Московский государственный технический университет им. Н.Э. Баумана
Факультет «Информатика и системы управления»
Кафедра «Системы обработки информации и управления»

Лабораторная работа №1
Разведочный анализ данных. Исследование и визуализация данных.

ИСПОЛНИТЕЛЬ:

Федорова Антонина Алексеевна
Группа ИУ5-24М

_____

"__"_____2021 г.

Москва 2021

Цель лабораторной работы: изучение различных методов визуализация данных.
Рекомендуемые инструментальные средства можно посмотреть здесь.

Задание:

- Выбрать набор данных (датасет). Вы можете найти список свободно распространяемых датасетов здесь.

Для лабораторных работ не рекомендуется выбирать датасеты большого размера.

- Создать ноутбук, который содержит следующие разделы:

1. Текстовое описание выбранного Вами набора данных.
2. Основные характеристики датасета.
3. Визуальное исследование датасета.
4. Информация о корреляции признаков.

- Сформировать отчет и разместить его в своем репозитории на github.

New York City Airbnb Open Data

1.1 Текстовое описание набора данных

В качестве набора данных мы будем использовать набор данные о показателях Airbnb в Нью-Йорке, США (2019 г.) - [https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data (https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data)](https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data)

С 2008 года гости и хозяева используют Airbnb, чтобы расширить возможности путешествий и представить более уникальный и индивидуальный способ познания мира. Этот набор данных описывает активность и показатели листинга в Нью-Йорке, штат Нью-Йорк, за 2019 год. Этот файл данных включает всю необходимую информацию, чтобы узнать больше о хостах, географической доступности, необходимых показателях, чтобы делать прогнозы и делать выводы.

1.2 На какие вопросы можно ответить с помощью этого датасета:

- Что мы можем узнать о разных хозяевах и территориях?
- Какие признаки больше всего влияют на стоимость жилья?
- Как минимальное количество ночей влияет на частоту съема?

1.3 Какие данные входят в датасет:

- id - идентификатор объявления
- name - наименование объявления
- host ID - идентификатор хозяина
- host_name - имя хозяина
- neighbourhood_group - расположение
- neighbourhood - зона
- latitude - широта
- longitude - долгота
- room_type - тип жилья
- price - цена
- minimum_nights - минимальное количество ночей
- number_of_reviews - количество отзывов
- last_review - дата последнего отзыва
- reviews_per_month - количество отзывов за месяц
- calculated_host_listings_count - количество объявлений на хозяина
- availability_365 - количество дней, когда жилье было доступно для съема

Импорт библиотек

In [2]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
```

Чтение данных

In [3]:
```
data = pd.read_csv('/Users/a.fedorova/Desktop/AB_NYC_2019.csv')
```

In [4]:
```
data.head(5)
```

Out[4]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latit
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79

Рассмотрим основные характеристики датасета

In [5]:
```
# Размер датасета - 48895 строк, 16 колонок
data.shape
```

Out[5]:
```
(48895, 16)
```

In [6]:
```
# Список колонок
data.columns
```

Out[6]:
```
Index(['id', 'name', 'host_id', 'host_name', 'neighbourhood_group'
,
       'neighbourhood', 'latitude', 'longitude', 'room_type', 'pri
ce',
       'minimum_nights', 'number_of_reviews', 'last_review',
       'reviews_per_month', 'calculated_host_listings_count',
       'availability_365'],
      dtype='object')
```

In [7]:
```
# Список колонок с типами данных
data.dtypes
```

Out[7]:
```
id                                 int64
name                              object
host_id                            int64
host_name                         object
neighbourhood_group               object
neighbourhood                     object
latitude                         float64
longitude                        float64
room_type                         object
price                              int64
minimum_nights                     int64
number_of_reviews                  int64
last_review                       object
reviews_per_month                float64
calculated_host_listings_count     int64
availability_365                   int64
dtype: object
```

In [8]:
```
# Проверим наличие пустых значений
# Цикл по колонкам датасета
for col in data.columns:
    # Количество пустых значений - все значения заполнены
    temp_null_count = data[data[col].isnull()].shape[0]
    print('{} - {}'.format(col, temp_null_count))
```

