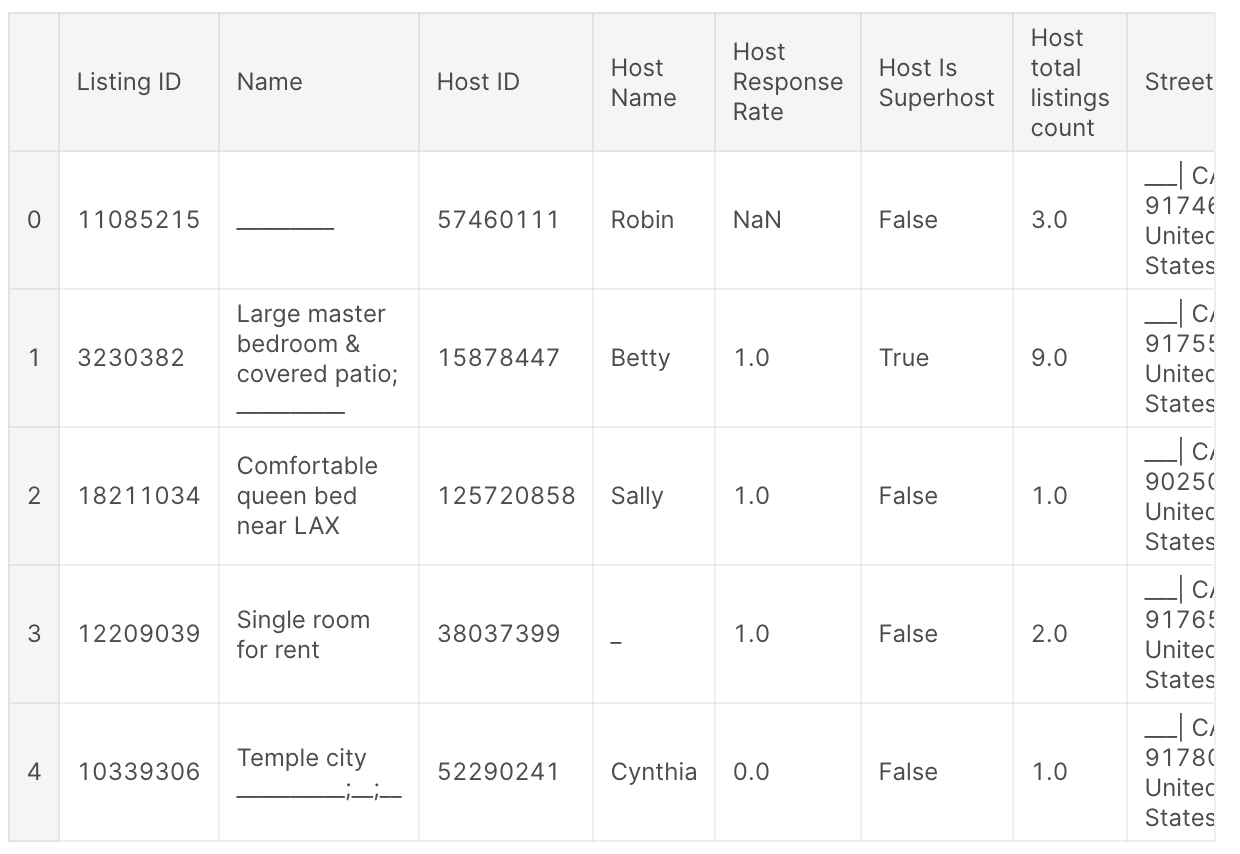
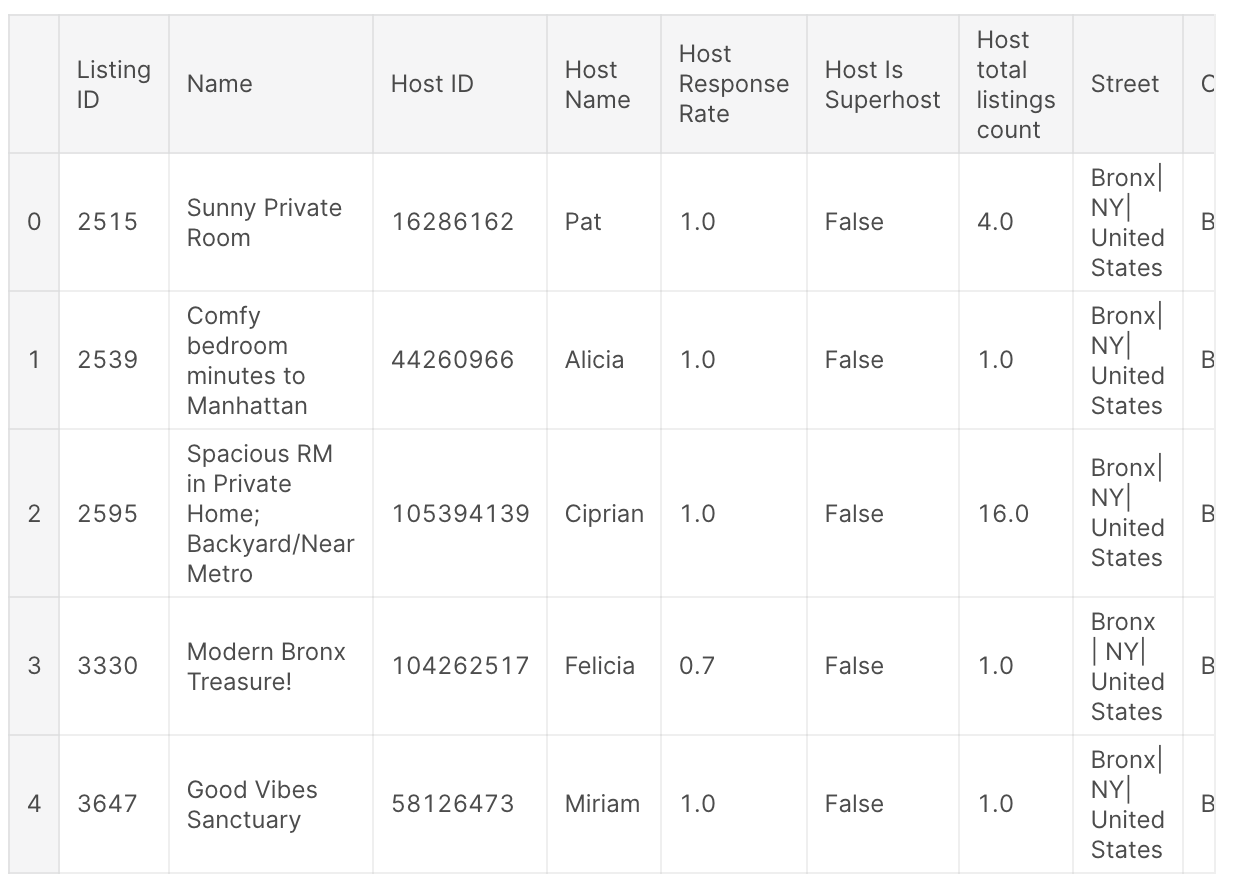
To begin the **exploratory data analysis,** we need to do the data preparation first.

