Data Exploration

2024-09-12

Loading the data - will be changed to the data from the drive

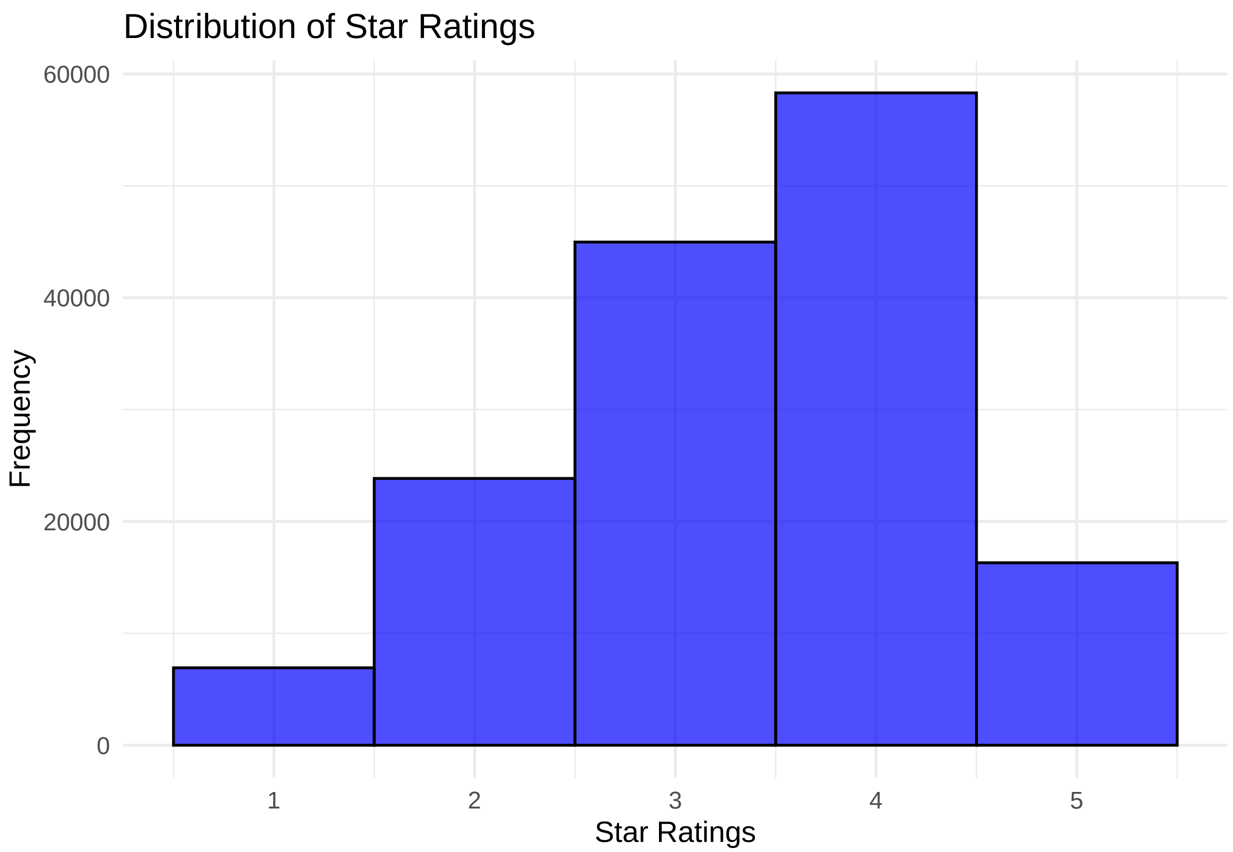
Summary Statistics

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



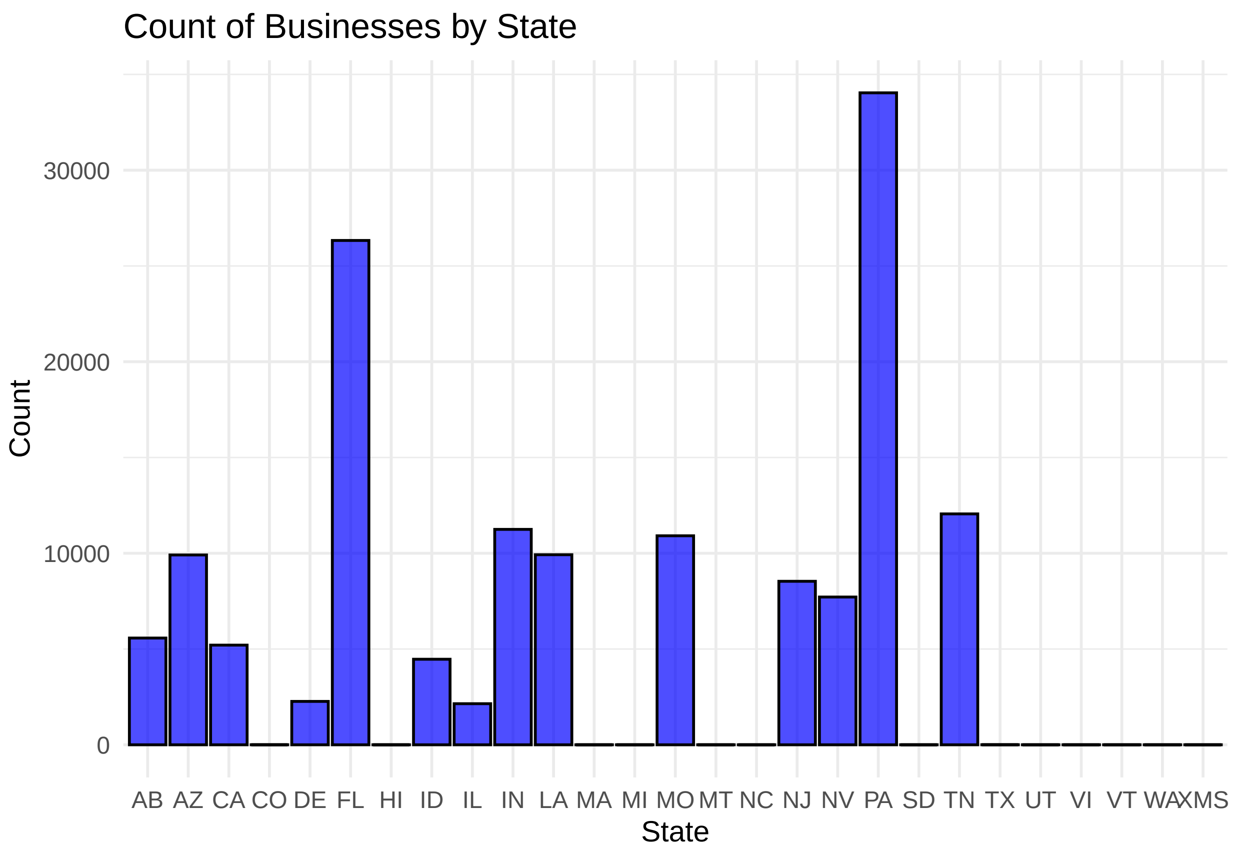
As depicted in the graph the most common rating obtained by the business on Yelp is of 4 stars. On the other hand, 1 star ratings are the least common.

