جامعة الملك سعود
King Saud University

College of Business Administration
Management Information Systems

King Saud University

College of Business Administration

First Semester

2021-2022

MIS419 Group Project

Instructor: Dr. Nora Alkhamees

	Student Name	ID	SN
1	Atheer Alhamidi		27
2	Mzoon Alnghamush		19
3	Alanoud Aldawsari		25
4	Raniah Alsehaibany		6
5	Sumayah Alwarthan		26
6	Latifa Alzughaibi		24

Section: 58713

Introduction:

The dataset is about a hotel booking demand, it comprises booking information for a city hotel and a resort hotel and it includes information such as when the booking was made, the length of the stay, the number of the adults, children and or babies and the number of available parking information. Our objective from this dataset is that we want to give an insight for the hotel reservation system to simplify for the hotel the customers reservations through using classification algorithm, and how can they treat their customers based on their purpose, such as if they have reserved from the booking agency they are coming for a vacation so the services that will be provided for them will be based on their purpose. And if they have reserved from the website then another indicator will be considered to give the customers the appropriate service.

Data exploration and Description: Alanoud

-The attributes before the preprocessing:

The number of attributes in the dataset before the preprocessing is 32 and 119390 row attributes is:hotel,is_canceled,lead_time,arrival_date_year,arrival_date_month,arrival_date_week_number,arrival_date_day_of_month,stays_in_weekend_nights,stays_in_week_nights,adults,children,babies,meal,country,market_segment,distribution_channel,is_repeated_guest,previous_cancellations,previous_bookings_not_canceled,reserved_room_type,assigned_room_type,booking_changes,deposit_type,agent,company,days_in_waiting_list,customer_type,adr,required_car_parking_spaces,total_of_special_requests,reservation_status,reservation_status_date

-Metadata:

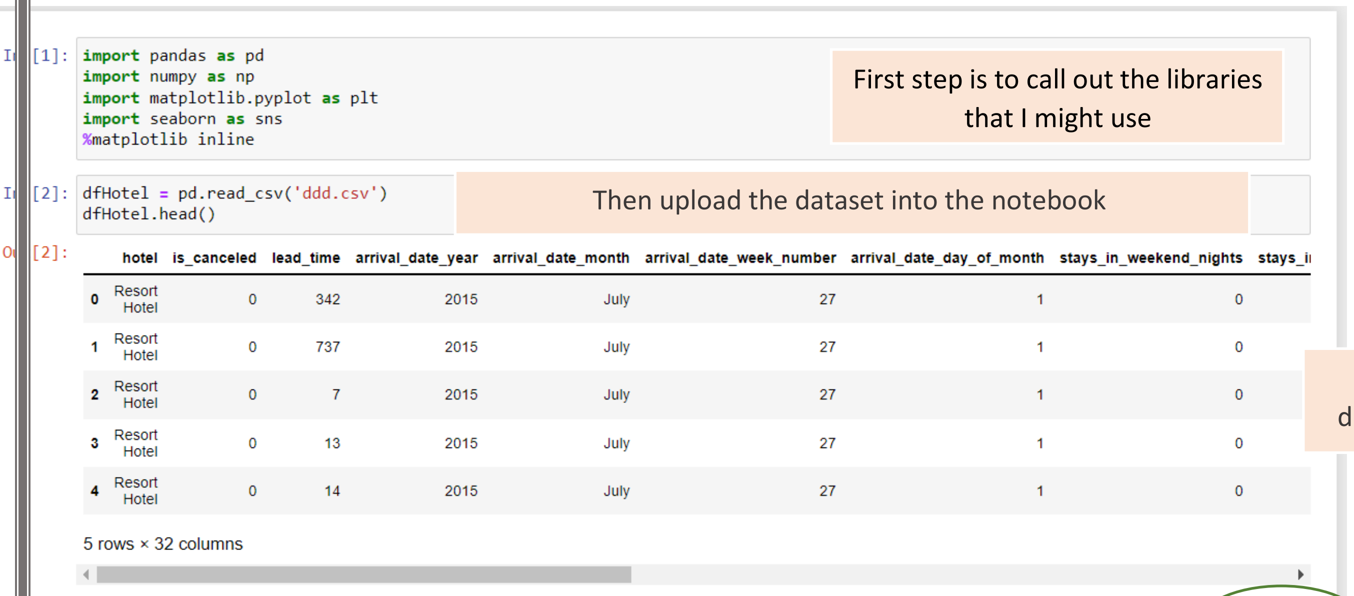
attribute	Type	Description
Adults	Integer	Number of adults in the reserved room
Children	Integer	Number of children in the reserved room
Babies	Integer	Number of babies in the reserved room
Agent	Categorical	ID of the online travel intermediary that helped with the reservation
Company	Categorical	ID of the company organization that responsible for handling the expenses the hotel
ArrivalDateDayOfMonth	Integer	The Day of the month of the arrival to the hotel
ArrivalDateMonth	Categorical	The Month of arrival to the hotel with 12 categories :January to December
ArrivalDateYear	Integer	The Year of arrival to the hotel

IsRepeatedGuest	Binary	The value indicates if the booking was from a repeated/regular guest (1) is yes and (0) is no
DaysInWaitingList	Integer	The Number of days the reservation was in the waiting list before it was approved to the visitor
Country	Categorical	The Country Categories are represented in the ISO 3155–3:2013 format
DistributionChannel	Categorical	The reservation distribution channel which is an offline company the categorizes is: -TA\TO : the TA means Travel Agents and TO means Tour Operators which mean Corporate,direct,GDS,undefiend
ReservationStatus	Categorical	The last Reservation status which includes of three categories: Canceled – reservation was canceled by the visitor Check-Out – the visitor has checked from the hotel roc No-Show – the visitor has not showed up to the reserv
StaysInWeekendNights	Integer	The Number of weekend nights the visitor stayed or reserved to stay at the hotel
StaysInWeekNights	Integer	The Number of weeknights the visitor stayed or reserved to stay at the hotel
IsCanceled	Binary	The Values indicates if the reservation was canceled (1) for yes and (0) in no
DepositType	Categorical	Indication on if the visitor made a deposit for the reservation the categories are: No Deposit – there is no deposit Non Refund – a deposit was made as the total cost of the reservation Refundable – a deposit was made under the total cost of the reservation
hotel	binary	This indicates the type of hotel and the values of it: City hotel, Resort hotel

This metadata includes the most relevant to the preprocessing or viewed as important attribute for more information about the other attributes click here: https://www.sciencedirect.com/science/article/pii/S2352340918315191

-Attribute values:

What helped me in finding out the values for the attribute is using Jupyter Notebook which is denote documents that contain both code and rich text elements, such as figures, links, equations



```
In [1]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns
        %matplotlib inline
```

First step is to call out the libraries that I might use

```
In [2]: dfHotel = pd.read_csv('ddd.csv')
        dfHotel.head()
```

Then upload the dataset into the notebook

Out[2]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_i
0	Resort Hotel	0	342	2015	July	27	1	0	
1	Resort Hotel	0	737	2015	July	27	1	0	
2	Resort Hotel	0	7	2015	July	27	1	0	
3	Resort Hotel	0	13	2015	July	27	1	0	
4	Resort Hotel	0	14	2015	July	27	1	0	

5 rows × 32 columns

The dataset

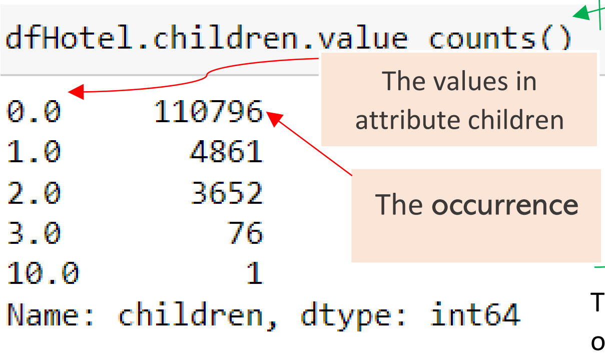
code

Attribute children:

by doing this code "value counts"

this will get the attribute values and the number of the occurrence of that attribute value.

The 0.0 indicates that there are no children in the reservation and the rest of numbers indicates that there is a child in the reservation.



```
dfHotel.children.value_counts()

0.0     110796
1.0       4861
2.0       3652
3.0         76
10.0         1
Name: children, dtype: int64
```

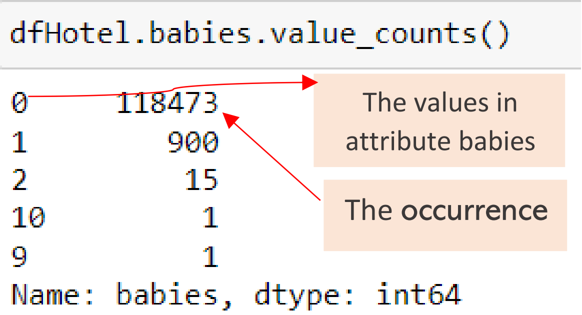
The values in attribute children

The occurrence

The outcome of the code

Attribute babies:

The 0 indicates that there are no babies in the reservation and the rest of numbers indicates that there is a baby in the reservation.



```
dfHotel.babies.value_counts()

0     118473
1        900
2         15
10         1
9          1
Name: babies, dtype: int64
```

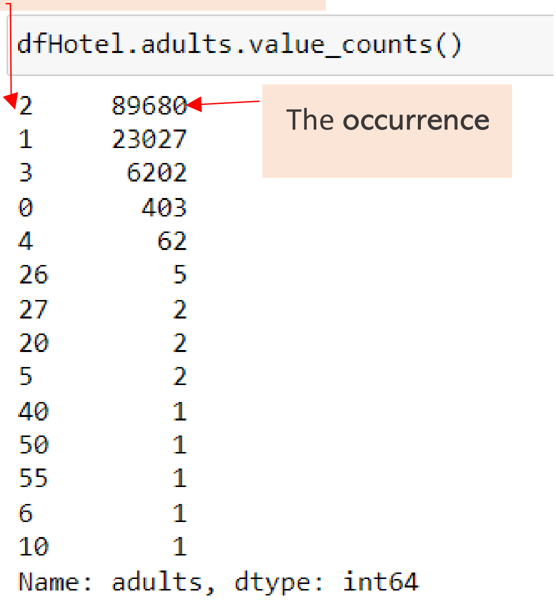
The values in attribute babies

The occurrence

Attribute adult

this indicates the number of adults in a reservation for a room the number 0 indicates that there is no adult which mean no reservation, or the room is for children which we will provide more information after the preprocessing take a place.

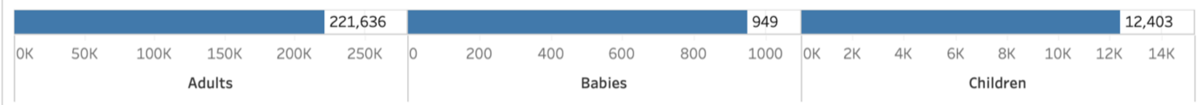
The values in attribute adult



```
dfHotel.adults.value_counts()
```

```
2       89680
1       23027
3        6202
0         403
4          62
26          5
27          2
20          2
5           2
40          1
50          1
55          1
6           1
10          1
Name: adults, dtype: int64
```

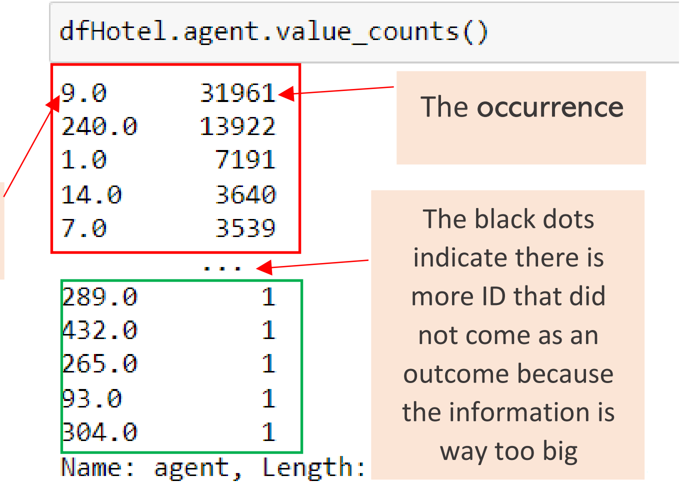
The occurrence



Attribute agent;

the attribute agent has IDs as its values there is some unique IDs like the ones in green box and some repeated IDs like the IDs in a red box

The values in attribute agent

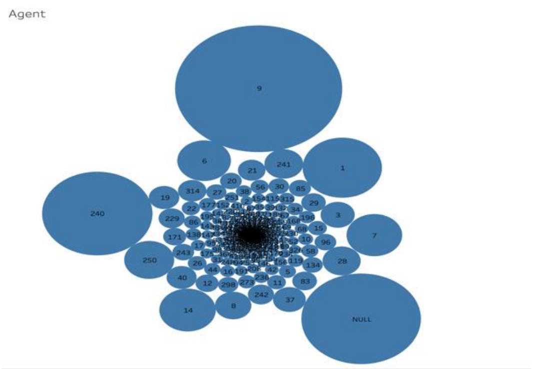


```
dfHotel.agent.value_counts()
```

```
9.0       31961
240.0     13922
1.0        7191
14.0       3640
7.0        3539
           ...
289.0         1
432.0         1
265.0         1
93.0          1
304.0         1
Name: agent, Length:
```

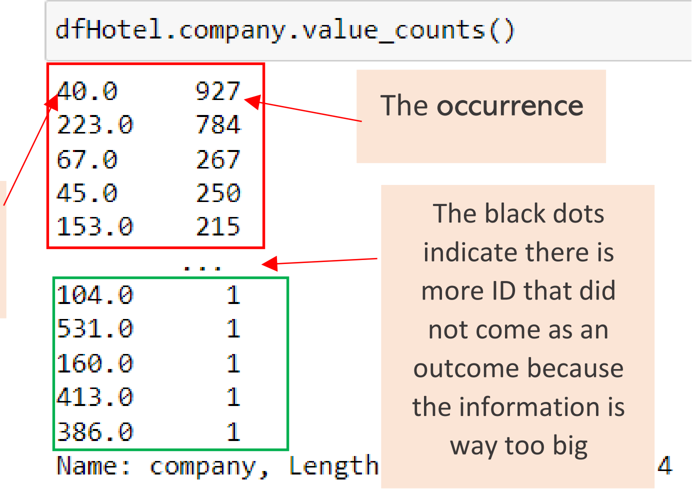
The occurrence

The black dots indicate there is more ID that did not come as an outcome because the information is way too big



Attribute company

the attribute company has IDs as its values there is some unique IDs like the ones in green box and some repeated IDs like the IDs in a red box

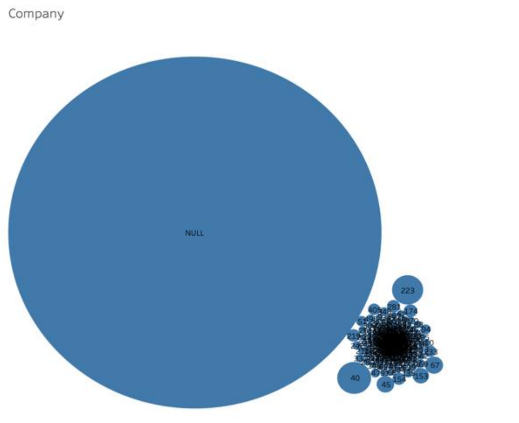


```
dfHotel.company.value_counts()
40.0      927
223.0     784
67.0      267
45.0      250
153.0     215
          ...
104.0       1
531.0       1
160.0       1
413.0       1
386.0       1
Name: company, Length                    4
```

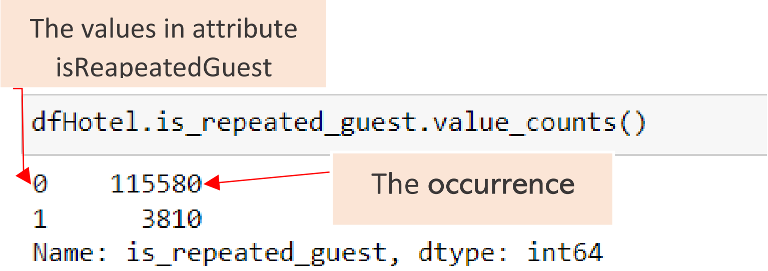
The values in attribute company

The occurrence

The black dots indicate there is more ID that did not come as an outcome because the information is way too big



