```
In [211]:
```

```
import pandas as pd
import os
import sys
from scipy import stats
import matplotlib.pyplot as plt
import seaborn as sns
```

## **Variables**

```
In [212]:
```

```
file_path = 'C:/Users/Admin/Desktop/project/uni_project/files'
step_two_df_headings = ['location','no_of_businesses','total_reviews','percentage_of_review
pd.set_option('display.max_rows', None)
```

#### In [213]:

```
os.chdir(file_path)
category_list = []
meta_data_df = pd.DataFrame(columns=['Category Name','Number of Listings'])

id = 19162
i=1
for file in os.listdir(os.getcwd()):
    if file.endswith('.csv'):
        file = file.split('_',1)
        file = file[0]
        if(file not in category_list):
            category_list.append(file)
            print(i,": ",file)
        i = i+1
```

```
1 : Beauty salon
2 : House Cleaning Service
3 : Meta Data.csv
4 : Photography
5 : Poultry farming
6 : Real estate Investment
7 : skill teaching
```

```
cat_number = input("Enter category by number: ")
```

Enter category by number: 1

In [214]:

```
In [215]:
```

```
category_to_continue_with = category_list[int(cat_number)-1]
category_to_continue_with
```

#### Out[215]:

'Beauty salon'

#### In [216]:

```
master_df = pd.DataFrame(columns=step_two_df_headings)
for file in os.listdir(os.getcwd()):
    if file.endswith('.csv') and category_to_continue_with in file:
        temp_df = pd.read_csv(file)
        total_rows = len(temp_df) #
        temp_df.replace(',','', regex=True, inplace=True)
        temp_df = temp_df[['no_of_reviews','rating']]
        #print(temp_df)
        temp_df.dropna(inplace=True)
        loc_temp = file.split('_',1)
        loc_temp = loc_temp[1]
        loc_temp = loc_temp.replace(".csv",'')
        loc_temp = loc_temp.replace("Lahore Pakistan",'')
        location = loc_temp
        temp_df['no_of_reviews'] = pd.to_numeric(temp_df['no_of_reviews'])
        total_reviews = temp_df['no_of_reviews'].sum()
        temp_dic = {'location':location,'no_of_businesses':total_rows,'total_reviews':total
        tdf = pd.DataFrame([temp_dic])
        master_df = pd.concat([master_df, tdf], ignore_index=True)
        #print(temp_dic)
#df = pd.read_csv(f'{file_path}/{category_to_continue_with}')
#df
```

## In [217]:

print(len(master\_df))
master\_df

58

## Out[217]:

	location	no_of_businesses	total_reviews	percentage_of_reviews
0	Airline Society	170	23711	139.476471
1	Aitchison Society	170	9766.0	57.447059
2	Ali Park	170	19761	116.241176
3	Allama Iqbal Town	166	9609.0	57.885542
4	Askrai Villa 3	170	21505	126.500000
5	Askrai Villa 6	170	21591	127.005882
6	Askrai Villa 9	157	23067	146.923567
7	Baghban Pura	170	14499	85.288235
8	Bagrian	170	22768	133.929412
9	Bedian Road	170	20064	118.023529
10	Bostan Colony	170	24196	142.329412
11	Burj Colony	170	24359	143.288235
12	Cantt	170	17795	104.676471
13	Cavalry Ground	168	19275	114.732143
14	Chachowali	170	24910	146.529412
15	DHA 1	168	13427	79.922619
16	DHA 4	170	11667.0	68.629412
17	DHA 7	170	19577	115.158824
18	Faisal Town	170	17209.0	101.229412
19	Faruque Ganj	170	24924	146.611765
20	Fransisi Town	156	0.0	0.000000
21	Gawal Mandi	170	18916	111.270588
22	Ghazi Road	170	19441	114.358824
23	Gor	170	26323	154.841176
24	Gulbahar Colony	170	20240	119.058824
25	Gulberg 3	170	19285	113.441176
26	Gulshan E Ravi	169	10143.0	60.017751
27	Hamdard Chowk	170	24784	145.788235
28	Herbuns Pura	122	21731	178.122951
29	Ittehad Colony	170	9416.0	55.388235
30	Johar Town	169	8921.0	52.786982
31	Karim Park	170	15533	91.370588
32	Lake City	170	22293	131.135294