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.

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.

Categories

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

Table 2: Count of Top 20 Categories

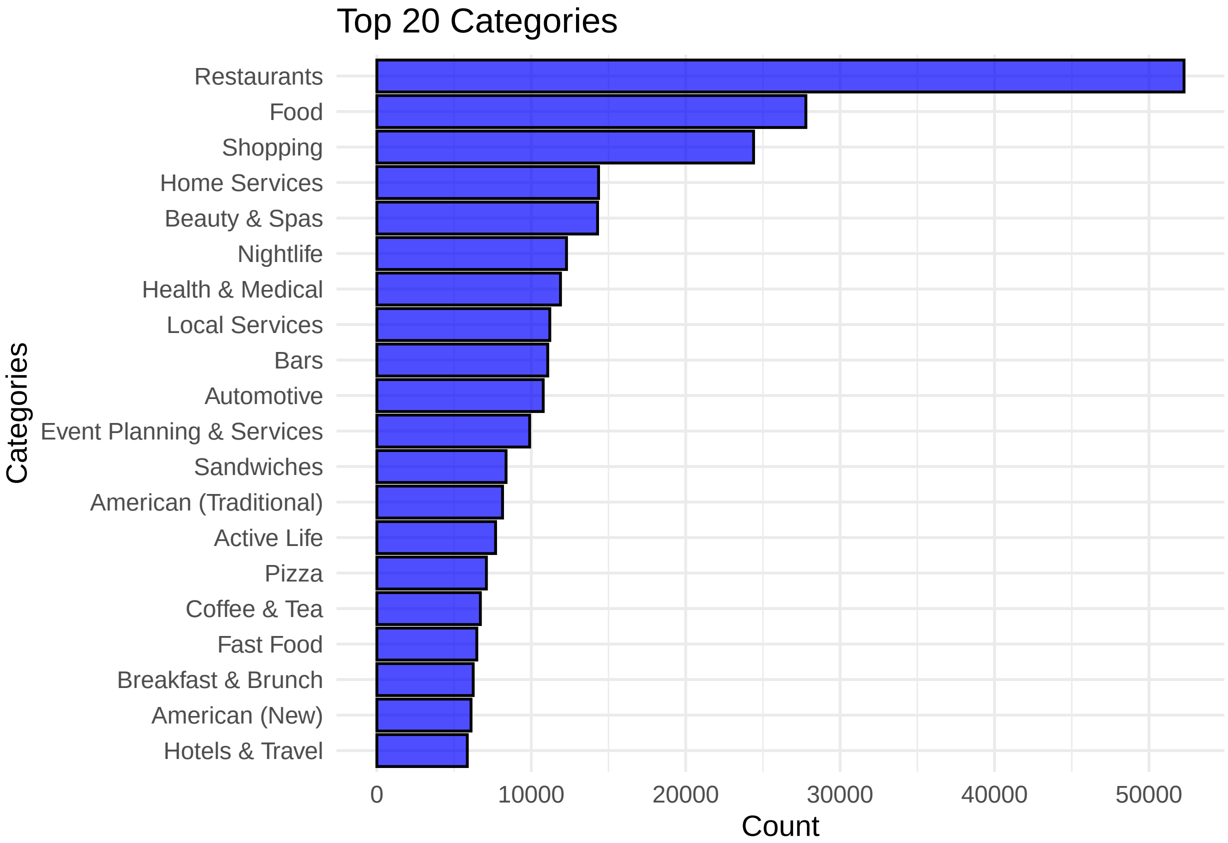
Category	Count
Restaurants	52268
Food	27781
Shopping	24395
Home Services	14356
Beauty & Spas	14292
Nightlife	12281
Health & Medical	11890
Local Services	11198
Bars	11065
Automotive	10773
Event Planning & Services	9895
Sandwiches	8366
American (Traditional)	8139
Active Life	7687
Pizza	7093

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.



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

“‘