Data Preparation Change Log

|  |  |  |  |
| --- | --- | --- | --- |
| No. | Variable Name | Issue | Action |
| 1 | host\_id | Data should not be in continuous modelling type. | Change to nominal data type. |
| 2 | host\_response\_rate | Data should not be character and nominal modelling type. | Change to numeric and continuous modelling type. |
| 3 | host\_acceptance\_rate | Data should not be character and nominal modelling type. | Change to numeric and continuous modelling type. |
| 4 | id | Data irrelevant to analysis. | Hide and exclude data. |
| 5 | last\_scraped | Data irrelevant to analysis. | Hide and exclude data. |
| 6 | listing\_url | Data irrelevant to analysis. | Hide and exclude data. |
| 7 | host\_url | Data irrelevant to analysis. | Hide and exclude data. |
| 8 | host\_name | Data irrelevant to analysis. | Hide and exclude data. |
| 9 | host\_location | Data irrelevant to analysis. | Hide and exclude data. |
| 10 | scrape\_id | Data irrelevant to analysis. | Hide and exclude data. |
| 11 | picture\_url | Data irrelevant to analysis. | Hide and exclude data. |
| 12 | host\_verifications | Data irrelevant to analysis. | Hide and exclude data. |
| 13 | host\_has\_profile\_pic | Data irrelevant to analysis. | Hide and exclude data. |
| 14 | host\_thumbnail\_url | Data irrelevant to analysis. | Hide and exclude data. |
| 15 | host\_picture\_url | Data irrelevant to analysis. | Hide and exclude data. |
| 16 | host\_neighbourhood | Data irrelevant to analysis. | Hide and exclude data. |
| 17 | host\_identity\_verified | Data irrelevant to analysis. | Hide and exclude data. |
| 18 | latitude | Data irrelevant to analysis. | Hide and exclude data. |
| 19 | longitude | Data irrelevant to analysis. | Hide and exclude data. |
| 20 | has\_availability | Data irrelevant to analysis considering data was taken during covid-19 period. | Hide and exclude data. |
| 21 | availability\_30 | Data irrelevant to analysis considering data was taken during covid-19 period. | Hide and exclude data. |
| 22 | availability\_60 | Data irrelevant to analysis considering data was taken during covid-19 period. | Hide and exclude data. |
| 23 | availability\_90 | Data irrelevant to analysis considering data was taken during covid-19 period. | Hide and exclude data. |
| 24 | availability\_365 | Data irrelevant to analysis considering data was taken during covid-19 period. | Hide and exclude data. |
| 25 | neighbourhood | Data irrelevant to analysis. | Hide and exclude data. |
| 26 | calendar\_updated | Data irrelevant to analysis. | Hide and exclude data. |
| 27 | calendar\_last\_scraped | Data irrelevant to analysis. | Hide and exclude data. |
| 28 | license | Data irrelevant to analysis. | Hide and exclude data. |
| 29 | name | Data irrelevant to analysis. | Hide and exclude data. |
| 30 | neighbourhood\_overview | Data irrelevant to analysis. | Hide and exclude data. |
| 31 | host\_about | Data irrelevant to analysis. | Hide and exclude data. |
| 32 | host\_id | Data irrelevant to analysis. | Hide and exclude data. |
| 33 | host\_since | Data irrelevant to analysis. | Hide and exclude data. |
| 34 | first review | Data irrelevant to analysis. | Hide and exclude data. |
| 35 | last review | Data irrelevant to analysis. | Hide and exclude data. |
| 36 | description | Data irrelevant to analysis. | Hide and exclude data. |
| 37 | instant\_bookable | Data irrelevant to analysis. | Hide and exclude data. |
| 38 | number\_of\_reviews\_ltm | Data irrelevant to analysis considering data was taken during covid-19 period. | Hide and exclude data. |
| 39 | number\_of\_review\_l30d | Data irrelevant to analysis considering data was taken during covid-19 period. | Hide and exclude data. |
| 40 | host\_neighbourhood | High correlatiion with neighbourhood\_cleansed | Hide and exclude data. Keep neighbourhood\_cleansed. |
