To: Airbnb Data Analytics Team
From: Thejaswini Paripally, Valentine Limauge

Dataset Description:

This dataset gives us a detailed look at what was happening on Airbnb in New York City during 2019. It encapsulates a wide range of data points that helps us understand things like who was renting out their places, where these places were located, and data we can use to make predictions and draw useful conclusions.

Key Areas of Exploration:

1. **Hosts and Geographic Insights:** Analyzing the dataset allows us to gain a deeper understanding of various hosts and the geographical distribution of their listings.

2. **Predictive Analysis:** This dataset provides valuable information for making predictions, including location trends, pricing factors, reviews, and more.

3. **Host Activity:** We can determine which hosts are the most active on the platform and explore the reasons behind their popularity.

4. **Geographical Differences:** By examining the data, we can identify variations in traffic and activity across different areas in NYC and investigate the underlying factors driving these distinctions.

Descriptive Analysis:

```
> summary(airbnb)
 neighbourhood_group neighbourhood        latitude       longitude       room_type
 Length:48895        Length:48895       Min.   :40.50   Min.   :-74.24   Length:48895
 Class :character    Class :character   1st Qu.:40.69   1st Qu.:-73.98   Class :character
 Mode  :character    Mode  :character   Median :40.72   Median :-73.96   Mode  :character
                                        Mean   :40.73   Mean   :-73.95
                                        3rd Qu.:40.76   3rd Qu.:-73.94
                                        Max.   :40.91   Max.   :-73.71

     price         minimum_nights    number_of_reviews last_review        reviews_per_month
 Min.   :    0.0   Min.   :   1.00   Min.   :  0.00    Length:48895       Min.   : 0.010
 1st Qu.:   69.0   1st Qu.:   1.00   1st Qu.:  1.00    Class :character   1st Qu.: 0.190
 Median :  106.0   Median :   3.00   Median :  5.00    Mode  :character   Median : 0.720
 Mean   :  152.7   Mean   :   7.03   Mean   : 23.27                       Mean   : 1.373
 3rd Qu.:  175.0   3rd Qu.:   5.00   3rd Qu.: 24.00                       3rd Qu.: 2.020
 Max.   :10000.0   Max.   :1250.00   Max.   :629.00                       Max.   :58.500
                                                                         NA's   :10052

 calculated_host_listings_count availability_365
 Min.   :  1.000                Min.   :  0.0
 1st Qu.:  1.000                1st Qu.:  0.0
 Median :  1.000                Median : 45.0
 Mean   :  7.144                Mean   :112.8
 3rd Qu.:  2.000                3rd Qu.:227.0
 Max.   :327.000                Max.   :365.0
```

Measures of location and central tendency:

When looking at the different tendencies from our data we can see that the average price and minimum nights is $152.72 and 7 nights. The median gives us the middle value that appears in the data set. In this case we can see that the middle values would be $106 a night and a minimum amount of nights per stay of 3. Lastly, we have the mode that gives us the value that appears the most in our data. In our case we can see that the minimum nights that are the most common is 1 at a price of $100.

	Minimum nights	Price
Mean	7.03	$152.72
Median	3	$106
Mode	1	$100

```
> #Mean
> means <- data.frame(Mean = c(mean(airbnb$price), mean(airbnb$minimum_nights)))
> row.names(means) <- c("Price", "Minimum nights")
> means
                    Mean
Price          152.720687
Minimum nights   7.029962
> #Median
> medians <- data.frame(Median = c(median(airbnb$price), median(airbnb$minimum_nights)))
> row.names(medians) <- c("Price", "Minimum nights")
> medians
               Median
Price             106
Minimum nights      3
> mod
               Mode
Price           100
Minimum nights    1
```

Measures of dispersion:

Looking at our data we can see that our standard deviation for the minimum nights is 20.51 and for the price it is $240.15. This tells us that there is a wide spread between the average and some of the data in our data set. The variance tells us a very similar story in the case of our data. With both variance being very high, we can say that our data is very spread out.