Attribute IsRepeatedGuest



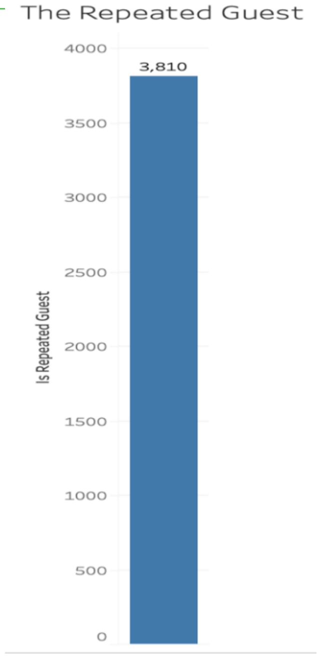
The values in attribute isReapeatedGuest

```
dfHotel.is_repeated_guest.value_counts()
0     115580
1       3810
Name: is_repeated_guest, dtype: int64
```

The occurrence

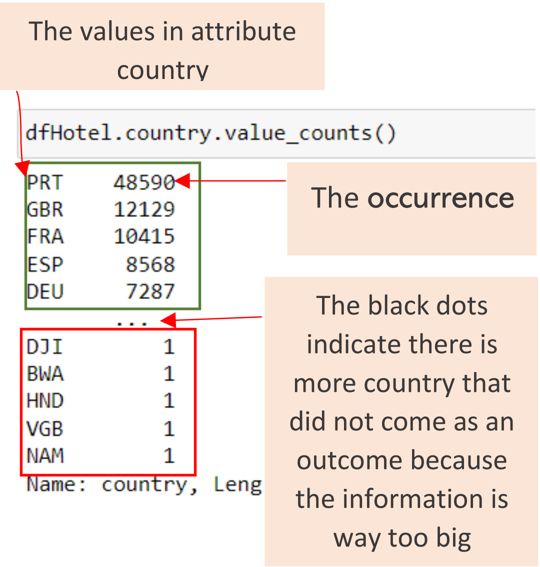
The zero indicate that the reservation was not by a repaeated guest and the one indicate that it is a reapeated guest this information could help in sending offers to the regluar guest and try to attrac and reatain the new gusets.

The percentage of the repatead guest 96.81% and 3.19% for the not repatead guest



Attribute country:

this attribute includes the country name, there is some countries that is repeated frequently like in the read box and some unique countries like in green box



The values in attribute country

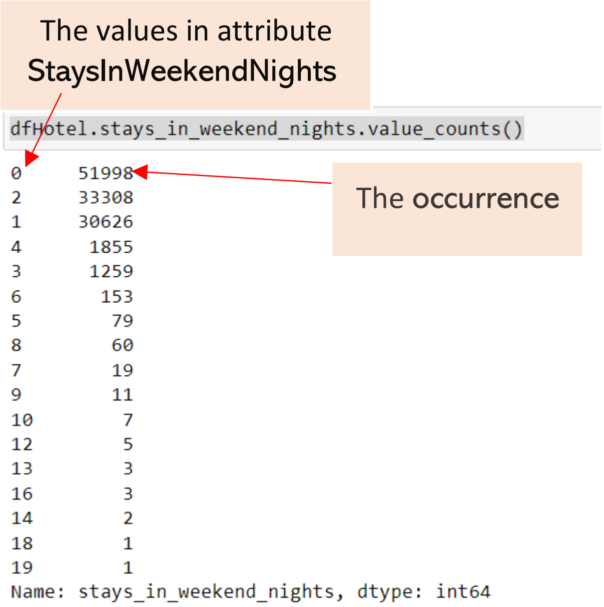
```
dfHotel.country.value_counts()

PRT     48590
GBR     12129
FRA     10415
ESP      8568
DEU      7287
         ...
DJI         1
BWA         1
HND         1
VGB         1
NAM         1
Name: country, Leng
```

The occurrence

The black dots indicate there is more country that did not come as an outcome because the information is way too big

Attribute StaysInWeekendNights:

The attribute stays_in_weekend_nights have the number of weekend nights the visitor stayed or reserved to stay at the hotel as its attribute values the percentage of visitors who stayed in weekends are 56.45% ano who did not 43.55%



The values in attribute StaysInWeekendNights

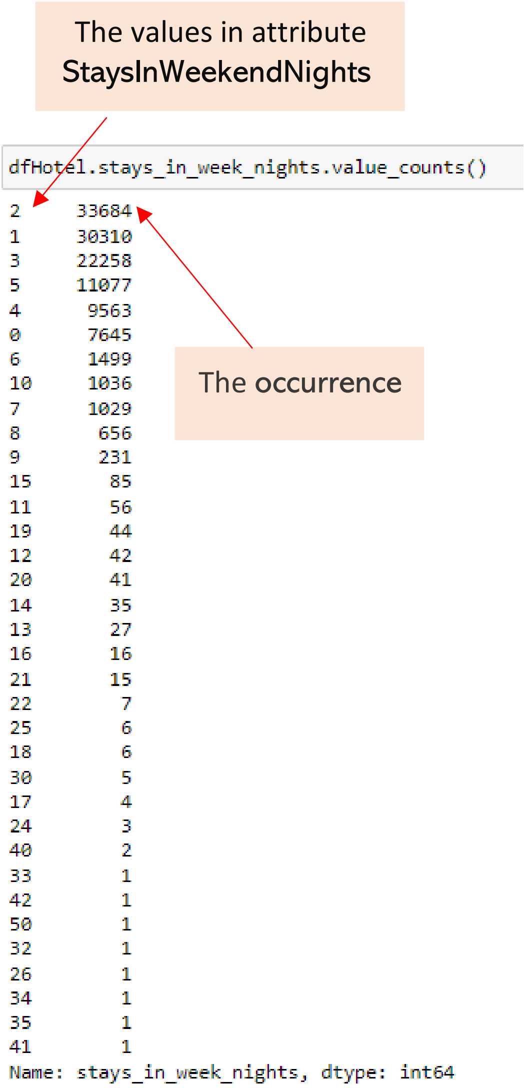
```
dfHotel.stays_in_weekend_nights.value_counts()

0     51998
2     33308
1     30626
4      1855
3      1259
6       153
5        79
8        60
7        19
9        11
10        7
12        5
13        3
16        3
14        2
18        1
19        1
Name: stays_in_weekend_nights, dtype: int64
```

The occurrence

Attribute StaysInWeekNights:

The attribute stays_in_week_nights have the number of weeknights the visitor stayed or reserved to stay at the hotel as its attribute values the percentage of visitors who

stayed for weeknights are 93.6% and who did not 6.4



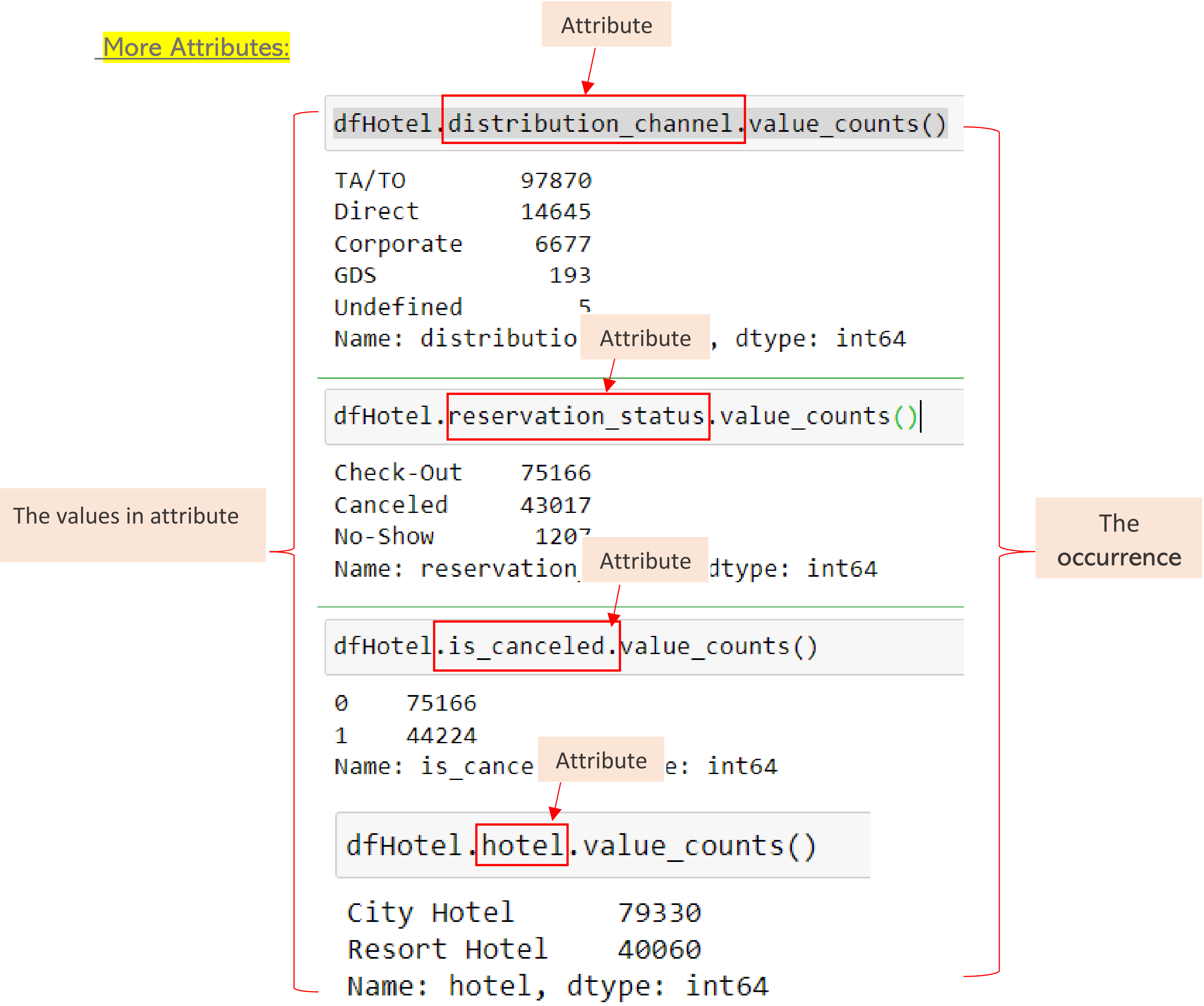
The values in attribute StaysInWeekendNights

The occurrence

```
dfHotel.stays_in_week_nights.value_counts()
```

```
2     33684
1     30310
3     22258
5     11077
4      9563
0      7645
6      1499
10     1036
7      1029
8       656
9       231
15       85
11       56
19       44
12       42
20       41
14       35
13       27
16       16
21       15
22        7
25        6
18        6
30        5
17        4
24        3
40        2
33        1
42        1
50        1
32        1
26        1
34        1
35        1
41        1
Name: stays_in_week_nights, dtype: int64
```

More Attributes:



```
dfHotel.distribution_channel.value_counts()
```

```
TA/TO        97870
Direct       14645
Corporate     6677
GDS            193
Undefined        5
Name: distributio , dtype: int64
```

```
dfHotel.reservation_status.value_counts()
```

```
Check-Out    75166
Canceled     43017
No-Show       1207
Name: reservation dtype: int64
```

```
dfHotel.is_canceled.value_counts()
```

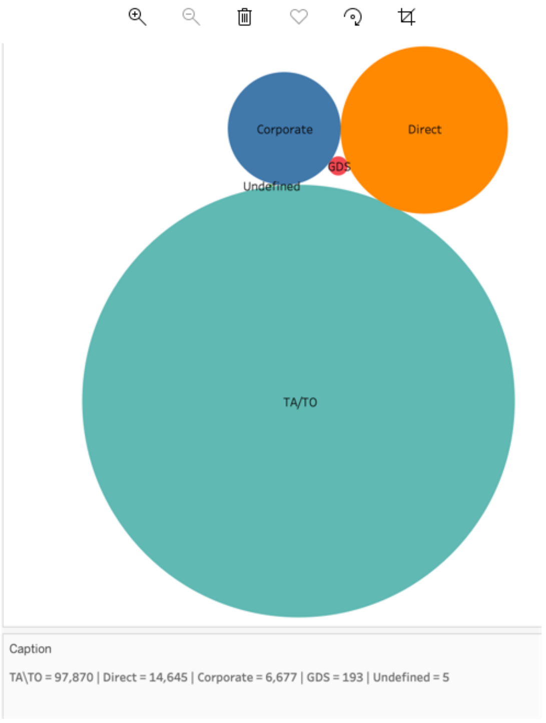
```
0    75166
1    44224
Name: is_cance e: int64
```

```
dfHotel.hotel.value_counts()
```

```
City Hotel      79330
Resort Hotel    40060
Name: hotel, dtype: int64
```

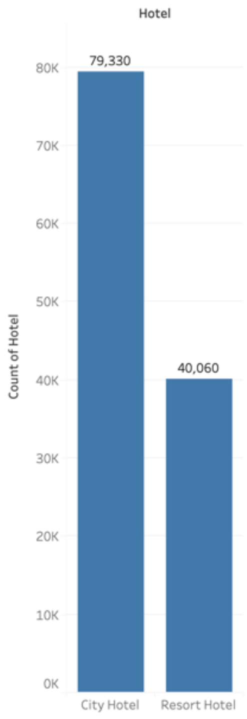
-for hotel attribute the city hotel is more than the resort hotel by 39270 differences ,the percentage of the city hotel is 66.45% and the resort hotel 33.55%

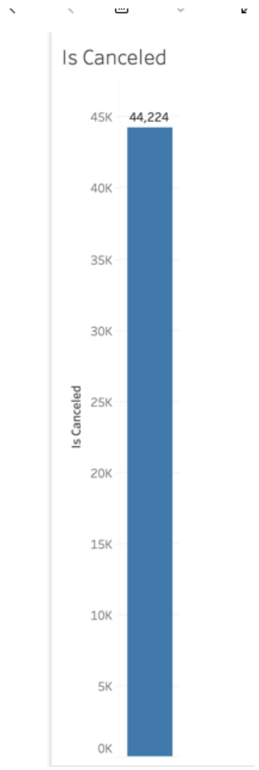
- for is canceled attribute the visitors seem to not cancel the reservation more than canceling it, the percentage for the guests who canceled the reservation 37.04% and who didn't cancel 62.96%



Caption

TA\TO = 97,870 | Direct = 14,645 | Corporate = 6,677 | GDS = 193 | Undefined = 5





-The attributes after the preprocessing:

Now the dataset includes 28 attribute and 119210 row and the attributes is(hotel,is_canceled,lead_time,arrival_date_week_number,stays_in_weekend_nights,stays_in_week_nights,meal,country,market_segment,distribution_channel,is_repeated_guest,previous_cancellations,previous_bookings_not_canceled,reserved_room_type,assigned_room_type,booking_changes,deposit_type,agent,purpous,days_in_waiting_list,customer_type,adr,required_car_parking_spaces,total_of_special_requests,reservation_status,reservation_status_date,date_of_arrival)

The reasons behind the preprocessing:

The look of the data was scattered because of a lot of dimensionalities and there is a lot of missing data in the attributes, which led to thinking of solutions to make it ready for data mining:

1-The column of dates (ArrivalDateDayOfMonth, ArrivalDateYear, ArrivalDateMonth) the best solution for this attributes is to collect them into single attribute which is date_of_arrival because the attributes being separate will not imply more information so it's the same (the type of the data preprocessing: Aggregation).

2-The country attribute have missing values and the number of the missing values is 407 out 119390 not missing values, which led to put the most repeated country due to the small amount of missing data in this attribute (handling missing values with Mode)

3-The Attributes of children ,adults ,babies was separated so it is better to put them together to make the data more useful so we created the attribute status which have adults ,family, groups , couples ,reservation without adult as its attribute value and also there is no reservation value which we got rid of because it's indicates there is no adult or children or babies which it doesn't make sense since the other attribute have data of the reservation so it might be just a missing value and since it's only 180 out of 119390 we chose to delete it (the type of data prepossessing: Aggregation, Feature creation and handling missing data by eliminate data object).