We will import, combine, and filter the data we need and output an csv file for the further use.

Изображение выглядит как текст

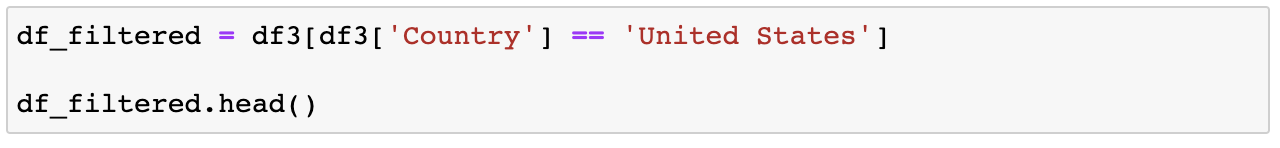
Автоматически созданное описаниеИзображение выглядит как текст

Автоматически созданное описание

Изображение выглядит как текст

Автоматически созданное описание

As the prices varies a lot in different countries It is reasonably to consider only USA listings:

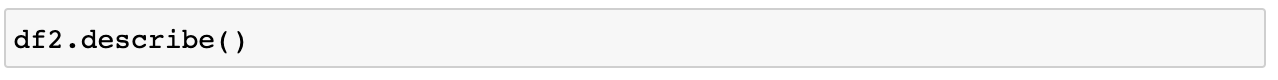


Изображение выглядит как стол

Автоматически созданное описание

Now, let's get more infomation with our datasets.