	Minimum nights	Price
Standard deviation	20.51	$240.15
Variance	420.68	$57674.02

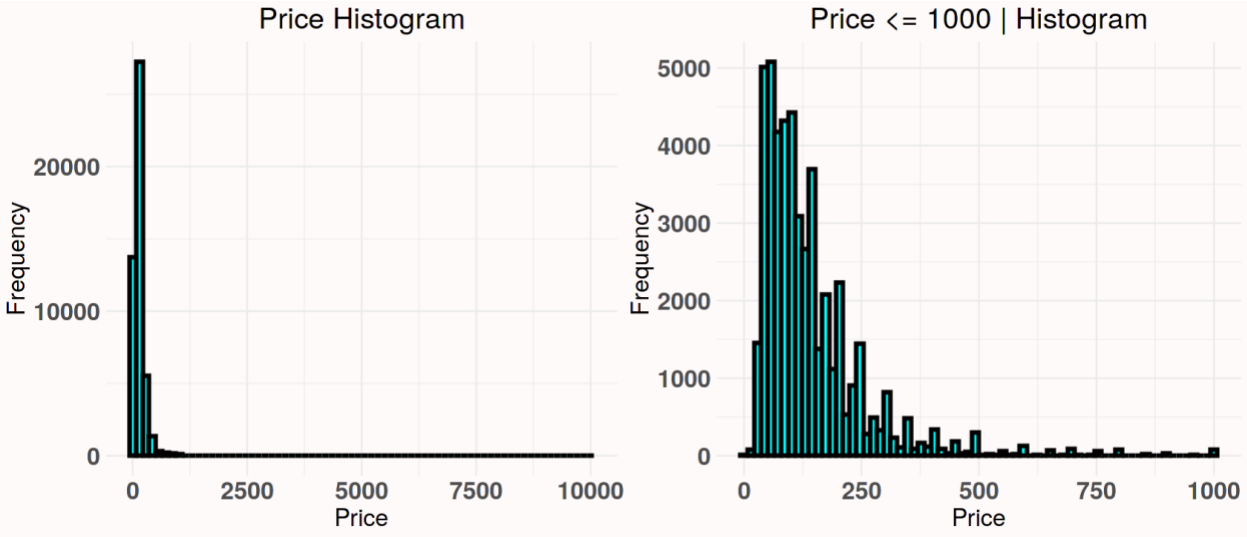
```
> #Variance
> v <- data.frame("Variance" = c(var(airbnb$price), var(airbnb$minimum_nights)))
> row.names(v) <- c("Price", "Minimum nights")
> v
                Variance
Price          57674.0252
Minimum nights   420.6826
> #Standard deviation
> std <- data.frame("Standard deviation" = c(sqrt(var(airbnb$price)), sqrt(var(airbnb$minimum_nights))))
> row.names(std) <- c("Price", "Minimum nights")
> std
               Standard.deviation
Price                    240.15417
Minimum nights            20.51055
```

Data distribution charts:

