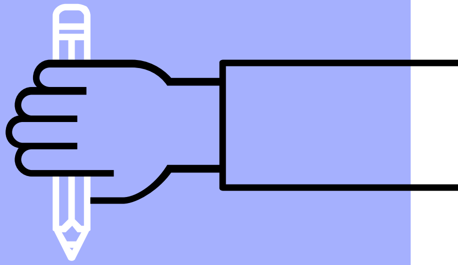
airbnb Listings

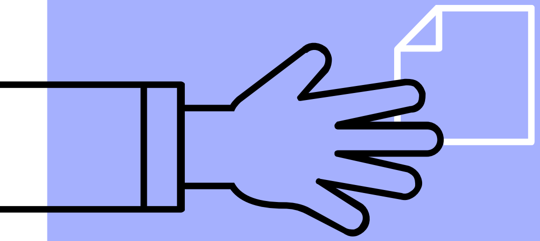
By: Chima, Harsha, Ritesh, Sangeetha, Tchamy

Agenda

- Airbnb Data Background
- Objective
- Data Consolidation and Cleaning
- Data Analysis
- Conclusions/Recommendations
- Team Project Process
- Open-ended questions

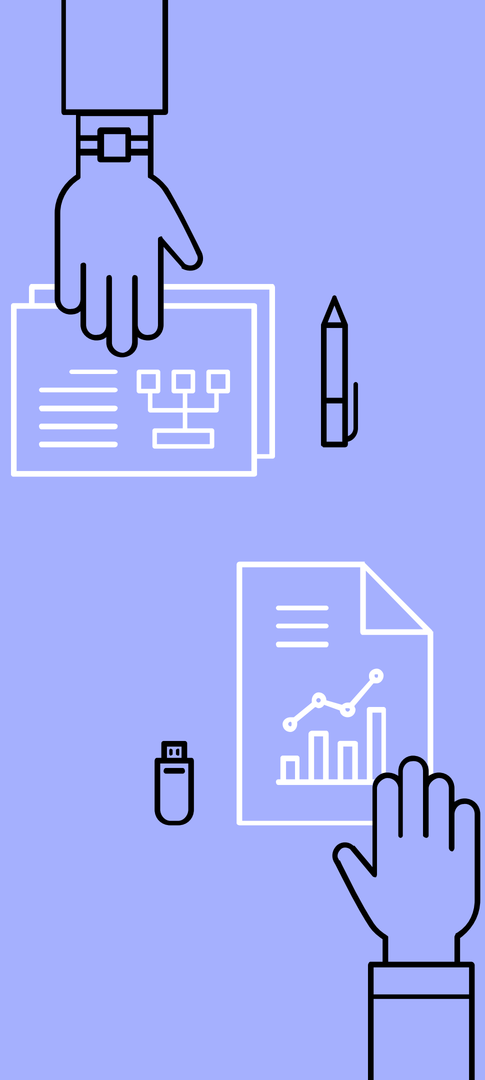


1. Airbnb Data Background



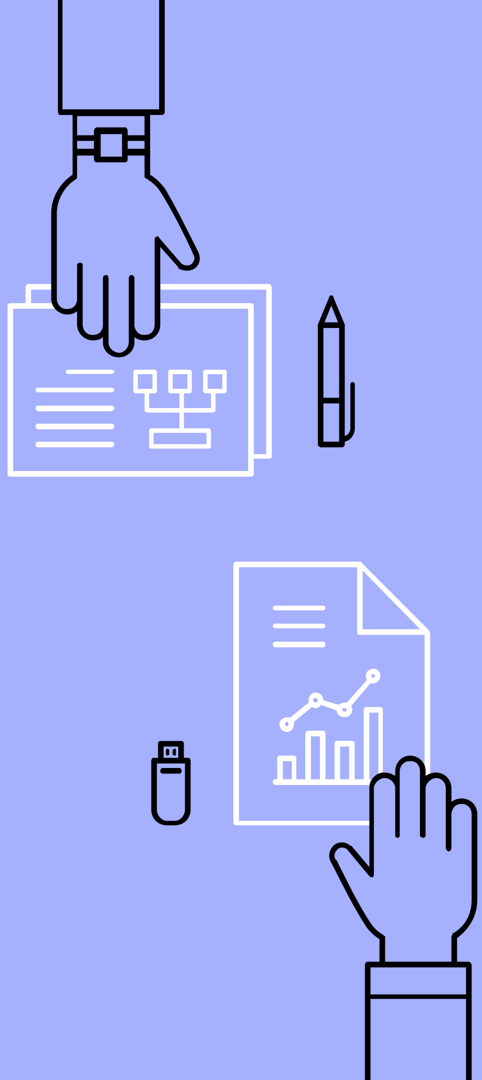
About Airbnb

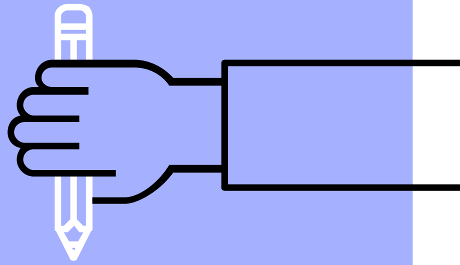
- Founded in 2008
- Airbnb is an online marketplace which allows private people rent out their space
- Over 7 million listings worldwide in over 100K cities and 190 countries



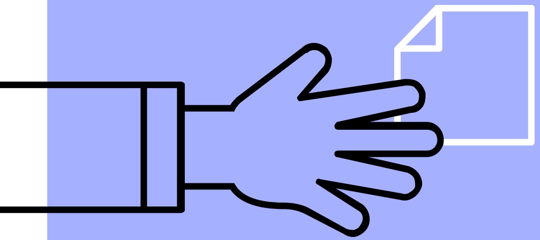
Airbnb Data Background

- Data has been published on Opendatasoft.
- The data has most recently been refreshed in July 2017.
- The dataset was compiled using a tool called Inside Airbnb
- Original dataset:
  - Included 494,594 records
  - Included measurement of important attributes like Host Response Rate, City, Host Response Time, Room Type, Bed Type, etc.



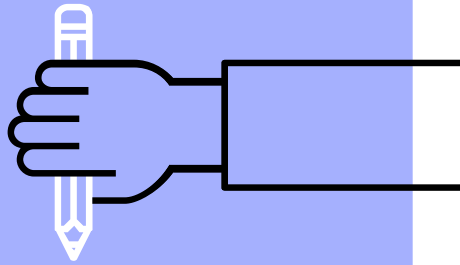


2.
Objective

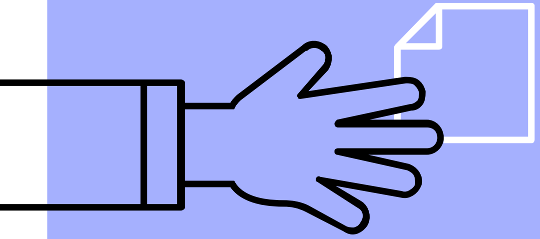


What recommendations can we make for a new host based on the following factors:

- Reviews Per Month*
- Number of Reviews*
- (Revenue)*
- Property Type
- Number of Listings
- Host Identification
- Host Popularity
- Cancellation Policy
- Response Time

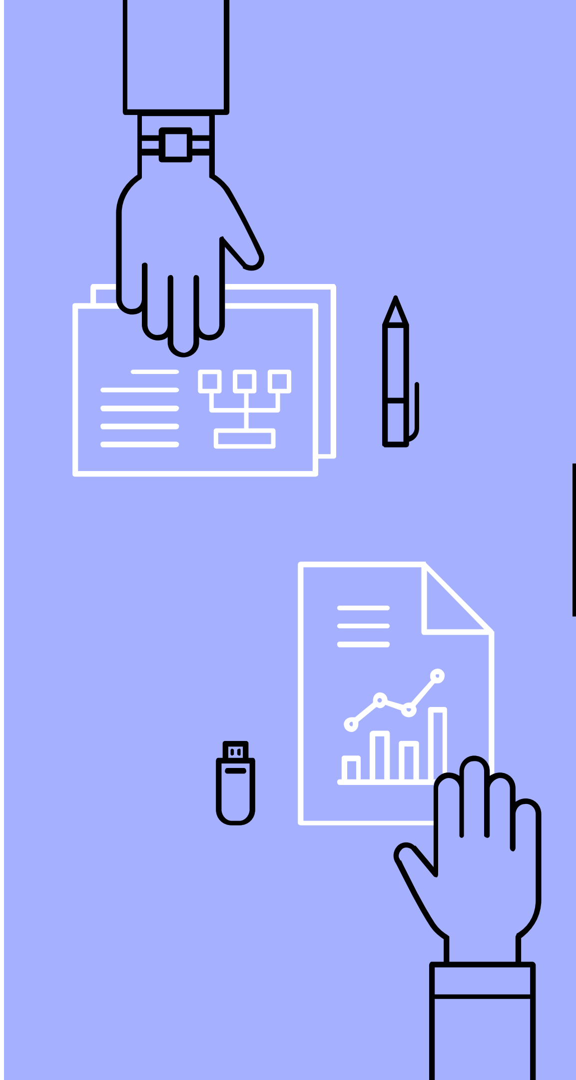


3.
Data Consolidation & Cleaning



Data Consolidation & Cleaning Steps

- Deleted columns that we deemed to be irrelevant
- Looked at the data types of all columns
    - Factorized columns by category (either ordinal/nominal)
    - Numbers became numeric
- Consolidated various category values for different column attributes to make a more comprehensive data set.
    - States- arranged all states to follow same capitalization format
    - Cancellation policy- standardized the format of categories (Strict, Moderate, Flexible)



R Code for Data Cleaning

Deleted columns that we deemed to be irrelevant

```r
airbnb = subset(airbnb, select = -c(license,host_verifications, house_rules, host_name,neighbourhood_cleansed,country,amenities,minimum_minimum_nights,maximum_minimum_nights,
                                    minimum_maximum_nights,maximum_maximum_nights,minimum_nights_avg_ntm,maximum_nights_avg_ntm,calendar_updated,has_availability,
                                    availability_30,availability_60,availability_90,calendar_last_scraped, number_of_reviews_ltm,first_review,last_review,requires_license,
                                    calculated_host_listings_count_entire_homes,calculated_host_listings_count_private_rooms,calculated_host_listings_count_shared_rooms) )
```

Consolidated several values to have less unique entries in each column

```r
#Limit cancellation policy to three unique entries

cancellation=gsub("super_strict_30", "strict",airbnb$cancellation_policy)
cancellation = data.table(cancellation)
airbnb$cancellation_policy = cancellation

cancellation=gsub("super_strict_60", "strict",airbnb$cancellation_policy)
cancellation = data.table(cancellation)
airbnb$cancellation_policy = cancellation
```

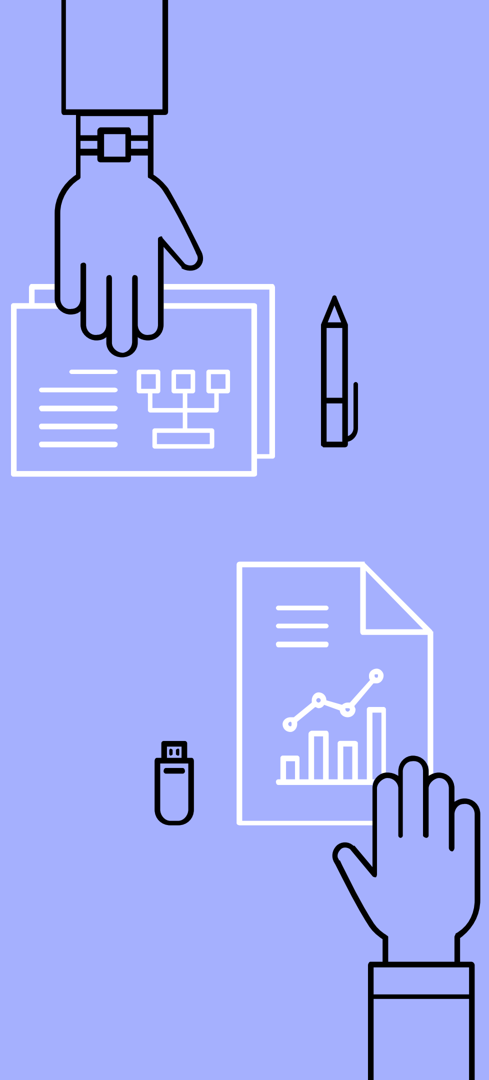
```r
#consolidate all individual states
airbnb$state = gsub("Ma","MA",airbnb$state)
airbnb$state = gsub("Hi","HI",airbnb$state)
airbnb$state = gsub("Ny","NY",airbnb$state)
airbnb$state = gsub("ny","NY",airbnb$state)
airbnb$state = gsub("New York","NY",airbnb$state)
```