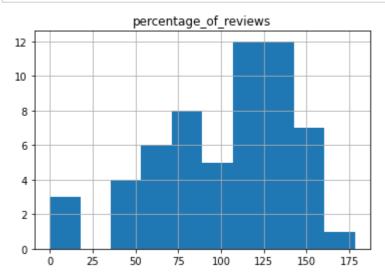
	location	no_of_businesses	total_reviews	percentage_of_reviews
33	Mazang	170	22509	132.405882
34	Nespak Society	170	24073	141.605882
35	NFC Town	170	20172	118.658824
36	Nishtar Colony	170	21295	125.264706
37	Officers Colony	170	21309	125.347059
38	Opf Housing Scheme	170	11502.0	67.658824
39	PASCO Housing Society	170	0.0	0.000000
40	PCSIR Housing Scheme	170	12812.0	75.364706
41	Punjab Society Near Dha	149	5521.0	37.053691
42	Punjab University Hostel	170	17196	101.152941
43	Punjab University	170	22116	130.094118
44	Qila Gujarsing	170	7680.0	45.176471
45	Railway Housing Society	117	9547	81.598291
46	Ram Garh	170	20913	123.017647
47	Samanabad	170	12528.0	73.694118
48	Sanat Nagar	170	20549	120.876471
49	Sanda Khurd	75	3958.0	52.773333
50	Sarshar Town	170	26862	158.011765
51	Shabir Town	170	17166	100.976471
52	Shad Bagh	168	12471	74.232143
53	Shadman Colony	169	19965	118.136095
54	Shalimar Town	170	14899	87.641176
55	Sukh Chain Society	16	0.0	0.000000
56	Suparco Colony	170	13713	80.664706
57	Wafaqi Colony	170	22721	133.652941

# Find averge of % review column by Removing outliers from dataframe below Rafy

Note: use a temporary dataframe

#### In [218]:

```
temp_DataFrame = master_df
temp_DataFrame.hist(column="percentage_of_reviews")
plt.show()
```

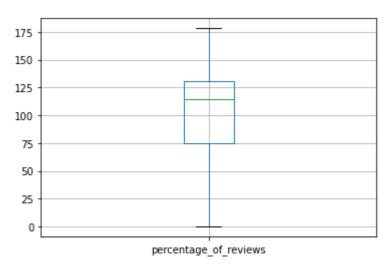


#### In [219]:

```
temp_DataFrame.boxplot(column='percentage_of_reviews')
```

#### Out[219]:

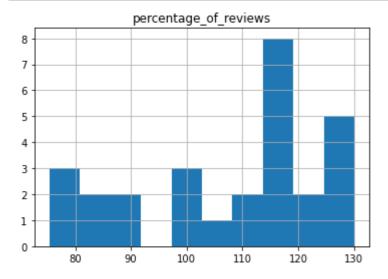
#### <AxesSubplot:>



#### In [220]:

```
q_low = temp_DataFrame["percentage_of_reviews"].quantile(0.25)
q_hi = temp_DataFrame["percentage_of_reviews"].quantile(0.75)

temp_DataFrame = temp_DataFrame[(temp_DataFrame["percentage_of_reviews"] < q_hi) & (temp_DataFrame.hist(column="percentage_of_reviews")
plt.show()</pre>
```

