Kumanan G
PGP - DSBA Online 2020
6 March 2021

Tourism Package

EDA and Business Insights

Introduction of the business problem

Problem Statement

The tourism company has plans to launch its new long term tourism package. The product manager wants to analyse the existing data of its customers to find out who are likely to buy the new long term tourism package.

Need of the Study

As with any business decision, analytics can add lots of values in terms of what would a company can make out of its new idea in this case tourism package. Globalisation and internet access changed the interaction between travellers and agency. The internet made almost all information available to all. Which is a good thing to customers but it may not be for travel agencies. Also the internet bargains made the travel companies to provide highly competitive priced package. Because there is always someone trying hard to earn more customers by giving attractive offers and benefits. This makes the concept "loyal customers" an obsolete. So it is important for a tourism company to know what type of customers are having business with them and what do they expect. This knowledge can make the company to provide a meaningful tourism package which will attract more customers and ultimately benefit the business.

Understanding business opportunity

Analysing past customer's data can give us solutions to the problem a business is facing. Analytics can tell us which segment of customers travels more and their expectation. If a company knows that information it cut down the marketing cost by concentrating on its targeted potential customers. Analytics can also tell us whether the current business strategy is working out or not. So analytics makes business more targeted and eliminates what a company doesn't need.

Data Report

Understanding Data Collection

The dataset has information about the customers and also the salesman's input about the individual customer. The data is collected over a month's period with customer's historical travel data and their preferences.

Visual inspection of data

The dataset has 20 different attributes of 4888 customers (4888 rows and 20 columns) . Of 20 variables 6 of them are in object type (categorical). Remaining 14 columns are in either integer or in float types in other words numbers. Every customer is identified with unique customer ID. No two customers share a common customer ID.

Understanding of attributes

The dataset has the information about customers like their age, gender, marital status, monthly income, occupation, city tier they live in, and salesman's perspective like **Pitch satisfaction score** which ranges from 1 to 5. The scores were given by the sales man who pitched the different products to the customer while the enquiry.

Prod pitched is another attribute which tells us about what tourism package has been pitched to the customer. There are 4 products available to choose namely Standard, King, Deluxe, Super Deluxe. But there are many customers asked about multiple products. Also we don't have information on which product the customers finally chose.

Number of follow-ups is one of the 20 variable which tells us about how many times the salesman has reached out to the customer after the sales pitch.

