**HOLIDAY PACKAGE PREDICTION**

***ABOUT THE PROJECT-***

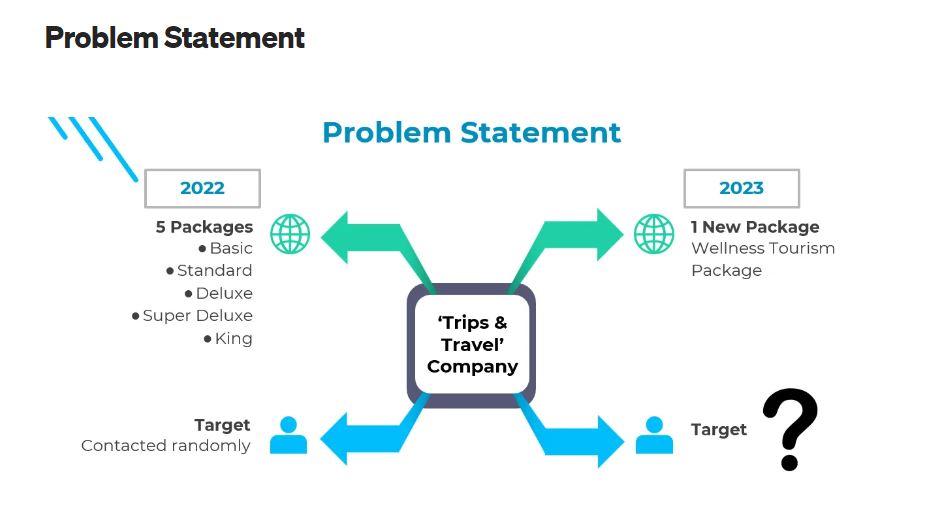
‘’Trips & Travel.Com" a U.S based travel company wants to enable and establish a viable business model to expand the customer base. One of the ways to expand the customer base is to introduce a new offering of packages.

Currently, there are 5 types of packages the company is offering - Basic, Standard, Deluxe, Super Deluxe, and King. Looking at the data of the last year, we observed that 18% of the customers purchased the packages.

However, the marketing cost was quite high because customers were contacted at random without looking at the available information.

The company is now planning to launch a new product i.e. Wellness Tourism Package. Wellness Tourism is defined as Travel that allows the traveller to maintain, enhance or kick-start a healthy lifestyle, and support or increase one's sense of well-being.

However, this time company wants to harness the available data of existing and potential customers to make the marketing expenditure more efficient.



OBJECTIVES-

* Create a system that can help companies predict which customer is more likely to purchase the newly introduced travel package.
* Knowing the characteristics of potential customers to target for new product offerings in order to a) increase conversion rates and reduce marketing costs and b) find out who will buy vacation packages to find new customers.



ABOUT THE INDUSTRY-

***EXPLORATORY DATA ANALYSIS***

DATA DICTIONARY-

The dataset used is a csv file originating from [Kaggle](https://www.kaggle.com/susant4learning/holiday-package-purchase-prediction). The dataset consists of 4888 rows and 20 columns with the following attributes.



***CustomerID****: Unique customer ID*

***ProdTaken****: Product taken or not (0: No, 1: Yes)*

***Age****: Age of customer*

***TypeofContact****: How customer was contacted (Company Invited or Self Inquiry)*

***CityTier****: City tier depends on the development of a city, population, facilities, and living standards. The categories are ordered i.e.*

***DurationOfPitch****: Duration of the pitch by a salesperson to the customer*

***Occupation****: Occupation of customer*

***Gender****: Gender of customer*

***NumberOfPersonVisiting****: Total number of persons planning to take the trip with the customer*

***NumberOfFollowups****: Total number of follow-ups has been done by the salesperson after the sales pitch*

***ProductPitched****: Product pitched by the salesperson*

***PreferredPropertyStar****: Preferred hotel property rating by customer*

***MaritalStatus****: Marital status of customer*

***NumberOfTrips****: Average number of trips in a year by customer*

***Passport****: The customer has a passport or not (0: No, 1: Yes)*

***PitchSatisfactionScore****: Sales pitch satisfaction score*

***OwnCar****: Whether the customers own a car or not (0: No, 1: Yes)*

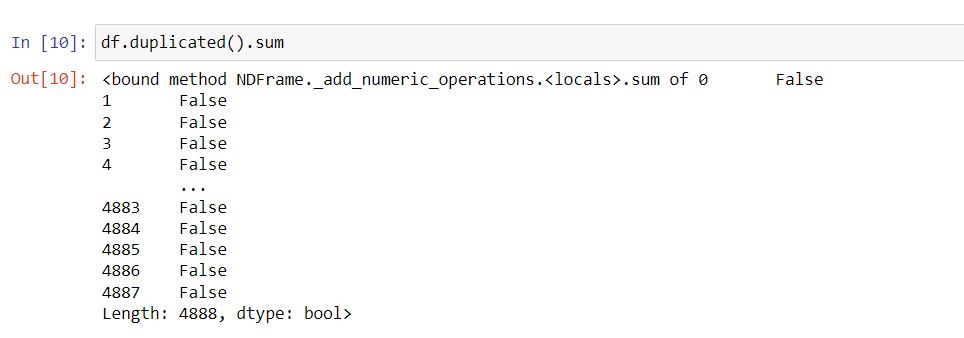
***NumberOfChildrenVisiting****: Total number of children with age less than 5 planning to take the trip with the customer*