```
id - 0
name - 16
host_id - 0
host_name - 21
neighbourhood_group - 0
neighbourhood - 0
latitude - 0
longitude - 0
room_type - 0
price - 0
minimum_nights - 0
number_of_reviews - 0
last_review - 10052
reviews_per_month - 10052
calculated_host_listings_count - 0
availability_365 - 0
```

In [9]:
```
# Основные статистические характеристки набора данных
data.describe()
```

Out[9]:

host_id	latitude	longitude	price	minimum_nights	number_of_reviews
4.889500e+04	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000
6.762001e+07	40.728949	-73.952170	152.720687	7.029962	23.274466
7.861097e+07	0.054530	0.046157	240.154170	20.510550	44.550582
2.438000e+03	40.499790	-74.244420	0.000000	1.000000	0.000000
7.822033e+06	40.690100	-73.983070	69.000000	1.000000	1.000000
3.079382e+07	40.723070	-73.955680	106.000000	3.000000	5.000000
1.074344e+08	40.763115	-73.936275	175.000000	5.000000	24.000000
2.743213e+08	40.913060	-73.712990	10000.000000	1250.000000	629.000000

Рассмотрим визуальные характеристики датасета

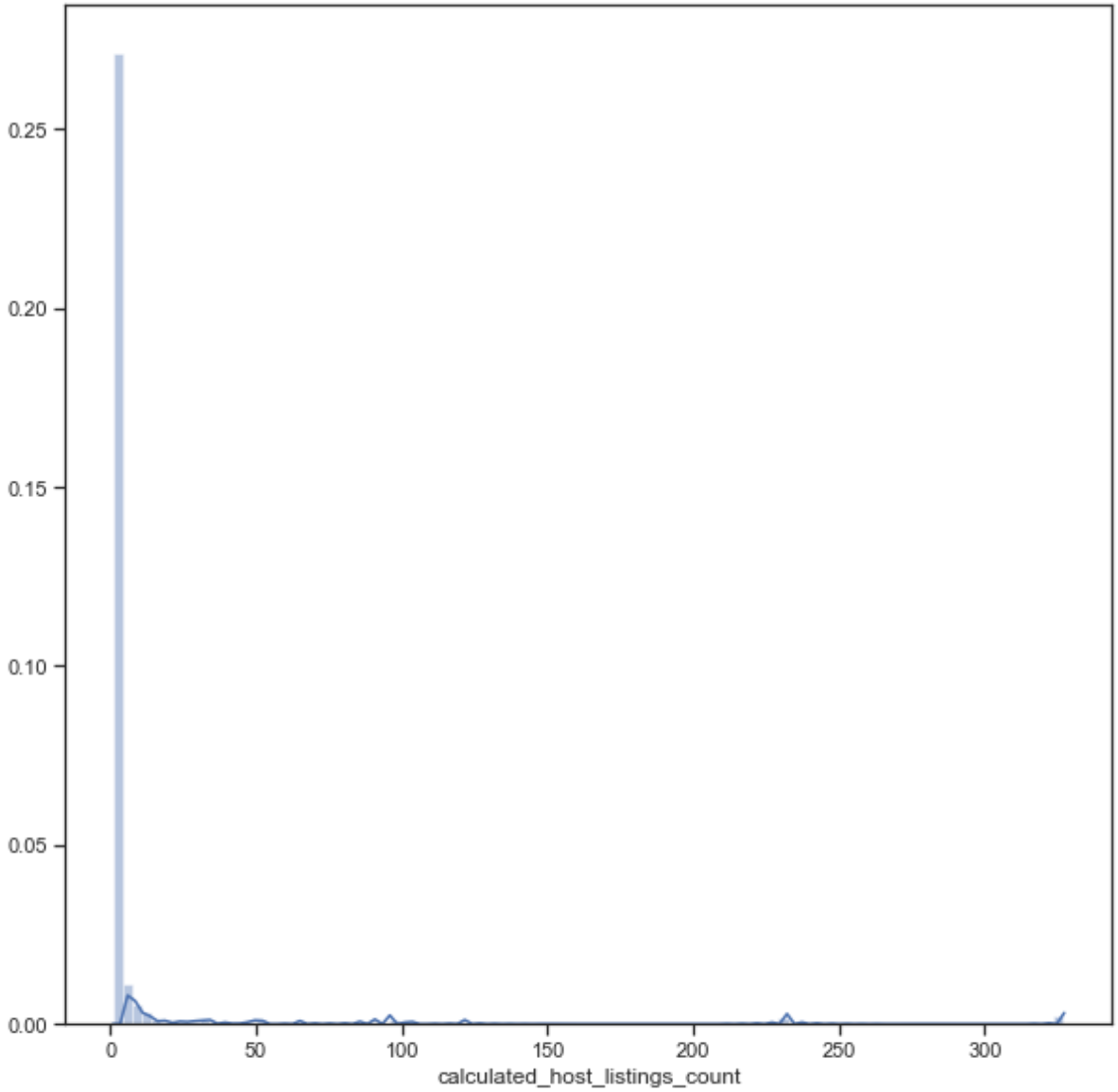
Чтобы понять, какие графики необходимо построить - попробуем ответить на вопросы, поставленные в начале работы

- Что мы можем узнать о разных хозяевах и территориях?

Для начала рассмотрим, как часто люди сдают по 1 квартире, по несколько. Это поможет понять, как много пользователей используют этот способ заработка, как основной, и как много пользователей используют его, как источник дополнительного заработка?

In [10]:
```
fig, ax = plt.subplots(figsize=(10,10))
sns.distplot(data['calculated_host_listings_count'], bins=100)
```

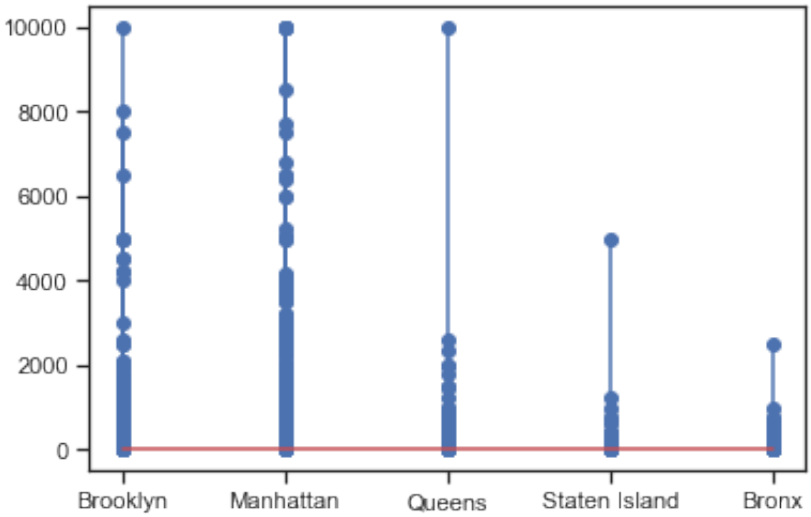
Out[10]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fb2ef4f9210>
```