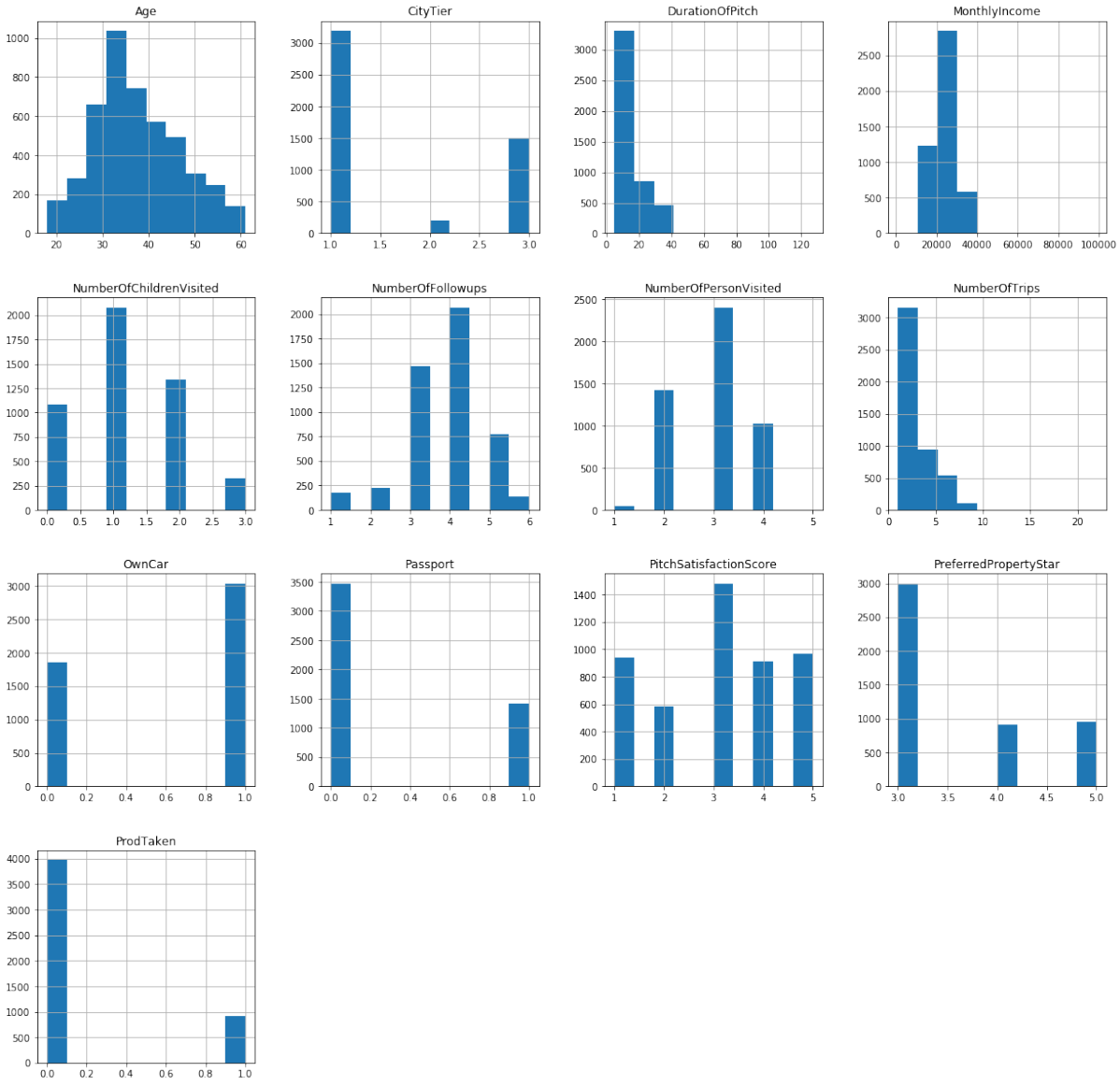
Preferred property star has values from 3 to 5. We can assume that higher the preference means higher the luxury level.

Prod taken is the Target variable to work on further with Machine Learning models. The target variable has two values 0 and 1. 0 means the customer did not purchase the tourism package. 1 represents successful purchase of the package.

Preferred login device is the variable which tells us about whether the customer self-enquired about products or Company invited.

All the attributes's names have mix of both upper and lower cases. So all the names have changed to small case for easier access. And Gender attribute have 3 categories which is due to spelling error. There were two female categories and those needed renaming to make as one female category.

