




CUSTOMER SEGMENTATION

a Hotel at Lisbon Portugal

by: Camelia Fitrianty

BACKGROUND




This real-world customer data set with 31 variables describes 83,590 instances (customers) of a hotel in Lisbon, Portugal. In addition to personal and behavioral information, the dataset also contains demographic and geographic information.

www.kaggle.com/datasets/nantonio/a-hotels-customers-dataset



PROBLEM DEFINITION



Lots of customer data at a hotel in Lisbon, Portugal (2015-2018), will make it easier for the scope of data science to group customers according to personality, behavior, demographics. The goal is to make marketing more targeted with different customer groups.



WORKFLOW



1. Data Cleaning

2. Data Preprocessing

- delete unnecessary columns
- add new column ('Total Revenue')
- statistical Summary
- heatmap correlation
- clean Outliers

3. Explanatory Data

4. Normalization data (using MinMaxScaler)

5. Hyperparameter tuning

- Yellow brick method, Elbow Plot

6. Modelling (using KMeans Clustering)

- Reduction PCA for visualization 2 dimension

7. Insight

ABOUT DATASET

ID	83510	int64	BookingsCheckedIn	83510	int64	SRBathtub	83510	int64
Nationality	83510	object	PersonsNights	83510	int64	SR Shower	83510	int64
Age	79811	float64	RoomNights	83510	int64	SRCrib	83510	int64
DaysSinceCreation	83510	int64	DaysSinceLastStay	83510	int64	SRKingSizeBed	83510	int64
NameHash	83510	object	DaysSinceFirstStay	83510	int64	SRTwinBed	83510	int64
DocIDHash	83510	object	DistributionChannel	83510	object	SRNearElevator	83510	int64
AverageLeadTime	83510	int64	MarketSegment	83510	object	SRAwayFromElevator	83510	int64
LodgingRevenue	83510	float64	SRHighFloor	83510	int64	SRNoAlcoholInMiniBar	83510	int64
OtherRevenue	83510	float64	SRLowFloor	83510	int64	SRQuietRoom	83510	int64
BookingsCanceled	83510	int64	SRAccessibleRoom	83510	int64			
BookingsNoShowed	83510	int64	SRMediumFloor	83510	int64			

DATA CLEANING

- Check Missing Value

ID	0	BookingsCheckedIn	0	SRBathtub	0
Nationality	0	PersonsNights	0	SR Shower	0
Age	3779	RoomNights	0	SRCrib	0
DaysSinceCreation	0	DaysSinceLastStay	0	SRKingSizeBed	0
NameHash	0	DaysSinceFirstStay	0	SRTwinBed	0
DocIDHash	0	DistributionChannel	0	SRNearElevator	0
AverageLeadTime	0	MarketSegment	0	SRAwayFromElevator	0
LodgingRevenue	0	SRHighFloor	0	SRNoAlcoholInMiniBar	0
OtherRevenue	0	SRLowFloor	0	SRQuietRoom	0
BookingsCanceled	0	SRAccessibleRoom	0		
BookingsNoShowed	0	SRMediumFloor	0		

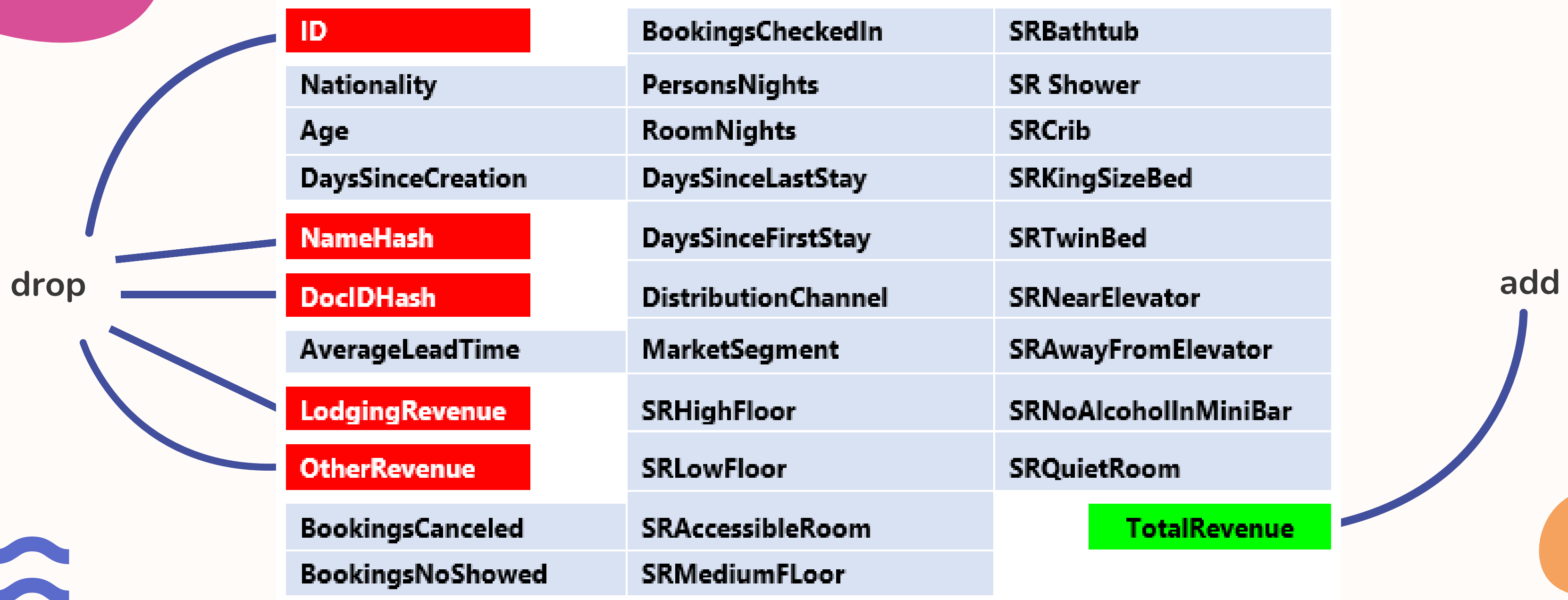
- Check Duplicated Rows = 0

DATA CLEANING

- After Cleaning

ID	79811	int64	BookingsCheckedIn	79811	int64	SRBathtub	79811	int64
Nationality	79811	object	PersonsNights	79811	int64	SR Shower	79811	int64
Age	79811	float64	RoomNights	79811	int64	SRCrib	79811	int64
DaysSinceCreation	79811	int64	DaysSinceLastStay	79811	int64	SRKingSizeBed	79811	int64
NameHash	79811	object	DaysSinceFirstStay	79811	int64	SRTwinBed	79811	int64
DocIDHash	79811	object	DistributionChannel	79811	object	SRNearElevator	79811	int64
AverageLeadTime	79811	int64	MarketSegment	79811	object	SRAwayFromElevator	79811	int64
LodgingRevenue	79811	float64	SRHighFloor	79811	int64	SRNoAlcoholInMiniBar	79811	int64
OtherRevenue	79811	float64	SRLowFloor	79811	int64	SRQuietRoom	79811	int64
BookingsCanceled	79811	int64	SRAccessibleRoom	79811	int64			
BookingsNoShowed	79811	int64	SRMediumFloor	79811	int64			

