# Holiday Package Prediction for Travelling

Name: Amit Acharekar

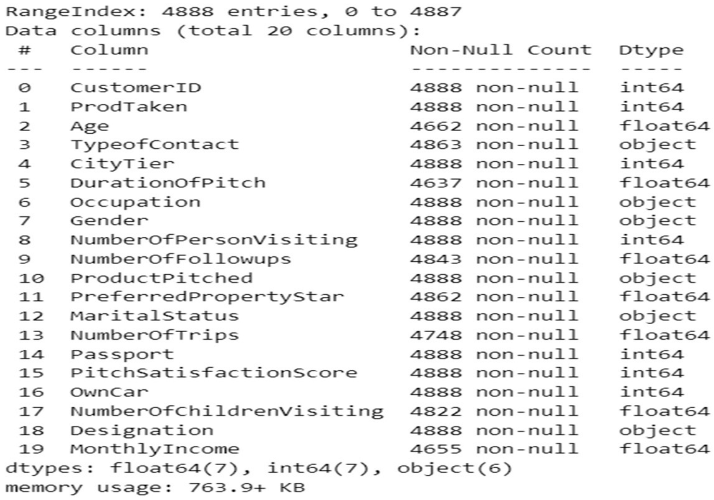
## Holiday Package Prediction for Travelling

## Problem Statement: Predict which type of customer  have maximum chance buying our holiday package.

**Code For following analysis can be found in:** [https://github.com/AmitAcharekar/Fyenn_Labs_Internship/tree/main/Holiday-Package-Prediction](https://github.com/AmitAcharekar/Fyenn_Labs_Internship/tree/main/Holiday-Package-Prediction)

### Data Description

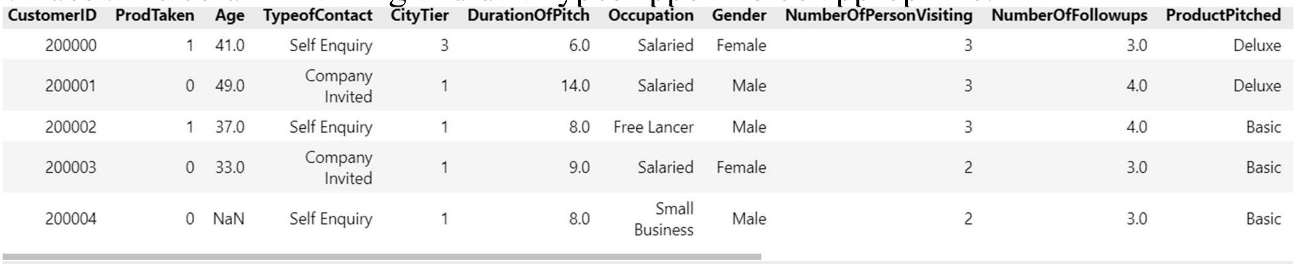
**CustomerID**: Unique customer ID
**ProdTaken**: Product taken or not (0: No, 1: Yes)
**Age**: Age of customer
**TypeofContact**: How customer was contacted (Company Invited or Self Inquiry)
**CityTier**: City tier depends on the development of a city, population, facilities, and living standards. The 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.



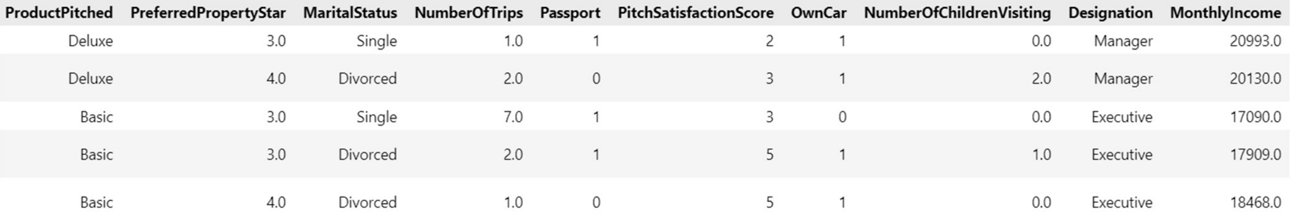
```
RangeIndex: 4888 entries, 0 to 4887
Data columns (total 20 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   CustomerID                4888 non-null   int64  
 1   ProdTaken                 4888 non-null   int64  
 2   Age                       4662 non-null   float64
 3   TypeofContact             4863 non-null   object 
 4   CityTier                  4888 non-null   int64  
 5   DurationOfPitch           4637 non-null   float64
 6   Occupation                4888 non-null   object 
 7   Gender                    4888 non-null   object 
 8   NumberOfPersonVisiting    4888 non-null   int64  
 9   NumberOfFollowups         4843 non-null   float64
 10  ProductPitched            4888 non-null   object 
 11  PreferredPropertyStar     4862 non-null   float64
 12  MaritalStatus             4888 non-null   object 
 13  NumberOfTrips             4748 non-null   float64
 14  Passport                  4888 non-null   int64  
 15  PitchSatisfactionScore    4888 non-null   int64  
 16  OwnCar                    4888 non-null   int64  
 17  NumberOfChildrenVisiting  4822 non-null   float64
 18  Designation               4888 non-null   object 
 19  MonthlyIncome             4655 non-null   float64
dtypes: float64(7), int64(7), object(6)
memory usage: 763.9+ KB
```

From this data it is known that:

The data consists of 4888 rows and 20 columns. It appears that several columns have null/missing values .The column naming and data types appear to be appropriate.



