Title: Classification of Airbnb Listing Prices in New York City

EE418 Project Proposal

November 11, 2025

Members:

Esma Kadrić, Amina Hrustić, Asja Bašović

Table of Contents

I Introduction

The aim of the project is the classify prices of Airbnb properties in New York City. The dataset used contains real-life data, hence the project proposed is useful, realistic, and challenging.

- Main Task: Supervised learning | Classification
- Target : Price | 'low', 'medium', 'high'

  The main task we will be working on is supervised learning, classification. We will define target as price: 'low', 'medium', 'high' using quantile-based bins or possibly cluster-based bins, derived using unsupervised learning, to explore underlying trends.

- Dataset Name: New York City, New York, United States (listings.csv.gz)
- Dataset URL: https://insideairbnb.com/get-the-data/
- Dataset collected: Monthly (Scraped)

  In the dataset we can see the exact day the page was scraped that month.

  All Scrape Dates:

04 November 2024	04 December 2024	03 January 2025
01 February 2025	01 March 2025	01 April 2025
01 May 2025	17 June 2025	01 July 2025
01 August 2025	01 September 2025	01 October 2025

One limitation of the project is that there is a lack of multi-year data, as it was not available on the website. However, the twelve different months still provide useful and valuable temporal data.

II Dataset Details

Database Dimensionality

- Initial dimensionality: (443898, 79)
- After dropping unnecessary columns (443898, 35)

Available features

['id', 'last_scraped', 'host_id', 'host_response_time',

'host_response_rate', 'host_acceptance_rate', 'host_is_superhost',

'host_has_profile_pic', 'host_identity_verified',

'neighbourhood_cleansed', 'property_type', 'room_type', 'accommodates',

'bathrooms', 'bedrooms', 'beds', 'amenities', 'price', 'minimum_nights',

'maximum_nights', 'has_availability', 'availability_30',

'availability_365', 'number_of_reviews', 'number_of_reviews_l30d',

'review_scores_rating', 'review_scores_accuracy',

'review_scores_cleanliness', 'review_scores_checkin',

'review_scores_communication', 'review_scores_location',

'review_scores_value', 'instant_bookable',

'calculated_host_listings_count', 'reviews_per_month']

Dropped Features

Metadata and text data:

```
drop_cols = [
    "listing_url",
    "scrape_id",
    "source",
    "name",
    "description",
    "neighborhood_overview",
    "picture_url",
    "host_url",
    "host_name",
    "host_about",
    "host_location",
    "host_thumbnail_url",
    "host_picture_url",
    "host_neighbourhood",
    "host_verifications",
```

"license",

"calendar_updated",

"calendar_last_scraped",

"bathrooms_text"

]

Data which is repeating, calculated, and not related to price:

optional_drop = [

"neighbourhood_group_cleansed",

"number_of_reviews_ltm",

"minimum_minimum_nights",

"maximum_minimum_nights",

"minimum_maximum_nights",

"maximum_maximum_nights",

"host_since",

"first_review",

"last_review",

"minimum_nights_avg_ntm",

"maximum_nights_avg_ntm",

"availability_eoy",

"number_of_reviews_ly",

'estimated_occupancy_l365d',

'estimated_revenue_l365d',

'neighbourhood',

'latitude', 'longitude','availability_60',

'availability_90', 'calculated_host_listings_count_entire_homes',

'calculated_host_listings_count_private_rooms',

'calculated_host_listings_count_shared_rooms',host_listings_count', 'host_total_listings_count',

]

Output variable

Output variable needs to be created as it is currently not ordinal data. – price (‘low’, ‘medium’, ‘high’)