DATA PREPROCESSING



DATA PREPROCESSING

- Summary of Categorical Data

	Nationality	DistributionChannel	MarketSegment
count	79811	79811	79811
unique	188	4	7
top	FRA	Travel Agent/Operator	Other
freq	12422	65692	46204

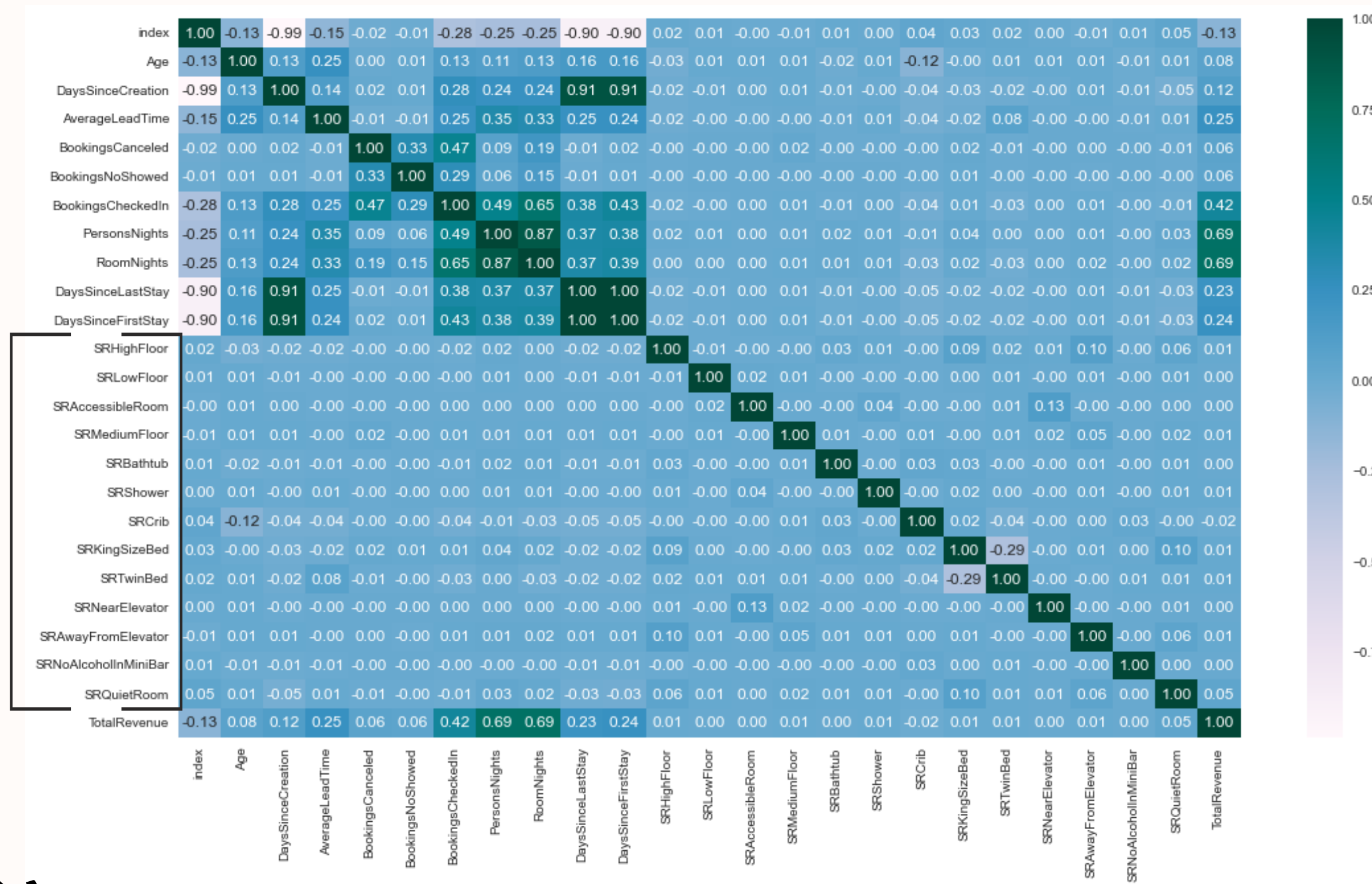
DATA PREPROCESSING

Summary of Numerical Data ●

	Age	DaysSinceCreation	AverageLeadTime	BookingsCanceled	BookingsNoShowed	BookingsCheckedIn	PersonsNights	RoomNights
count	79811.000000	79811.000000	79811.000000	79811.000000	79811.000000	79811.000000	79811.000000	79811.000000
mean	45.398028	446.483267	66.809663	0.001842	0.000576	0.792948	4.698337	2.376276
std	16.572368	310.620996	87.990086	0.065912	0.028312	0.690435	4.587289	2.196953
min	-11.000000	0.000000	-1.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	34.000000	174.000000	0.000000	0.000000	0.000000	1.000000	1.000000	1.000000
50%	46.000000	385.000000	30.000000	0.000000	0.000000	1.000000	4.000000	2.000000
75%	57.000000	703.000000	104.000000	0.000000	0.000000	1.000000	8.000000	4.000000
max	122.000000	1095.000000	588.000000	9.000000	3.000000	66.000000	116.000000	116.000000

