Fundamentals of Data Analytics and Statistics

NYC AirBnB Dataset Analysis And Visualization

INSTRUCTOR : HAMID RAJAEE

BY : VANDANA MURALIDHARAN

UNDERSTANDING BUSINESS

Here I have a dataset of AirBnB collected from New York city. Airbnb is an online platform that connects people who want to rent out their homes for people looking for accommodations in that locality for a cheaper price with more facilities than a regular hotel. NYC is the most popular city in United States for tourism and business, and one of the most populated cities known.

Nowadays, Airbnb's are becoming very popular service and more people have started to prefer such accommodations rather than hotels. Therefore, the Data analysis of these data's have become an essential element for Airbnb companies that provide millions of listings through out the year. These data's can be analyzed to make useful business decisions, like understanding customers behavioral pattern in choosing their accommodation, current market trends etc. These analysis can be used to improve the business based on customer needs. It can be achieved by implementing innovative additional services, managing marketing strategies, and much more.

In this project my dataset consists of NYC AirBnB listing details from the year 2015 to 2019 , with 37788 observations and 14 variables. Our goal is to use this data of 5 years to Perform EDA , analyze and visualize it to identify customer behavioral pattern in all the locations to find out their preference based on their reviews and detect areas that need improvement to increase the business. And suggest strategies to improve the customer base in coming years.

BUSINESS QUESTIONS

What are the major locations in NYC and how are the distribution of AirBnB in these locations?

How prices of AirBnB vary based on these locations?

What are the different types of properties available around NYC and what kind of accommodation is preferred by customers?

Are the demand and prices of the rentals correlated?

What is the requirement of minimum night stay based on each location ?

What localities are highly rated by the customers?

Has the popularity of AirBnB increased over the years?

Which season is the busiest time for business?

Most popular hosts in the NYC AirBnB ?

ATTRIBUTES

- **Host_ID** – Online generated identification number for host
- **Host_Name** – Name of the host who has listed the property
- **Neighbourhood_group-** Locations in NYC
- **Neighbourhood** – places inside each neighbourhood group
- **Latitude** – Description of location
- **Longitude** – Description of location
- **Property_type** – Type of listed property
- **Room_type** – Type of accommodation offered
- **Price** – Cost of accommodation per night
- **Minimum_nights** – Minimum night required to stay
- **Review_Score_Rating** – Rating score based on stay
- **Number_of_Review** – Number of reviews for each host
- **Last_review** – Date a review has been given
- **Listing_Count** – Number of listing by host

ANALYZING AND GETTING FAMILIAR WITH DATA

Importing the Dataset

```
PROC IMPORT OUT= PROJECT.AirBnB
     DATAFILE= "C:\Users\vanda\OneDrive\Desktop\PROJECT DATASETS\AirBnBdataset.csv"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
      DATAROW=2;
           guessingrows = max;
RUN;
```

```
NOTE: PROJECT.AIRBNB data set was successfully created.
NOTE: The data set PROJECT.AIRBNB has 37788 observations and 14 variables.
NOTE: PROCEDURE IMPORT used (Total process time):
      real time           6.14 seconds
      cpu time            6.20 seconds
```

VIEWTABLE: Project.Airbnb

	host_id	hoSt_name	neighbourHood_group	neighbourhood	latitude	longitude	property_type	rooM_type	price
1	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Apartment	Private room	149
2	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Apartment	Entire home/apt	225
3	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.9419	Apartment	Private room	150
4	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Apartment	Entire home/apt	89
5	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	House	Entire home/apt	80
6	7322	Chris	Manhattan	Murray Hill	40.74767	-73.975	Apartment	Entire home/apt	200
7	7356	Garon	Brooklyn	Bedford-Stuyvesant	40.68688	-73.95596	Condominium	Private room	60
8	8967	Shunichi	Manhattan	Hell's Kitchen	40.76489	-73.98493	Apartment	Private room	79
9	7549	Ben	Manhattan	Chinatown	40.71344	-73.99037	Apartment	Entire home/apt	150
10	7702	Lena	Manhattan	Upper West Side	40.80316	-73.96545	Apartment	Entire home/apt	135
11	7989	Kate	Manhattan	Hell's Kitchen	40.76076	-73.98867	Apartment	Private room	85
12	9744	Laurie	Brooklyn	South Slope	40.66829	-73.98779	Apartment	Private room	89
13	11528	Claudio	Manhattan	Upper West Side	40.79826	-73.96113	Apartment	Private room	85
14	11975	Alina	Manhattan	West Village	40.7353	-74.00525	Loft	Entire home/apt	120
15	15991	Allen & Irina	Brooklyn	Williamsburg	40.70837	-73.95352	Apartment	Entire home/apt	140
16	17571	Jane	Brooklyn	Fort Greene	40.69169	-73.97185	Apartment	Entire home/apt	215
17	18946	Doti	Manhattan	Chelsea	40.74192	-73.99501	Apartment	Private room	140
18	20950	Adam And Charity	Brooklyn	Crown Heights	40.67592	-73.94694	Apartment	Entire home/apt	99
19	22486	Lisel	Brooklyn	Park Slope	40.68069	-73.97706	Apartment	Private room	130
20	22486	Lisel	Brooklyn	Park Slope	40.67989	-73.97798	Apartment	Private room	80
21	22486	Lisel	Brooklyn	Park Slope	40.68001	-73.97865	Apartment	Private room	110
22	25183	Nathalie	Brooklyn	Bedford-Stuyvesant	40.68371	-73.94028	Apartment	Entire home/apt	120
23	25326	Gregory	Brooklyn	Windsor Terrace	40.65599	-73.97519	Apartment	Private room	60

Output - (Untitled) | Log - (Untitled) | AirBnB project * | Editor - Untitled2 * | VIEWTABLE: Project.Ai...

Browsing The Description Portion

```
PROC CONTENTS DATA=PROJECT.AirBnB order=varnum ;
RUN;
```

The CONTENTS Procedure

Data Set Name	PROJECT.AIRBNB	Observations	37788
Member Type	DATA	Variables	14

Variables in Creation Order

#	Variable	Type	Len	Format	Informat
1	host_id	Num	8	BEST12.	BEST32.
2	hoSt_name	Char	82	$82.	$82.
3	neighbourHood_group	Char	13	$13.	$13.
4	neighbourhood	Char	28	$28.	$28.
5	latitude	Num	8	BEST12.	BEST32.
6	longitude	Num	8	BEST12.	BEST32.
7	property_type	Char	18	$18.	$18.
8	rooM_type	Char	15	$15.	$15.
9	price	Num	8	BEST12.	BEST32.
10	minimum_nights	Num	8	BEST12.	BEST32.
11	review_scores_rating	Num	8	BEST12.	BEST32.
12	number_of_reviews	Num	8	BEST12.	BEST32.
13	last_review	Num	8	MMDDYY10.	MMDDYY10.
14	listings_count	Num	8	BEST12.	BEST32.

FINDING HEAD OF DATA FOR 10 OBSERVATION

```
proc print data= PROJECT.AirBnB (obs=10);
run;
```

Obs	host_id	hoSt_name	neighbourHood_group	neighbourhood	latitude	longitude	property_type	rooM_type	price	minimum_nights	review_scores_rating	number_of_reviews	last_review	listings_count
1	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Apartment	Private room	149	1	9	9	10/19/2018	6
2	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Apartment	Entire home/apt	225	1	9	45	05/21/2019	2
3	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.9419	Apartment	Private room	150	3	10	0	07/21/2017	1
4	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Apartment	Entire home/apt	89	1	9	270	07/05/2019	1
5	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	House	Entire home/apt	80	10	7	9	11/19/2018	1
6	7322	Chris	Manhattan	Murray Hill	40.74767	-73.975	Apartment	Entire home/apt	200	3	10	74	06/22/2019	1
7	7356	Garon	Brooklyn	Bedford-Stuyvesant	40.68688	-73.95596	Condominium	Private room	60	45	10	49	10/05/2017	1
8	8967	Shunichi	Manhattan	Hell's Kitchen	40.76489	-73.98493	Apartment	Private room	79	2	9	430	06/24/2019	1
9	7549	Ben	Manhattan	Chinatown	40.71344	-73.99037	Apartment	Entire home/apt	150	1	10	160	06/09/2019	4
10	7702	Lena	Manhattan	Upper West Side	40.80316	-73.96545	Apartment	Entire home/apt	135	5	9	53	06/22/2019	1

FINDING TAIL OF DATA FOR LAST 10 OBSERVATION

```
proc print data= PROJECT.AirBnB(obs=37788  firstobs=37779);
run;
```

Obs	host_id	hoSt_name	neighbourHood_group	neighbourhood	latitude	longitude	property_type	rooM_type	price	minimum_nights	review_scores_rating	number_of_reviews	last_review	listings_count
37779	5288346	Celine	Manhattan	Upper East Side	40.7603	-73.96225	Apartment	Entire home/apt	110	1	8	1	07/01/2019	1
37780	46232598	Schmid	Manhattan	Upper East Side	40.77001	-73.94915	Condominium	Entire home/apt	33	1	10	1	07/01/2019	1
37781	271844440	Chris	Queens	Rockaway Beach	40.59029	-73.81277	Apartment	Entire home/apt	45	1	8	1	07/04/2019	3
37782	271885652	Kailey	Queens	Rockaway Beach	40.5879	-73.81269	Apartment	Entire home/apt	150	1	10	1	07/06/2019	4
37783	1409706	Phoenix	Brooklyn	Fort Greene	40.68889	-73.97632	Apartment	Private room	550	6	6	1	07/02/2019	2
37784	258998574	Alejandro	Brooklyn	Bushwick	40.70384	-73.92232	Apartment	Private room	129	2	10	1	07/03/2019	1
37785	208514239	Rehana	Brooklyn	Williamsburg	40.71232	-73.9422	Apartment	Entire home/apt	45	1	10	1	07/07/2019	3
37786	3850264	Linda Lou	Manhattan	Harlem	40.80658	-73.95736	Apartment	Private room	235	3	10	1	07/03/2019	2
37787	256197494	Farina	Brooklyn	Cypress Hills	40.68042	-73.88978	Apartment	Private room	100	1	10	1	07/05/2019	1
37788	272557707	Prince	Staten Island	Rosebank	40.6075	-74.07979	Apartment	Private room	30	1	10	1	07/05/2019	1

