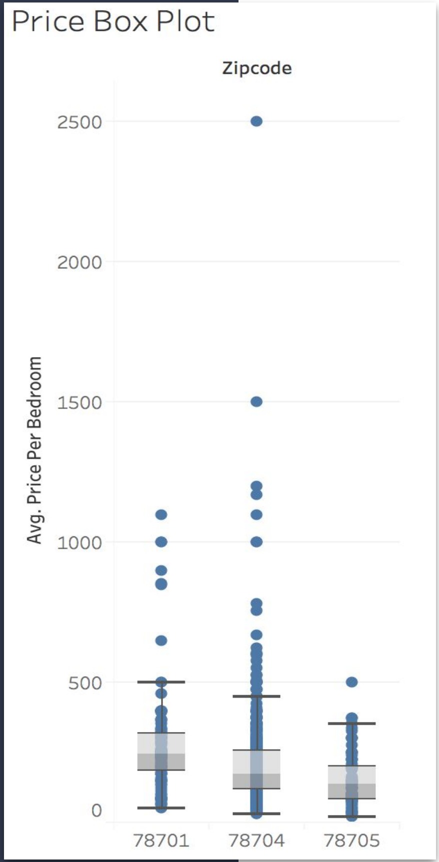
Overview

Listing price lift when there is an event (ACL, LSU Game)*



*Data pulled on the date of July 12, 2019

How much should you charge your guest during ACL weekends?



Average price per bedroom during ACL in 3 zipcodes

For our study, we summarized the data through following steps:

- Date cut-off at December 31, 2019 due to a selection bias affecting dates after the end of the year
- Listing features are combined to 5 groups:
  - Group 1: 1 bedroom, entire house
  - Group 2: 1 bedroom, shared house
  - Group 3: 2 bedrooms 1 bathroom
  - Group 4: 2 bedrooms 2 bathrooms
  - Group 5: 3-4 bedrooms
- Price is calculated as price per bedroom.
- Zipcode is limited to *Zilker (78704)*, *Downtown (78701)* and *West Campus (78705)*

Let’s try linear regressions...

lm1<-lm(formula=price~**bedrooms+bathrooms+zilker+entire_home**, data=data8)

FIRST REGRESSION

```
lm1<-lm(formula=price~bedrooms+bathrooms+zilker+entire_home, data=data8)
summary(lm1)
```

```
##
## Call:
## lm(formula = price ~ bedrooms + bathrooms + zilker + entire_home,
##     data = data8)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -373.85  -95.50  -45.44   55.72  732.66
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -40.6305     1.6671 -24.373   <2e-16 ***
## bedrooms     81.8407     0.8260  99.083   <2e-16 ***
## bathrooms    88.8248     0.9986  88.953   <2e-16 ***
## zilker      -10.0576     1.1312  -8.891   <2e-16 ***
## entire_home  50.4638     1.4367  35.124   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 164.3 on 107855 degrees of freedom
```

Attributes:

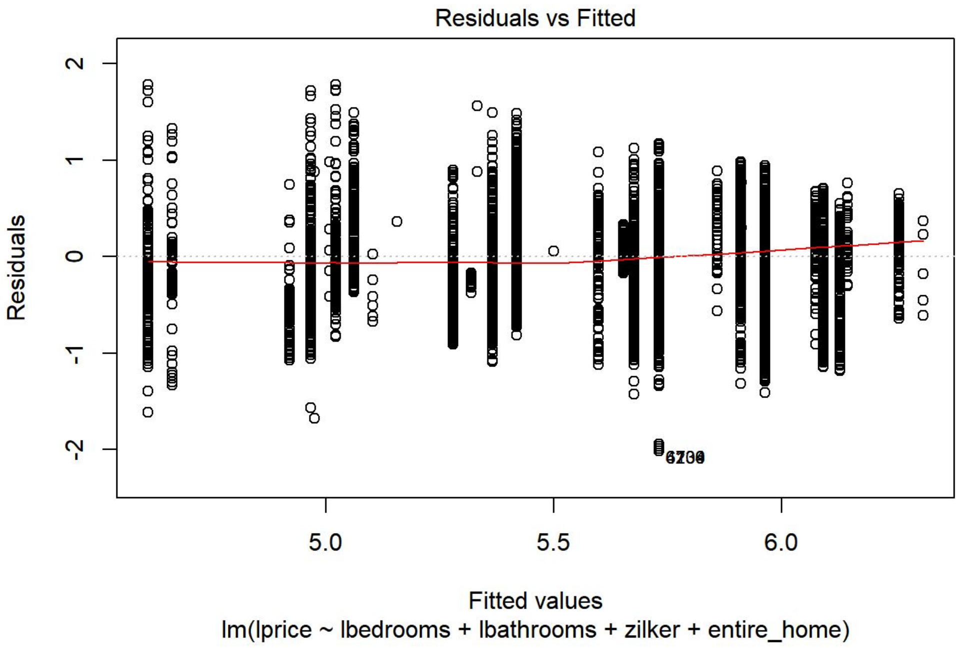
- Number of bedrooms
- Number of bathrooms
- Zilker
- Entire home

The residuals don’t look great, but our t-values are all significant.

Let’s try the log-log model.

A log-log regression

Question: where is date in this model?



This model still shows signs of heteroskedasticity and patterns in the residuals, but based on the t-scores and coefficients, we proceed to build a **random effects** predicting price subject to date and listing features with the hypothesis that bedrooms, bathrooms, zipcode, and home_type are all significant determinants of price.

17 Groups to 5 Groups...

333815	89	1	78704	1	3	0	0	89.0	2019
333856	109	0	78704	1	3	0	0	109.0	2019
354334	850	0	78704	2	3	1	0	425.0	2019
356837	799	0	78701	2	3	1	0	399.5	2019
412290	47	0	78704	1	2	0	0	47.0	2019
433996	109	1	78704	1	2	1	0	109.0	2019
916328	350	1	78704	2	3	1	0	175.0	2019
929974	95	1	78704	1	2	0	0	95.0	2019
1080879	59	0	78704	1	2	0	0	59.0	2019

1-10 of 5,832 rows | 1-10 of 21 columns Previous 1 2 3 4 5 6 ... 584 Next

```
filter(lmer_data, bedrooms==4,entire_home==0)
```

0 rows | 1-10 of 21 columns

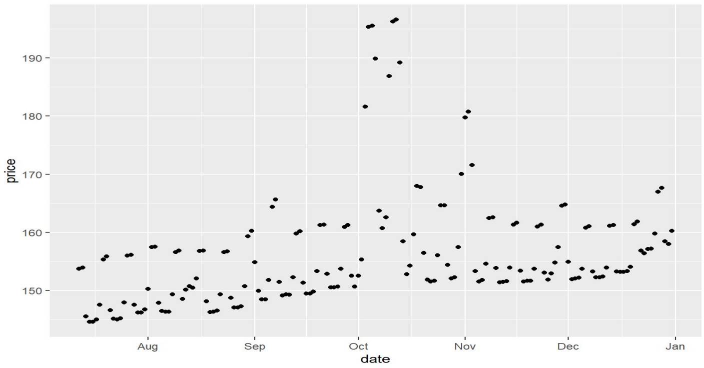
```
filter(lmer_data, bedrooms==3,entire_home==0)
```

*Tests: correlation test between number of bedrooms and bathrooms; the size of each groups.