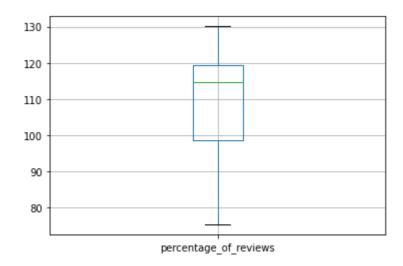


#### In [221]:

```
temp_DataFrame.boxplot(column='percentage_of_reviews')
```

#### Out[221]:

#### <AxesSubplot:>



## In [222]:

print(len(temp\_DataFrame))
temp\_DataFrame

28

## Out[222]:

	location	no_of_businesses	total_reviews	percentage_of_reviews
2	Ali Park	170	19761	116.241176
4	Askrai Villa 3	170	21505	126.500000
5	Askrai Villa 6	170	21591	127.005882
7	Baghban Pura	170	14499	85.288235
9	Bedian Road	170	20064	118.023529
12	Cantt	170	17795	104.676471
13	Cavalry Ground	168	19275	114.732143
15	DHA 1	168	13427	79.922619
17	DHA 7	170	19577	115.158824
18	Faisal Town	170	17209.0	101.229412
21	Gawal Mandi	170	18916	111.270588
22	Ghazi Road	170	19441	114.358824
24	Gulbahar Colony	170	20240	119.058824
25	Gulberg 3	170	19285	113.441176
31	Karim Park	170	15533	91.370588
35	NFC Town	170	20172	118.658824
36	Nishtar Colony	170	21295	125.264706
37	Officers Colony	170	21309	125.347059
40	PCSIR Housing Scheme	170	12812.0	75.364706
42	Punjab University Hostel	170	17196	101.152941
43	Punjab University	170	22116	130.094118
45	Railway Housing Society	117	9547	81.598291
46	Ram Garh	170	20913	123.017647
48	Sanat Nagar	170	20549	120.876471
51	Shabir Town	170	17166	100.976471
53	Shadman Colony	169	19965	118.136095
54	Shalimar Town	170	14899	87.641176
56	Suparco Colony	170	13713	80.664706

#### In [223]:

```
total_percentage_reviews = temp_DataFrame['percentage_of_reviews'].sum()
total_rows_after_outliers = temp_DataFrame.shape[0]
threshhold_for_selection = total_percentage_reviews/total_rows_after_outliers
threshhold_for_selection
```

#### Out[223]:

108.10969643281375

## Prepare a list of selected files which have greater % reviews than average Rafy

#### In [224]:

```
sel_list=[]
selected = master_df[(master_df["percentage_of_reviews"] > threshhold_for_selection)]
selected = selected[["location"]]
sel_list = selected['location'].tolist()
#sel_list.append(sel_list)
print(len(sel_list))
print(sel_list)
```

```
['Airline Society ', 'Ali Park ', 'Askrai Villa 3 ', 'Askrai Villa 6 ', 'Ask rai Villa 9 ', 'Bagrian ', 'Bedian Road ', 'Bostan Colony ', 'Burj Colony ', 'Cavalry Ground ', 'Chachowali ', 'DHA 7 ', 'Faruque Ganj ', 'Gawal Mandi ', 'Ghazi Road ', 'Gor ', 'Gulbahar Colony ', 'Gulberg 3 ', 'Hamdard Chowk ', 'Herbuns Pura ', 'Lake City ', 'Mazang ', 'Nespak Society ', 'NFC Town ', 'N ishtar Colony ', 'Officers Colony ', 'Punjab University ', 'Ram Garh ', 'San at Nagar ', 'Sarshar Town ', 'Shadman Colony ', 'Wafaqi Colony ']
```

## its my turn...

#### In [138]:

32

['Beauty salon\_Airline Society Lahore Pakistan.csv', 'Beauty salon\_Ali Park Lahore Pakistan.csv', 'Beauty salon\_Askrai Villa 3 Lahore Pakistan.csv', 'Be auty salon\_Askrai Villa 6 Lahore Pakistan.csv', 'Beauty salon\_Askrai Villa 9 Lahore Pakistan.csv', 'Beauty salon\_Bagrian Lahore Pakistan.csv', 'Beauty sa lon\_Bedian Road Lahore Pakistan.csv', 'Beauty salon\_Bostan Colony Lahore Pak istan.csv', 'Beauty salon\_Burj Colony Lahore Pakistan.csv', 'Beauty salon\_Ca valry Ground Lahore Pakistan.csv', 'Beauty salon\_Chachowali Lahore Pakistan. csv', 'Beauty salon\_DHA 7 Lahore Pakistan.csv', 'Beauty salon\_Faruque Ganj L ahore Pakistan.csv', 'Beauty salon\_Gawal Mandi Lahore Pakistan.csv', 'Beauty salon\_Ghazi Road Lahore Pakistan.csv', 'Beauty salon\_Gor Lahore Pakistan.cs v', 'Beauty salon\_Gulbahar Colony Lahore Pakistan.csv', 'Beauty salon\_Gulber g 3 Lahore Pakistan.csv', 'Beauty salon\_Hamdard Chowk Lahore Pakistan.csv', 'Beauty salon\_Herbuns Pura Lahore Pakistan.csv', 'Beauty salon\_Lake City Lah ore Pakistan.csv', 'Beauty salon\_Mazang Lahore Pakistan.csv', 'Beauty salon\_ Nespak Society Lahore Pakistan.csv', 'Beauty salon\_NFC Town Lahore Pakistan. csv', 'Beauty salon\_Nishtar Colony Lahore Pakistan.csv', 'Beauty salon\_Offic ers Colony Lahore Pakistan.csv', 'Beauty salon\_Punjab University Lahore Paki stan.csv', 'Beauty salon\_Ram Garh Lahore Pakistan.csv', 'Beauty salon\_Sanat Nagar Lahore Pakistan.csv', 'Beauty salon\_Sarshar Town Lahore Pakistan.csv', 'Beauty salon\_Shadman Colony Lahore Pakistan.csv', 'Beauty salon\_Wafaqi Colo ny Lahore Pakistan.csv']