1. **Histogram:**

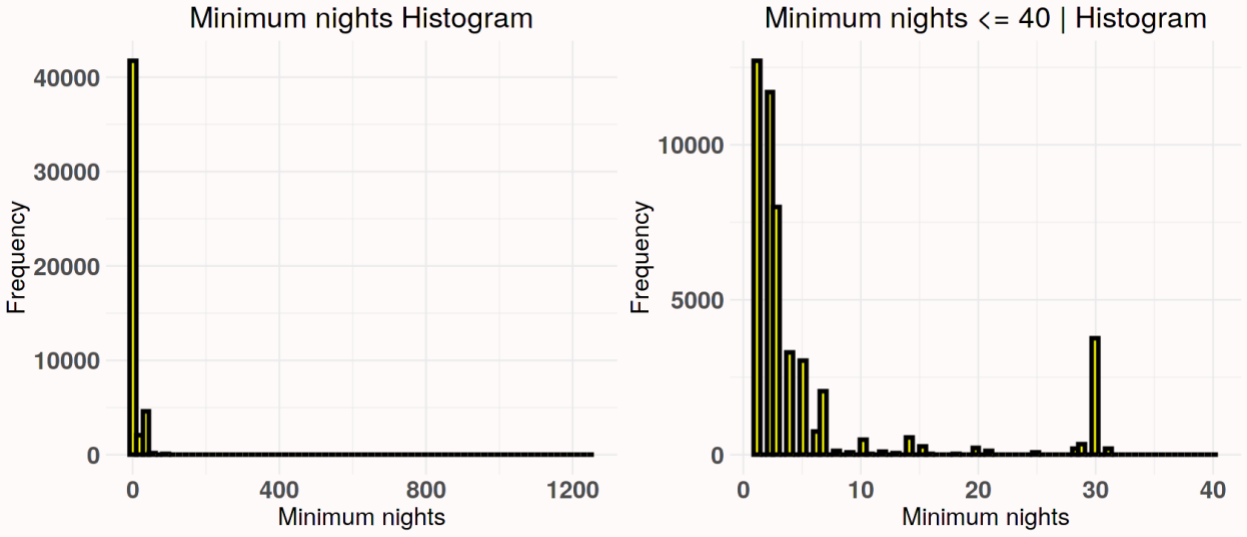
From the histogram, we can see that the majority of prices cluster below the $200 mark. Specifically, the table indicates that 98.15% of reservations fall within the price range of $0 to $58,823.

```
> freq_price
                      Frequency      Percent
      0  –  58,823       47994 98.157275795
 58,823  –  117,647        691  1.413232437
117,647  –  176,470         89  0.182022702
176,470  –  235,294         43  0.087943553
235,294  –  294,117         21  0.042949177
294,117  –  352,941         16  0.032723182
352,941  –  411,764         10  0.020451989
411,764  –  470,588          5  0.010225994
470,588  –  529,411          8  0.016361591
529,411  –  588,235          0  0.000000000
588,235  –  647,058          3  0.006135597
647,058  –  705,882          4  0.008180796
705,882  –  764,705          2  0.004090398
764,705  –  823,529          2  0.004090398
823,529  –  882,352          1  0.002045199
882,352  –  941,176          0  0.000000000
941,176  –  10 000           6  0.012271193
```



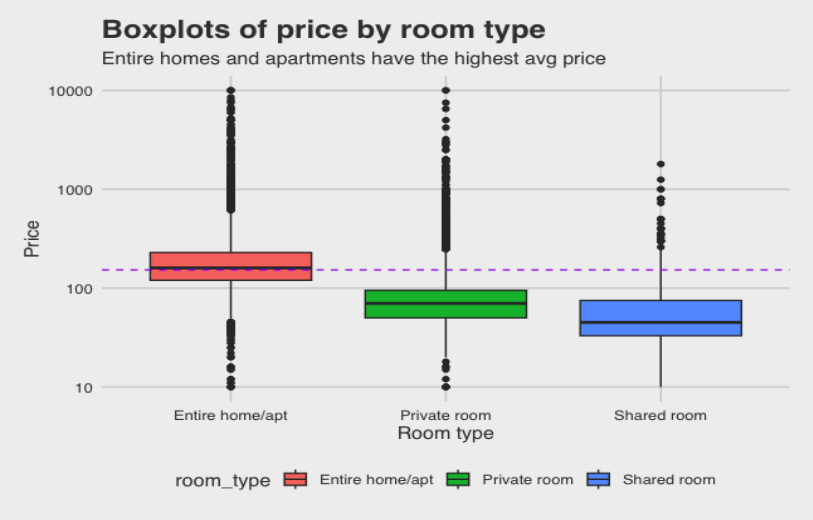
Similarly using the "minimum_nights" variable, we observe a lot of reservations with minimum nights below 10, along with a minor peak at 30. From the frequency table, we can see that 99.34% of Airbnb reservations fall within the range of 1 to 73 nights.

```
> freq_nights
                    Frequency      Percent
      0  -  58,823      48577 99.349626751
 58,823  -  117,647       184  0.376316597
117,647  -  176,470        67  0.137028326
176,470  -  235,294        10  0.020451989
235,294  -  294,117        44  0.089988751
294,117  -  352,941         2  0.004090398
352,941  -  411,764         6  0.012271193
411,764  -  470,588         0  0.000000000
470,588  -  529,411         0  0.000000000
529,411  -  588,235         0  0.000000000
588,235  -  647,058         0  0.000000000
647,058  -  705,882         0  0.000000000
705,882  -  764,705         0  0.000000000
764,705  -  823,529         4  0.008180796
823,529  -  882,352         0  0.000000000
882,352  -  941,176         0  0.000000000
941,176  -  10 000          1  0.002045199
```