Number Of Unique/Distinct Values In All Variables

```
Proc freq data=PROJECT.AirBnB nlevels;
   ods exclude onewayfreqs;
run;
```

Number of Variable Levels	
Variable	Levels
host_id	29515
hoSt_name	9730
neighbourHood_group	5
neighbourhood	218
latitude	17227
longitude	13493
property_type	6
rooM_type	3
price	576
minimum_nights	87
review_scores_rating	9
number_of_reviews	394
last_review	1559
listings_count	47

Finding Unique Value in Each Variable Group

```
proc freq data = PROJECT.AirBnB;
table neighbourhood_group room_type property_type
neighbourhood /nopercent nocum;
run;
```

neighbourHood_group	Frequency
Bronx	849
Brooklyn	16020
Manhattan	16173
Queens	4443
Staten Island	303

rooM_type	Frequency
Entire home/apt	19769
Private room	17195
Shared room	824

property_type	Frequency
Apartment	31176
Condominium	913
House	2800
Loft	1182
Serviced apartment	474
Townhouse	1243

neighbourhood	Frequency
Allerton	37
Arden Heights	4
Arrochar	20
Arverne	65
Astoria	689
Bath Beach	15
Battery Park City	33
Bay Ridge	115
Bay Terrace	5
Bay Terrace, Staten Island	2
Baychester	6
Bayside	30
Bayswater	9
Bedford-Stuyvesant	3063
Belle Harbor	5
Bellerose	8
Belmont	20
Bensonhurst	59
Bergen Beach	8
Boerum Hill	145

Data Cleaning

DROPPING THE VARIABLE LATITUDE AND LONGITUDE AS THEY DONT ADD ANY MEANINGFUL INFORMATION

```
DATA PROJECT.AirBnB1(DROP = latitude longitude);
SET PROJECT.AirBnB;
RUN;
PROC CONTENTS DATA=PROJECT.AirBnB1 order=varnum ;
RUN;
```

Variables in Creation Order

#	Variable	Type	Len	Format	Informat
1	host_id	Num	8	BEST12.	BEST32.
2	hoSt_name	Char	82	$82.	$82.
3	neighbourHood_group	Char	13	$13.	$13.
4	neighbourhood	Char	28	$28.	$28.
5	property_type	Char	18	$18.	$18.
6	rooM_type	Char	15	$15.	$15.
7	price	Num	8	BEST12.	BEST32.
8	minimum_nights	Num	8	BEST12.	BEST32.
9	review_scores_rating	Num	8	BEST12.	BEST32.
10	number_of_reviews	Num	8	BEST12.	BEST32.
11	last_review	Num	8	MMDDYY10.	MMDDYY10.
12	listings_count	Num	8	BEST12.	BEST32.

CHANGING ALL THE VARIABLE NAMES TO UPPERCASE

```
option VALIDVARNAME=UPCASE;
proc contents data= PROJECT.AirBnB1
out=PROJECT.AirBnBNY order = varnum; run;
```

Variables in Creation Order

#	Variable	Type	Len	Format	Informat
1	HOST_ID	Num	8	BEST12.	BEST32.
2	HOST_NAME	Char	82	$82.	$82.
3	NEIGHBOURHOOD_GROUP	Char	13	$13.	$13.
4	NEIGHBOURHOOD	Char	28	$28.	$28.
5	PROPERTY_TYPE	Char	18	$18.	$18.
6	ROOM_TYPE	Char	15	$15.	$15.
7	PRICE	Num	8	BEST12.	BEST32.
8	MINIMUM_NIGHTS	Num	8	BEST12.	BEST32.
9	REVIEW_SCORES_RATING	Num	8	BEST12.	BEST32.
10	NUMBER_OF_REVIEWS	Num	8	BEST12.	BEST32.
11	LAST_REVIEW	Num	8	MMDDYY10.	MMDDYY10.
12	LISTINGS_COUNT	Num	8	BEST12.	BEST32.

RENAMING VARIABLE NEIGHBOURHOOD_GROUP TO LOCATION

```
DATA PROJECT.AirBnB3;
SET PROJECT.AirBnB1(RENAME=(NEIGHBOURHOOD_GROUP = LOCATION));
run;
proc print data= PROJECT.AirBnB3 (OBS=5);* Viewing first 5 rows of the dataset;
run;
```

Obs	HOST_ID	HOST_NAME	LOCATION	NEIGHBOURHOOD	PROPERTY_TYPE	ROOM_TYPE	PRICE	MINIMUM_NIGHTS	REVIEW_SCORES_RATING	NUMBER_OF_REVIEWS	LAST_REVIEW	LISTINGS_COUNT
1	2787	John	Brooklyn	Kensington	Apartment	Private room	149	1	9	9	10/19/2018	6
2	2845	Jennifer	Manhattan	Midtown	Apartment	Entire home/apt	225	1	9	45	05/21/2019	2
3	4632	Elisabeth	Manhattan	Harlem	Apartment	Private room	150	3	10	0	07/21/2017	1
4	4869	LisaRoxanne	Brooklyn	Clinton Hill	Apartment	Entire home/apt	89	1	9	270	07/05/2019	1
5	7192	Laura	Manhattan	East Harlem	House	Entire home/apt	80	10	7	9	11/19/2018	1

CHECKING FOR MISSING VALUES

```
proc sql;
    select nmiss(host_name) as HOST_NAME, nmiss(location) as LOCATION,nmiss(property_type) as
PROPERTY_TYPE,
nmiss(room_type) as ROOM_TYPE
    from PROJECT.AirBnB3;
quit;
proc means data= PROJECT.AirBnB3 nmiss;
run;
```

The SAS System

The MEANS Procedure

Variable	N Miss
host_id	0
latitude	0
longitude	0
price	0
minimum_nights	0
number_of_reviews	0
last_review	0
listings_count	0

DROPPING DUPLICATE OBSERVATIONS

```
PROC SORT DATA = PROJECT.AirBnB3 OUT = PROJECT.AirBnBdata
NODUPKEY;
BY _ALL_;
RUN;
```

```
NOTE: There were 37788 observations read from the data set PROJECT.AIRBNB3.
NOTE: 1 observations with duplicate key values were deleted.
NOTE: The data set PROJECT.AIRBNBDATA has 37787 observations and 12 variables.
NOTE: PROCEDURE SORT used (Total process time):
      real time             0.05 seconds
      cpu time              0.03 seconds
```

DESCRIPTIVE ANALYSIS OF CONTINOUS VARIABLES

```
TITLE "DESCRIPTIVE ANALYSIS OF CONTINUOUS VARIABLES";
PROC MEANS DATA=PROJECT.AirBnBdata  N NMISS MIN Q1 MEDIAN  Q3 MAX qrange
mean std cv clm maxdec=2  ;
VAR PRICE
MINIMUM_NIGHTS
REVIEW_SCORES_RATING
NUMBER_OF_REVIEWS
LISTINGS_COUNT ;
RUN;
title;
```

DESCRIPTIVE ANALYSIS OF CONTINUOUS VARIABLES

The MEANS Procedure

Variable	N	N Miss	Minimum	Lower Quartile	Median	Upper Quartile	Maximum	Quartile Range	Mean	Std Dev	Coeff of Variation	Lower 95% CL for Mean	Upper 95% CL for Mean
PRICE	37787	0	15.00	69.00	101.00	170.00	10000.00	101.00	142.28	196.66	138.22	140.30	144.26
MINIMUM_NIGHTS	37787	0	1.00	1.00	2.00	4.00	999.00	3.00	5.80	16.06	276.91	5.64	5.96
REVIEW_SCORES_RATING	37787	0	2.00	9.00	10.00	10.00	10.00	1.00	9.59	0.83	8.69	9.58	9.60
NUMBER_OF_REVIEWS	37787	0	0.00	3.00	10.00	34.00	629.00	31.00	29.61	48.54	163.94	29.12	30.10
LISTINGS_COUNT	37787	0	1.00	1.00	1.00	2.00	327.00	1.00	5.18	26.34	508.39	4.91	5.45