#### In [225]:

```
dataframes_list = []

# append datasets into the list
for i in range(len(prepared_list)):
    temp_df = pd.read_csv(file_path+'/'+prepared_list[i])
    temp_df = temp_df[['company','no_of_reviews','rating','address1']]
    temp_df.replace(',','', regex=True, inplace=True)
    temp_df.dropna(inplace=True)
    temp_df['no_of_reviews'] = pd.to_numeric(temp_df['no_of_reviews'])
    temp_df['rating'] = pd.to_numeric(temp_df['rating'])
    dataframes_list.append(temp_df)
```

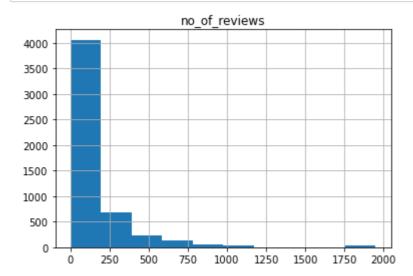
#### In [227]:

```
master_df2 = pd.DataFrame(columns=step_three_df_headings)
combining_dfs = [df2.set_index('company') for df2 in dataframes_list]
master_df2 = pd.concat(combining_dfs, axis=0)
print(len(master_df2))
master_df2
```

5223

#### In [228]:

```
master_df2_temp = master_df2
master_df2_temp.hist(column="no_of_reviews")
plt.show()
```

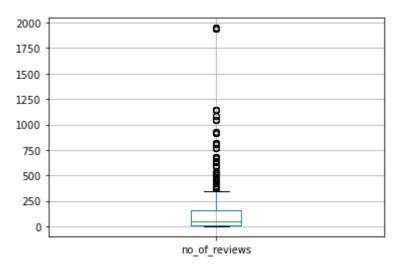


#### In [229]:

```
master_df2_temp.boxplot(column='no_of_reviews')
```

#### Out[229]:

### <AxesSubplot:>

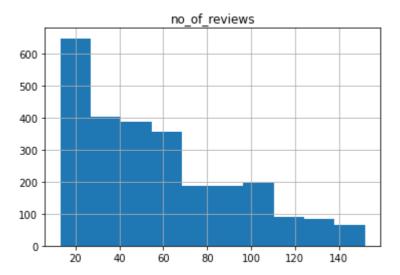


#### In [230]:

```
q_low = master_df2["no_of_reviews"].quantile(0.25)
q_hi = master_df2["no_of_reviews"].quantile(0.75)
```

#### In [231]:

```
master_df2 = master_df2[(master_df2["no_of_reviews"] < q_hi) & (master_df2["no_of_reviews"]
master_df2.hist(column="no_of_reviews")
plt.show()</pre>
```

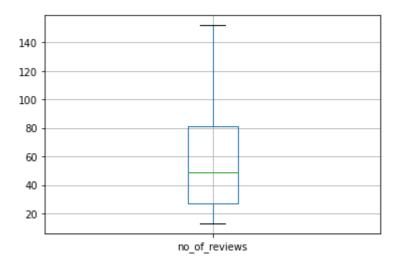


#### In [232]:

```
master_df2.boxplot(column='no_of_reviews')
```

#### Out[232]:

#### <AxesSubplot:>



#### In [233]:

```
# Average reviews
total_reviews = master_df2_temp['no_of_reviews'].sum()
total_rows_after_outliers = master_df2_temp.shape[0]
threshhold_for_selection_of_reviews = total_reviews/total_rows_after_outliers
threshhold_for_selection_of_reviews
```

#### Out[233]:

134.75090943901972

#### In [234]:

```
#Average rating
total_rating = master_df2_temp['rating'].sum()
total_rows_of_rating = len(master_df2_temp)
threshhold_for_selection_of_rating = total_rating/total_rows_of_rating
threshhold_for_selection_of_rating
```

