



## **Customer Retention (Analytics based project)**

Submitted by:

Amit Kumar Tiwari

### Index

1. Acknowledgement
2. Introduction
  - a) Problem Statement
  - b) Data
3. Project life cycle
  - a) Pre-Processing
    - I. Null/missing value removal
    - II. Duplicate row & column analysis & removal
    - III. Outlier analysis & removal
    - IV. Data Distribution check-up & visualization
4. Conclusion

## ACKNOWLEDGMENT

Here are a few of the resources which I referred during the project implementation cycle. I used TDS, TAI, Analytics Vidya along with the knowledge that I acquired from during my curriculum.

1. For data visualization –

<https://towardsai.net/ai/data-visualization>  
<https://klib.readthedocs.io/en/latest/>  
<https://towardsdatascience.com/speed-up-your-exploratory-data-analysis-with-pandas-profiling-88b33dc53625>

## INTRODUCTION

- Problem Statement

E-retail factors for customer activation and retention: A case study from Indian e-commerce customers

Customer satisfaction has emerged as one of the most important factors that guarantee the success of online store; it has been posited as a key stimulant of purchase, repurchase intentions and customer loyalty. A comprehensive review of the literature, theories and models have been carried out to propose the models for customer activation and customer retention. Five major factors that contributed to the success of an e-commerce store have been identified as: service quality, system quality, information quality, trust, and net benefit. The research furthermore investigated the factors that influence the online customers repeat purchase intention. The combination of both utilitarian value and hedonistic values are needed to affect the repeat purchase intention (loyalty) positively. The data is collected from the Indian online shoppers. Results indicate the e-retail success factors, which are very much critical for customer satisfaction.

- Data

Data was provided in CSV file. The data format is as follows:

In [27]: `import pandas as pd`  
`Data_Day = pd.read_excel(r'C:\Users\Admin\Documents\customer_retention_dataset.xlsx')`  
`Data_Day.head()`

Out[27]:

	1 Gender of respondent	2 How old are you?	3 Which city do you shop online from?	4 What is the Pin Code of where you shop online from?	5 Since How Long You are Shopping Online ?	6 How many times you have made an online purchase in the past 1 year?	7 How do you access the internet while shopping on-line?	8 Which device do you use to access the online shopping?	9 What is the screen size of your mobile device?	10 What is the operating system (OS) of your device?	...	Longer time to get logged in (promotion, sales period)	Longer time in displaying graphics and photos (promotion, sales period)	Late declaration of price (promotion, sales period)	Longer loading (prom sales pt
0	0	3	Delhi	110009	5	4	4	3	5	1	...	Amazon.in	Amazon.in	Flipkart.com	Flipkar
1	1	2	Delhi	110030	5	5	2	1	2	3	...	Amazon.in, Flipkart.com	Myntra.com	snapdeal.com	Snapdea
2	1	2	Greater Noida	201308	4	5	3	1	4	2	...	Myntra.com	Myntra.com	Myntra.com	Myntra
3	0	2	Karnal	132001	4	1	3	1	4	3	...	Snapdeal.com	Myntra.com, Snapdeal.com	Myntra.com	Paytm
4	1	2	Bangalore	530068	3	2	2	1	2	3	...	Flipkart.com, Paytm.com	Paytm.com	Paytm.com	Paytm

5 rows x 71 columns

Here, on having a look into the dataset we found that there are 71 variables.

Column names are pretty much self-explainable.

After treating the dates column, we found the following datatypes in various variables.

## Project life cycle

- Data pre-processing

Data pre-processing is an essential step before we go for the model development & training process. It is very important because the model will be performing in the we have trained. If we have trained our model with false data, it's results would also be faulty because model have learned something wrong due to false pattern of the data that it has created into it.

**Null value removal/imputation** – Null value removal is a process of removing the values where the data is null. We may also fill the null values with imputation process.

**Duplicate row & column removal** – Here, I used pandas profiling for the for duplicate column & row visualization.

**Data type check & conversion** – Here, I converted the existing data types to more suitable data types before further analysis & pre-processing.