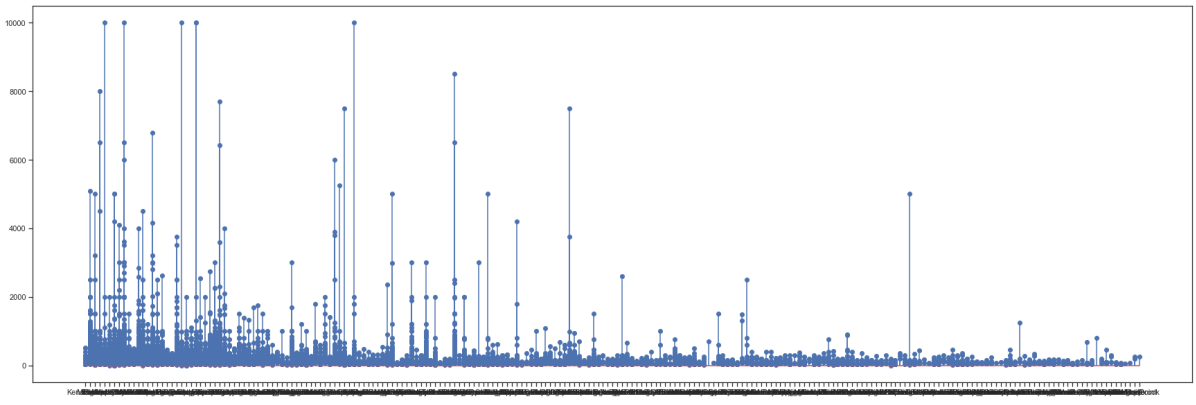
По данному графику можно заметить, что в данной выборке большая часть пользователей размещает объявления в количестве 1-3. Но максимальное значение всей выборки при этом составляет 327 объявлений на одного человека. Можно предположить, что большинство людей просто сдают для туристов свою квартиру, которая в данный момент не нужна. Но есть выбросы с большим числом объявлением, что говорит о том, что данным сайтом также пользуются и крупные арендодатели. Можно предположить, что некоторые сети отелей также используют данный сайт для сдачи жилья на короткие сроки.

Теперь рассмотрим зависимость стоимость проживания от местоположения сдаваемого жилья

In [11]:
```
plt.stem(data['neighbourhood_group'], data['price'], use_line_colle
plt.show()
```