#### Out[234]:

4.324736741336396

#### In [235]:

## In [236]:

## print(len(selected)) selected

64

## Out[236]:

	no_of_reviews	rating	address1
company			
Adeeb Hair Salon	139	3.8	Air Line Housing Scheme Airline Blvd Lahore Ai
Let's Makeup	143	4.3	MUST Acadamy 25. College Road Township Ahmed A
New Kashee Beauty Parlor	145	3.9	H72J+V9F E Block Block E Gulshan-e- Ravi Lahore
Arammish Gulberg Branch	135	4.3	Park 66-b Syed Maratib Ali Road F.C.C. Lahore
New Kashee Beauty Parlor	145	3.9	H72J+V9F E Block Block E Gulshan-e- Ravi Lahore
New Kashee Beauty Parlor	145	3.9	H72J+V9F E Block Block E Gulshan-e- Ravi Lahore
New Kashee Beauty Parlor	145	3.9	H72J+V9F E Block Block E Gulshan-e- Ravi Lahore
Depilex Beauty Clinic   Johar Town Lahore	137	3.6	House #42 Allah Hoo Chowk near Gourmet Bakers
New Kashee Beauty Parlor	145	3.9	H72J+V9F E Block Block E Gulshan-e- Ravi Lahore
New Kashee Beauty Parlor	145	3.9	H72J+V9F E Block Block E Gulshan-e- Ravi Lahore
Sun In Hair Salon	150	4.2	Central Commercial Market 20C Model Town Lah
New Kashee Beauty Parlor	145	3.9	H72J+V9F E Block Block E Gulshan-e- Ravi Lahore
Arammish Gulberg Branch	135	4.3	Park 66-b Syed Maratib Ali Road F.C.C. Lahore
New Kashee Beauty Parlor	145	3.9	H72J+V9F E Block Block E Gulshan-e- Ravi Lahore
New Kashee Beauty Parlor	145	3.9	H72J+V9F E Block Block E Gulshan-e- Ravi Lahore
Sun In Hair Salon	150	4.2	Central Commercial Market 20C Model Town Lah
Depilex Beauty Clinic   Johar Town Lahore	137	3.6	House #42 Allah Hoo Chowk near Gourmet Bakers
New Kashee Beauty Parlor	145	3.9	H72J+V9F E Block Block E Gulshan-e- Ravi Lahore
Arammish Gulberg Branch	135	4.3	Park 66-b Syed Maratib Ali Road F.C.C. Lahore
New Kashee Beauty Parlor	145	3.9	H72J+V9F E Block Block E Gulshan-e- Ravi Lahore
Let's Makeup	143	4.3	MUST Acadamy 25. College Road Township Ahmed A

	no_ot_reviews	rating	address1
company			
Let's Makeup	143	4.3	MUST Acadamy 25, College Road Township Ahmed A
Arammish Gulberg Branch	135	4.3	Park 66-b Syed Maratib Ali Road F.C.C. Lahore
New Kashee Beauty Parlor	145	3.9	H72J+V9F E Block Block E Gulshan-e- Ravi Lahore
Depilex Beauty Clinic   Johar Town Lahore	137	3.6	House #42 Allah Hoo Chowk near Gourmet Bakers
Sun In Hair Salon	150	4.2	Central Commercial Market 20C Model Town Lah
Bobby Hair Saloon	137	4.1	Ghalib Rd opposite pace opposite hafeez center
Arammish Gulberg Branch	135	4.3	Park 66-b Syed Maratib Ali Road F.C.C. Lahore
New Kashee Beauty Parlor	145	3.9	H72J+V9F E Block Block E Gulshan-e- Ravi Lahore
Let's Makeup	143	4.3	MUST Acadamy 25. College Road Township Ahmed A
Arammish Gulberg Branch	135	4.3	Park 66-b Syed Maratib Ali Road F.C.C. Lahore
Depilex Beauty Clinic   Johar Town Lahore	137	3.6	House #42 Allah Hoo Chowk near Gourmet Bakers
SMARTCUT - Faisal Town Lahore	151	4.3	52 Qazi Muhammad Isa Rd Block B Faisal Town La
New Kashee Beauty Parlor	145	3.9	H72J+V9F E Block Block E Gulshan-e- Ravi Lahore
Depilex Beauty Clinic   Johar Town Lahore	137	3.6	House #42 Allah Hoo Chowk near Gourmet Bakers
Let's Makeup	143	4.3	MUST Acadamy 25. College Road Township Ahmed A
New Kashee Beauty Parlor	145	3.9	H72J+V9F E Block Block E Gulshan-e- Ravi Lahore
Let's Makeup	143	4.3	MUST Acadamy 25. College Road Township Ahmed A
New Kashee Beauty Parlor	145	3.9	H72J+V9F E Block Block E Gulshan-e- Ravi Lahore
Depilex Beauty Clinic   Johar Town Lahore	137	3.6	House #42 Allah Hoo Chowk near Gourmet Bakers
New Kashee Beauty Parlor	145	3.9	H72J+V9F E Block Block E Gulshan-e- Ravi Lahore
Depilex Beauty Clinic   Johar Town Lahore	137	3.6	House #42 Allah Hoo Chowk near Gourmet Bakers
Let's Makeup	143	4.3	MUST Acadamy 25، College Road Township Ahmed A
Depilex Beauty Clinic   Johar Town Lahore	137	3.6	House #42 Allah Hoo Chowk near Gourmet Bakers
New Kashee Beauty Parlor	145	3.9	H72J+V9F E Block Block E Gulshan-e- Ravi Lahore