A Random Effects Model

Fixed effect:

- Date: because we are interested in how price varies over time, we use date as a fixed effect in our analyses and test its interaction with our random effects



Random effects:

- Feature group: by dividing the listings into these groups, we will be able to see to which extent each combination explains the price variation

Feature Group	Group Details	Count
1	1 bedroom, entire house	26,134
2	1 bedroom , shared house	15,743
3	2 bedrooms, 1 bathroom, entire house	17,819
4	2 bedrooms, 2 bathrooms, shared house	14,178
5	3-4 bedrooms	16,040

- Zipcode: similarly, we can see how impactful the location of the listing is in explaining the price variation

We will use the R package Lme4 to build our model:
https://cran.r-project.org/web/packages/lme4/vignettes/lmer.pdf

fm<-lmer(formula<-y~x+(1|random1)+(1|random2)

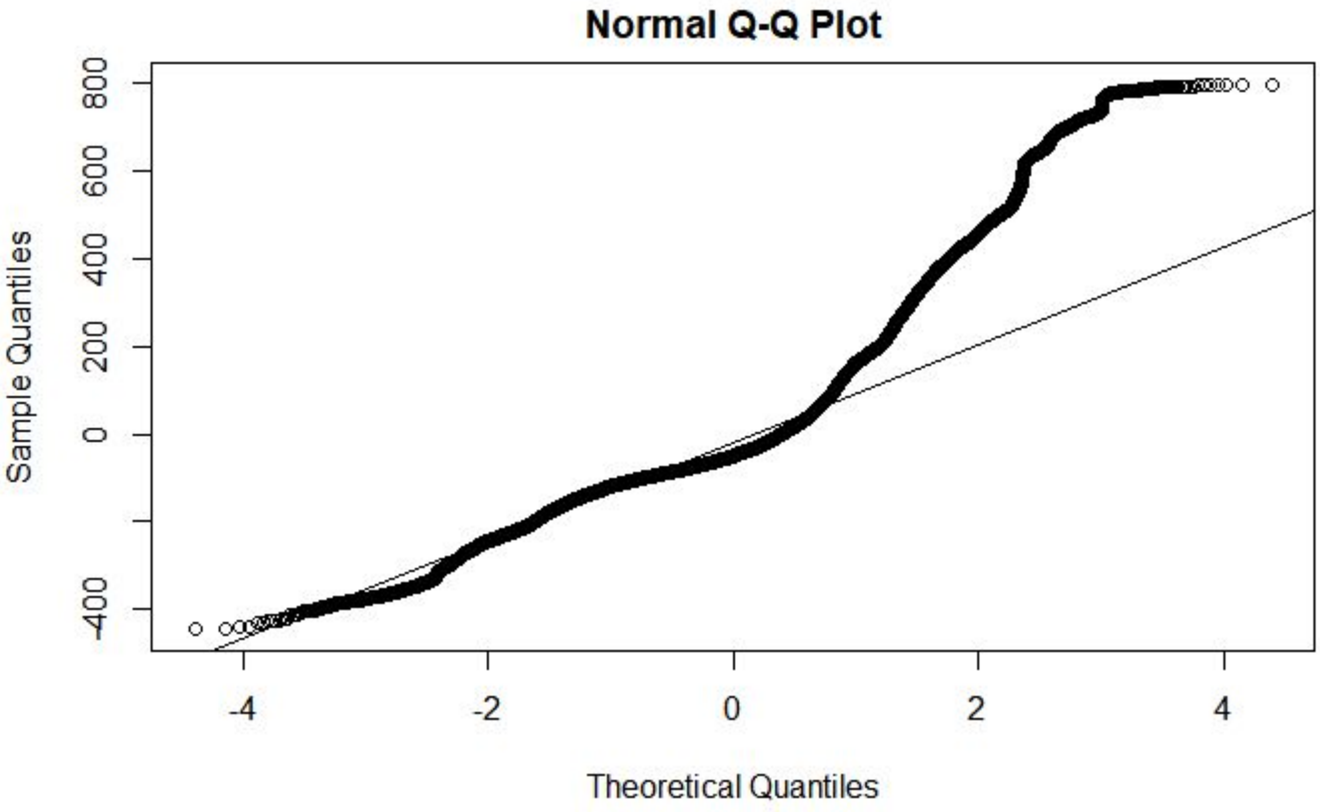
lmer1 <- lmer (formula = price ~ date_f + (1|feature_group)+(1|zipcode), data=lmer_data5)

FIXED EFFECTS

```
REML criterion at convergence: 163186.8

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.3613 -0.6734 -0.1092  0.6133  3.1211

Random effects:
 Groups        Name        Variance Std.Dev.
 feature_group (Intercept) 0.26474  0.51453
 zipcode       (Intercept) 0.00708  0.08414
 Residual                  0.35627  0.59688
Number of obs: 89914, groups:  feature_group, 5; zipcode, 3

Fixed effects:
                 Estimate Std. Error t value
(Intercept)      5.3463986  0.2366368  22.593
date_f2019-07-13  0.0019763  0.0370172   0.053
date_f2019-07-14 -0.0929871  0.0369994  -2.513
```

RANDOM EFFECTS

```
$feature_group
  (Intercept)
1 -0.296441986
2 -0.645781193
3 -0.009126066
4  0.250552855
5  0.700796370

$zipcode
      (Intercept)
78701  0.09125282
78704 -0.01711276
78705 -0.07414006
```


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
800
600
400
200
0
-400
-4
-2
0
2
4

Observation/Improvement

Note that...

- Not all dates have a statistically significant t-score
- QQplot shows heteroscedasticity
- All ACL dates have large positive and statistically significant coefficients - $40 to $50 intercept bump!

Next...

- Log-log model to improve the fit
- Add interaction between date and zipcode
- Add interaction between date and feature group

```
lmer2<-lmer(formula=lprice~date_f+(1|feature_group:date_f)+(1|zipcode:date_f),data=lmer_data5)
```

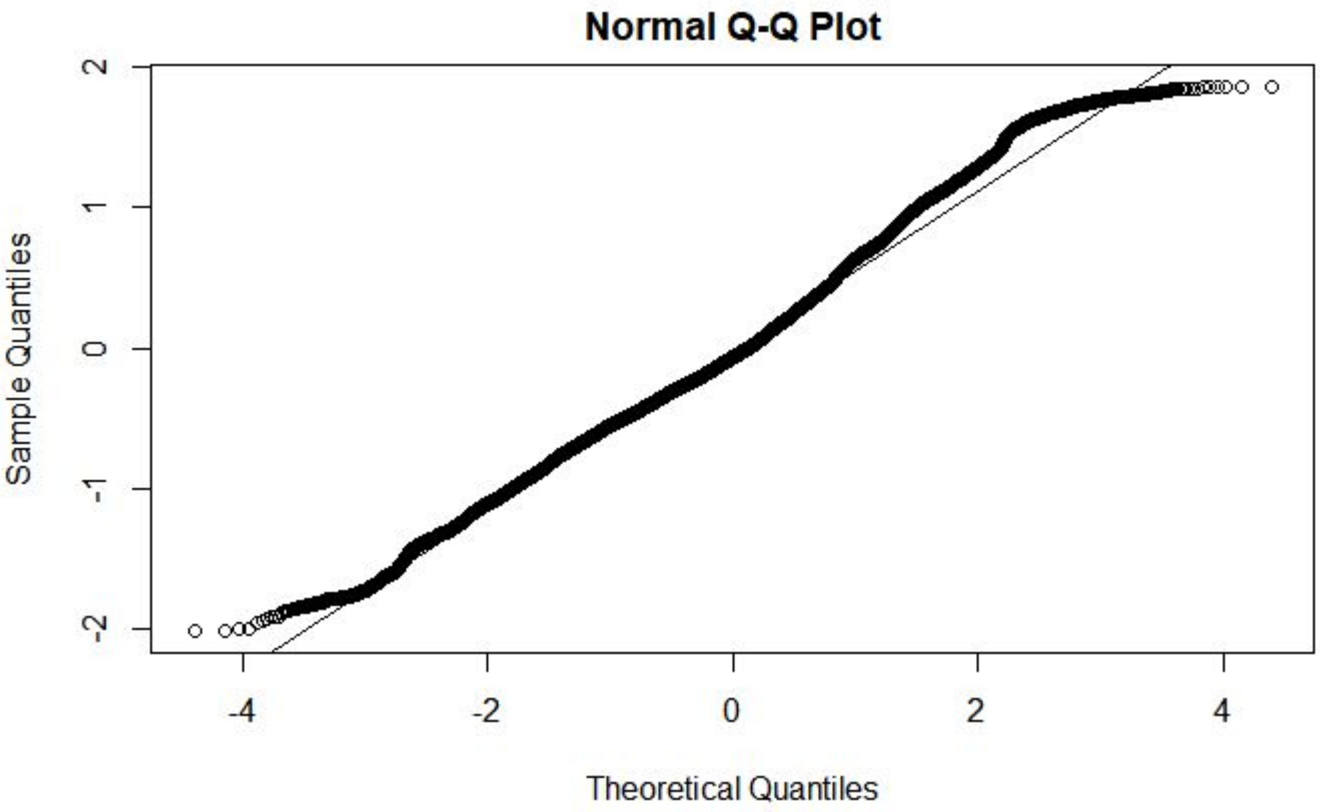
The Log Model: log_lmer1<-lmer(formula=lprice~date_f+(1|feature_group)+(1|zipcode),data =lmer_data5)