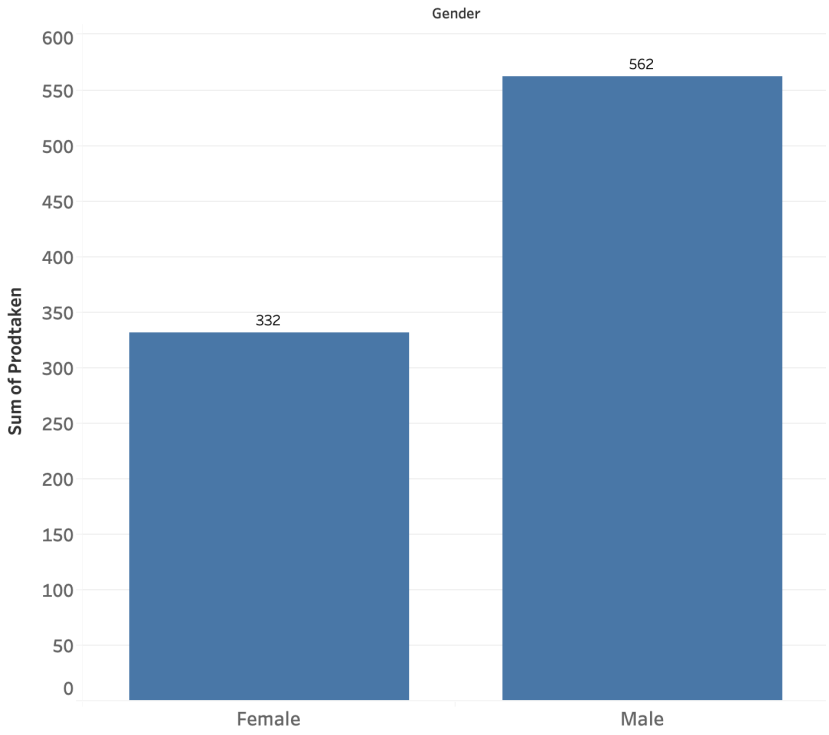
Exploratory Data Analysis



Insights :

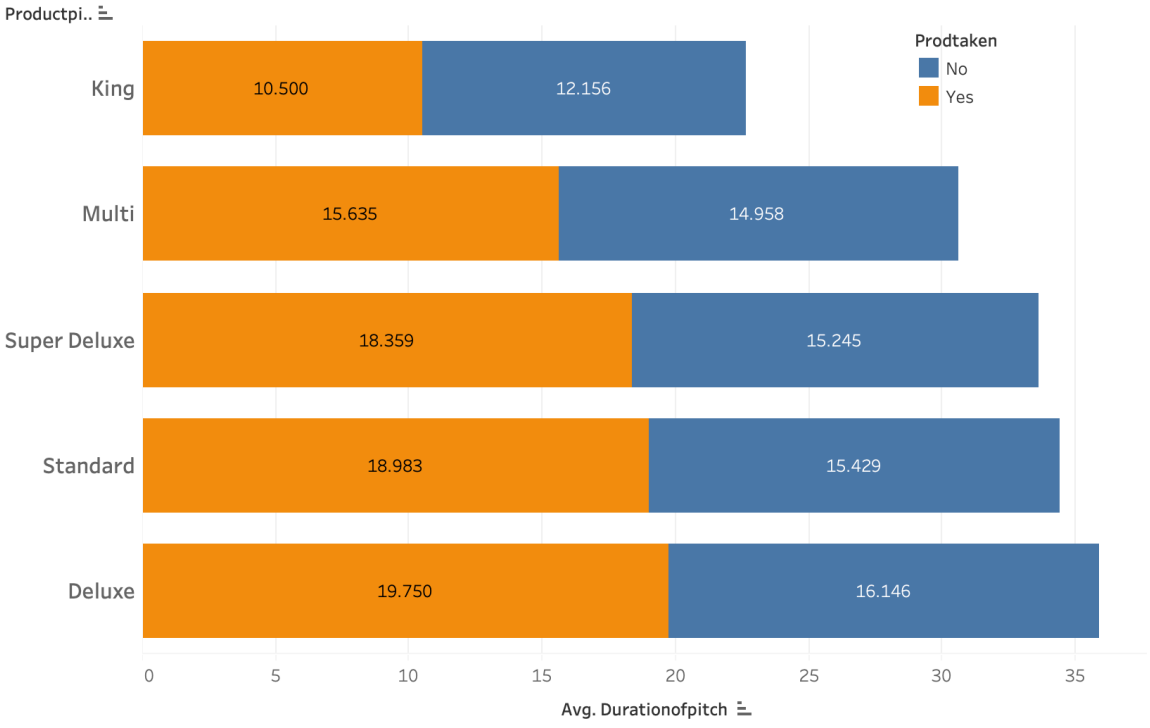
1. Age : Most of sample distribution has age between 25 and 45.
2. City tier : more than 3000 people are from Tier 1 cities which is little over 60% , and 1500 people are from tier 3 city.
3. Duration of Pitch : Mean duration pitch is 15 minutes per customer. Couple of customers got pitched for 2 hours.
4. Monthly Income: Average income is 23619 it's because most of the customers are executive level employees who tend to get lower salary than VPs and AVPs.
5. Number of children visited: out of 4888 customers around 1000 of them did not brought children with them during the sales pitch. Most people brought at least 1 child with them.
6. Number of follow-ups: Over 90% of customers received 3 or more times of follow-ups regarding their package enquiry from salesman.
7. Own Car : Over 80% of the customers owns a car.
8. Pitch Satisfaction score : This score is given by salesman themselves. For 1500 customers the score is 1 and 2 out of 5 which considerably low. Most of the scores are 3 to 5.
9. Preferred Property : 3000 customers prefers 3 star rated accommodation which is around 61% of customers. Remaining prefers 4 and 5 star accommodations.