***Designation****: Designation of the customer in the current organization*

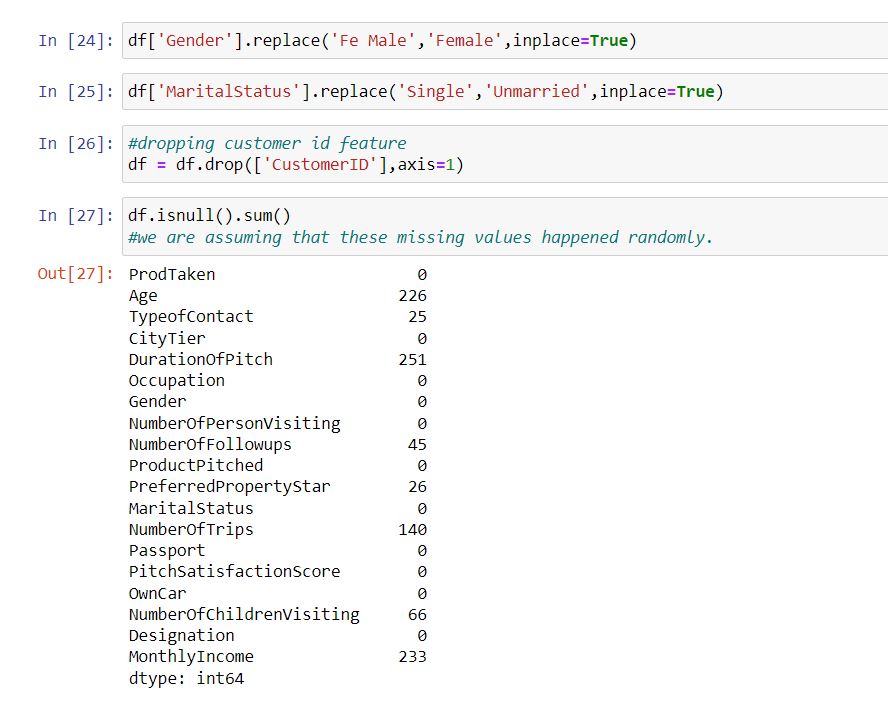
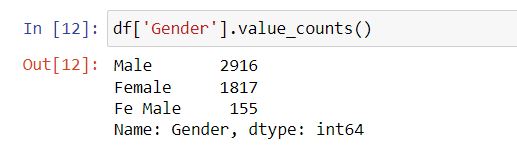
***MonthlyIncome****: Gross monthly income of the customer*

Of the 20 columns that will be selected as targets are columns ProdTaken while the other columns are used as features and will be re-selected which ones are suitable.

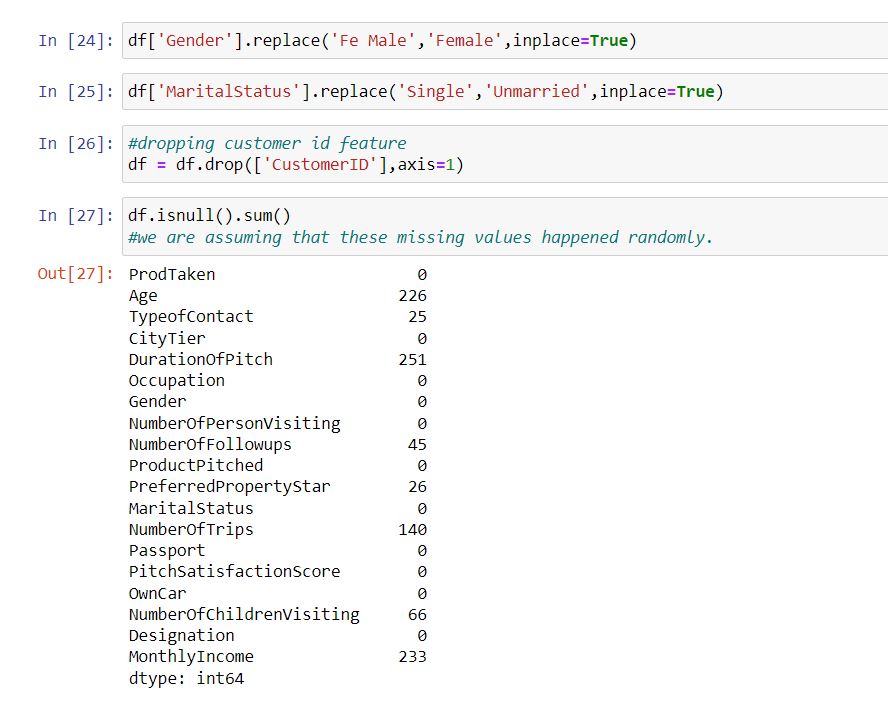
***DATA CLEANING***



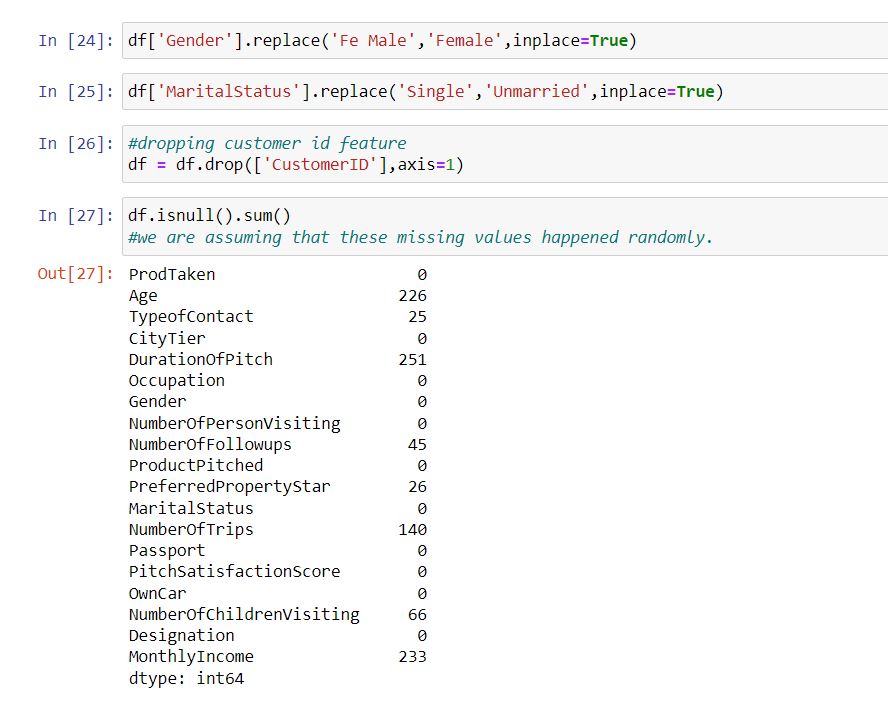
* There are 4888 distinct observations in the sample because there are no duplicate rows.
* It was discovered while analysing the data that the feature Gender had three categories: ‘Male’, ‘Female’, and ‘Fe Male’. Clearly, there is an error made in ‘Fe male’, therefore all ‘Fe Male’ were replaced by ‘Female’.