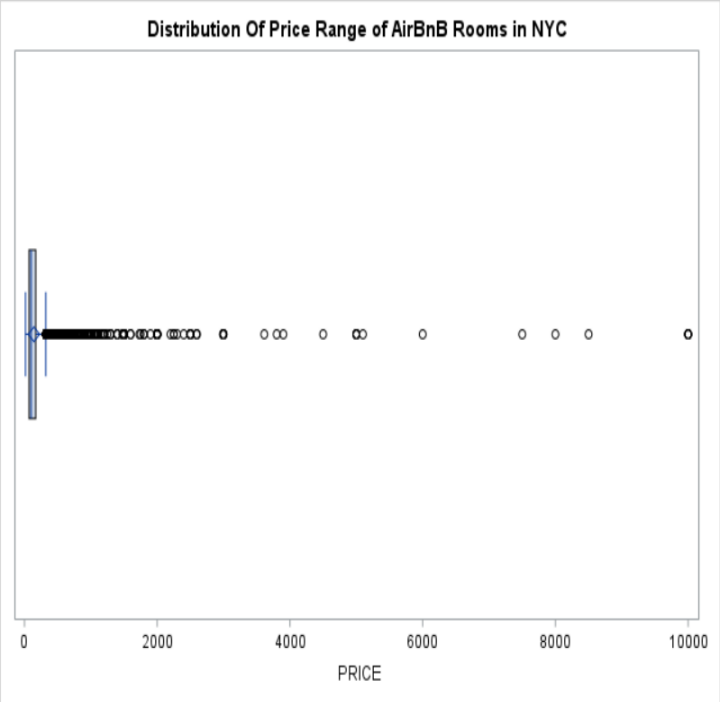
Outliers →

Lower fence Q1 – (1.5* IQR) →-82.5

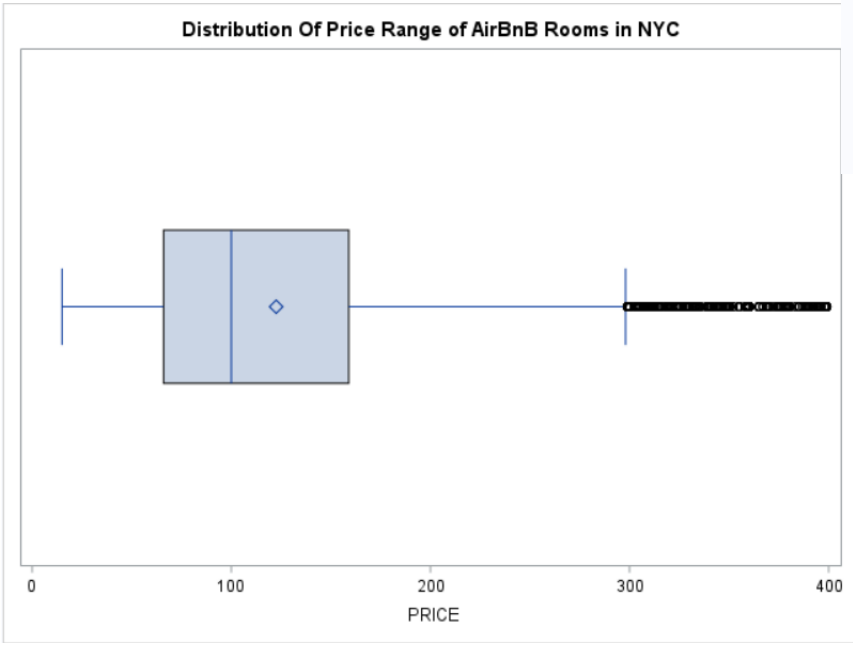
Upper fence Q3 + (1.5* IQR) →321.5

WHAT IS THE DISTRIBUTION OF PRICE RANGE OF AIRBNB ROOMS?

```
Title "Distribution Of Price Range of AirBnB Rooms in NYC";
proc sgplot data=PROJECT.AirBnBNyc;
 hbox Price;
 run;
```




```
DATA PROJECT.AirBnBT;
SET PROJECT.AirBnBDATA;
WHERE PRICE <400;
run;
proc sgplot data=PROJECT.AirBnBT;
 hbox Price;
 run;
```



The UNIVARIATE Procedure
Variable: PRICE

Moments			
N	36559	Sum Weights	36559
Mean	122.542794	Sum Observations	4480042
Std Deviation	73.2356979	Variance	5363.46744
Skewness	1.20483324	Kurtosis	1.23326835
Uncorrected SS	745074506	Corrected SS	196077643
Coeff Variation	59.7633656	Std Error Mean	0.38302372

Basic Statistical Measures			
Location		Variability	
Mean	122.5428	Std Deviation	73.23570
Median	100.0000	Variance	5363
Mode	150.0000	Range	384.00000
		Interquartile Range	93.00000

Quantiles (Definition 5)	
Level	Quantile
100% Max	399
99%	350
95%	275
90%	225
75% Q3	159
50% Median	100
25% Q1	66
10%	49
5%	40
1%	30
0% Min	15

Outliers →

Lower fence Q1 – (1.5* IQR) →-82.5

Upper fence Q3 + (1.5* IQR) →321.5

we can see that the first 25 percentile of **minimum_nights** are 1, the median is 2 and the 75% percentile is 4, So a sensible way of categorization would be one night, two nights, three nights, four nights and five nights ,six nights ,seven nights or more

```
data PROJECT.AirBnBMinNght;
set PROJECT.AirBnBdata;
length MIN_NIGHT $50;
if MINIMUM_NIGHTS =1 then
do;
MIN_NIGHT = "One Night";
end;
else if MINIMUM_NIGHTS = 2 then
do;
MIN_NIGHT = "Two Nights";
end;
else if MINIMUM_NIGHTS = 3 then
do;
MIN_NIGHT = "Three Nights";
end;
else if MINIMUM_NIGHTS = 4 then
do;
MIN_NIGHT = "Four Nights";
end;
else if MINIMUM_NIGHTS = 5 then
do;
MIN_NIGHT = "Five Nights";
end;
else if MINIMUM_NIGHTS =6 then
do;
MIN_NIGHT = "Six Nights";
end;
else if MINIMUM_NIGHTS = 7 then
do;
MIN_NIGHT = "Seven Nights";
end;
else
do;
MIN_NIGHT = "Seven Nights or More";
end;
proc print data= PROJECT.AirBnBMinNght(OBS=35);
run;
```

Variable	N	N Miss	Minimum	Lower Quartile	Median	Upper Quartile	Maximum	Quartile Range	Mean	Std Dev	Coeff of Variation	Lower 95% CL for Mean	Upper 95% CL for Mean
MINIMUM_NIGHTS	37787	0	1.00	1.00	2.00	4.00	999.00	3.00	5.80	16.06	276.91	5.64	5.96

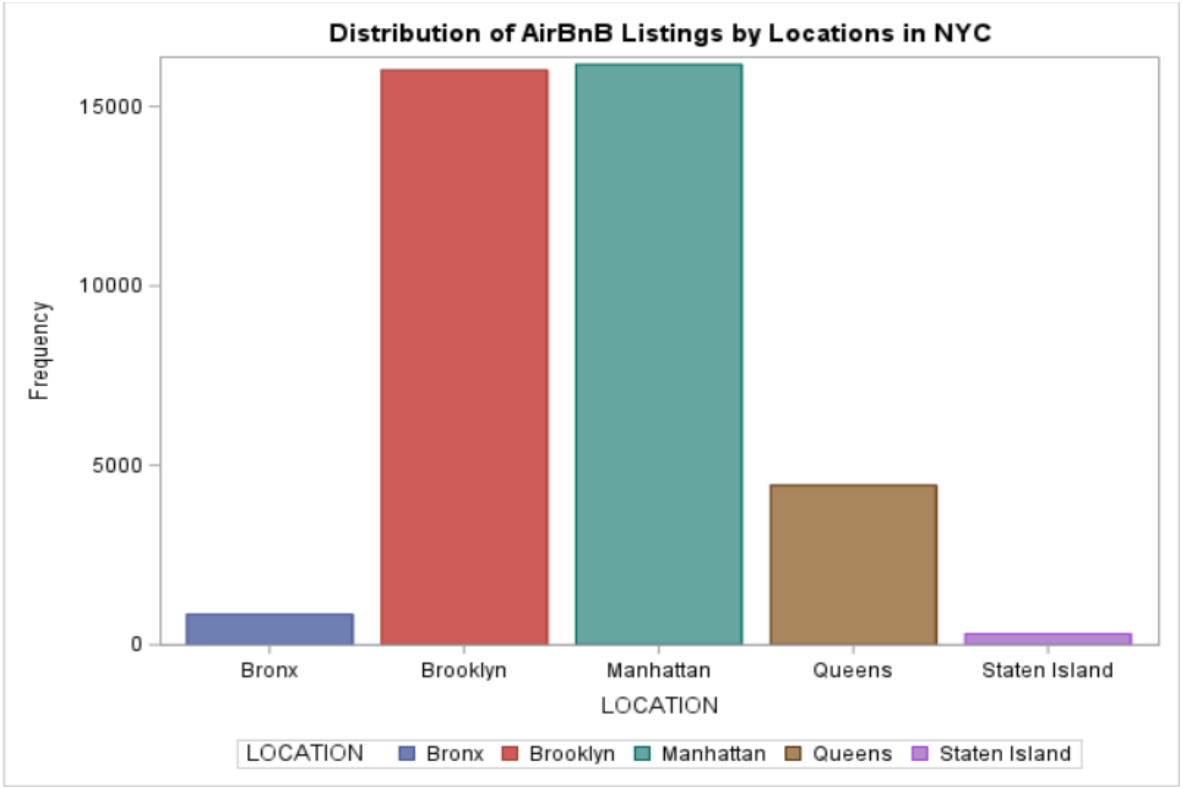
Private room	75	1	10	1	05/26/2019	1	One Night
Private room	59	3	10	4	06/02/2019	1	Three Nights
Entire home/apt	79	4	10	60	06/25/2019	2	Four Nights
Entire home/apt	99	3	9	48	06/23/2019	2	Three Nights
Private room	95	1	10	80	07/02/2019	2	One Night
Private room	95	1	9	92	06/29/2019	2	One Night
Private room	99	1	10	106	06/21/2019	2	One Night
Private room	74	1	10	2	11/06/2016	2	One Night
Private room	150	3	10	0	07/21/2017	1	Three Nights
Entire home/apt	89	1	9	270	07/05/2019	1	One Night
Entire home/apt	151	2	9	43	07/01/2019	1	Two Nights
Private room	47	2	10	14	01/02/2018	1	Two Nights
Entire home/apt	160	5	9	4	05/09/2019	1	Five Nights
Entire home/apt	80	10	7	9	11/19/2018	1	Seven Nights or More
Entire home/apt	220	3	10	108	06/21/2019	1	Three Nights
Entire home/apt	90	14	9	1	01/02/2019	1	Seven Nights or More
Entire home/apt	200	3	10	74	06/22/2019	1	Three Nights
Private room	85	2	10	23	06/10/2019	3	Two Nights

QUESTION 1. WHAT ARE THE DIFFERENT LOCATIONS OF AIRBNB LISTINGS IN NYC?

```
title "Distribution of AirBnB Listings by Locations in NYC";

PROC SGPLOT DATA= PROJECT.AirBnBNyc ;
 VBAR location / GROUP=location ;
RUN ;

proc freq data =PROJECT.AirBnBNyc ;
table Location /nopercent nocum;
run;
```