for more information about this table, kindly check my
jupyter notebook 😊

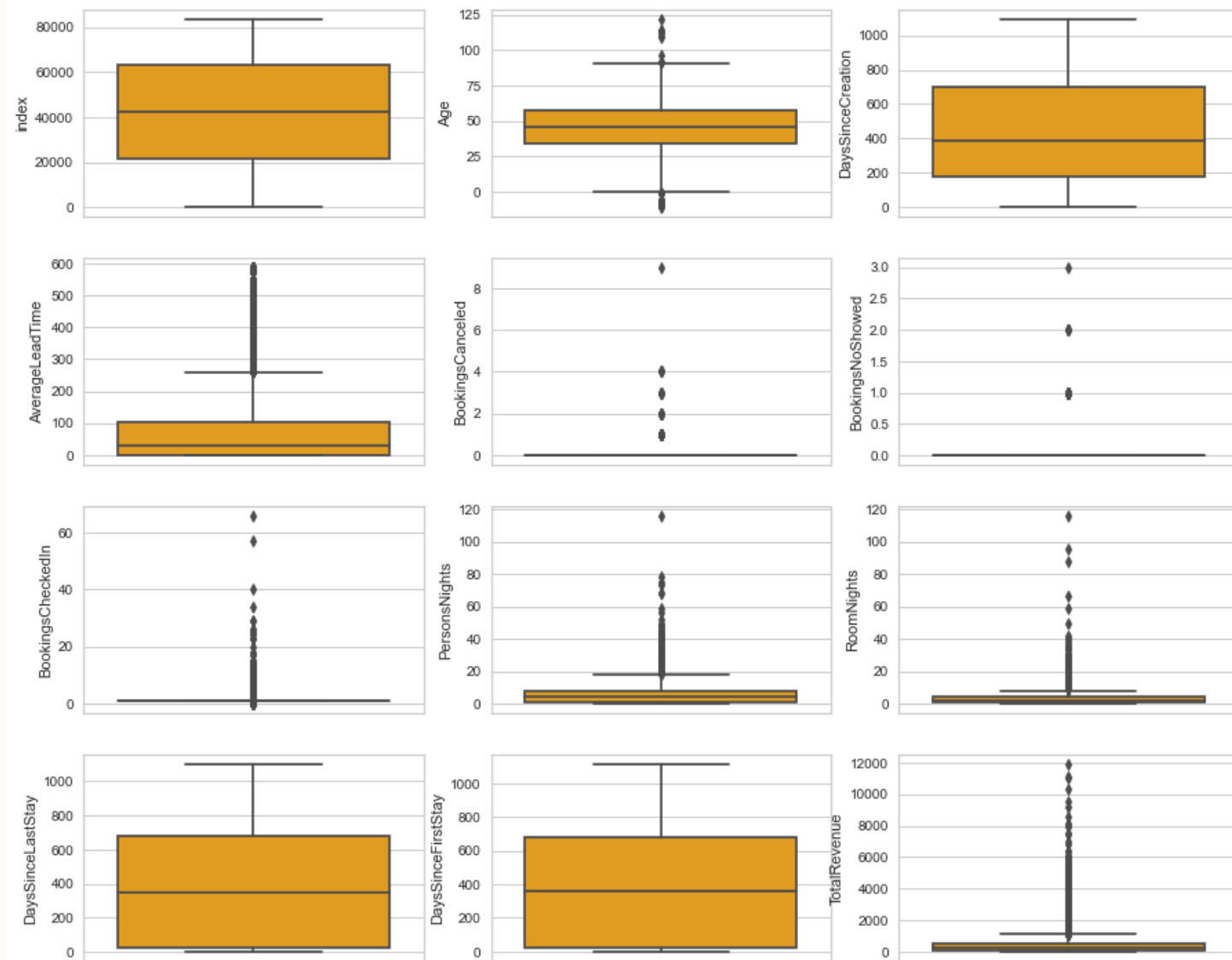
DATA PREPROCESSING



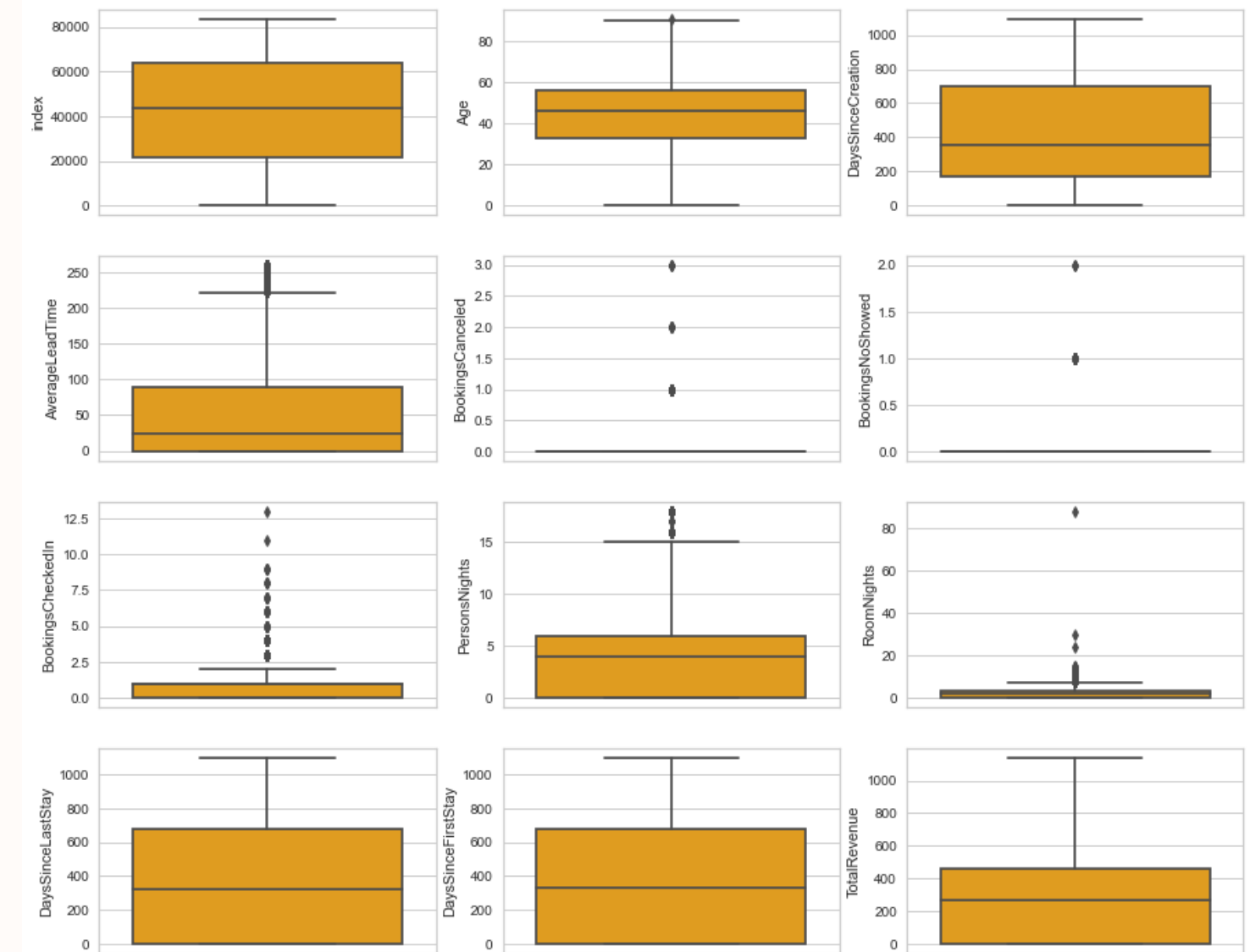
HEATMAP
CORRELATION

Drop columns who have low correlation