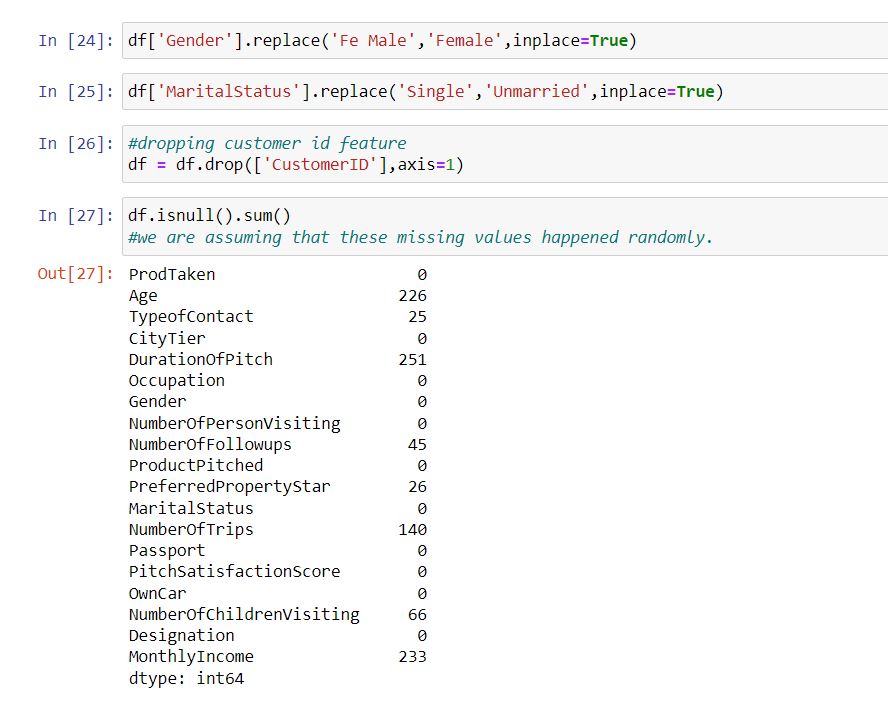
* Additionally, the two categories: "Single" and "Unmarried," in marital status are virtually interchangeable. As a result, "Single" is changed to "Unmarried."



* Customer ID is no longer considered a feature because it is just used for identification purposes and has no bearing on ProdTaken.



* ***NULL VALUES or MISSING VALUES***- Identification and Imputation of null values is a very important step of Data Cleaning process. This is because most of the ML model algorithms require the imputation of missing values before fitting them and ensures the preservation of all cases/observations. There are many null values; therefore deleting them will highly comprise our study. So we will use various imputation methods according to the distribution and domain knowledge of the variables.

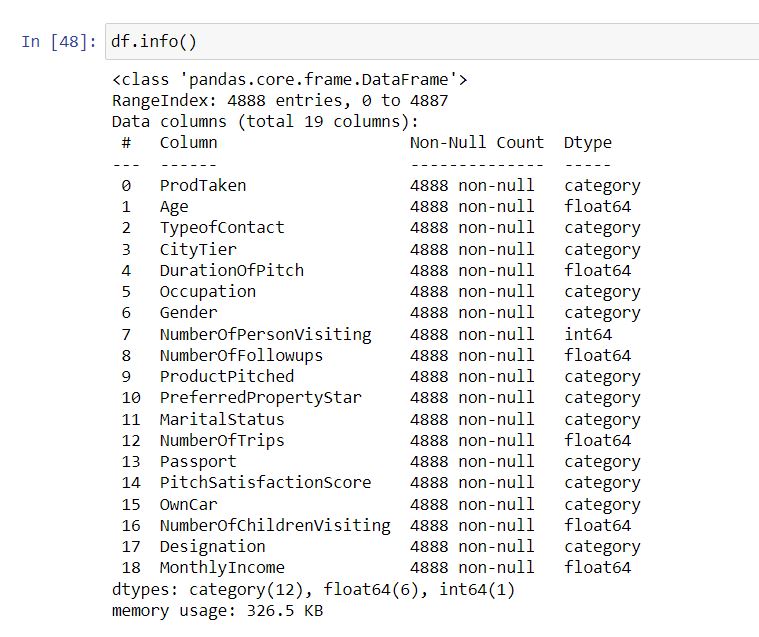


IMPUTATION OF MISSING VALUES

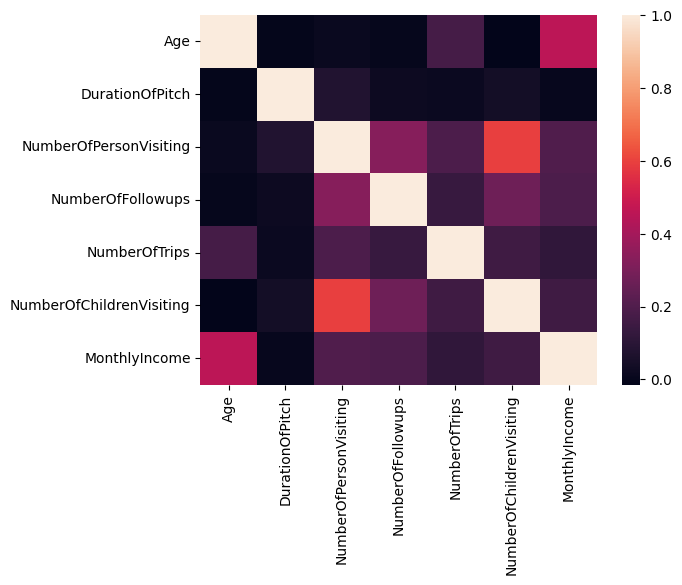
1. Mean imputation is used for Age because it is approximately normally distributed with most of our observations around the mean value. Also, the mean age comes out to be 37 yrs which also lies in the largest age group of adults aged between 30 – 45 years.
2. Missing value in type of contact is replaced with Unkwown
3. Median imputation is used for MonthlyIncome, DurationOfPitch and NumberOfTrips because they are highly skewed so using mean imputation can produce unrealistic values.
4. Imputation of number of follow ups and number of children vising with 0 because we are assuming in case of no or zero follow ups and no children, the data was not entered and left null.
5. #3 star rating is the most repeated value which implies 3 is the mode of this categorical feature

* Changing the Data type of the variables as category-‘TypeofContact', 'CityTier', 'Occupation', 'Gender','ProductPitched', 'PreferredPropertyStar', 'MaritalStatus', 'Passport', 'PitchSatisfactionScore', 'OwnCar', 'Designation','ProdTaken'