4-The Agent attribute indicates that it’s the ID of the online travel intermediary that helped with the booking So It was IDs and null values as its attribute values and the number of IDs was 103050 and the null values was 16340, so it's a lot of missing values which we need to deal with so the solution is to say to whom there are IDs are the visitors who used Booking website as a way to reserve a room and for the null values are the visitors who used the hotel website as a way to reserve a room (the type of data prepossessing: Feature Creation).

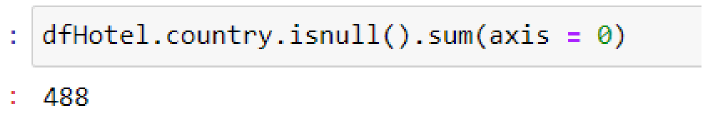
5-The company attribute indicating IDs of the company that are responsible for handling the expenses of the hotel so it's containing IDs and null values as its attribute values, the number of IDs is 6797 and the number of null values is 112593, it's a lot of missing values that we need to deal with, the solution is the visitors who have IDs are on a business trip because the company paid for them for any reason either for checking other branches or a conference in that country. For those who have no IDs then they came for entertainment and vacation (the type of data prepossessing: Feature Creation).

6-is_canceled and is_repeated_guest we changed their attribute values from 0,1 to yes and no to make it more useful (the type of data prepossessing: Feature Creation).

The prepressing steps:

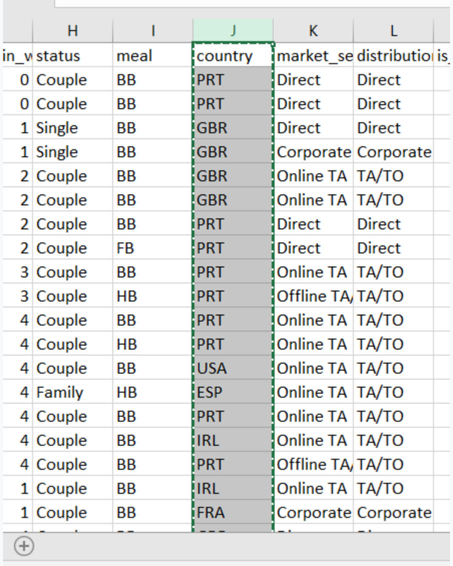
country attribute:

1-first I knew that there is missing values because I did this code to every column, and this is the output when I tried it on the column country

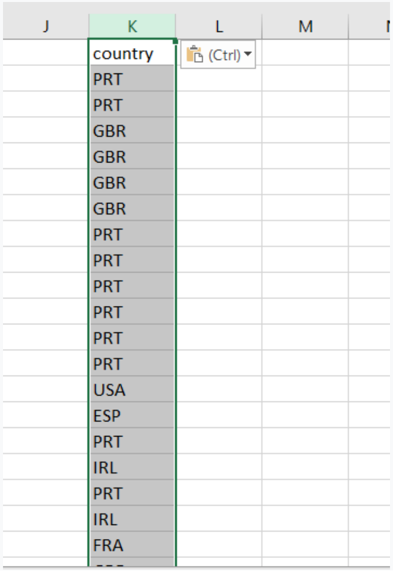


```
: dfHotel.country.isnull().sum(axis = 0)

: 488
```

So there is 488 Null value out of 119390 row which is small amount so I can use the most frequent country to solve the missing value, so I copied the column and took it to empty excel sheet, so I find the data and replace more accurately without the danger of replacing any other values in other columns

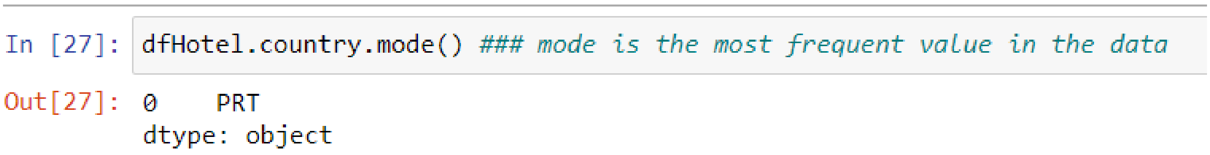


	H	I	J	K	L	
in_w	status	meal	country	market_se	distributio	is_
0	Couple	BB	PRT	Direct	Direct	
0	Couple	BB	PRT	Direct	Direct	
1	Single	BB	GBR	Direct	Direct	
1	Single	BB	GBR	Corporate	Corporate	
2	Couple	BB	GBR	Online TA	TA/TO	
2	Couple	BB	GBR	Online TA	TA/TO	
2	Couple	BB	PRT	Direct	Direct	
2	Couple	FB	PRT	Direct	Direct	
3	Couple	BB	PRT	Online TA	TA/TO	
3	Couple	HB	PRT	Offline TA,	TA/TO	
4	Couple	BB	PRT	Online TA	TA/TO	
4	Couple	HB	PRT	Online TA	TA/TO	
4	Couple	BB	USA	Online TA	TA/TO	
4	Family	HB	ESP	Online TA	TA/TO	
4	Couple	BB	PRT	Online TA	TA/TO	
4	Couple	BB	IRL	Online TA	TA/TO	
4	Couple	BB	PRT	Offline TA,	TA/TO	
1	Couple	BB	IRL	Online TA	TA/TO	
1	Couple	BB	FRA	Corporate	Corporate	



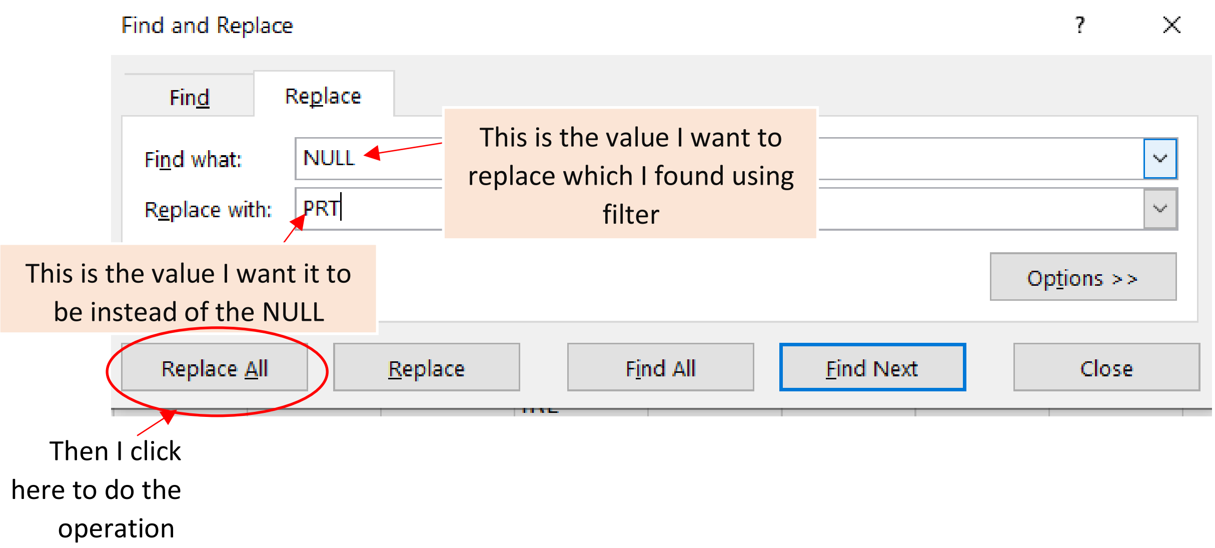
J	K	L	M
	country	(Ctrl)	
	PRT		
	PRT		
	GBR		
	GBR		
	GBR		
	GBR		
	PRT		
	PRT		
	PRT		
	PRT		
	PRT		
	PRT		
	USA		
	ESP		
	PRT		
	IRL		
	PRT		
	IRL		
	FRA		

2- then I need to know the most frequent country so I can use find and replace feature in excel to replace the null value with it, so I did the code below and the output is the most frequent country in the country column.



```
In [27]: dfHotel.country.mode() ### mode is the most frequent value in the data

Out[27]: 0    PRT
         dtype: object
```

Then since now I know the most frequent value now I will simply use find and replace by clicking Cntrl and F



Find what: NULL

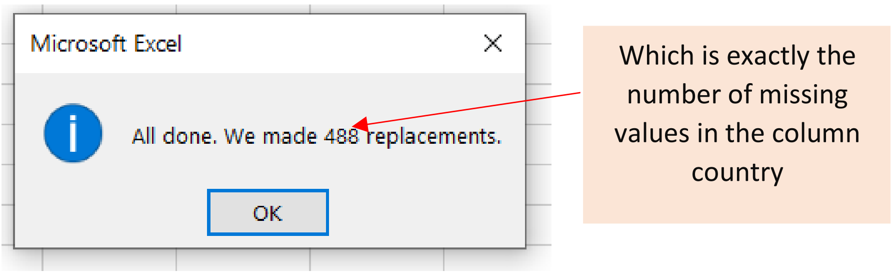
Replace with: PRT

This is the value I want to replace which I found using filter

This is the value I want it to be instead of the NULL

Then I click here to do the operation

3-then be clicking on replace all ,all the null values are now PRT country

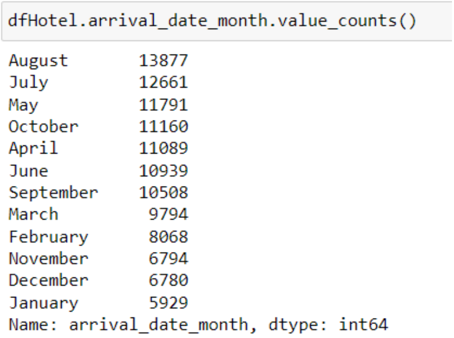


All done. We made 488 replacements.

Which is exactly the number of missing values in the column country

Date of arrival attribute:

1-first to fix this all we need is Date formula, but I have a problem which is the ArrivalDateMonth attribute because it's a categorical data type and include the name of month as its attribute value as u can see here

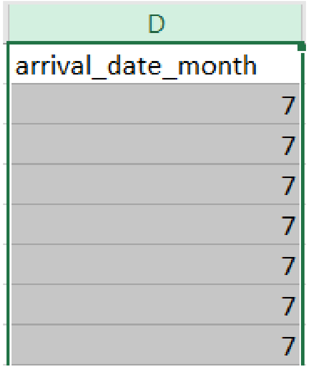


```
dfHotel.arrival_date_month.value_counts()
```

```
August       13877
July         12661
May          11791
October      11160
April        11089
June         10939
September    10508
March         9794
February      8068
November      6794
December      6780
January       5929
Name: arrival_date_month, dtype: int64
```

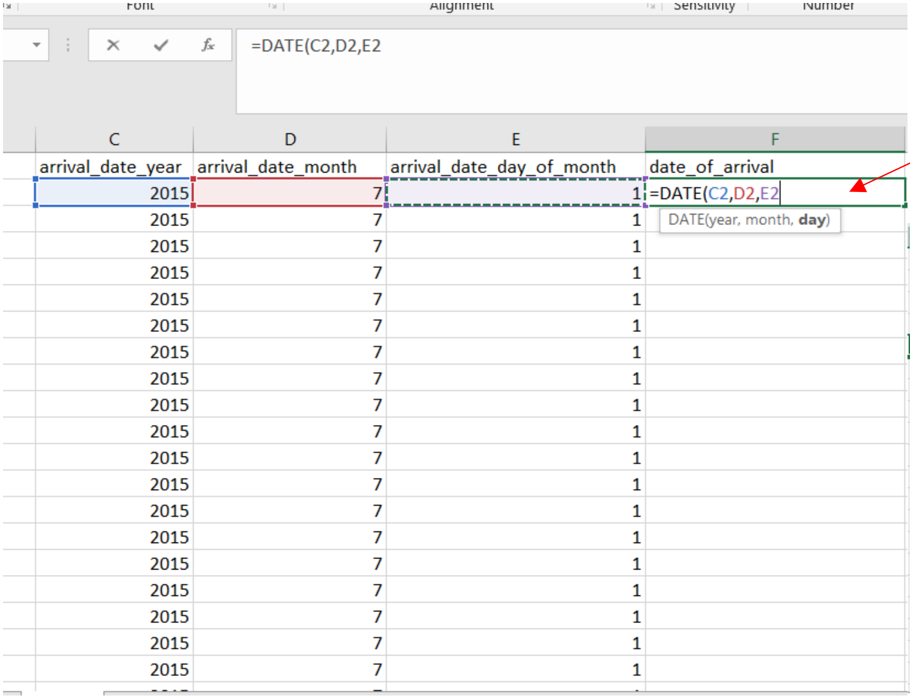
and this doesn't work with the Date formula, so I need to change the data type from categorical to integer like for example if the attribute value is January then it's should be 1

2-so I used the find and replace feature in excel to fix this and this is the outcome the 7 indicate that it's July



D
arrival_date_month
7
7
7
7
7
7
7

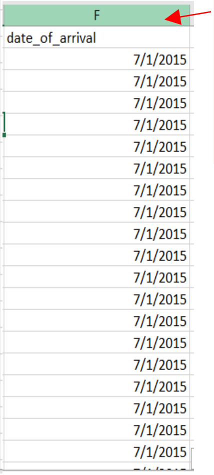
3-then I can now use the date formula without any problems



=DATE(C2,D2,E2

C	D	E	F
arrival_date_year	arrival_date_month	arrival_date_day_of_month	date_of_arrival
2015	7	1	=DATE(C2,D2,E2
2015	7	1	DATE(year, month, **day**)
2015	7	1	
2015	7	1	
2015	7	1	
2015	7	1	
2015	7	1	
2015	7	1	
2015	7	1	
2015	7	1	
2015	7	1	
2015	7	1	
2015	7	1	
2015	7	1	
2015	7	1	
2015	7	1	
2015	7	1	
2015	7	1	
2015	7	1	

The formula

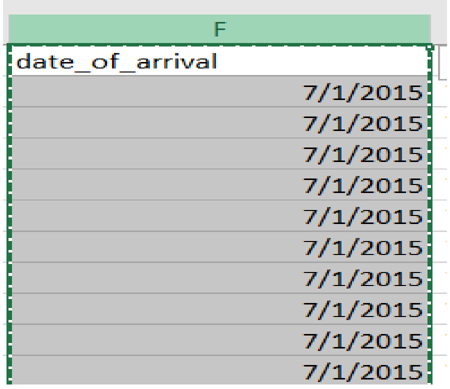


F
date_of_arrival
7/1/2015
7/1/2015
7/1/2015
7/1/2015
7/1/2015
7/1/2015
7/1/2015
7/1/2015
7/1/2015
7/1/2015
7/1/2015
7/1/2015
7/1/2015
7/1/2015
7/1/2015
7/1/2015
7/1/2015
7/1/2015
7/1/2015