Summarization is done using proc freq to create a frequency table

Distribution of AirBnB Listings by Locations in NYC

The FREQ Procedure

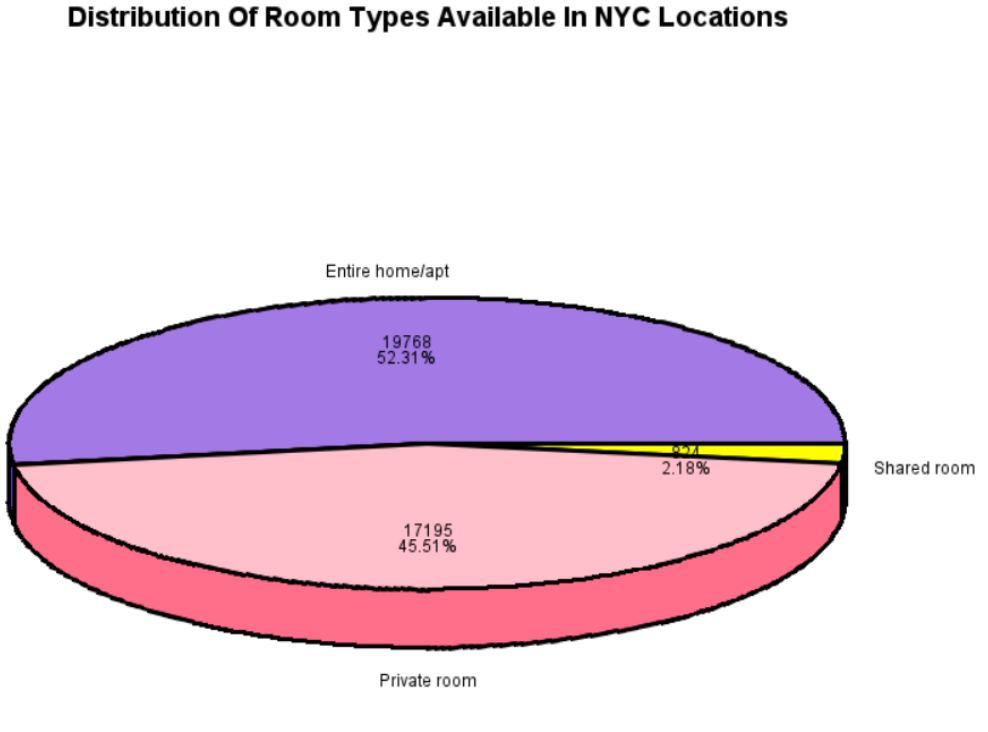
LOCATION	Frequency
Bronx	849
Brooklyn	16020
Manhattan	16172
Queens	4443
Staten Island	303

Barchart is used for visualization. From the above diagram we can see that Bronx,Brooklyn,Manhattan,Queens and Staten island are the available locations around NYC
With Manhattan region having the highest distribution of listings followed by Brooklin . Staten Island have the lowest listing

QUESTION 2. WHAT ARE THE DIFFERENT ROOM TYPES AVAILABLE IN THESE LOCATIONS ?

```
title "Distribution Of Room Types Available In NYC
Locations";
pattern1 color=VLIPB;
pattern2 color=PINK;
pattern3 color=YELLOW;

proc gchart data=PROJECT.AirBnBNyc;
 pie3d ROOM_TYPE /
 noheading percent=inside
 slice=outside value=inside
 coutline=black woutline=2;
run;
```



Summarization of available room types

Distribution Of Room Types Available In NYC Locations

The FREQ Procedure

ROOM_TYPE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Entire home/apt	19768	52.31	19768	52.31
Private room	17195	45.51	36963	97.82
Shared room	824	2.18	37787	100.00

From the pie chart we can see that there are 3 different room types available. With Entirehome/apt being highest with 19768 and shared room being the lowest available with 824 listing

QUESTION 3. WHAT IS THE DISTRIBUTION OF LISTING COUNT BETWEEN YEAR 2015 TO 2019?

```
Title "Distribution Of Listing counts Available Between the Year
2015 to 2019";
proc univariate data = PROJECT.AirBnBNyc plot normal;
var listings_count;
run;

Proc sgplot data=PROJECT.AirBnBNyc;
histogram listings_count/ binwidth=30 binstart=0
datalabel=percent;
density listings_count;
density listings_count/ type = kernal;
run;
quit;
```

Summarization is done with proc univariate

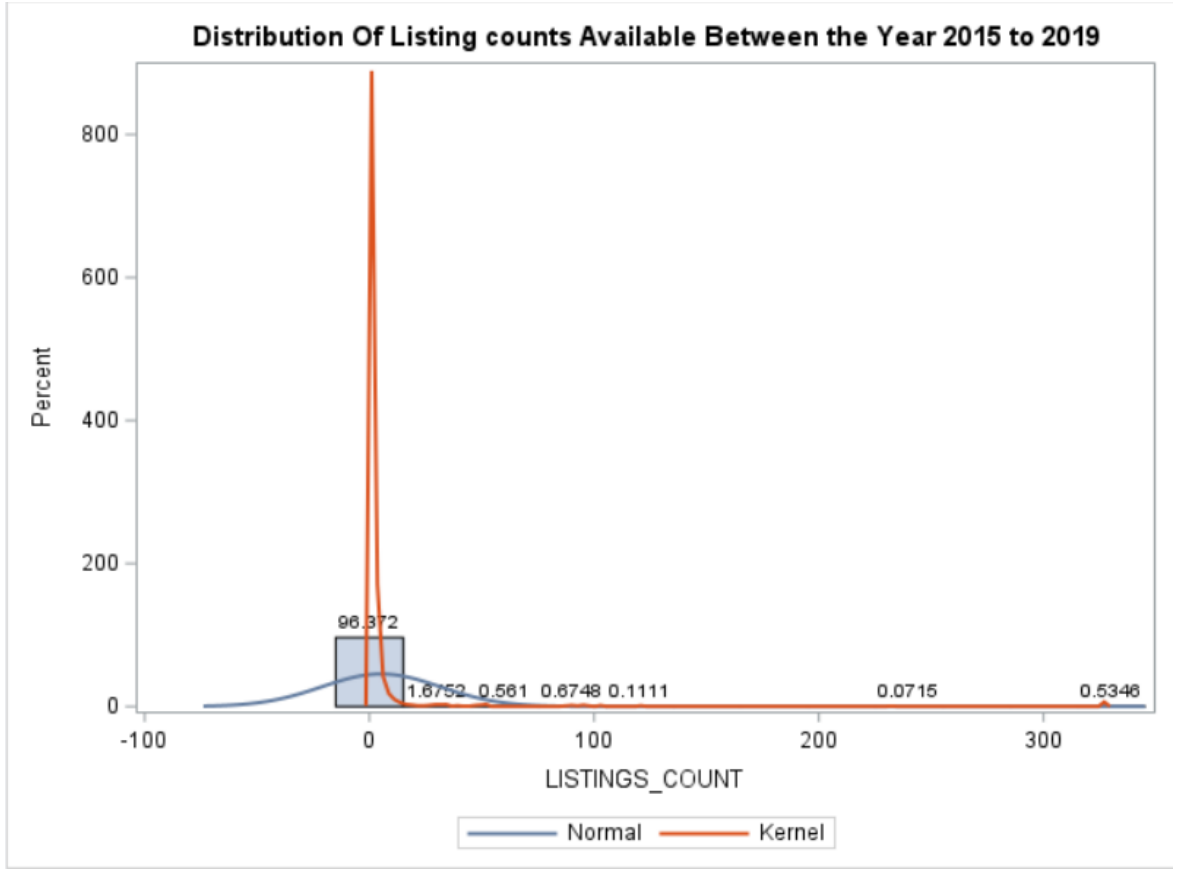
Distribution Of Listing counts Available Between the Year 2015 to 2019

The UNIVARIATE Procedure
Variable: LISTINGS_COUNT

Moments			
N	37787	Sum Weights	37787
Mean	5.18040596	Sum Observations	195752
Std Deviation	26.3365307	Variance	693.612851
Skewness	10.6099682	Kurtosis	121.529575
Uncorrected SS	27222930	Corrected SS	26208855.2
Coeff Variation	508.387391	Std Error Mean	0.1354838

Basic Statistical Measures			
Location		Variability	
Mean	5.180406	Std Deviation	26.33653
Median	1.000000	Variance	693.61285
Mode	1.000000	Range	326.00000
		Interquartile Range	1.00000

Quantiles (Definition 5)	
Level	Quantile
100% Max	327
99%	96
95%	9
90%	5
75% Q3	2
50% Median	1
25% Q1	1
10%	1
5%	1
1%	1
0% Min	1