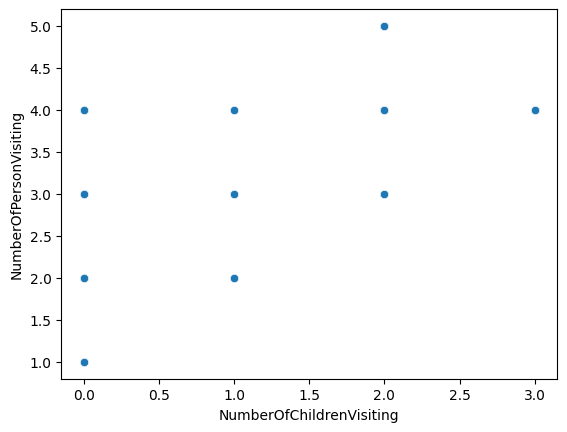
So, now we have 12 categorical variables, 6 float and one integer. This can be seen in the picture below-



* CORRELATION BETWEEN VARIABLES- to check for multicollinearity



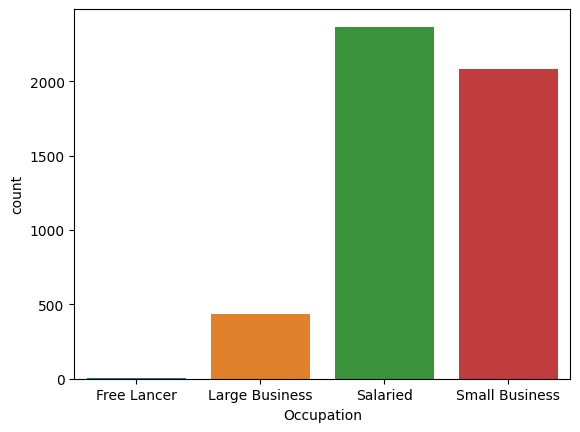
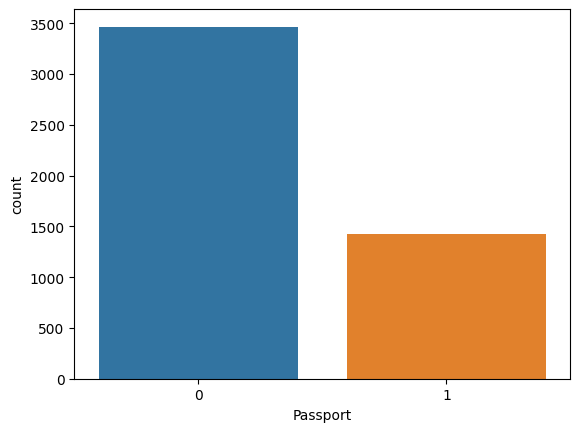
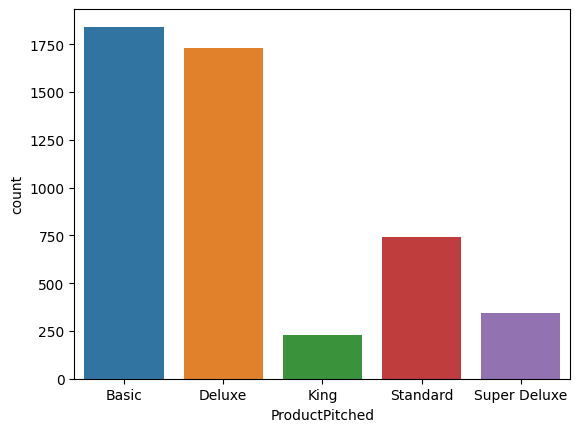
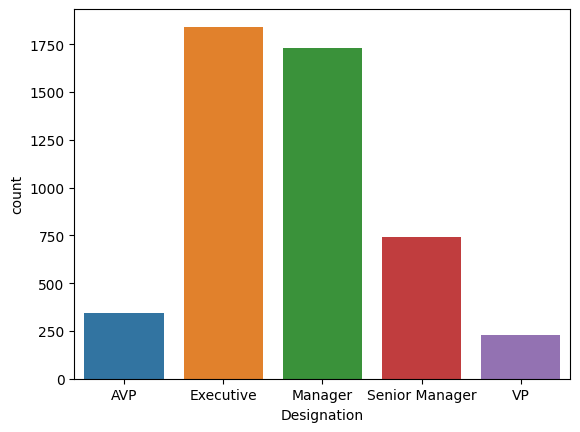
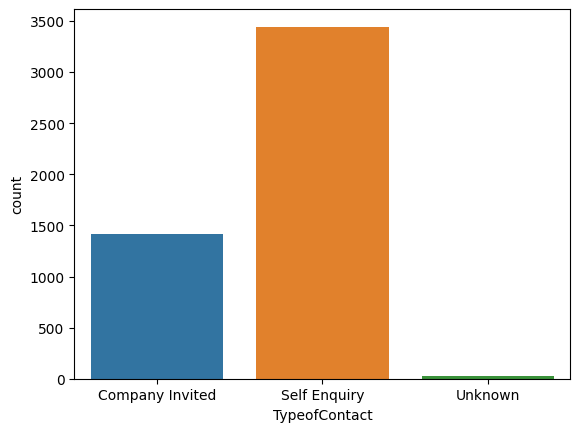
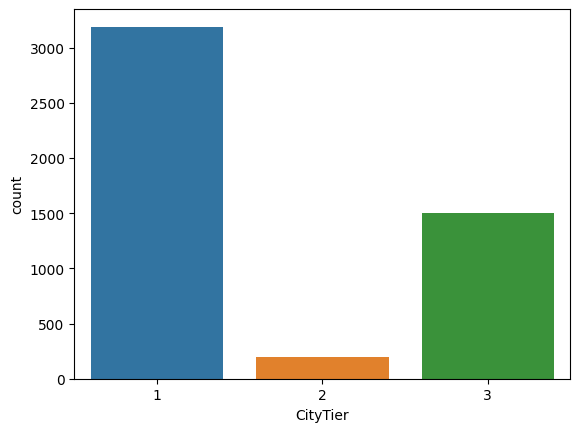
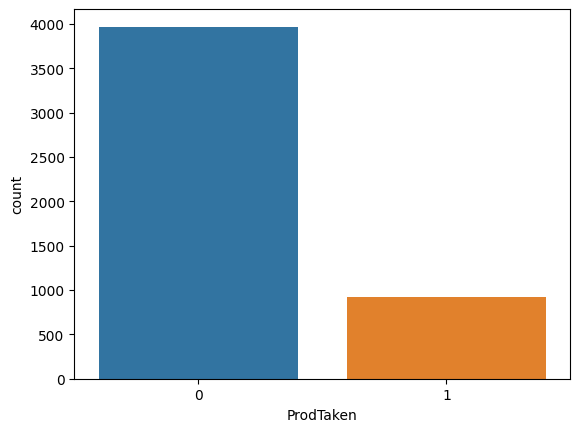
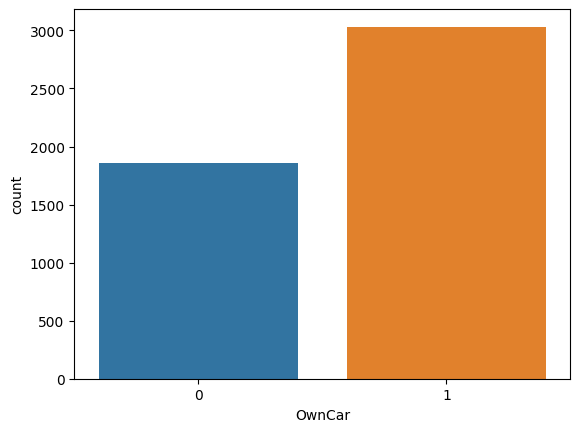
NumberOfPersonVisiting and NumberOfChildrenVisiting are positively associated, as shown by the heatmap above. Since it makes intuitive sense that NumberOfPersonVisiting includes NumberOfChildrenVisiting, modelling that incorporates both features will only result in repetition and overfitting. So, just before getting ready for our model, we will remove NumberOfChildrenVisiting. NumberOfPersonVisiting was chosen since it is a more inclusive characteristic.

