KAGGLE PRESENTATIONS

AirBnB Price Prediction

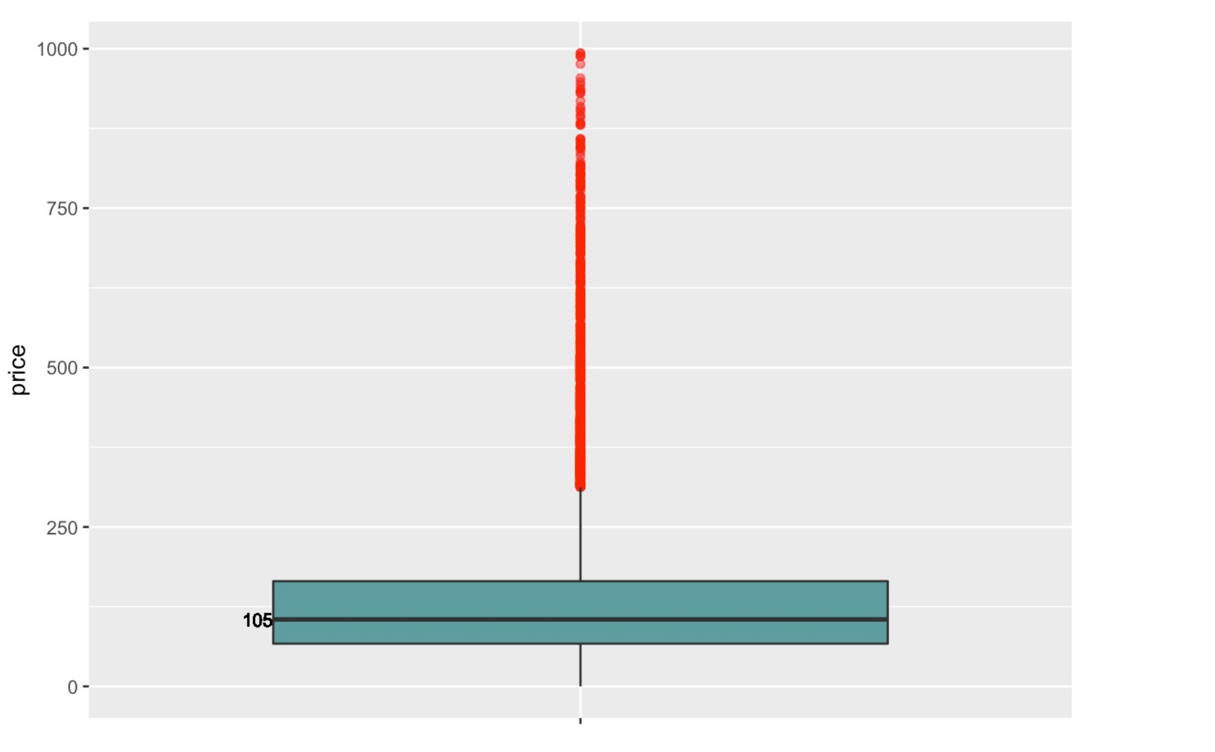
By: Romauli Butarbutar

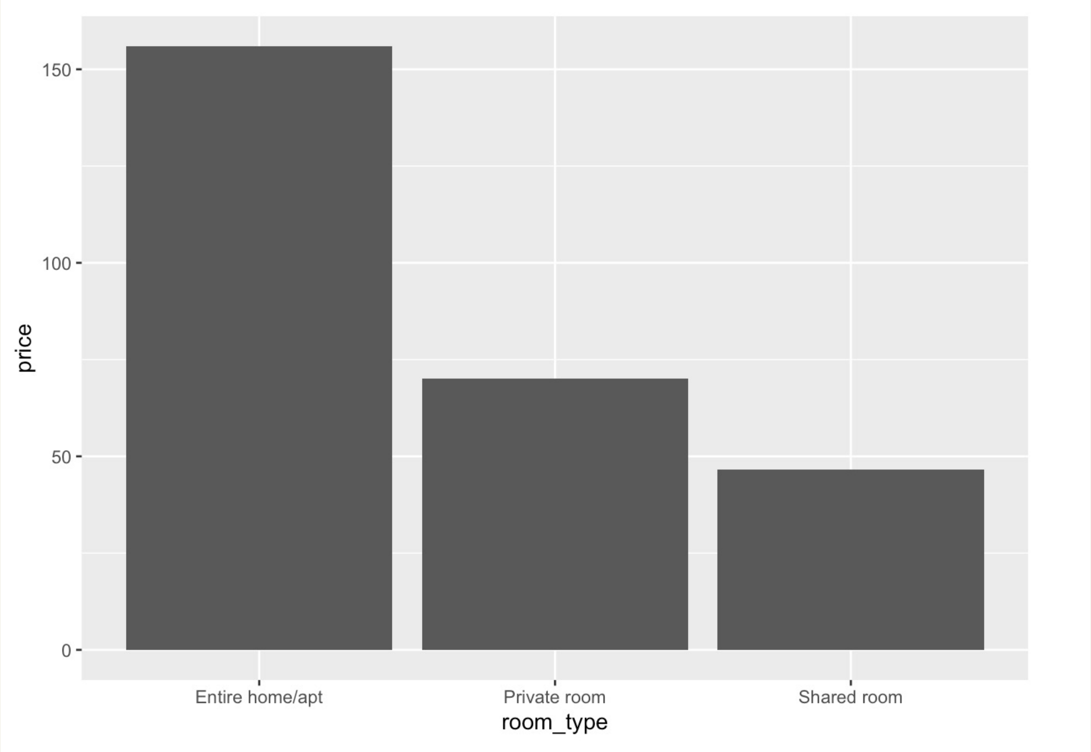
What I did right with the analysis

Part 1 - The steps of data analysis

- Data Category for better understanding
  - *Listing/URL, Host, Location, Property, Price, Term of condition, Additional descriptors ***
  - *Ignore insignificant variable: Null value, Id, Additional descriptor, Listing/URL descriptors ***
  - *Examine outliers by visualizing the data*
- Data Wrangling, Cleaning and Tidying
  - *Reformat data, convert data type : date, character*
  - *Exclude data with country_code = 'UY' (for UK as most of the data for US)*
  - *Check blank data and impute missing value for both train and test*
  - *Word count for character data types e.g. summary, description*
  - *Check different types of amenities using Regular Expressions functions*
  - *Levelling multiple factoral variable and imputing the average price by group e.g. neighbourhood_group_cleansed*

Examine outliers by visualizing the data





What I did wrong with the analysis

Part 2 – Creating The Models

- **Linear Regression** per descriptor category, the least RSME is about 67.89162
- **Feature Selection** :

  ***Corrplot, Best Subset Selection, Forward and Hybrid Selection, the least RSME is about 67.90036*

  ***After perform shrinkage, the least RSME using Lasso method is about 68.38622,*
- **Tree Model**

  ***Simple Regression Tree, the least RSME is about 72.12098*

  ***Regression Tree Complex, the least RSME is about 69.49447,*

  ***Advanced Tree, the least RSME is about 51.00288 : but with note for file submission is not working (errors)*

  ***Tree with Tuning, using 5-fold cross-validation. The least RSME is about 64.24447.*

  When I try to perform Random Forest, Tuned Random Forest, Forest with Ranger and Boosting wit cross-validation and Boosting with XGBoost it took so long time, so I decided to cut the process by terminating R.
- **Final Boosting Models**

  After perform Data Cleaning Complexity, and include some wordcount and amenities, this is my FINAL MODEL with BOOSTING METHOD.

  ## predict train dataset : RMSE 37.50427

  ON PUBLIC LEADERBOARD: 58.71

  ON PRIVATE LEADERBOARD : 62.57

Model	RMSE
Linear Regression	67.89162
Feature Selection - Corrplot, Best Subset Selection, Forward and Hybrid Selection - Lasso method	67.90036 68.38622
Tree Model - Simple Regression Tree - Regression Tree Complex - Advanced Tree - Tree with Tuning, 5-fold cross-validation	72.12098 69.49447 51.00288 64.24447
- Final Boosting Method	37.50427
- ON PUBLIC LEADERBOARD	58.71
