# Be a Wise Host in Airbnb

HAPPY DATA

**Four aspects to analyze the topic:**
Find 1) a worthy neighborhood, 2) the best house-type, 3) the proper price to lease, and 4) good descriptions to maximize hosts' profits in Airbnb.
Our dataset: reviews_detail(1050548x6), listings_detailed(50041X96), rolling_sales(82145x20)

## 1. Data Cleaning (Dataset from Airbnb & NYU Rolling Sales Data)

- **Handle 'bad' data**: we import the **re library** to transform price-related columns into int or float type. For example, by **re.findall(r'[0-9]+, string)**, change '$ 100000'(str) into 100000(int).
- **Remove 'missing' data**: Use **dropna(inplace=True)** remove all NA rows.

## 2. Machine Learning

- **Categorical Variables in regression**

To answer question 2&3, we try to fit a pricing model for our dataset. Check all decision variables firstly:

***accommodates (float), bathrooms (float), bedrooms(float), review_scores_rating (float), security_deposite (float), cleaning_fee (float), beds (float), number_of_reviews (float), host_listings_count(float), neighbourhood (string), room_type (string).***

We find variables 'neighbourhood' and 'room type' are two string columns, which is difficult for our regression analysis. In order to solve it, we come up with 2 methods:

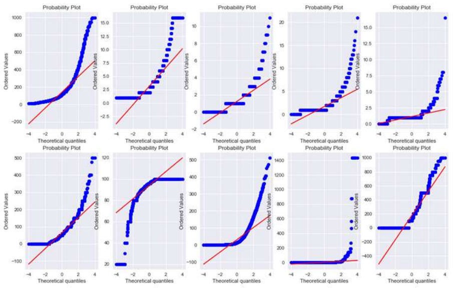
1) ***Transfer string into float, assign an integer value to every distinct string.***
2) ***Keep them as categorical variables.***

With method 1), there are a large number of regression models we can apply, such as Linear regression, Lasso regression, and Ridge regression. However, this method is not suitable when we dig into the actual meaning of variables. For example, variable 'neighbourhood' means the location information of a house. It is meaningless to compare the value of this variable after we simply transform it from string type into float type. With method 2), we can cluster the data by variable 'neighborhood', which is more reasonable. In order to include categorical variables in the regression, we apply a method named categorical regression, which is a model based on ordinary linear regression.

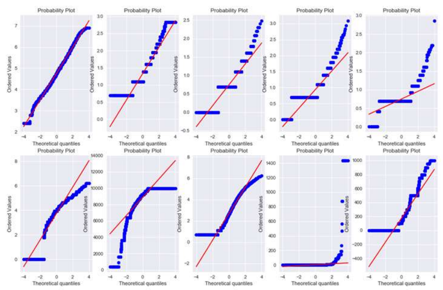
- **Modify Categorical Regression**

As ordinary linear regression always has a large bias, we find a method to minimize the error. In fact, the linear regression on variables satisfying Normal distribution always performs better than other cases. Therefore, we transform the original dataset using square and log function to remove the skewness of all variables:

QQ Plot before transformation



QQ Plot after transformation



From the QQ plot, we can conclude we successfully make all variables approximately satisfy normal distribution. What's more, the MSE of the model with transformation decreases from 0.41 to 0.33 compared with the model without transformation. Thus, our transformation increases the prediction accuracy a lot.

To make our model more reality-oriented, we scale the variables. As all know, coefficients in linear regression can reflect the importance of corresponding variables if and only if all the variables are on the same scale. For example, variable 'review score rating' belongs to interval $(0, 100)$ while variable 'accommodates' in interval $(0, 15)$. As a result, larger parameter of 'review score rating' dosen't represent it is more important than 'accommodates'. For the purpose of fixing this problem,, we scale all variables into interval $(0, 1)$. Then the larger the coefficient of a variable, the more important of this factor.

