Airbnb

Q3 BOOKING PREDICTION

KBK Consulting:
Kristina Dolan
Bryan Garcia
Kavya Rajesh

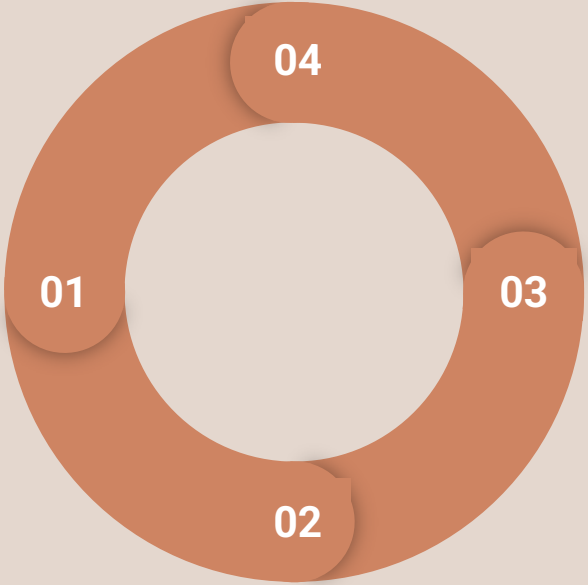
Data Process

01 Data Cleaning

Replaced missing values with the median

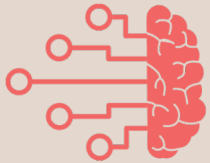
02 Feature Engineering

Created a feature that was the Q1 and Q2 average price for each property



04 Modeling

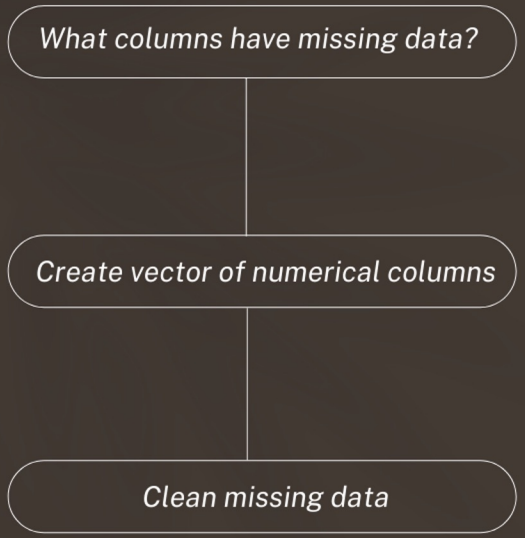
Utilized Linear Regression and Regression Trees



03 Exploratory Data Analysis

Used simple table functions and histograms to explore/understand the data

R Implementation - Important Code




```r
# Gives TRUE/FALSE values for each feature whether they have ANY missing value
na_checker <- apply(property_info, 2, function(x) any(is.na(x)))
# Provides ONLY columns that have missing value
col_with_na <- names(na_checker[as.numeric(na_checker) == 1])

# Remove any non-numerical columns from previous list
col_vec_toclean <- col_with_na[!col_with_na %in% c("Neighborhood","Superhost")]


# Replace missing value with the median
for (i in col_vec_toclean)
{
 property_info[[i]][is.na(property_info[[i]])] <- median(property_info[[i]], na.rm = TRUE)
}
```

Feature Selection

Correlation Matrix

01

Relationship With Target Variable

Used Correlation Matrix to find features that had a relationship with the target variable

02

Features Not Highly-Skewed

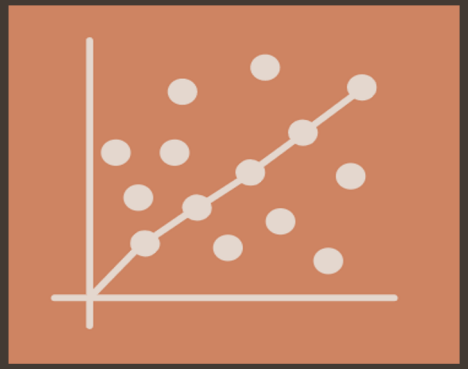
Did not utilize features that had highly-skewed data. (e.g. rating features were all 10)

03

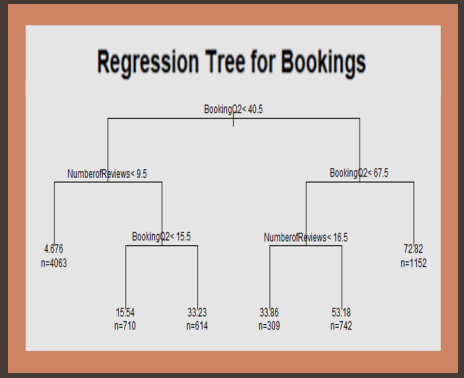
Statistically Significant Features

Through trial-and-error, we removed features that were statistically insignificant in our model.(based on p-value)

Prediction Models and Methods



Linear Regression



Regression Trees

Data Cleaning | Feature Selection | Modeling | Reflection

R Implementation - Model

Linear Regression

```
Call:
lm(formula = NumReserveDays2016Q3 ~ Superhost + Longitude + BookingQ2 +
    AvgPriceQ2 + NumberofReviews + BlockedQ1 + MaxGuests + BlockedQ2,
    data = property_info_train)
```

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  1.372e+03  3.771e+02   3.638 0.000276 ***
SuperhostTRUE                4.256e+00  6.743e-01   6.312 2.90e-10 ***
Superhostunknown host type -7.211e+00  1.220e+00  -5.909 3.60e-09 ***
Longitude                    1.849e+01  5.099e+00   3.625 0.000291 ***
BookingQ2                    6.821e-01  1.015e-02  67.194  < 2e-16 ***
AvgPriceQ2                  -4.583e-03  1.181e-03  -3.881 0.000105 ***
NumberofReviews              1.704e-01  7.375e-03  23.100  < 2e-16 ***
BlockedQ1                   -1.303e-01  7.870e-03 -16.561  < 2e-16 ***
MaxGuests                    6.262e-01  1.133e-01   5.528 3.34e-08 ***
BlockedQ2                    8.055e-02  9.016e-03   8.934  < 2e-16 ***
```

15.84	8	0.73
MSE	**Statistically Significant Predictors**	**R^2**

Data Cleaning | Feature Selection | Modeling | Reflection

Reflection on the Prediction Challenge

Data Visualization

Data Cleaning

Cross Validation

Data Cleaning | Feature Selection | Modeling | Reflection

Biggest Challenge

Data Cleaning | Feature Selection | Modeling | Reflection