prodtaken by gender

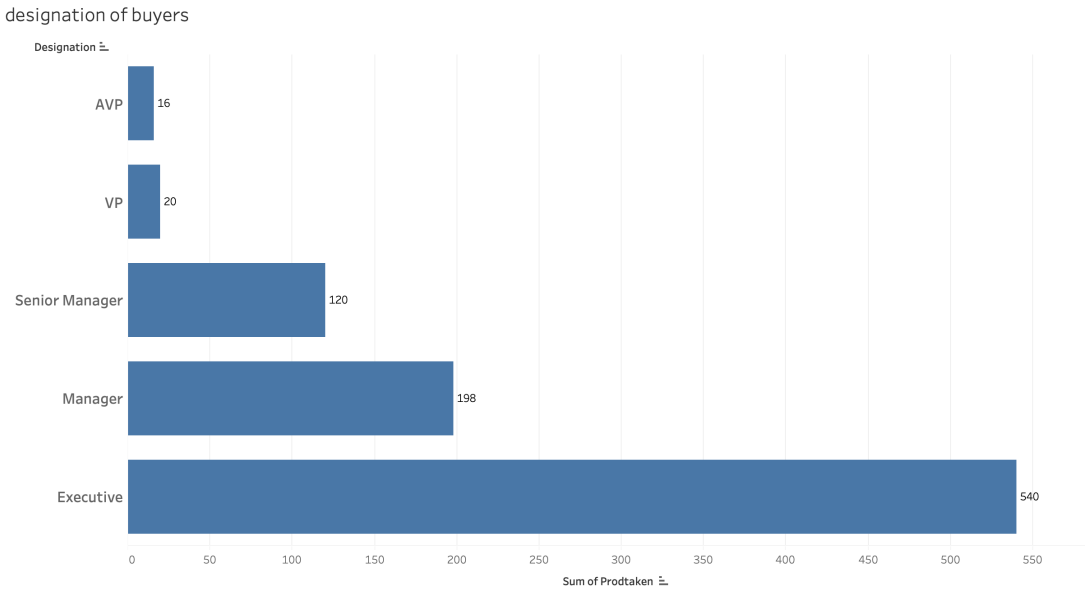


Out of 4888 customers 894 chose a tour package. And 562 of them are male and 332 are female.