FIXED EFFECTS

```
Linear mixed model fit by REML ['lmerMod']
Formula: lprice ~ date_f + (1 | feature_group) + (1 | zipcode)
   Data: lmer_data5

REML criterion at convergence: 163186.8

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.3613 -0.6734 -0.1092  0.6133  3.1211

Random effects:
 Groups        Name        Variance Std.Dev.
 feature_group (Intercept) 0.26474  0.51453
 zipcode       (Intercept) 0.00708  0.08414
 Residual                  0.35627  0.59688
Number of obs: 89914, groups:  feature_group, 5; zipcode, 3

Fixed effects:
                  Estimate Std. Error t value
(Intercept)       5.3463986  0.2366368  22.593
date_f2019-07-13  0.0019763  0.0370172   0.053
date_f2019-07-14 -0.0929871  0.0369994  -2.513
date_f2019-07-15 -0.1064446  0.0369994  -2.877
date_f2019-07-16 -0.1062999  0.0369994  -2.873
```

RANDOM EFFECTS

```
$feature_group
   (Intercept)
1 -0.296441986
2 -0.645781193
3 -0.009126066
4  0.250552855
5  0.700796370

$zipcode
      (Intercept)
78701  0.09125282
78704 -0.01711276
78705 -0.07414006

with conditional variances for “feature_group” “zipcode”
Analysis of Variance Table
        Df Sum Sq Mean Sq F value
date_f 172 439.71  2.5565  7.1757
```


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
2
1
0
-1
-2
-4
-2
0
2
4

Random Effects Model

$feature_group

(Intercept)
1   -81.60254
2 -128.08494
3   -35.62097
4   44.53983
5   200.76862

$zipcode
     (Intercept)
78701   23.579224
78704   -1.594414
78705 -21.984810

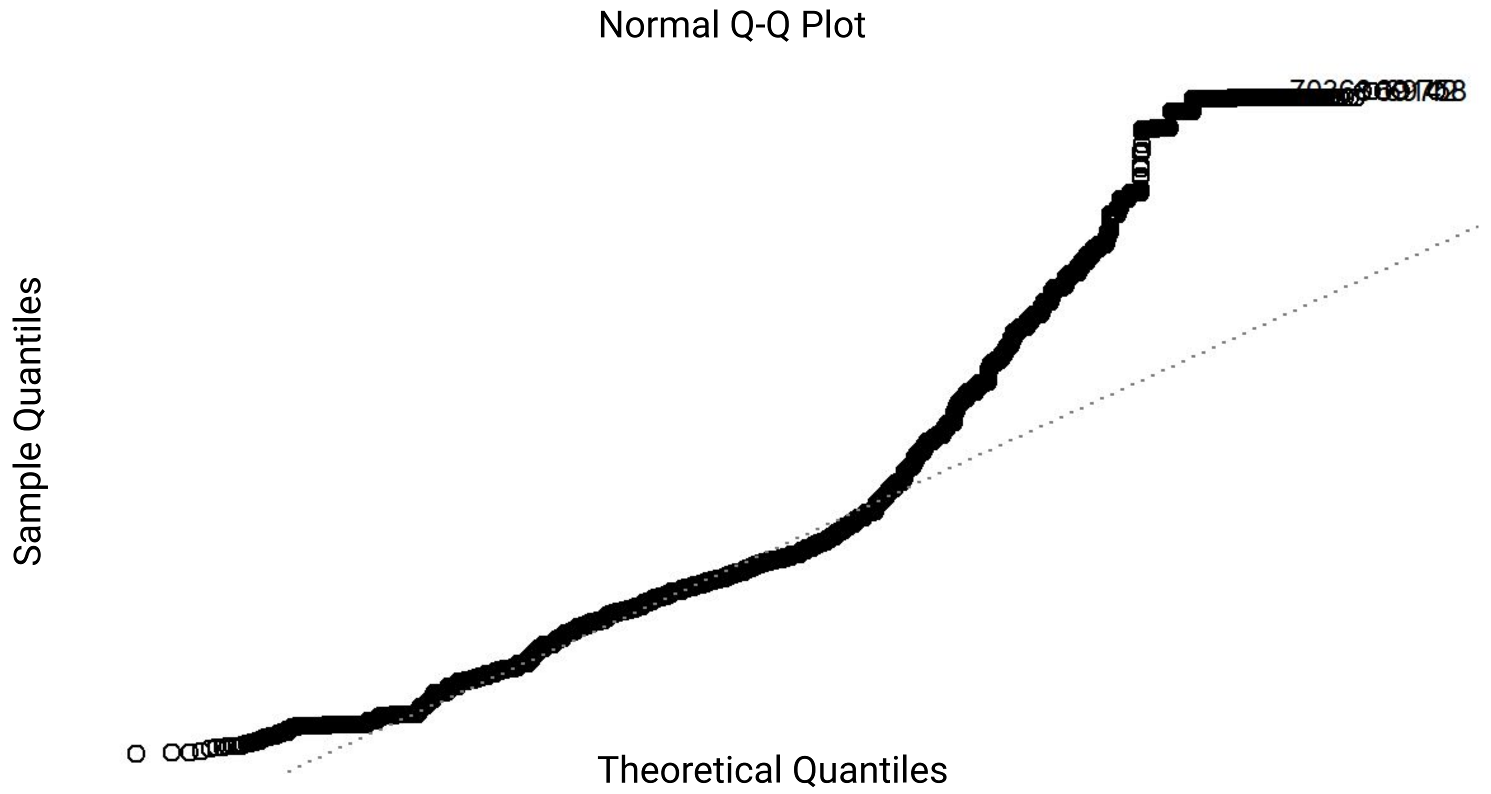
- **Feature groups**: groups 1,2 and 3 have negative effects on price
- As you go to higher feature groups, number of bedrooms increases - A price increase for these groups makes sense
- Groups 1 and 2 both have just 1 bedroom, but as a Group 2 is a shared listing, it’s price is even lower on average

- **Location**: houses in Zilker and West Campus have a have a negative effect on price
- Indicates higher WTP for listings downtown


Percent Increase in Price Over Time
30.00%
25.00%
20.00%
15.00%
10.00%
5.00%
0.00%
-5.00%
-10.00%
-15.00%
7/12/2019
8/12/2019
9/12/2019
10/12/2019
11/12/2019
12/12/2019


expected daily price
$300.00
$250.00
$200.00
$150.00
$100.00
$50.00
$0.00
7/12/2019
8/12/2019
9/12/2019
10/12/2019
11/12/2019
12/12/2019

Model Improvement...



