Data Exploration

2024-09-12

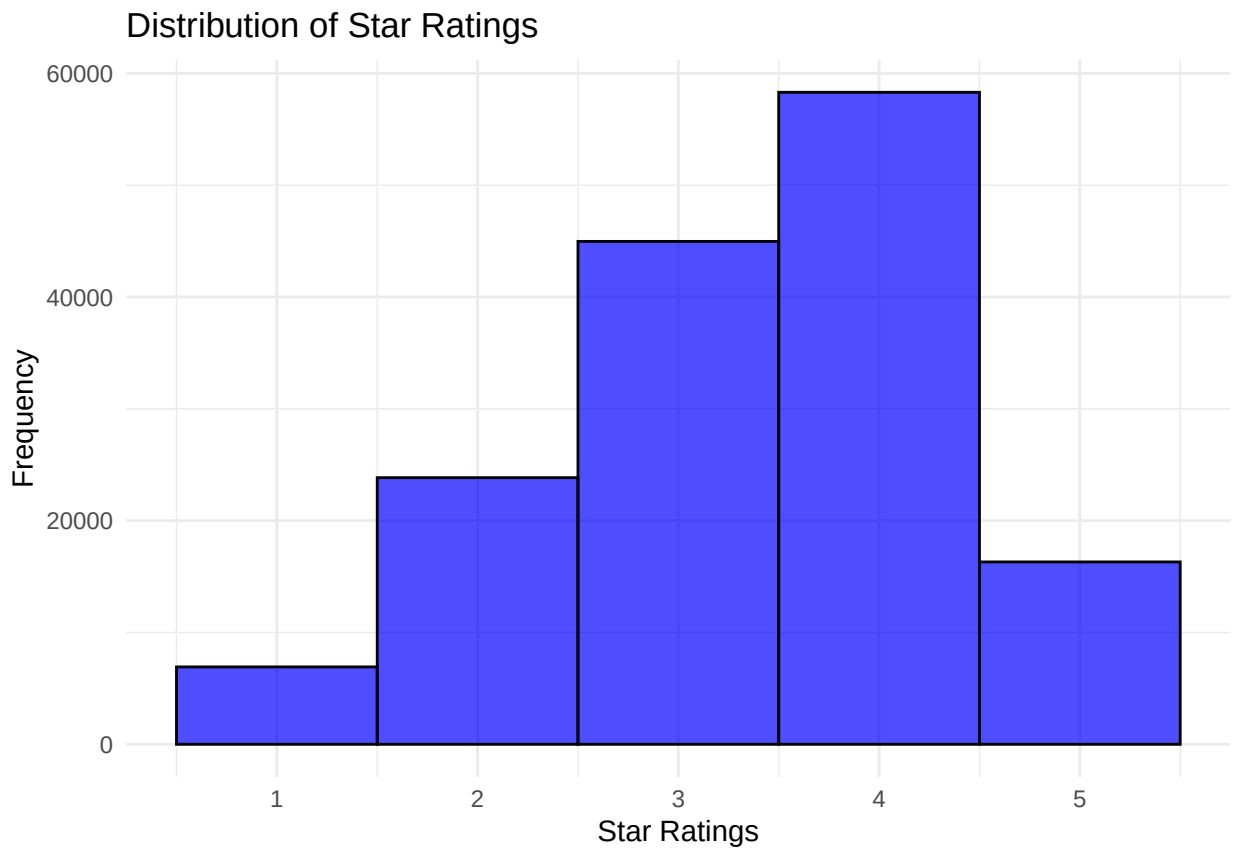
Summary Statistics

Variable: Stars

This section will explain the key statistics for the stars column as well as depict a plot of this column for a better understanding of the data. As our research will focus on the impact on these ratings, it is important to have a good understanding of this variable.