Изображение выглядит как стол

Автоматически созданное описание

Изображение выглядит как стол

Автоматически созданное описание

Изображение выглядит как стол

Автоматически созданное описание

Now, let's combine those three datasets into one:

Изображение выглядит как текст

Автоматически созданное описание

Изображение выглядит как стол

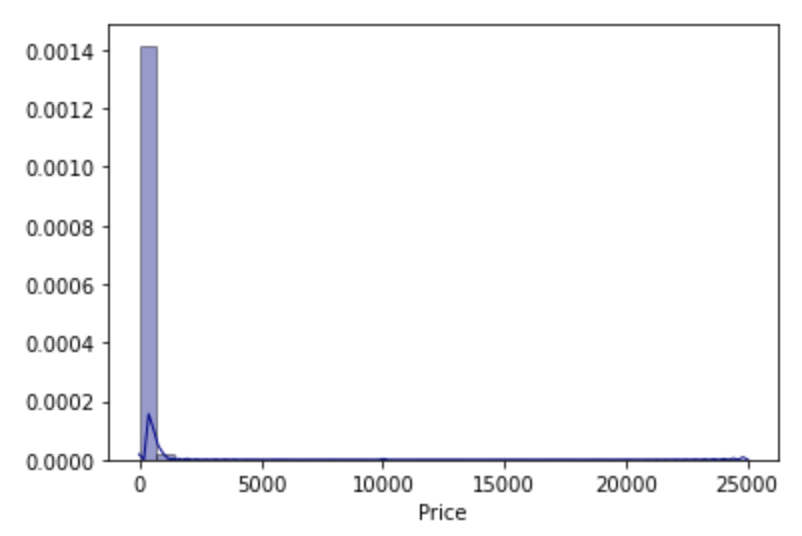
Автоматически созданное описание

Now, 'df\_final' has 295,452 lines of data and ready to use.

Let's do exploratory data analysis by examine the correlation between **Price** with **number of bedrooms**, **bathrooms** and **review scores**.

Изображение выглядит как текст

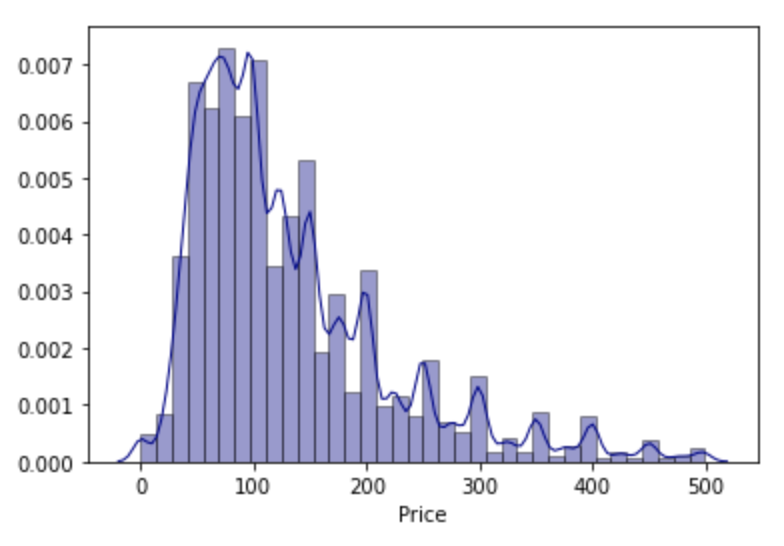
Автоматически созданное описание



From the graph we can see, because of the large range of 'Price', we need to filter those unnessary data which could influence our analysis. After observation, we found set the range from 0 to 500 is appropriate.