DATA PREPROCESSING



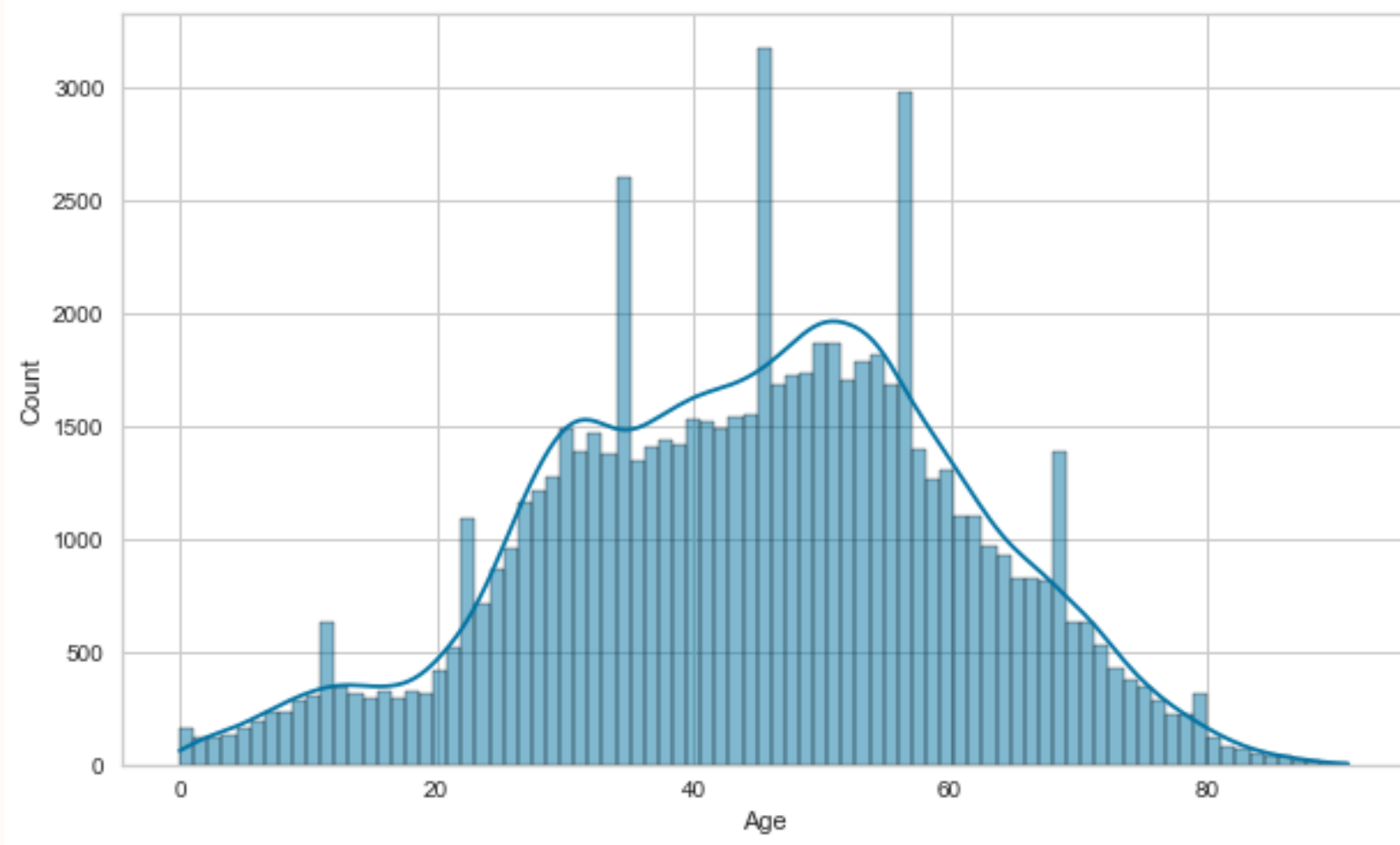
Before



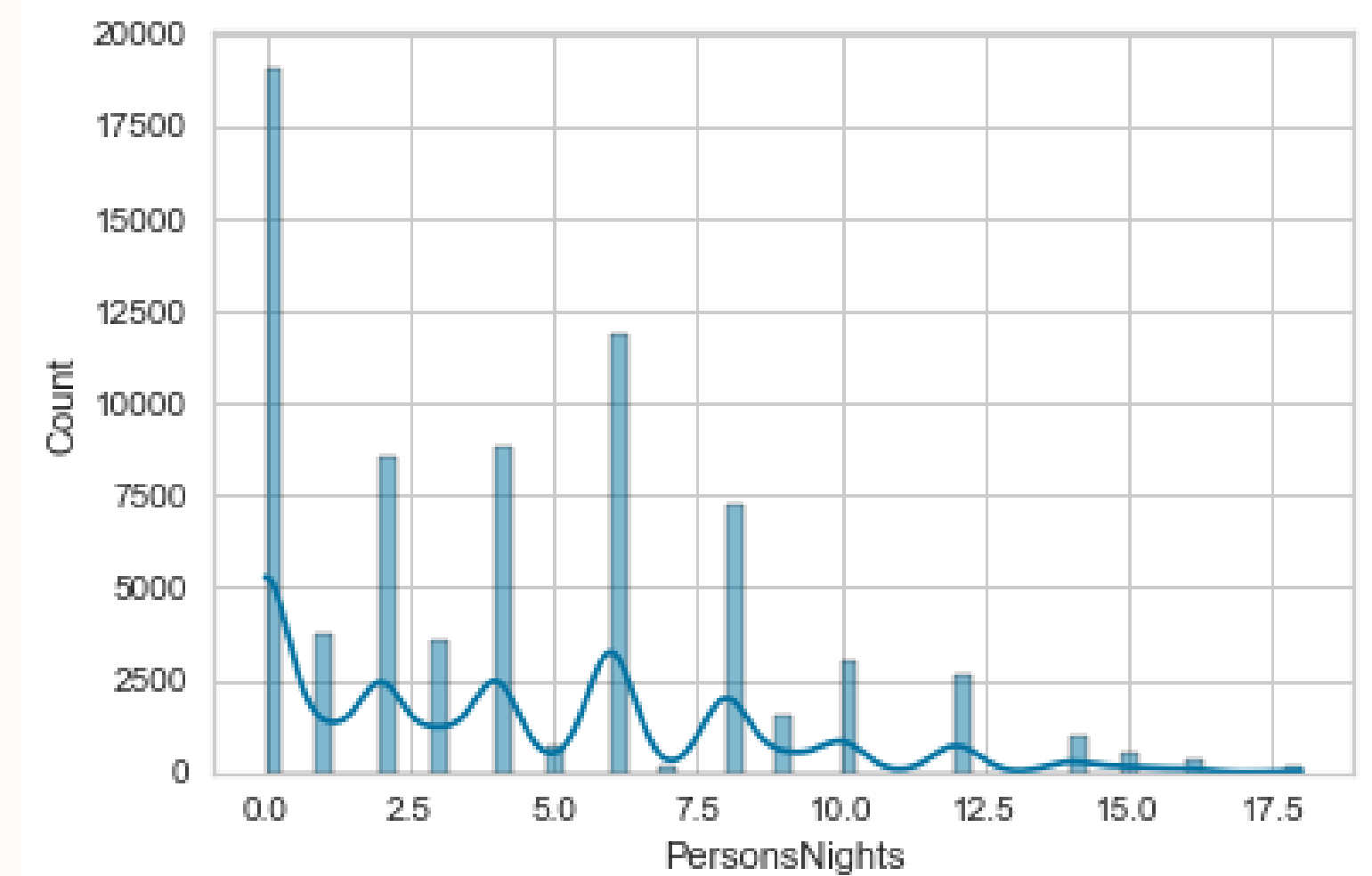
After

Total of rows before filtering outliers = 79811
Total of rows after filtering outliers = 72907