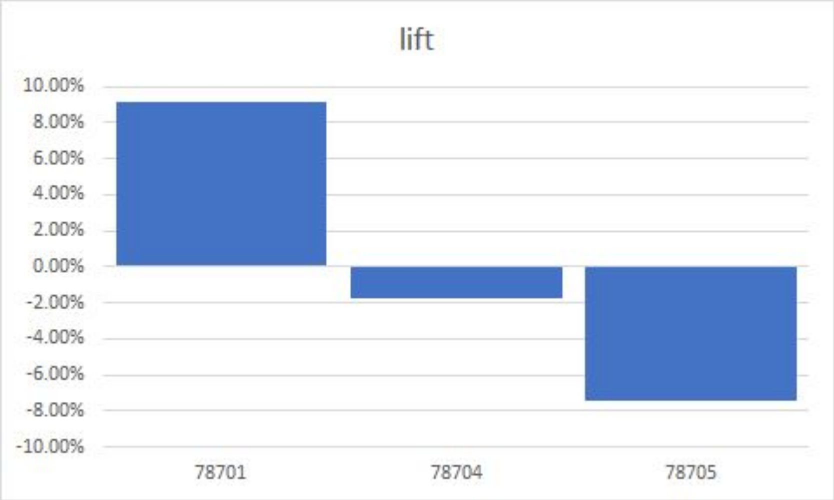

```
qqnorm(resid(lmer1))
qqline(resid(lmer1))
```

Listing Group

group	lift	price
1	-29.64%	$1.00
2	-64.58%	$1.00
3	-0.91%	$1.00
4	25.06%	$1.00
5	70.08%	$1.00

Zipcode

zipcode	lift	price
78701	9.13%	$229.90
78704	-1.71%	$206.29
78705	-7.41%	$194.86



Fact Check

Booker and host behavioral analysis

Price ~ Availability

- Hosts:
  - Will list rooms well ahead of time for ACL dates (observation: more listings)
- Bookers
  - Will book rooms relatively ahead of time and reserve the *cheaper* listings (observation: higher priced listings left)

```
cor.test(lmer_data$price_bed,lmer_data$available)
```

```

	Pearson's product-moment correlation

data:  lmer_data$price_bed and lmer_data$available
t = -24.999, df = 89912, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.08956949 -0.07658702
sample estimates:
        cor
-0.08308178
```

Date ~ Availability

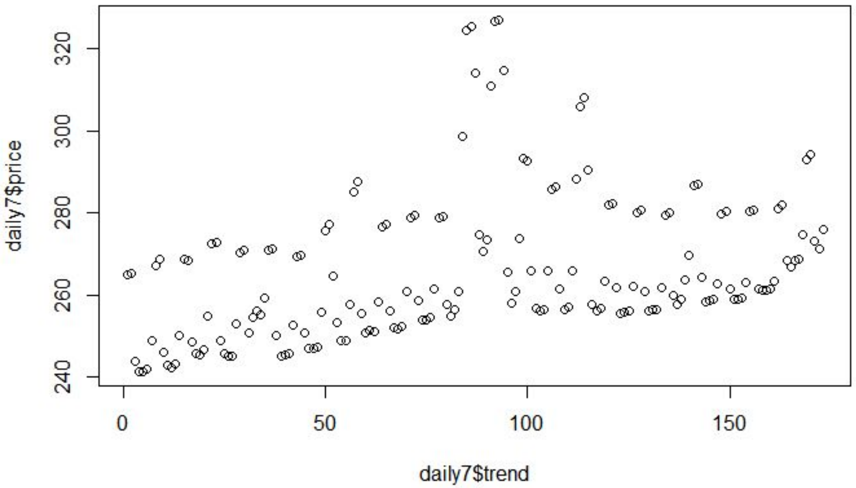
- Available=mean(available)
- Trend=seq.int(nrow(daily4))
- Correlation test: cor.test(daily7$available,daily7$trend)

```
##  Pearson's product-moment correlation
##
## data:  daily7$available and daily7$trend
## t = 11.959, df = 171, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5845339 0.7486898
## sample estimates:
##       cor
## 0.6748765
```

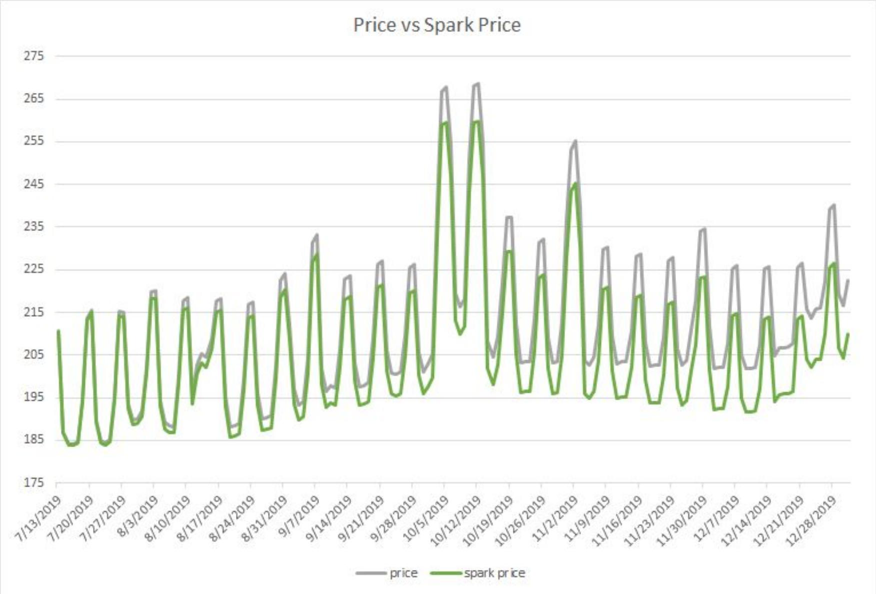
Price ~Time (Trend)

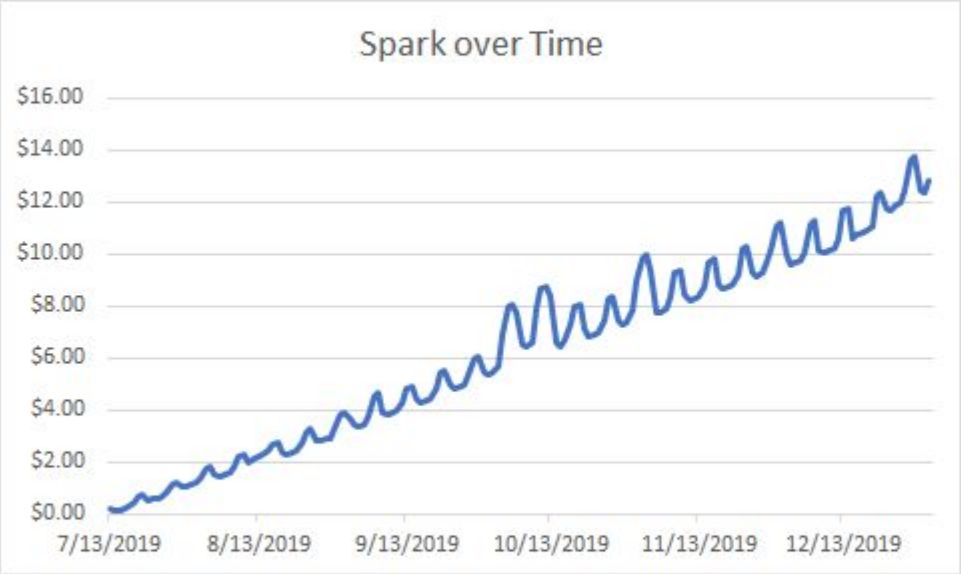
Some explanations for log-multiplier

The correlation between price and time is significant, positive, and very strong. This is logical because it tells us that listers are more likely to list far in advance than bookers are to book far in advance, so the number of available listings is lower in the near future than distant future. It’s a good indicator of differences in lister and booker behavior, and reinforces the notion that we have a selection bias in our data. Since price is the variable we are most interested in exploring, we return to the relationship between price and date. Rather than diluting our random effects model with a variable accounting for availability over time, we will create a multiplier to apply to the model output, i.e., a multiplier we draw from a simple log model.