**Data distribution check-up & scaling** – Here, I used pandas profiling for plotting the data distribution & found some of the data are not normal based on the characteristics of the variables & can be scaled down so that data distribution is normal. Let's jump into the visualization report now

**Visualization report:** The link given here can be used to check visualization report.

<http://localhost:8888/view/Documents/Customer%20Retention.html>

I'll give you a guidance as how to read the report to assess the dataset in a glimpse

```
In [11]: import pandas_profiling
report = pandas_profiling.ProfileReport(Data_Day)
report
```



The output shows three progress bars for the pandas\_profiling tasks:

Task	Progress	Details
Summarize dataset	100%	85/85 [00:47<00:00, 1.67s/it, Completed]
Generate report structure	100%	1/1 [00:29<00:00, 29.71s/it]
Render HTML	100%	1/1 [00:05<00:00, 5.35s/it]

Running this single line of code will create an HTML EDA report of your data. The code displayed above will create an inline output of the result; however, you could also choose to save your EDA report as an HTML file to be able to share it more easily.

## Overview

Overview		Alerts 249	Reproduction
Dataset statistics		Variable types	
Number of variables	71	Categorical	70
Number of observations	269	Numeric	1
Missing cells	0		
Missing cells (%)	0.0%		
Duplicate rows	61		
Duplicate rows (%)	22.7%		
Total size in memory	149.3 KiB		
Average record size in memory	568.5 B		

The first part of the HTML EDA report will contain an overview section providing you with basic information (number of observations, number of variables, etc.). It will also output a list of warnings telling you were to double-check the data and potentially focus your cleaning efforts on. Since we don't have any missing/null values in the data set therefore it won't reflect here in this case.

Majorly it has three elements marked as below in the snapshot

## Overview

Overview	Alerts 249	Reproduction
Dataset statistics		Variable types
Number of variables	71	Categorical 70
Number of observations	269	Numeric 1
Missing cells	0	
Missing cells (%)	0.0%	
Duplicate rows	61	
Duplicate rows (%)	22.7%	
Total size in memory	149.3 KiB	
Average record size in memory	568.5 B	