- ON PRIVATE LEADERBOARD	62.57

Report Summarizing & Lesson Learned

- The price of Airbnb rental affected by room type neighbourhood_group_cleansed, amenities, cleaning_fee,review, rating etc.
- Before doing a deep analysis : the most significant independent variable mainly LOCATION
- After modelling : detail description of apartment e.g summary, rating, neighborhood_overview etc
- Suggestion : the detail description/ summary in the listing to gain more users and popularity.
- The failed steps or missteps along the way
  - **So many errors for the first kick!*
  - **Some of my models like XGBoosting Model, Tuning the Tree are not working and still confuse with the error.*
  - **When I perform Dimension Reduction Technique, I found some errors that I decided to cut the process.*
  - **Performing several technique of Forward selection and Hybrid selection would result insignificant difference, so I think we just need to choose one for time efficiency.*

“It is through science that we prove. But it is through intuition that we discover”
— Henri Poincare —

- More importantly, I realize that the level of complexity of the model and variables probably could lead to overfitting problem.
- 70% of the time is for data cleaning, wrangling and tidying
- It is important to use the common knowledge and use a good intuition how to logically select the relevant variables for a model, like the quotes.

```
boostModelFinal = gbm(price ~ meanPrice+meanPriceGC + level_nc + bedrooms + room_type + property_type + bathrooms + beds
                  + accommodates + cleaning_fee + monthly_price + security_deposit + minimum_nights + maximum_nights  + neighbourhood_group_cleansed
                  + host_is_superhost + availability_30 + availability_60 + availability_90 + availability_365
                  + review_scores_rating + number_of_reviews + last_review_days + first_review_days + review_scores_cleanliness + review_scores_accuracy
                  + wc_transit + wc_summary+ wc_description+ wc_host_about + wc_neighborhood_overview #word count
                  + host_listings_count + host_since_days + reviews_per_month + host_has_profile_pic
                  + extra_people + guests_included + cancellation_policy
                  + Airconditioning + Dryer + Elevator + Familykidfriendly + Freestreetparking #amenities
                  + Hairdryer + Iron + Oven + Refrigerator + Shampoo + Selfcheckin  #amenities
            ,data = train, distribution = "gaussian",
            n.trees = 30000,
            interaction.depth = 5,
            shrinkage = 0.005,
            n.minobsinnode = 5)
```