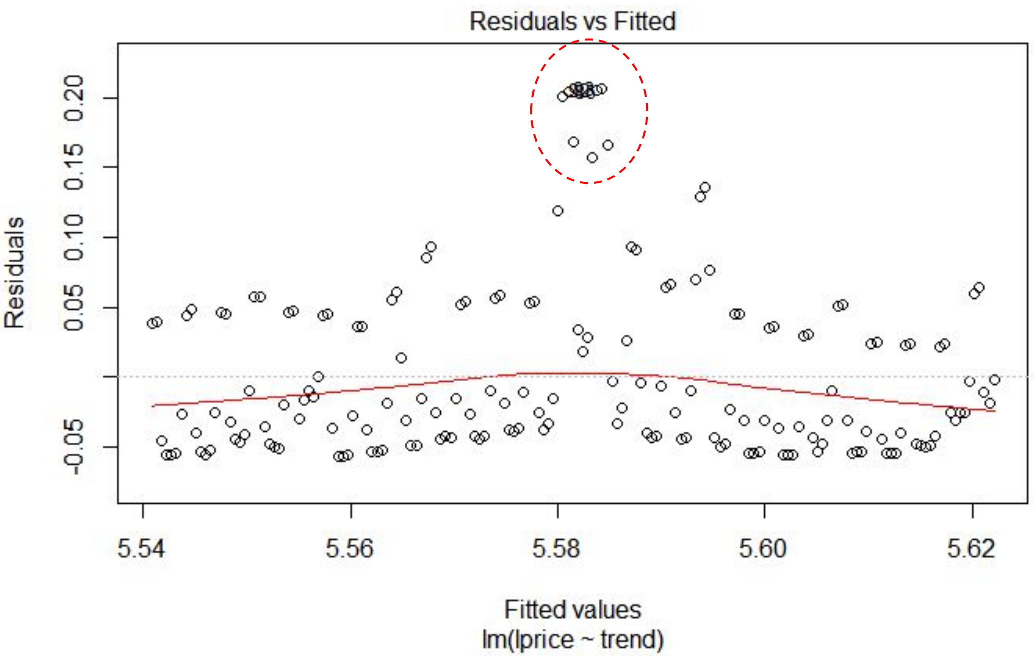
Spark Model





*SparkPrice=price-spark*price/bedrooms-.0005*trend*price/bedrooms*

```
log_multiplier<-lm(formula=lprice~trend, data=daily7)

summary(log_multiplier)

plot(log_multiplier)

spark<-resid(log_multiplier)

daily8<-cbind(spark,daily7,deparse.level=1)

spark_data<-merge(lmer_dec,daily8, by=NULL)

spark_data1<-mutate(spark_data, spark_bed=spark*bedrooms)

spark_data2<-mutate(spark_data1,

 spark_price=(price.x-spark_bed-.0005*trend*(price.x/bedrooms)))

spark_data3<-mutate(spark_data2,make_date(date.y))

spark_data4<-filter(spark_data3,month(date.y)==month,

day(date.y)==day)

spark_data5<-mutate(spark_data4,lspark_price=log(spark_price))
```



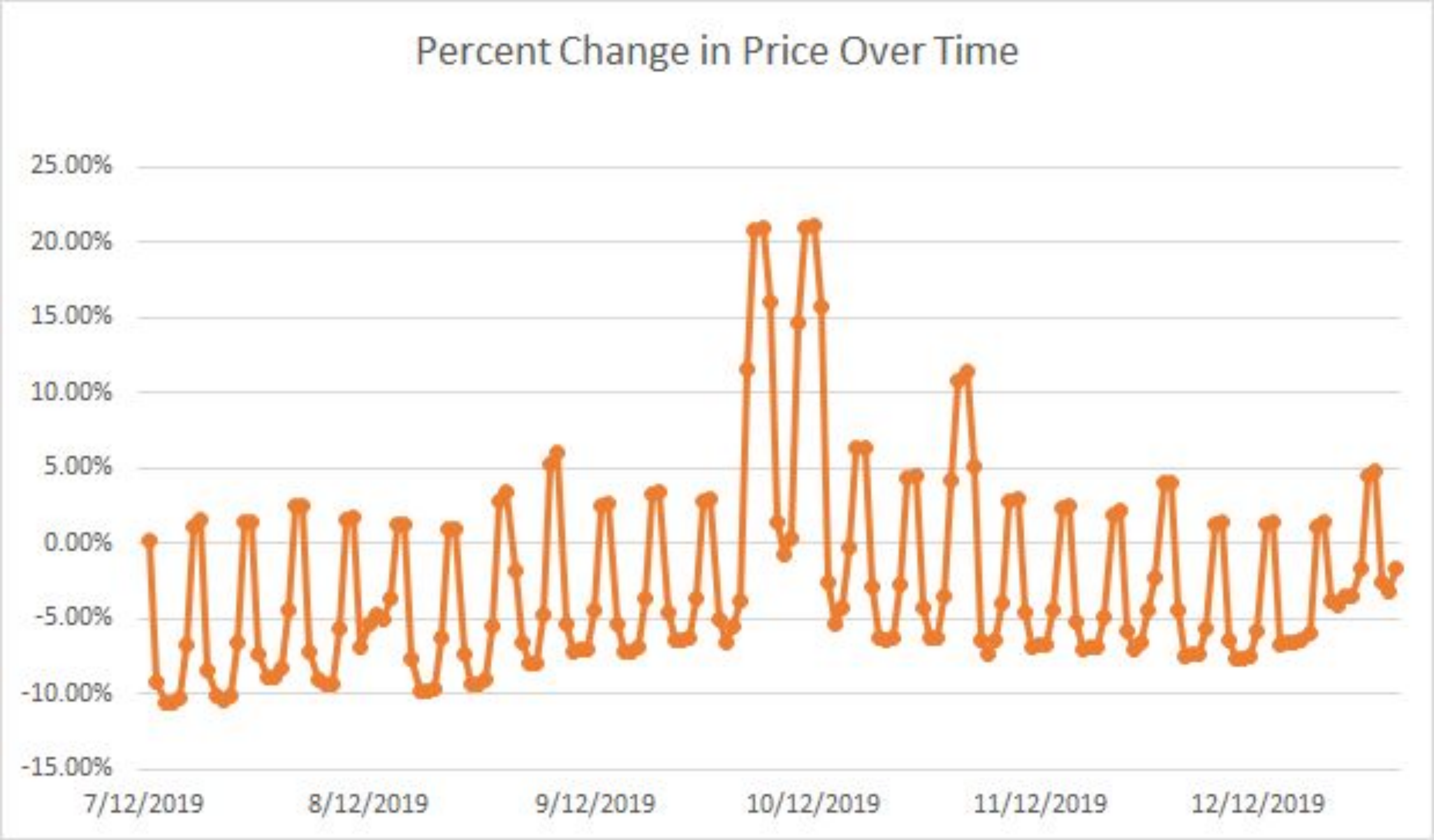
log_lmer1_spark<-lmer(formula=lspark_price~date_f+(1|feature_group)+(1|zipcode), data=spark_data5)