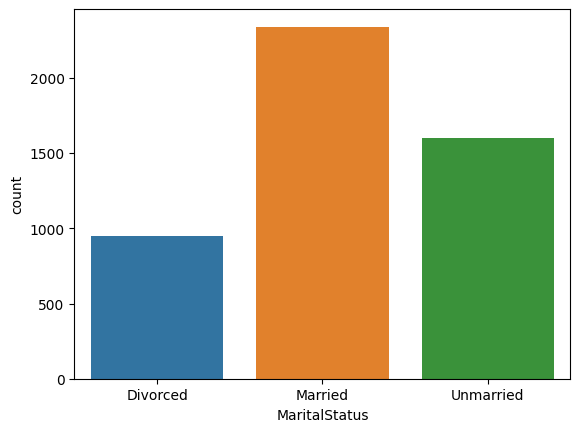
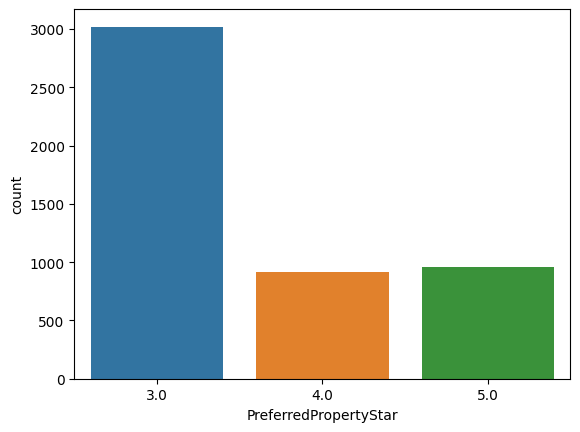


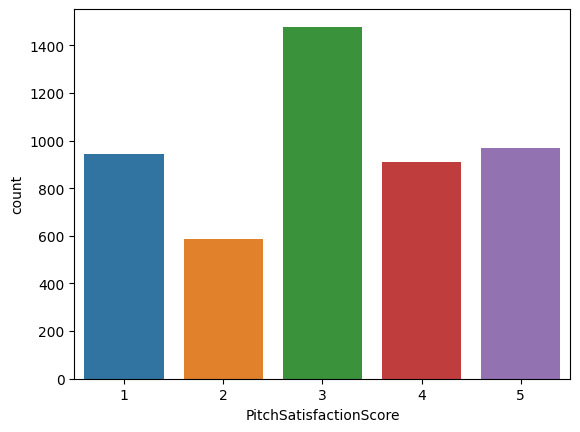
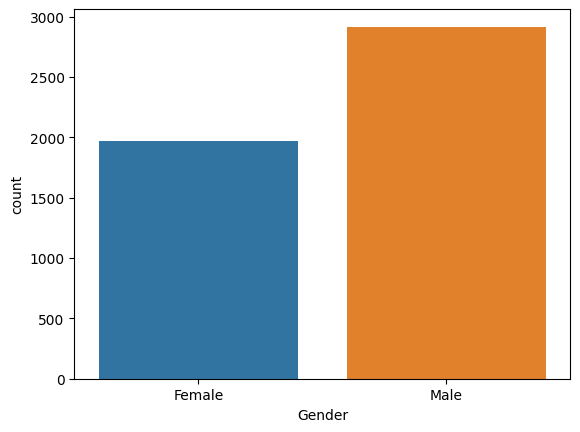
The graph also shows a positive relation between the two variables.