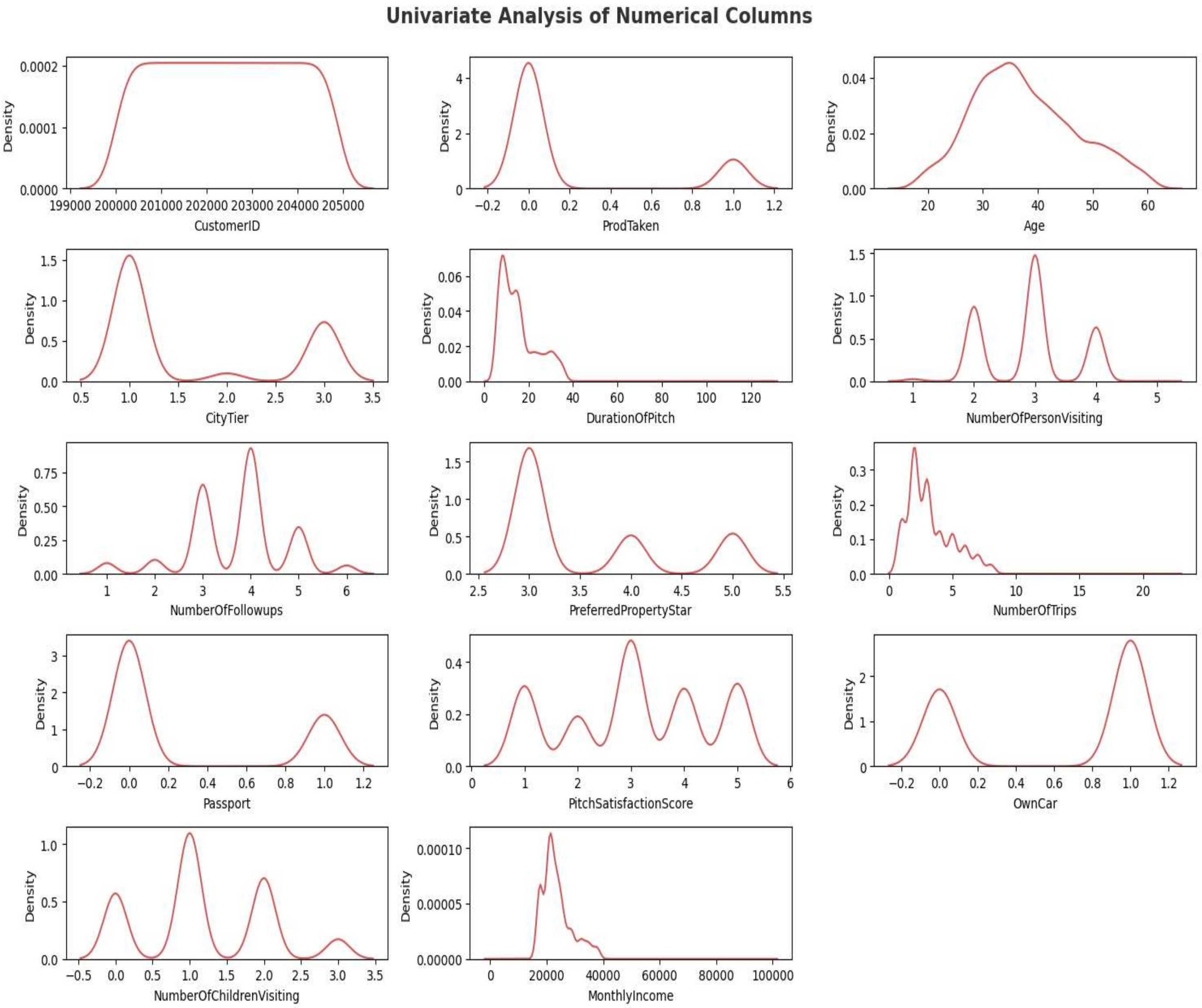
| CustomerID | ProdTaken | Age | TypeofContact | CityTier | DurationOfPitch | Occupation | Gender | NumberOfPersonVisiting | NumberOfFollowups | ProductPitched |
|---|---|---|---|---|---|---|---|---|---|---|
| 200000 | 1 | 41.0 | Self Enquiry | 3 | 6.0 | Salaried | Female | 3 | 3.0 | Deluxe |
| 200001 | 0 | 49.0 | Company Invited | 1 | 14.0 | Salaried | Male | 3 | 4.0 | Deluxe |
| 200002 | 1 | 37.0 | Self Enquiry | 1 | 8.0 | Free Lancer | Male | 3 | 4.0 | Basic |
| 200003 | 0 | 33.0 | Company Invited | 1 | 9.0 | Salaried | Female | 2 | 3.0 | Basic |
| 200004 | 0 | NaN | Self Enquiry | 1 | 8.0 | Small Business | Male | 2 | 3.0 | Basic |



| ProductPitched | PreferredPropertyStar | MaritalStatus | NumberOfTrips | Passport | PitchSatisfactionScore | OwnCar | NumberOfChildrenVisiting | Designation | MonthlyIncome |
|---|---|---|---|---|---|---|---|---|---|
| Deluxe | 3.0 | Single | 1.0 | 1 | 2 | 1 | 0.0 | Manager | 20993.0 |
| Deluxe | 4.0 | Divorced | 2.0 | 0 | 3 | 1 | 2.0 | Manager | 20130.0 |
| Basic | 3.0 | Single | 7.0 | 1 | 3 | 0 | 0.0 | Executive | 17090.0 |
| Basic | 3.0 | Divorced | 2.0 | 1 | 5 | 1 | 1.0 | Executive | 17909.0 |
| Basic | 4.0 | Divorced | 1.0 | 0 | 5 | 1 | 0.0 | Executive | 18468.0 |