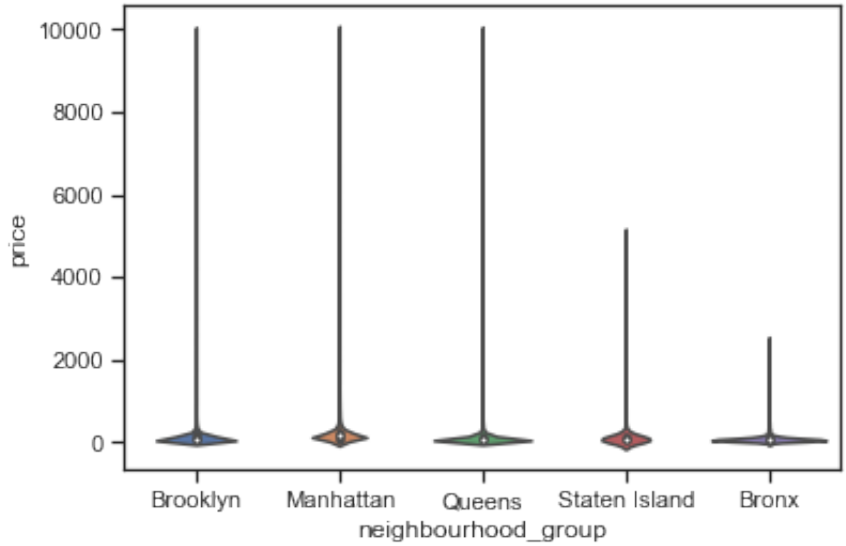


In [12]:
```
# Пыталась вывести информацию по зоне жилья, но слишком много улиц
fig, ax = plt.subplots(figsize=(30,10))
plt.stem(data['neighbourhood'], data['price'], use_line_collection
plt.show()
```



In [13]:
```
sns.violinplot( x=data["neighbourhood_group"], y=data["price"] )
```