***UNIVARIATE, BIVARIATE and MULTIVARIATE ANALYSIS-***

1. CATEGORICAL VARIABLES-



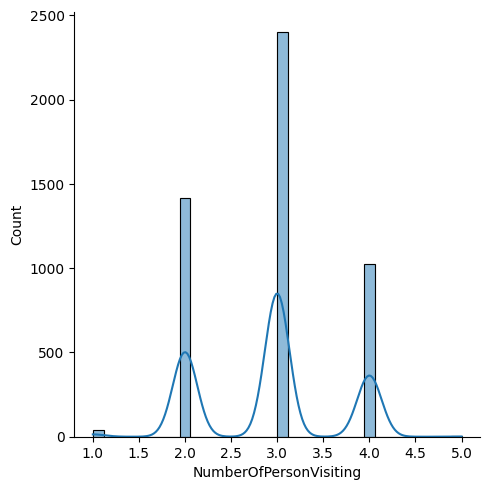
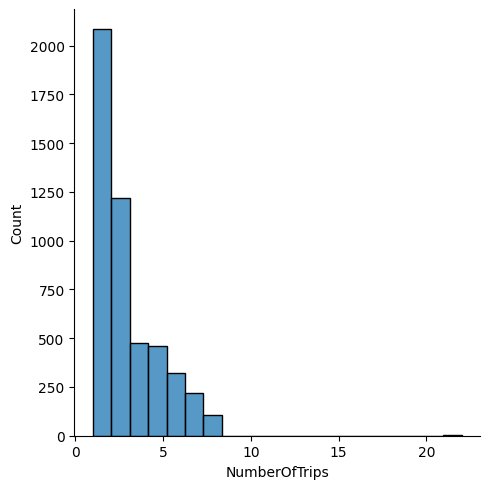
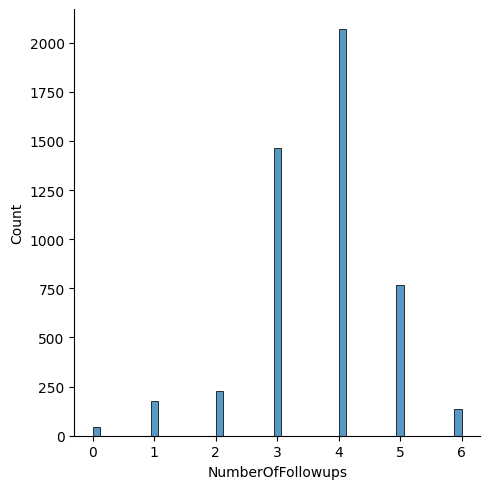
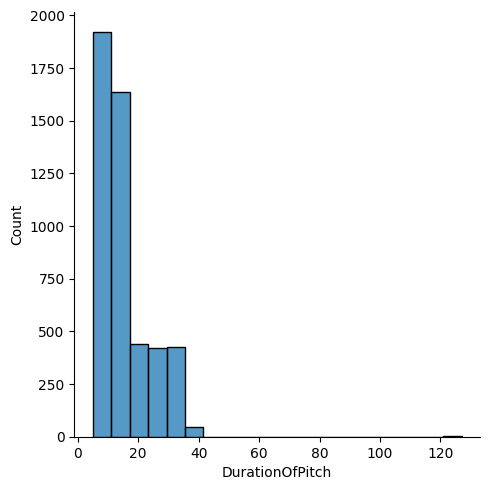
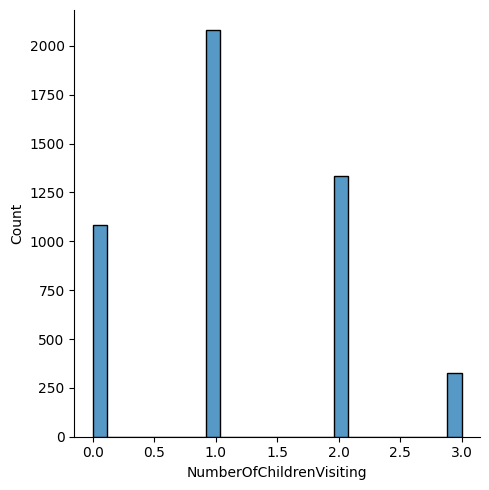
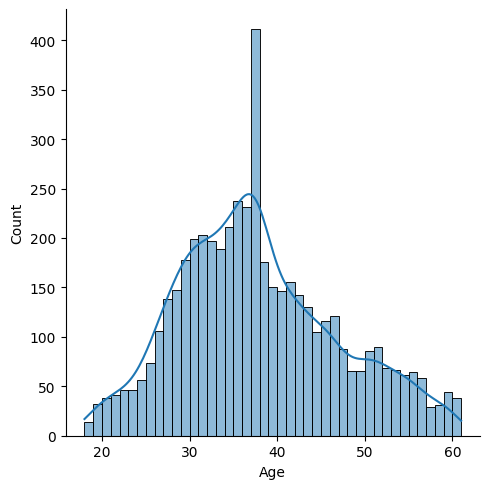


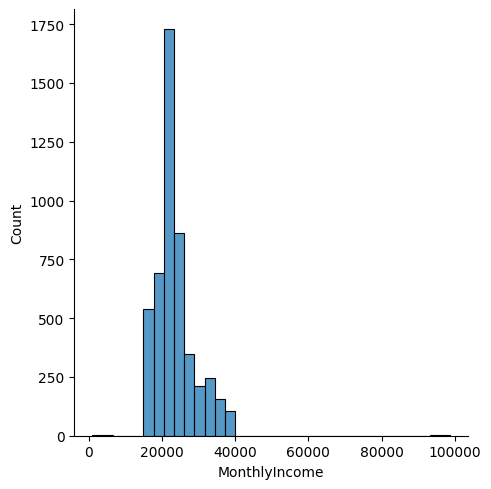


OBSERVATIONS 1-

* ProdTaken is the dependent variable. We that only 18.8% of the total customers purchased any of the travel packages. The plot shows heavy imbalance in the dataset
* Self-Enquiry is the most preferred contact method by the customers at 71%
* 65.3% of customers are from Tier 1 cities and Tier3 cities come second at 30.7%.
* 48.4% of customers are Salaried, i.e. work for an organization and customers with Small Business are the next highest in Occupation at 42.6%.
* Male observations (59.7%) are higher than Female observations (40.3%)
* 49.1% of observations plan to take atleast 3 persons with them during trip. Around 29% customers want to take 2 people and 21% customers want to take 4 additional persons with them during their travel
* Basic and deluxe are the most popular travel packages. The next slightly popular one is the Standard Travel package.
* About 61.8% of our observations prefer a three star hotel rating compared to four (18.7%) and five (19.6%) star rating hotels
* Married customers form the bulk of the data at 47.9% with Divorced (19.4%) and Single (18.7%) coming in close second and Unmarried(with partners) customers form 14% of the data
* Only 29.1% of observations have a passport and almost 62% of customers own a car
* Only 30.2% of observations rated the Sales Pitch with a score of 3. Even though 18.7% customers rated at 4 and 19.8% rated a pitch score of 5, we also see that 19.3% rated the Sales pitch score at 1. This shows a need for improvement in this area
* Around 43.9% of observations have atleast one child under age Five, planning to accompany them in the travels
* Executive (37.7%) and Manager(35.4%) are the highest Designations of the observations in the dataset