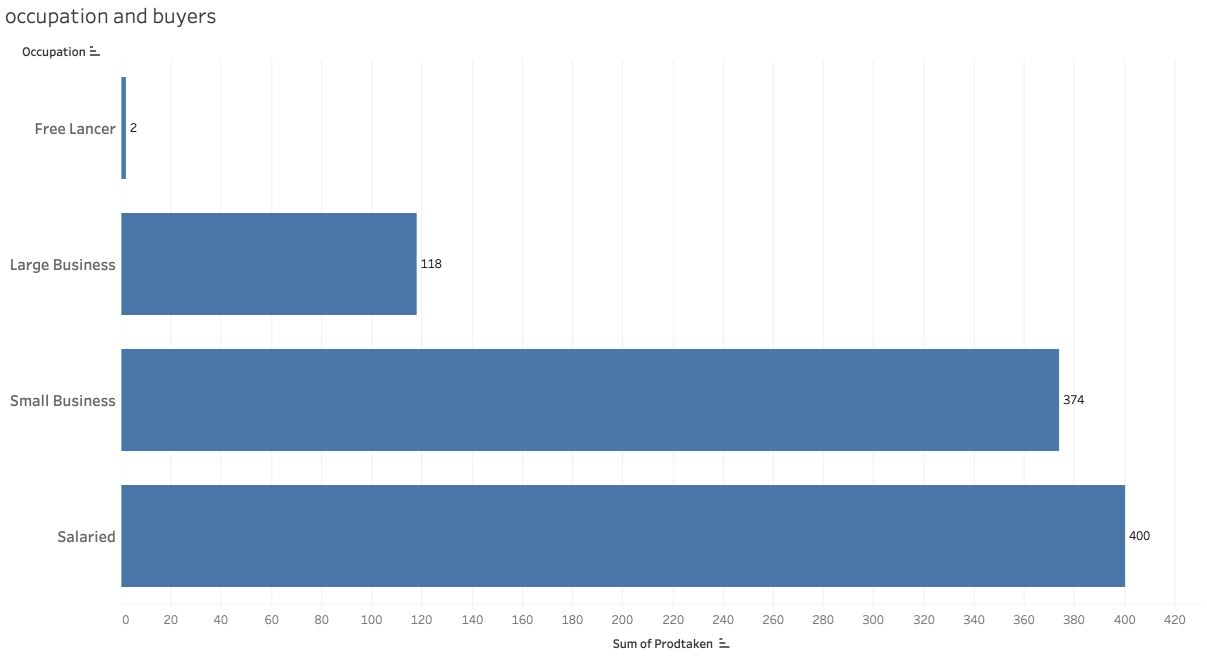
products and pitch duration

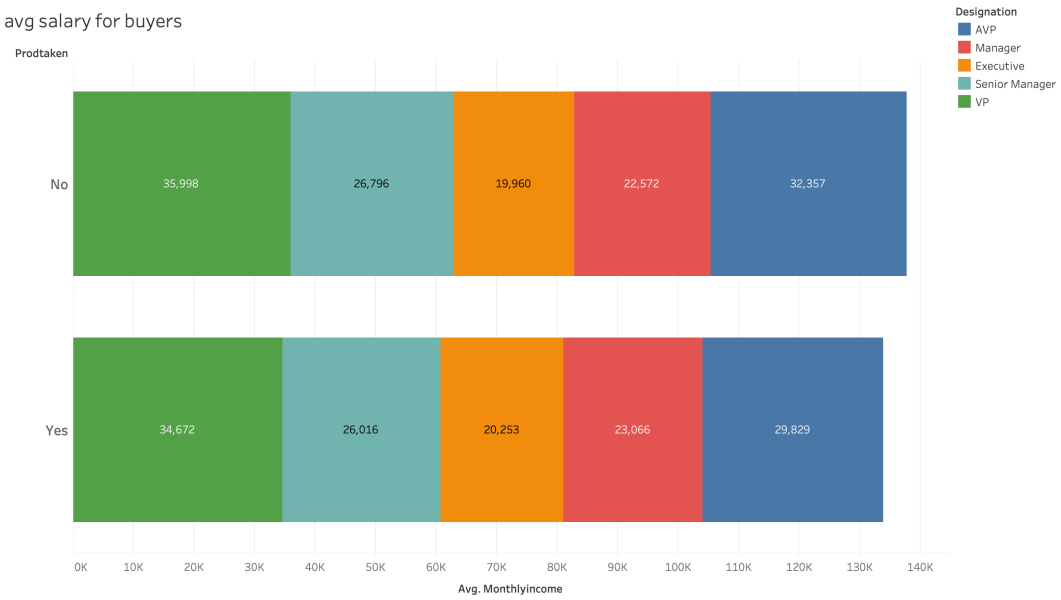


Deluxe package got pitched more than any other package. And king is the lowest amount of time pitched to the customers. And the average time of pitch which also turned into a successful conversion for Deluxe is 19.7 minutes.