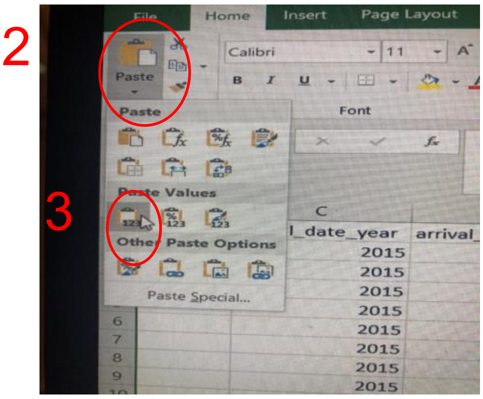
This is when the formula is applied

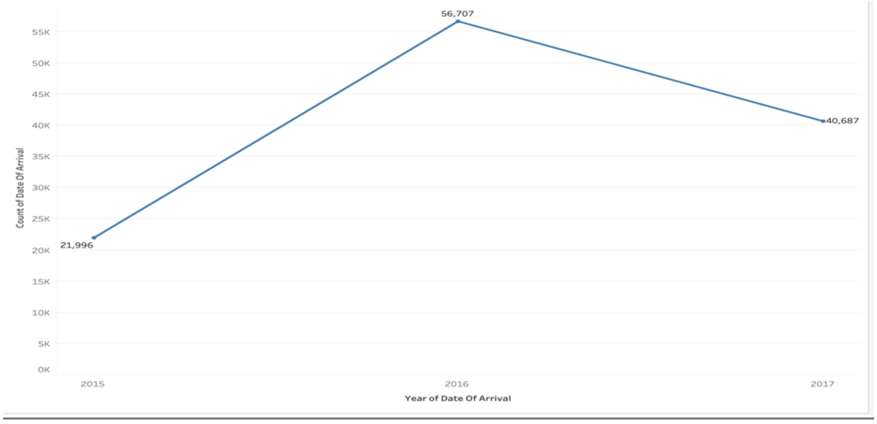
4-then I deleted the attributes (ArrivalDateDayOfMonth, ArrivalDateYear, ArrivalDateMonth) and just kept the DateOfArrival but before that I need to make the column independent from the other column so the formula still works when I delete them this will be done by selecting the column DateOfArrival and click on copy thin click paste

1



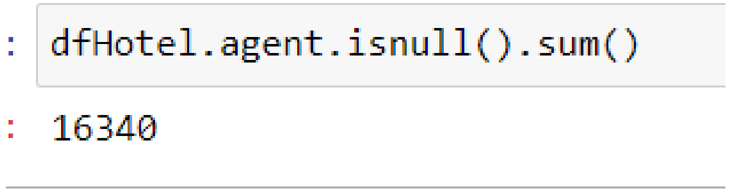
F
date_of_arrival
7/1/2015
7/1/2015
7/1/2015
7/1/2015
7/1/2015
7/1/2015
7/1/2015
7/1/2015
7/1/2015
7/1/2015





Agent attribute:

1-snice this attribute only includes IDs and missing values as its attribute value the first thing that I did is to figure out how many missing values in this column which is 16340 out of 119390 row and the percentage 13.69% which is a lot so it's a must to deal with

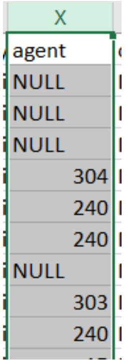


```
: dfHotel.agent.isnull().sum()
```

```
: 16340
```

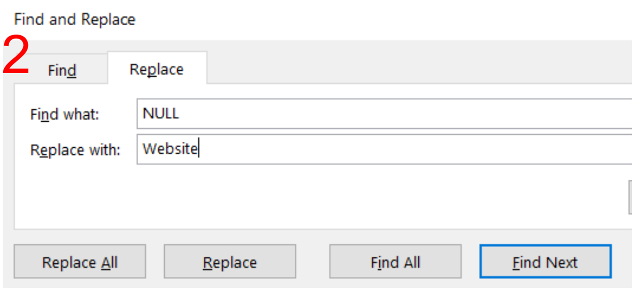
Then I went to use find and replace then I replaced the NULL with website

1



```
X
agent
NULL
NULL
NULL
 304
 240
 240
NULL
 303
 240
```

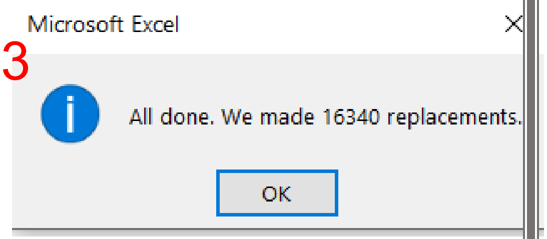
2



Find what: NULL

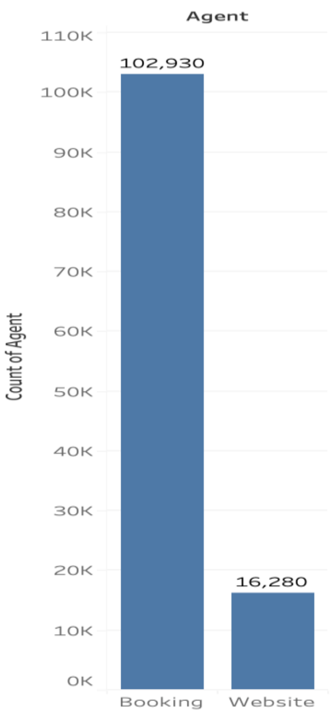
Replace with: Website

3



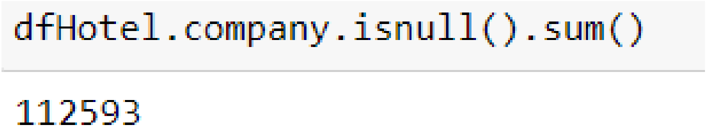
All done. We made 16340 replacements.

2-then I figure out what the IDs that are in the column using filter and then to replace it with Booking using find and replace feature in excel.



Purpose attribute:

1-since the attribute company only includes IDs of the company and missing values as its attribute values the first thing that I did is to figure out how many missing values in this column which is 112593 out of 119390 rows and the percentage 94.33% so it's a lot and we must deal with it.

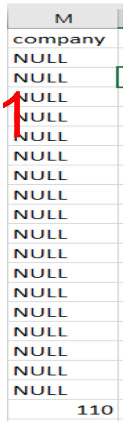


```
dfHotel.company.isnull().sum()
```

```
112593
```

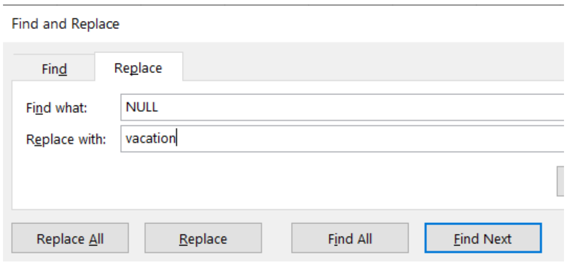
Then I went to use find and replace then I replaced the NULL with vacation.

1



```
M
company
NULL
NULL
NULL
NULL
NULL
NULL
NULL
NULL
NULL
NULL
NULL
NULL
NULL
NULL
NULL
NULL
NULL
NULL
        110
```

2

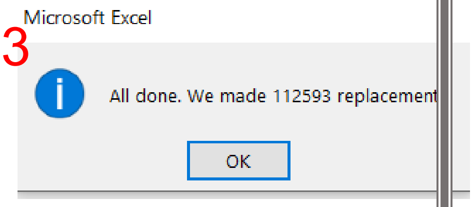


Find and Replace

Find what: NULL

Replace with: vacation

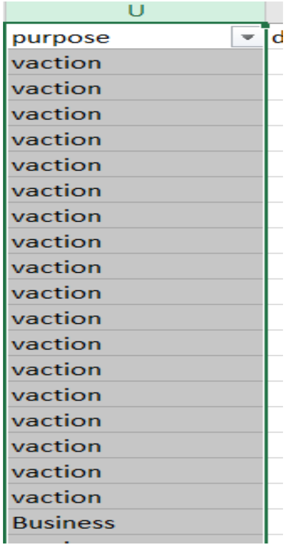
3



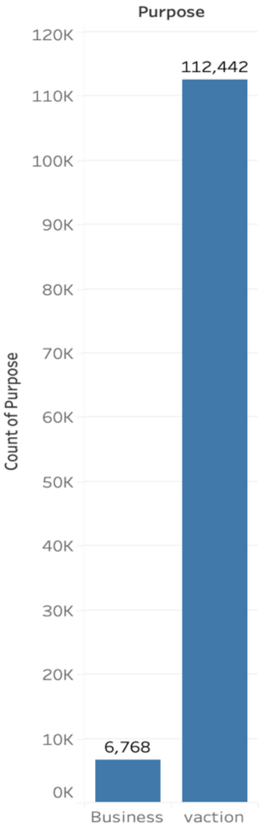
Microsoft Excel

All done. We made 112593 replacement

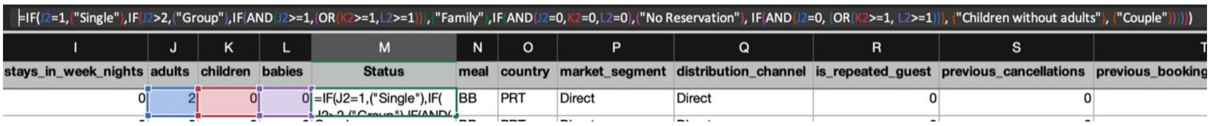
2-then I figure out what the IDs that are in the column using filter and then to replace it with Business using find and replace feature in excel and I change the name of the column from company to purpose to make more sense.



U
purpose
vaction
vaction
vaction
vaction
vaction
vaction
vaction
vaction
vaction
vaction
vaction
vaction
vaction
vaction
vaction
vaction
vaction
vaction
Business



Status Attribute:



```
=IF(J2=1,("Single"),IF(J2>2,("Group"),IF(AND(J2>=1,(OR(K2>=1,L2>=1))),"Family",IF(AND(J2=0,K2=0,L2=0),("No Reservation"), IF(AND(J2=0, (OR(K2>=1, L2>=1))), ("Children without adults"), ("Couple"))))))
```

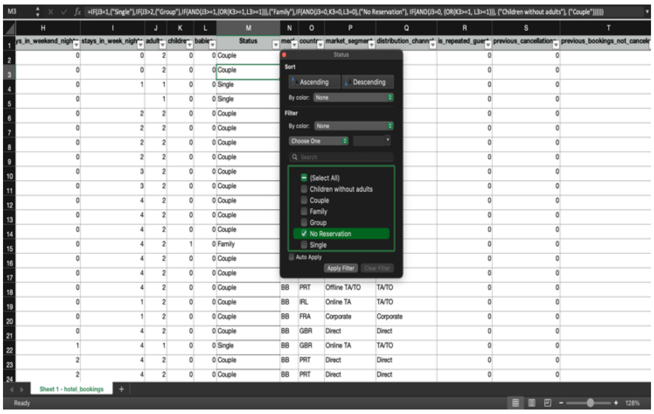
I	J	K	L	M	N	O	P	Q	R	S	T
stays_in_week_nights	adults	children	babies	Status	meal	country	market_segment	distribution_channel	is_repeated_guest	previous_cancellations	previous_booking
0	2	0	0	=IF(J2=1,("Single"),IF(	BB	PRT	Direct	Direct	0	0	

At the beginning of the IF statements, we mentioned that if the column of the adult is equal to 1, the result will be a Single, but if it is greater than 2, it will be a group.

And then we mentioned two conditions in one IF statement, which is if the adult column is greater than or equal to 1 and also if the children or babies column is greater than or equal to 1 then the result will be a family.

And then in the case of all the columns for adult, children, and babies, it was equal to zero, which means that there is no reservation, but after reviewing the rows in which this option appeared, it was found that there is other information such as the time of arrival and so on, and this is illogical, and their total was only 180 rows. So, we decide to delete the rows to not affect the dataset with counterproductive result. How did we delete it?

1- First we choose the Sort and filter option then we choose to just show the rows with No reservation.



M3 =IF(J3=1,("Single"),IF(J3>2,("Group"),IF(AND(J3>=1,(OR(K3>=1,L3>=1))),("Family"),IF(AND(J3=0,K3=0,L3=0),("No Reservation"), IF(AND(J3=0, (OR(K3>=1, L3>=1))), ("Children without adults"), ("Couple")))))

	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	ys_in_weekend_nigh	stays_in_week_nigh	adult	childre	babie	Status	mea	countr	market_segmen	distribution_chann	is_repeated_gues	previous_cancellation	previous_bookings_not_cancele
2	0	0	2	0	0	Couple					0	0	
3	0	0	2	0	0	Couple					0	0	
4	0	1	1	0	0	Single					0	0	
5	0		1	0	0	Single					0	0	
6	0	2	2	0	0	Couple					0	0	
7	0	2	2	0	0	Couple					0	0	
8	0	2	2	0	0	Couple					0	0	
9	0	2	2	0	0	Couple					0	0	
10	0	3	2	0	0	Couple					0	0	
11	0	3	2	0	0	Couple					0	0	
12	0	4	2	0	0	Couple					0	0	
13	0	4	2	0	0	Couple					0	0	
14	0	4	2	0	0	Couple					0	0	
15	0	4	2	1	0	Family					0	0	
16	0	4	2	0	0	Couple					0	0	
17	0	4	2	0	0	Couple					0	0	
18	0	4	2	0	0	Couple	BB	PRT	Offline TA/TO	TA/TO	0	0	
19	0	1	2	0	0	Couple	BB	IRL	Online TA	TA/TO	0	0	
20	0	1	2	0	0	Couple	BB	FRA	Corporate	Corporate	0	0	
21	0	4	2	0	0	Couple	BB	GBR	Direct	Direct	0	0	
22	1	4	1	0	0	Single	BB	GBR	Online TA	TA/TO	0	0	
23	2	4	2	0	0	Couple	BB	PRT	Direct	Direct	0	0	
24	2	4	2	0	0	Couple	BB	PRT	Direct	Direct	0	0	

Status

Sort

Ascending | Descending

By color: None

Filter

By color: None

Choose One

☑ (Select All)
☐ Children without adults
☐ Couple
☐ Family
☐ Group
☑ No Reservation
☐ Single

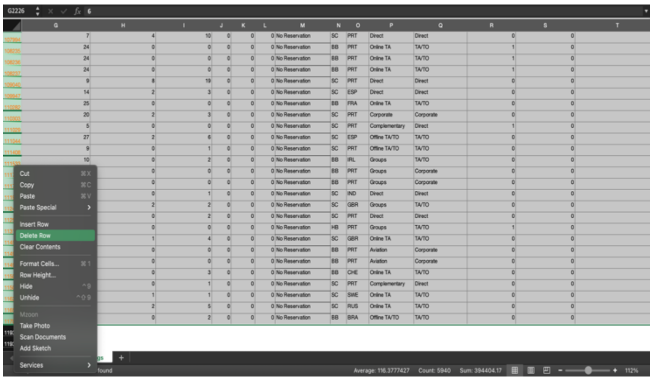
☐ Auto Apply

Apply Filter | Clear Filter

Sheet 1 - hotel_bookings

Ready 128%

2- Then we select all the 180 rows and choose the Delete row option.