FIXED EFFECTS

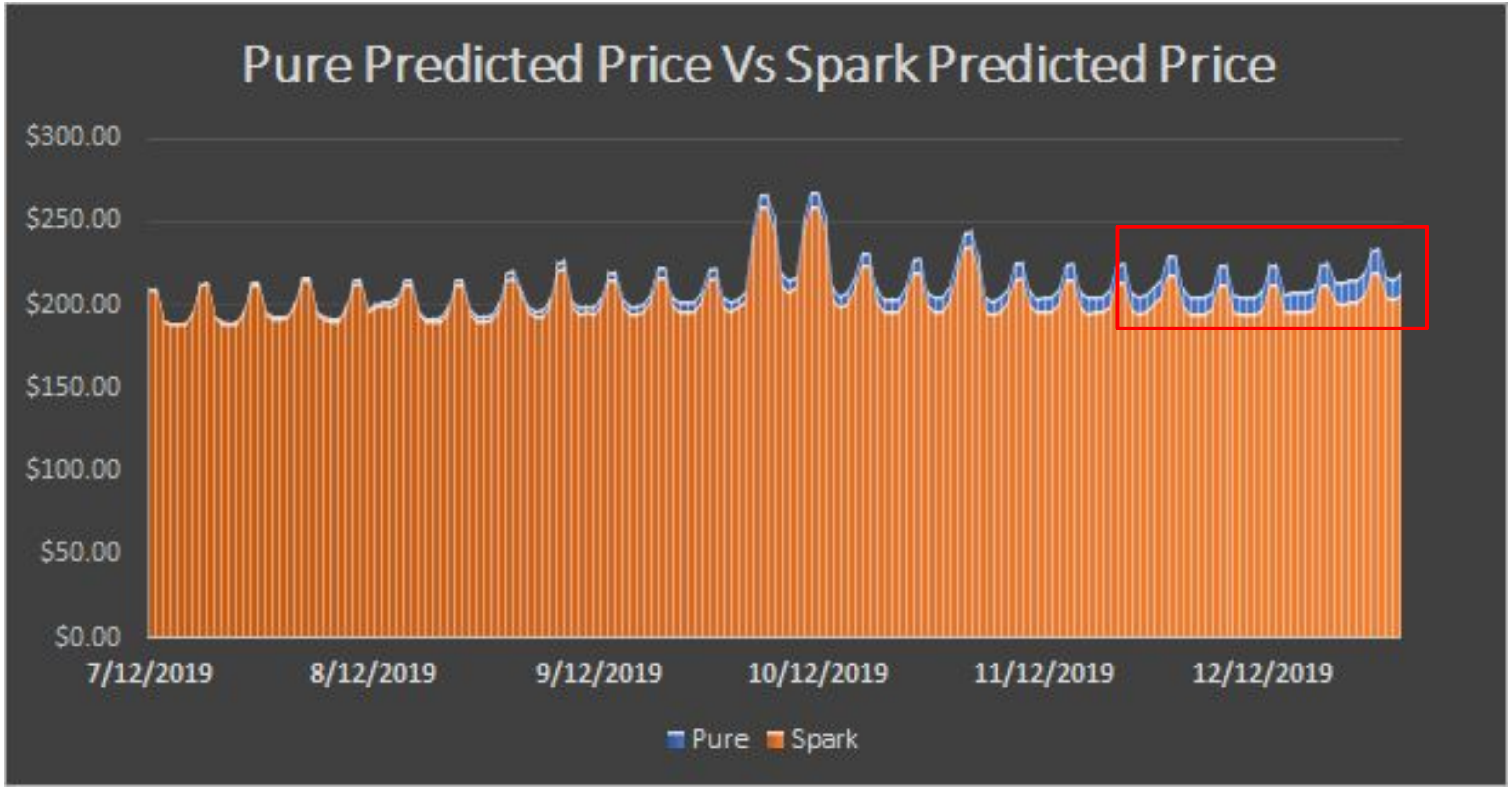
```
REML criterion at convergence: 1172006

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.6458 -0.5678 -0.2921  0.3150  4.8442

Random effects:
 Groups        Name        Variance Std.Dev.
 feature_group (Intercept) 16701.7  129.23
 zipcode       (Intercept)   473.8   21.77
 Residual                  27157.3  164.79
Number of obs: 89914, groups:  feature_group, 5; zipcode, 3

Fixed effects:
                Estimate Std. Error t value
(Intercept)     274.2578    59.5878   4.603
date_f2019-07-13   0.2478    10.2201   0.024
date_f2019-07-14 -21.0668    10.2152  -2.062
date_f2019-07-15 -23.4185    10.2152  -2.293
```

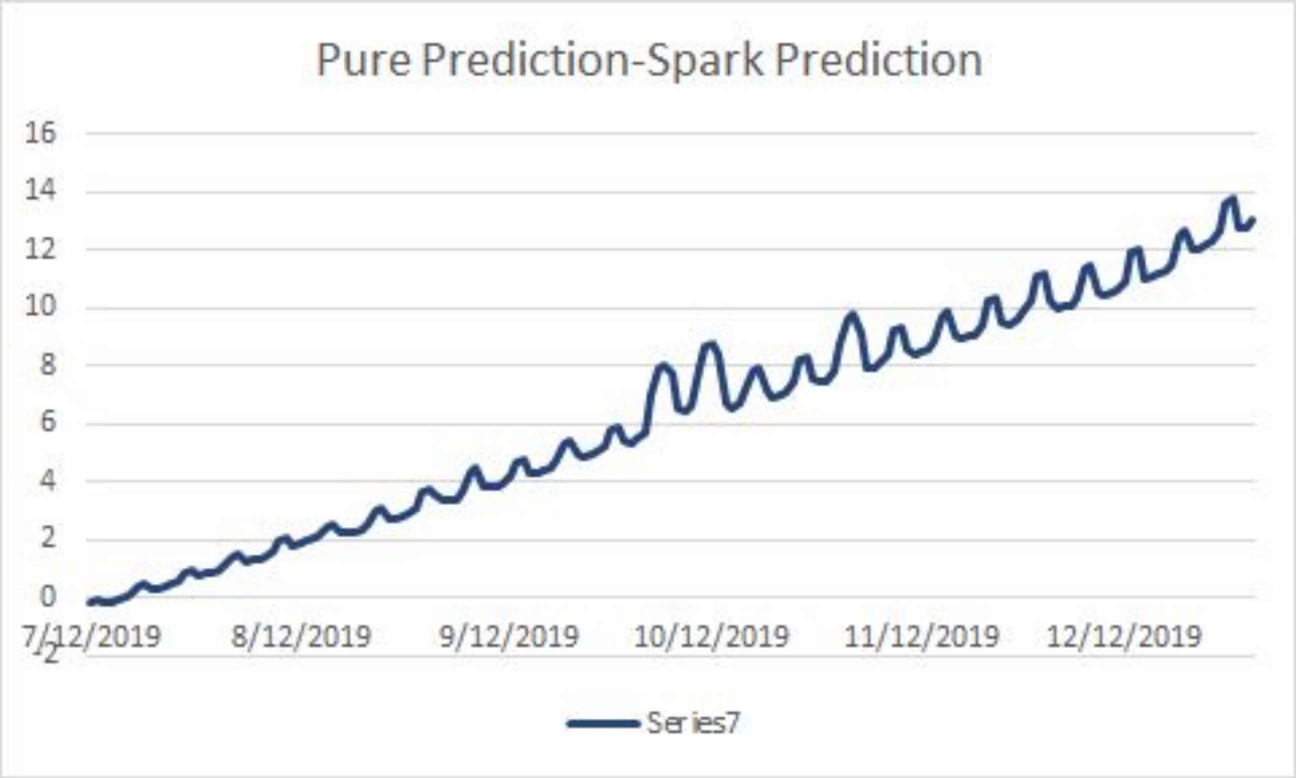
RANDOM EFFECTS

```
Analysis of Variance Table
        Df Sum Sq Mean Sq F value
date_f 172 373.52  2.1717  6.0904
$feature_group
  (Intercept)
1 -0.311754169
2 -0.661068319
3 -0.001626554
4  0.258002320
5  0.716446647

$zipcode
      (Intercept)
78701  0.09128013
78704 -0.01712680
78705 -0.07415333
```


Percent Change in Price Over Time
25.00%
20.00%
15.00%
10.00%
5.00%
0.00%
-5.00%
-10.00%
-15.00%
7/12/2019
8/12/2019
9/12/2019
10/12/2019
11/12/2019
12/12/2019


Listing Price Over Time
$300.00
$250.00
$200.00
$150.00
$100.00
$50.00
$0.00
7/12/2019
8/12/2019
9/12/2019
10/12/2019
11/12/2019
12/12/2019


Pure Predicted Price Vs Spark Predicted Price
$300.00
$250.00
$200.00
$150.00
$100.00
$50.00
$0.00
7/12/2019
8/12/2019
9/12/2019
10/12/2019
11/12/2019
12/12/2019
Pure
Spark



Total Over-Prediction by Pure Model: $559,565.10

Experimental: an Interactive Model

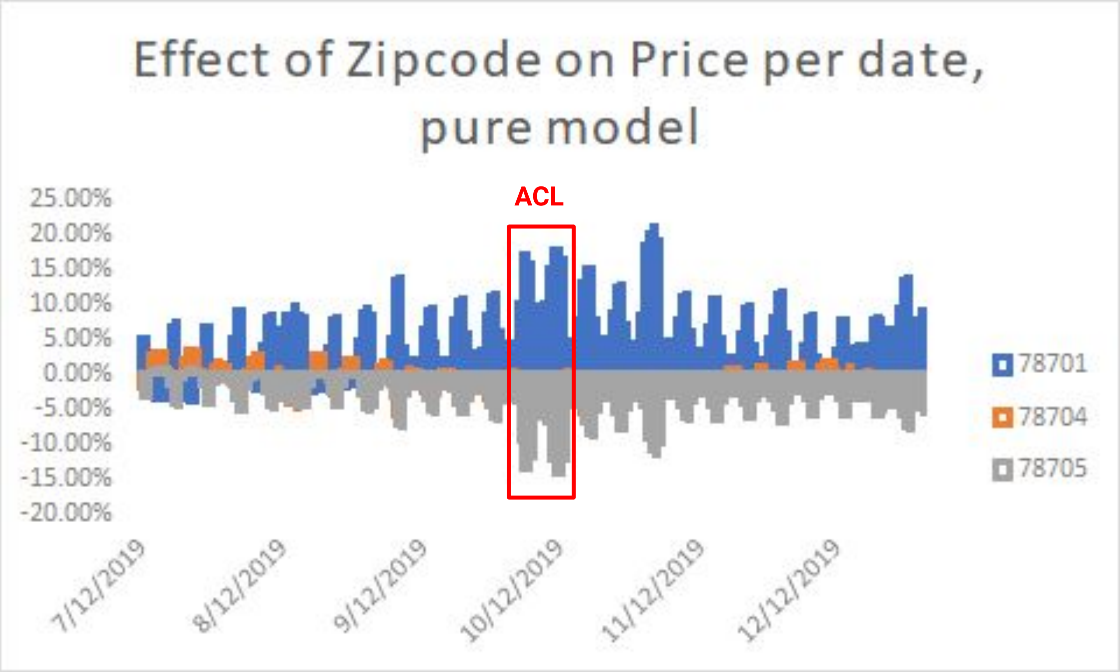
fm<-lmer(formula=y~x+(1|x:random1)+(1|x:random2)