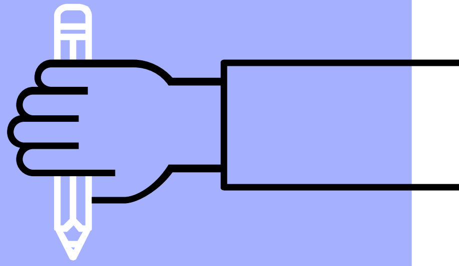
Turned appropriate columns to ordinal or nominal factors

```r
airbnb$cancellation_policy = factor(airbnb$cancellation_policy, levels = c('flexible','moderate','strict') ordered = T)
#Turning columns to factors
airbnb$host_is_superhost = factor(airbnb$host_is_superhost)
airbnb$host_has_profile_pic = factor(airbnb$host_has_profile_pic)
```

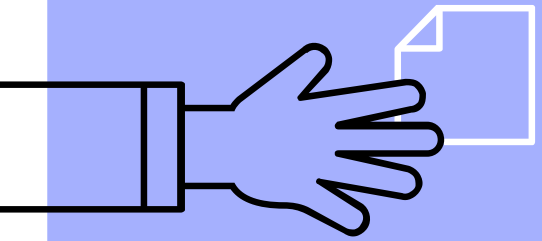
Revenue Calculation

```
#How many bookings does a host have per year?
bookings_year = rev_airbnb$reviews_per_month*12*2

#Adding the bookings per year to the data table
rev_airbnb[,bookings_year:=bookings_year]

#Calculating how many nights a year a host rents out their place
occupancy = rev_airbnb$bookings_year*3

#Adding the occupancy as a column
rev_airbnb[,occupancy_year := occupancy]

#Multiplying the per night price with the occupied nights to get revenue per year
revenue = rev_airbnb$price*rev_airbnb$occupancy_year

#Adding the revenue as a new column
rev_airbnb$revenue = revenue
```





4.
Data Analysis



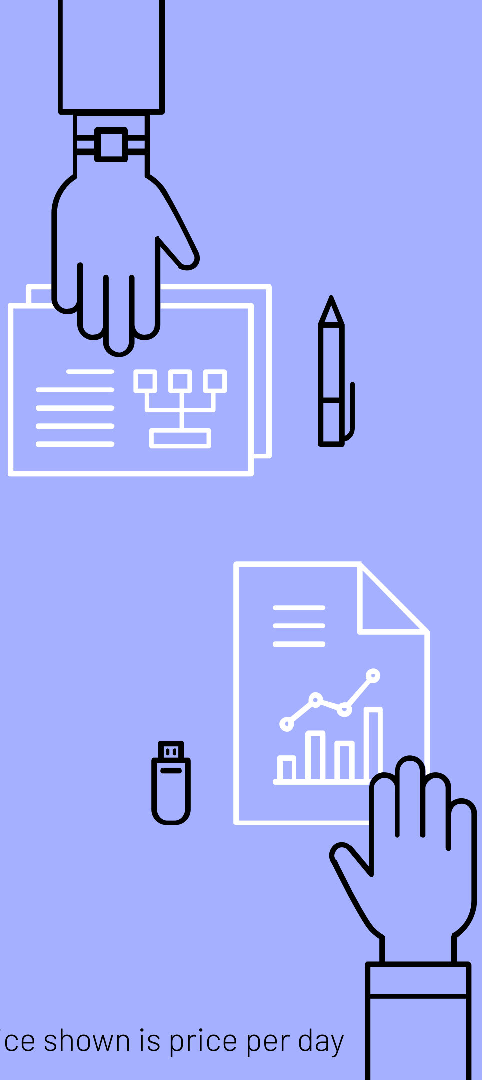
Property Type

Property Type Popularity

Property Type	
Apartment	6,940,017
Condominium	2,943,126
House	862,883
Serviced apartment	263,030
Townhouse	163,128
Resort	121,793
Other	92,288
Hotel	86,531
Villa	55,138
Guest suite	54,448
Boutique hotel	50,719
Loft	33,425
Cottage	21,424
Guesthouse	21,055
Hostel	18,411
Aparthotel	16,035
Bungalow	12,743
Bed and breakfast	12,345
Cabin	4,566
Camper/RV	2,426
Tiny house	1,869
Tent	1,413
Boat	1,196
Farm stay	819

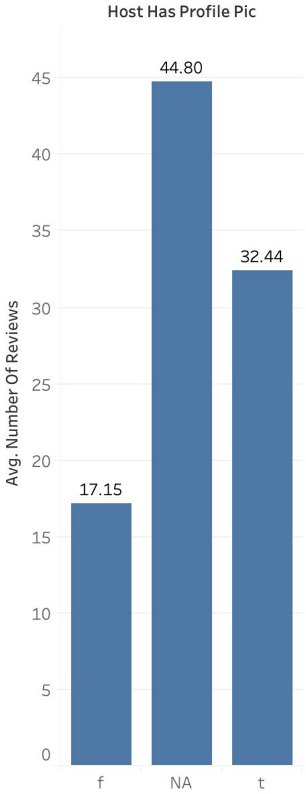
Median Price by Property Type

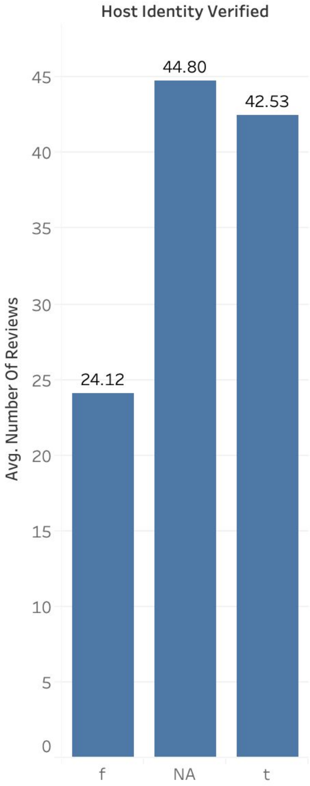
Property Type	
Villa	400.0
Resort	260.0
Serviced apartment	198.5
Condominium	174.0
Boutique hotel	174.0
Hotel	155.0
Other	150.0
Cottage	150.0
Loft	145.0
Townhouse	120.0
House	120.0
Bungalow	120.0
Apartment	120.0
Bed and breakfast	114.0
Guesthouse	100.0
Guest suite	95.0

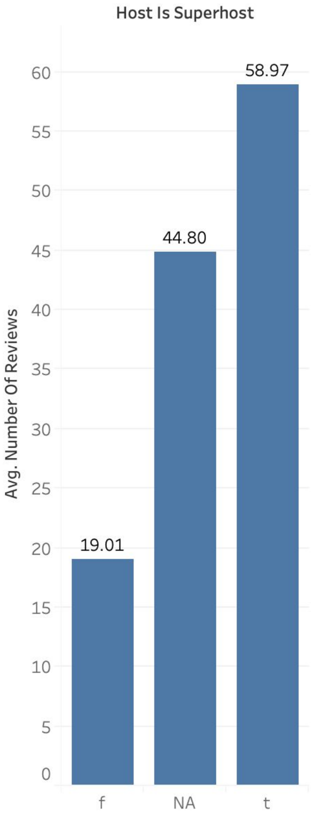


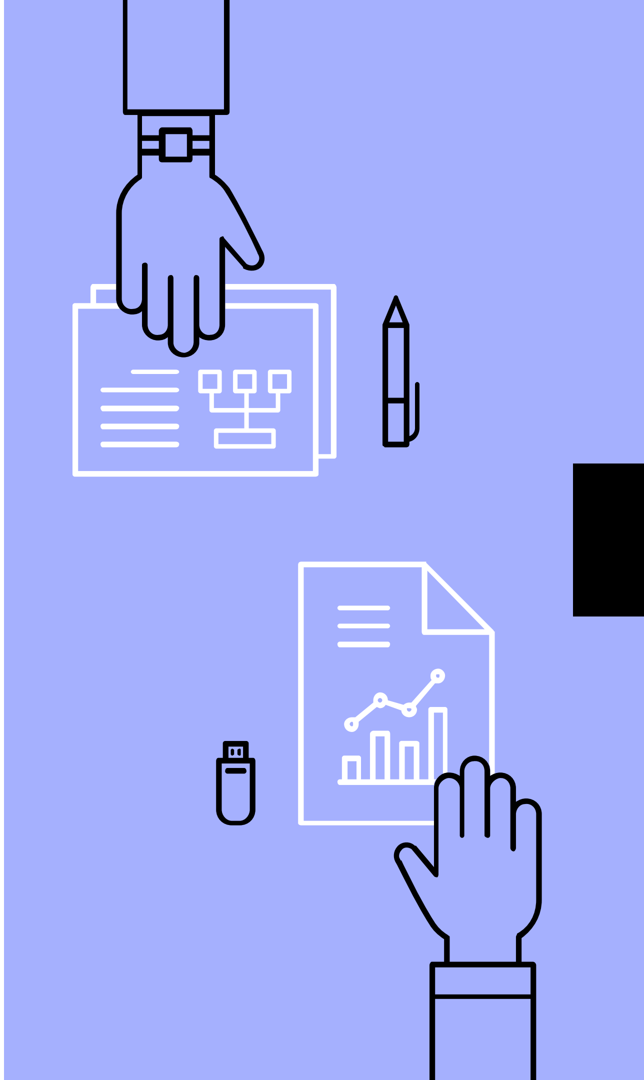
*Price shown is price per day

Host Data

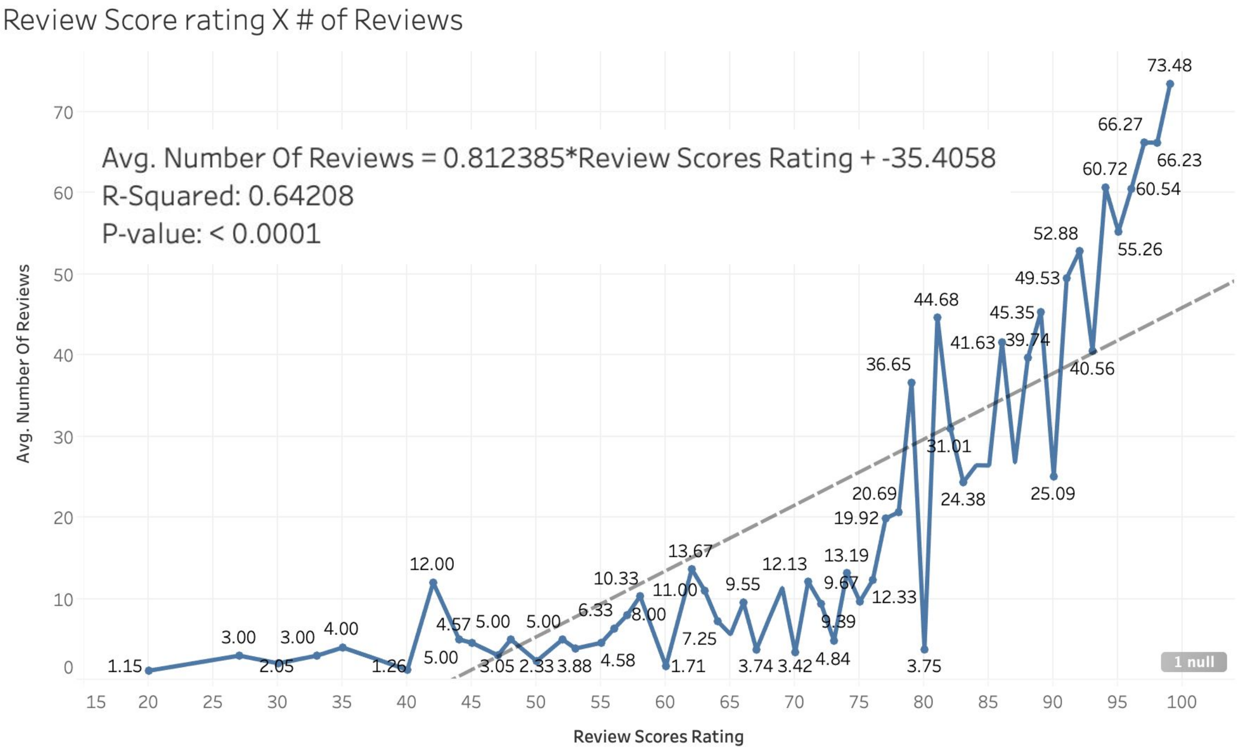

Host Has Profile Pic
Avg. Number Of Reviews
0
5
10
15
20
25
30
35
40
45
17.15
44.80
32.44
f
NA
t