EXPLANATORY DATA

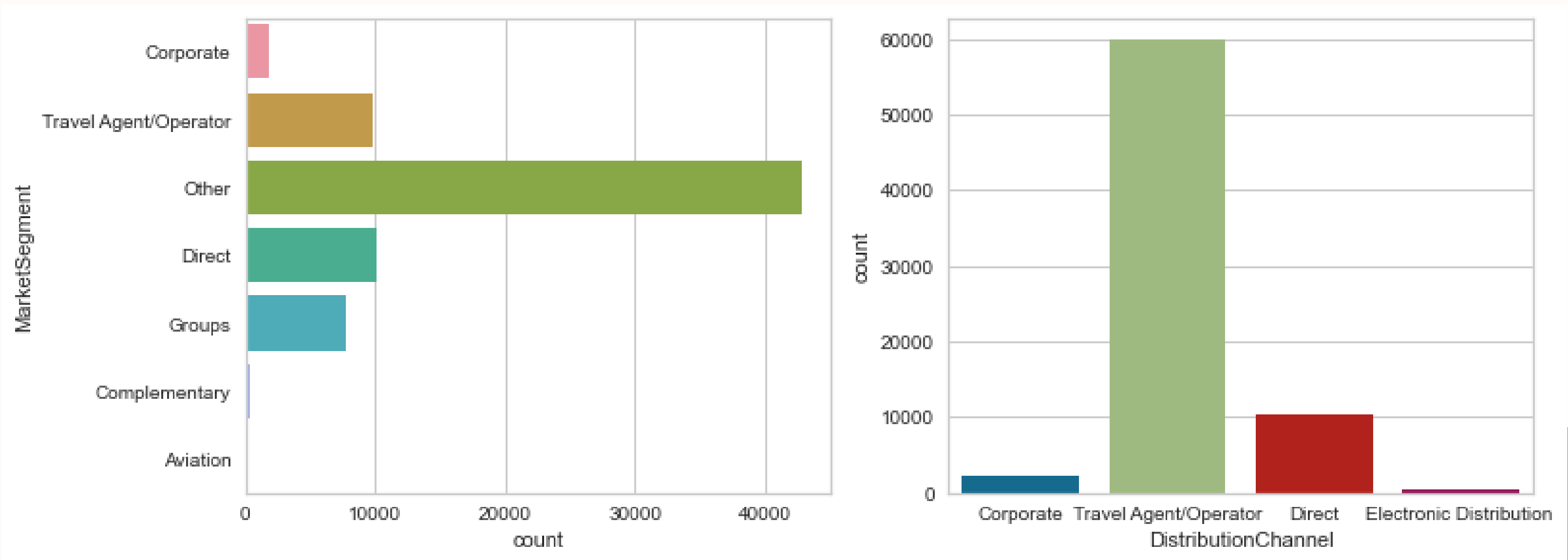


Distribution of Age



Distribution of Total Person Nights

EXPLANATORY DATA

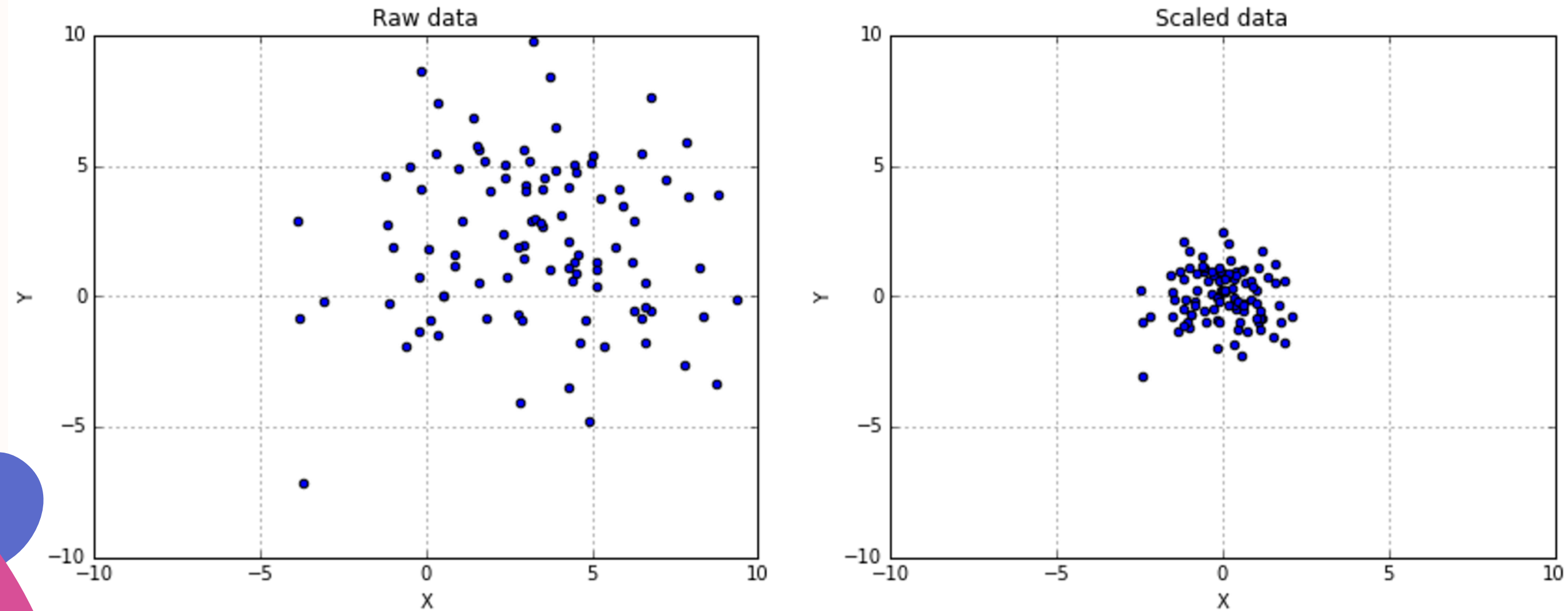


Distribution of Market Segment and Distribution channel. as we can see, this hotel lacks marketing in the corporate