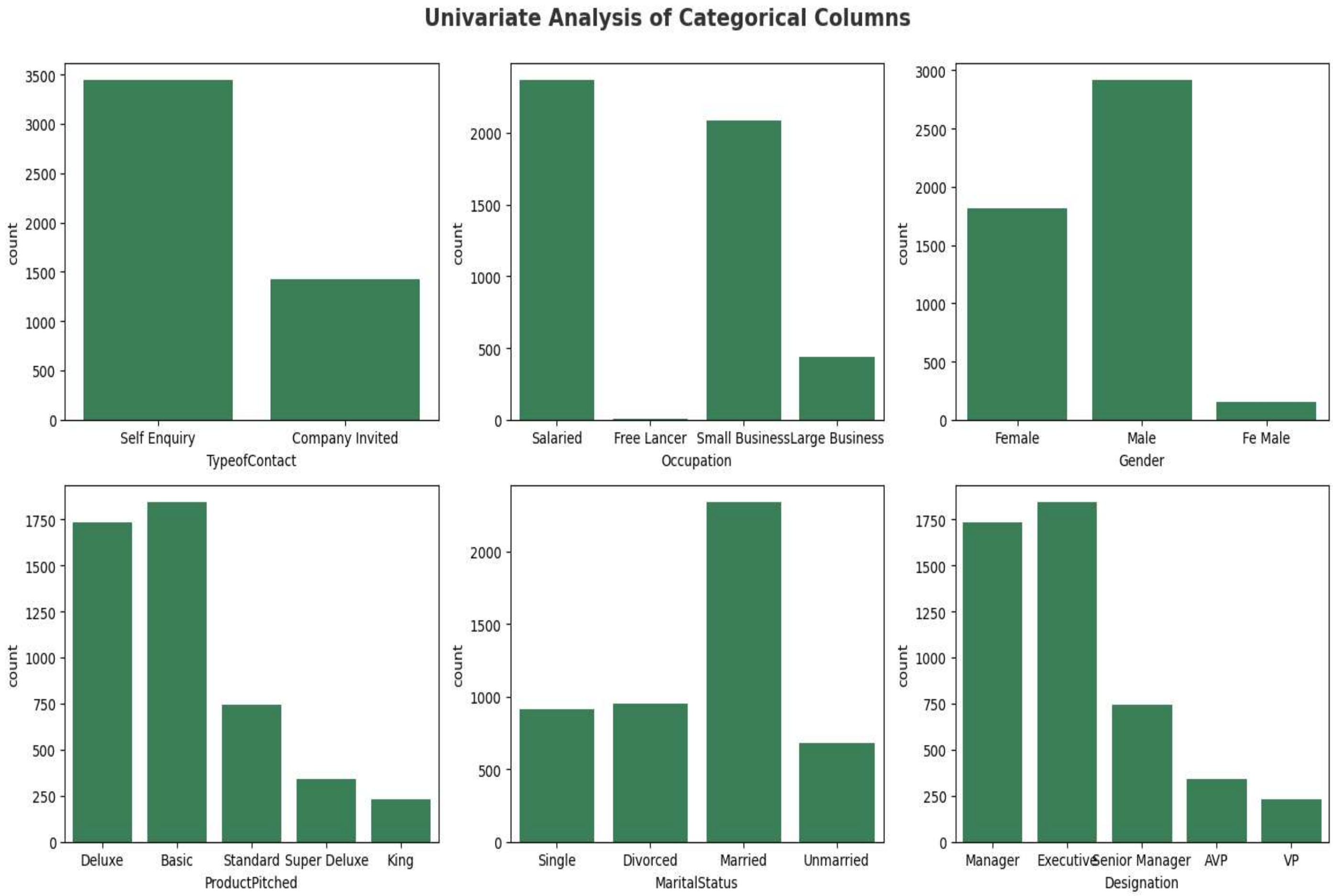
Based on the 5 data samples, it is known that:

The contents of the Gender column contain an error in writing Fe Male which should be Female.
The contents of the MaritalStatus column contain marital status containing Single which has the same meaning as Unmarried.
The contents of the other columns are appropriate.



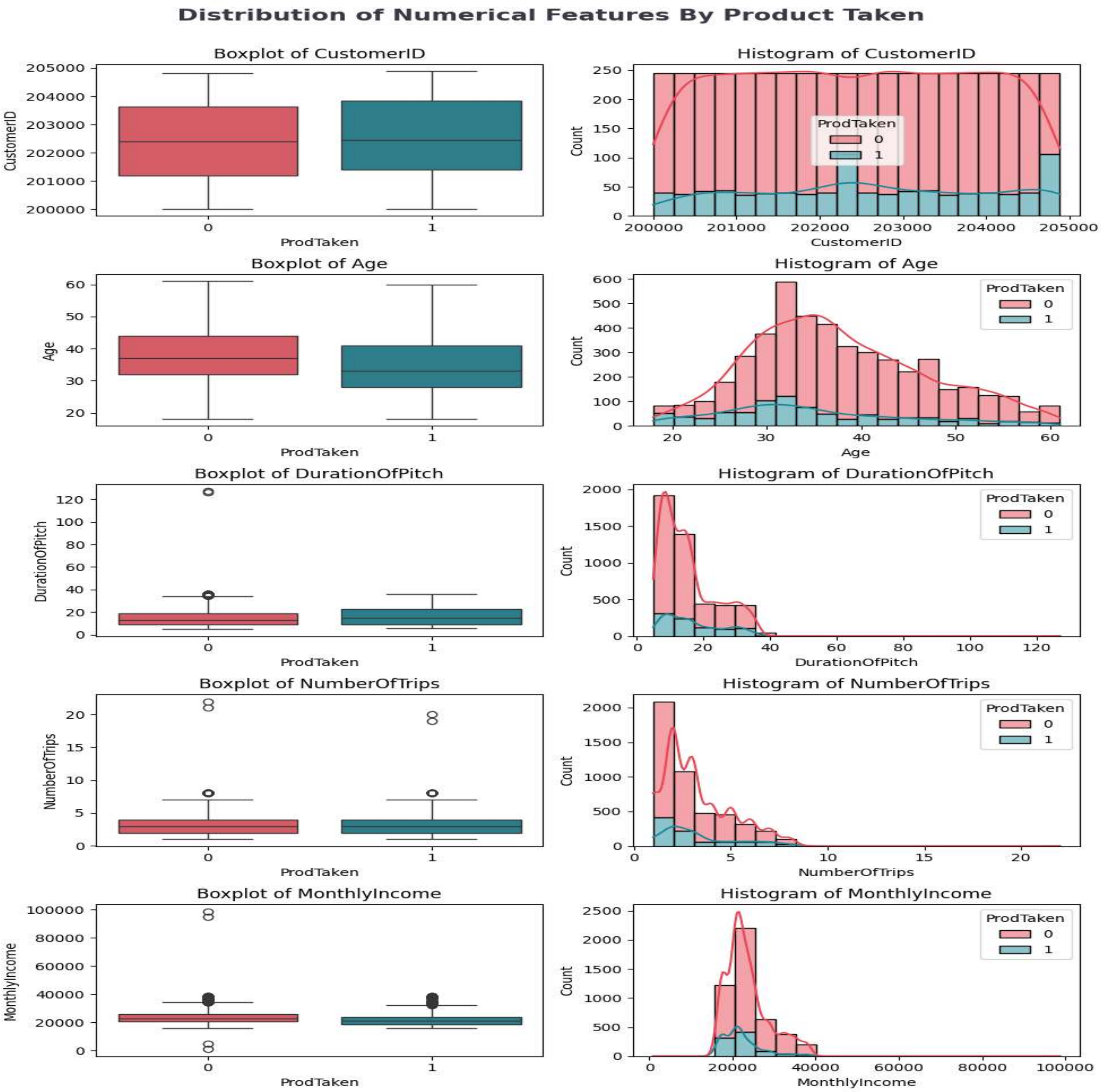
From the results of this visualization it can be concluded that:

The CustomerID column has too much data spread, perhaps because the customers in each row are always unique so this column can be deleted later.The Age column seems to approximate a normal distribution.The DurationOfPitch, NumberOfTrips, and MonthlyIncome columns seem to have a positively skewed data distribution (leaning to the right) which indicates there are outliers. Other columns can be ignored because they are actually discrete or ordinal data types.



From the visualization results above, it can be concluded that:

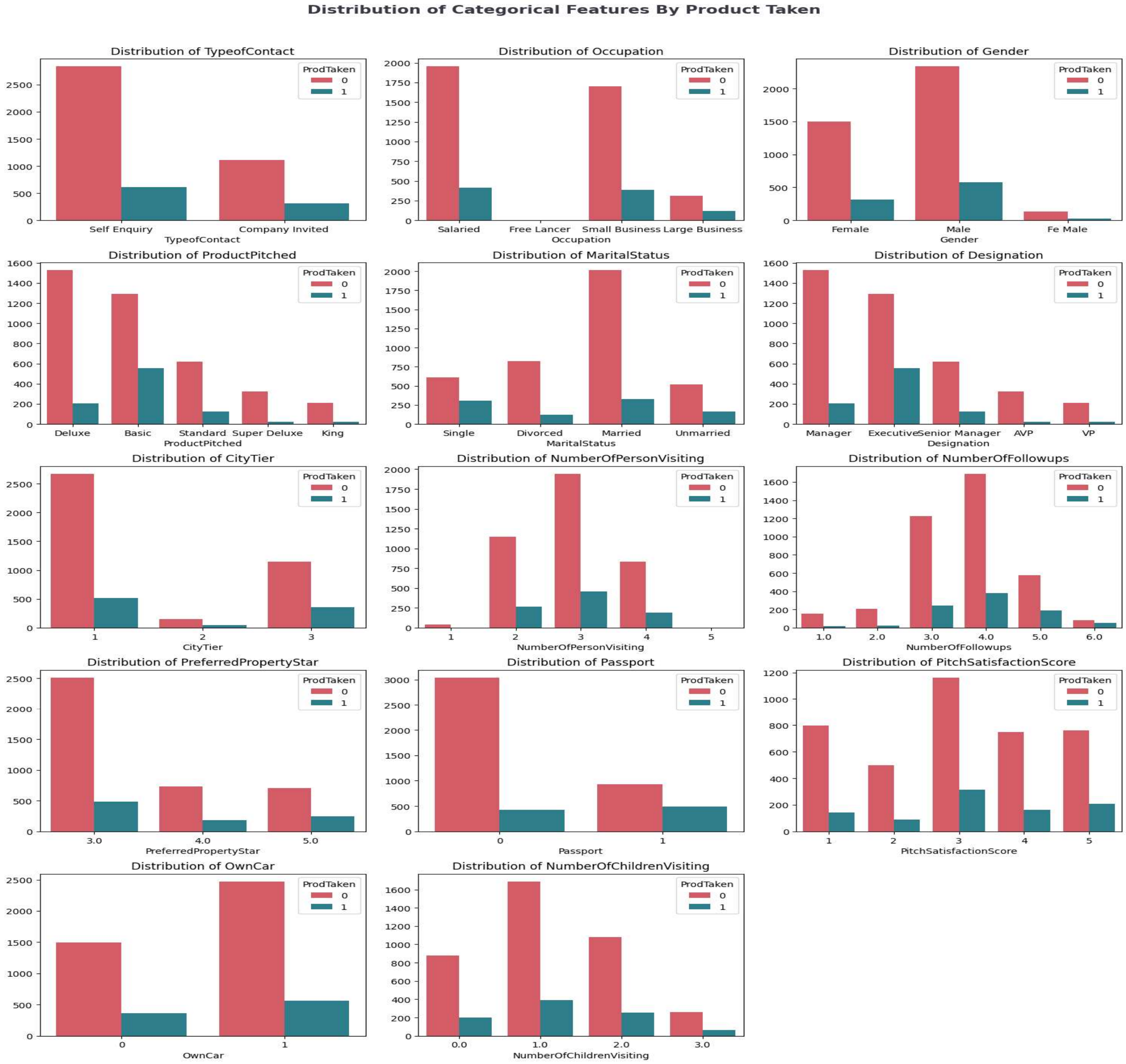
The TypeofContact column is dominated (data proportion more than 50%) by Self Inquiry.
The Occupation column is dominated by Salaried and Small Business, but the number of Free Lancers is too few so they can be deleted later.
In the Gender column there are more males than females. Apart from that, there was an error in writing the Fe Male category which should have been Female.
In the ProductPitched and Designation columns, 2 categories dominate.
The MaritalStatus column is dominated by Married status. Then the statuses Single and Unmarried can be interpreted the same way so they can be combined.



From the results of the above visualization it can be concluded that:

In the CustomerID column, because the data has a unique value for each row, it can be ignored as a feature.
Customers who buy travel packages are mostly aged 20-35.
The duration of the sales pitch is under 20 minutes, allowing customers to buy travel packages.
Most customers take no more than 5 trips in a year.