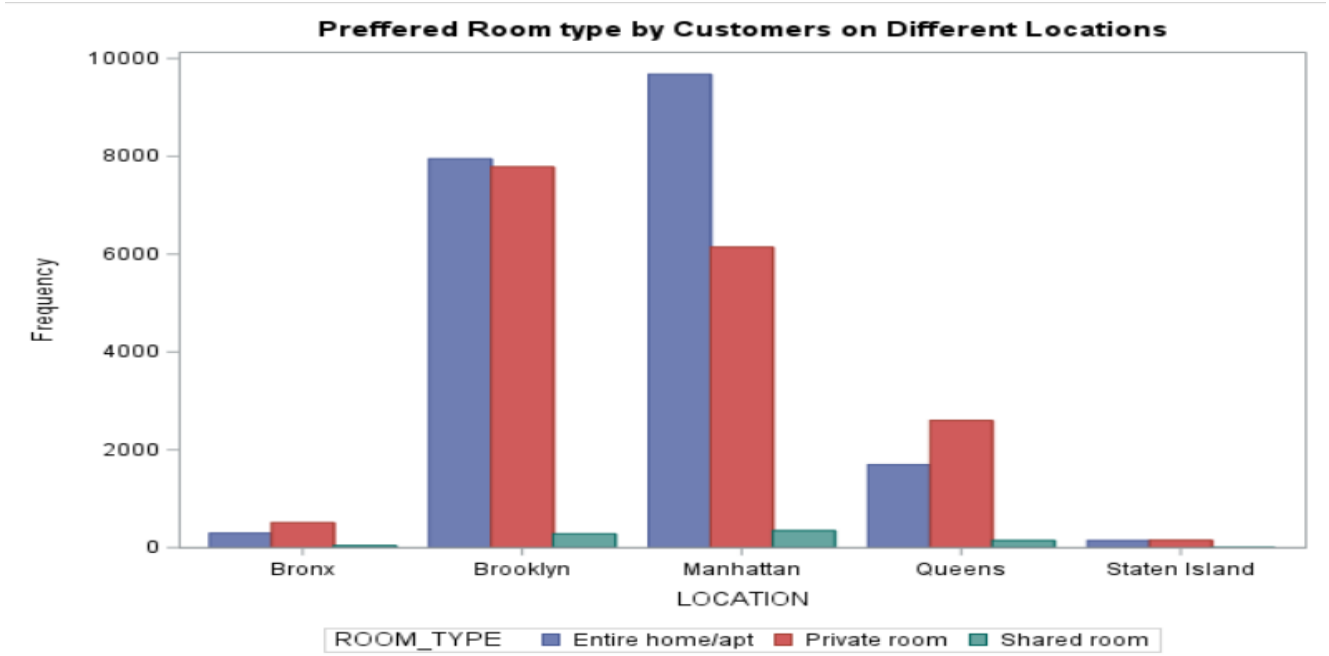
Visualization is done using histogram and density plot. From the histogram we can see that the distribution of listings available between 2015 to 2019 .The highest range falls between which is around 96 percentage

QUESTION 4. WHICH IS THE PREFFERED ROOM TYPE BY CUSTOMERS BASED ON LOCATION?

Chi-square Test

```
PROC FREQ DATA = PROJECT.AirBnBNyc;
TABLE ROOM_TYPE*LOCATION/chisq;
run;

Title "Preffered Room type by Customers on Different
Locations";
proc sgplot data =PROJECT.AirBnBNyc;
vbar location/ group = room_type groupdisplay = cluster;
run;
quit;
```



Summarization of preferred room type based on location is done using the contingency table as shown below.

The FREQ Procedure

Frequency / Percent / Row Pct / Col Pct	Table of ROOM_TYPE by LOCATION					
	LOCATION					
ROOM_TYPE	Bronx	Brooklyn	Manhattan	Queens	Staten Island	Total
Entire home/apt	295 0.78 1.49 34.75	7951 21.04 40.22 49.63	9680 25.62 48.97 59.86	1696 4.49 8.58 38.17	146 0.39 0.74 48.18	19768 52.31
Private room	512 1.35 2.98 60.31	7787 20.61 45.29 48.61	6144 16.26 35.73 37.99	2600 6.88 15.12 58.52	152 0.40 0.88 50.17	17195 45.51
Shared room	42 0.11 5.10 4.95	282 0.75 34.22 1.76	348 0.92 42.23 2.15	147 0.39 17.84 3.31	5 0.01 0.61 1.65	824 2.18
Total	849 2.25	16020 42.40	16172 42.80	4443 11.76	303 0.80	37787 100.00

Visualization is done using grouped bar chart . From the above bar chart we can see that in every location people prefer entirehome/apartment as their accommodation. And the least preferred type is shared room.

Statistics for Table of ROOM_TYPE by LOCATION

Statistic	DF	Value	Prob
Chi-Square	8	930.1636	<.0001
Likelihood Ratio Chi-Square	8	930.2887	<.0001
Mantel-Haenszel Chi-Square	1	0.0090	0.9243
Phi Coefficient		0.1569	
Contingency Coefficient		0.1550	
Cramer's V		0.1109	

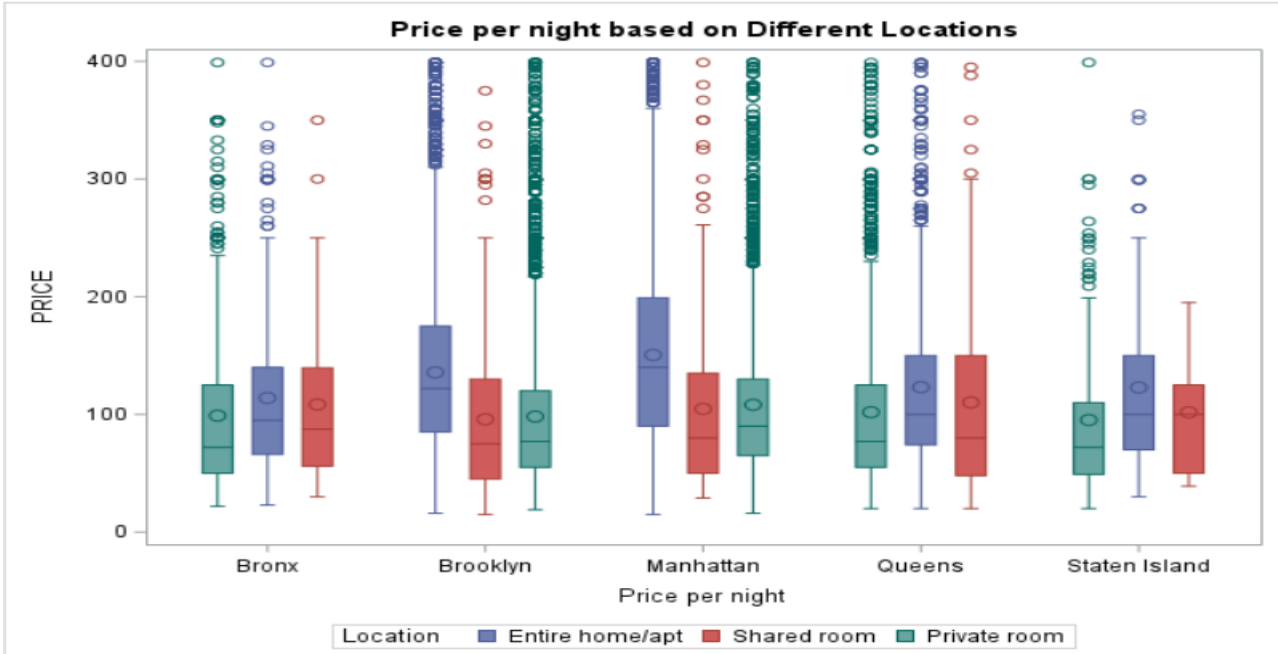
Sample Size = 37787

Test of Hypothesis

Here from chi-square test we can see the p value .0001 which is <0.05
As p value is less than 5% ,we reject null hypothesis and can say that there is statistical association between room type and the AirBnB location.

QUESTION 5. PRICE PER NIGHT BASED ON LOCATION AROUND NYC?

```
Title "Price per night based on Different Locations";
proc sgplot data=PROJECT.AirBnBT;
   vbox PRICE / category=location group=room_type;
   xaxis label="Price per night";
   keylegend / title="Location";
run;
```



Summarization

Summarization is done using proc univariate.

Class Level Information		
Class	Levels	Values
LOCATION	5	Bronx Brooklyn Manhattan Queens Staten Island

Moments			
N	293	Sum Weights	293
Mean	108.443686	Sum Observations	31774
Std Deviation	69.7968756	Variance	4871.60384
Skewness	1.40373322	Kurtosis	1.74928346
Uncorrected SS	4868198	Corrected SS	1422508.32
Coeff Variation	64.3623231	Std Error Mean	4.07757701

Basic Statistical Measures			
Location		Variability	
Mean	108.4437	Std Deviation	69.79688
Median	85.0000	Variance	4872
Mode	75.0000	Range	379.00000
		Interquartile Range	77.00000

Test of Normality is done using Shapiro-Wilk.

Since p value is less than 0.0001 which is <0.05. As the p value is less than 5% we conclude that data is not normally distributed

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.863284	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.165892	Pr > D	<0.0100
Cramer-von Mises	W-Sq	2.200691	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	12.58342	Pr > A-Sq	<0.0050

Test of Equality of variance

Levene's Test for Homogeneity of Variances is produced by PROC GLM to test if variances are considered equal across all groups . From Levene's test result we find that p value is <0.0001. As the P value is less than 5%, we reject null hypothesis and conclude variances are not equal.

Price per night based on Different Locations

The GLM Procedure

Levene's Test for Homogeneity of PRICE Variance
ANOVA of Absolute Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
LOCATION	4	289392	72347.9	35.73	<.0001
Error	36554	74015198	2024.8		