segment and sales through electronic media. maybe we can give a recommendation business in the end slide.

FEATURE SELECTION

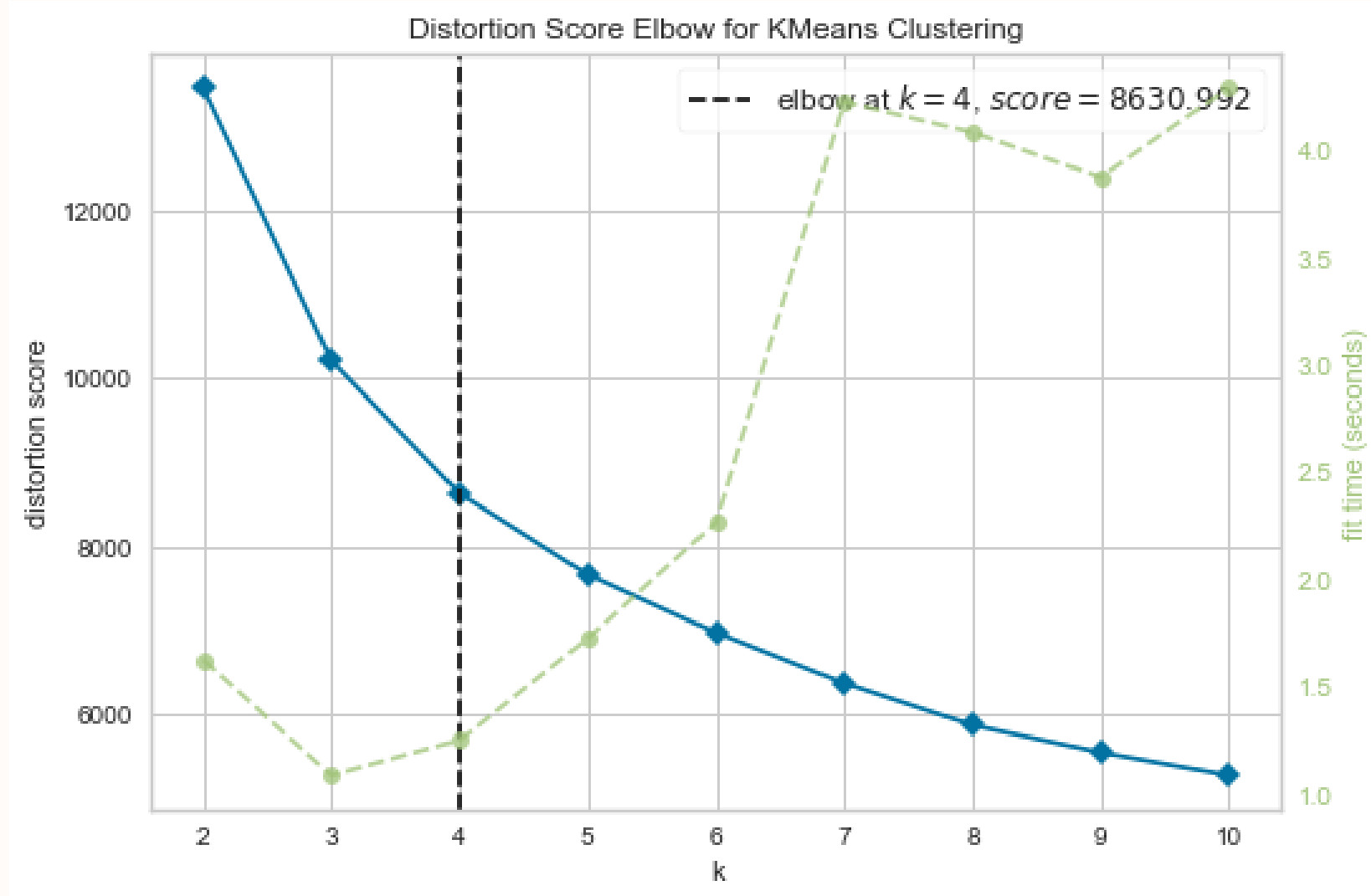
Age	72907	float64
DaysSinceCreation	72907	int64
AverageLeadTime	72907	int64
PersonsNights	72907	int64
TotalRevenue	72907	float64