Table 1: Summary of Star Ratings

Statistic	Value
Mean Star Rating	3.596724
Rounded Mean Star Rating	3.600000
Median Star Rating	3.500000
Maximum Star Rating	5.000000
Minimum Star Rating	1.000000



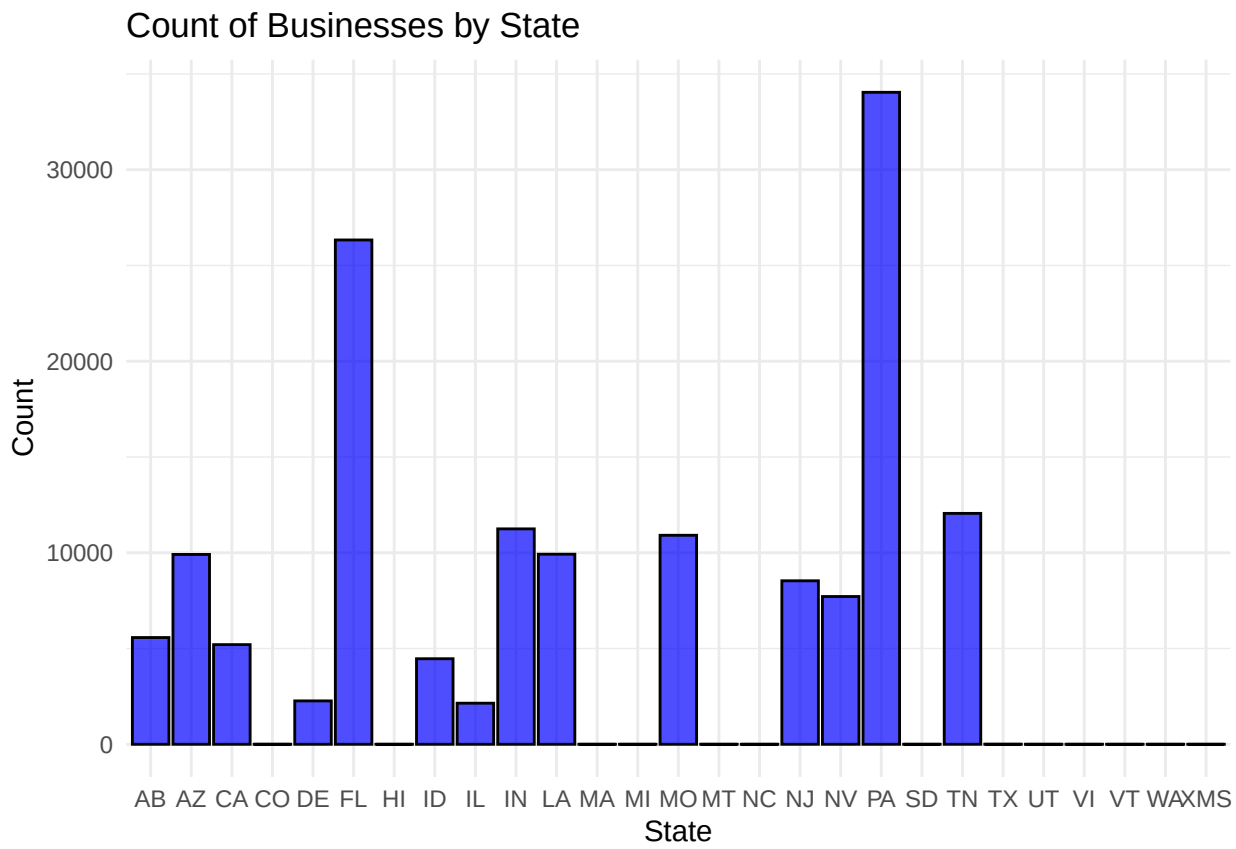
As depicted in the graph the most common rating obtained by the business on Yelp is of 4 stars. On the other hand, 1 star ratings are the least common.

Useful Information

In the following section, when referring to the "Count" of a variable, it refers to the amount of times that this specific variable is present throughout the businesses in the Yelp data set.

Variable: States

This section will depict the location distribution of business among the different states in the USA as the Yelp Reviews are from these location.



This figure allows us to better understand the geographical distribution of the businesses, which might of interest when assesing the reviews and ratings.

Variable: Categories

Only the top 20 categories are depicted in the following table for illustrative purposes.

