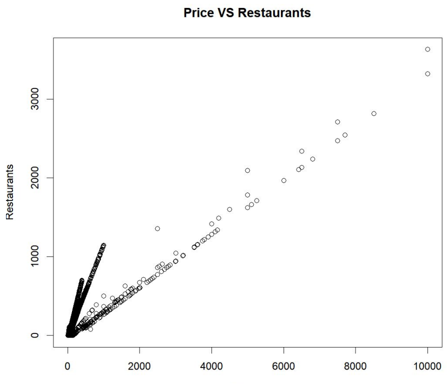
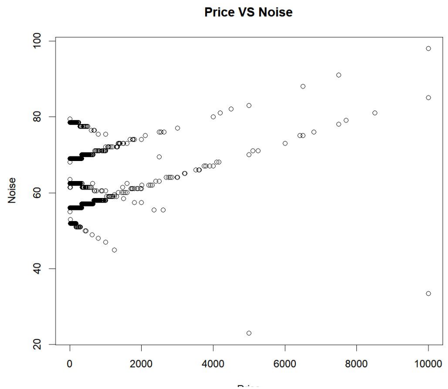
Price Prediction

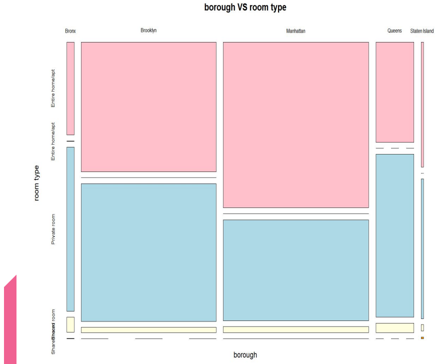
Tanvi Pathak

Interesting Relationships

- I wanted to make sure I can reflect all of the relationships observed in the data with the models.
- I thought that room type, borough type, and number of restaurants played the biggest role in determining airbnb prices.
- I noticed that the most popular places(that is the area with the most restaurants) had the priciest airbnbs.
- Interestingly, there were also differences within boroughs. So certain places were more popular than others within the borough.
- I did notice that the most popular borough was Manhattan.
- Further, since I saw a difference amongst room types between different boroughs, I assumes that there would be other differences like restaurant concentration or noise level differences between boroughs as well.







Making the Model

- I wanted to include the most important variables.
- In my opinion this was number of restaurants, room type, noise levels, and neighbourhood group.
- I included neighbourhood group so we can see relationships between each group.
- My initial model was just a multiple regression model with all the interesting variables to see what I was working with.
- I made the categories factors so that the model doesn’t treat the categories as ordinal when they aren’t. Plus it enhances interpretability.
- This was the original model.
- Not the best but at least it gives us an idea of the relationship between the variables.

```
Call:
lm(formula = price ~ as.factor(neighbourhood_group) + as.factor(room_type) +
    noise.dB. + resturants, data = nyc_data_final)

Residuals:
    Min      1Q  Median      3Q     Max
-1673.3   -12.7     1.8    19.9 12031.9

Coefficients:
                                          Estimate Std. Error  t value Pr(>|t|)
(Intercept)                             -1.325e+04  7.686e+01 -172.354  < 2e-16 ***
as.factor(neighbourhood_group)Brooklyn   1.496e+03  1.033e+01  144.808  < 2e-16 ***
as.factor(neighbourhood_group)Manhattan  3.730e+03  2.258e+01  165.153  < 2e-16 ***
as.factor(neighbourhood_group)Queens     2.684e+03  1.637e+01  163.982  < 2e-16 ***
as.factor(neighbourhood_group)Staten Island 4.535e+03 2.734e+01 165.903 < 2e-16 ***
as.factor(room_type)Entire home/apt      1.532e+03  6.795e+01   22.542  < 2e-16 ***
as.factor(room_type)Private room         2.101e+01  1.397e+00   15.037  < 2e-16 ***
as.factor(room_type)Shared room          2.375e+01  4.017e+00    5.913 3.38e-09 ***
as.factor(room_type)Shared room         -2.188e+02  5.875e+01   -3.724 0.000197 ***
noise.dB.                                1.698e+02  9.818e-01  172.943  < 2e-16 ***
resturants                               8.493e-01  5.373e-03  158.056  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 117.2 on 39132 degrees of freedom
Multiple R-squared:  0.7682,	Adjusted R-squared:  0.7681
F-statistic: 1.297e+04 on 10 and 39132 DF,  p-value: < 2.2e-16
```