Screenshot of the dataset

Due to the size of the dataset, only the first 42 rows are included and, for visibility, the dataset is shown across two pictures.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	id	last_scra	host_id	host_res	host_res	host_acc	host_is_s	host_has	host_ide	neighbou	property_	room_typ	accommo	bathroon	bedroom	beds	ame
2	39572	1/3/2025	169927	within a	100%		f	t	t	Hell's Kit	Private ro	Private ro	2		1	1	["Ha
3	39593	1/3/2025	110506	within a	100%	60%	f	t	t	Sunnysid	Private ro	Private ro	1	1	1	1	["Ha
4	39704	1/3/2025	170510				f	t	t	Bedford-	Entire to	Entire ho	4		1		["Ha
5	42300	1/3/2025	184755	within a	100%	100%	f	t	t	Lower Ea	Entire lot	Entire ho	5	2	3	4	["Ha
6	42729	1/3/2025	11481	within a	67%	33%	f	t	t	Carroll G	Private ro	Private ro	2	1	1	1	["Ha
7	43105	1/3/2025	188082	within a	100%	67%	f	t	t	Midtown	Entire rei	Entire ho	2	1	1	1	["Ha
8	44229	1/3/2025	181376				f	t	t	Fort Gree	Private ro	Private ro	2	1	1	1	["Wi
9	44973	1/3/2025	198411	within ar	100%	33%	f	t	t	Williams	Entire rei	Entire ho	2	1	0	0	["Cof
10	45542	1/3/2025	202249				f	t	f	Harlem	Entire rei	Entire ho	4		2		["Ha
11	2595	1/3/2025	2845	within a	79%	18%	f	t	t	Midtown	Entire rei	Entire ho	1	1	0	1	["Ha
12	6848	1/3/2025	15991	within a	100%	100%	t	t	t	Williams	Entire rei	Entire ho	3	1	2	1	["Ha
13	6872	1/3/2025	16104	within a	75%	33%	f	t	f	East Harl	Private ro	Private ro	1	1	1	1	["Ha
14	6990	1/3/2025	16800	within a	100%	90%	t	t	t	East Harl	Private ro	Private ro	1	1	1	1	["Cof
15	7064	1/3/2025	17297			0%	f	t	t	Williams	Private ro	Private ro	2		1		["Ha
16	7097	1/3/2025	17571	within ar	100%	100%	t	t	t	Fort Gree	Private ro	Private ro	2	1	1	2	["Ha
17	7801	1/3/2025	21207	within ar	100%	96%	t	t	t	Williams	Entire pla	Entire ho	2	1	0	1	["Ha
18	8490	1/3/2025	25183	within ar	100%		f	t	t	Bedford-	Entire lot	Entire ho	5	1	1	4	["Ha
19	45910	1/3/2025	204539	within a	100%	20%	f	t	t	Ridgewo	Entire to	Entire ho	16	2.5	5	9	["Cof
20	45935	1/3/2025	204586				f	t	t	Mott Hav	Private ro	Private ro	1	1	1	1	["Ha
21	45936	1/3/2025	867225	within a	100%	91%	t	t	t	Mornings	Private ro	Private ro	2	1	1	1	["Wi
22	46544	1/3/2025	8198			50%	f	t	t	Park Slop	Entire rei	Entire ho	2		1		["Ha
23	46911	1/3/2025	210746	within ar	100%	100%	t	t	t	Park Slop	Private ro	Private ro	2	1	1	1	["Ha
24	49048	1/3/2025	35935	within a	100%	77%	t	t	t	Bedford-	Private ro	Private ro	3	2	1	2	["Ha
25	51438	1/3/2025	236421				f	t	t	Upper Ea	Private ro	Private ro	1				["Ele
26	52689	1/3/2025	244071	within ar	100%	100%	t	t	t	Ditmars	Private ro	Private ro	2	1	2	2	["Ha
27	53469	1/3/2025	204539	within a	100%	20%	f	t	t	Middle V	Entire gu	Entire ho	4	1	1	1	["TV"
28	9357	1/3/2025	30193	within a	83%	0%	f	t	t	Hell's Kit	Entire rei	Entire ho	2				["Ha
29	10452	1/3/2025	35935	within a	100%	77%	t	t	t	Bedford-	Private ro	Private ro	2	1	1	2	["Pai
30	11943	1/3/2025	45445				f	t	f	Flatbush	Private ro	Private ro	1		1		["Wi
31	12192	1/3/2025	46978	within a	90%	80%	f	t	f	East Villa	Private ro	Private ro	2	1	1	1	["Ha
32	12940	1/3/2025	50148	within a	91%	91%	t	t	t	Bedford-	Entire rei	Entire ho	2	1	1	1	["Ha
33	14314	1/3/2025	56246	within ar	100%	43%	t	t	t	Greenpo	Entire rei	Entire ho	2	1	1	2	["Ha
34	15341	1/3/2025	60049			100%	t	t	t	Lower Ea	Entire co	Entire ho	3	1	1	2	["Ha
35	15385	1/3/2025	60252	within a	100%	100%	t	t	f	Williams	Private ro	Private ro	1	1	1	1	["Wi
36	15396	1/3/2025	60278	within a	83%	42%	t	t	t	Chelsea	Entire rei	Entire ho	4		2		["Ha
37	16580	1/3/2025	64442	within ar	100%	100%	t	t	t	East Villa	Private ro	Private ro	1	1	1	1	["Ha
38	53470	1/3/2025	204539	within a	100%	20%	f	t	t	Ridgewo	Entire rei	Entire ho	7	1	2	3	["Wi
39	53477	1/3/2025	204539	within a	100%	20%	f	t	t	Middle V	Entire to	Entire ho	12	2	4	8	["Wi
40	54466	1/3/2025	253385				f	t	t	Harlem	Private ro	Private ro	2				["Wi
41	54508	1/3/2025	210746	within ar	100%	100%	t	t	t	Park Slop	Private ro	Private ro	2	1.5	1	1	["Ha
42	54544	1/3/2025	256161	within ar	80%	78%	t	t	t	Harlem	Entire rei	Entire ho	5	1	1	4	["Ha