| 41 | host\_listing\_counts | High correlatiion with host\_total\_listing\_counts | Hide and exclude data. Keep host\_total\_listing\_counts. |
| 42 | minimum\_minimum\_nights | High correlation with minimum\_nights | Hide and exclude data. Keep minimum\_nights |
| 43 | maximum\_minimum\_nights | High correlation with minimum\_nights | Hide and exclude data. Keep minimum\_nights |
| 44 | minimum\_avg\_ntm | High correlation with minimum\_nights | Hide and exclude data. Keep minimum\_nights |
| 45 | minimum\_maximum\_nights | High correlation with maximum\_nights | Hide and exclude data. Keep maximum\_nights. |
| 46 | maximum\_maximum\_nights | High correlation with maximum\_nights | Hide and exclude data. Keep maximum\_nights. |
| 47 | maximum\_avg\_ntm | High correlation with maximum\_nights | Hide and exclude data. Keep maximum\_nights. |
| 48 | bathrooms | Data is missing | Hide and exclude data. |
| 49 | review\_scores\_rating | Data has 1811/4221 missing rows. | Hide and exclude data. |
| 50 | review\_scores\_accuracy | Data has 1866/4221 missing rows. | Hide and exclude data. |
| 51 | review\_scores\_cleanliness | Data has 1865/4221 missing rows. | Hide and exclude data. |
| 52 | review\_scores\_checkin | Data has 1866/4221 missing rows. | Hide and exclude data. |
| 53 | review\_scores\_communication | Data has 1865/4221 missing rows. | Hide and exclude data. |
| 54 | review\_scores\_location | Data has 1867/4221 missing rows. | Hide and exclude data. |
| 55 | review\_scores\_value | Data has 1867/4221 missing rows. | Hide and exclude data. |
| 56 | reviews\_per\_month | Data has 1811/4221 missing rows. | Hide and exclude data. |
| 57 | host\_response\_rate | Data has 773/4221 missing rows. | Hide and exclude data. |
| 58 | host\_acceptance\_rate | Data has 1052/4221 missing rows. | Hide and exclude data. |
| 59 | property\_type | Similar to room type | Hide and exclude property\_type. Keep room\_type. |
| 60 | number\_of\_reviews | Data is disproportionately skewed-left, with majority having less reviews. Number of reviews without review score irrelevant to analysis. | Hide and exclude data. |
| 61 | calculated\_host\_listings  \_count | Repeated from total\_host\_listings\_count | Hide and exclude data. |
| 62 | calculated\_host\_listings  \_count\_entire\_homes | Data is disproportionately skewed left. Use total\_host\_listings\_count instead. | Hide and exclude data. |
| 63 | calculated\_host\_listings  \_count\_private\_rooms | Data is disproportionately skewed left. Use total\_host\_listings\_count instead. | Hide and exclude data. |
| 64 | calculated\_host\_listings  \_count\_shared\_rooms | Data is disproportionately skewed left. Use total\_host\_listings\_count instead. | Hide and exclude data. |
| 65 | bathrooms\_text | Contains too many categories -45, may be difficult for analysis. | Recode into 'private' and 'shared' in 'bathroom\_type' column |
| 66 | amenities | Contains 3109 categories, will be difficult for analysis | Using text frequency analysis, choose top amenities and recoded into 16 Boolean columns. |
| 67 | host\_is\_superhost | Contains 8 rows of missing data | Hide and exclude rows with missing data |
| 68 | host\_total\_listings\_count | Contains 8 rows of missing data | Hide and exclude rows with missing data |
| 69 | bathrooms\_text | Contains 29 rows of missing data | Hide and exclude rows with missing data |
| 70 | bedrooms | Contains 451 rows of missing data | Hide and exclude rows with missing data |
| 71 | beds | Contains 78 rows of missing data | Hide and exclude rows with missing data |
| 72 | neighbourhoods\_cleansed | Several neighbourhoods such as Mandai, Pionner contains < 50 counts of data. May not be sufficient for analysis. | Hide and exclude categories with < 50 counts of data. |
| 73 | neighbourhood\_groups\_cleansed | Contains close correlation of estimates to neighbourhoods\_cleansed data, resulting in biased in model. | Hide and exclude data. Keep neighbourhoods\_cleansed. |