Vysoká škola ekonomická v Praze

Fakulta informatiky a statistiky

****

Airbnb: Predikce ceny v Praze

4IT439 – Data-X – Aplikované analytické datové modely v reálných úlohách

Semestrální práce

Tým 3: Tomáš Mikulenka, Lukáš Kuthan, Adéla Smrčková, Jana Štolcová

Datum: 30. 4. 2024

Obsah

[1 Data Understanding 4](#_Toc164002798)

[2 Data Preparation 5](#_Toc164002799)

[3 Data Visualization 6](#_Toc164002800)

[4 Modelling 7](#_Toc164002801)

[5 Model interpretation 8](#_Toc164002802)

[6 Bonus tasks 9](#_Toc164002803)

Seznam obrázků

**Nenalezena položka seznamu obrázků.**

# Data Understanding

Data jsme prošli pomocí funkcí v pythonu. Zde jsou zajímavé poznatky ze zkoumání dat:

* price – nutno převést na float64, nyní se jedná o text se znakem dolaru
* bathroom\_text – text, obsahuje počet koupelen včetně informace o sdílení
* bathroom – prázdný sloupec
* neighbourhood – nepročištěný sloupec, data jsou ale již vyčištěná ve sloupci neighbourhood\_cleansed
* amenities – výčet vybavení, mohl by být použit v modelu

Pro účel tohoto reportu jsme v pythonu zpracovali následující tabulku se všemi sloupcemi:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Column name** | **Data Type** | **Non-null Count** | **Unique Count** | **Min Value** | **Max Value** | **Mean Value** |
| id | int64 | 8949 | 8949 | 3884 | 9,81E+17 | 3,21E+17 |
| listing\_url | object | 8949 | 8949 |  |  |  |
| scrape\_id | int64 | 8949 | 1 | 20230917033704 | 20230917033704 | 20230917033704 |
| last\_scraped | object | 8949 | 1 |  |  |  |
| source | object | 8949 | 2 |  |  |  |
| name | object | 8949 | 6728 |  |  |  |
| description | object | 8785 | 7599 |  |  |  |
| neighborhood\_overview | object | 4904 | 3126 |  |  |  |
| picture\_url | object | 8949 | 8627 |  |  |  |
| host\_id | int64 | 8949 | 3180 | 3128,00 | 537229454,00 | 167351187,56 |
| host\_url | object | 8949 | 3180 |  |  |  |
| host\_name | object | 8949 | 1383 |  |  |  |
| host\_since | object | 8949 | 2144 |  |  |  |
| host\_location | object | 7025 | 191 |  |  |  |
| host\_about | object | 5259 | 1479 |  |  |  |
| host\_response\_time | object | 7860 | 4 |  |  |  |
| host\_response\_rate | object | 7860 | 47 |  |  |  |
| host\_acceptance\_rate | object | 8284 | 92 |  |  |  |
| host\_is\_superhost | object | 8637 | 2 |  |  |  |
| host\_thumbnail\_url | object | 8949 | 3048 |  |  |  |
| host\_picture\_url | object | 8949 | 3048 |  |  |  |
| host\_neighbourhood | object | 8197 | 132 |  |  |  |
| host\_listings\_count | int64 | 8949 | 68 | 0,00 | 2562,00 | 21,96 |
| host\_total\_listings\_count | int64 | 8949 | 83 | 0,00 | 5324,00 | 36,50 |
| host\_verifications | object | 8949 | 6 |  |  |  |
| host\_has\_profile\_pic | object | 8949 | 2 |  |  |  |
| host\_identity\_verified | object | 8949 | 2 |  |  |  |
| neighbourhood | object | 4904 | 236 |  |  |  |
| neighbourhood\_cleansed | object | 8949 | 52 |  |  |  |
| neighbourhood\_group\_cleansed | float64 | 0 | 0 |  |  |  |
| latitude | float64 | 8949 | 5334 | 49,95 | 50,17 | 50,08 |
| longitude | float64 | 8949 | 6193 | 14,27 | 14,68 | 14,43 |
| property\_type | object | 8949 | 60 |  |  |  |
| room\_type | object | 8949 | 4 |  |  |  |
| accommodates | int64 | 8949 | 16 | 1,00 | 16,00 | 3,93 |
| bathrooms | float64 | 0 | 0 |  |  |  |
| bathrooms\_text | object | 8933 | 34 |  |  |  |
| bedrooms | float64 | 7030 | 15 | 1,00 | 34,00 | 1,57 |
| beds | float64 | 8833 | 26 | 1,00 | 50,00 | 2,61 |
| amenities | object | 8949 | 8040 |  |  |  |
| price | object | 8949 | 3121 |  |  |  |
| minimum\_nights | int64 | 8949 | 63 | 1,00 | 1115,00 | 4,97 |
| maximum\_nights | int64 | 8949 | 128 | 1,00 | 9000,00 | 590,46 |
| minimum\_minimum\_nights | int64 | 8949 | 61 | 1,00 | 1115,00 | 4,43 |
| maximum\_minimum\_nights | int64 | 8949 | 66 | 1,00 | 1115,00 | 9,64 |
| minimum\_maximum\_nights | int64 | 8949 | 97 | 1,00 | 3333,00 | 693,11 |
| maximum\_maximum\_nights | int64 | 8949 | 97 | 1,00 | 3333,00 | 751,94 |
| minimum\_nights\_avg\_ntm | float64 | 8949 | 231 | 1,00 | 1115,00 | 6,16 |
| maximum\_nights\_avg\_ntm | float64 | 8949 | 298 | 1,00 | 3333,00 | 734,00 |
| calendar\_updated | float64 | 0 | 0 |  |  |  |
| has\_availability | object | 8949 | 2 |  |  |  |
| availability\_30 | int64 | 8949 | 31 | 0,00 | 30,00 | 9,90 |
| availability\_60 | int64 | 8949 | 61 | 0,00 | 60,00 | 25,93 |
| availability\_90 | int64 | 8949 | 91 | 0,00 | 90,00 | 43,06 |
| availability\_365 | int64 | 8949 | 366 | 0,00 | 365,00 | 159,07 |
| calendar\_last\_scraped | object | 8949 | 1 |  |  |  |
| number\_of\_reviews | int64 | 8949 | 499 | 0,00 | 1693,00 | 63,26 |
| number\_of\_reviews\_ltm | int64 | 8949 | 121 | 0,00 | 385,00 | 17,72 |
| number\_of\_reviews\_l30d | int64 | 8949 | 19 | 0,00 | 23,00 | 1,50 |
| first\_review | object | 8066 | 2585 |  |  |  |
| last\_review | object | 8066 | 932 |  |  |  |
| review\_scores\_rating | float64 | 8066 | 146 | 0,00 | 5,00 | 4,69 |
| review\_scores\_accuracy | float64 | 8050 | 132 | 1,00 | 5,00 | 4,74 |
| review\_scores\_cleanliness | float64 | 8050 | 164 | 1,00 | 5,00 | 4,68 |
| review\_scores\_checkin | float64 | 8050 | 133 | 1,00 | 5,00 | 4,79 |
| review\_scores\_communication | float64 | 8051 | 129 | 1,00 | 5,00 | 4,79 |
| review\_scores\_location | float64 | 8050 | 124 | 1,00 | 5,00 | 4,77 |
| review\_scores\_value | float64 | 8050 | 136 | 1,00 | 5,00 | 4,66 |
| license | float64 | 0 | 0 |  |  |  |
| instant\_bookable | object | 8949 | 2 |  |  |  |
| calculated\_host\_listings\_count | int64 | 8949 | 53 | 1,00 | 96,00 | 17,06 |
| calculated\_host\_listings\_count\_entire\_homes | int64 | 8949 | 50 | 0,00 | 96,00 | 14,30 |
| calculated\_host\_listings\_count\_private\_rooms | int64 | 8949 | 22 | 0,00 | 59,00 | 1,82 |
| calculated\_host\_listings\_count\_shared\_rooms | int64 | 8949 | 7 | 0,00 | 77,00 | 0,78 |
| reviews\_per\_month | float64 | 8066 | 782 | 0,01 | 27,28 | 2,00 |