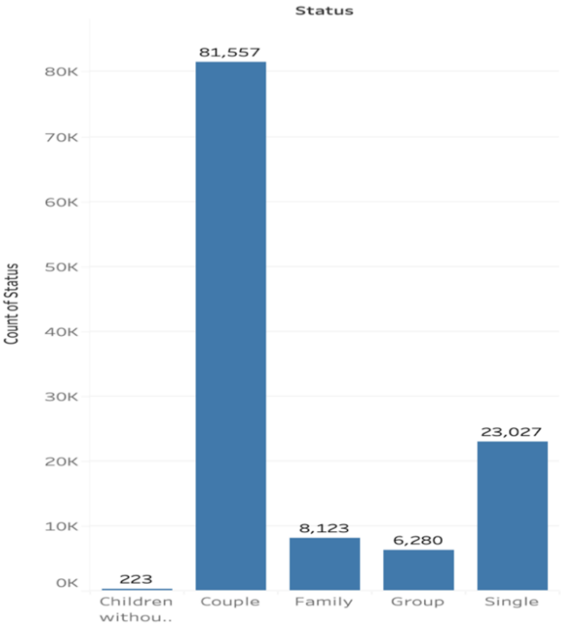


G2226 6

	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
107994	7	4	10	0	0	0	No Reservation	SC	PRT	Direct	Direct	0	0	
108235	24	0	0	0	0	0	No Reservation	BB	PRT	Online TA	TA/TO	1	0	
108236	24	0	0	0	0	0	No Reservation	BB	PRT	Online TA	TA/TO	1	0	
108237	24	0	0	0	0	0	No Reservation	BB	PRT	Online TA	TA/TO	1	0	
109040	9	8	19	0	0	0	No Reservation	SC	PRT	Direct	Direct	0	0	
109947	14	2	3	0	0	0	No Reservation	SC	ESP	Direct	Direct	0	0	
110282	25	0	0	0	0	0	No Reservation	BB	FRA	Online TA	TA/TO	0	0	
110303	20	2	3	0	0	0	No Reservation	SC	PRT	Corporate	Corporate	0	0	
111029	5	0	0	0	0	0	No Reservation	SC	PRT	Complementary	Direct	1	0	
111044	27	2	6	0	0	0	No Reservation	SC	ESP	Offline TA/TO	TA/TO	0	0	
111408	9	0	1	0	0	0	No Reservation	SC	PRT	Offline TA/TO	TA/TO	0	0	
111533	10	0	2	0	0	0	No Reservation	BB	IRL	Groups	TA/TO	0	0	
111		0	0	0	0	0	No Reservation	BB	PRT	Groups	Corporate	0	0	
111		0	0	0	0	0	No Reservation	BB	PRT	Groups	Corporate	0	0	
111		0	1	0	0	0	No Reservation	SC	IND	Direct	Direct	0	0	
112		2	2	0	0	0	No Reservation	SC	GBR	Groups	TA/TO	0	0	
112		0	2	0	0	0	No Reservation	SC	PRT	Direct	Direct	0	0	
113		0	0	0	0	0	No Reservation	HB	PRT	Groups	TA/TO	1	0	
114		1	4	0	0	0	No Reservation	SC	GBR	Online TA	TA/TO	0	0	
114		0	0	0	0	0	No Reservation	BB	PRT	Aviation	Corporate	0	0	
114		0	0	0	0	0	No Reservation	BB	PRT	Aviation	Corporate	0	0	
115		0	3	0	0	0	No Reservation	BB	CHE	Online TA	TA/TO	0	0	
115		0	1	0	0	0	No Reservation	SC	PRT	Complementary	Direct	0	0	
116		1	1	0	0	0	No Reservation	SC	SWE	Online TA	TA/TO	0	0	
116		2	5	0	0	0	No Reservation	SC	RUS	Online TA	TA/TO	0	0	
117		0	2	0	0	0	No Reservation	BB	BRA	Offline TA/TO	TA/TO	0	0	

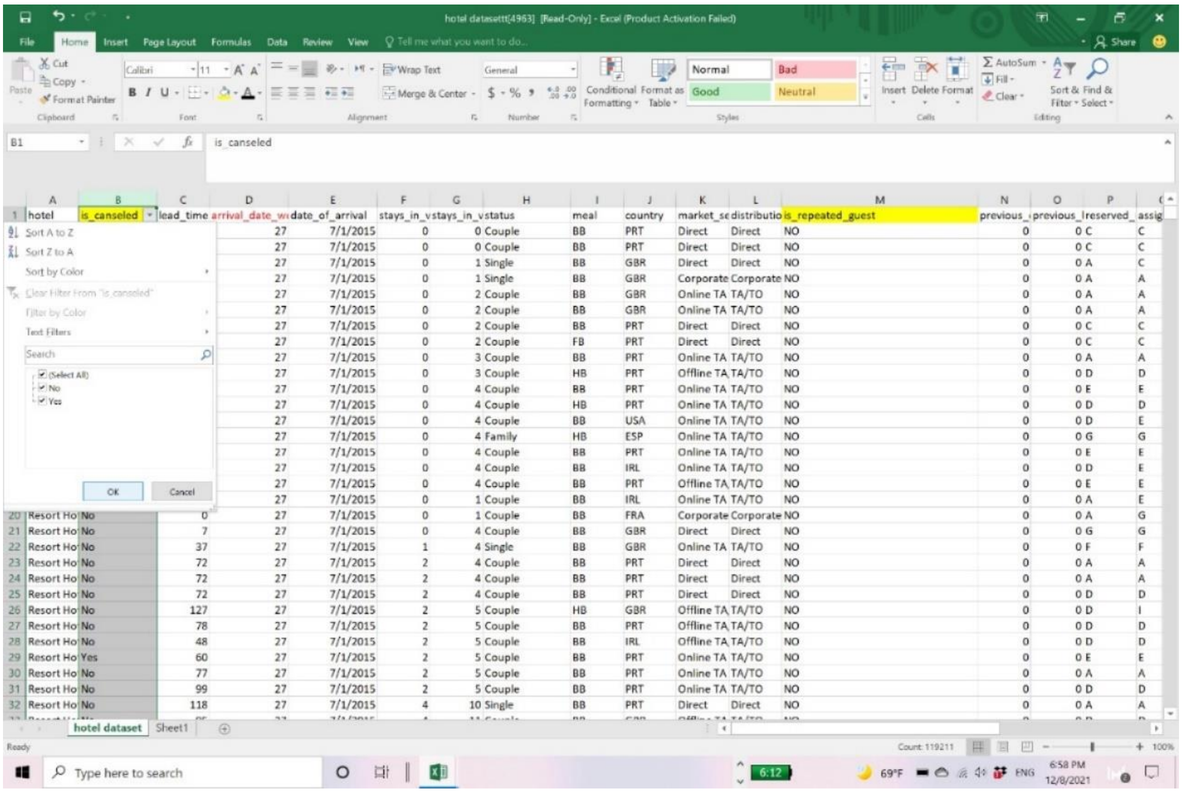
found Average: 116.3777427 Count: 5940 Sum: 394404.17 112%

In the last case, we also combined two conditions in one IF statement, which states that if the adult column is equal to 0 and the children column or babies is equal to or greater than 1, the result will be children without adult and if it does not stipulate one of the options mentioned, it will of course be couple.



Is canceled Attribute:

1-first step is to copy the column and then go to an empty excel to do the preprocessing.

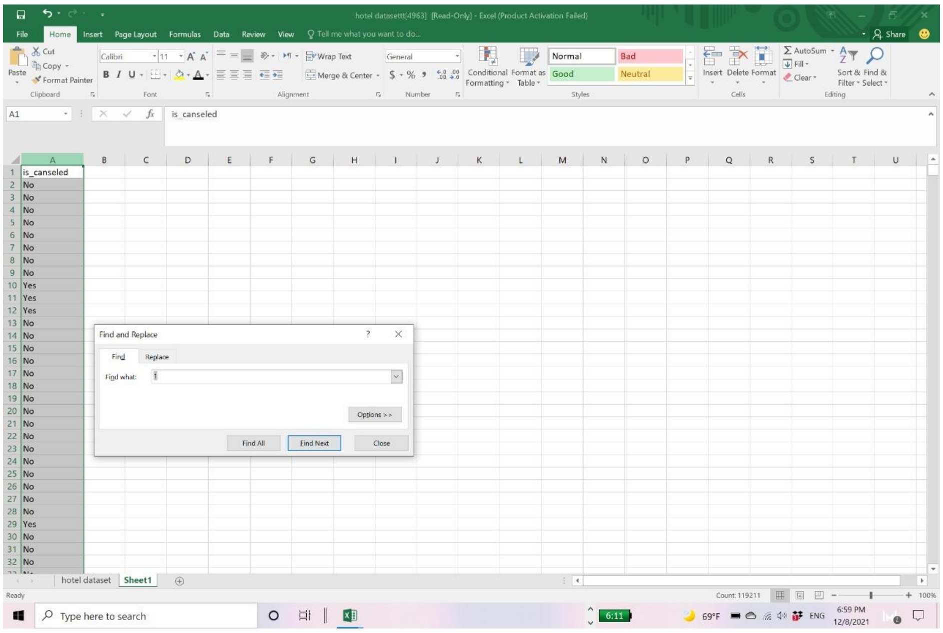


B1: is_canseled

☑ (Select All)
☑ No
☑ Yes

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	hotel	is_canseled	lead_time	arrival_date_w	date_of_arrival	stays_in_v	stays_in_v	status	meal	country	market_se	distributio	is_repeated_guest	previous_	previous_	reserved_	assig
2				27	7/1/2015	0	0	Couple	BB	PRT	Direct	Direct	NO	0	0	C	C
3				27	7/1/2015	0	0	Couple	BB	PRT	Direct	Direct	NO	0	0	C	C
4				27	7/1/2015	0	1	Single	BB	GBR	Direct	Direct	NO	0	0	A	C
5				27	7/1/2015	0	1	Single	BB	GBR	Corporate	Corporate	NO	0	0	A	A
6				27	7/1/2015	0	2	Couple	BB	GBR	Online TA	TA/TO	NO	0	0	A	A
7				27	7/1/2015	0	2	Couple	BB	GBR	Online TA	TA/TO	NO	0	0	A	A
8				27	7/1/2015	0	2	Couple	BB	PRT	Direct	Direct	NO	0	0	C	C
9				27	7/1/2015	0	2	Couple	FB	PRT	Direct	Direct	NO	0	0	C	C
10				27	7/1/2015	0	3	Couple	BB	PRT	Online TA	TA/TO	NO	0	0	A	A
11				27	7/1/2015	0	3	Couple	HB	PRT	Offline TA	TA/TO	NO	0	0	D	D
12				27	7/1/2015	0	4	Couple	BB	PRT	Online TA	TA/TO	NO	0	0	E	E
13				27	7/1/2015	0	4	Couple	HB	PRT	Online TA	TA/TO	NO	0	0	D	D
14				27	7/1/2015	0	4	Couple	BB	USA	Online TA	TA/TO	NO	0	0	D	E
15				27	7/1/2015	0	4	Family	HB	ESP	Online TA	TA/TO	NO	0	0	G	G
16				27	7/1/2015	0	4	Couple	BB	PRT	Online TA	TA/TO	NO	0	0	E	E
17				27	7/1/2015	0	4	Couple	BB	IRL	Online TA	TA/TO	NO	0	0	D	E
18				27	7/1/2015	0	4	Couple	BB	PRT	Offline TA	TA/TO	NO	0	0	E	E
19				27	7/1/2015	0	1	Couple	BB	IRL	Online TA	TA/TO	NO	0	0	A	E
20	Resort Ho	No	0	27	7/1/2015	0	1	Couple	BB	FRA	Corporate	Corporate	NO	0	0	A	G
21	Resort Ho	No	7	27	7/1/2015	0	4	Couple	BB	GBR	Direct	Direct	NO	0	0	G	G
22	Resort Ho	No	37	27	7/1/2015	1	4	Single	BB	GBR	Online TA	TA/TO	NO	0	0	F	F
23	Resort Ho	No	72	27	7/1/2015	2	4	Couple	BB	PRT	Direct	Direct	NO	0	0	A	A
24	Resort Ho	No	72	27	7/1/2015	2	4	Couple	BB	PRT	Direct	Direct	NO	0	0	A	A
25	Resort Ho	No	72	27	7/1/2015	2	4	Couple	BB	PRT	Direct	Direct	NO	0	0	D	D
26	Resort Ho	No	127	27	7/1/2015	2	5	Couple	HB	GBR	Offline TA	TA/TO	NO	0	0	D	I
27	Resort Ho	No	78	27	7/1/2015	2	5	Couple	BB	PRT	Offline TA	TA/TO	NO	0	0	D	D
28	Resort Ho	No	48	27	7/1/2015	2	5	Couple	BB	IRL	Offline TA	TA/TO	NO	0	0	D	D
29	Resort Ho	Yes	60	27	7/1/2015	2	5	Couple	BB	PRT	Online TA	TA/TO	NO	0	0	E	E
30	Resort Ho	No	77	27	7/1/2015	2	5	Couple	BB	PRT	Online TA	TA/TO	NO	0	0	A	A
31	Resort Ho	No	99	27	7/1/2015	2	5	Couple	BB	PRT	Online TA	TA/TO	NO	0	0	D	D
32	Resort Ho	No	118	27	7/1/2015	4	10	Single	BB	PRT	Direct	Direct	NO	0	0	A	A

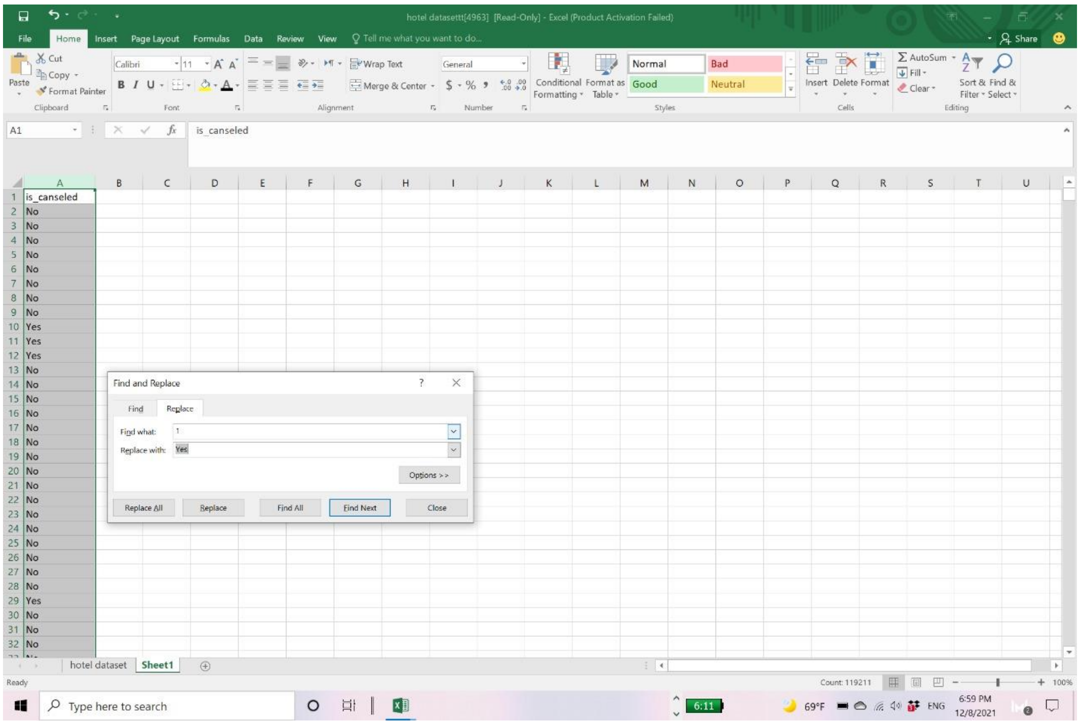
2-we will use find and replace feature to make 0 mean no and 1 mean yes.