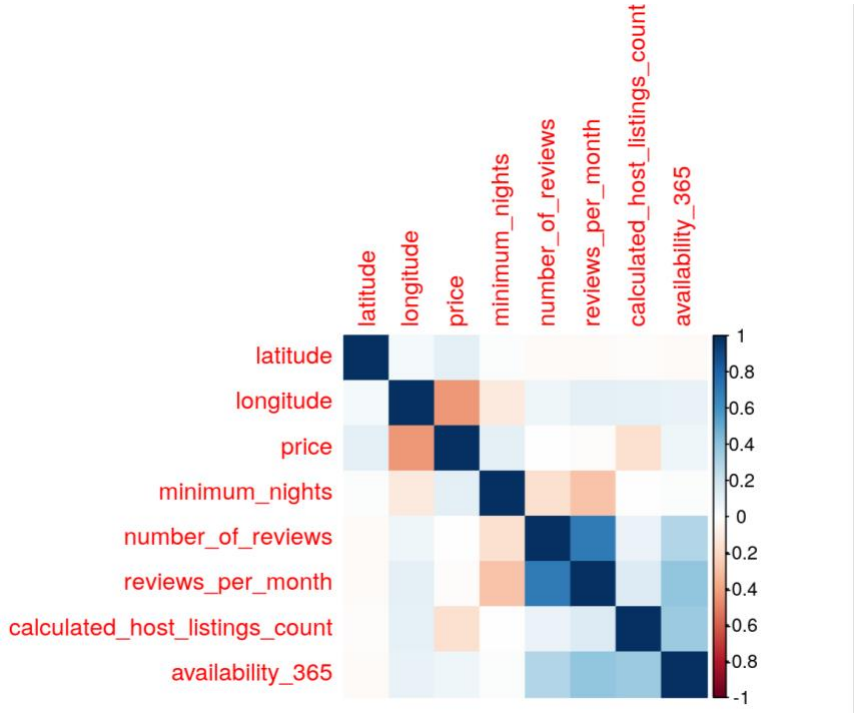


2. **Box plot:**

When investigating price by room type as anticipated, the "entire home or apartment" category boasts the highest average price, while shared rooms, as expected, offer more budget-friendly options than private rooms.



Correlation plot:



Analysis of outliers or high-leverage data:

In our analysis of the Airbnb dataset, we identified outliers in the "price" variable using two methods: z-scores and the 1.5 times the Interquartile Range (IQR) method. The identified outliers are data points that significantly deviate from the typical price range. These outliers can potentially influence statistical analyses and predictions. Retaining outliers can provide insights into unique cases, while removing them may lead to a more robust and generalizable analysis.

```
> print(outliers)
   [1]   497   763   947  1106  1415  1481  1554  1863  1900  2019  2041  2156  2216  2237  2356
  [16]  2387  2521  2524  2699  2773  2897  3132  3307  3334  3346  3421  3538  3576  3595  3599
  [31]  3623  3637  3638  3671  3685  3690  3696  3701  3703  3721  3722  3724  3728  3731  3732
  [46]  3733  3756  3759  3762  3775  3783  3785  3786  3789  3794  3805  3813  3814  3818  4128
  [61]  4346  4377  4378  4483  4731  5433  5500  5757  5802  5840  5862  5943  5957  6278  6334
  [76]  6398  6502  6512  6531  6621  6716  6988  7089  7097  7191  7478  7486  7514  7542  7847
  [91]  7983  8523  8531  8728  8806  8916  9036  9093  9152  9604  9884 10334 10342 10432 10521
 [106] 11022 11240 11265 11340 11370 11395 11561 12330 12343 12741 12801 12847 12879 13725 13789
 [121] 13931 14167 14381 14386 14460 14574 14575 14630 15092 15108 15155 15391 15471 15561 15577
 [136] 15678 15834 15838 15891 15914 16041 16123 16546 16876 17171 17653 17666 17693 17812 17978
 [151] 18357 18520 18633 19061 19237 19271 19305 19475 19685 19803 19804 19877 19906 20012 20095
 [166] 20220 20513 20552 20820 20890 21002 21177 21221 21370 21822 21956 22354 22374 22408 22473
 [181] 22636 22977 22993 23397 23412 23487 23546 23585 23625 23695 23866 24220 24287 24478 24541
 [196] 24994 25065 25259 25403 25517 25826 25902 25948 26444 26618 26740 26782 27259 27513 27591
 [211] 27786 27923 28728 28859 28946 28947 28948 28953 29065 29238 29239 29662 29663 29664 29665
 [226] 29666 29667 29674 29682 29684 30081 30258 30260 30261 30269 30825 30858 30917 31106 31339
 [241] 31508 31533 31867 31955 32004 32042 32373 32440 32545 32572 32797 33138 33431 33572 33756
 [256] 34118 34119 34245 34311 34614 34845 34852 35696 35921 36713 36834 36843 37084 37180 37195
 [271] 37254 37451 37571 37776 37865 38001 38123 38221 38224 38317 38326 38359 38457 38499 38592
 [286] 38948 39140 39159 39244 39737 39743 39767 39815 39847 39862 39913 40033 40087 40153 40380
 [301] 40413 40434 40587 40598 40639 40735 40765 41182 41215 41216 41217 41225 41284 41401 41581
 [316] 41584 41585 41586 41648 41716 41771 41894 42173 42524 42543 42681 42737 42913 42916 43010
 [331] 43071 43131 43202 43267 43564 43671 44035 44163 44225 44430 44477 45032 45115 45171 45186
 [346] 45415 45573 45611 45667 45711 45867 45868 45870 45887 45892 45902 45968 46141 46292 46299
 [361] 46338 46378 46389 46393 46407 46439 46534 46597 46615 46711 46851 46896 46966 47042 47351
 [376] 47392 47549 47671 47870 48044 48051 48081 48283 48302 48305 48306 48524 48536    62    86
 [391]   104   115   122   159   182   234   243   264   300   305   325   328   329   346   366
 [406]   396   419   420   431   461   468   474   495   501   502   517   627   634   638   654
 [421]   662   663   680   685   690   692   712   727   743   771   781   800   814   894   906
 [436]   917   921   954   984  1020  1053  1059  1112  1130  1131  1142  1160  1162  1196  1202
 [451]  1203  1204  1206  1226  1244  1274  1282  1316  1363  1401  1427  1434  1436  1452  1460
 [466]  1495  1528  1530  1561  1568  1577  1589  1595  1622  1631  1651  1659  1661  1680  1685
 [481]  1712  1760  1788  1809  1824  1881  1883  1895  1896  1897  1911  2005  2010  2053  2069