Figure 1, The first part of the dataset

Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ
amenitie	price	minimun	maximun	has_avai	availabil	availabil	number_	number_	review_s	review_s	review_s	review_s	review_s	review_s	review_s	instant_l	calculate	reviews_per_month	
["Hanger	$139.00	30	1125	t	30	365	44	0	4.86	4.93	4.83	4.98	5	4.98	4.86	f	2	0.25	
["Hanger	$78.00	31	150	t	0	248	34	0	4.93	4.89	4.89	4.96	4.96	4.79	4.93	f	1	0.2	
["Hangers", "Coffe		30	30	t	0	0	337	0	4.71	4.86	4.82	4.89	4.92	4.38	4.72	f	2	1.93	
["Hanger	$750.00	30	240	t	0	87	70	0	4.8	4.72	4.69	4.85	4.87	4.57	4.62	f	1	0.4	
["Hanger	$185.00	30	60	t	29	364	220	0	4.66	4.78	4.81	4.71	4.73	4.58	4.64	f	4	1.26	
["Hanger	$165.00	30	365	t	30	365	4	0	5	5	5	4.75	5	5	5	f	1	0.04	
["Wifi", "	$110.00	30	30	t	30	365	232	0	4.69	4.85	4.82	4.83	4.72	4.83	4.71	f	2	1.34	
["Coffee	$98.00	30	120	t	7	216	124	0	4.64	4.84	4.72	4.86	4.93	4.55	4.69	f	1	0.71	
["Hangers", "Wifi"		30	1125	t	0	0	18	0	4.78	4.83	4.56	5	4.83	4.44	4.67	f	1	0.23	
["Hanger	$240.00	30	1125	t	30	365	49	0	4.68	4.73	4.63	4.77	4.8	4.81	4.4	f	3	0.27	
["Hanger	$81.00	30	120	t	0	196	195	0	4.58	4.59	4.85	4.85	4.8	4.69	4.58	f	1	1.03	
["Hanger	$65.00	30	180	t	23	83	1	0	5	5	5	5	5	5	5	f	2	0.03	
["Coffee	$70.00	30	365	t	8	8	251	0	4.88	4.82	4.95	4.96	4.94	4.85	4.83	f	1	1.36	
["Hangers", "Coffe		30	45	t	0	0	13	0	4.91	5	4.91	5	5	5	5	f	2	0.07	
["Hanger	$205.00	2	1125	t	15	261	395	3	4.89	4.91	4.89	4.96	4.93	4.95	4.82	t	2	2.17	
["Hanger	$310.00	30	120	t	1	30	13	0	4.91	4.78	4.67	4.89	4.78	5	4.89	f	1	0.07	
["Hanger	$170.00	30	1125	t	10	157	190	0	4.77	4.83	4.74	4.88	4.88	4.67	4.76	f	2	1.02	
["Coffee	$500.00	30	730	t	30	365	13	0	4.42	4.64	4.36	4.82	5	4.82	4.55	f	9	0.08	
["Hanger	$60.00	30	40	t	23	83	0	0								t	1		
["Wifi", "	$75.00	31	730	t	15	274	134	0	4.66	4.61	4.48	4.87	4.87	4.89	4.63	f	2	0.79	
["Hangers", "Coffe		30	31	t	0	0	57	0	4.71	4.81	4.38	4.85	4.94	4.89	4.7	f	1	0.33	