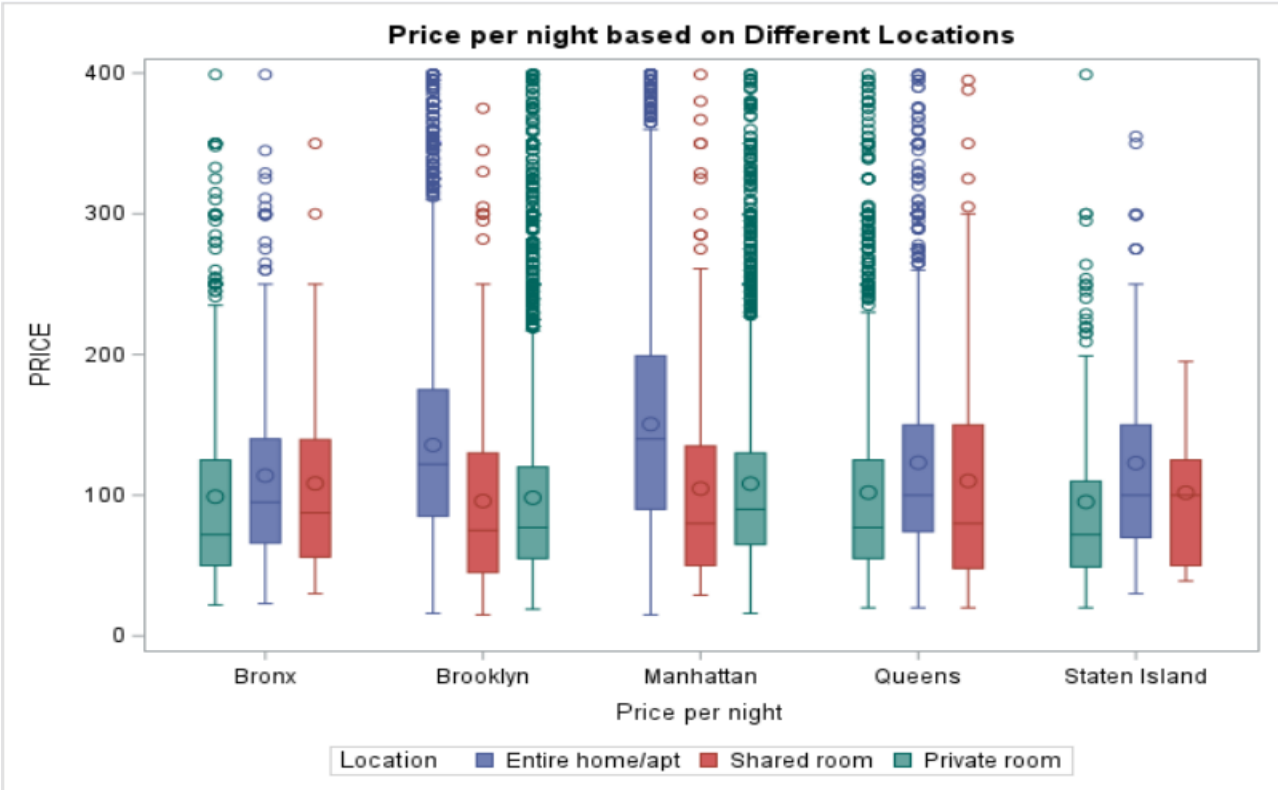
Test of Independency using Welch Anova

Since the Levene's test p-value < 0.05, we use a Welch's ANOVA

Welch's ANOVA for PRICE

Source	DF	F Value	Pr > F
LOCATION	4.0000	153.49	<.0001
Error	1687.8		

Based on the result from Welch Anova, the p value is 0.0001 , as it is less than 5 % we reject null hypothesis and accept alternate hypothesis that there is a statistical relationship between price and location.



Visualization

From the visualization of grouped box plot we can see that the price per night range is more in Manhattan and cheapest is the Staten island

.

QUESTION 6. HAS AIRBnB'S BECOME POPULAR OVER THE YEARS?

```
title "Popularity of AirBnB Over the
years";
data project.year;
            set PROJECT.AirBnBNyc;
            my_year = year(last_review);
run;
proc print data= project.year(obs = 10);
run;
title "Popularity of AirBnB Over the
years";
proc sgplot data=project.year;
    scatter x = last_review  y =
number_of_reviews;
xaxis label= "Year";
yaxis label ="Reviews";
run;
```

MBER_OF_REVIEWS	LAST_REVIEW	LISTINGS_COUNT	MIN_NIGHT	MY_YEAR
1	03/17/2018	1	Seven Nights or More	2018
27	05/21/2019	1	Seven Nights or More	2019
15	09/29/2018	6	One Night	2018
24	05/11/2019	6	One Night	2019
17	06/26/2019	6	One Night	2019
9	10/19/2018	6	One Night	2018
21	10/27/2018	6	One Night	2018
19	06/08/2019	6	One Night	2019
45	05/21/2019	2	One Night	2019
1	07/18/2018	2	One Night	2018

Pearson Correlation

Popularity of AirBnB Over the years

The CORR Procedure

2 Variables:	LAST_REVIEW NUMBER_OF_REVIEWS

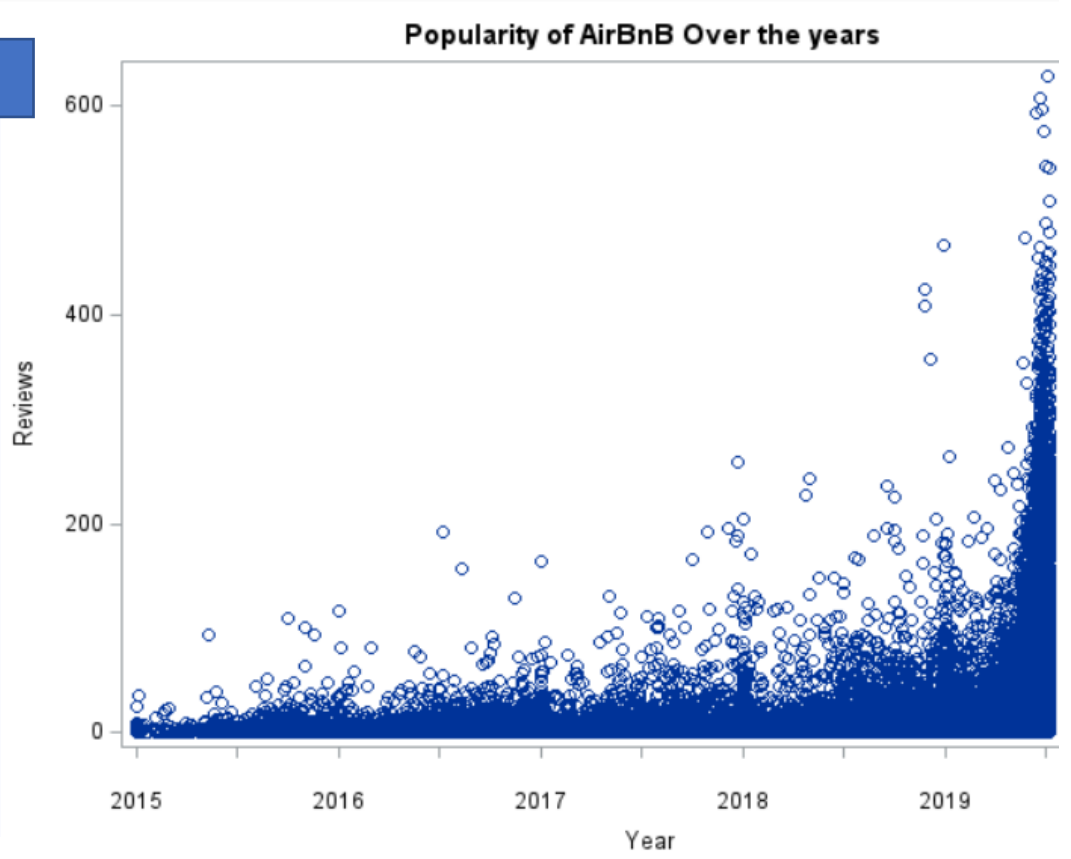
Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
LAST_REVIEW	37787	21472	390.29470	811356382	20089	21738
NUMBER_OF_REVIEWS	37787	29.60688	48.53614	1118755	0	629.00000

Pearson Correlation Coefficients, N = 37787
Prob > |r| under H0: Rho=0

	LAST_REVIEW	NUMBER_OF_REVIEWS
LAST_REVIEW	1.00000	0.28810 <.0001
NUMBER_OF_REVIEWS	0.28810 <.0001	1.00000

From the above table we can see that Correlation coefficient is 0.288. and the p value is <0.0001. That means year and number of review have a statistically significant linear relationship (r = 0.2881, p < .001).

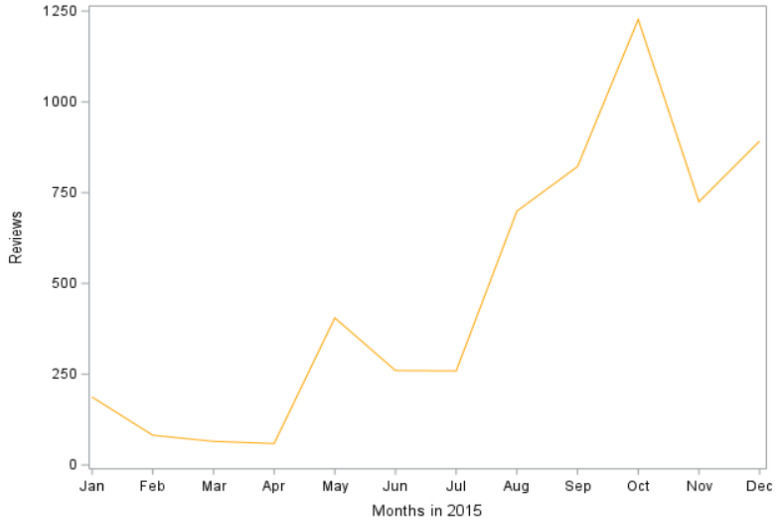


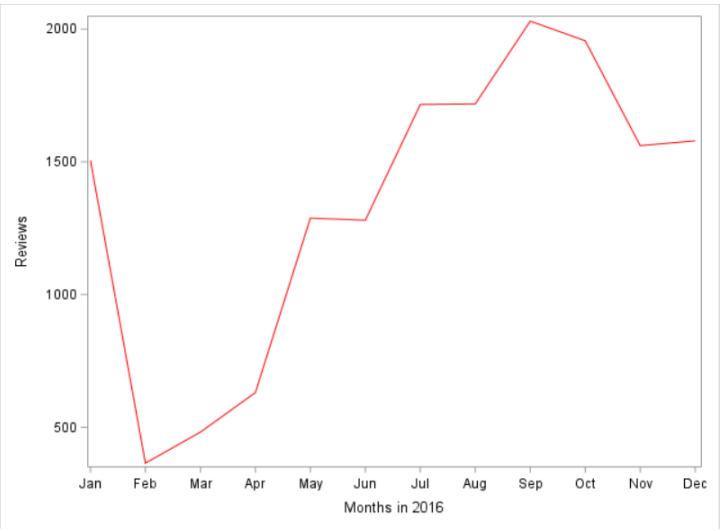
Visualization

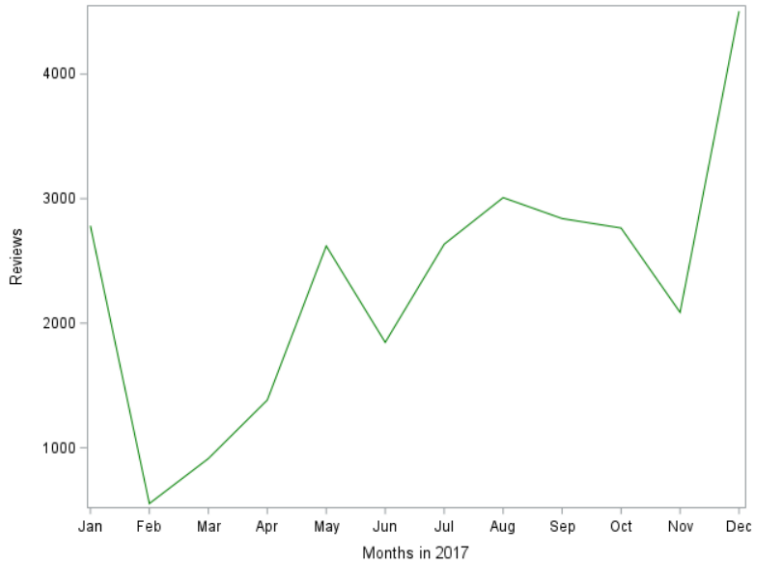
From the above scatter plot we can see that there is an increase in the amount of people preferring to stay in Airbnb's . From this we can see that Airbnb' s are becoming popular over the years. But has this affected the price?