A1: is_canseled

Find what: 1

	A
1	is_canseled
2	No
3	No
4	No
5	No
6	No
7	No
8	No
9	No
10	Yes
11	Yes
12	Yes
13	No
14	No
15	No
16	No
17	No
18	No
19	No
20	No
21	No
22	No
23	No
24	No
25	No
26	No
27	No
28	No
29	Yes
30	No
31	No
32	No

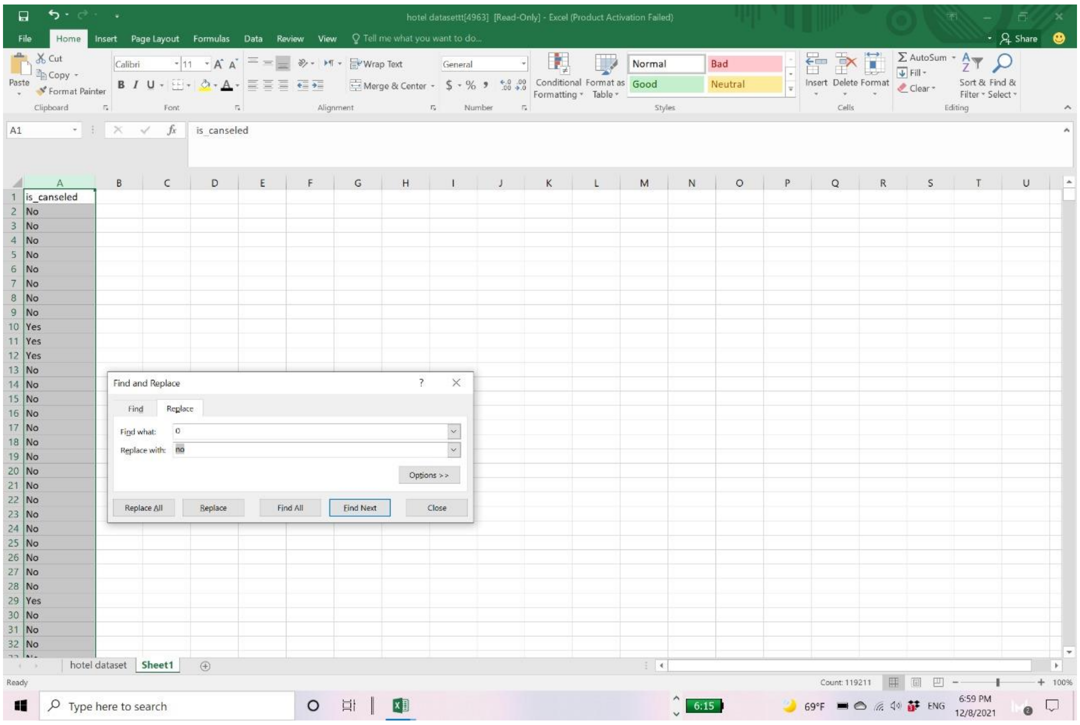

hotel datasett(4963] [Read-Only] - Excel (Product Activation Failed)
File Home Insert Page Layout Formulas Data Review View
Find and Replace
Find Replace
Options >>
Replace All Replace Find All Find Next Close
hotel dataset Sheet1
Ready
Count: 119211
Type here to search
6:59 PM
12/8/2021

A1 fx is_canseled

	A
1	is_canseled
2	No
3	No
4	No
5	No
6	No
7	No
8	No
9	No
10	Yes
11	Yes
12	Yes
13	No
14	No
15	No
16	No
17	No
18	No
19	No
20	No
21	No
22	No
23	No
24	No
25	No
26	No
27	No
28	No
29	Yes
30	No
31	No
32	No

Find what: 1

Replace with: Yes


hotel datasett(4963] [Read-Only] - Excel (Product Activation Failed)
File Home Insert Page Layout Formulas Data Review View
Find and Replace
Find Replace
Options >>
Replace All Replace Find All Find Next Close
hotel dataset Sheet1
Ready
Count: 119211
Type here to search
6:59 PM
12/8/2021

A1 fx is_canseled

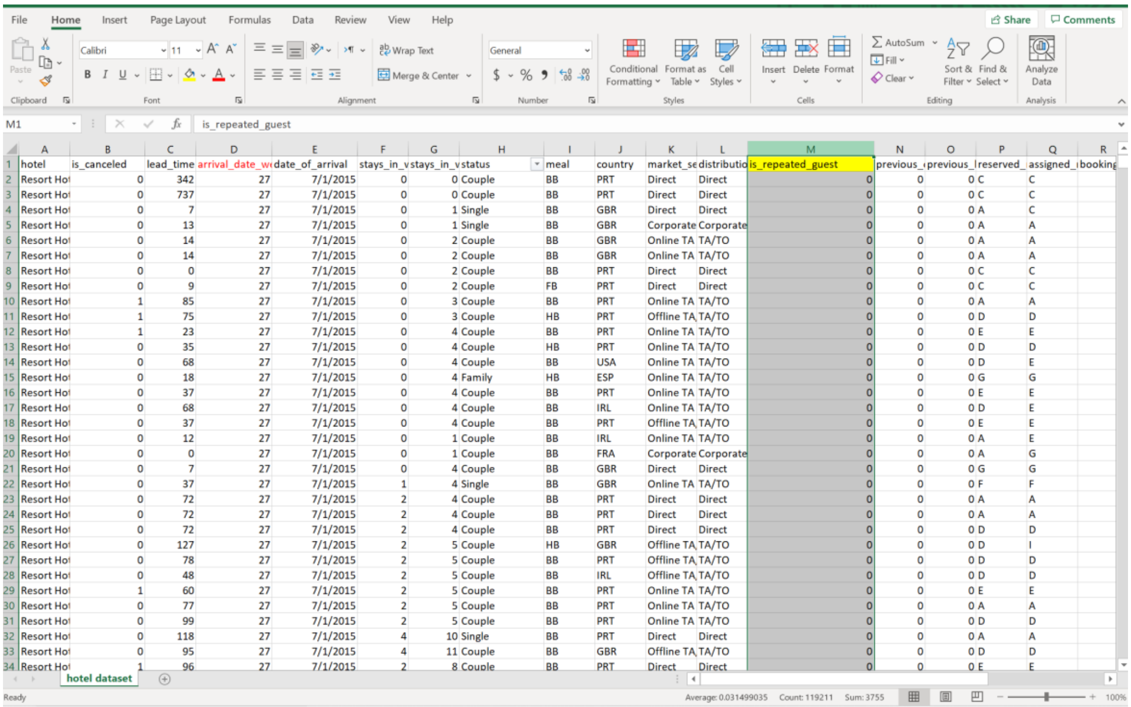
	A
1	is_canseled
2	No
3	No
4	No
5	No
6	No
7	No
8	No
9	No
10	Yes
11	Yes
12	Yes
13	No
14	No
15	No
16	No
17	No
18	No
19	No
20	No
21	No
22	No
23	No
24	No
25	No
26	No
27	No
28	No
29	Yes
30	No
31	No
32	No

Find what: 0

Replace with: no

IsRepeatedGuest Attribute:

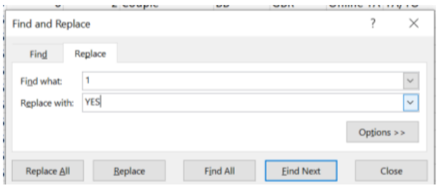
1-first step is to copy the column and then go to an empty excel to do the preprocessing



M1 | is_repeated_guest

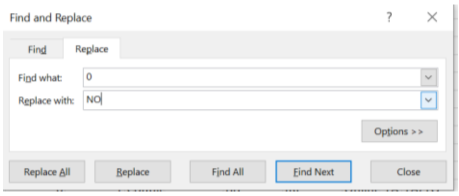
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	hotel	is_canceled	lead_time	arrival_date_w	date_of_arrival	stays_in_v	stays_in_v	status	meal	country	market_se	distributio	is_repeated_guest	previous_	previous_	reserved_	assigned_	booking
2	Resort Hot	0	342	27	7/1/2015	0	0	Couple	BB	PRT	Direct	Direct	0	0	0	C	C	
3	Resort Hot	0	737	27	7/1/2015	0	0	Couple	BB	PRT	Direct	Direct	0	0	0	C	C	
4	Resort Hot	0	7	27	7/1/2015	0	1	Single	BB	GBR	Direct	Direct	0	0	0	A	C	
5	Resort Hot	0	13	27	7/1/2015	0	1	Single	BB	GBR	Corporate	Corporate	0	0	0	A	A	
6	Resort Hot	0	14	27	7/1/2015	0	2	Couple	BB	GBR	Online TA	TA/TO	0	0	0	A	A	
7	Resort Hot	0	14	27	7/1/2015	0	2	Couple	BB	GBR	Online TA	TA/TO	0	0	0	A	A	
8	Resort Hot	0	0	27	7/1/2015	0	2	Couple	BB	PRT	Direct	Direct	0	0	0	C	C	
9	Resort Hot	0	9	27	7/1/2015	0	2	Couple	FB	PRT	Direct	Direct	0	0	0	C	C	
10	Resort Hot	1	85	27	7/1/2015	0	3	Couple	BB	PRT	Online TA	TA/TO	0	0	0	A	A	
11	Resort Hot	1	75	27	7/1/2015	0	3	Couple	HB	PRT	Offline TA	TA/TO	0	0	0	D	D	
12	Resort Hot	1	23	27	7/1/2015	0	4	Couple	BB	PRT	Online TA	TA/TO	0	0	0	E	E	
13	Resort Hot	0	35	27	7/1/2015	0	4	Couple	HB	PRT	Online TA	TA/TO	0	0	0	D	D	
14	Resort Hot	0	68	27	7/1/2015	0	4	Couple	BB	USA	Online TA	TA/TO	0	0	0	D	E	
15	Resort Hot	0	18	27	7/1/2015	0	4	Family	HB	ESP	Online TA	TA/TO	0	0	0	G	G	
16	Resort Hot	0	37	27	7/1/2015	0	4	Couple	BB	PRT	Online TA	TA/TO	0	0	0	E	E	
17	Resort Hot	0	68	27	7/1/2015	0	4	Couple	BB	IRL	Online TA	TA/TO	0	0	0	D	E	
18	Resort Hot	0	37	27	7/1/2015	0	4	Couple	BB	PRT	Offline TA	TA/TO	0	0	0	E	E	
19	Resort Hot	0	12	27	7/1/2015	0	1	Couple	BB	IRL	Online TA	TA/TO	0	0	0	A	E	
20	Resort Hot	0	0	27	7/1/2015	0	1	Couple	BB	FRA	Corporate	Corporate	0	0	0	A	G	
21	Resort Hot	0	7	27	7/1/2015	0	4	Couple	BB	GBR	Direct	Direct	0	0	0	G	G	
22	Resort Hot	0	37	27	7/1/2015	1	4	Single	BB	GBR	Online TA	TA/TO	0	0	0	F	F	
23	Resort Hot	0	72	27	7/1/2015	2	4	Couple	BB	PRT	Direct	Direct	0	0	0	A	A	
24	Resort Hot	0	72	27	7/1/2015	2	4	Couple	BB	PRT	Direct	Direct	0	0	0	A	A	
25	Resort Hot	0	72	27	7/1/2015	2	4	Couple	BB	PRT	Direct	Direct	0	0	0	D	D	
26	Resort Hot	0	127	27	7/1/2015	2	5	Couple	HB	GBR	Offline TA	TA/TO	0	0	0	D	I	
27	Resort Hot	0	78	27	7/1/2015	2	5	Couple	BB	PRT	Offline TA	TA/TO	0	0	0	D	D	
28	Resort Hot	0	48	27	7/1/2015	2	5	Couple	BB	IRL	Offline TA	TA/TO	0	0	0	D	D	
29	Resort Hot	1	60	27	7/1/2015	2	5	Couple	BB	PRT	Online TA	TA/TO	0	0	0	E	E	
30	Resort Hot	0	77	27	7/1/2015	2	5	Couple	BB	PRT	Online TA	TA/TO	0	0	0	A	A	
31	Resort Hot	0	99	27	7/1/2015	2	5	Couple	BB	PRT	Online TA	TA/TO	0	0	0	D	D	
32	Resort Hot	0	118	27	7/1/2015	4	10	Single	BB	PRT	Direct	Direct	0	0	0	A	A	
33	Resort Hot	0	95	27	7/1/2015	4	11	Couple	BB	GBR	Offline TA	TA/TO	0	0	0	D	D	
34	Resort Hot	1	96	27	7/1/2015	2	8	Couple	BB	PRT	Direct	Direct	0	0	0	E	E	

2-we will use find and replace feature to make 0 mean no and 1 mean yes.



Find what: 1

Replace with: YES



Find what: 0

Replace with: NO

The metadata after the preprocessing:

-this metadata includes only the attribute that been through preprocessing

attribute	Type	Description
purpose	Binary	This attribute indicates the purpose of the trip And have this two attribute value: Vacation: indicates that the guest is staying for vacation Business: this indicates that the guest is staying for work
Status	Categorical	This attribute indicates the room was reserved to what of this categorizes:

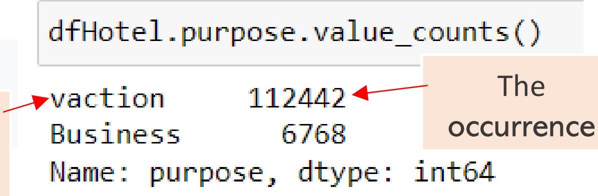
		Single: the room is reserved for only one adult without children or babies Couples: the room is reserved for only two adults without children or babies Groups: the room is reserved for more than two adults without babies or children Family: the room is reserved to one or two or more adults with children and babies Children without adults: the room is reserved to children without adults
Agent	Binary	This attribute indicates what are the online intermediary that the guest used to reserve in the hotel And have this two attribute value: Website: indicates that the guest used the hotel website to reserve in the hotel Booking: indicates that the guest used the Booking website to reserve in the hotel
DateOfArrival	Integer	This attribute indicates the day, month, year of arrival
IsCanceled	Binary	The Values indicates if the reservation was canceled the values is yes,no
IsRepeatedGuest	Binary	The value indicates if the booking was from a repeated/regular guest and the values is yes,no

-Attribute values:

-this includes only the attribute values that are relevant and need more explanation

Purpose Attribute:

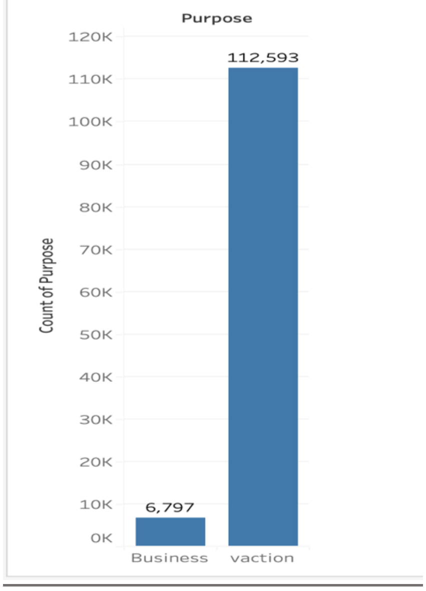
-now instead of company and its values it's more useful now like we can see if the guest is coming for a business trip so we can put them on a room away from disturbance and make an offer to them, the percentage for the Vacation is 94.33% and for the business trip 5.67%.



```
dfHotel.purpose.value_counts()

vaction       112442
Business        6768
Name: purpose, dtype: int64
```

The values in attribute purpose

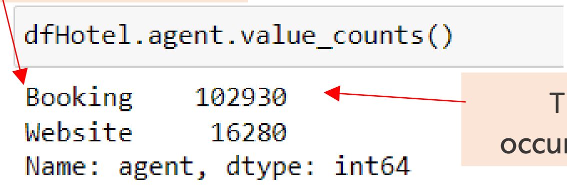
The occurrence



Agent Attribute:

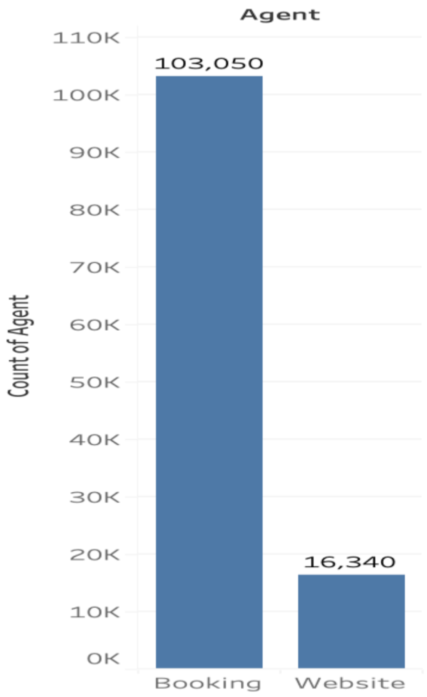
-now instead of the old values now it's more useful, like the number of visitors using booking are more than the visitors who use the hotel website which means maybe the website need more work for it to be recognized by the visitors, the percentage for Booking is 86.21% and for the website 13.64%.