["Hanger	$100.00	30	1125	t	2	214	74	0	4.72	4.79	4.53	4.82	4.85	4.81	4.64	f	3	0.43	
["Hanger	$95.00	30	365	t	0	253	28	0	4.44	4.33	4.22	4.26	4.37	4.41	4.3	f	6	0.17	
["Elevator", "Wash		30	31	t	0	0	0	0								f	2		
["Hanger	$150.00	4	120	t	15	138	102	1	4.97	4.98	4.91	4.97	4.96	4.88	4.9	f	1	0.6	
["TV", "Fr	$130.00	30	365	t	30	365	33	0	4.34	4.41	4.3	4.59	4.78	4.44	4.35	f	9	0.19	
["Hangers", "Smok		30	120	t	27	43	58	0	4.52	4.68	4.16	4.97	5	4.95	4.58	f	1	0.31	
["Paid dr	$90.00	30	730	t	15	350	81	0	4.66	4.49	4.62	4.74	4.85	4.42	4.67	f	6	0.45	
["Wifi", "Breakfast		30	730		0	0	0	0								f	1		
["Hanger	$73.00	30	1125	t	0	123	316	0	4.4	4.6	4.06	4.81	4.85	4.71	4.53	f	2	1.71	
["Hanger	$114.00	30	365	t	18	235	79	0	4.55	4.64	4.45	4.73	4.57	4.11	4.41	f	1	0.43	
["Hanger	$115.00	30	1125	t	0	248	179	0	4.82	4.88	4.91	4.91	4.93	4.8	4.79	f	1	0.98	
["Hanger	$165.00	30	1125	t	30	90	40	0	4.59	4.68	4.45	4.93	4.83	4.75	4.4	f	1	0.29	
["Wifi", "	$96.00	30	30	t	0	301	59	0	4.86	4.82	4.82	4.86	4.86	4.86	4.68	f	1	0.33	
["Hangers", "Babys		180	730	t	0	129	5	0	5	4.8	4.8	5	5	4.8	4.8	f	1	0.05	
["Hanger	$102.00	1	40	t	4	139	608	2	4.8	4.84	4.78	4.79	4.79	4.74	4.69	f	2	4.39	
["Wifi", "	$248.00	30	730	t	30	365	5	0	4	4.75	4.25	4.5	5	4.25	4.25	f	9	0.03	
["Wifi", "	$350.00	30	730	t	30	365	16	0	4.71	4.5	4.44	4.81	4.63	4.69	4.44	f	9	0.1	
["Wifi", "Single lev		30	60	t	30	365	0	0								f	1		
["Hanger	$103.00	3	1125	t	2	261	194	4	4.78	4.85	4.65	4.94	4.91	4.95	4.83	f	3	1.12	
["Hanger	$100.00	30	90	t	3	58	143	0	4.65	4.71	4.58	4.53	4.47	4.49	4.68	f	4	0.84	

Figure 2, The second part of the dataset

Data Dictionary

The data dictionary embedded below is collected from the Insisdeairbnb website. The file contains each feature of the original dataset and its description. In case the embedding does not work, the data dictionary has also been submitted alongside the report.