**Overview** gives a descriptive analytics/insight about the data set as in the above snapshot.

**Alerts** gives a quick summary about each variables & the data point.

## Overview

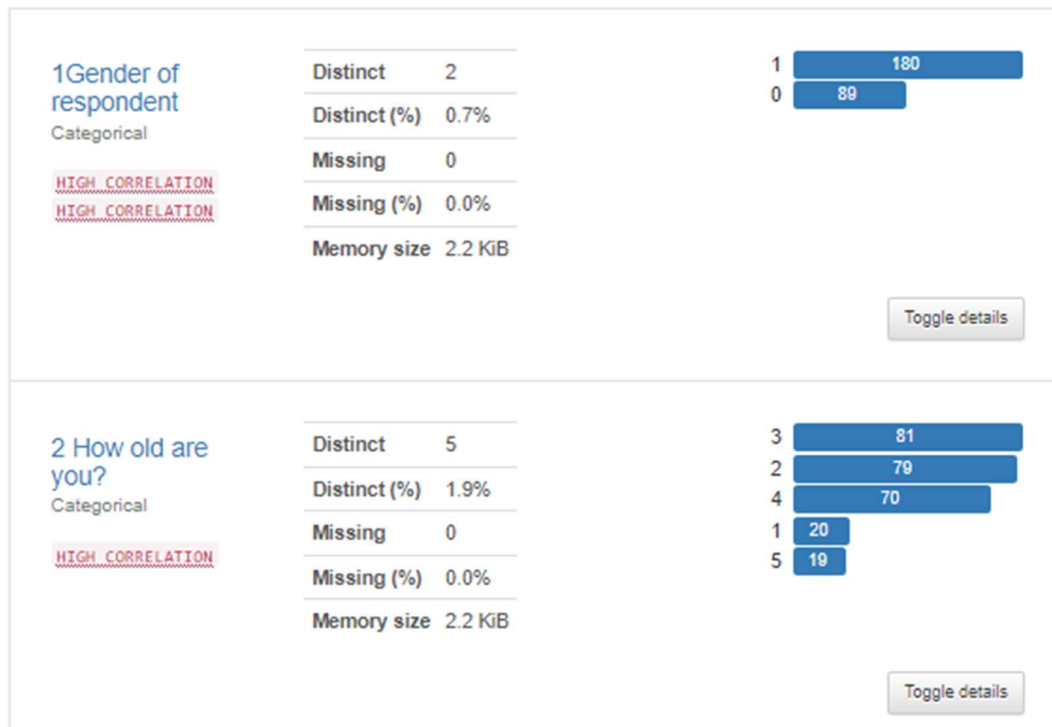
Overview	Alerts 249	Reproduction
Alerts		
Dataset has 61 (22.7%) duplicate rows		Duplicates
8 Which device do you use to access the online shopping? is highly correlated with 9 What is the screen size of your mobile device? and 5 other fields		High correlation
9 What is the screen size of your mobile device? is highly correlated with 8 Which device do you use to access the online shopping? and 4 other fields		High correlation
10 What is the operating system (OS) of your device? is highly correlated with 8 Which device do you use to access the online shopping? and 5 other fields		High correlation
11 What browser do you run on your device to access the website? is highly correlated with 10 What is the operating system (OS) of your device? and 4 other fields		High correlation
12 Which channel did you follow to arrive at your favorite online store for the first time? is highly correlated with 11 What browser do you run on your device to access the website? and 2 other fields		High correlation

**Reproduction** gives a quick summary about execution as well as variables. Overview is the summary of reproduction & alerts sections.

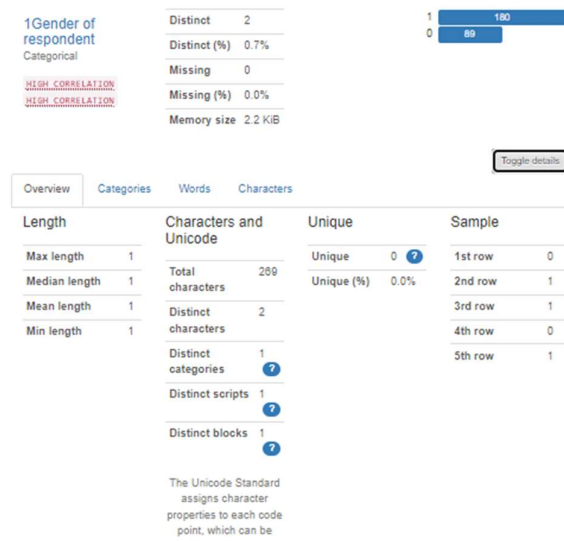
Overview	Alerts 249	Reproduction
Reproduction		
Analysis started	2021-12-31 06:34:01.055590	
Analysis finished	2021-12-31 06:34:48.121020	
Duration	47.07 seconds	
Software version	pandas-profiling v3.1.1	
Download configuration	config.json	

**Variable-Specific EDA** - Following the overview, the EDA report provides you with helpful insights for each specific variable. These also include a small visualization describing the distribution of each variable:

# Variables



Now upon toggling the data in the HTML files will give a further key insight of that variable.



Remark -I'm again posting the link to the HTML report for insights. Report is very self-explanatory.

<http://localhost:8888/view/Documents/Customer%20Retention.html>

## **CONCLUSION**

- **Learning Outcomes of the Study in respect of Data Science**

The project is good as far as the learning of the data cleaning or more precisely working with numerical data.

- **Limitations of this work and Scope for Future Work**

This is very specific to the insight of the review's forms filled by various users. This very project is very specific to the provide analytics using data that we must take immediate actions but can not predict what is going to happen next. Also, it is talking about where the sites are lacking heavily to provide a better user experience.