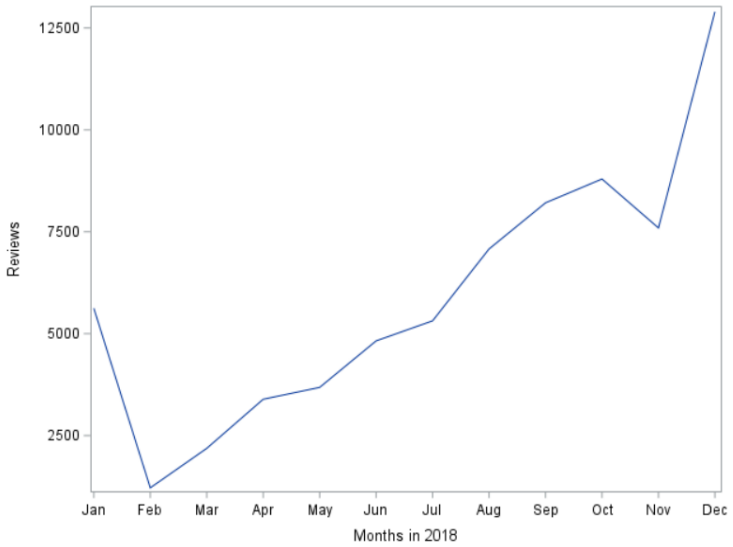
Question 7. WHICH SEASON IS THE BUSIEST TIME FOR BUSINESS?

```
title "Seasonal demand of airbnb";
data project.seasonal;
              set PROJECT.AirBnBNyc;
  month_year = intnx('month',last_review,0,'b');
  my_year = year(last_review);
run;
proc print data= project.seasonal(obs = 25);
format month_year  monname3.;
run;
*for year 2018;
DATA PROJECT.AirBnBmonth;
   SET project.seasonal;
   KEEP MONTH_YEAR MY_YEAR number_of_reviews ;
   where MY_YEAR = 2018;
RUN;
proc print data = PROJECT.AirBnBmonth (obs=25);
format month_year  monname3.;
run;
proc sql;
create table myseason as
select put (month_year,monname3.)as month,
sum(number_of_reviews) as review_sum
from PROJECT.AirBnBmonth
group by month_year;
run;
 data PROJECT.AirBnBmyseason;
 set myseason;
 run;
 proc print data =  PROJECT.AirBnBmyseason;
 run;
proc sgplot data= PROJECT.AirBnBmyseason;
    series x= month y=review_sum;
     xaxis label= "Months in 2018";
yaxis label ="Reviews";
run;
```









Visualization

From the time series line graph we can see that October through January is usually the busiest month for Airbnb booking . And February to April is the least busiest.

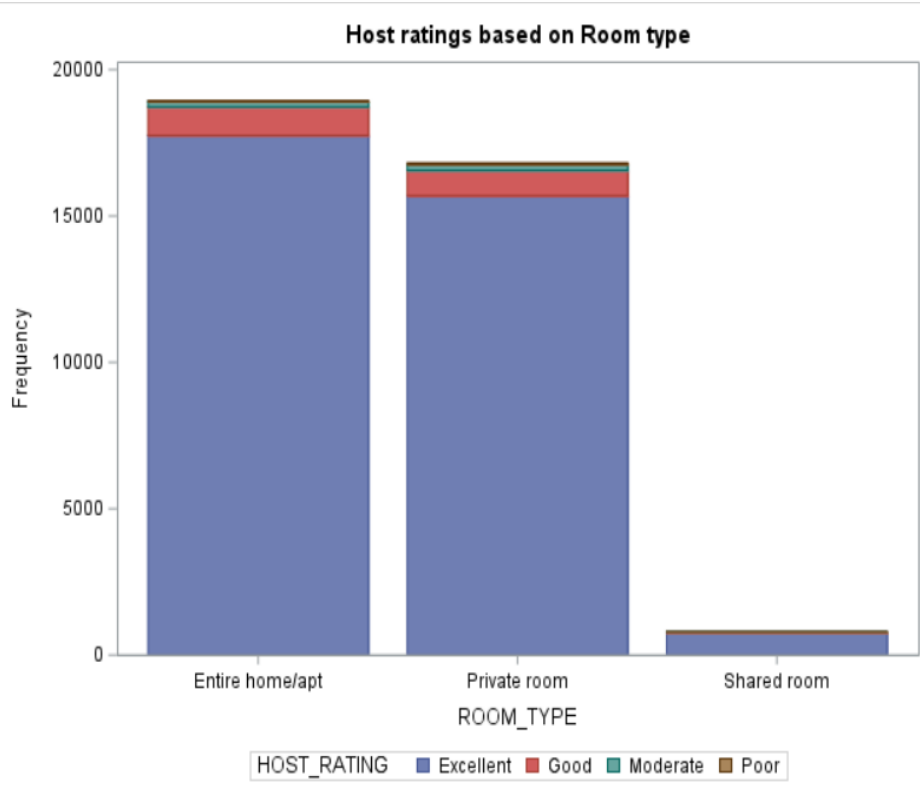
QUESTION 8. IS THERE A RELATIONSHIP BETWEEN HOST RATING AND ROOM TYPE?

CREATING A NEW COLUMN AS HOST_RATING BASED ON THE COLUMN REVIEW_SCORES_RATING TO MAKE 4 LEVELS OF DATA FOR EASY ANALYZATION, CHANGE THE DATATYPE OF THE NEW COLUMN TO CHARACTER;

ROOM_TYPE	PRICE	MINIMUM_NIGHTS	REVIEW_SCORES_RATING	NUMBER_OF_REVIEWS	LAST_REVIEW	LISTINGS_COUNT	HOST_RATING
Entire home/apt	75	45	10	1	03/17/2018	1	Excellent
Entire home/apt	182	9	8	27	05/21/2019	1	Good
Shared room	79	1	10	15	09/29/2018	6	Excellent
Private room	149	1	10	24	05/11/2019	6	Excellent
Private room	79	1	10	17	06/26/2019	6	Excellent
Private room	149	1	9	9	10/19/2018	6	Excellent
Private room	99	1	9	21	10/27/2018	6	Excellent
Private room	58	1	8	19	06/08/2019	6	Good
Entire home/apt	225	1	9	45	05/21/2019	2	Excellent
Shared room	60	1	10	1	07/18/2018	2	Excellent

```
data PROJECT.AirBnB2;
set PROJECT.AirBnB1;
length HOST_RATING $20;
if REVIEW_SCORES_RATING <=3 then
do;
HOST_RATING = "Poor";
end;
else if 4<=REVIEW_SCORES_RATING <=6 then
do;
HOST_RATING = "Moderate";
end;
else if 7<=REVIEW_SCORES_RATING<=8 then
do;
HOST_RATING = "Good";
end;
else
do;
HOST_RATING = "Excellent";
end;
proc print data= PROJECT.AirBnB2(OBS=25);
run;
```

```
Title "Host ratings based on Room type";
proc sgplot data =PROJECT.AirBnBrating;
vbar room_type/ group = host_rating
groupdisplay = stack;
run;
```



Summarization

Frequency
Percent
Row Pct
Col Pct

Table of ROOM_TYPE by HOST_RATING

ROOM_TYPE	HOST_RATING				
	Excellent	Good	Moderate	Poor	Total
Entire home/apt	17749 48.55 93.69 51.92	978 2.68 5.16 51.88	182 0.50 0.96 46.79	36 0.10 0.19 36.73	18945 51.82
Private room	15691 42.92 93.33 45.90	866 2.37 5.15 45.94	199 0.54 1.18 51.16	57 0.16 0.34 58.16	16813 45.99
Shared room	747 2.04 93.26 2.19	41 0.11 5.12 2.18	8 0.02 1.00 2.06	5 0.01 0.62 5.10	801 2.19
Total	34187 93.51	1885 5.16	389 1.06	98 0.27	36559 100.00

Test of Independency

Statistic	DF	Value	Prob
Chi-Square	6	15.5801	0.0162
Likelihood Ratio Chi-Square	6	14.7670	0.0221
Mantel-Haenszel Chi-Square	1	6.6819	0.0097
Phi Coefficient		0.0206	
Contingency Coefficient		0.0206	
Cramer's V		0.0146	

p value is 0.01 which is <0.05 As p value is less than 5% ,we reject null hypothesis and can say that there is statistical association between host rating and room type