Inside Airbnb Data Dictionary - listings.cs

III Previous Work

Many research papers and projects have been published on Airbnb price prediction using machine learning. Most of them focus on specific cities or combine data from several cities, usually for short periods of about 30 days. These studies often apply regression models to predict exact prices and use open data collected from the Inside Airbnb platform. For our

work, we looked for research related specifically to **New York City**, since it offers one of the largest and most detailed Airbnb datasets. By studying existing approaches, we identified methods and challenges that helped us design our own project, which analyzes a full year of data and approaches the task as a classification problem rather than regression.

Case A: Predicting New York City Rent through Machine Learning — Based on Airbnb Data (Gong, 2025, ACM IoTML Conference)

- Paper URL: https://dl.acm.org/doi/10.1145/3749566.3749594
- **Summary:**

  This study focuses on predicting Airbnb rental prices in New York City using twelve monthly datasets from 2024. The author compares multiple machine learning models - Ridge Regression, Decision Tree, Random Forest, and XGBoost and performs detailed feature engineering, one-hot encoding, and hyperparameter tuning.

- Our contribution:

  Our project extends this work by using a **newer time range** (from **November 2024 to October 2025**) to include seasonal effects and capture market variation across months. Unlike Gong's regression-based approach, we perform **price classification** (*low, medium, high*) to simplify analysis and interpretation. We also improve data cleaning and encoding to handle changes between monthly scrapes and analyze whether the most important factors (e.g., room type, location) remain consistent throughout the year.

Case B: Predicting US Airbnb Listing Prices by Machine Learning Models (Yang, 2024, EMFRM Conference Proceedings)

- Paper URL: https://drpress.org/ojs/index.php/HBEM/article/view/16483
- **Summary:**

  This paper studies how machine learning can predict Airbnb prices in the U.S. using open data from 2023. It compares three models (Linear Regression, Random Forest, and XGBoost) and evaluates them with R^2, MSE, and RMSE metrics. After data cleaning and parameter tuning, the XGBoost model performed best, showing that ensemble methods can capture complex price patterns more accurately than basic models. The author emphasizes that good data quality and model selection are key for trustworthy price predictions.

- Our contribution:

  Our project extends this idea by focusing on **New York City listings across 12 months (Nov 2024 – Oct 2025)** instead of a single dataset. Rather than predicting exact prices, we use **classification ('low', 'medium', 'high')** to group listings by price range. We also handle real-world data issues like missing values, inconsistent formats, and class imbalance caused by monthly scraping. This makes our approach more robust and better suited for understanding seasonal and market patterns throughout the year.

IV Potential Challenges

After the initial review of the dataset, while being careful to not explore the data too much to avoid bias and data leakage, several main challenges have been identified. The dataset is complex and offers a great project opportunity for this course.

Challenge: Combining the Datasets

Description: The website lists each month as a separate .csv file, we have combined all available months to get a full year of records (specifically November 2024 – October 2025). The same listing may appear monthly with different prices, reviews, and availability. Treating these as duplicates would remove valuable temporal variation. But not all listings are scraped on the same day within one month.

- Solution idea: Keep all observations and use 'last_scraped' to create time-aware features (month/season)

Challenge: Messy Real-World Data

- Description: Missing values, incomplete values, outliers, complex text fields, symbols in numerical fields.
- Solution idea: Investigate optimal method of handling missing data/outlies, and use it, due to the dataset size, there is no restriction on which method can be used. Most complex text fields have been dealt with already by removing the feature completely. Clean and convert columns with symbols to numeric values.

Challenge : Data Leakage

- Description: Because the data spans 12 months (November 2024 – October 2025) and listings appear multiple times, a random split could mix future data into the training set and cause data leakage. The challenge is to create a realistic test set that reflects unseen, future listings or months.
- Solution idea: Split by ID which would ensure that a property (across all months) would completely be in either train or test set. This would ensure that, for example, the train data does not contain data about property X from June while the test contains data about property X from July. This would give a false sense of accuracy.