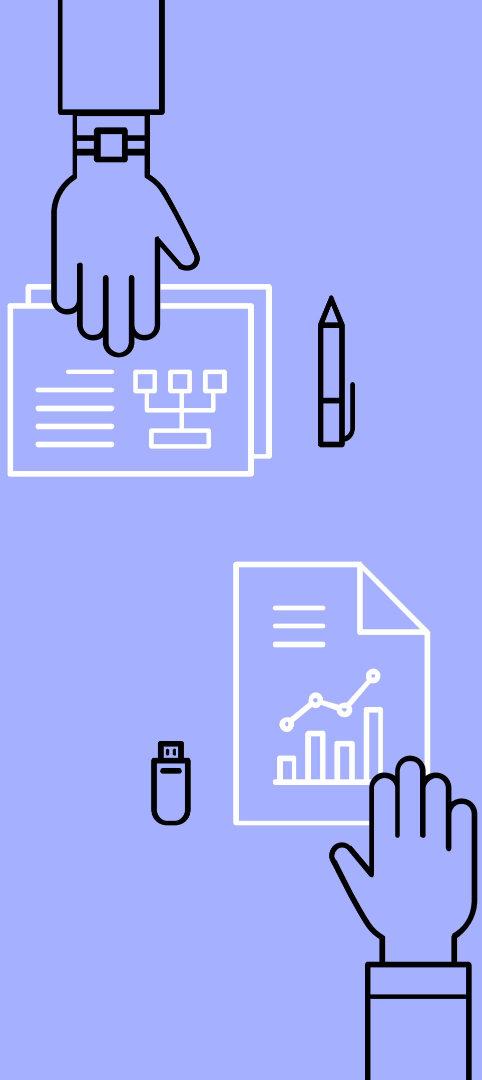

Host Identity Verified
Avg. Number Of Reviews
0
5
10
15
20
25
30
35
40
45
24.12
44.80
42.53
f
NA
t


Host Is Superhost
Avg. Number Of Reviews
0
5
10
15
20
25
30
35
40
45
50
55
60
19.01
44.80
58.97
f
NA
t

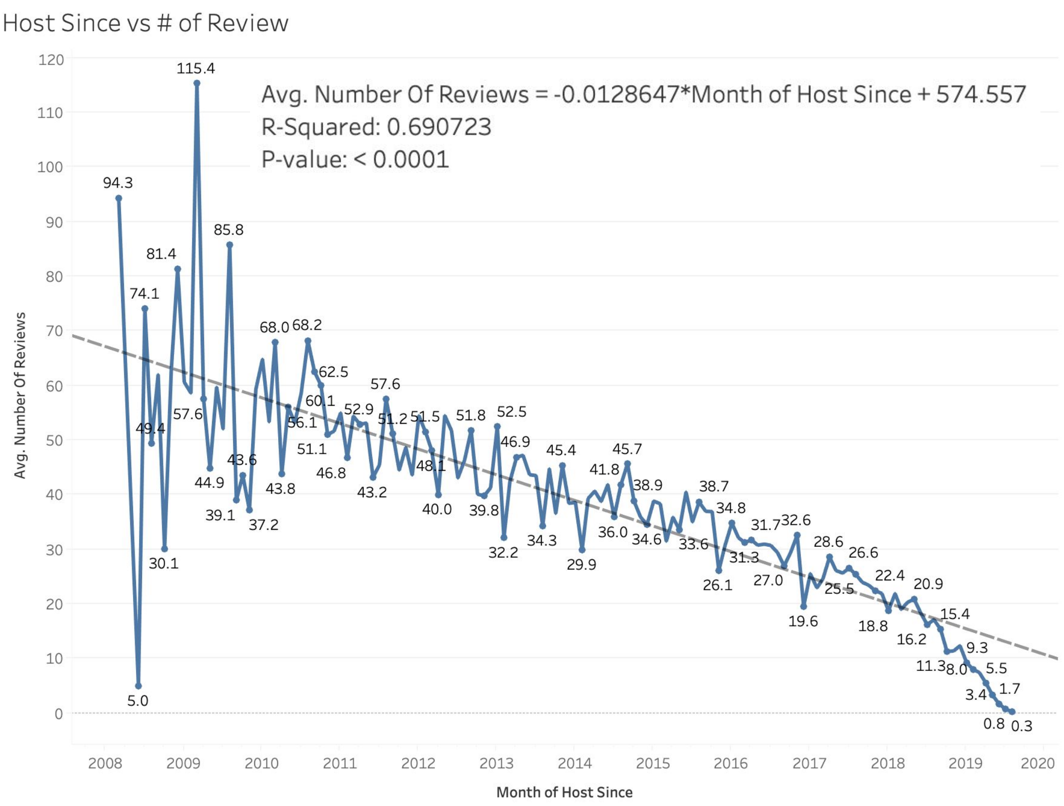


Review Score Rating x # of Reviews



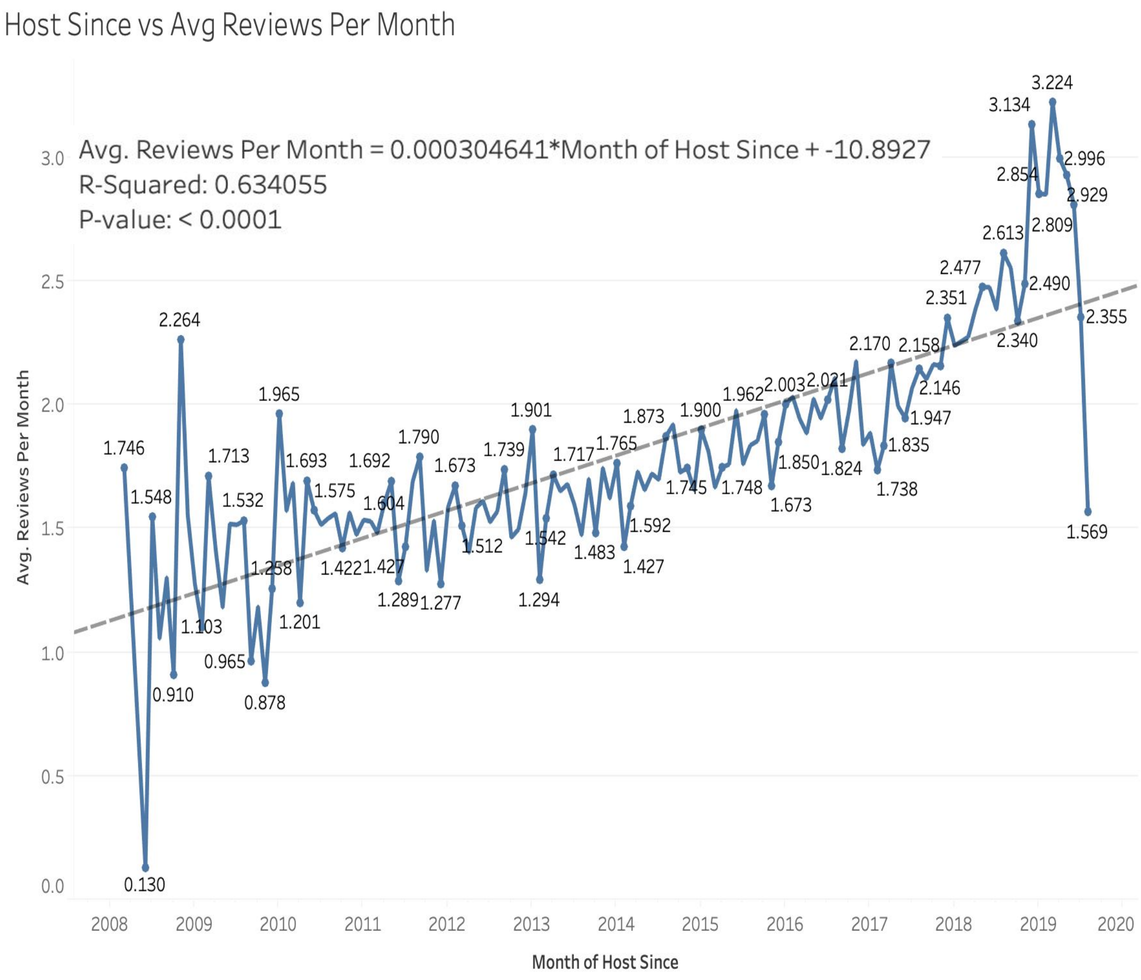


Time as a Host vs # of Reviews



Summary

Count:	136
AVG(Number Of Revie...	
Sum:	5,359.9
Average:	39.1
Minimum:	0.3
Maximum:	115.4
Median:	38.9



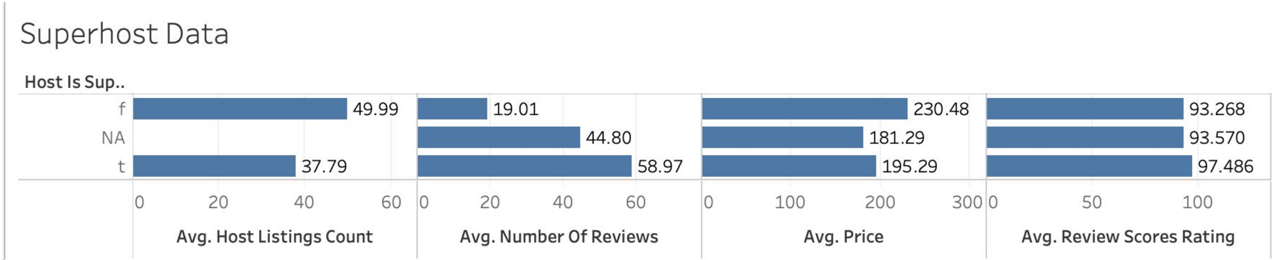
Summary

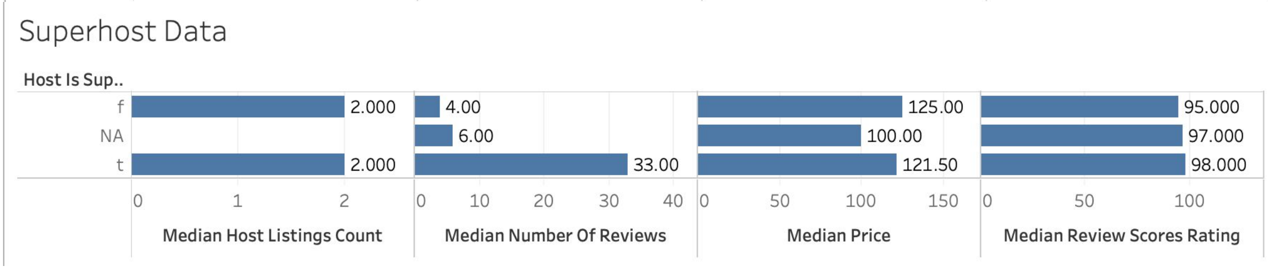
Count:	136
AVG(Reviews Per Mont...	
Sum:	244.640
Average:	1.786
Minimum:	0.130
Maximum:	3.224
Median:	1.713

This may look contrary to what some would believe however there is an explanation