 [496]  2096  2101  2143  2147  2149  2164  2175  2182  2200  2218  2231  2238  2249  2297  2300
 [511]  2357  2370  2450  2462  2480  2484  2485  2507  2508  2512  2538  2543  2568  2572  2619
 [526]  2639  2642  2727  2746  2752  2760  2812  2825  2845  2855  2858  2883  2914  2962  2974
 [541]  2998  3005  3009  3023  3050  3056  3059  3079  3085  3091  3121  3123  3135  3136  3181
 [556]  3192  3198  3199  3222  3234  3241  3266  3284  3311  3319  3323  3337  3366  3372  3381
 [571]  3386  3399  3405  3414  3418  3436  3452  3456  3479  3483  3511  3513  3522  3530  3558
 [586]  3598  3602  3612  3616  3628  3636  3645  3682  3688  3697  3698  3710  3715  3717  3730
 [601]  3736  3737  3742  3751  3753  3754  3765  3772  3773  3779  3787  3791  3793  3797  3808
 [616]  3809  3817  3826  3827  3846  3867  3886  3889  3898  3916  3933  3934  3978  4002  4003
 [631]  4005  4035  4049  4076  4140  4167  4199  4206  4242  4253  4293  4321  4333  4415  4425
 [646]  4428  4429  4435  4444  4504  4531  4538  4546  4553  4562  4570  4582  4588  4615  4622
 [661]  4638  4665  4720  4736  4737  4770  4807  4816  4819  4826  4844  4861  4881  4894  4899
 [676]  4926  4931  4942  4957  4972  5113  5116  5128  5131  5187  5222  5239  5296  5316  5352
 [691]  5455  5459  5504  5518  5530  5583  5648  5686  5733  5737  5763  5769  5775  5804  5892
 [706]  5917  5923  5966  6032  6049  6062  6078  6103  6108  6113  6117  6200  6250  6280  6281
 [721]  6308  6319  6323  6332  6339  6378  6397  6406  6422  6427  6452  6493  6497  6527  6539
 [736]  6614  6671  6785  6874  6875  6899  6934  6956  6961  6972  6979  6985  7005  7018  7019
 [751]  7024  7058  7109  7116  7217  7257  7274  7330  7357  7371  7387  7391  7431  7465  7505
 [766]  7516  7546  7581  7645  7707  7717  7828  7834  7851  7852  7908  7909  7912  7923  7960
 [781]  7973  8007  8012  8056  8061  8139  8207  8235  8238  8292  8293  8305  8310  8312  8366
 [796]  8380  8407  8420  8433  8439  8462  8463  8472  8491  8501  8504  8522  8549  8560  8574
 [811]  8593  8613  8638  8677  8701  8718  8732  8745  8793  8865  8880  8899  8945  8984  8994
 [826]  9040  9046  9057  9089  9140  9156  9177  9209  9212  9216  9222  9237  9267  9275  9287
 [841]  9360  9361  9366  9369  9376  9381  9391  9413  9423  9429  9455  9495  9504  9515  9521
 [856]  9537  9547  9606  9645  9658  9678  9680  9716  9719  9734  9745  9802  9822  9857  9858
 [871]  9880  9881  9898  9915  9925  9936  9987 10038 10042 10122 10132 10186 10200 10201 10207
 [886] 10208 10213 10235 10253 10265 10336 10380 10383 10399 10401 10405 10411 10429 10442 10452
 [901] 10477 10526 10527 10589 10639 10659 10723 10726 10751 10779 10793 10800 10838 10855 10858
 [916] 10899 10908 10930 10953 10995 10997 11068 11071 11090 11102 11106 11109 11134 11147 11199
 [931] 11235 11248 11250 11272 11290 11295 11313 11337 11361 11364 11433 11439 11455 11477 11478
 [946] 11558 11590 11744 11791 11819 11829 11833 11845 11901 11914 11925 11926 11948 11954 12011
 [961] 12030 12065 12077 12152 12175 12194 12214 12250 12257 12260 12263 12271 12274 12282 12285
 [976] 12325 12375 12386 12396 12428 12459 12464 12478 12487 12507 12524 12562 12572 12604 12605
 [991] 12625 12653 12654 12706 12709 12712 12781 12789 12793 12795
 [ reached getOption("max.print") -- omitted 1972 entries ]
>
```

Missing data:

Upon examining the dataset and the accompanying graph, it becomes evident that both the "reviews_per_month" and "last_review" columns exhibit an identical percentage of missing values, approximately 20.56%. This correlation is logical since the absence of information regarding the last review date prevents the calculation of reviews per month.