Изображение выглядит как текст

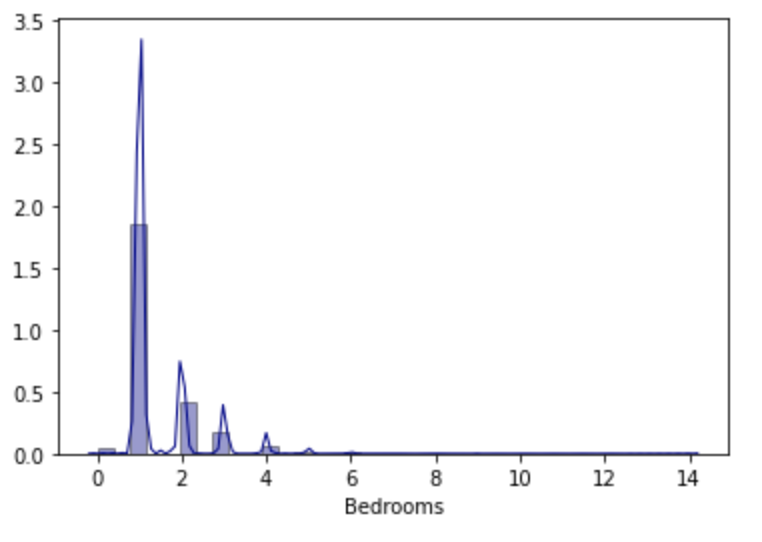
Автоматически созданное описание



Now, let's see the distribution of numbers of Bedrooms:

Изображение выглядит как текст

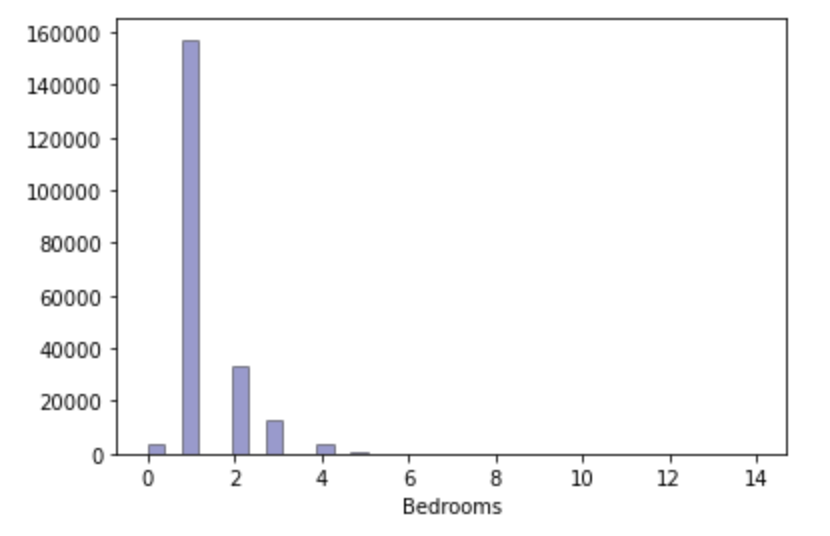
Автоматически созданное описание



We can see that most houses have bedrooms from 0 to 6, so let's filter the data:

Изображение выглядит как текст

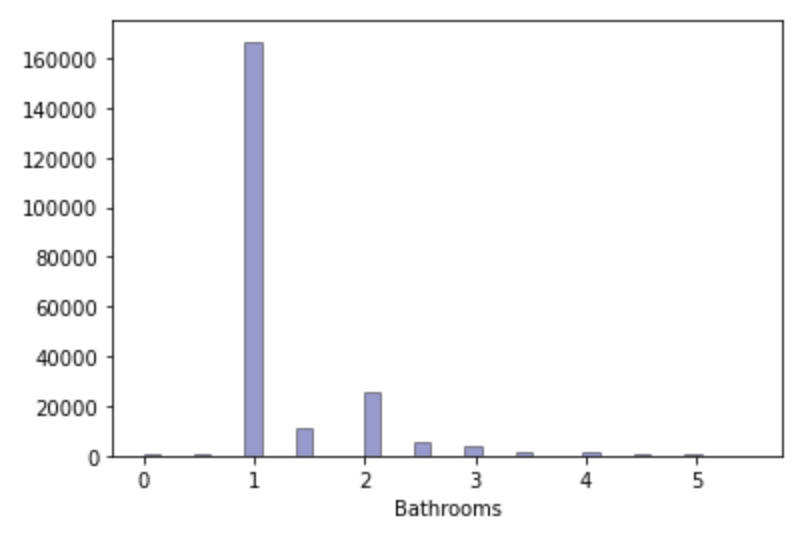
Автоматически созданное описание



From this graph, we can see that the shape of the distributions of 'Numbers of Bedrooms' and the distributions of 'Price' are very similar, which indicates the possibilities between them, and we will do further investigations later. Before that, let's do more distribution graph on other variables.

Изображение выглядит как текст

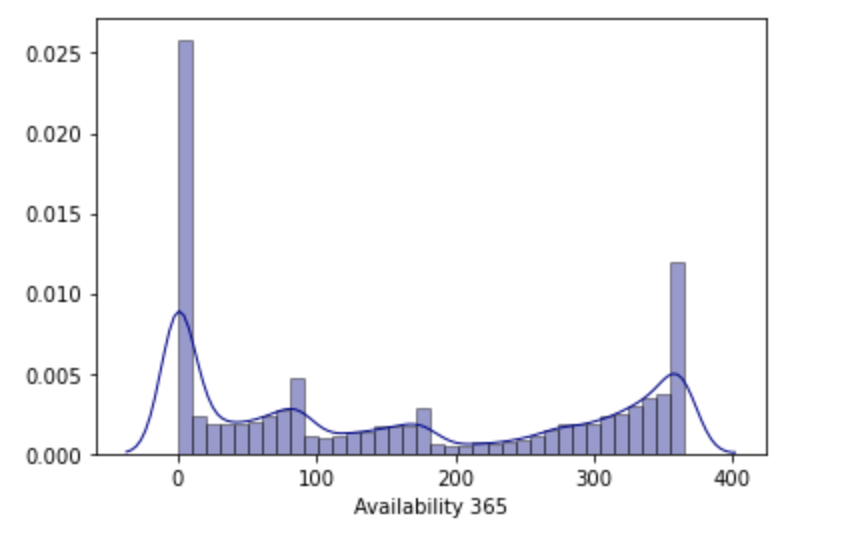
Автоматически созданное описание



Let's see the distribution with more varibles:

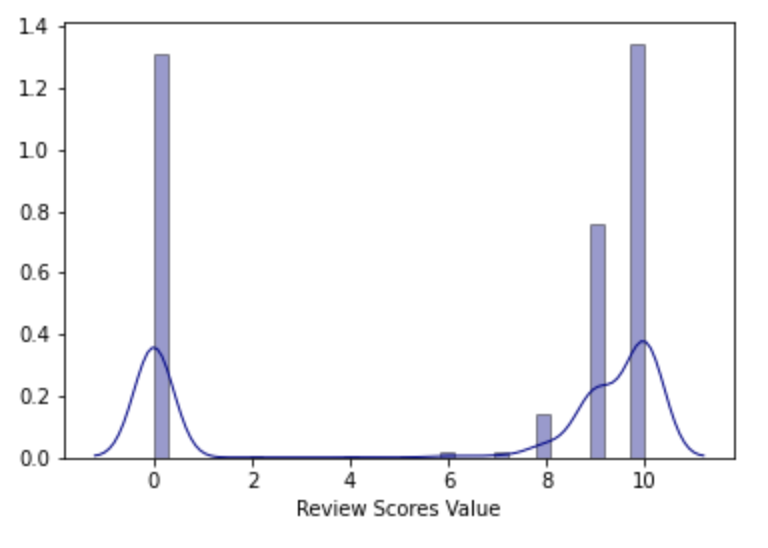
Изображение выглядит как текст

Автоматически созданное описание

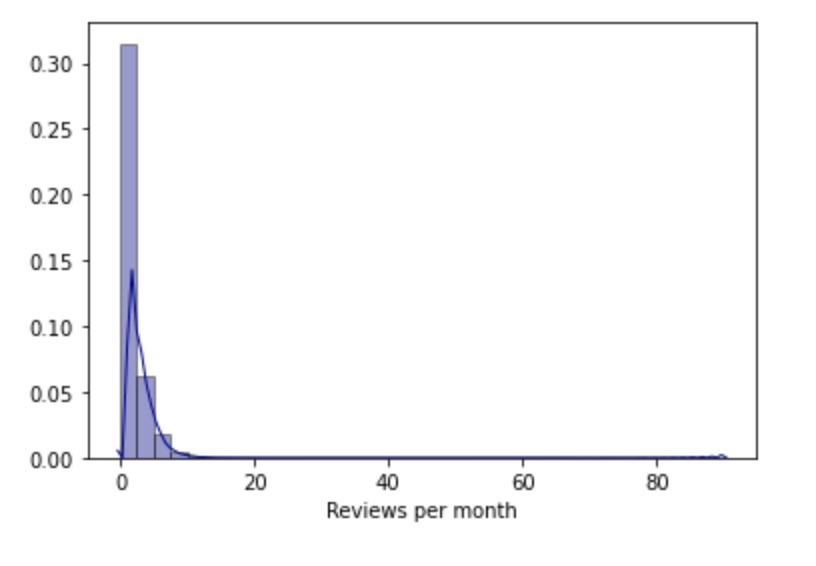


Изображение выглядит как текст

Автоматически созданное описание



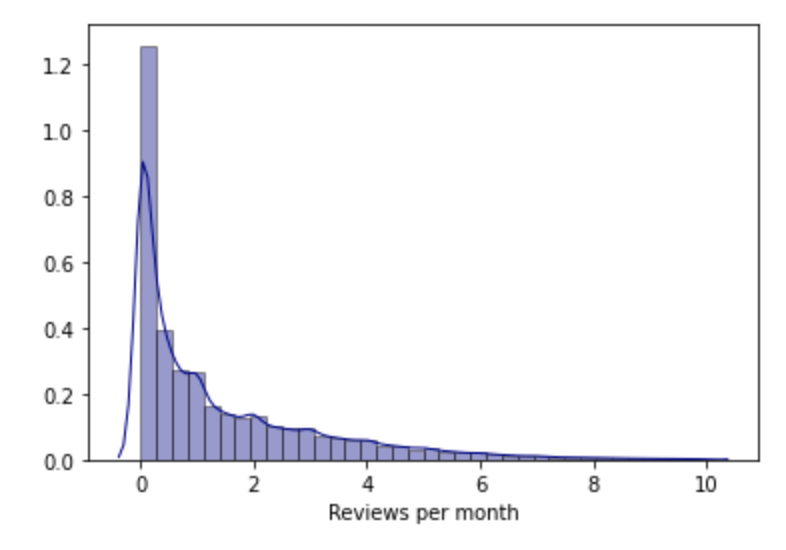
Изображение выглядит как текст

Автоматически созданное описание

Let's filter the Varible:

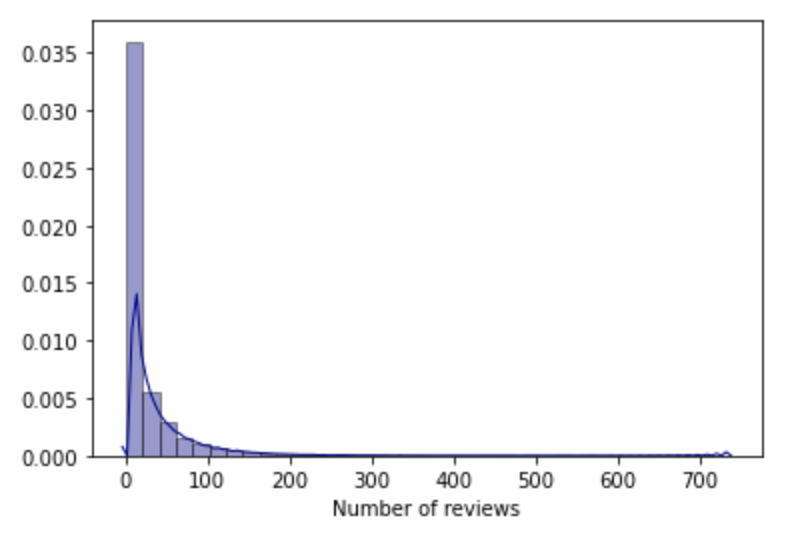
Изображение выглядит как текст

Автоматически созданное описание

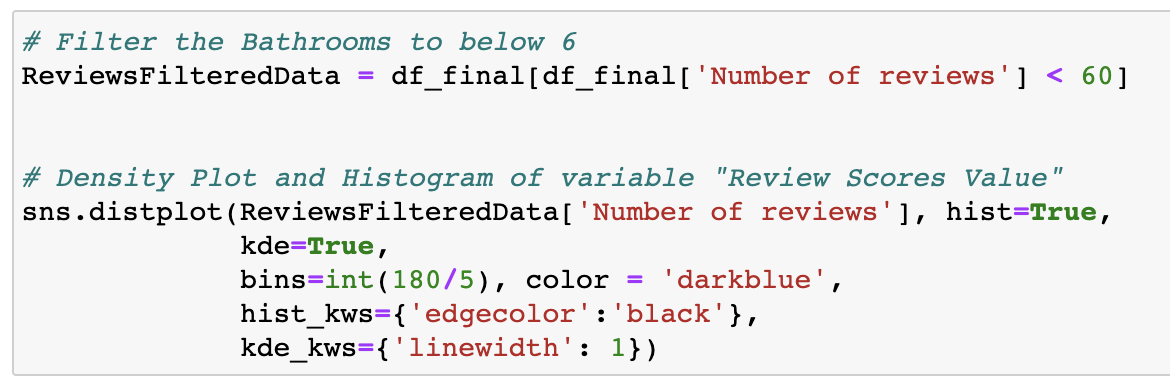


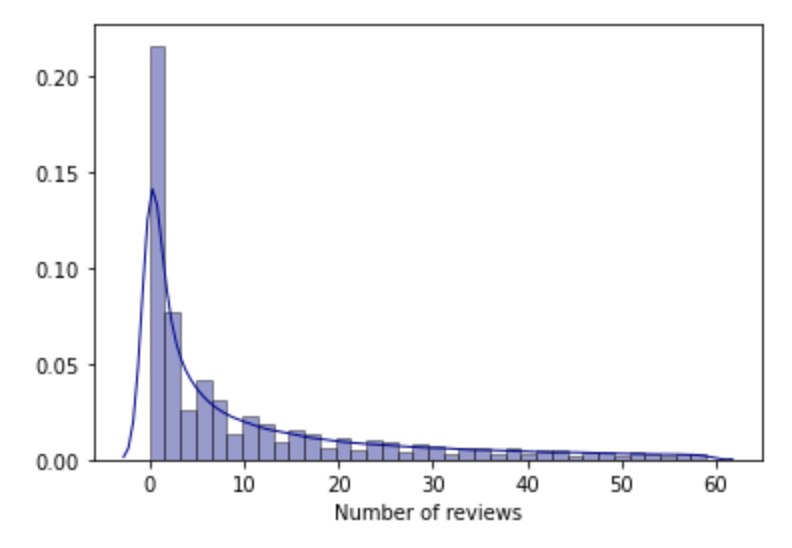
Изображение выглядит как текст

Автоматически созданное описание



Filter the data:





We can see that most listing have 0 - 10 reviews.

Next, let's do the correlation test to find out the potential correlation with *Spearman correlation test*:

Изображение выглядит как текст

Автоматически созданное описание

Изображение выглядит как текст, газета

Автоматически созданное описание

Изображение выглядит как стол

Автоматически созданное описание

Изображение выглядит как стол

Автоматически созданное описание

**Conclusions**

From the result table, we found that **'Price'** and **'Accommodates'** have a correlation coefficient of 0.55, which indicates they are *moderately* correlated, and **'number of Bedrooms'** has a correlation coefficient of 0.46 with **'Price'**, which is the second highest value in all variables, which can be understand, because more bedrooms a house has, the higher the price can be, and more people a house can accommodates, more expensive it will be.