Finally, we can fit the model:

$$\begin{aligned}\text{price} = {} & 0.195 \times \text{accommodates} + 0.170 \times \text{bathrooms} + 0.123 \times \text{bedrooms} \\ & + 0.080 \times \text{review scores rating} + 0.036 \times \text{cleaning fee} + 0.035 \times \text{security deposit}\end{aligned}$$

$$+0.028 \times \text{beds} - 0.012 \times \text{numbers of reviews} + 0.001 \times \text{host listings count}$$

$$+\sum_{i=0}^{n} a_i I(\text{neighbourhood} = \text{neighbourhood}_\text{i}) + \sum_{i=0}^{n} b_i I(\text{room type} = \text{room type}_\text{i})$$

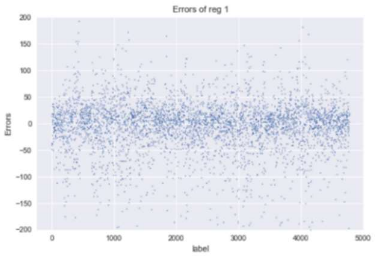
where: 1) $I(\text{x} = \text{c}) = \begin{cases} 1, & \text{x} = \text{c} \\ 0, & \text{otherwise} \end{cases}$

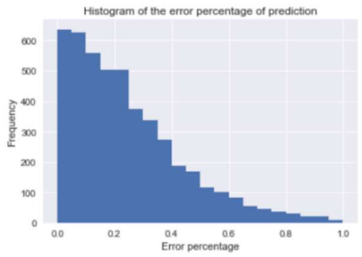
2) $a_i$ is the coefficient of $\text{neighbourhood}_\text{i}$, $b_i$ is the coefficient of room $\text{type}_\text{i}$.

- **Model Accuracy**

In order to test the accuracy of our model, we include the definition of error percentage.

**Definition (Error percentage)**: The floating rate of prediction with the true value. For example, if the true price is 100 and the estimation is 120, the error percentage is 0.2.





*From Histogram of error percentage, we know that the error percentage of 70% houses in the testing set is below 33%. Therefore, our prediction model is reasonable.*

- **Importance of all factors**

From the coefficients of the model we fit, the influence of all factors on price can be ordered as:

***accommodates > bathrooms > bedrooms > review_scores_rating > security_deposite > cleaning_fee > beds > number_of_reviews > host_listings_count***

## 3. Text Mining

- **Wordcloud**

For question 4, we want to use technicals from word cloud to find the strategy of writing a description of the Airbnb house. After conducting comparisons several times, we find the relationship between description and popularity of the houses. We choose two groups with the sample in same size:

**Group one**(popular): review_scores_rating > 90(high_quality house) & reviews_per_month > 7

**Group two**(unpopular): review_scores_rating > 90(high_quality house) & reviews_per_month < 0.04

In order to distinguish the two group, we write a random_color_func to define a different color for the two pictures. As a result, we find that the biggest difference between the description two group is that the description of popular house emphasis on making the guest feel at home, which is more emotional than just describing the house and neighborhood and attracts customers more.

- **Sentiment Analysis**

For the second Part, we choose two group to analysis what emotion affects the rating scores.

**Group one**: number_of_reviews > 100 & top 50 highest review_scores_rating houses

**Group two**: number_of_reviews > 100 & top 50 lowest review_scores_rating houses

We calculate the average of the result from emotion analysis and visualize it. We find the most important emotion that differs guests' rating is joy, the second is trust.

## 4. Visualization

To provide a clearer view of our analysis, we use five graphs visualizing our data, function and results.

- **Heatmap** to compare each neighborhood's average rental price
- **Scatter chart** to show the difference between rental price and house price in each neighborhood
- **Markercluster map** from Markercluster library to show the distribution of all Airbnb houses in NYC
- **WordCloud** to reflect how effective coefficients of variables are in the Regression model.

## 5. Results

- Find a worthy neighborhood: use the scatter chart showing differences between the average rental price in Airbnb and house sale price, finding the most worthy neighborhood -- Belle Harbor.
- Find the best house-type: use the linear regression model to find the factor that influences the price most, which is accommodate.
- Price the House: Apply Categorical regression on pricing and adjust the price referring to the Markercluster ma
- Write the Description: Use text mining to find the best way to write a description to attracts customers. As a result, emotional description attracts most