New Airbnb hosts are more likely to be active

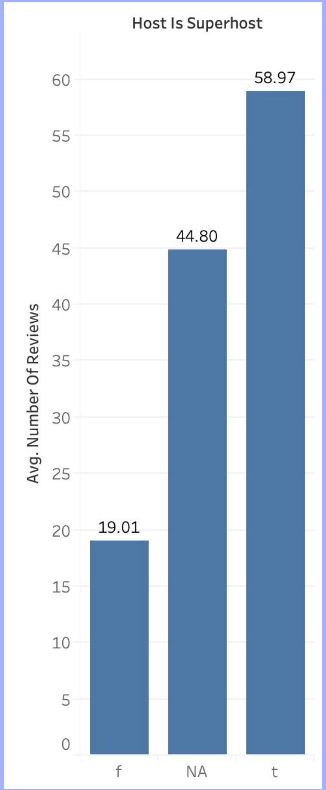
Superhost Data Comparison

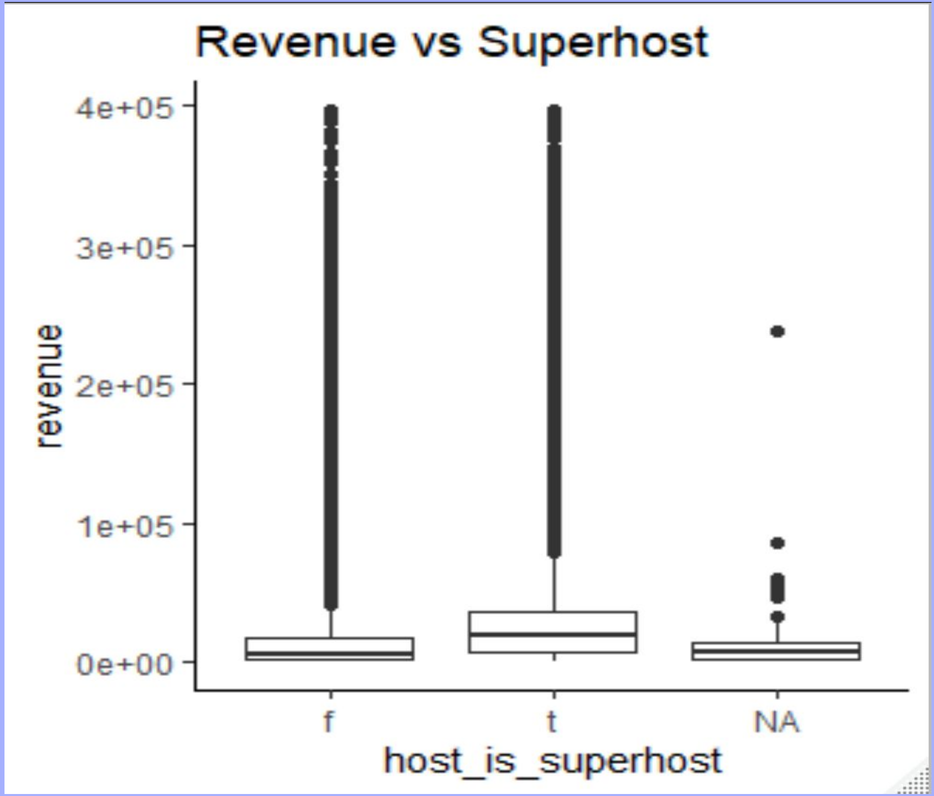




Above shows some key differences concerning the superhost delineation both the average and the median.

To the right shows the average # of reviews separated by status (Superhost vs non superhost)



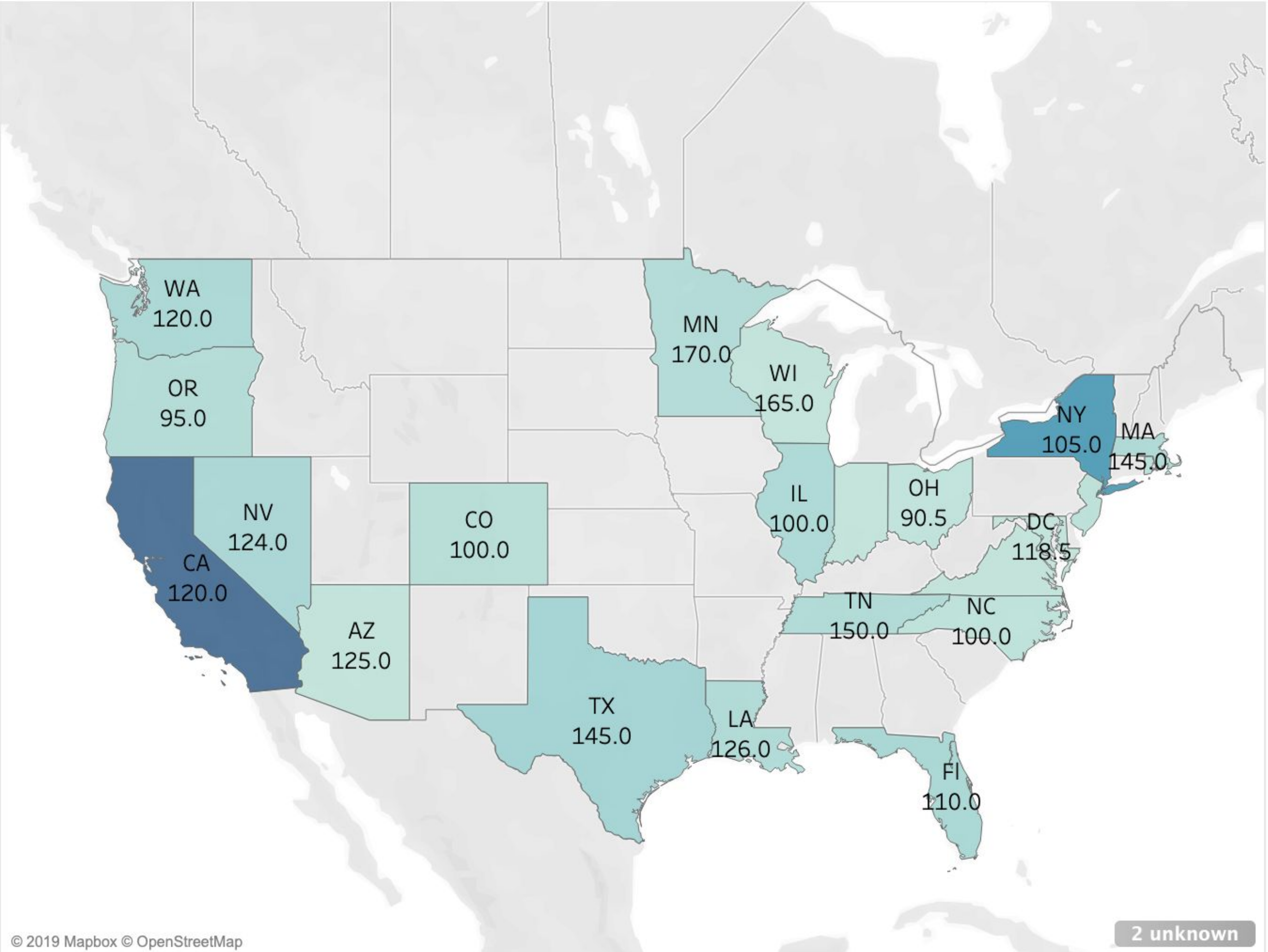



```
> aggregate(corairbnb$revenue, list(Superhost = corairbnb$host_is_superhost), summary)
  Superhost      x.Min.  x.1st Qu.   x.Median     x.Mean  x.3rd Qu.     x.Max.    x.NA's
1         f        0.00    1958.40    6426.00   15356.24   17790.48 4557600.00  44036.00
2         t        0.00    8152.56   19152.00   30304.47   36432.00 6271200.00   5385.00
```

C.A

TABLEAU DEMONSTRATION

Map Breakdown - Median Price Shown



Summary

Count:	25
AVG(Accommodates)	
Sum:	125.072
Average:	4.632
Minimum:	2.000
Maximum:	8.000
Median:	4.526
AVG(Bedrooms)	
Sum:	47.156
Average:	1.747
Minimum:	1.000
Maximum:	3.000
Median:	1.667
AVG(Square Feet)	
Sum:	19,565.4
Average:	1,029.8
Minimum:	695.3
Maximum:	1,630.0
Median:	944.9
MEDIAN(Price)	
Sum:	4,835.0
Average:	179.1
Minimum:	90.0
Maximum:	789.0
Median:	124.0
SUM(Number of Recor...	
Sum:	256,814
Average:	9,511.63
Minimum:	1
Maximum:	77,880
Median:	5,833.00

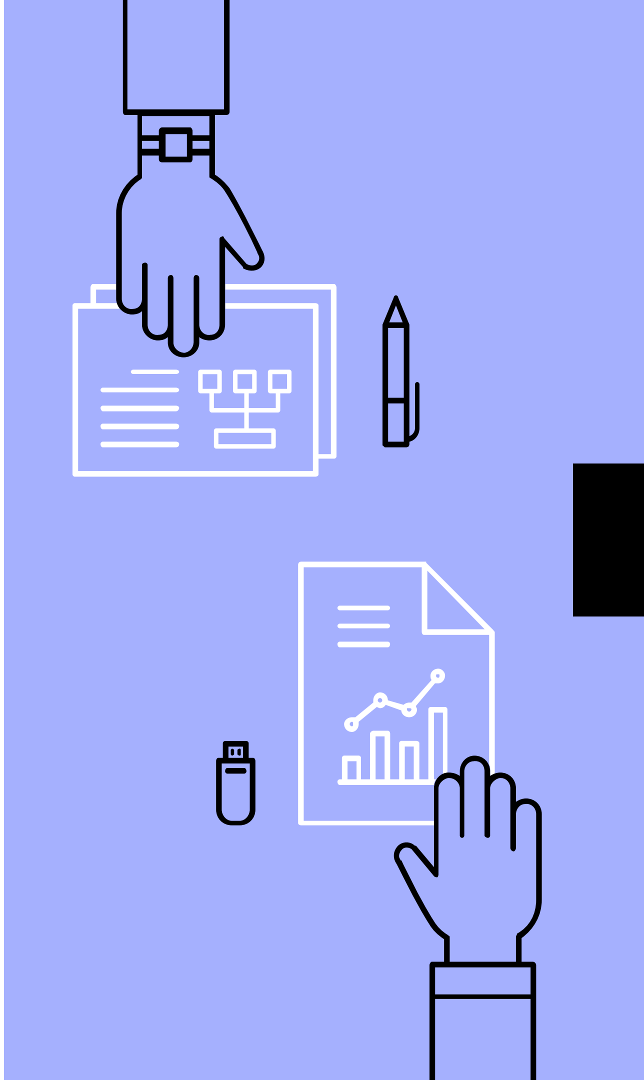
SUM(Number of Recor...

20

1 77,880

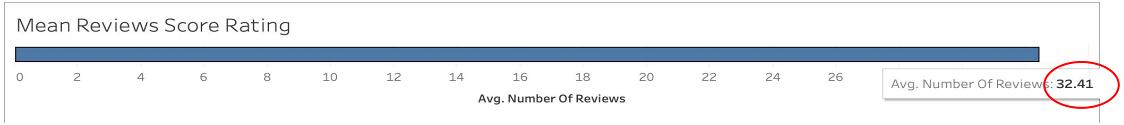
Host Popularity

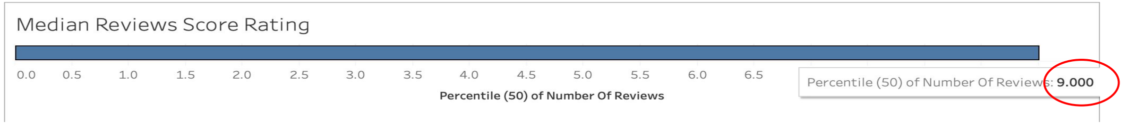
- How do we define what's "popular"?
  - Hosts with high review rating vs. number of reviews
  - Develop a baseline threshold number of reviews based on our dataset to determine what can be considered a "meaningful" review rating
- What are we trying to conclude from this?
  - How many people do you want to try and host for your review score to be significant enough?

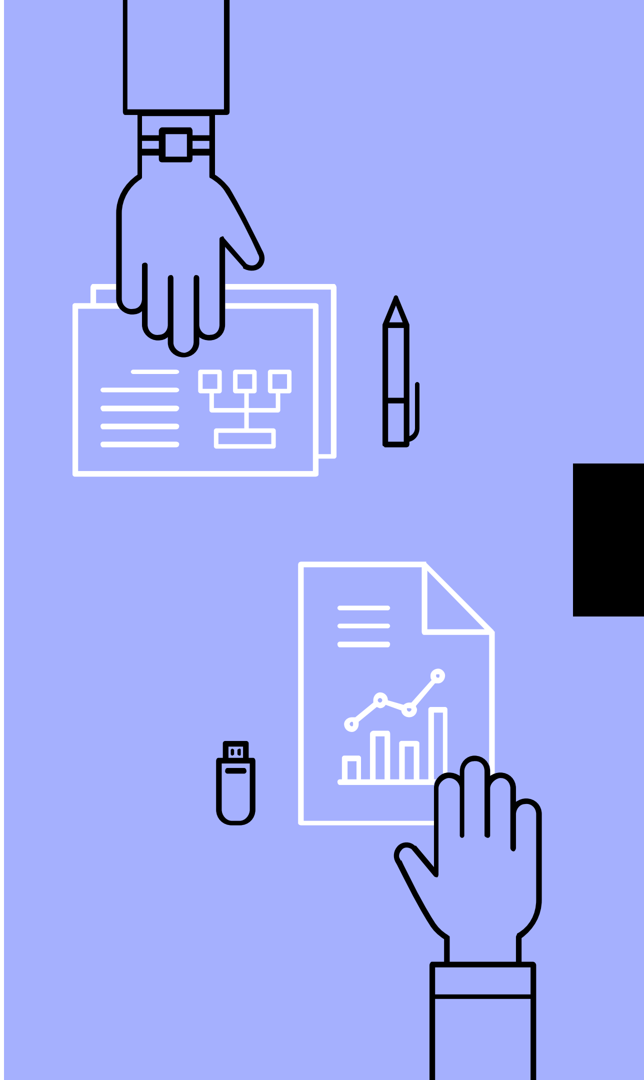


Host Popularity (cont.)