Making the Model

- In order to make the model better, I thought I could add some interaction terms in order to highlight the relationship between certain predictor variables.
- I decided to observe the relationships between each borough and the other predictors.
  - For example, how noise levels affect property prices differently in various boroughs
- The model has way too many categories and relationships so there are definitely some issues with it.
- My R^2 was pretty decent and the MSE score was also pretty low.
- However, this doesn’t mean it's a good model(addressed later).

```
:all:
lm(formula = price ~ as.factor(neighbourhood_group) * noise.dB. +
    as.factor(neighbourhood_group) * as.factor(room_type) + as.factor(neighbourhood_group) *
    resturants, data = nyc_data_final)

Residuals:
     Min       1Q   Median       3Q      Max
-1636.08   -11.14    -1.86    12.40  2170.38

Coefficients: (4 not defined because of singularities)
                                                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                                                           -3.588e+03  1.412e+03  -2.541 0.011057 *
as.factor(neighbourhood_group)Brooklyn                                -1.648e+04  1.413e+03 -11.663  < 2e-16 ***
as.factor(neighbourhood_group)Manhattan                               -1.266e+04  1.412e+03  -8.960  < 2e-16 ***
as.factor(neighbourhood_group)Queens                                   2.053e+04  1.418e+03  14.479  < 2e-16 ***
as.factor(neighbourhood_group)Staten Island                            9.589e+03  1.434e+03   6.686 2.32e-11 ***
noise.dB.                                                              4.688e+01  1.798e+01   2.607 0.009149 **
as.factor(room_type)Entire home/apt                                    6.663e+02  2.924e+02   2.279 0.022682 *
as.factor(room_type)Private room                                      -3.334e+01  3.543e+00  -9.411  < 2e-16 ***
as.factor(room_type)Shared room                                       -4.420e+01  7.406e+00  -5.968 2.42e-09 ***
as.factor(room_type)Shared room                                        7.686e+02  2.924e+02   2.629 0.008574 **
resturants                                                             2.154e+00  1.187e-01  18.151  < 2e-16 ***
as.factor(neighbourhood_group)Brooklyn:noise.dB.                       2.446e+02  1.800e+01  13.588  < 2e-16 ***
as.factor(neighbourhood_group)Manhattan:noise.dB.                      2.449e+02  1.799e+01  13.611  < 2e-16 ***
as.factor(neighbourhood_group)Queens:noise.dB.                        -3.169e+02  1.809e+01 -17.515  < 2e-16 ***
as.factor(neighbourhood_group)Staten Island:noise.dB.                 -1.606e+02  1.862e+01  -8.623  < 2e-16 ***
as.factor(neighbourhood_group)Brooklyn:as.factor(room_type)Entire home/apt   2.004e+03  2.963e+02   6.763 1.37e-11 ***
as.factor(neighbourhood_group)Manhattan:as.factor(room_type)Entire home/apt -1.010e+03  2.963e+02  -3.410 0.000651 ***
as.factor(neighbourhood_group)Queens:as.factor(room_type)Entire home/apt            NA         NA      NA       NA
as.factor(neighbourhood_group)Staten Island:as.factor(room_type)Entire home/apt     NA         NA      NA       NA
as.factor(neighbourhood_group)Brooklyn:as.factor(room_type)Private room      1.510e+01  3.668e+00   4.116 3.86e-05 ***
as.factor(neighbourhood_group)Manhattan:as.factor(room_type)Private room    -9.324e+00  3.644e+00  -2.559 0.010510 *
as.factor(neighbourhood_group)Queens:as.factor(room_type)Private room        2.119e+01  3.968e+00   5.342 9.27e-08 ***
as.factor(neighbourhood_group)Staten Island:as.factor(room_type)Private room -2.874e+01  7.063e+00  -4.068 4.75e-05 ***
as.factor(neighbourhood_group)Brooklyn:as.factor(room_type)Shared room       1.917e+01  7.915e+00   2.422 0.015445 *
as.factor(neighbourhood_group)Manhattan:as.factor(room_type)Shared room     -1.270e+01  7.820e+00  -1.624 0.104358
as.factor(neighbourhood_group)Queens:as.factor(room_type)Shared room         2.162e+01  8.467e+00   2.553 0.010684 *
as.factor(neighbourhood_group)Staten Island:as.factor(room_type)Shared room  1.949e+01  2.004e+01   0.972 0.330989
as.factor(neighbourhood_group)Brooklyn:as.factor(room_type)Shared room              NA         NA      NA       NA
as.factor(neighbourhood_group)Manhattan:as.factor(room_type)Shared room     -2.208e+03  2.963e+02  -7.454 9.27e-14 ***
as.factor(neighbourhood_group)Queens:as.factor(room_type)Shared room                NA         NA      NA       NA
as.factor(neighbourhood_group)Staten Island:as.factor(room_type)Shared room  8.919e+02  3.027e+02   2.947 0.003215 **
as.factor(neighbourhood_group)Brooklyn:resturants                     -1.769e+00  1.187e-01 -14.900  < 2e-16 ***
as.factor(neighbourhood_group)Manhattan:resturants                    -1.871e+00  1.187e-01 -15.759  < 2e-16 ***
as.factor(neighbourhood_group)Queens:resturants                       -1.545e+00  1.192e-01 -12.959  < 2e-16 ***
as.factor(neighbourhood_group)Staten Island:resturants                -1.375e+00  1.350e-01 -10.188  < 2e-16 ***
---
Residual standard error: 47.7 on 39112 degrees of freedom
Multiple R-squared:  0.9616,    Adjusted R-squared:  0.9616
F-statistic: 3.266e+04 on 30 and 39112 DF,  p-value: < 2.2e-16
```