Table 2: Count of Top 20 Categories

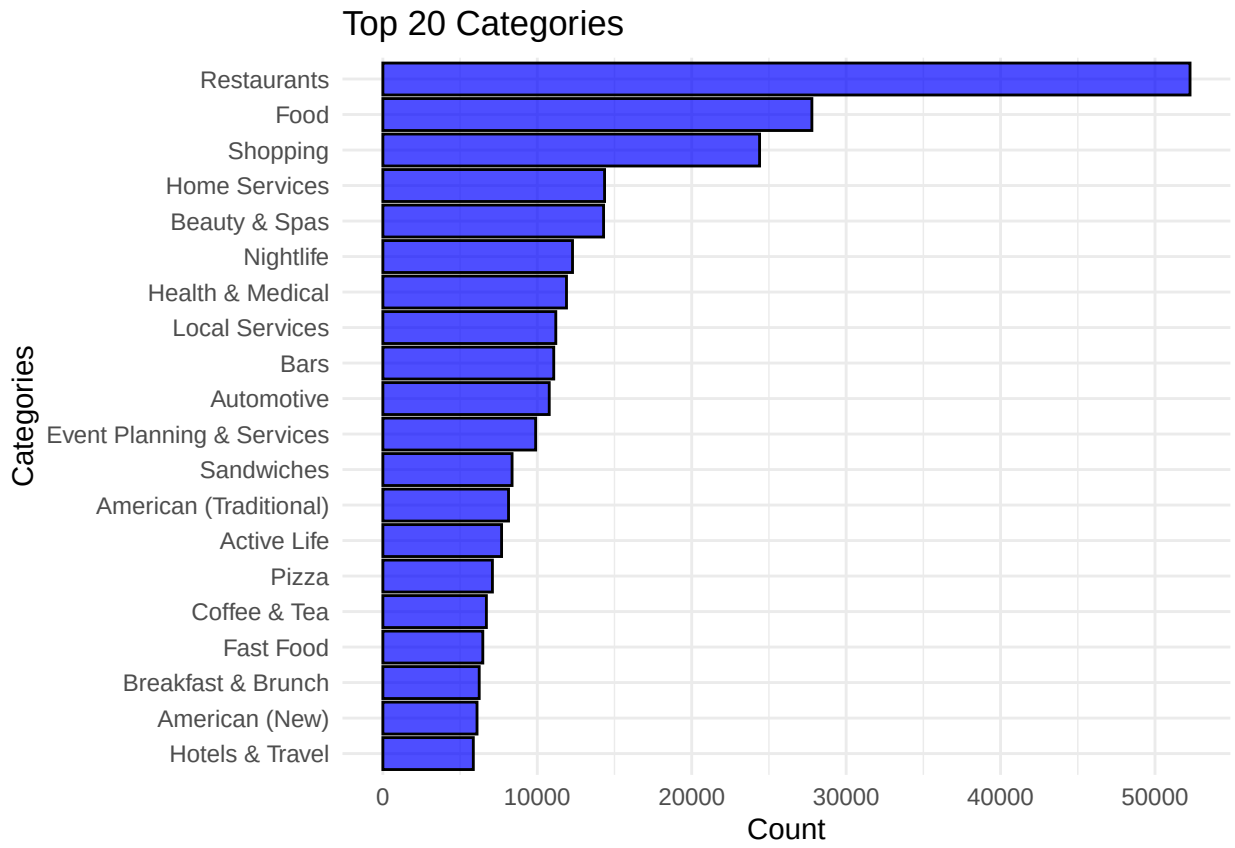
Category	Count
Restaurants	52268
Food	27781
Shopping	24395
Home Services	14356
Beauty & Spas	14292
Nightlife	12281
Health & Medical	11890
Local Services	11198
Bars	11065
Automotive	10773
Event Planning & Services	9895
Sandwiches	8366
American (Traditional)	8139
Active Life	7687
Pizza	7093

Category	Count
Coffee & Tea	6703
Fast Food	6472
Breakfast & Brunch	6239
American (New)	6097
Hotels & Travel	5857

The code to plot all business categories can be found below hiding.

This figure represents the **Top 20 categories** of businesses that appear more on Yelp.

To obtain a better illustrative depiction of the categories only the 20 top categories are depicted on this plot.