Out[13]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fb2f0677a10>
```



Можно заметить, что есть большая разница в стоимости жилья между районом Бруклина, Манхэттена и Бронкса. Из чего можно сделать вывод, что при сдаче квартиры очень сильно на стоимость влияет район, где она находится.

- Какие признаки больше всего влияют на стоимость жилья?

Лучше всего на этот вопрос можно ответить с помощью корреляции признаков

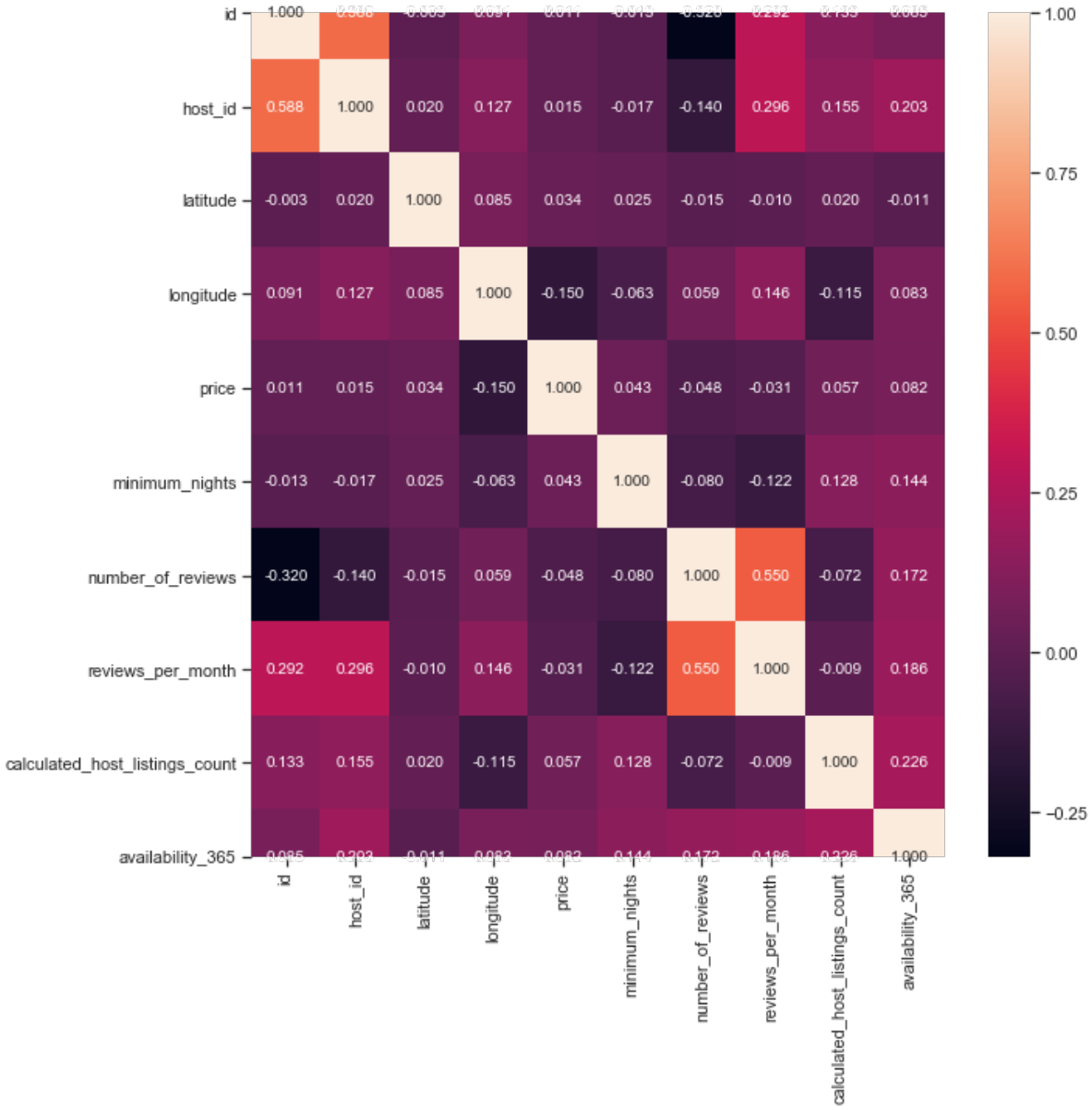
In [14]:
```
data.corr()
```

Out[14]:

	id	host_id	latitude	longitude	price	minimum
id	1.000000	0.588290	-0.003125	0.090908	0.010619	-0
host_id	0.588290	1.000000	0.020224	0.127055	0.015309	-0
latitude	-0.003125	0.020224	1.000000	0.084788	0.033939	0
longitude	0.090908	0.127055	0.084788	1.000000	-0.150019	-0
price	0.010619	0.015309	0.033939	-0.150019	1.000000	0
minimum_nights	-0.013224	-0.017364	0.024869	-0.062747	0.042799	1
number_of_reviews	-0.319760	-0.140106	-0.015389	0.059094	-0.047954	-0
reviews_per_month	0.291828	0.296417	-0.010142	0.145948	-0.030608	-0
calculated_host_listings_count	0.133272	0.154950	0.019517	-0.114713	0.057472	0
availability_365	0.085468	0.203492	-0.010983	0.082731	0.081829	0

In [15]:
```
fig, ax = plt.subplots(figsize=(10,10))
sns.heatmap(data.corr(), annot=True, fmt='.3f')