Improving the Model

- One of the biggest issues with the model is the use of too many variables.
- I decided to only focus on the most meaningful predictor and relationships.
- I thought I can take a look at just neighbourhood group, number of restaurants, and noise levels.
- Since Airbnbs would be more expensive where it is more popular, I assumed that noise and number of restaurants can be a good indicator of popularity the spot.
    - The more crowded a place, the noisier it is.
    - Plus if there are more restaurants there are probably more tourists.
- This model was also pretty good with an adjusted R^2 of 0.9537.
- However the mean squared error was 2740.104 which isn't bad but was greater than the MSE for the previous model.
- Hence, I decided to stick with the previous model.

```
Call:
lm(formula = price ~ as.factor(neighbourhood_group) * noise.dB. +
    as.factor(neighbourhood_group) * resturants, data = nyc_data_final)

Residuals:
     Min       1Q   Median       3Q      Max
-1590.12   -10.67    -1.40    12.49  2604.53

Coefficients:
                                                        Estimate Std. Error  t value Pr(>|t|)
(Intercept)                                           -7.107e+01  1.803e+02   -0.394    0.693
as.factor(neighbourhood_group)Brooklyn                -1.934e+04  1.910e+02 -101.247  < 2e-16 ***
as.factor(neighbourhood_group)Manhattan               -1.562e+04  1.839e+02  -84.923  < 2e-16 ***
as.factor(neighbourhood_group)Queens                   1.665e+04  2.200e+02   75.668  < 2e-16 ***
as.factor(neighbourhood_group)Staten Island            1.507e+03  2.171e+02    6.939 4.02e-12 ***
noise.dB.                                              1.779e+00  2.298e+00    0.774    0.439
resturants                                             1.897e+00  3.332e-02   56.932  < 2e-16 ***
as.factor(neighbourhood_group)Brooklyn:noise.dB.       2.799e+02  2.473e+00  113.163  < 2e-16 ***
as.factor(neighbourhood_group)Manhattan:noise.dB.      2.796e+02  2.389e+00  117.020  < 2e-16 ***
as.factor(neighbourhood_group)Queens:noise.dB.        -2.662e+02  3.055e+00  -87.141  < 2e-16 ***
as.factor(neighbourhood_group)Staten Island:noise.dB. -2.853e+01  3.263e+00   -8.743  < 2e-16 ***
as.factor(neighbourhood_group)Brooklyn:resturants     -1.457e+00  3.349e-02  -43.509  < 2e-16 ***
as.factor(neighbourhood_group)Manhattan:resturants    -1.526e+00  3.349e-02  -45.576  < 2e-16 ***
as.factor(neighbourhood_group)Queens:resturants       -1.240e+00  3.499e-02  -35.445  < 2e-16 ***
as.factor(neighbourhood_group)Staten Island:resturants  2.300e-02  4.777e-02    0.481    0.630
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 52.36 on 39128 degrees of freedom
Multiple R-squared:  0.9537,     Adjusted R-squared:  0.9537
F-statistic: 5.762e+04 on 14 and 39128 DF,  p-value: < 2.2e-16
```

Issues with the Model

```
Warning message:
In predict.lm(model_5, test_data) :
  prediction from a rank-deficient fit may be misleading
```

- One of the biggest issues with the model I chose was multicollinearity.
- When I was predicting the new prices using the test data, I got the following warning.
- Rank deficiency means that some predictors can be expressed as linear combinations of others.
- This can result in issues with estimation leading to larger standard errors or inflated coefficients.
- The best way to address this problem is to determine what predictors are highly correlated and either trying to simplify the model further or trying something like a PCA or factor analysis to take care of the multicollinearity problem.