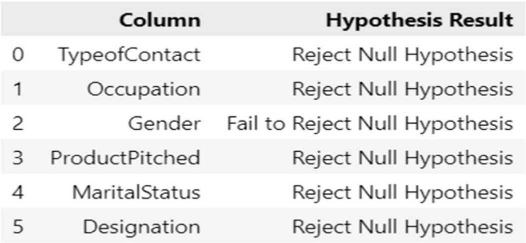
Customers with monthly incomes ranging from 150000-250000 may be more interested in buying travel packages.


Distribution of Categorical Features By Product Taken

From the visualization results above, it can be concluded that:

Customers with the Self Inquiry contact type are more likely to purchase travel packages than Company Invited.
Customers with Occupation Salaried and Small Business are more interested in buying travel packages, and Free Lancers are definitely buying travel packages.

More male customers buy travel packages than women.
The basic product types offered by sales to customers are purchased more often.
Married, single or unmarried customers are more likely to buy travel packages.
More travel packages offered to Executive customers are purchased.
Customers in City Tier 1 and 3 are more interested in buying travel packages.
The number of people who will travel is 2-4 more buying travel packages.
Customers who were followed up >=3 times after the pitch bought more travel packages.
Customers who give a rating of 3 to hotel properties that will be used during their holidays buy more travel packages.
Customers who have passports are slightly more likely to buy travel packages.
Customers who give a satisfaction score >=3 buy more travel packages.
Customers who own cars buy more travel packages.
Customers with children under 5 years are 0 or 1 more likely to buy travel packages



| | Column | Hypothesis Result |
|---|---|---|
| 0 | TypeofContact | Reject Null Hypothesis |
| 1 | Occupation | Reject Null Hypothesis |
| 2 | Gender | Fail to Reject Null Hypothesis |
| 3 | ProductPitched | Reject Null Hypothesis |
| 4 | MaritalStatus | Reject Null Hypothesis |
| 5 | Designation | Reject Null Hypothesis |

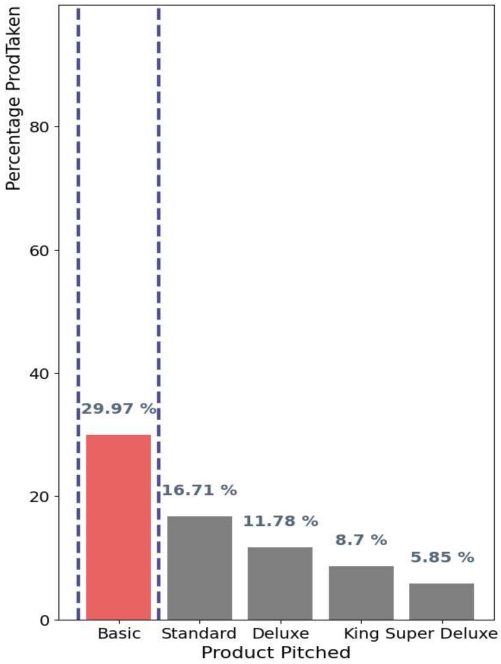
From the above test results it can be concluded that:

The Gender column may not be included as a feature because it Fails to Reject Null Hypothesis (meaning the column is not correlated with the target).The TypeofContact, Occupation, ProductPitched, MaritalStatus, and Designation columns can be included as features because they Reject Null Hypothesis (meaning the columns are correlated with the target).

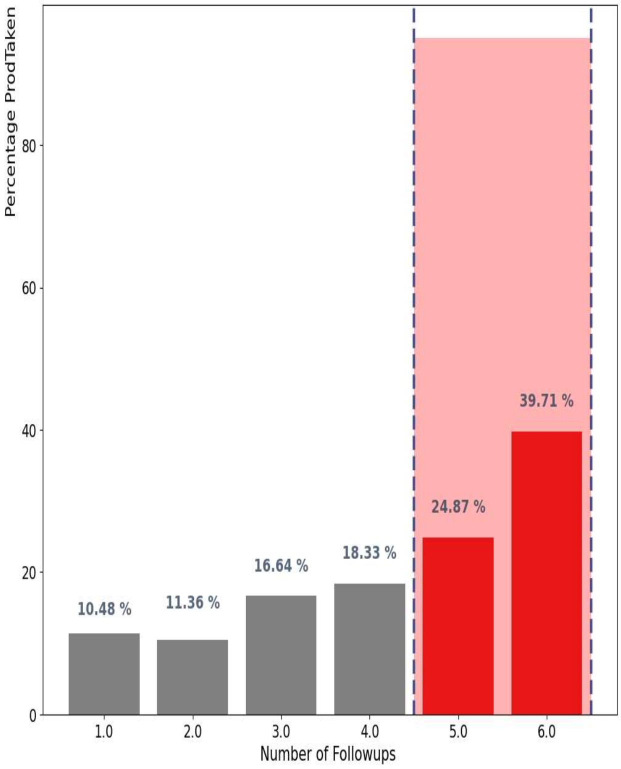
The percentage of customers who did not purchase a travel package was 81.2%The percentage of customers who purchased travel packages was 18.8%.