The values in attribute Agent



```
dfHotel.agent.value_counts()

Booking     102930
Website      16280
Name: agent, dtype: int64
```

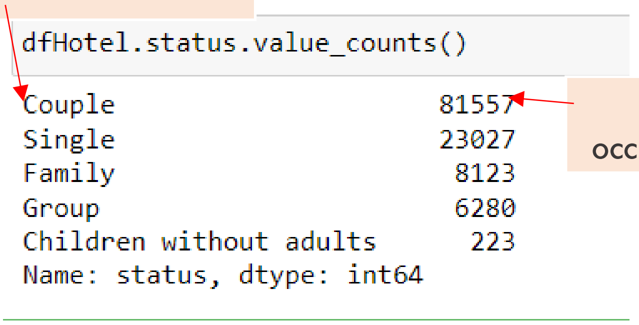
The occurrence



Status Attribute:

-now the attribute and its values are more useful, the children without adults could indicates that their parent reserved the room for their kids only ,there was a no reservation but since it just 180 row out of 119390 we deleted it

The values in attribute Status



```
dfHotel.status.value_counts()

Couple                      81557
Single                      23027
Family                       8123
Group                        6280
Children without adults       223
Name: status, dtype: int64
```

The occurrence

The percentage for status:

Couple:68.31%

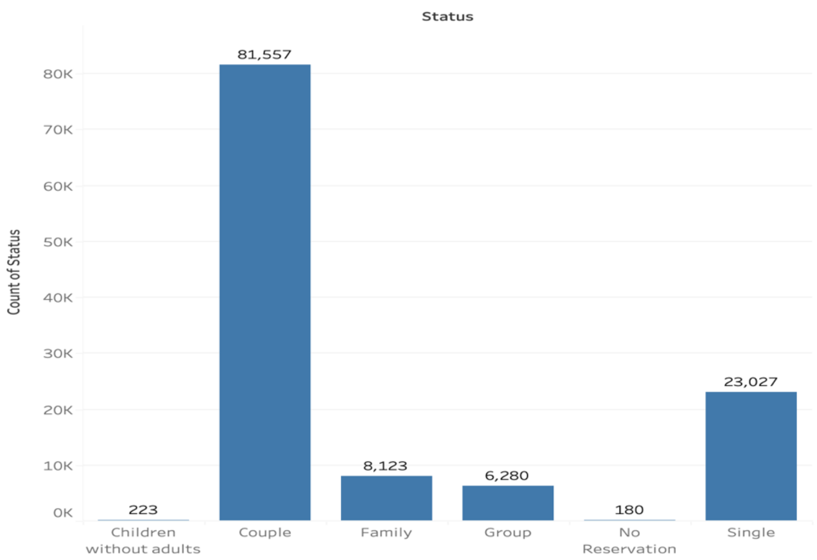
Single: 19.29%

Family:6.8%

Group:5.26%

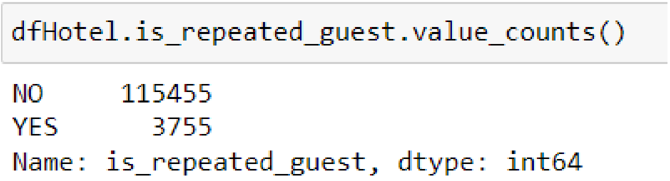
Children without adults:1.87%

So the most status of the reservation is reserved for couples.



IsRepeatedGuest Attribute:

this attribute includes yes and no as its attribute values the no mean it's not a repeated guest and yes means it's repeated guest the percentage of the repeated guest is 3.15% and 96.85% is not a repeated guest.

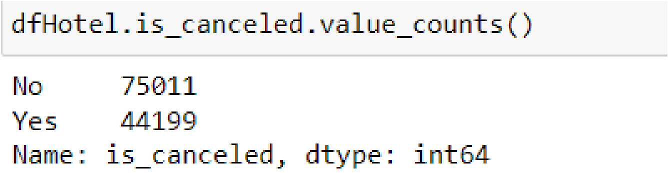


```
dfHotel.is_repeated_guest.value_counts()
```

```
NO     115455
YES      3755
Name: is_repeated_guest, dtype: int64
```

IsCanceled Attribute:

this attribute includes yes and no as its attribute values the no mean it's not a canceled and yes means it's canceled percentage of the cancelation 37.04% and who didn't cancel 62.96%

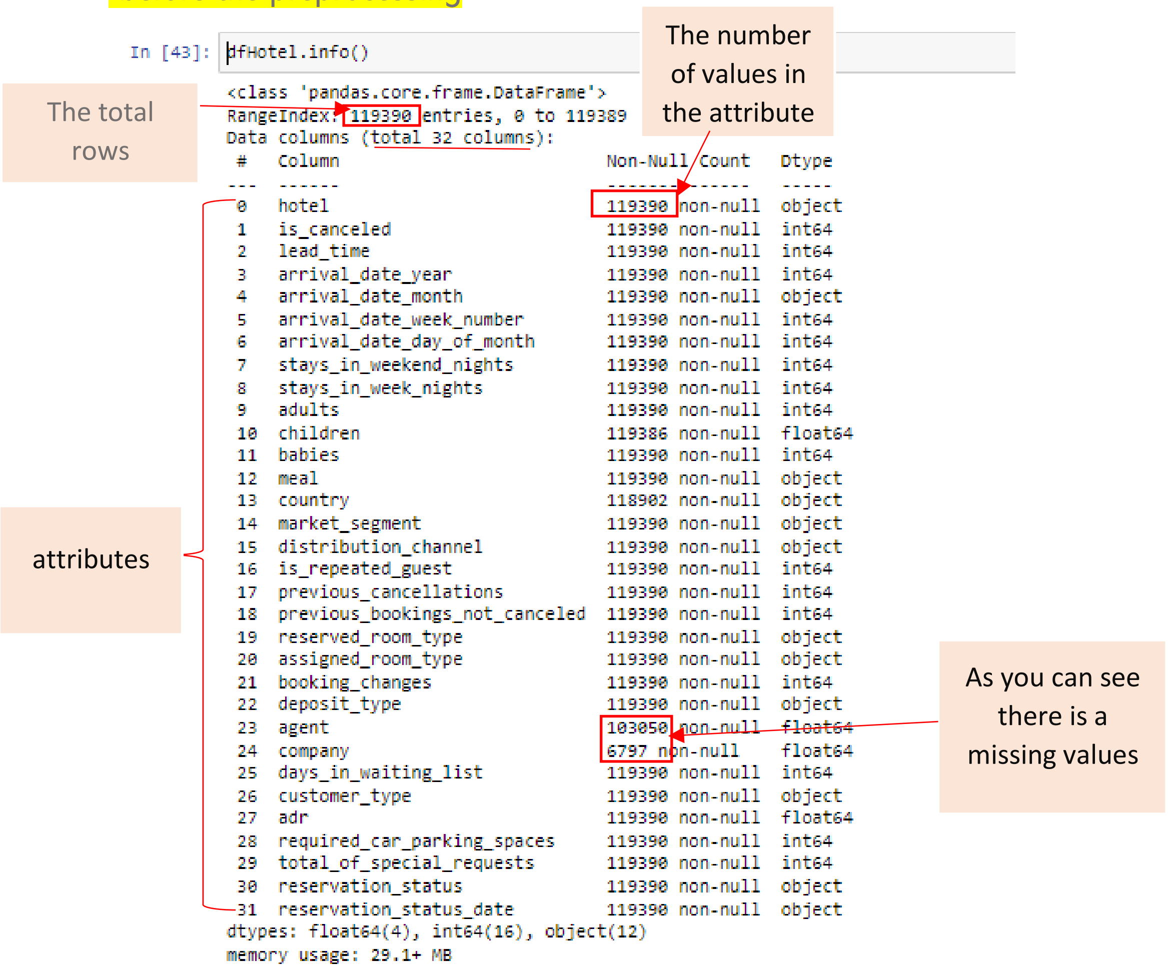


```
dfHotel.is_canceled.value_counts()
```

```
No     75011
Yes    44199
Name: is_canceled, dtype: int64
```

-The dataset before and after the prepossessing:

-before the preprocessing



```
In [43]: dfHotel.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 32 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           119390 non-null  object
 1   is_canceled                     119390 non-null  int64
 2   lead_time                       119390 non-null  int64
 3   arrival_date_year               119390 non-null  int64
 4   arrival_date_month              119390 non-null  object
 5   arrival_date_week_number        119390 non-null  int64
 6   arrival_date_day_of_month       119390 non-null  int64
 7   stays_in_weekend_nights         119390 non-null  int64
 8   stays_in_week_nights            119390 non-null  int64
 9   adults                          119390 non-null  int64
 10  children                        119386 non-null  float64
 11  babies                          119390 non-null  int64
 12  meal                            119390 non-null  object
 13  country                         118902 non-null  object
 14  market_segment                  119390 non-null  object
 15  distribution_channel            119390 non-null  object
 16  is_repeated_guest               119390 non-null  int64
 17  previous_cancellations          119390 non-null  int64
 18  previous_bookings_not_canceled  119390 non-null  int64
 19  reserved_room_type              119390 non-null  object
 20  assigned_room_type              119390 non-null  object
 21  booking_changes                 119390 non-null  int64
 22  deposit_type                    119390 non-null  object
 23  agent                           103050 non-null  float64
 24  company                         6797 non-null    float64
 25  days_in_waiting_list            119390 non-null  int64
 26  customer_type                   119390 non-null  object
 27  adr                             119390 non-null  float64
 28  required_car_parking_spaces     119390 non-null  int64
 29  total_of_special_requests       119390 non-null  int64
 30  reservation_status              119390 non-null  object
 31  reservation_status_date         119390 non-null  object
dtypes: float64(4), int64(16), object(12)
memory usage: 29.1+ MB
```

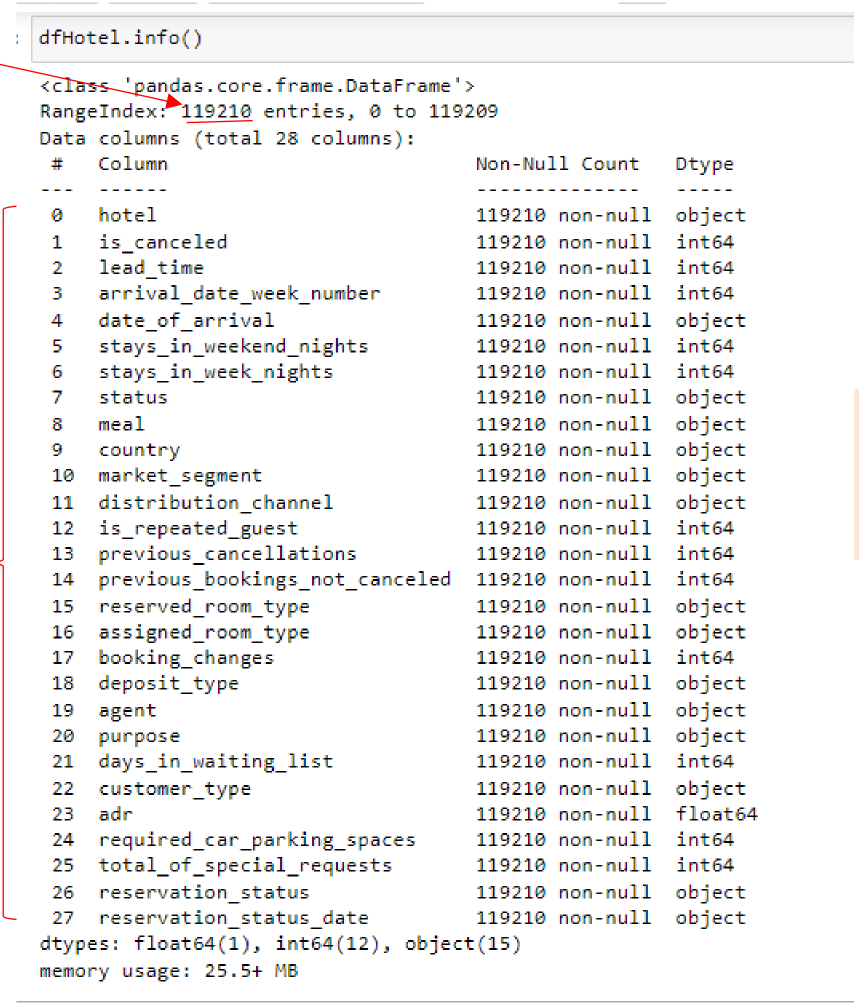
The number of values in the attribute

The total rows

attributes

As you can see there is a missing values

-After the preprocessing



```
dfHotel.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119210 entries, 0 to 119209
Data columns (total 28 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           119210 non-null  object
 1   is_canceled                     119210 non-null  int64
 2   lead_time                       119210 non-null  int64
 3   arrival_date_week_number        119210 non-null  int64
 4   date_of_arrival                 119210 non-null  object
 5   stays_in_weekend_nights         119210 non-null  int64
 6   stays_in_week_nights            119210 non-null  int64
 7   status                          119210 non-null  object
 8   meal                            119210 non-null  object
 9   country                         119210 non-null  object
 10  market_segment                  119210 non-null  object
 11  distribution_channel            119210 non-null  object
 12  is_repeated_guest               119210 non-null  int64
 13  previous_cancellations          119210 non-null  int64
 14  previous_bookings_not_canceled  119210 non-null  int64
 15  reserved_room_type              119210 non-null  object
 16  assigned_room_type              119210 non-null  object
 17  booking_changes                 119210 non-null  int64
 18  deposit_type                    119210 non-null  object
 19  agent                           119210 non-null  object
 20  purpose                         119210 non-null  object
 21  days_in_waiting_list            119210 non-null  int64
 22  customer_type                   119210 non-null  object
 23  adr                             119210 non-null  float64
 24  required_car_parking_spaces     119210 non-null  int64
 25  total_of_special_requests       119210 non-null  int64
 26  reservation_status              119210 non-null  object
 27  reservation_status_date         119210 non-null  object
dtypes: float64(1), int64(12), object(15)
memory usage: 25.5+ MB
```

The total rows

attributes

As you can see there is no missing values

- The limitation:

1- The first limitation in the data set were the Is Canceled column and the Is Repeated guest column were represented as Integer 0,1 and we converted them to Categorical Yes,No to make it easier to understand, especially in the decision tree.

2- The second limitation were in the column of the arrival date year, arrival date month, and the arrival date day of month, each of them was in a separate column, and this is considered illogical to some extent, so we collected them in one column to make it easier to read.

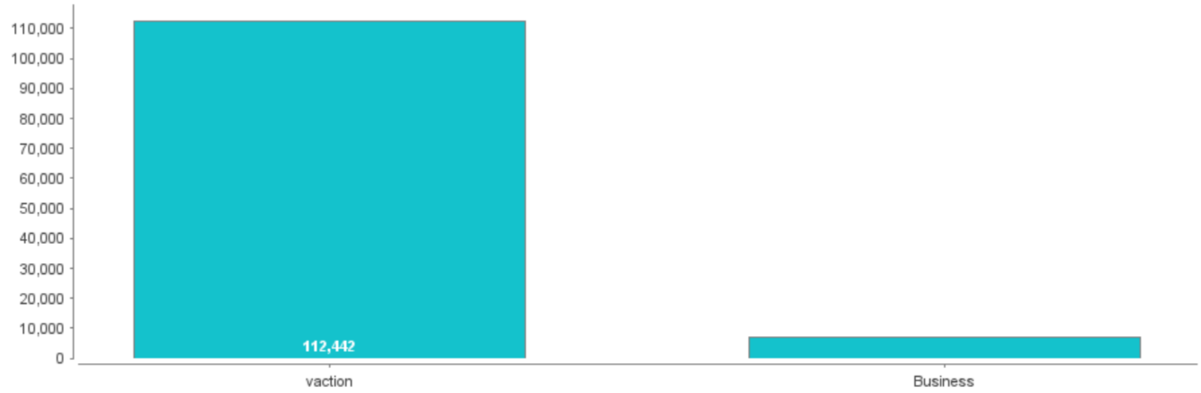
3- The third limitation was in the adult, Children and Babies column. It was difficult to read the columns and determine the type of reservation, so we made a new column that combines these three columns so that it is classified if they are single, couple , family, groups or children without adult.

4- According to the third limitation, 180 rows were found. Adult, children, and babies all contained zero, with other information in other attributes, and this is also illogical. After discussing the limitation with the project members, it was decided to delete the rows because there is no information that can help us determinate the status, and it constitutes a very small percentage of 100,000 row that we have.

5- The fourth limitation were in the Agent and company column it's contained many codes and nulls as well, so they were collected in one column to show whether the booking was through the website or through Booking itself.