- Summary Statistics- number of reviews
  - We want to determine which number of review scores can be considered “valid” enough.
    - Ex. if a property has a high review score, but only 1 rating, how do we interpret that?
  - Median number of reviews is more accurate of a measure than the mean number of reviews
    - Why? Mean is too heavily skewed to accurately represent the data.
      - Mean = 32.41
      - Median = 9 (50th percentile)
      - ***Analyze all scores with high ratings, that have at least 9 reviews.***

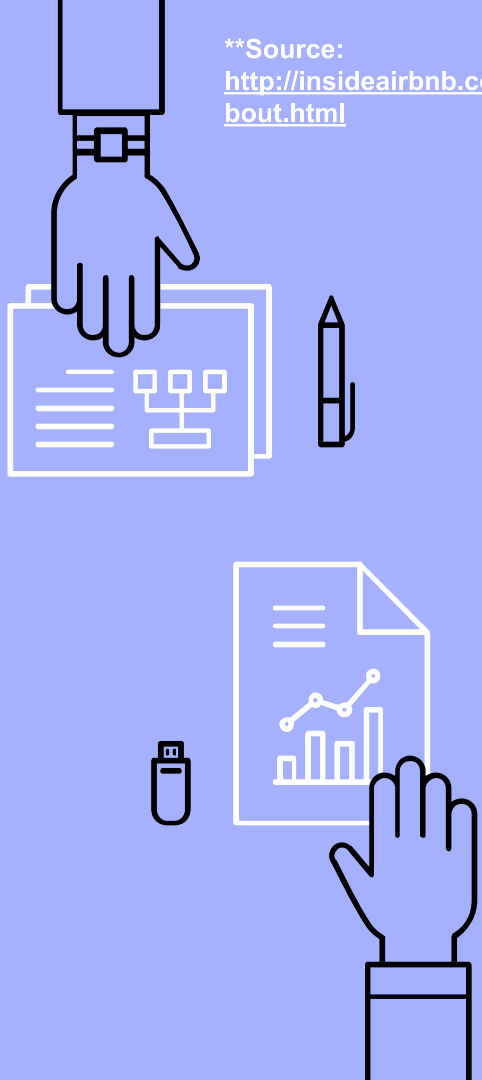






Assumptions

- “A Review Rate of 50% is used to convert **reviews** to **estimated bookings**”**
    - 50% of the reviews that are collected indicates that the estimated bookings is double the number listed.
        - If you have 10 reviews, then the assumption is that 20 people have stayed at the location.



**Source: http://insideairbnb.com/about.html

Summary Statistics

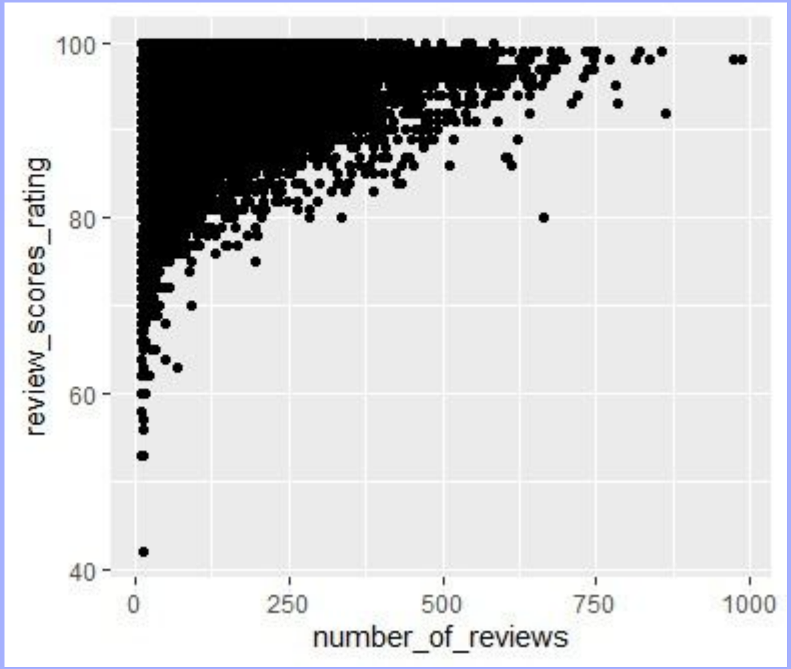
Coding Output for Number of Reviews

```
count    256816.000000
mean         32.405606
std          57.528697
min           0.000000
25%           1.000000
50%           9.000000
75%          38.000000
max         987.000000
Name: number_of_reviews, dtype: float64
```

Coding Output for Review Scores Rating

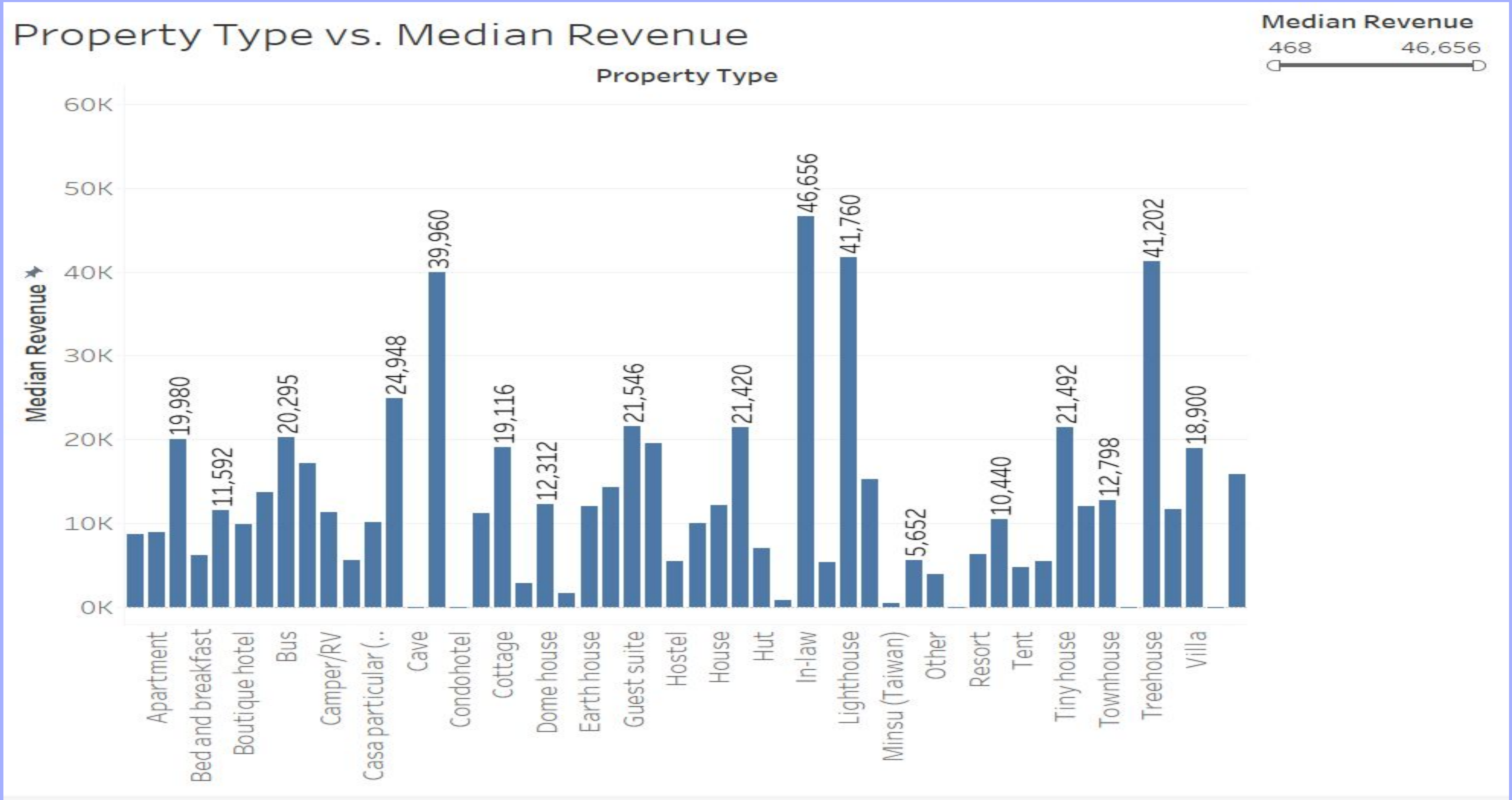
```
count    204455.000000
mean         94.930381
std           7.710480
min          20.000000
25%          93.000000
50%          97.000000
75%         100.000000
max         100.000000
Name: review_scores_rating, dtype: float64
```

Scatter Plot



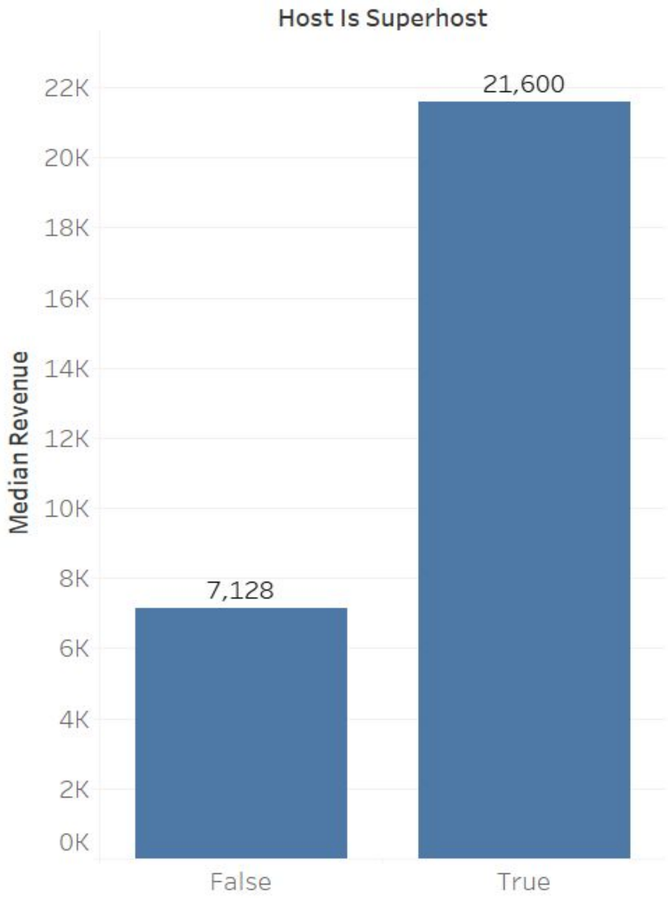
This scatter plot shows the review scores rating by the number of reviews, with a filter of (number of reviews > 9). This plot is very similar in comparison to the same plot without the filter, just with much less outliers.

We chose 9, the median number of reviews, as our threshold to determine significance.