1. NUMERICAL FEATURES-



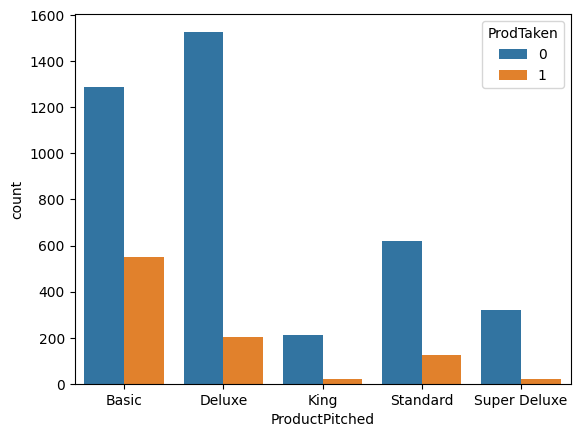


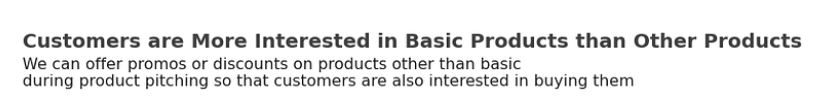
OBSERVATIONS 2-

* Age variable is almost normally distributed with no outliers. We can observe that most customers are in the age brackets 30- 45 yrs.
* DurationofPitch is slightly right-skewed. We see that most pitch duration was under 20 mins. We also see few outliers at 40 minutes and at 120+ minutes.
* The highest number of followups is 4.0 followed by 3.0.
* NumberofTrips is right-skewed a little and majority of the observations seem to take atleast 3 trips per year. We also see very few outliers in the higher end
* MonthlyIncome is also right-skewed. However, we see that the majority of observations are between income bracket 20K dollars and 30K dollars. We also see two outliers in the low end and on the highest end. There are several outliers after the approximately 35K dollars income level.
* The average number of visitors is usually 3, with one outlier at 5 which cannot be ignored because otherwise there will be loss of data since my data only have around 4880 observations.
* NumberOfChildrenVisiting is mostly 1 with no outlier.

We will look at the outlier using boxplot to get more descriptive idea in the next section.

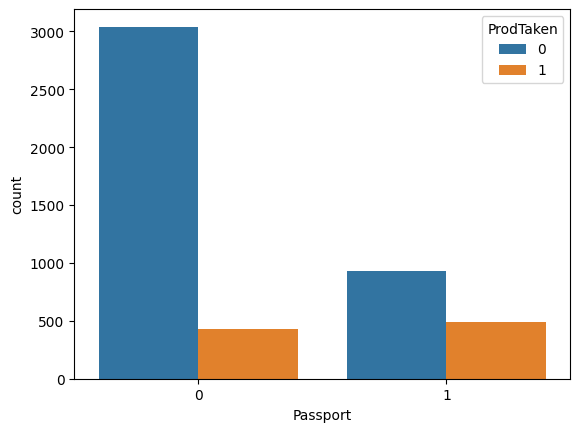
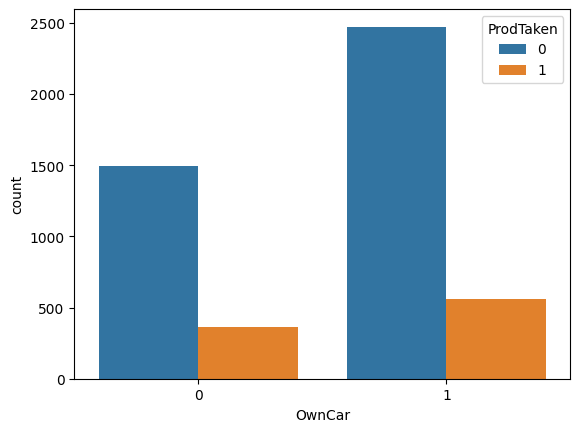
1. SOME INTERESTING OBSERVATIONS-