We test for correlation between our random effects and fixed effect *lmer2<-lmer(formula=lprice~date_f+(1|feature_group:date_f)+(1|zipcode:date_f),data=lmer_data5)*

```
Analysis of Variance Table
       Df Sum Sq  Mean Sq F value
date_f 172 5.7997 0.033719  0.0945


$`feature_group:date_f`
             (Intercept)
1:2019-07-12 -0.2998089377
1:2019-07-13 -0.2983868739
1:2019-07-14 -0.2477463354


$`zipcode:date_f`
                 (Intercept)
78701:2019-07-12  4.986867e-02
78701:2019-07-13  4.851789e-02
78701:2019-07-14 -2.966720e-02
78701:2019-07-15 -3.681704e-02
```

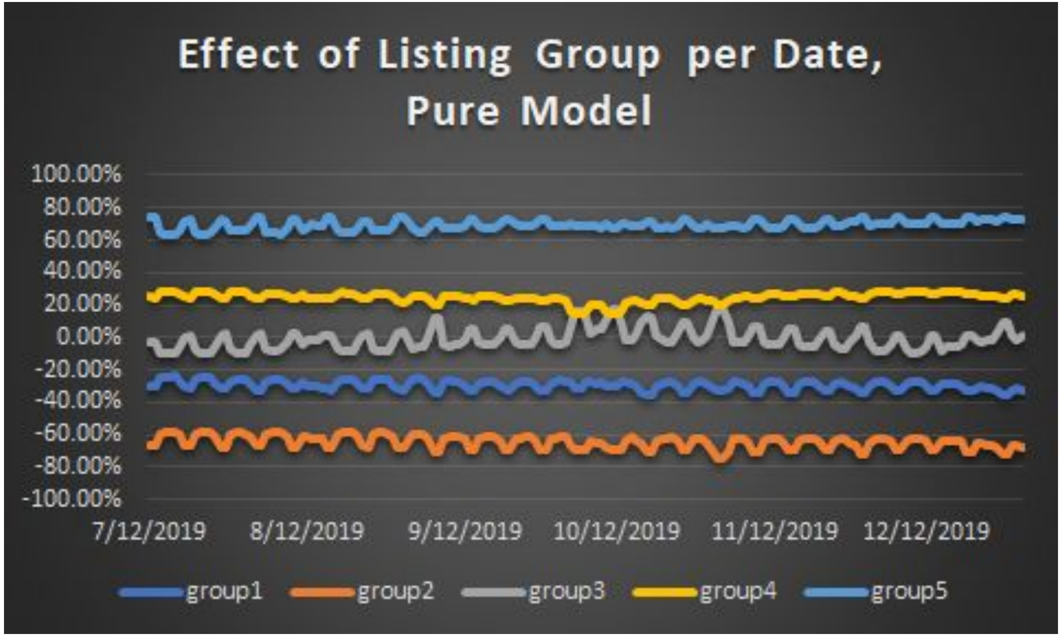
lmer2_spark<-lmer(formula=lspark_price~date_f+(1|feature_group:date_f)+(1|zipcode:date_f),data=spark_data5)

```
Analysis of Variance Table
       Df Sum Sq  Mean Sq F value
date_f 172 7.0027 0.040713  0.1141


$`feature_group:date_f`
             (Intercept)
1:2019-07-12 -0.2995626024
1:2019-07-13 -0.2979763028
1:2019-07-14 -0.2470013474

$`zipcode:date_f`
                 (Intercept)
78701:2019-07-12  4.986867e-02
78701:2019-07-13  4.851789e-02
78701:2019-07-14 -2.966720e-02
78701:2019-07-15 -3.681704e-02
```



Column1	78701	78704	78705
average	5.89%	-1.57%	-4.32%
acl	16.87%	-3.49%	-13.38%

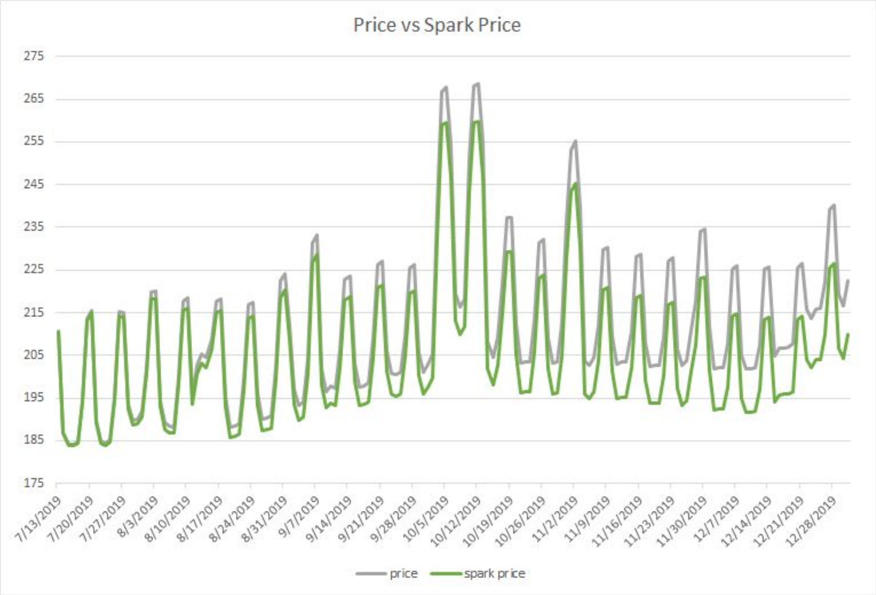
*Without Spark - model does not account for the interaction between date~groups/zipcode

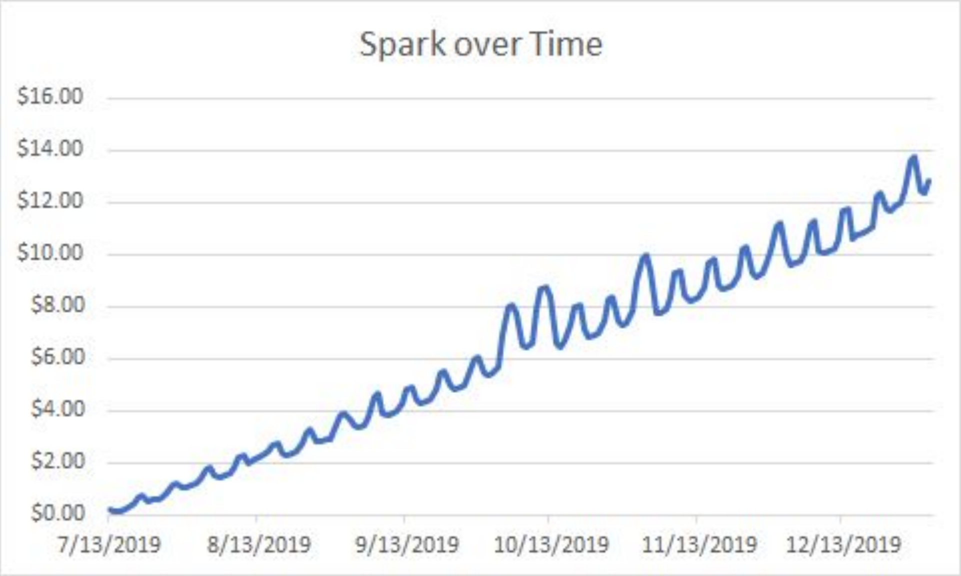


Column1	group1	group2	group3	group4	group5
average	-0.29298	-0.64393	-0.00643	0.248242	0.695095
acl	-0.29471	-0.69064	0.14645	0.15169	0.687209