```

Out[15]: <matplotlib.axes._subplots.AxesSubplot at 0x7fb2f5516410>

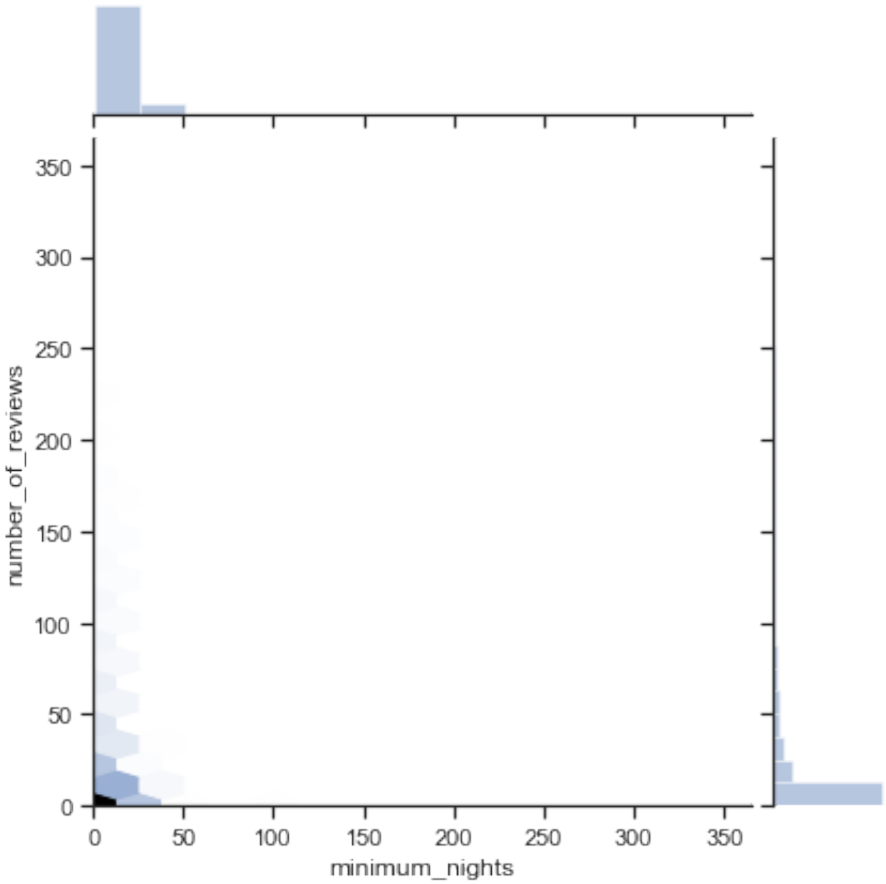


Можно отметить, что не наблюдается никаких сильно влияющих зависимостей между ценой и другими числовыми признаками. Как было отмечено ранее на цену сильное влияние оказывает именно район, в котором и располагается квартира.

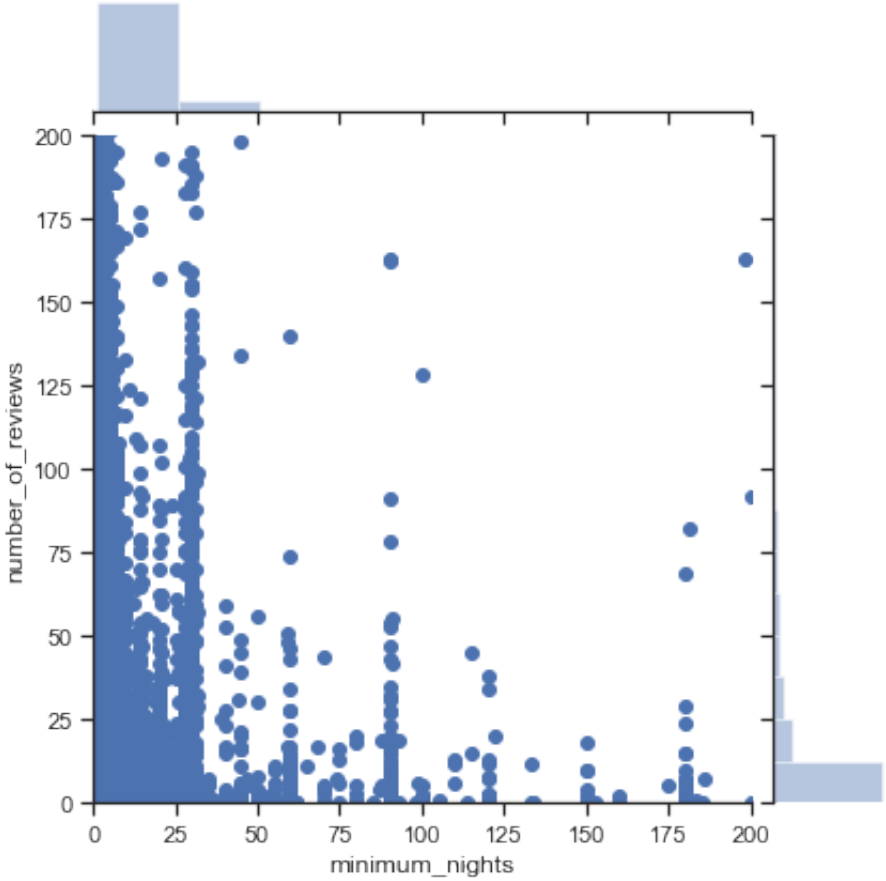
- Как минимальное количество ночей влияет на частоту съема и стоимость квартир?

В качестве определения частоты съема будем использовать количество отзывов всего и количество отзывов за месяц

In [18]:
```
plot = sns.jointplot(x=data["minimum_nights"], y=data["number_of_re
plot.ax_marg_x.set_xlim(0, 365)
plot.ax_marg_y.set_ylim(0, 365)
plt.show()
```



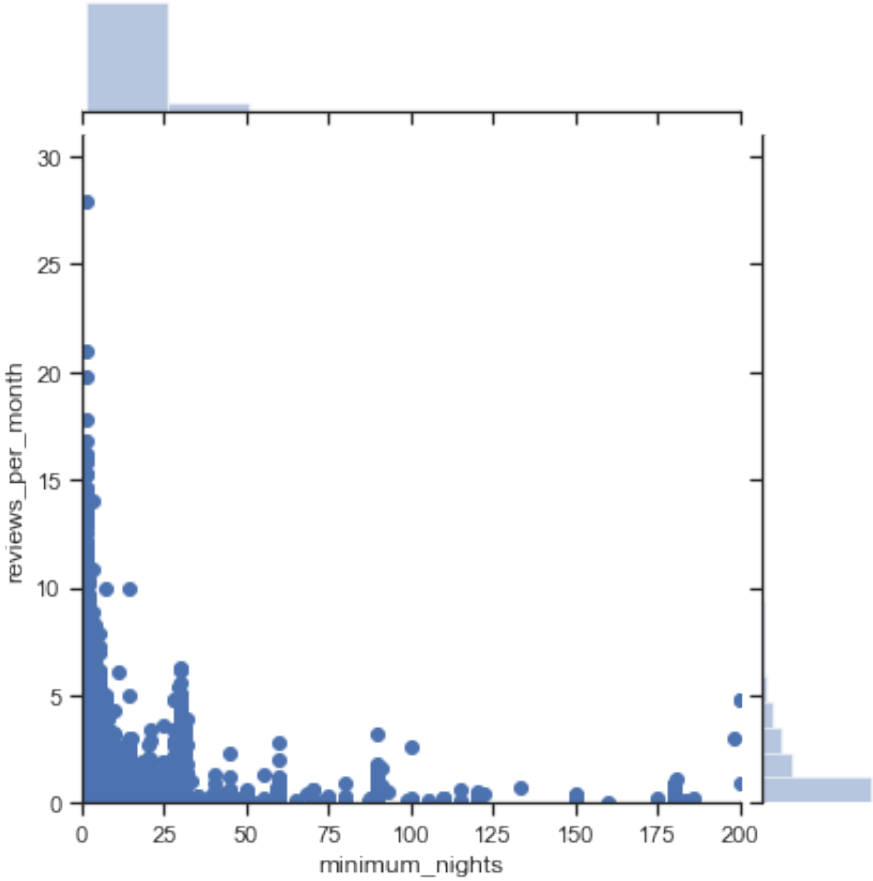
In [25]:
```
plot = sns.jointplot(x=data["minimum_nights"], y=data["number_of_re
plot.ax_marg_x.set_xlim(0, 200)
plot.ax_marg_y.set_ylim(0, 200)
plt.show()
```



Можно заметить, что чаще снимают квартиры, в которых не прописаны ограничения по количеству минимальных ночей. Это может быть по 2 причинам:

- квартира гораздо дольше по времени занята, поэтому шансов снять ее гораздо меньше
- многие приезжают в Нью-Йорк в короткий отпуск и не хотят снимать квартиру на долгий срок, поэтому большее количество людей снимают квартиры, в которых нет таких ограничений