no\_of\_reviews rating

address1

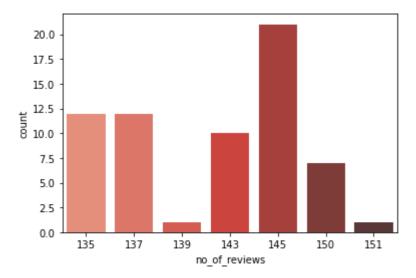
	no_of_reviews	rating	address1
company			
Sun In Hair Salon	150	4.2	Central Commercial Market 20C₊ Model Town₁ Lah
Arammish Gulberg Branch	135	4.3	Park 66-b Syed Maratib Ali Road F.C.C. Lahore
Arammish Gulberg Branch	135	4.3	Park 66-b Syed Maratib Ali Road F.C.C. Lahore
Deedar Beauty Salon	135	4.1	Kasbah Towe College Rd Sector C1 Township T
Bobby Hair Saloon	137	4.1	Ghalib Rd opposite pace opposite hafeez center
Sun In Hair Salon	150	4.2	Central Commercial Market 20C Model Town Lah
Let's Makeup	143	4.3	MUST Acadamy 25. College Road Township Ahmed A
Depilex Beauty Clinic   Johar Town Lahore	137	3.6	House #42 Allah Hoo Chowk near Gourmet Bakers
New Kashee Beauty Parlor	145	3.9	H72J+V9F E Block Block E Gulshan-e- Ravi Lahore
Deedar Beauty Salon	135	4.1	Kasbah Towe، College Rd، Sector C1 Township، T
Arammish Gulberg Branch	135	4.3	Park 66-b Syed Maratib Ali Road F.C.C. Lahore
Arammish Spa DHA	143	4.0	FC67+CFH St 5 Sector C Phase 5 D.H.A Lahore Pu
New Kashee Beauty Parlor	145	3.9	H72J+V9F E Block Block E Gulshan-e- Ravi Lahore
Sun In Hair Salon	150	4.2	Central Commercial Market 20C Model Town Lah
Arammish Gulberg Branch	135	4.3	Park 66-b Syed Maratib Ali Road F.C.C. Lahore
Let's Makeup	143	4.3	MUST Acadamy 25- College Road Township Ahmed A
Depilex Beauty Clinic   Johar Town Lahore	137	3.6	House #42 Allah Hoo Chowk near Gourmet Bakers
New Kashee Beauty Parlor	145	3.9	H72J+V9F E Block Block E Gulshan-e- Ravi Lahore
Sun In Hair Salon	150	4.2	Central Commercial Market 20C Model Town Lah

#### In [248]:

```
sns.set_palette("husl",3)
sns.countplot(x = "no_of_reviews", data = selected , palette = "Reds_d")
```

#### Out[248]:

<AxesSubplot:xlabel='no\_of\_reviews', ylabel='count'>



#### In [249]:

```
sns.set_palette("husl",3)
sns.countplot(x = "rating", data = selected , palette = "Reds_d")
```

#### Out[249]:

<AxesSubplot:xlabel='rating', ylabel='count'>