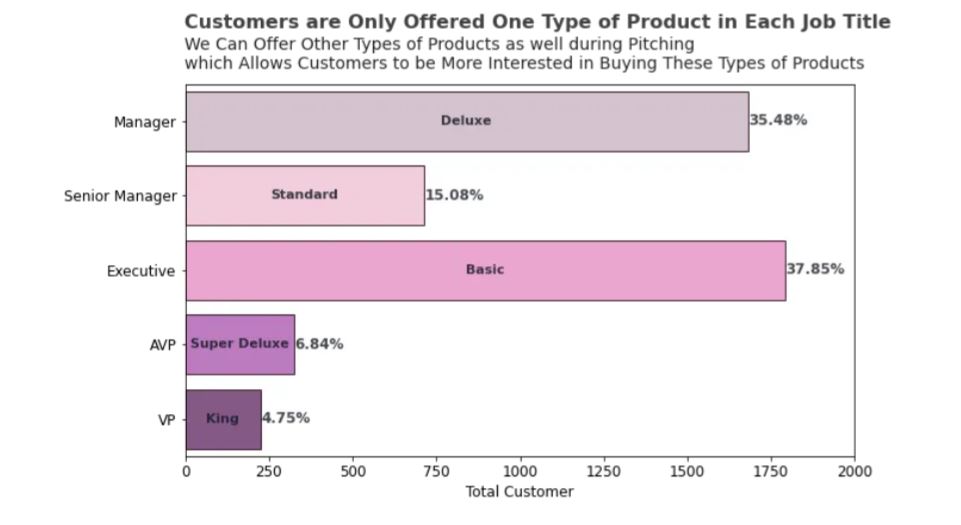




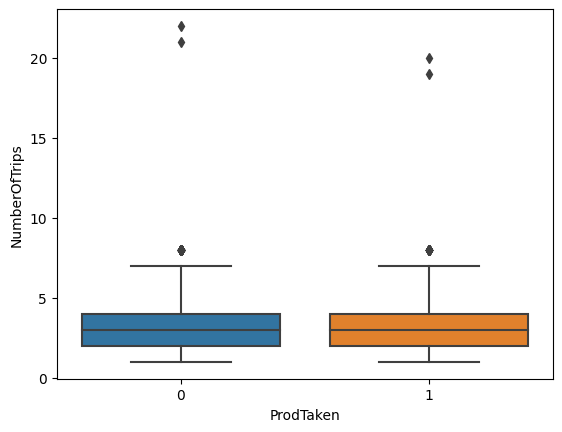
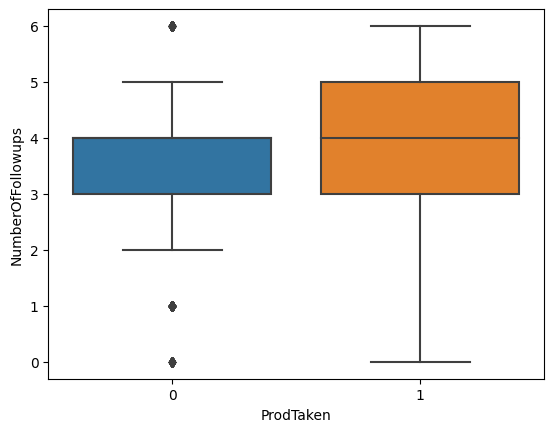
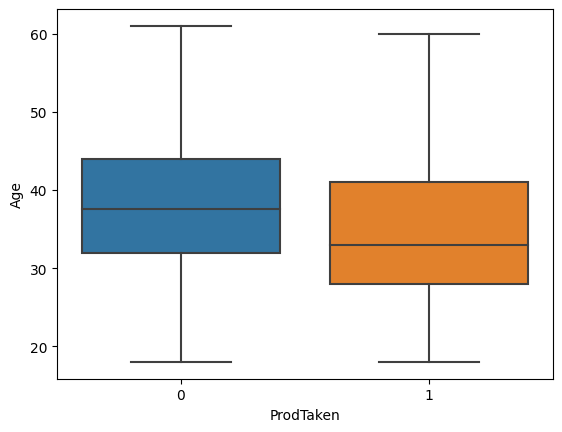
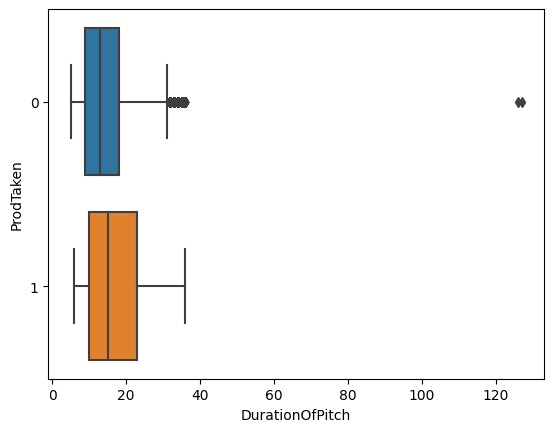
EVEN THOUGH IT WAS OBSERVED THAT BOTH EXECUTIVE AND MANAGERS ARE IN HIGH NUMBERS IN OUR DATASET, THE MORE NUMBER OF EXECUTIVES HAVE TAKEN OUR PRODUCT. HENCE, OUR TARGET SEGMENT SHOULD BE EXECUTIVES.



From the graphs above, it can be seen that customers who refuse the offer of holiday packages are dominated by customers who do not have passports and who do not own a car.



***OUTLIERS DETECTION***

******

Observations 4 -

The mean Age for customers who purchased any Product is slightly less than those who did not. We also see that Age variable does not have any outliers. Customers who purchased the product are mostly in age range of 25 to 35.

The mean DurationofPitch for both classed of ProdTaken is almost equal. We see there are many outliers in Class '0' of ProdTaken, suggesting that longer pitch duration’s doesn’t lead to product purchase. Interestingly, Customers who purchased the packages had an average of atleast four followups, compared to customers who did’nt.

The Averages for NumberofTrips and MonthlyIncome;for both Classes of ProdTaken is almost equal. Number of trips feature has some outliers ranging of 17 to 20.

MonthlyIncome variable has several outliers in the higher end for both ProdTaken classes and very few in low end of Class '0'.Customers who purchased product has monthly income in average of 18000 to 23000 is likely to purchase the travel package

Number of person visiting distribution is slightly higher for customer who had not taken the product.

Customers located in City Tier 1 and 3 are more interested in purchasing travel packages The number of people visiting ranging of 2 to 4 are more likely to purchase travel packages Customers who have a car are more likely to purchase travel packages

***BUSINESS INSIGHTS THAT CAN BE DRAWN FROM ABOVE ANALYSIS-***

* There is a negative correlation between the Age variable to ProdTaken and the MonthlyIncome variable to ProdTaken, which means that the smaller the age or monthly income value, the greater the product taken value. In other words, young/small age customers are more likely to buy holiday package offers compared to older customers, and low-paid customers tend to buy holiday package offers compared to high-paid customers.
* we recommend is to create a campaign to attract new customers who already have passports, so that the chances of receiving a vacation package offer are higher.
* Customer Targeting : Encouraging customers to acquire passports by making special offers. For example, with the same cost as domestic holiday packages, customers who have passports are given the option of traveling abroad with a longer vacation duration or better accommodation.
* The business recommendation that we recommend is to create a campaign to attract new, younger users, so that later the chances of receiving a vacation package offer will be higher.
* Companies can also offer thematic holiday packages that are specifically designed for young people to be an attraction. For example, extreme adventure vacation packages, creative culinary tours, or music and festival trips. Customize holiday packages with the interests and lifestyle of the younger generation.
* The business recommendations that we suggest are the same as before, namely creating a campaign to attract young customers because young customers do not have as much salary as older customers.
* With more follow ups customers have higher chances of buying the product so we should make atleast 4 follow ups with them.
* Also, it was observed that certain job titles are offered only certain type of product, this should not happen. All customers irrespective of their job title should be presented with all the available offers to provide them with options.

PREPROCESSING

We want to predict the customers who will purchase the newly introduced travel package. Hence the Customer Interaction Data for the previous existing travel packages will not add any information to the models. So we will be dropping them for further model building and analysis process.

As the two products are different and hence will require different interaction to sell. We can take insights from such data but it is not very useful in the prediction.

df.drop(['PitchSatisfactionScore','ProductPitched','NumberOfFollowups','DurationOfPitch','Gender'],axis=1)

After resampling and hyperparameter tuning it was found that KNN classifier turns out to be the best predicting model for our dataset with high accuracy, f1 score, precision and recall.