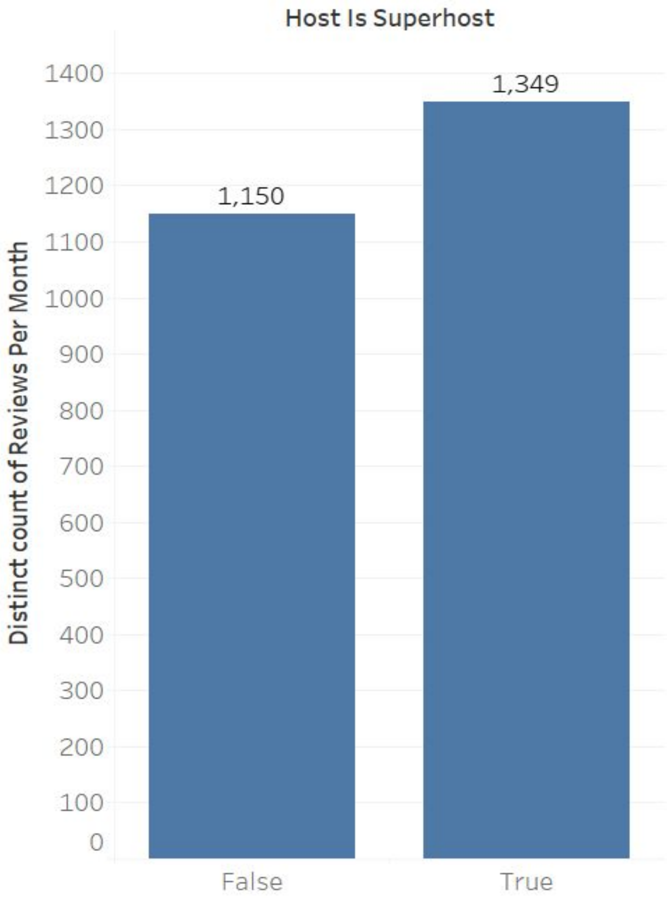


Property Type vs. Median Revenue
Median Revenue
468
46,656
Property Type
Median Revenue
60K
50K
40K
30K
20K
10K
0K
19,980
11,592
20,295
24,948
39,960
19,116
12,312
21,546
21,420
46,656
41,760
5,652
10,440
21,492
12,798
41,202
18,900
Apartment
Bed and breakfast
Boutique hotel
Bus
Camper/RV
Casa particular (..
Cave
Condohotel
Cottage
Dome house
Earth house
Guest suite
Hostel
House
Hut
In-law
Lighthouse
Minsu (Taiwan)
Other
Resort
Tent
Tiny house
Townhouse
Treehouse
Villa

Host Is Superhost vs. Median Revenue



Host Is Superhost vs Reviews per month

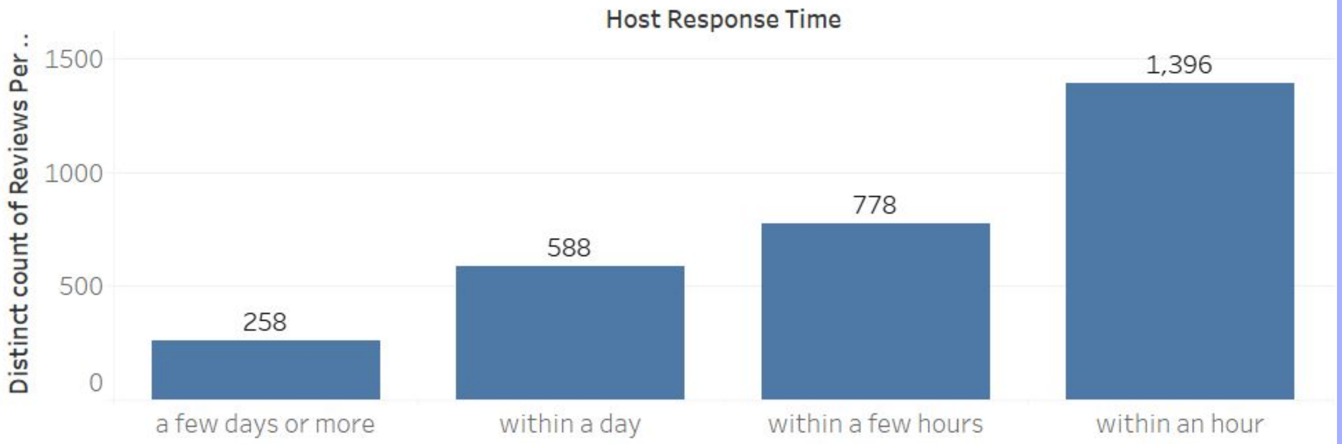


Definite more median revenue if host is superhost.

Superhost has more reviews per month.

So how to increase reviews per month?

Reviews per month vs. Host Response Time



Reviews per month vs. Cancellation Policy

Directly proportional.

Sooner the response, more distinct reviews per month.

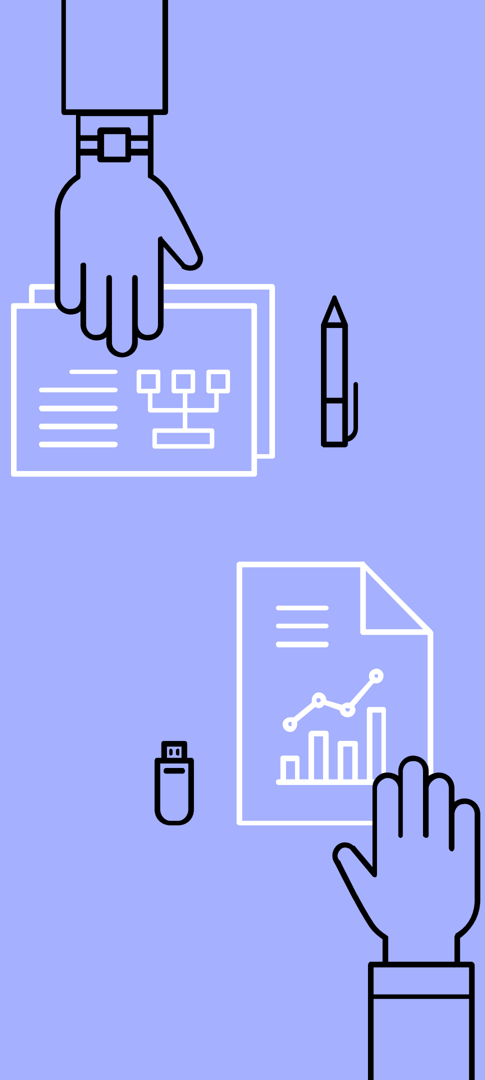
Not strict for sure.

Moderate works the best.

Hypothesis Testing

H_o -> Reviews per month directly affect Host's revenue.

H_A -> Reviews per month do not directly affect Host's revenue.



Simple Linear Regression

```
> summary(model2)

Call:
lm(formula = revenue ~ reviews_per_month, data = airbnb)

Residuals:
    Min      1Q  Median      3Q     Max 
-235460   -8461   -4147      77 6188576 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        4004.31     221.79   18.05   <2e-16 ***
reviews_per_month  9026.40      79.69  113.27   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 73760 on 207376 degrees of freedom
  (49438 observations deleted due to missingness)
Multiple R-squared:  0.05826,	Adjusted R-squared:  0.05826 
F-statistic: 1.283e+04 on 1 and 207376 DF,  p-value: < 2.2e-16
```

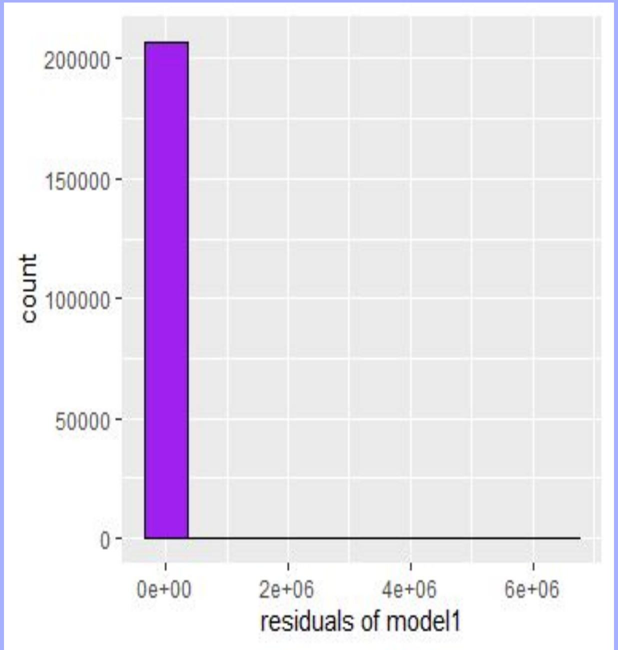
$Y = 4004 + 9026X$
0 reviews per month, 4004 revenue.
With 1 increment review per month, revenue would increase by 9026.

P-value < 0.05

There is statistically significant evidence to reject our Ho.

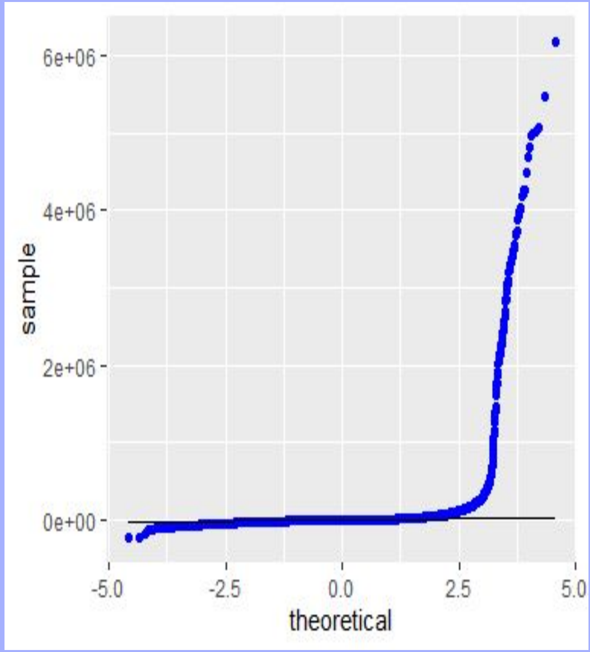
Check for LINE assumptions:
L- Linearity
I- Independence
N- Normality
E- equal variance

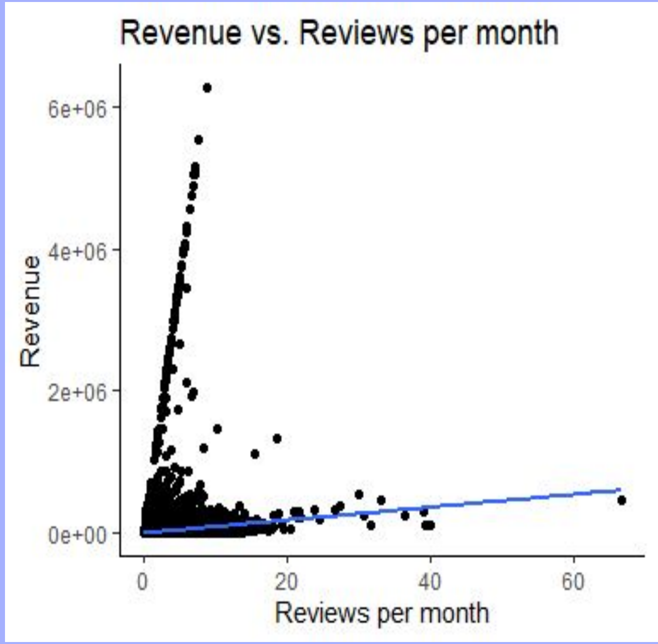
Linearity and Independence violated




```
> cor.test(airbnb$reviews_per_month,airbnb$revenue)

	Pearson's product-moment correlation

data:  airbnb$reviews_per_month and airbnb$revenue
t = 113.27, df = 207376, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2373180 0.2454244
sample estimates:
      cor
0.2413754
```

Normality and Equal Variance violated





Simple linear regression cannot be performed on this data set.

Reviews per month doesn't linearly affect Host's revenue.

Can we predict the Revenue for a new Host? YES!