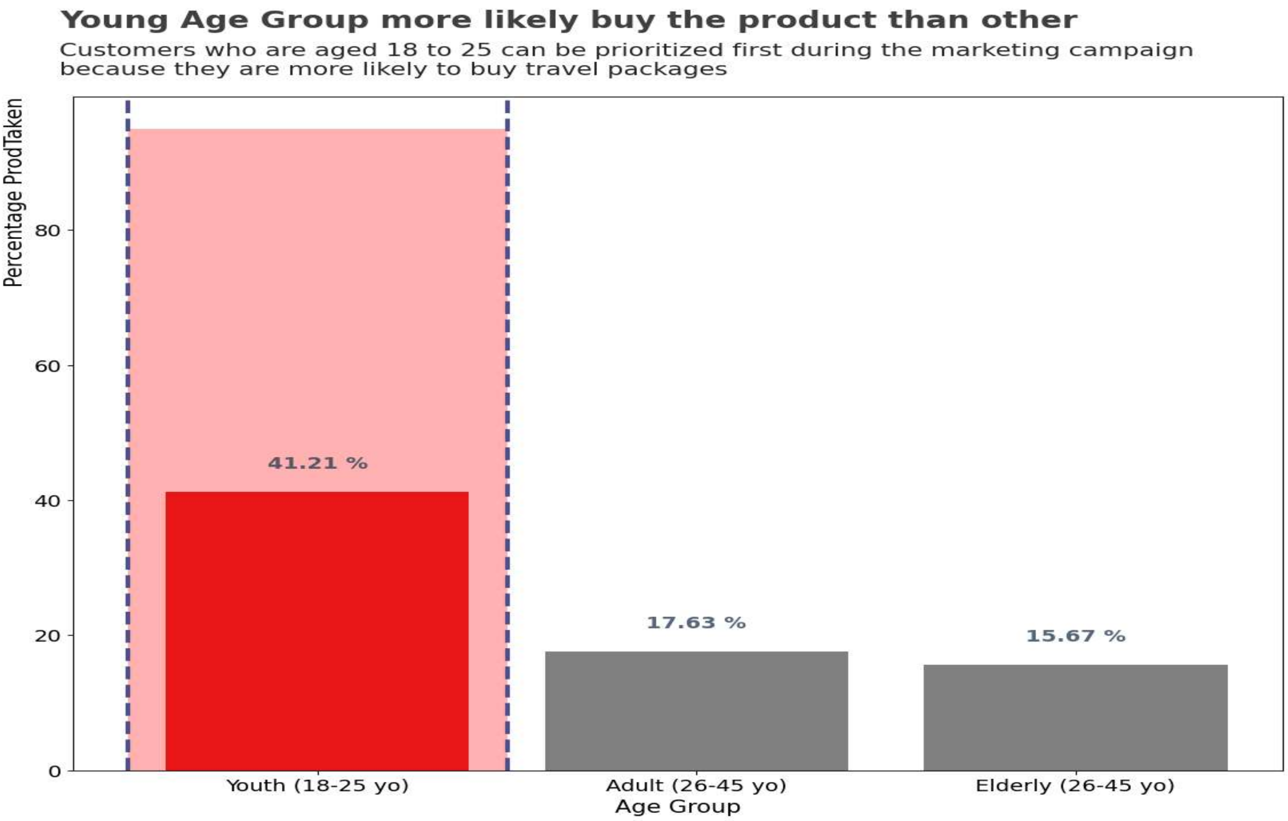
## Customers are More Interested in Basic Products than Other Products

We can offer promos or discounts on products other than basic product pitching so that customers are also interested in buying them



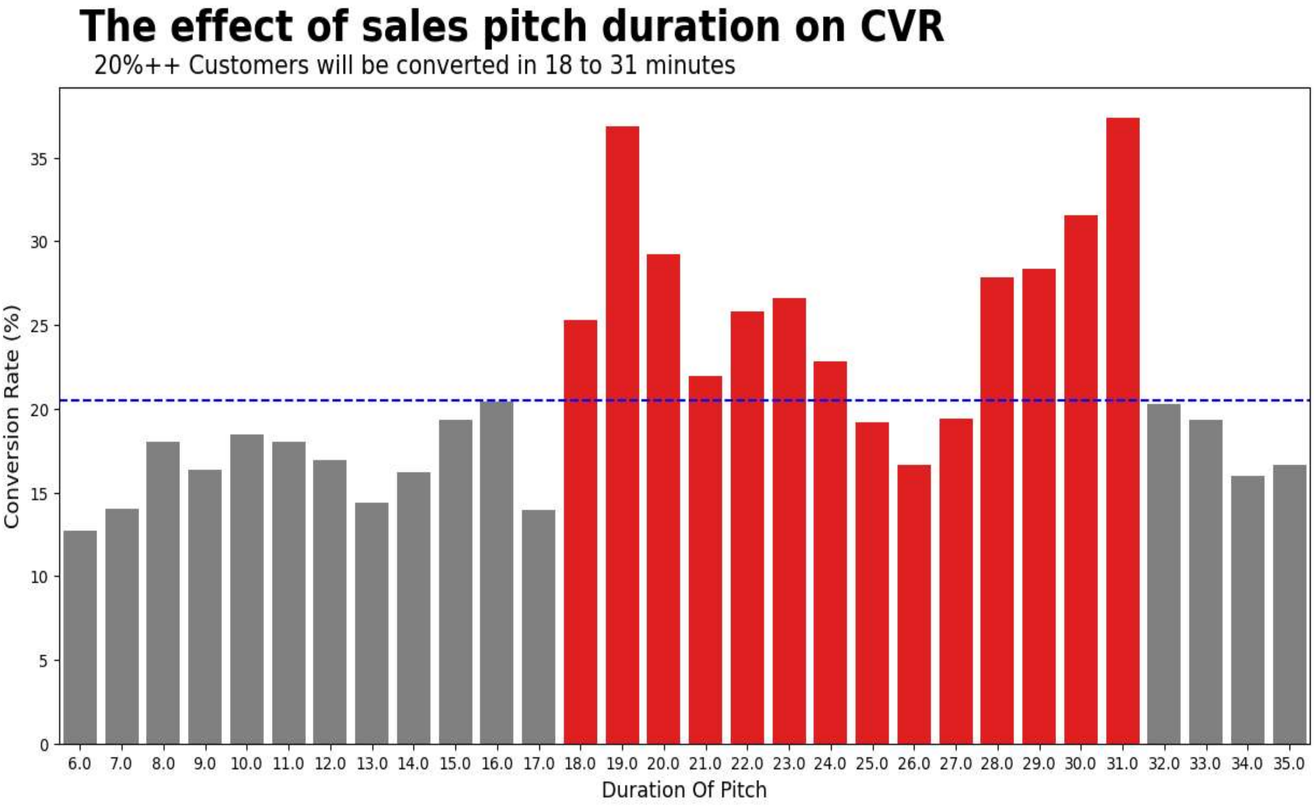
## Follow Up After Pitching 5 Times or More the Product at Least With More Follow-Ups, Customers to Buy Travel Packages Have a Higher Chance