Visualization

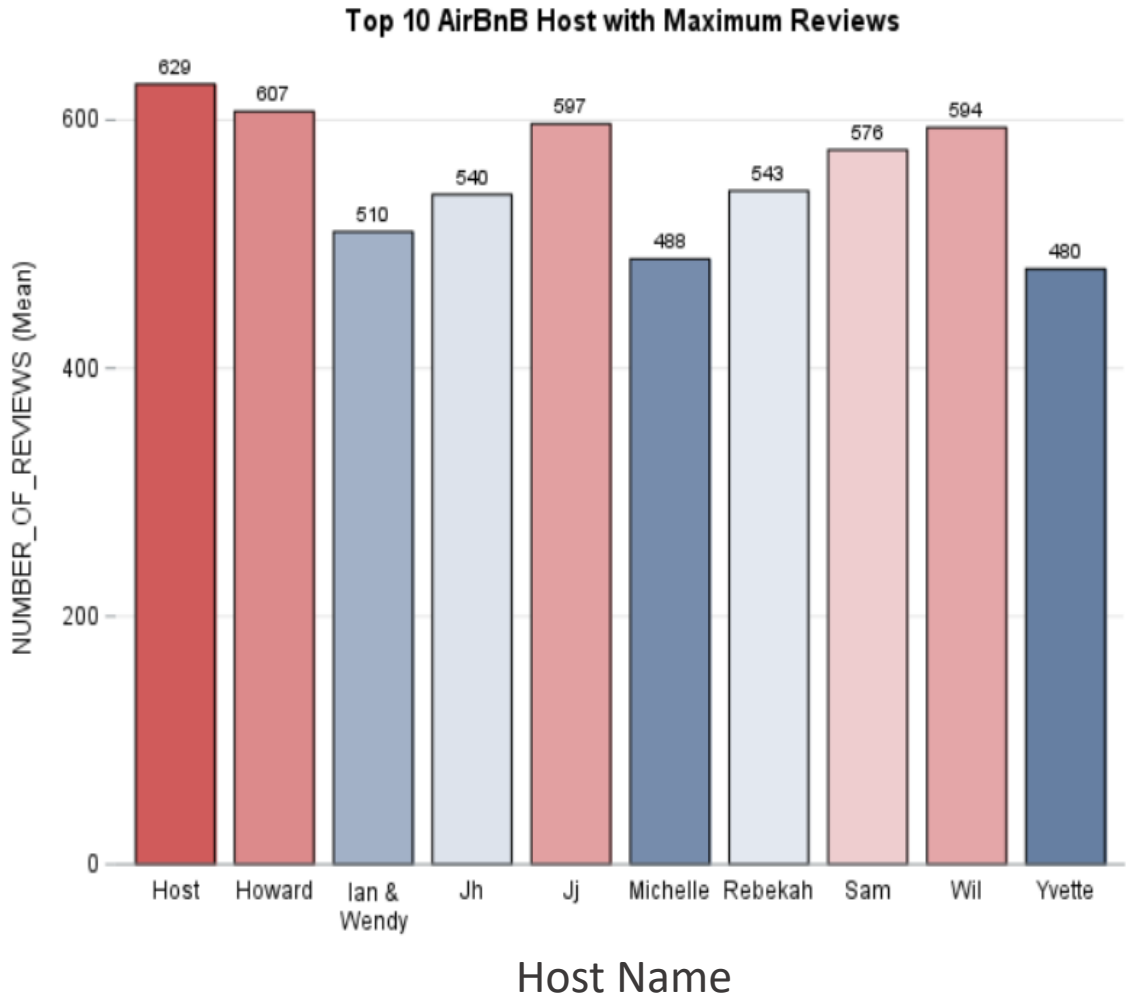
This graph shows that AirBnB is more popular among customers who prefer entire house/apt and private rooms. Also we can see that most of the reviews were either excellent or good irrespective of the category

Question 9. TOP 10 SUPER HOSTS BASED ON THE NUMBER OF REVIEWS ?

```
proc sort data = PROJECT.AirBnBTop10 out=PROJECT.AirBnBTop10out;
by  Number_of_reviews;
run;
proc print data= PROJECT.AirBnBTop10out (obs=37788
firstobs=37777);
run;
Title "Top 10 AirBnB Host with Maximum Reviews";
data project.airbnbtopoutt;
set PROJECT.AirBnBTop10out;
where Number_of_reviews >= 480;
run;
proc print data= PROJECT.AirBnBTopoutt;
run;
```

Top 10 AirBnB Host with Maximum Reviews

Obs	HOST_NAME	NUMBER_OF_REVIEWS
1	Host	629
2	Howard	607
3	Jj	597
4	Wil	594
5	Sam	576
6	Rebekah	543
7	Jh	540
8	Ian & Wendy	510
9	Michelle	488
10	Yvette	480



QUESTION 10. HAS THE POPULARITY OF AIRBNB AFFECTED THE PRICE OF ACCOMODATION OVER THE YEARS?

```
title  "Variation in AirBnB price over the Years";
proc sgplot data=PROJECT.AirBnBT;
    scatter x = last_review y = price/colorresponse=price;
xaxis label= "Year( 2015 -2019 )";
yaxis label ="Price";
run;
```

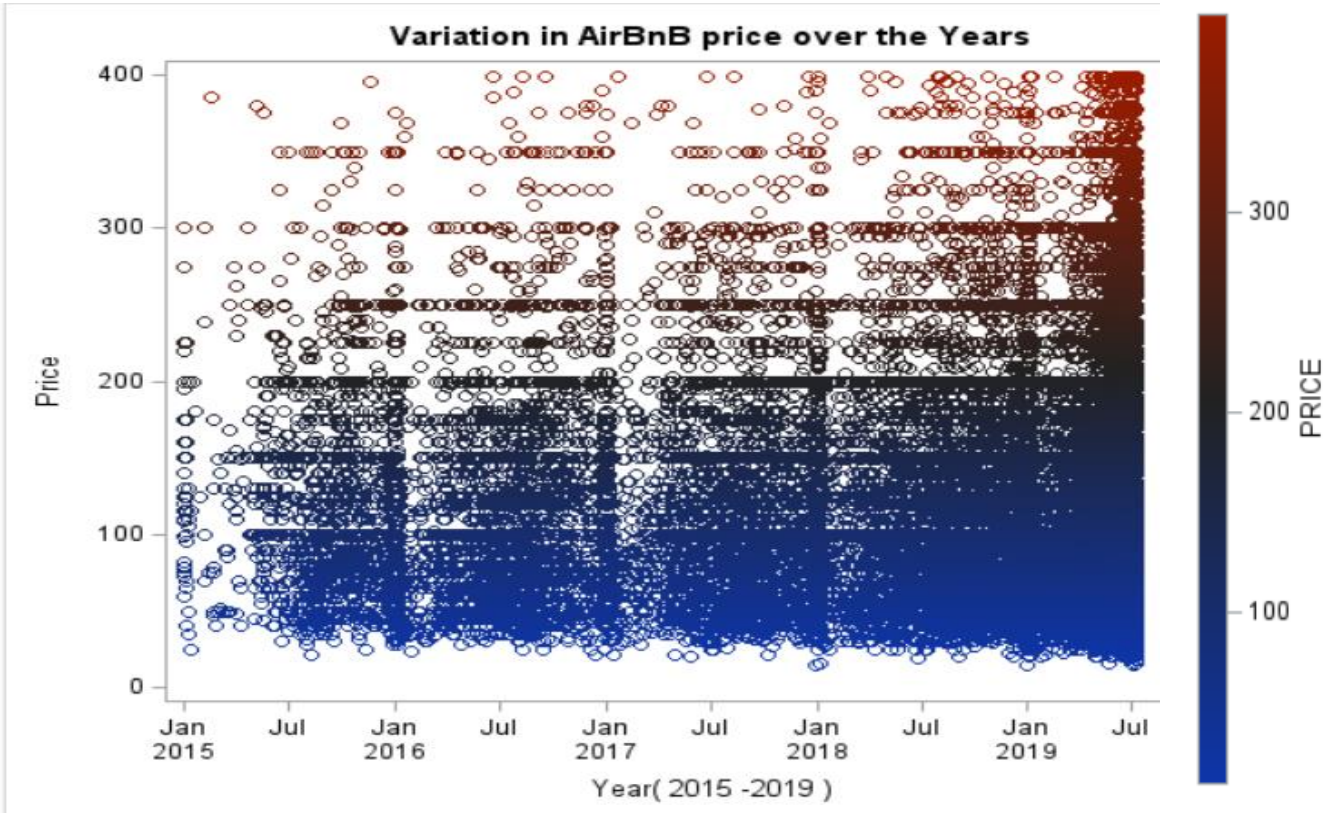
Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
LAST_REVIEW	36559	21471	390.27885	784968450	20089	21738
PRICE	36559	122.54279	73.23570	4480042	15.00000	399.00000

Pearson Correlation Coefficients, N = 36559
Prob > |r| under H0: Rho=0

	LAST_REVIEW	PRICE
LAST_REVIEW	1.00000	0.00968 0.0643
PRICE	0.00968 0.0643	1.00000

From the above table we can see that Correlation coefficient is 0.0096. and the p value is 0.065 >5%. Therefore, we fail to reject null hypothesis and conclude that there is no statistical relation between the variables.



Visualization

This graph shows that over the years number of listings have increased. Prices of listings are over a wider range than before. Which means people have more options to choose now compared to previous years , be it low priced or high priced based on their needs

CONCLUSION AND RECOMMENDATION

Based on the data analysis and visualizations, the following conclusions can be made,

- Manhattan region had the highest distribution of listings followed by Brooklin.
- Majority of the customers prefer to take the entire home or apartment, followed by a private room.
- There is statistical association between room type and the AirBnB location, and there is a statistical relationship between price and location.
- Price per night range is more in Manhattan and cheapest is in Staten island.
- There is an increase in the amount of people preferring to stay in Airbnb's and it has become more popular over the years.
- October through January is usually the busiest month for Airbnb booking, and February to April is the least busiest.
- Most of the reviews were either excellent or good irrespective of the category of room type.

As a recommendation, promotions or deals can be offered during the least busiest months from February to April to improve the bookings.