Variable: Attributes

This section explores in more detail the variable "Attributes" as it is one of the key elements for this research. As it is a key element in the research, the 30 most used attributes are depicted below.

The variable attribute includes elements categorized as "True" or "False". If an element is indicated as "True" this means that said business has that attribute present while if it is indicated as "False" it indicates that that attribute is not present in that specific business.

As we want to asses the impact of that the attributes have on ratings, firstly, the top 30 attributes present in most businesses will be depicted:

Table 3: Count of Top 30 Attributes

Attributes	Count
'BusinessAcceptsCreditCards': 'True'	68183
'RestaurantsTakeOut': 'True'	44368
'BikeParking': 'True'	42301
'GoodForKids': 'True'	35100
'RestaurantsGoodForGroups': 'True'	33949
'HasTV': 'True'	29623
'RestaurantsDelivery': 'True'	24083
{'BusinessAcceptsCreditCards': 'True'	23349
'WheelchairAccessible': 'True'	20403
'Caters': 'True'	19837
'OutdoorSeating': 'True'	19590
'RestaurantsReservations': 'True'	13263
'BusinessAcceptsCreditCards': 'True'}	12750
'RestaurantsTableService': 'True'	11419
{'BusinessAcceptsCreditCards': 'True'}	9385
'HappyHour': 'True'	8271
'BikeParking': 'True'}	6894
'RestaurantsDelivery': 'True'}	6343
{'BikeParking': 'True'	5835
{'GoodForKids': 'True'	5667
'ByAppointmentOnly': 'True'	5421
'DogsAllowed': 'True'	4784
{'ByAppointmentOnly': 'True'	4756
'RestaurantsTakeOut': 'True'}	4283
{'RestaurantsTakeOut': 'True'	4233
'ByAppointmentOnly': 'True'}	3834
'WheelchairAccessible': 'True'}	3822
'DriveThru': 'True'	3430
{'RestaurantsGoodForGroups': 'True'	2997
'HasTV': 'True'}	2866

As the lack of a an attribute can also have an impact on the rating of a business, the top 30 attributes less present on businesses will also be depicted below:

Table 4: Count of Top 30 Attributes

Attributes	Count
'RestaurantsReservations': 'False'	26670
'OutdoorSeating': 'False'	21581
'RestaurantsDelivery': 'False'	17082
'Caters': 'False'	15701
'ByAppointmentOnly': 'False'	14972
'BikeParking': 'False'	13482
'BusinessAcceptsBitcoin': 'False'	11824
'DogsAllowed': 'False'	10495
'HasTV': 'False'	9777
'GoodForKids': 'False'	7112
'RestaurantsTableService': 'False'	6513
'ByAppointmentOnly': 'False'}	5641

Attributes	Count
{‘ByAppointmentOnly’: ‘False’	5366
‘RestaurantsGoodForGroups’: ‘False’	5231
‘HappyHour’: ‘False’	4903
‘CoatCheck’: ‘False’	4753
‘RestaurantsTakeOut’: ‘False’	3633
‘GoodForDancing’: ‘False’	3466
‘BusinessAcceptsBitcoin’: ‘False’}	3235
‘BYOB’: ‘False’	3048
‘BusinessAcceptsCreditCards’: ‘False’	2683
‘BikeParking’: ‘False’}	2418
‘DriveThru’: ‘False’	2297
{‘RestaurantsReservations’: ‘False’	2257
‘WheelchairAccessible’: ‘False’	2237
‘Corkage’: ‘False’	2156
{‘BusinessAcceptsBitcoin’: ‘False’	1891
{‘OutdoorSeating’: ‘False’	1837
‘RestaurantsDelivery’: ‘False’}	1674
{‘BikeParking’: ‘False’	1612

Variable: Review Count

The purpose of this section is to explore the key statistics of the review count column as well as depicting plots of this column for a better understanding of the data. As our research will distinguish between businesses that have a low number of reviews versus businesses that have a high number of reviews, it is crucial to have a good understanding of this variable.

