DA3 Assignment 1

Airbnb prediction models: New York

Introduction

This is a report for Data Analysis 3: Assignment 1. The task is to perform prediction using the knowledge we obtained in the course.

As a first step, we define a business question: to help a company operating small and mid-size apartments hosting 2-6 guests. The company is set to price their new apartments not on the market. For this, I built a price prediction model similarly to how we did in our case study for London (Hackney borough). Also, discussion of my modeling decisions and comparison of my results is provided at the end. In the next step, I define the data for analysis (sample design). For this, I obtained data from insideairbnb.com website. Data is for New York, it was compiled on 10 December, 2020.

Label Engineering

We have a quantitative y - price in USD dollars. That's why I leave y as it is (without any transformation).

Feature engineering

Data cleaning

The original dataset that was obtained from the website contained 36923 rows and 62 columns. I decided to remove several columns that are not used in the analysis (see Appendix).

Then dollar signs ("$") were removed from price column. Also, false (f) and true (t) values were replaced with 1 and 0.

Amenities column consisted of several values. These values in the row were pivoted as a column of dummy variable. List of amenities is given in Appendix.

Code for data cleaning is provided in **data-cleaning-NY.R** file.

Data preparation

When dataset was inspected, it was clear that "Entire apartment", "Private room in apartment", and "Entire condominium" were the most frequent property type (Appendix). Therefore, I decided to filter out the dataset and leave only observations that have these property types.

Dummy variables were created from amenities column (that starts with 'd'). I kept dummy variables, numerical and factorial (categorical) columns and removed all other columns that are not used for the analysis. After that, I left with 49 columns (variables).

The observations where the value for price column is missing were dropped as our analysis will be prediction of price.

I decided to narrow down my focus and filtered dataset so that it includes observations only from Manhattan neighborhood. After that, 28402 observations decreased to 14170.

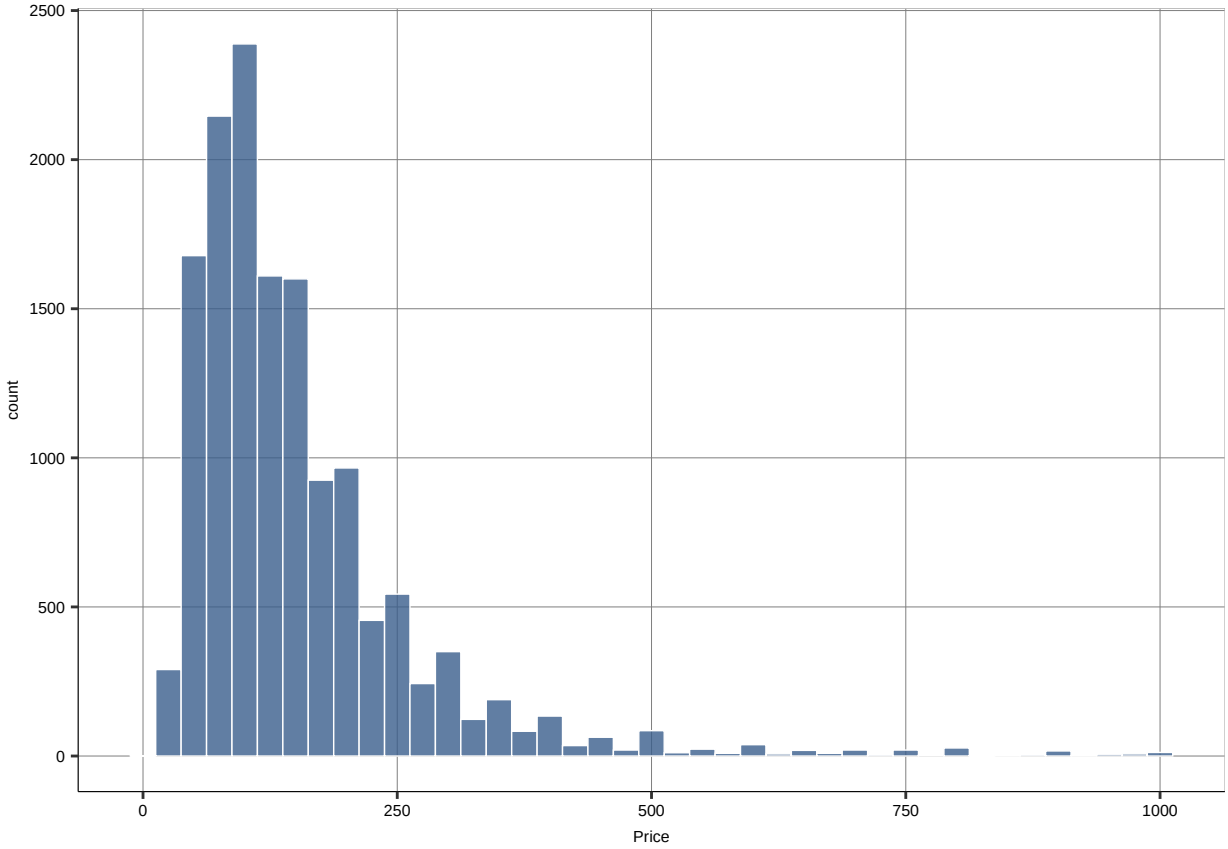
Exploratory data analysis

We can see that minimum price is 10 dollars and maximum price is 999 dollars. Average price is 152.1 dollars.

Table 1: Summary of prices

	V1		V1		V1		V1		V1		V1
Min.	10	1st Qu.	80	Median	120	Mean	152	3rd Qu.	186	Max.	999

The distribution (normal) of log-price shown below.



Appendix

Data Cleaning

Removed columns:

“host_thumbnail_url”,“host_picture_url”,“listing_url”,“thumbnail_url”,“medium_url”,“picture_url”,“xl_picture_url”,“h
“experiences_offered”, “neighborhood_overview”, “notes”, “transit”, “access”, “interaction”, “house_rules”,
“host_about”, “host_response_time”, “name”, “summary”, “space”, “host_location”

Amenities:

“conditioner”, “stove”, “hdtv”, “tv”, “conditioning”, “sound system”, “refrigerator”, “shampoo”, “soap”,
“oven”, “toiletries”, “speaker”, “fan”, “heating”, “breakfast”, “table”, “dishwasher”, “dryer”, “elevator”,

“fitness”, “parking”, “garage”, “wifi”, “game”, “garden”, “gym”, “restaurant”, “bar”, “washer”, “barbeque”, “bbq”

Property types:

Table 2: Property types

	V1
Entire apartment	15577
Private room in apartment	12007
Private room in house	1871
Private room in townhouse	995
Entire condominium	978