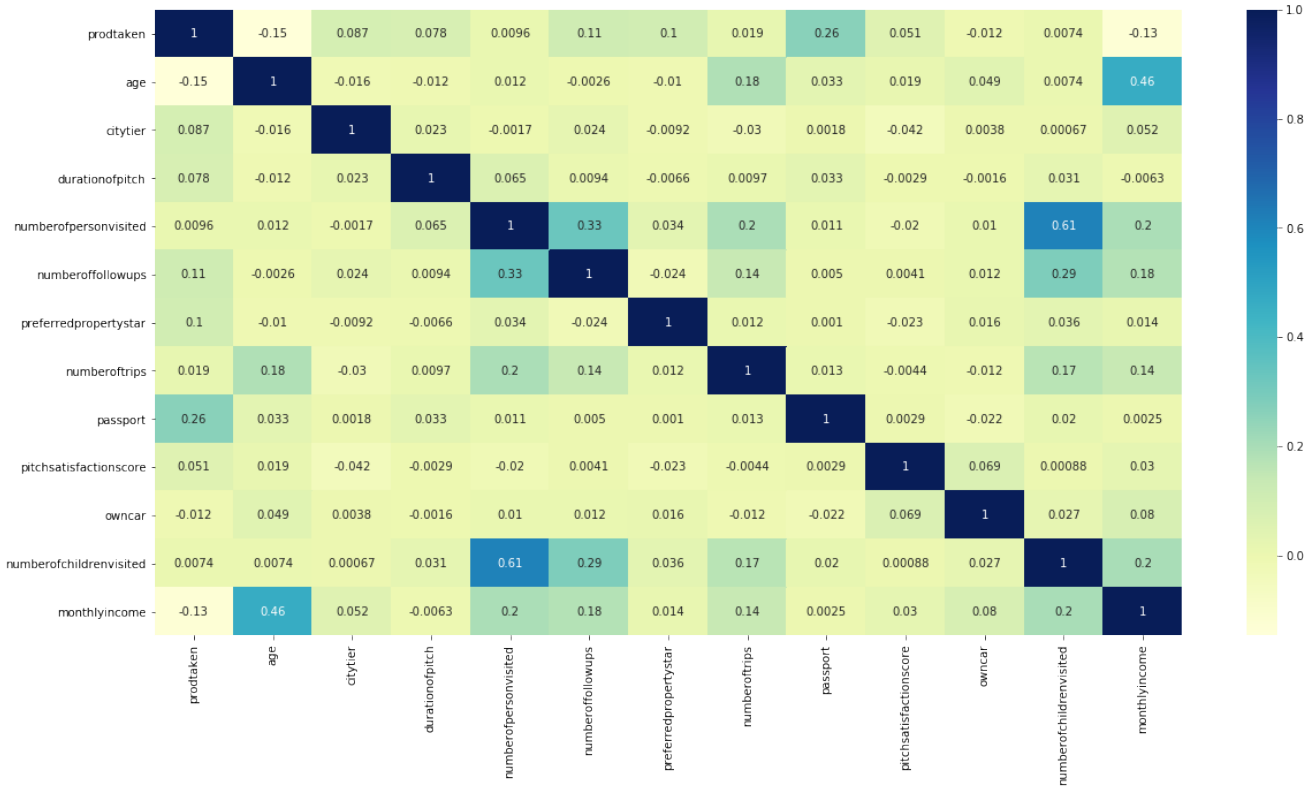


It is very clear that from above chart that most of the customers who bought the package are executive level employees followed by managers. Out of 894 customers 540 of them are executives. 400 of them are salaried and 374 of them who does small business.





Below Heatmap shows the correlation between the variables. Darker the shade more the correlation. Highly correlated attributes are not highly useful in further analysis. So it is important to identify them before proceeding with model building. But from the correlation analysis there are no variables highly correlated with each other apart from number of person visited and number of children visited. The correlation between them is natural and understandable.