Table 5: Summary of Review Count

Statistic	Value
Mean Review Count	44.86656
Rounded Mean Review Count	44.87000
Median Review Count	15.00000
Maximum Review Count	7568.00000
Minimum Review Count	5.00000

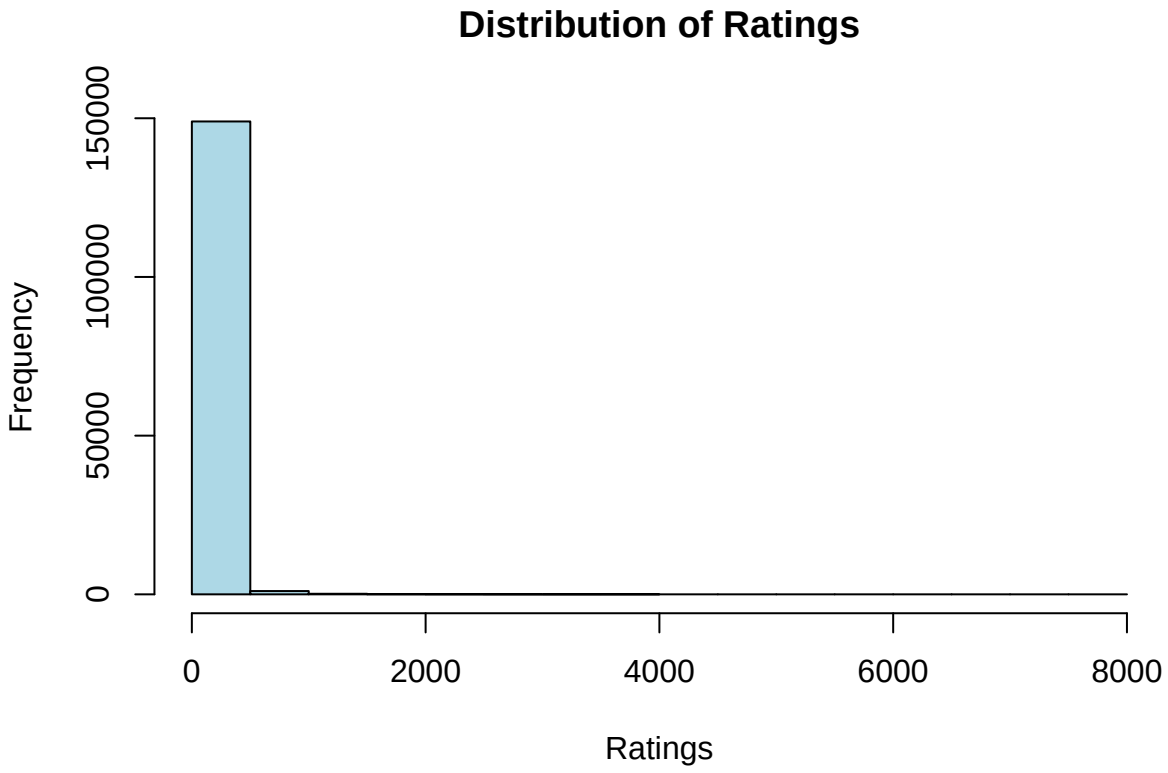
The variable ‘review count’ can be divided into 4 quartiles with the following ranges:

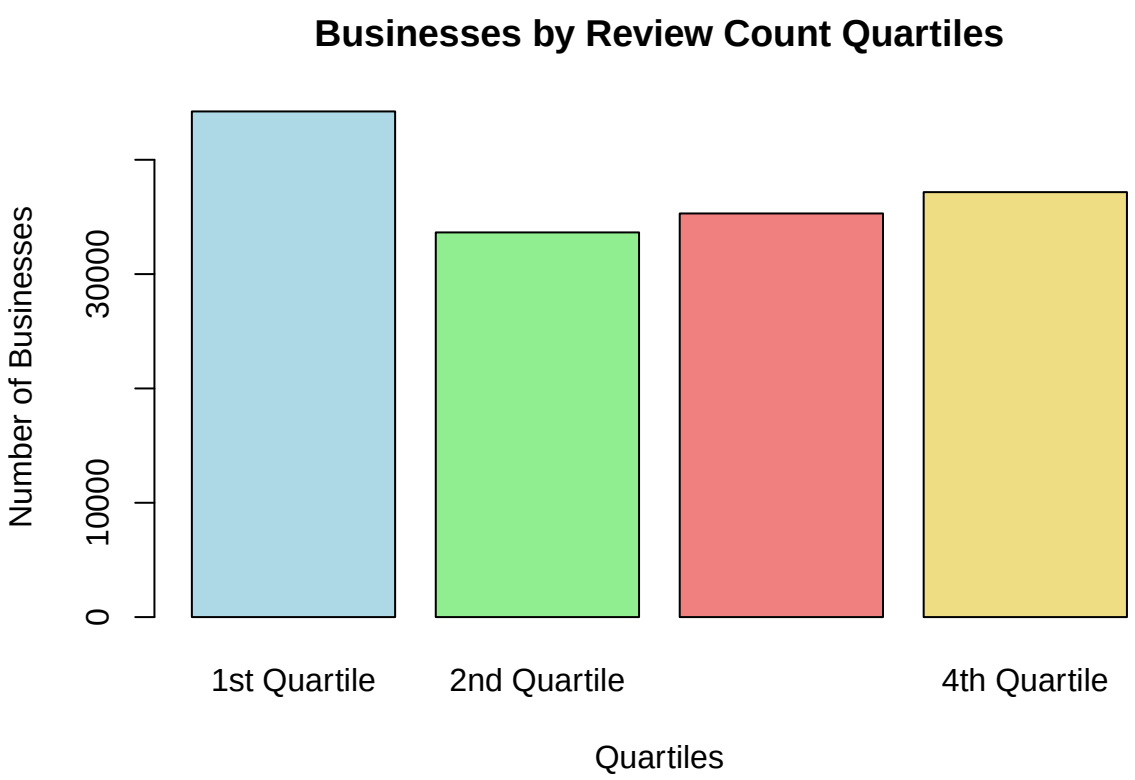
Review Count Quartile Ranges

Quartile	Lower Bound	Upper Bound
1st Quartile	5	8
2nd Quartile	8	15
3rd Quartile	15	37
4th Quartile	37	7,568

Each quartile represents a specific range of review counts, from the lowest 25% to the highest 25%. This segmentation helps us understand the distribution of review activity among businesses. The table depicted above implies that businesses that fall into the first quartile have between 5 and 8 reviews. Businesses that fall into the second quartile have between 8 and 15 reviews, businesses in the third quartile have between

15 and 37 reviews and businesses in the fourth quartile have between 37 and 7568 reviews. This is valuable information because it allows to distinguish between businesses that have a low review count versus businesses with a high review count.