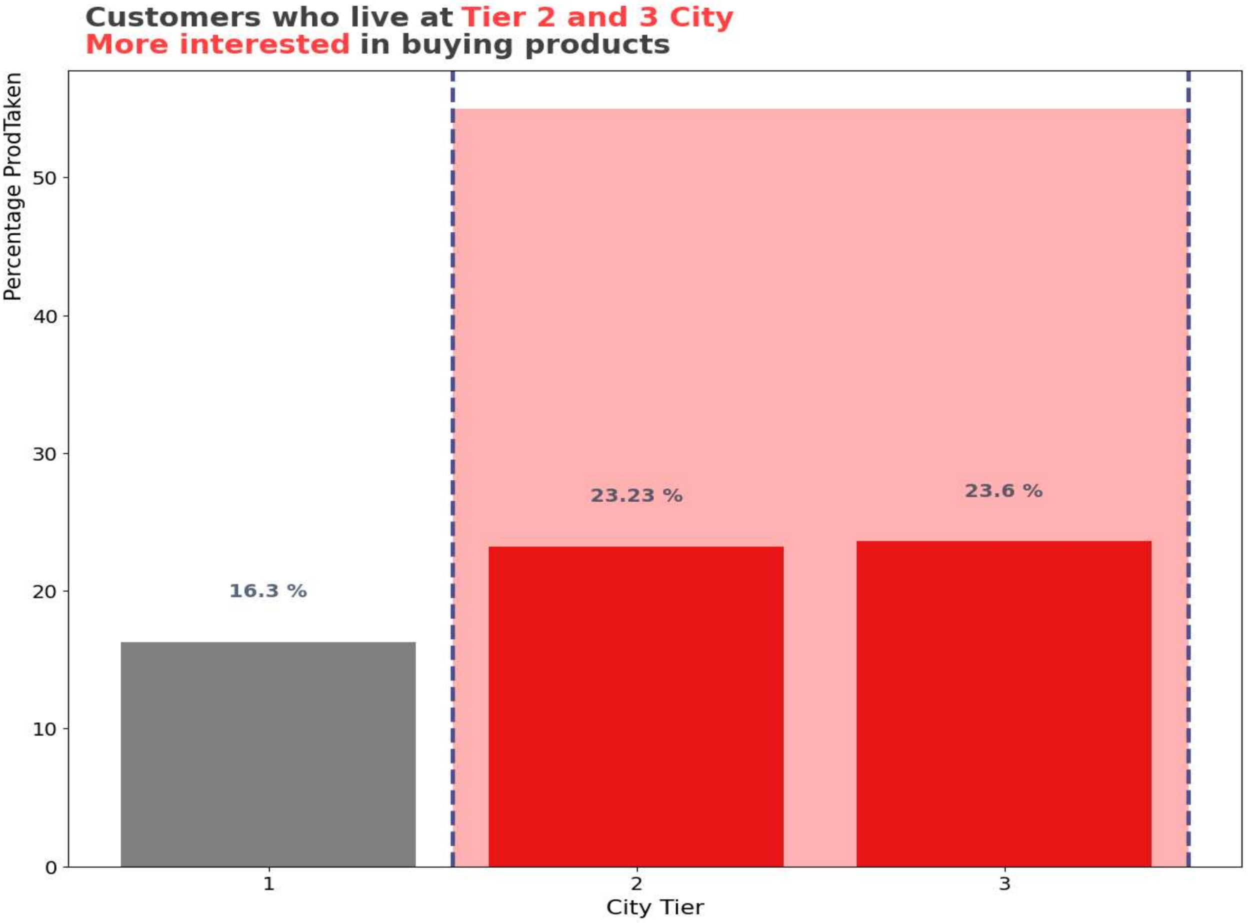
Young Age Group more likely buy the product than other

Customers who are aged 18 to 25 can be prioritized first during the marketing campaign because they are more likely to buy travel packages