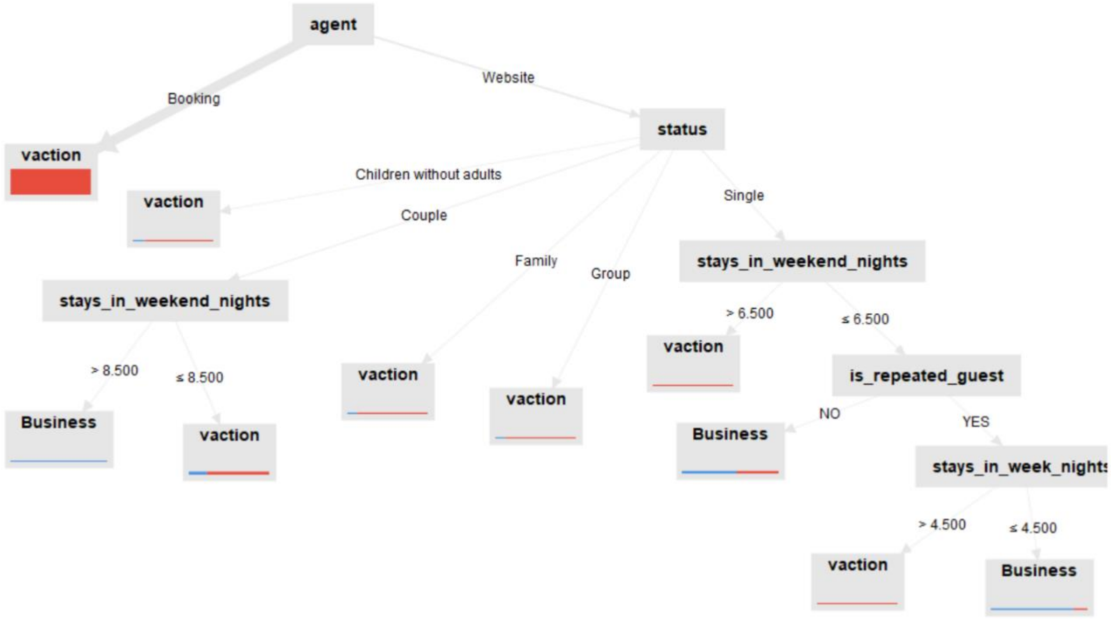
6- In the end, there were a number of attributes that were not sufficiently explained in the metadata, so we wish we know the owner of the dataset to ask him some questions that will help us.

Pattern discovery:



We build our model based on this chart comparison, choose "purpose" as class label so we want to demonstrate opportunities that hidden in purpose attribute.

The result:



We choose prediction model and decision tree method, our class label is Purpose and the class label indicate tow results from traveling 1(vacation) ,2 (Business)

Explanation:

Based on purpose class label we notice the following:

1- Who booked through Booking (as travel agent) the purpose from the travel is Vacation

2- Who booked through website will be based on another indicator which is (statues)

   Statues values are: Group, Family, Single, Couple, Children without adults

   A- If booked through website and statues is Children without adults, purpose is vacation

   B- If booked through website and statues is Family, purpose is vacation

   C- If booked through website and statues is Group, purpose is vacation

   If booked through website and statues is Couple, purpose is depending in (Stays in weekend night) Attribute

   Stays in weekend nights Attribute indicate: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel (integer)

   If the number of the gust stays in weekend nights was >8.500, purpose will be Business

   Else the number of the gust stays in weekend nights was≤ 8.500, purpose will be vacation

   D- If booked through website and statues is single

   is depending in (Stays in weekend night) Attribute

   If the number of the gust stays in weekend nights was >6.500, purpose will be Vacation
   Else the number of the gust stays in weekend nights was≤ 6.500, we will look to if the gust is repeated gust through attribute called (is_repeted_guest)
   So
   If guest not repeated guest, purpose will be business
   If the guest is repeated guest, we look up for the (stays in week night) attribute
   Stays in week night indicate: Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel (integer)
   If the number of the gust stays in week nights was >4.500, purpose will be Vacation
   Else the number of the gust stays in week nights was≤ 4.500, purpose will be Business

Our recommendation:

1-In scenario number “1”

Have special collaboration with booking to customize service to people who booked through booking in order to provide seasonal offers, special prices, customize the hotel service ex: massage, hair salon, gym for guest from booking

2-in scenario number “2”

A, B, C, D couple, single, have same purpose (vacation) with different statues segment: Customize hotel website

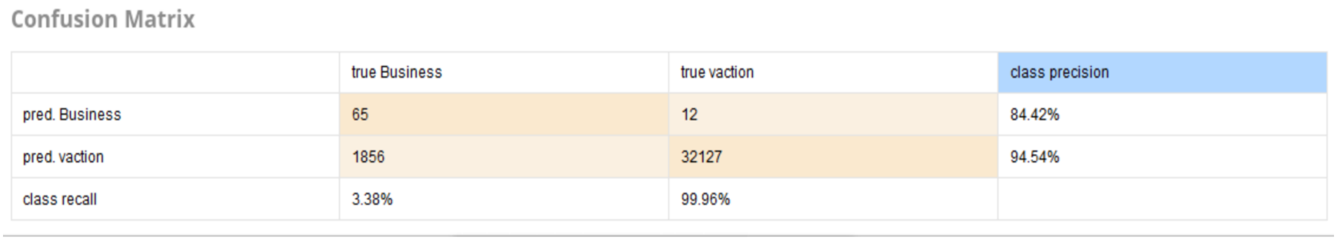
1-updated offers in real time
2-accesbility
3-easy to navigate
Provide special activates ex: choose specific room, activity for family group
Reservation for table in the restaurant for any kind of occasions
Other
4-provide partnership with rental car company to make the customers transportation easier.
5-Provide bus to guide tour across city or island

3- couple and single where their purpose is business

Track period of their business trips and:

- Provide separate section with strong internet connection
- Provide separate elevators used to reach designed (meetings room)
- Provide special cars to pick them form the airport to the hotel or any desired locations

Performance:



Confusion Matrix

	true Business	true vaction	class precision
pred. Business	65	12	84.42%
pred. vaction	1856	32127	94.54%
class recall	3.38%	99.96%	

Explanation:

Performance: F score
F score for Business: 0.06499
F score for Vacation: 0.9717

Based on confusion matrix:

We calculated F1 score (F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is **[0, 1]**. It tells you how precise your classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances). and the results was indicating increment in vacation as our model predicted

Numbers

Vacation: 0.9717 close to 1, our recommendation is that hotel offers remain stable for travelers with vacation purpose

Business: 0.0649 which indicate decrement in number of travelers with business purpose, with take in consideration that number of travelers with business purpose are most likely stays in city hotels

We target business with different statues, the hotel should look for numbers of business travelers in city hotels and manage to provide special offers, service for them including: meetings room, quite sections and areas.

Furthermore:

Based on confusion matrix:

we have high recall value; high recall indicates: there were very few false negatives (test result that is incorrect because the test failed to recognize an existing condition or finding make the condition absent .) in high recall the classifier is more permissive in the criteria for classifying something as positive, in other word our recall return most relevant information about purpose attribute, which indicate that we will focus on travelers with vacation purpose either our class predicted as true or false ,this measure gives our model more flexibility to choose more segment to serve

“32127” travelers have predicted as true vacation

“1856” travelers have predicted as false vacation (false negative)

Additionally:

We also have high precision whish mean that an algorithm returns more relevant results than irrelevant ones, with few false positive (test result which wrongly indicates that a particular condition or attribute is present) in our model if we chose to focus on high precision we will focus on travelers with correct business purpose only.

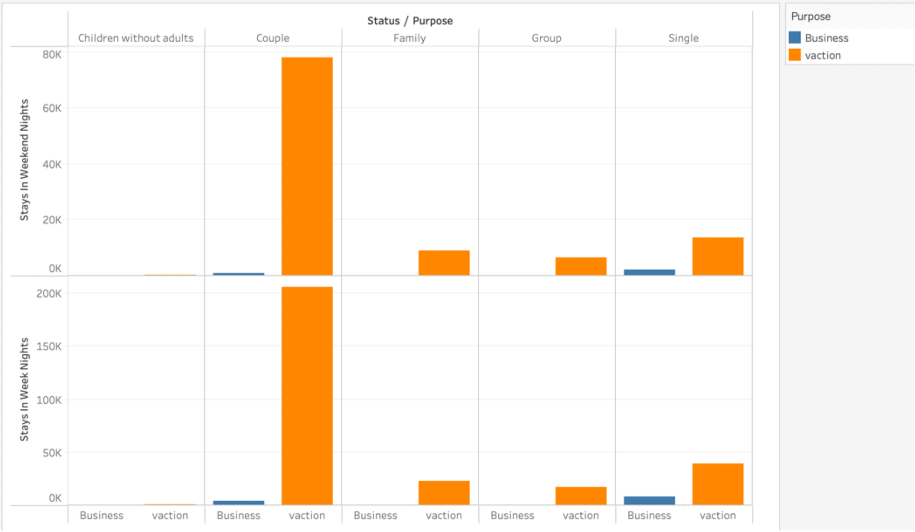
“65” travelers have predicted as true Business

“12” travelers have predicted as false Business (false positive)

Finally:

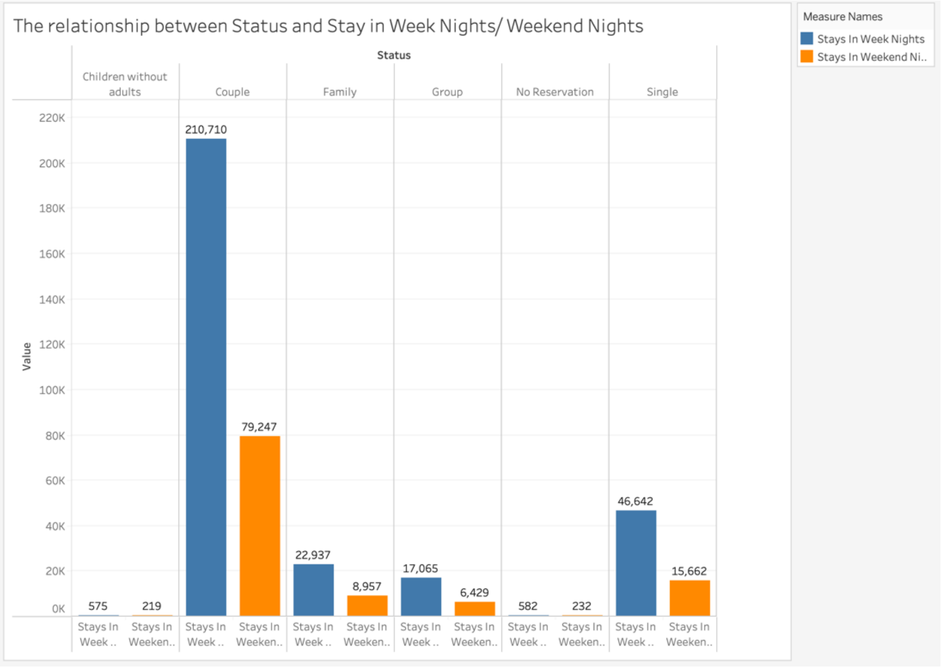
We have an ideal model with high precision and recall with all results predicted correctly.

Charts and comparisons

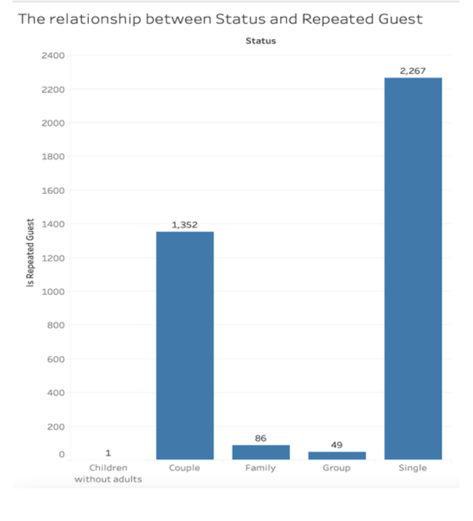


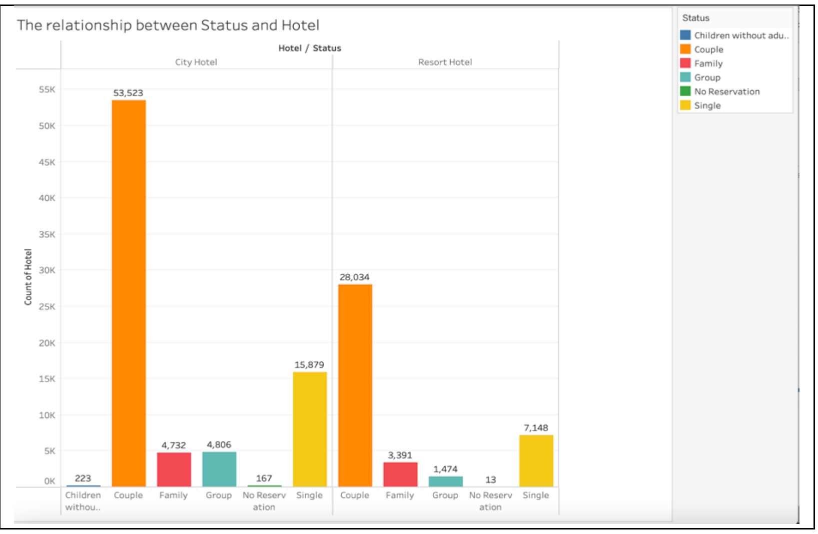
As we see above, when we are compering all Status attribute and the Purpose if it’s a cognize eBusiness or Vacation with if they stay in weekend nights or week nights, we r that all of type of Status is stays in Week nights/Weekend nights for vacation is more than business purpose.

So, based on that we should focuses on how to offer a good services for those who come for vacation because they are the majority.



Here we can see the relationship between “Status” and stay in Week nights/Weekend nights. We noticed the guests who stays in Week nights is more than Weekend, so because of that the hotels may focus on how to make special offer in price or Special events and activities in the Weekend to attract people.





We need to know which type of Status is repeating guest and what is the hotel they reserve.

Here is the relationship between Status and Repeated Guest, so we noticed Singles are the most repeated guest and based on the right picture they were reserved City hotel more than resort hotel.

Conclusion:

In conclusion, the result we had after applying classification algorithm in our dataset is the data have been classified to different groups , and as a result we could explain those groups such as giving the appropriate services for the different type of customers and for their different type of purpose. We reached out the final hidden pattern for our model, one of the most helping tool was confusion matrix and F1 score measurement which guide as to discover new segment or target to serve them in sufficient way.

Some of the limitations that we have faced during this project is that using excel for data preprocessing was not as efficient as using python, the data preprocessing require logic thinking and problem solving skills And using Rapidminer as main software there were difficulties during the mining part starting from building the model to the final result, maybe if we could use another software like Jupyter it could be easier and we could find more useful tutorials. And for working in this project we had a zoom meetings during the semester that have reached 15 hours.

References:

1- Antonio, N., Almeida, A. de, & Nunes, L. (2018, November 29). Hotel booking demand datasets. Data in Brief. Retrieved December 11, 2021, from https://www.sciencedirect.com/science/article/pii/S2352340918315191.

2- tutorial) Jupyter Notebook: The definitive guide. DataCamp) Community. (n.d.). Retrieved December 11, 2021, from https://www.datacamp.com/community/tutorials/tutorial-jupyter-notebook?utm_source=adwords_ppc&amp;utm_medium=cpc&amp;utm_campaignid=14989519638&amp;utm_adgroupid=127836677279&amp;utm_device=c&amp;utm_keyword=&amp;utm_matchtype=&amp;utm_network=g&amp;utm_adpostion=&amp;utm_creative=278443377086&amp;utm_targetid=aud-299261629574%3Adsa-473406581035&amp;utm_loc_interest_ms=&amp;utm_loc_physical_ms=9077053&amp;gclid=Cj0KCQiA47GNBhDrARIsAKfZ2rATmPUtVdmnPMf.UoLwn4VeB5fiNksftF_mCSE8Rr-W0BuW1V1RTzXAaAqDCEALw_wcB

3- Where developers learn, share, & build careers. Stack Overflow. (n.d.). Retrieved December 11, 2021, from https://stackoverflow.com/.

4- Mostipak, J. (2020, February 13). Hotel Booking Demand. Kaggle. Retrieved December 11, 2021, from https://www.kaggle.com/jessemostipak/hotel-booking-demand.

5- https://docs.google.com/spreadsheets/d/18UX9r57cW_0boUSLsF-xEMPU6RE9qXQO/edit?usp=sharing&ouid=115413546090402204534&rtpof=true&sd=true

For more charts or if the charts are not obvious CLICK HERE