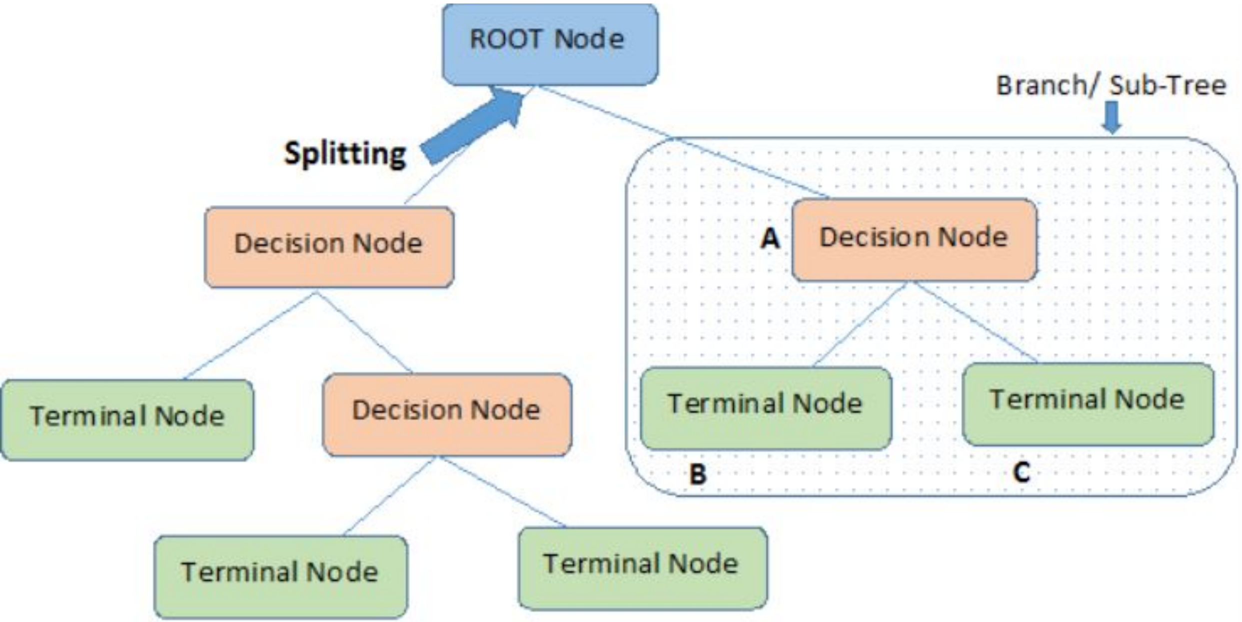
Using Machine Learning Algorithms!

- Fill the missing values in the Dataset
- Impute the NaNs using 0's or different strategies: mean imputation or over sampling
- Split the Dataset into Train and Test sets
- Fit the Train set and predict on the Test set.

VOILA!

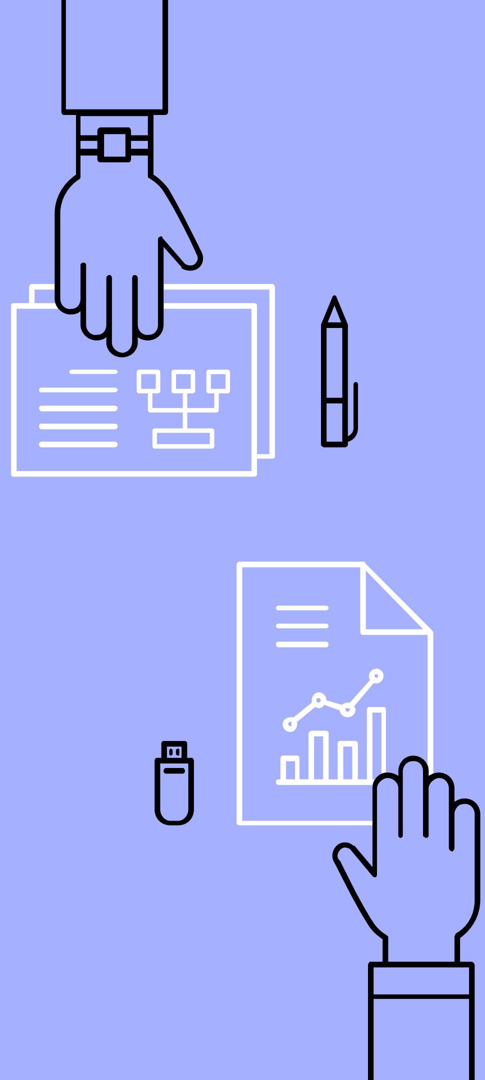
Only Dropped:'host_since','host_neighbourhood','city','state','zipcode', 'market'

Can we predict the Revenue for a new Host? YES!



$$\textit{Standard Deviation} = S = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}} = 9.3$$

$$\textit{Coeffeicient of Variation} = CV = \frac{S}{\bar{x}} * 100\%$$



Multivariate LR

	Actual Values	Predicted Values
0	44553.60	1555.775368
1	0.00	-28119.843336
2	2044.80	3253.240119
3	1800.00	-2498.618204
4	21600.00	18464.196943
5	1359.36	-1560.207537
6	13770.00	13066.305598
7	0.00	-14544.039844
8	74188.80	68767.048271
9	3733.20	2095.308750

LR r2 = 30.119202385213

Decision Trees

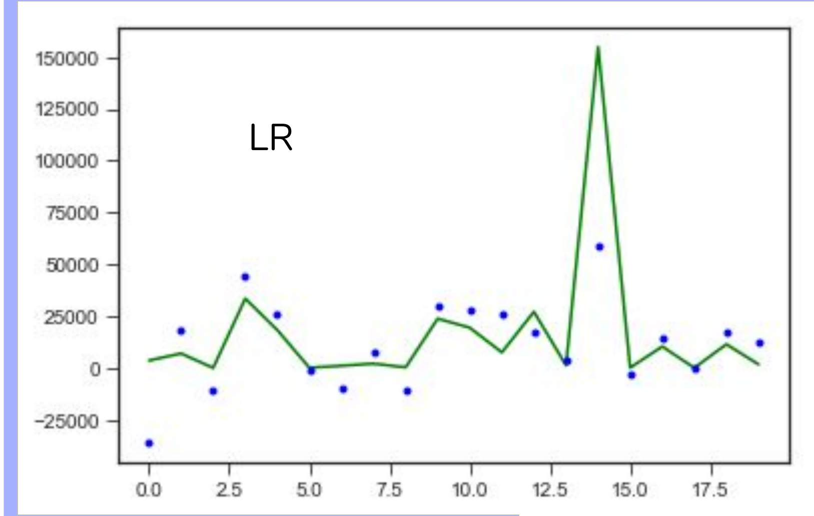
	Actual Values	Predicted Values
0	44553.60	46311.690000
1	0.00	0.000000
2	2044.80	1985.923636
3	1800.00	1800.000000
4	21600.00	21615.615000
5	1359.36	1408.338947
6	13770.00	13892.419459
7	0.00	0.000000
8	74188.80	72122.498182
9	3733.20	3690.595862

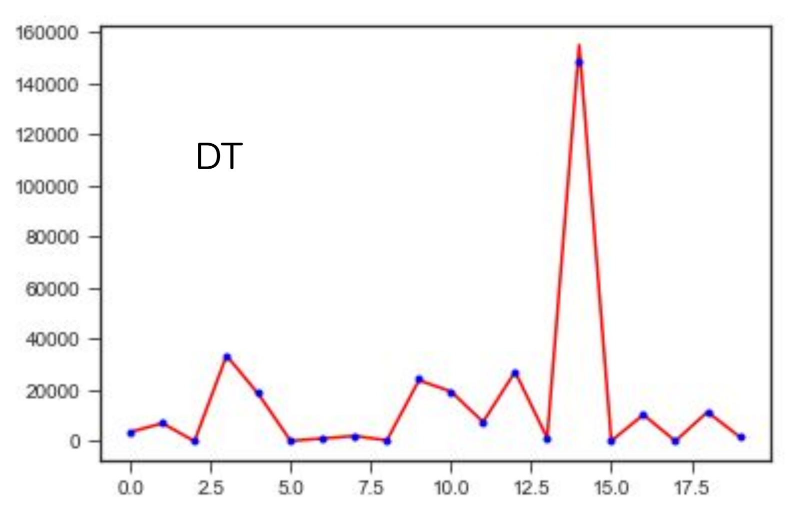
DT r2 =95.534158302556

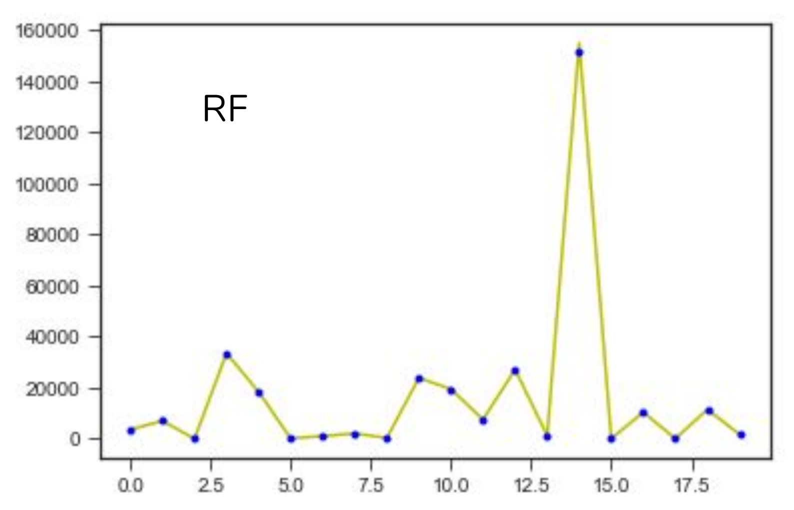
Random Forest

	Actual Values	Predicted Values
0	3598.56	3607.632
1	6996.24	7016.112
2	0.00	0.000
3	33480.00	33444.000
4	18316.80	18168.624
5	115.20	115.200
6	1069.20	1053.216
7	2070.00	2070.000
8	322.56	327.744
9	23803.20	23954.040

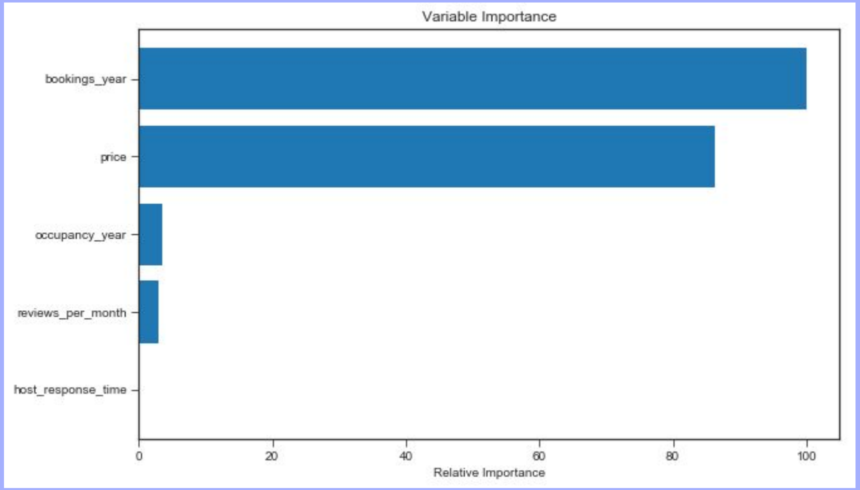
RF r2= 99.09899504600

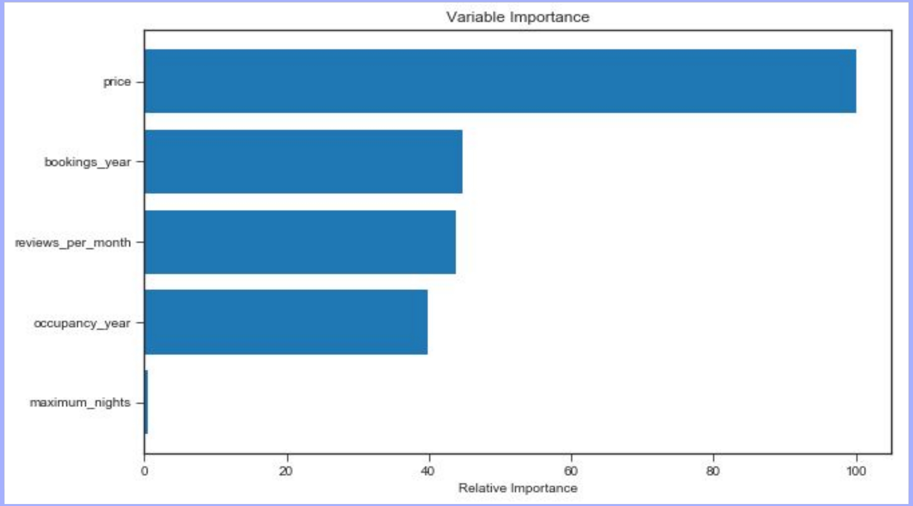

LR
150000
125000
100000
75000
50000
25000
0
-25000
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5


DT
160000
140000
120000
100000
80000
60000
40000
20000
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5

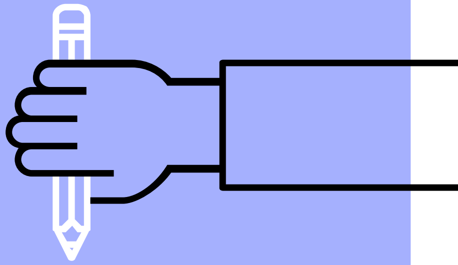

RF
160000
140000
120000
100000
80000
60000
40000
20000
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5

Variable Significance

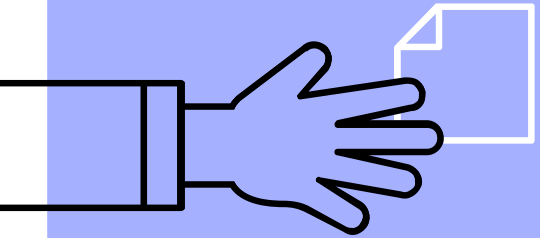




Reviews per month weigh more heavily in Random Forest.