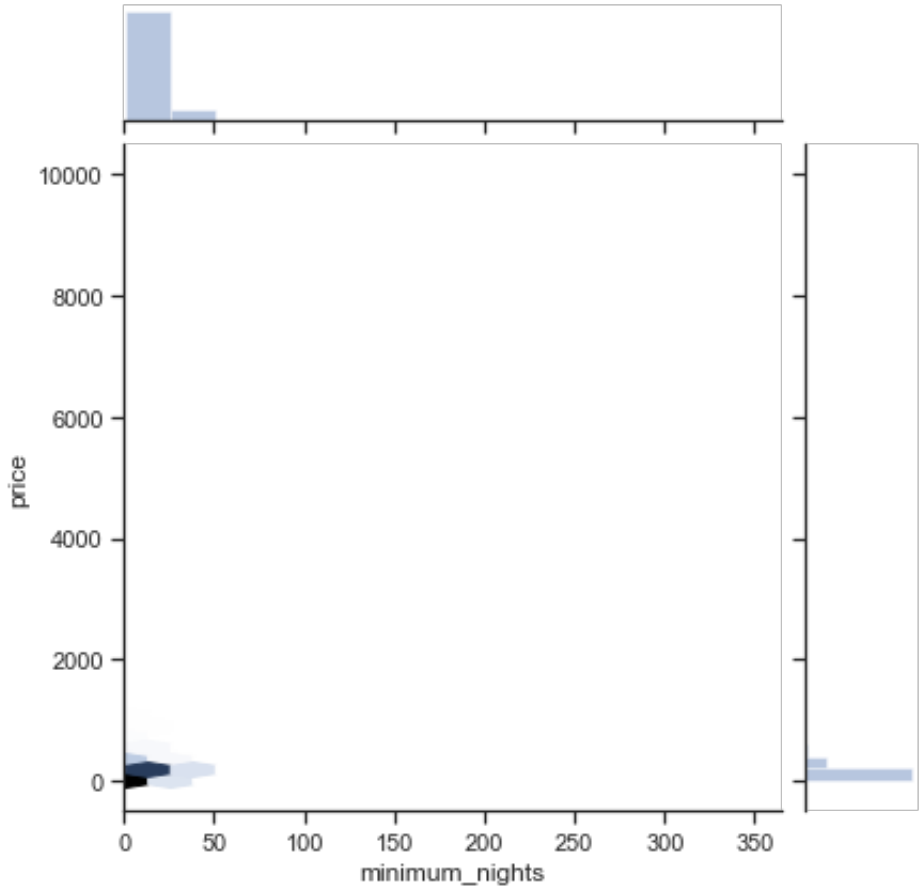
In [28]:
```
plot = sns.jointplot(x=data["minimum_nights"], y=data["reviews_per_
plot.ax_marg_x.set_xlim(0, 200)
plot.ax_marg_y.set_ylim(0, 31)
plt.show()
```



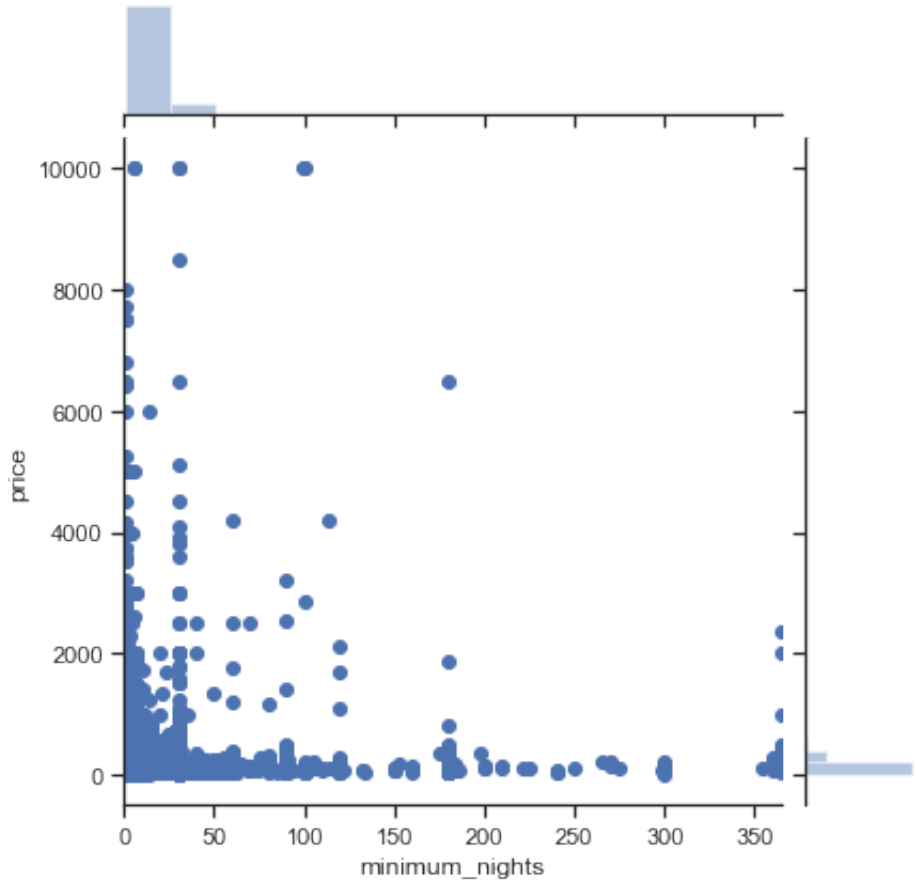
Тут ситуация та же, только теперь уже можно заметить, что некоторые квартиры, несмотря на ограничения все же сдают на меньшее число дней иногда, чем указано в информации. Это можно понять по тому, что в квартирах, которые сдаются минимум на 60 дней не может быть больше 2 отзывово за месяц, а здесь такая ситуация наблюдается и часто.

Теперь посмотрим на то, как зависит стоимость жилья от количества минимальных ночей. Есть предположение, что снимать на длительный период квартиру будет дешевле, чем снимать ее на короткий промежуток времени.

In [29]:
```
plot = sns.jointplot(x=data["minimum_nights"], y=data["price"], kin
plot.ax_marg_x.set_xlim(0, 365)
plt.show()
```



In [34]:
```
plot = sns.jointplot(x=data["minimum_nights"], y=data["price"])
plot.ax_marg_x.set_xlim(0, 365)
plt.show()
```



Можно заметить, что за исключением небольшого числа выбросов наблюдается уменьшение стоимости жилья с увеличением числа минимальных ночей в этом жилье. Тем самым можно сделать вывод, что гораздо выгоднее снимать жилье на длительный период времени.

Итоги

В ходе изучения данного датасета мы получили следующую информацию о данных:

- большинство людей просто сдают для туристов свою квартиру, которая в данный момент не нужна. Но есть выбросы с большим числом объявлением, что говорит о том, что данным сайтом также пользуеются и крупные арендодатели;
- есть большая разница в стоимости жилья между районом Бруклина, Манхэттена и Бронкса. Из чего можно сделать вывод, что при сдаче квартиры очень сильно на стоимость влияет район, где она находится;
- не наблюдается никаких сильно влияющих зависимостей между ценой и другими числовыми признаками. Как было отмечено ранее на цену сильное влияние оказывает именно район, в котором и располагается квартира;
- чаще снимают квартиры, в которых не прописаны ограничения по количеству минимальных ночей;
- некоторые квартиры, несмотря на ограничения все же сдают на меньшее число дней иногда, чем указано в информации;
- наблюдается уменьшение стоимости жилья с увеличением числа минимальных ночей в этом жилье.

Вывод

Данный датасет можно использовать для предсказания стоимости жилья, по его параметрам с целью определения оптимального дохода от сдачи с учетом частоты и продолжительности съема этого жилья.