Challenge : High Cardinality in Categorical Features

- Description: Variables such as 'neighbourhood_cleansed' have many unique categories.

- Solution idea: Use target encoding, frequency encoding, or group infrequent categories into an “Other” group.

Challenge : Creating Target Column

- Description: The dataset contains continuous price values in dollar format (e.g., "$139.00"), which need to be converted into categories(‘low’, ‘medium’, ‘high’) for classification. Raw prices vary by neighborhood, property type, and time, which can lead to imbalanced or misleading class boundaries if not carefully defined.
- Solution idea: As mentioned in the beginning of the document, use quantile-based binning (e.g., pd.qcut) to create balanced classes, or explore cluster-based binning (e.g., KMeans on price) to capture natural groupings.

Challenge C: Missing Values Per Listing

- Description: Missing data often occurs at the listing level, meaning if one field (like host_response_rate) is missing, many others (such as review_scores or beds) are missing too. This suggests certain listings are incomplete or inactive, reducing the overall reliability of their records.
- Solution idea: If more than ~80-90% of rows are complete, then dropping is the cleanest option and justified given the dataset size. (big)

Challenge : Data Quality Due To Scraping Issues

- Description: Since data is scraped monthly, certain listings may have incomplete or inconsistent fields, formatting errors (like string vs numeric), or stale information if the website didn’t update in time.
- Solution idea: Perform data cleaning and validation after combining files, standardize column formats, fix inconsistent data types, remove corrupted rows, and verify scraped data with logic.

Challenge : Preparing the ‘Amenities’ Feature

- Description: The amenities column is stored as a JSON-like, long text/list structure with inconsistent formatting. Each listing as a different number of tags, however the tags are consistent meaning, for example, for television there is only ‘TV’ not any other tag. (e.g., ["Lockbox", "Self check-in", "Wifi", "Kitchen"])
- Solution idea: Create binary features for key amenities, parse and create features like has_wifi, has_kitchen etc. Represent via TF-IDF(Term Frequency–Inverse Document Frequency) encoded vector. Create a count-based feature (number_of_amenities), or

further generalize into amenity categories (e.g., count of cleaning amenities, tech amenities, kitchen amenities)

Challenge : High Dimensionality

- Description: Right now this is not a problem. However, after encoding it can become a potential challenge and lead to overfitting and longer training times. A possibility exists that after expanding features (for example, by one-hot encoding neighbourhood_cleansed, property_type, and amenities), the feature count could increase into hundreds or even thousands.
- Solution idea: In case of this occurring use dimensionality reduction, feature selection, and to prevent this from occurring carefully chose which features are worth expanding/segmenting.

Challenge : Listings Age Discrimination

- Description: New listings with 0 reviews have null review scores, creating a distinct pattern that may bias classification toward listing age rather than quality. There is a fundamental difference between "new/unreviewed" vs "established but poorly rated".
- Solution idea: Separate "new listing" indicator, model then learns "new listings" are a separate pattern with different behaviour. Impute with priors, fill missing review scores using informed estimates based on equivalent listings, rather than global averages.

Challenge : Feature Scale Disparity

- Description: Numerical features range dramatically (bedrooms: 0-4 vs availability_365: 0-365 vs number_of_reviews: 0-2000) affecting distance-based algorithms.
- Solution idea: Use a scaler, e.g., StandardScaler, MinMaxScaler.

Challenge : Class Imbalance

- Description: Target variable distribution may be heavily skewed toward certain classes.
- Solution idea: Synthetic Minority Over-sampling Technique - SMOTE, class weighting, or stratified sampling when creating test set.

V Conclusion

The project proposed is a classification task of prices of Airbnb listings in New York City. The dataset was created by combining twelve datasets from the Insideairbnb website. Each

of the datasets contain scraped data from Airbnb for a different month, ranging from November 2024 to October 2025. Despite the lack of more data from different years, the dataset requires a lot of careful work. Mainly, the challenges posed are in regards to missing values, high dimensionality, data leakage, text data, creating the target feature and others.