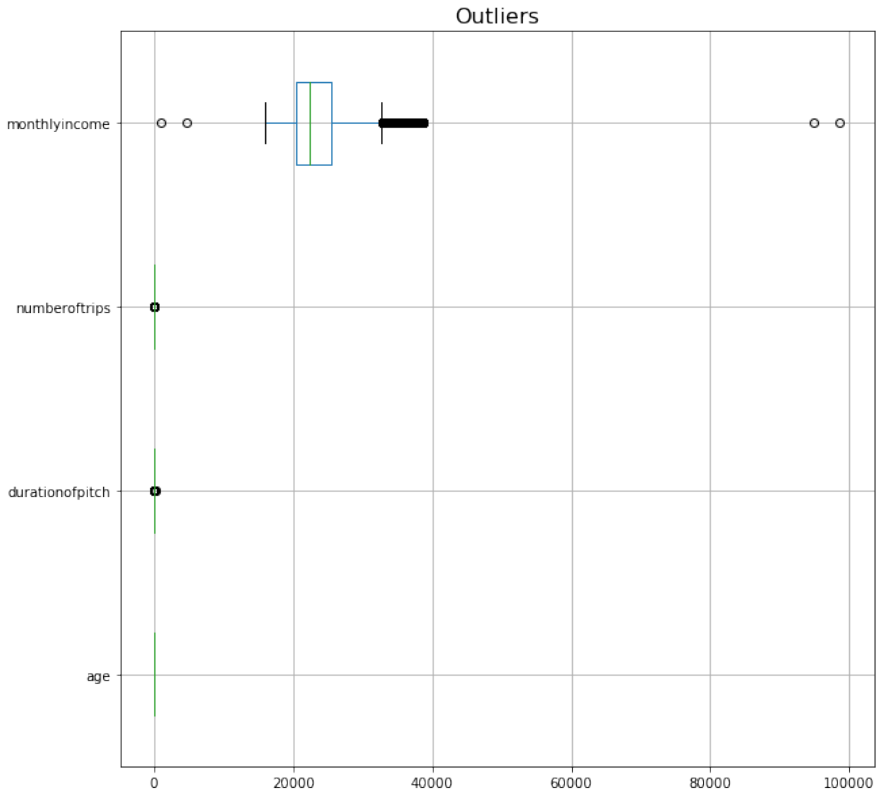
Removal of unwanted variables

From the data reading and EDA, it is found that customer ID will not be useful in further analysis

Duplicates and Missing Value treatment

There are 141 duplicated rows found in the dataset and same has been removed. In total 980 missing values found among different variables. Except Preferred login device every else missing value has been replaced with their respective median values. For example there were 224 missing values found in monthly income column and those replaced with median value of 22347. Since preferred login device is categorical attribute it is replaced with most occurring value which is self enquiry.

Outlier treatment



Box plot helps us to find outliers in the dataset. Monthly income has several outliers and it needs to be fixed. Those outliers were capped with IQR (InterQuantile Range) values.

Variable transformation

Several categorical variables (prod taken, city tier, passport, own car, satisfaction scores and etc.) are in either integer or float data type due to they are numerical in nature. So they converted to object data type.

Since the values in different variables are in different magnitudes. For example monthly incomes are in 1000s and all other variables are range 1 to 5. So these differences needs to be fixed. Scaling is the method which makes the dataset scaled as equally. Scikit's standard scaler is used to transform the data.

Business Insights from EDA

Data Balancing

Out of 4747 customers (after removing duplicates), only 894 of them (just 18.8% of customers) actually purchased the product which is a heavily skewed data to work on. To make the data more useful it needs to be balanced. And there are many balancing techniques available. One of them is SMOTE (Synthetic Minority Oversampling Technique). This technique makes or synthesises new examples (rows) on minority class in our case customers who bought the product.

Business Insights

Apart from insights shared previously shared in this document, python analysis gave more insights to share,

More follow-ups yields more business, 842 customers who bought the product have received 3 or more follow-ups which is 94% of the customers.

High pitch score represents high successful conversion. 25 % of customers bought the product when they received the scores between 3 and 5. While low scores yields 17% of the customers.

The mean duration of pitch is 15 minutes. 22.5% of People who got pitched above mean pitch time bought the product. While 15.9 % of people who got pitched below mean pitch time bought the product. So pitch duration is an indicator of successful conversion.

Owning a car is not an indicator of potential customer since 19% of people doesn't own a car but bought the product. 18% of people who does own a car also bought the product.

Thank You !!!