NORMALIZATION DATA

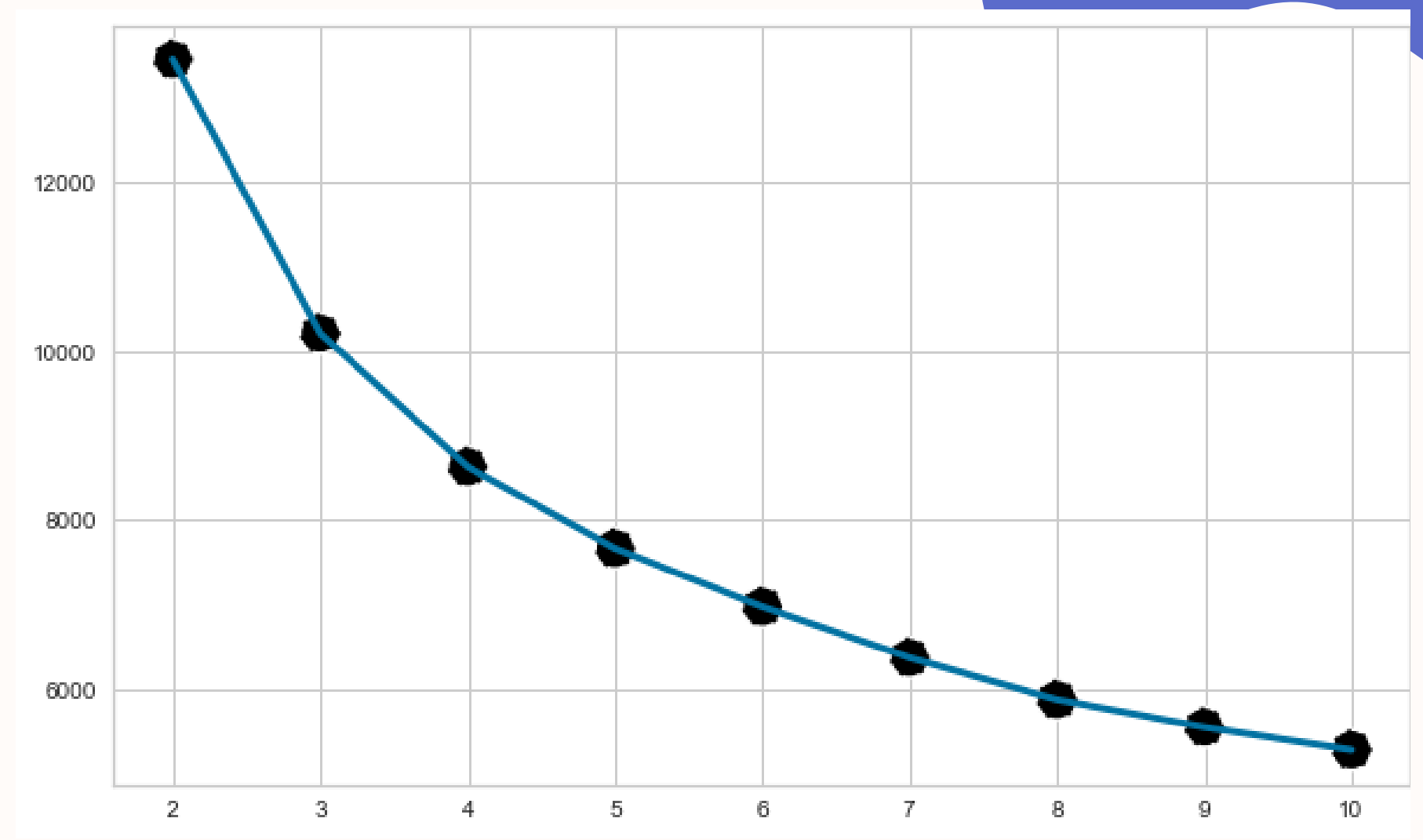


using `MinMaxScaler`

HYPERPARAMETER TUNNING



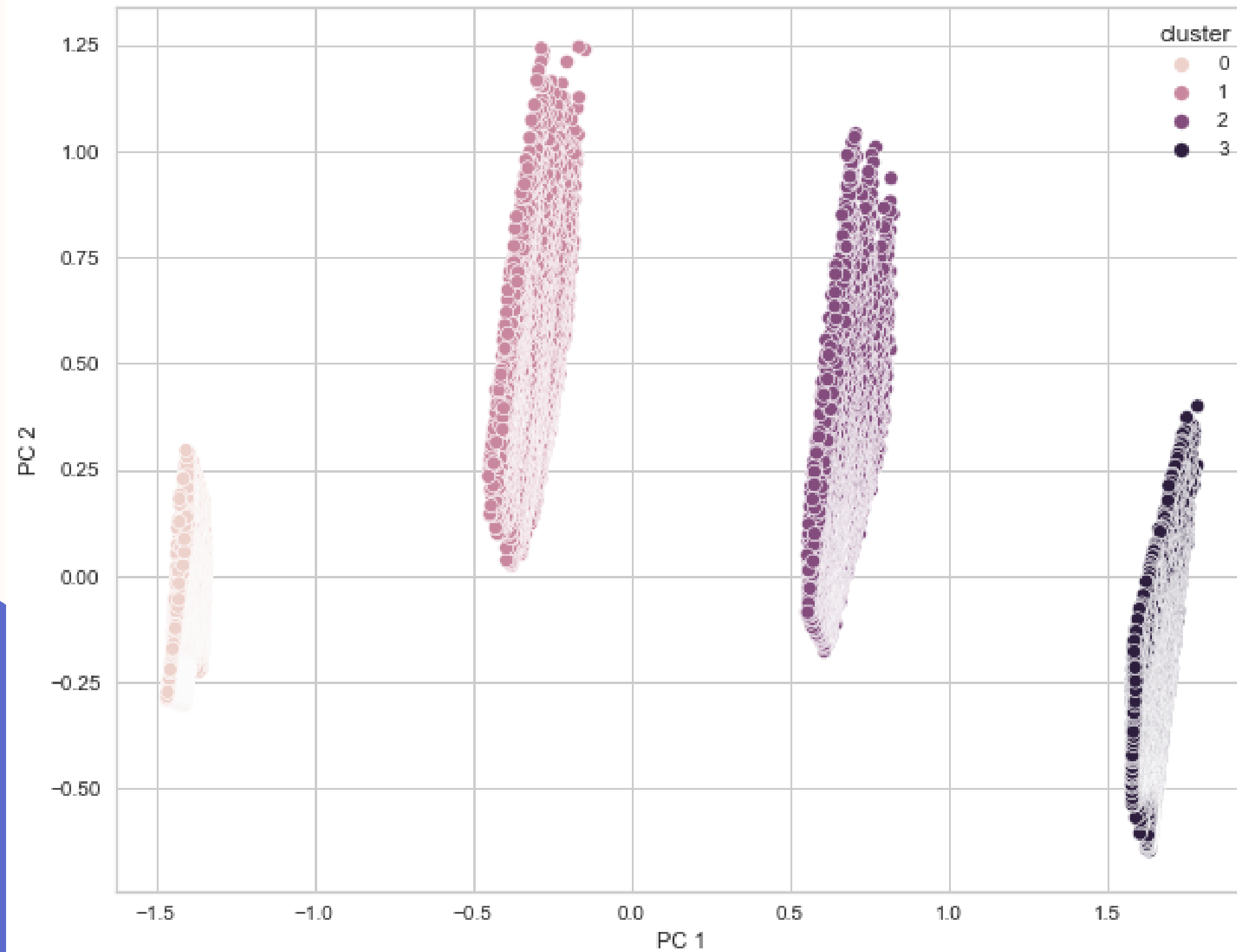
Yellowbrick method



Elbow Plot

x for total cluster
y for total error
we can choose $K=4$ for `n_clusters_`

RESULT



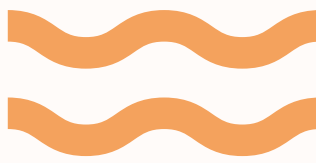
- Modelling using K-Means Clustering
- Reduction PCA for visualization 2 dimension