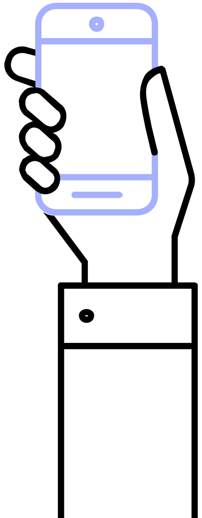


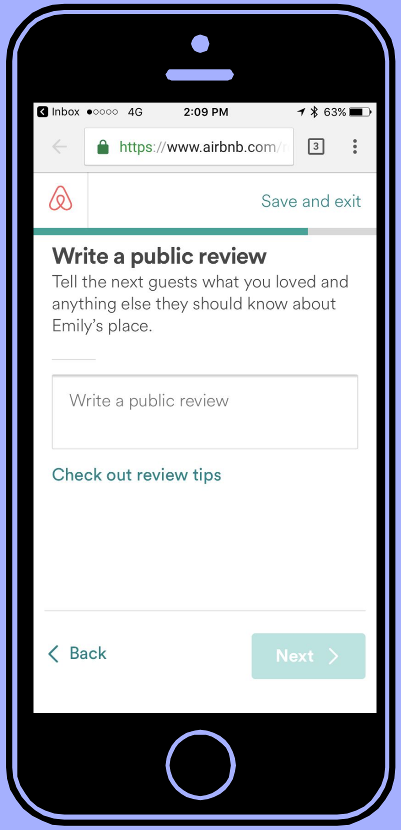
5.
Conclusions & Recommendations



Recommendations:

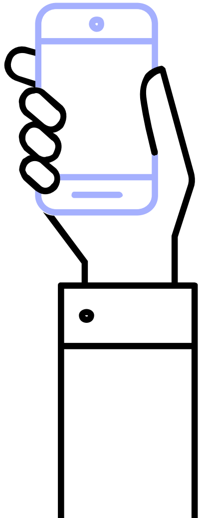
- **Encourage Positive Reviews**
  - More reviews makes you more reputable!

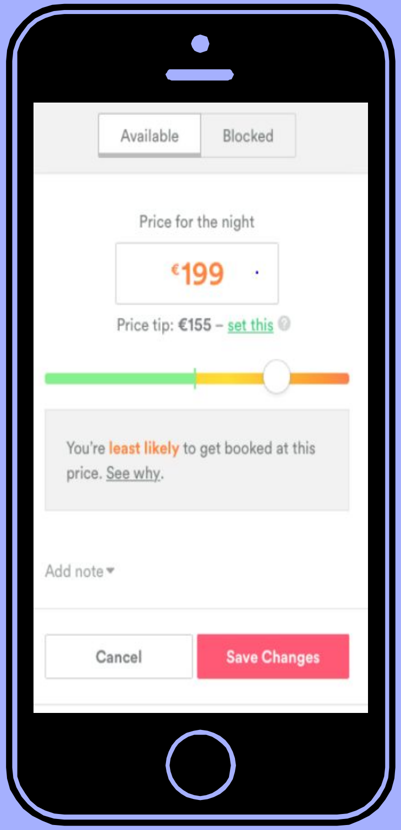




Recommendations:

- **Encourage Positive Reviews**
  - More reviews makes you more reputable!
- **Have Competitive Pricing**
  - Research your neighbors!



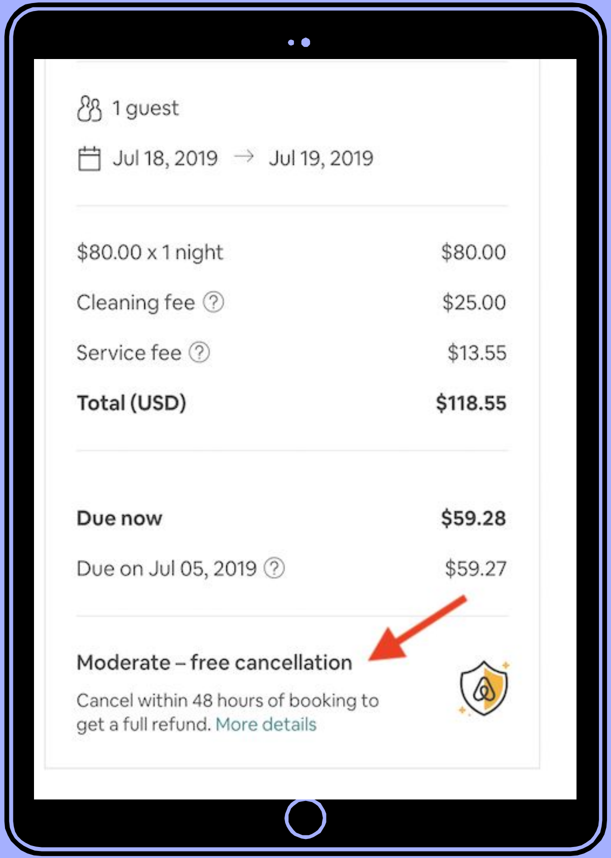


Recommendations:

- **Encourage Positive Reviews**
    - More reviews makes you more reputable!
- **Have Competitive Pricing**
    - Research your neighbors!
- **Aim to be a Superhost**
    - Superhosts get more reviews and make more revenue!

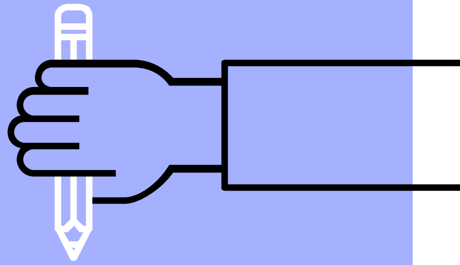
Recommendations:

- **Encourage Positive Reviews**
  - More reviews makes you more reputable!
- **Have Competitive Pricing**
  - Research your neighbors!
- **Aim to be a Superhost**
  - Superhosts get more reviews and make more revenue!
- **Moderate Cancellation Policy**
  - Don't be too strict, don't be too flexible!

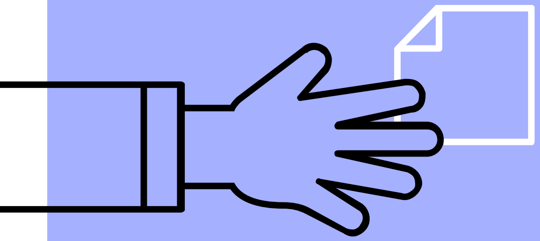


Recommendations:

- **Encouage Positive Reviews**
    - More reviews makes you more reputable!
- **Have Competitive Pricing**
    - Research your neighbors!
- **Aim to be a Superhost**
    - Superhosts get more reviews and make more revenue!
- **Moderate Cancellation Policy**
    - Don’t be too strict, don’t be too flexible!
- **Communicate**
    - More bookings if you respond to your guests in a timely manner!



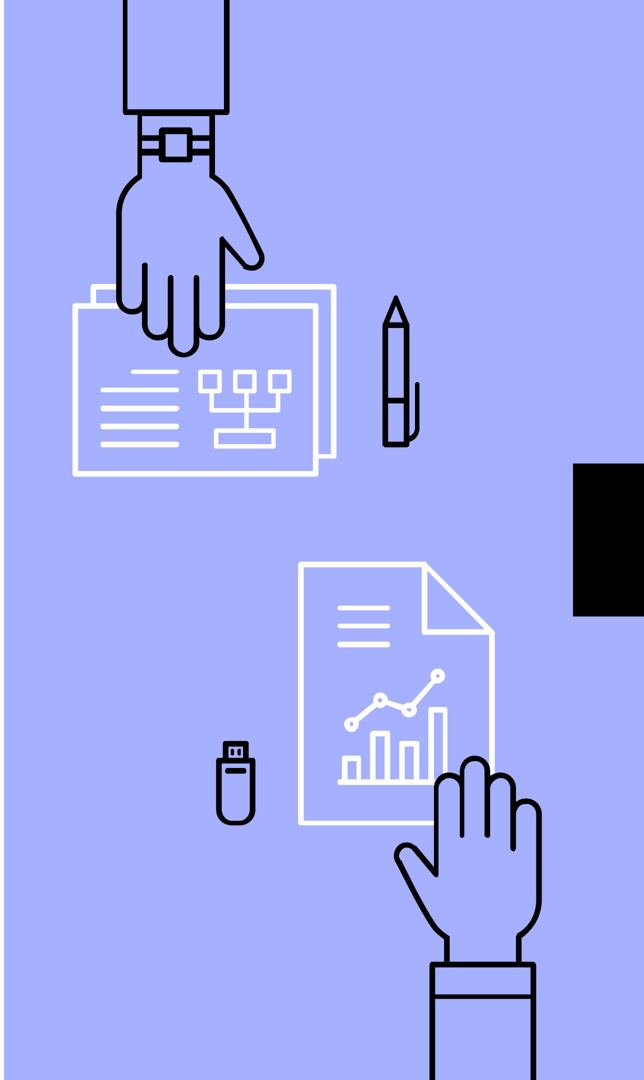
6.
Tentative: Team Project Process



Our Team Process

Describes the process of how we worked as a team to derive results.

- Charts & graphs defined our questions
- Began to allow our questions to drive our analysis
- Change of focus from many questions to one large question
  - This allowed us to vary our analysis and distribute different factors to different team members.
- Solutions provided to advise Airbnb hosts where to invest into a property based on the variety of analyzed factors.

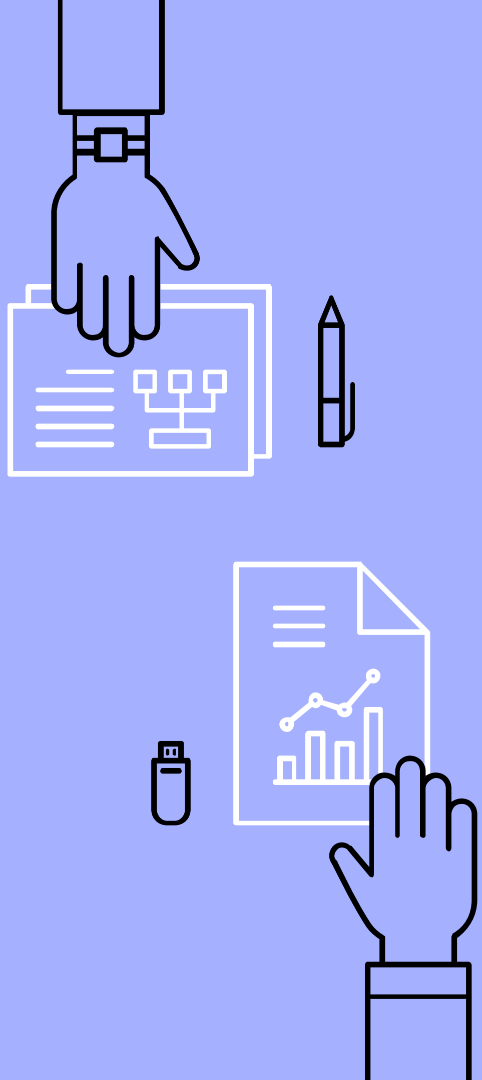


Open-Ended Questions & Lessons

- Is the revenue a good indicator for the success of a host?
- Can a host have consistent revenue?
- Restriction of time-independent data.
- Significance of team discussions.

Resources

- https://public.opendatasoft.com/explore/dataset/airbnb-listings/table/?disjunctive.host_verifications&disjunctive.amenities&disjunctive.features
- http://insideairbnb.com/about.html
- https://home.bt.com/lifestyle/travel/travel-advice/what-is-airbnb-11363981595930
- https://news.airbnb.com/fast-facts/



THANKS!

Any questions?