*Without Spark - model does not account for the interaction between date~groups/zipcode

Spark in The Interaction Model





Putting the Model To The Test

- We tested the model on 9885 listings from our dataset designated as the test group

SSE	**140578185.8**
variance	14221.36427
std_dev	119.2533617

Groups	Name	Variance	Std.Dev.
feature_group	(Intercept)	0.26474	0.51453
zipcode	(Intercept)	0.00708	0.08414
Residual		0.35627	0.59688

Putting the Model To The Test

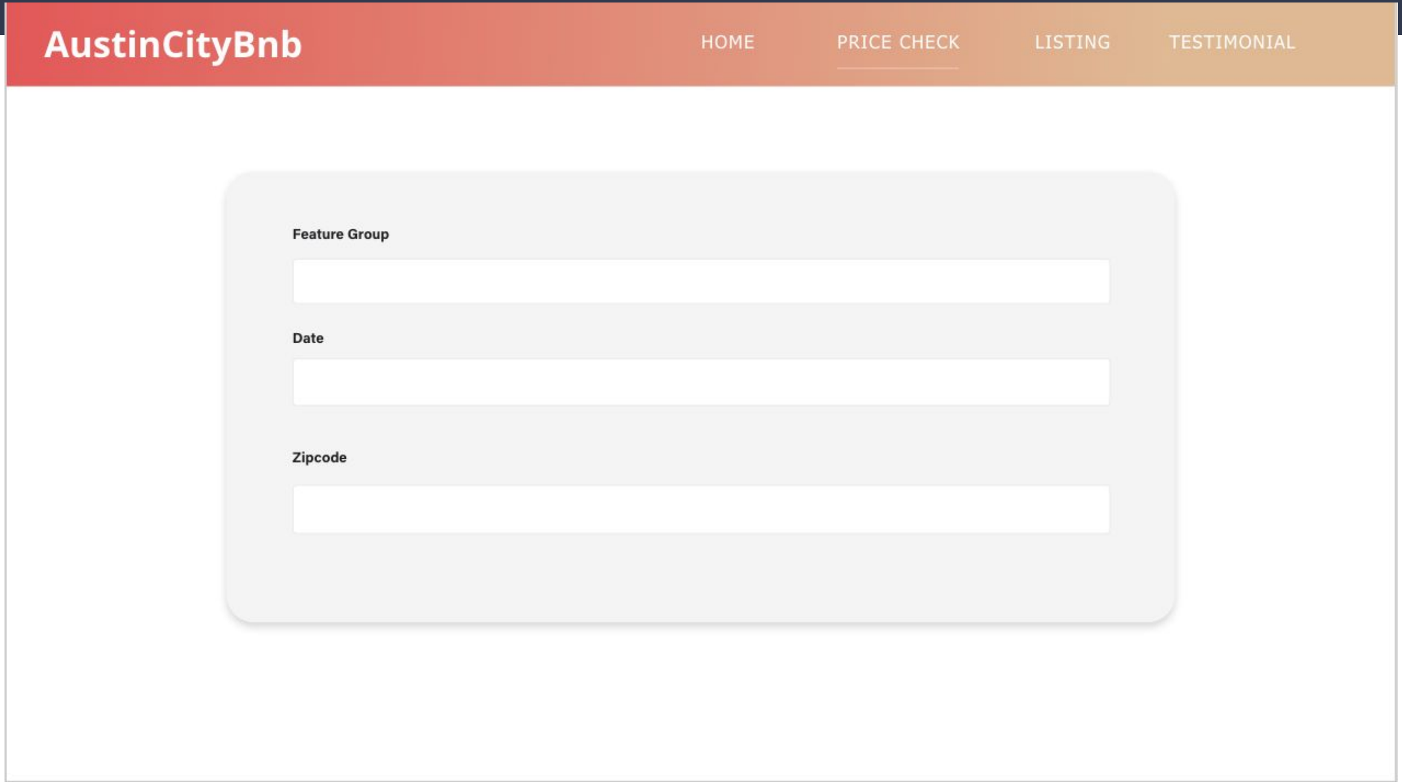
Column1	78,701	78,704	78,705	1	2	3	4	5
sse	13916804.66	122849415.61	3811965.50	9975721.89	4952579.56	27686560.55	10743054.71	87220269.06
variance	3730.52	11083.75	1952.43	3158.44	2225.44	5261.80	3277.66	9339.18
std dev	3.79	1.38	2.26	0.93	1.29	2.21	6.33	5.08

Dashboard

Host can check on the optimal price of the listings, if they choose to list for a day, a week, or a month ahead.

Our model will return the true price the host is getting from each day.

How much is your house worth during ACL?



Limitations

- Selection biases
- Six months of usable data
- Can’t test “spark” on real data
- Processing power
- Meaning of “available”
- My right arrow key doesn’t work

Key Insights

- Local events have a large monetary effect on a city through Airbnb activity
- Spark: Book now, save money

ACL Adds Value to Austin

Predicted Listing Price	NULL	ACL	ADDED VALUE PER LISTING	COUNT	ADDED VALUE BY GROUP	TOTAL ADDED VALUE
Overall	$229.52	$269.89	$40.37	4060	$163,902.20	$163,902
Group 1	$170.71	191.83	$21.12	1204	$25,428.48	
Group 2	$109.95	132.35	$22.40	728	$16,307.20	
Group 3	$207.88	256.82	$48.94	824	$40,326.56	
Group 4	$269.61	332.16	$62.55	652	$40,782.6	
Group 5	$422.85	521.72	$98.87	652	$64,463.24	$146,535
78701	$229.75	314.23	$84.48	657	$55,503.36	
78704	$206.23	271.64	$65.41	2891	$189,100.31	
78705	$194.80	203.07	$8.20	512	$4,198.40	$248,801

Supply-Side vs. Demand Side Behavior Matters

```
lmer1_coef<-fixef(log_lmer1)
lmer1price<-exp(lmer1_coef+5.346)
spark_coef<-fixef(log_lmer1_spark)
sparkprice1<-exp(spark_coef+5.346)

diff<-lmer1price-sparkprice1
as.data.frame(diff)
daily9<-cbind(diff, daily7)

summary(lm(formula=diff~trend, data=daily9))
```

```
Call:
lm(formula = diff ~ trend, data = daily9)

Residuals:
    Min      1Q  Median      3Q     Max
-38.689  -0.376   0.231   0.684   2.290

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.181460   0.465212   -2.54    0.012 *
trend        0.084607   0.004638   18.24   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.046 on 171 degrees of freedom
Multiple R-squared:  0.6606,	Adjusted R-squared:  0.6586
F-statistic: 332.8 on 1 and 171 DF,  p-value: < 2.2e-16
```

References

listings.csv.gz

calendar.csv.gz

https://stats.stackexchange.com/questions/13166/rs-lmer-cheat-sheet
https://ourcodingclub.github.io/2017/03/15/mixed-models.html#randomstr
https://cran.r-project.org/web/packages/lme4/vignettes/lmer.pdf
http://www.rensenieuwenhuis.nl/r-sessions-16-multilevel-model-specification-lme4/
https://freshbiostats.wordpress.com/2013/07/28/mixed-models-in-r-lme4-nlme-both/

Appendix

Dataset:

https://drive.google.com/open?id=1Apu2qeOJ1DnXllYHM6Py1n1xhByZNMsG

Cleaned Dataset:

https://drive.google.com/open?id=1jzfng3hZglqasmKnmeZSjd7-IX_1wLkT

Code:

https://drive.google.com/file/d/1QRPSawsEsY3kszKh7NQIM47q7hYv3ulp/view?usp=sharing