INSIGHT

	Age	Days Since Creation	Average Lead Time	Persons Nights	Total Revenue
	mean	mean	mean	mean	mean
cluster					
0	40.578397	193.626913	8.235495	0.757082	65.115519
1	49.893158	487.818873	174.140225	6.462568	421.012747
2	46.355814	354.904268	55.517755	8.051899	612.276833
3	46.039212	831.553850	37.935138	4.479737	290.420243

- Cluster 0 (solo traveler): a customer with an average age of 41 years, registered as a hotel customer approximately 6 months ago, this customer booked a hotel close to the day before he checked in with an average of 8 days before arrival, only 1 person who stays, and spends up to \$65.
- Cluster 1: customers with an average age of 50 years, have been registered for almost 1.5 years, these customers book hotel rooms 6 months before arrival, with an average of 6 people staying, while spending while at the hotel is \$421 .
- - Cluster 2: with an average age of 46 years, registered for 1 year, these customers booked hotels an average of 2 months before arrival with 8 people staying and spent an average of \$612.
- Cluster 3: is a customer with the age of 46 years, who has been registered for a long time, namely more than 2 years, and booked a hotel room a month before check-in, with an average of 4 people staying, and spent \$290.

RECOMMENDATION BUSINESS



The hotel can provide special prices for each agent who brings guests to stay at the hotel



The hotel can work with travellers to provide hotel stay packages that include accommodation, transportation and meals.



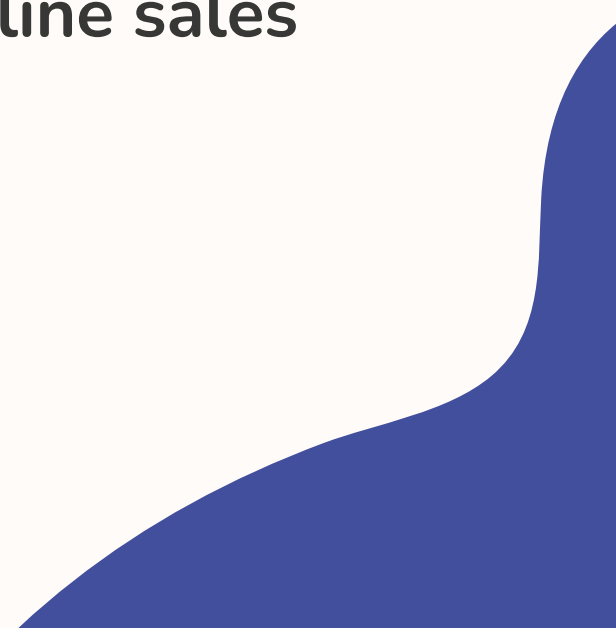
The hotel can work with companies for business expansion with cheap packages for half/full day meetings.



Always update hotel rooms or other stay packages on the website, so that online sales increase. You can also add it to other platforms.

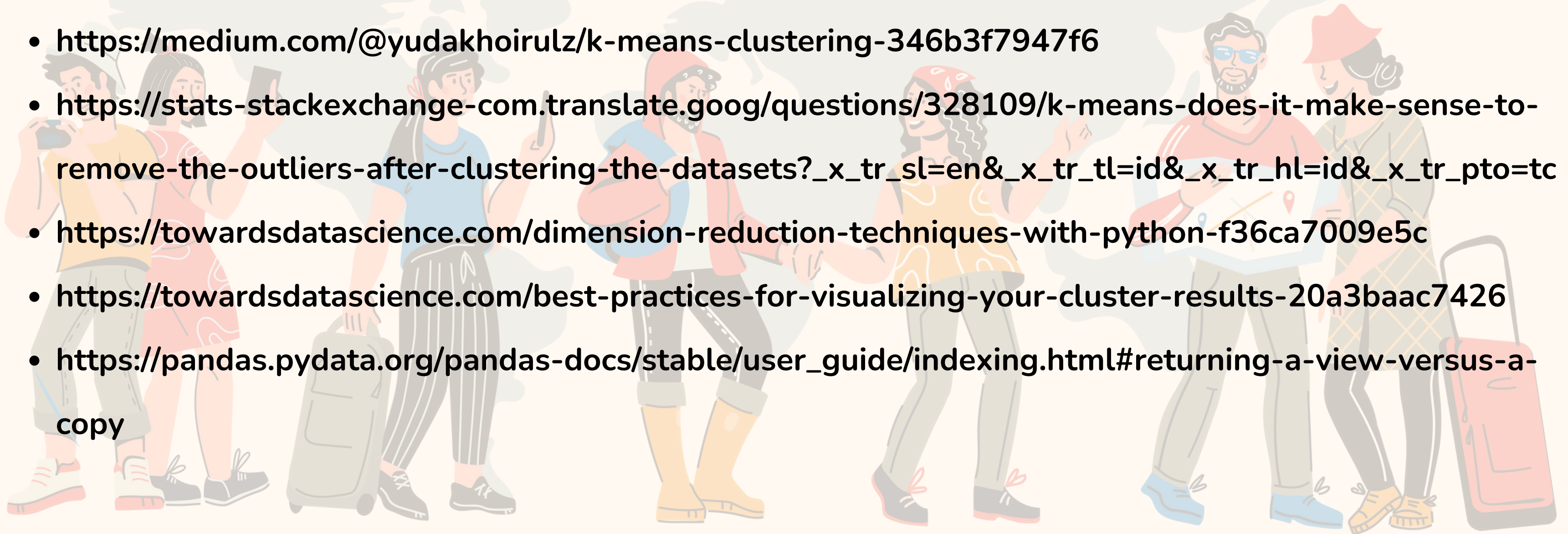


For every transaction above 500 dollars, the hotel can give free showcase or merchandise to customer.



REFERENCE

- <https://medium.com/@yudakhoirulz/k-means-clustering-346b3f7947f6>
- https://stats.stackexchange-com.translate.goog/questions/328109/k-means-does-it-make-sense-to-remove-the-outliers-after-clustering-the-datasets?_x_tr_sl=en&_x_tr_tl=id&_x_tr_hl=id&_x_tr_pto=tc
- <https://towardsdatascience.com/dimension-reduction-techniques-with-python-f36ca7009e5c>
- <https://towardsdatascience.com/best-practices-for-visualizing-your-cluster-results-20a3baac7426>
- https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy





THANK'S AND FEEL FREE TO CONECT



github.com/cfitrianty/



linkedin.com/in/camelia-fitrianty-12b340210/



cfitrianty@gmail.com