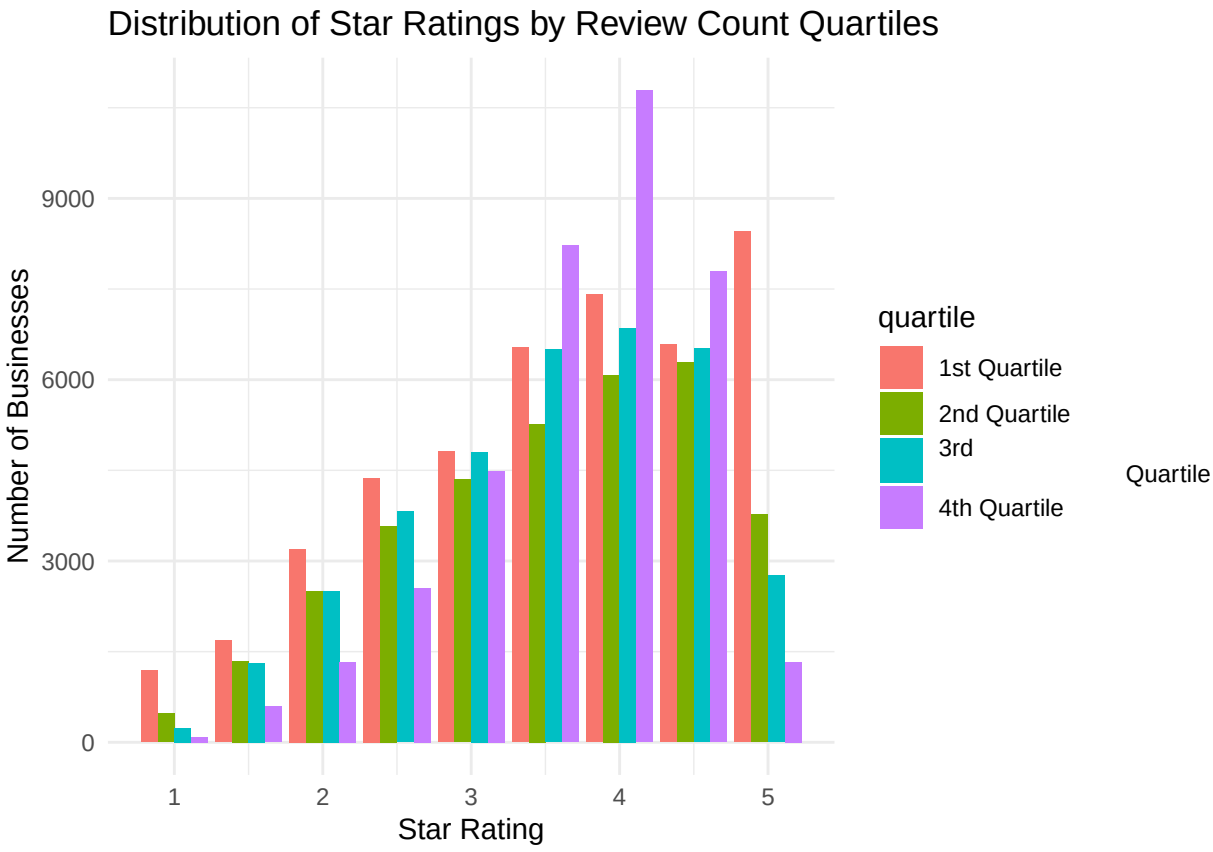
The graph above depicts how many businesses fall into each "category" of review count that is elaborated above:

- 1st Quartile: 44228 businesses
- 2nd Quartile: 33646 businesses
- 3rd Quartile: 35303 businesses
- 4th Quartile: 37169 businesses

Review Count and Stars

In this analysis, we created quartiles based on the review count variable to categorize businesses into four distinct groups representing different levels of review activity. We then examined the distribution of the 'stars' variable across these quartiles. This involved plotting the number of businesses within each quartile against their corresponding star ratings.

The quartiles allowed us to segment businesses into groups ranging from the lowest to the highest 25% of review counts. By plotting star ratings against these quartiles, we were able to visually assess how star ratings vary across different levels of review activity. The resulting plots illustrate any patterns or trends in star ratings relative to review count quartiles. This approach provides insights into the relationship between the volume of reviews and the star ratings assigned to businesses. The graph and table below depict our findings:



Summary of Businesses by Quartile and Star Rating

quartile	stars	n
1st Quartile	1.0	1187
1st Quartile	1.5	1682
1st Quartile	2.0	3194
1st Quartile	2.5	4375
1st Quartile	3.0	4814
1st Quartile	3.5	6534
1st Quartile	4.0	7407
1st Quartile	4.5	6584
1st Quartile	5.0	8451
2nd Quartile	1.0	478
2nd Quartile	1.5	1348
2nd Quartile	2.0	2502
2nd Quartile	2.5	3576
2nd Quartile	3.0	4345
2nd Quartile	3.5	5263
2nd Quartile	4.0	6076
2nd Quartile	4.5	6288
2nd Quartile	5.0	3770

3rd Quartile	1.0	236
3rd Quartile	1.5	1309
3rd Quartile	2.0	2507
3rd Quartile	2.5	3816
3rd Quartile	3.0	4805
3rd Quartile	3.5	6499
3rd Quartile	4.0	6854
3rd Quartile	4.5	6517
3rd Quartile	5.0	2760
4th Quartile	1.0	85
4th Quartile	1.5	593
4th Quartile	2.0	1324
4th Quartile	2.5	2549
4th Quartile	3.0	4489
4th Quartile	3.5	8223
4th Quartile	4.0	10788
4th Quartile	4.5	7792
4th Quartile	5.0	1326

“‘