* + Describe the dataset, report should include at minimum:
* Indication of missing values, type of values, unique values, outliers
* Descriptive statistics where applicable
* What can you conclude about the dataset after doing some data visualization?

# Data Preparation

* Délka textu – description
* Amenities – očíslovat všechny možné položky
* Pro modelování bych vynechala sloupce scrape\_id, last\_scraped, source, name, description, neighbourhood, has\_availability, reviews\_per\_month
* Sloupec bathrooms\_text bych převedla na numerickou proměnnou do sloupce bathrooms (současně prázdný) a ještě udělat flag jestli je shared nebo ne – podle slova shared
* Převést price na číslo – odstranit dolar
* Merging date datasets – three files merged using python into one, condition was to take always the newest records (e.g. from the june file it was only records up to 16.9.2023).
* Neighbourhoods – the same, kept only one
* Reviews – contains the same data, kept the newest with newest reviews
  + Prepare data to be ingested in the model of your choice. Provide a rationale for the features chosen as part of the training data. If you decide to remove or add columns or features explain why

# Data Visualization

* + Provide relevant plots that can be used to explain your model performance or rationale to choose predictors or features.
  + What can you conclude about the dataset after doing some data visualization?

# Modelling

* + Try different models and provide a rationale for your selected model choice and architecture. Describe your validation process. Your model report must include the following:n
* Model limitations and considerations
* Ideas to improve the model
* Explain how you chose the values for the hyper-parameters of your model

# Model interpretation

* + Use appropriate methods to interpret the impact of your features on the predictions.
  + Try to interpret main interactions of the most influential features.

# Bonus tasks

* + Analyse the relation between the sentiment and price. Were people who paid more also more satisfied?
  + What high seasons did you identify? How do the seasons differ for different locations and estate types?