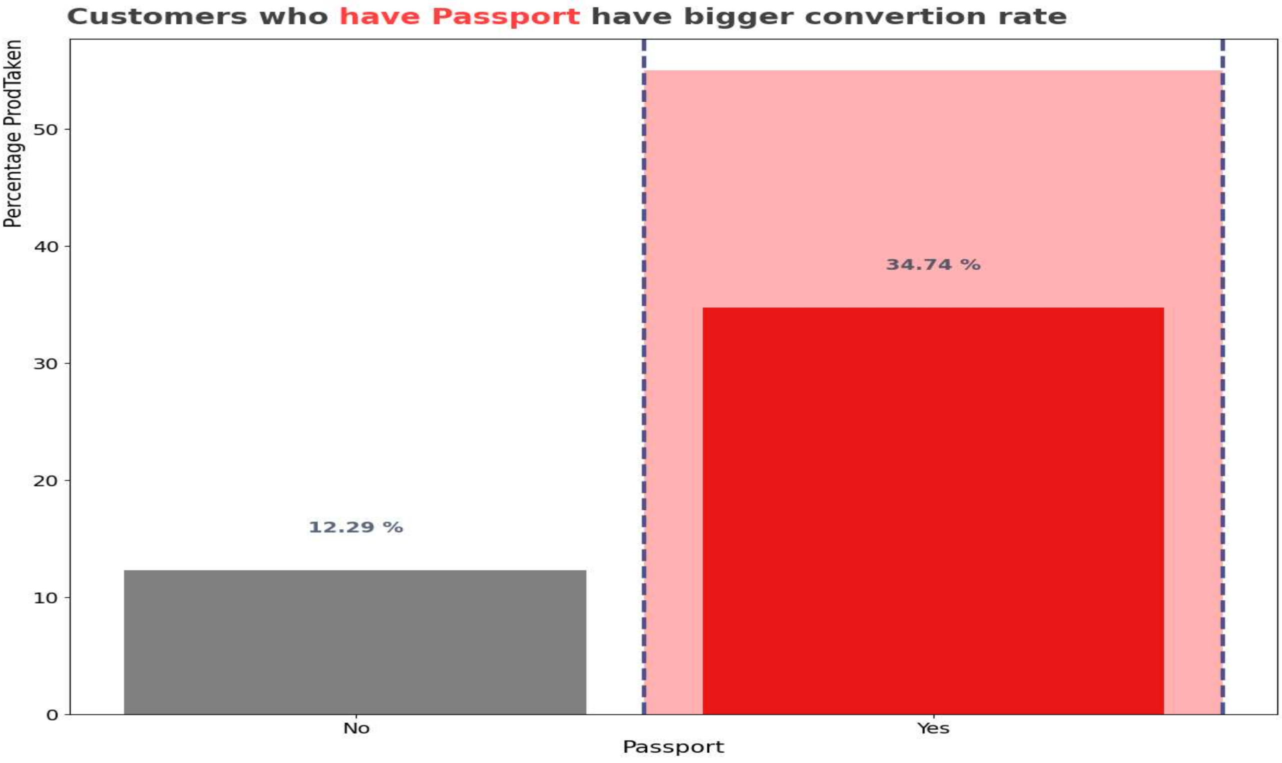


The effect of sales pitch duration on CVR

20%++ Customers will be converted in 18 to 31 minutes



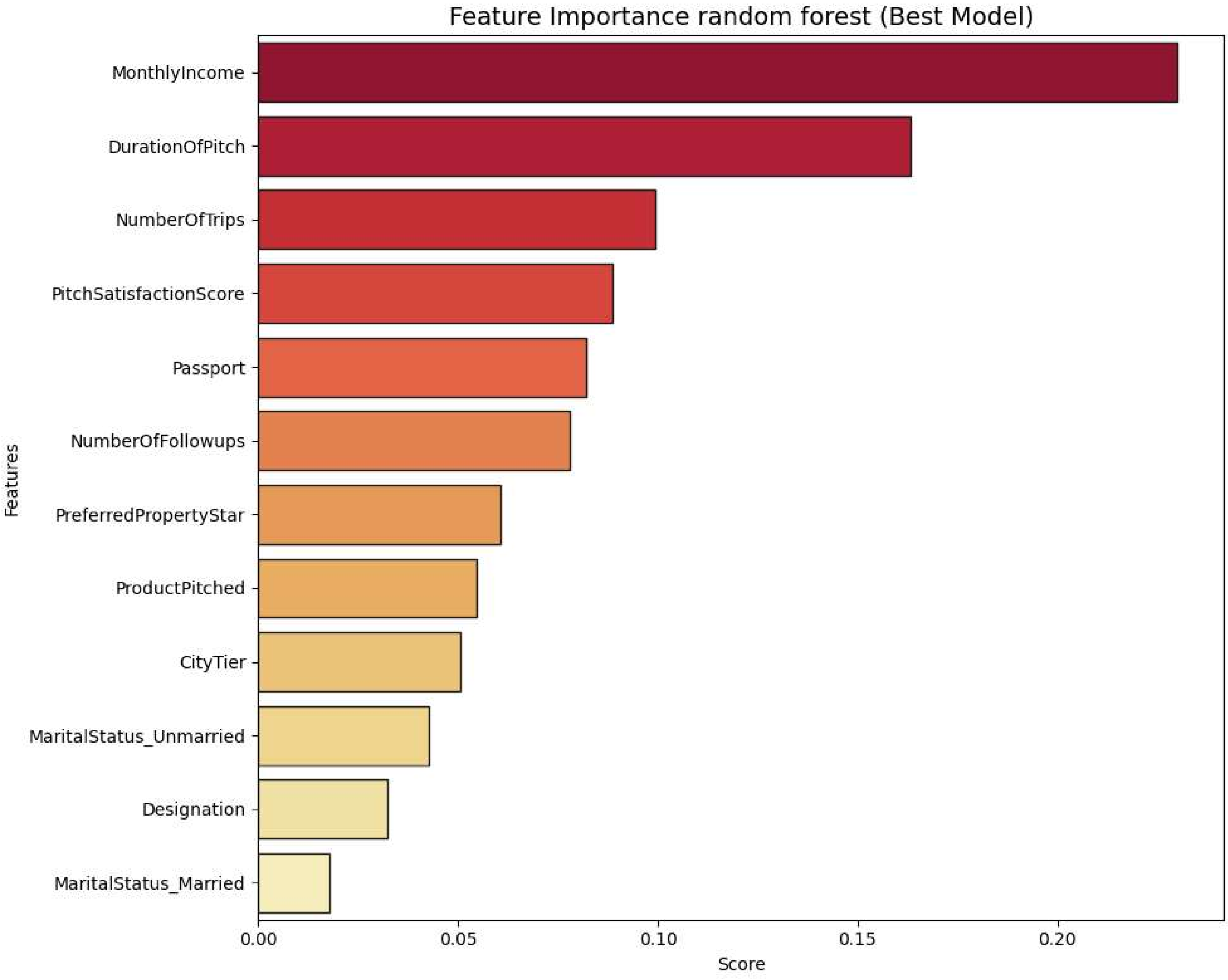
Customers who live at Tier 2 and 3 City More interested in buying products



Customers who have Passport have bigger convertion rate

For Prediction three models were used and their accuracy is:

```
Model Logistic Regression: 84.37%.
Model Decision Tree:  91.2%.
Model Random Forest: 91.87%.
```



We also found the cost required for marketing before and after the use of our model .

**Cost before = NumberOfFollowups * cost_per_person**
**Cost after = ProdTaken_Pred * cost_per_person**

Marketing costs required before using the model are 56340 INR
Marketing costs required after using the model are 7620 INR
So with the model we created we can reduce marketing costs by 48720 INR or decrease by 86.47 %

## Conclusion

Based on the results of the predictions and analysis that we have carried out, it can be concluded that:
Customers that we consider potential to purchase travel packages are:

Using Basic products
Get follow up above 3 times
Have a passport
Not married yet
Live in a Tier 3 City
Purchase type of company solicitation

With the model we created, we succeeded in reducing marketing costs by 47,400 INR.