Data-Driven Growth for O2R2 Mobile

Assignment - 5

Subject : Machine Learning

Group - 3

Objective :

- To predict which customer is more likely to purchase the newly introduced telecom plan.

Data Dictionary :

- **CustomerID :** Unique customer ID
- **PlanTaken :** Whether the customer has purchased the plan 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. Tier 1 > Tier 2 > Tier 3
- **Occupation :** Occupation of customer

- **Gender :** Gender of customer
- **NumberOfPersons :** Total number of persons planning to take the plan with the customer (since these are Friends and Family plans)
- **PreferredServiceStar :** Preferred service rating by customer
- **MaritalStatus :** Marital status of customer
- **NumberOfUpgrades :** Average number of upgrades in a year by customer
- **iPhone :** The customer has an iphone or not (0: No, 1: Yes)

- **PhoneContract :** Whether the customers has a contracted phone or not (0: No, 1: Yes)
- **NumberOfChildren :** Total number of children planning to take the plan with the customer
- **Designation :** Designation of the customer in the current organization
- **MonthlyIncome :** Gross monthly income of the customer
- **PitchSatisfactionScore :** Sales pitch satisfaction score

- **PlanPitched :** Plan pitched by the salesperson
- **NumberOfFollowups :** Total number of follow-ups has been done by the salesperson after the sales pitch
- **DurationOfPitch :** Duration of the pitch by a salesperson to the customer

Steps :

- Data Visualization
- Data Preprocessing Steps and Explanations
- Feature Engineering Steps and Explanation Logistic Regression w/ Regularization Model
- Building and Evaluation (Confusion Matrix, AUC) Steps and Explanation SVM Model
- Building and Evaluation Steps, Comparison with Logistic Regression and Explanation
- Model Tunning

Important Questions to ask the Dataset :

1. How big is the Data ?
2. How does the data look like ?
3. What is the data type of columns ?
4. Are there any missing values ?
5. How does the data look mathematically?
6. Are there duplicate values ?
7. How is the correlation between columns ?

1. How big is the Data ?

```
In [14]: df.shape
Out[14]: (4888, 20)
```

2. How does the data look like ?

```
In [5]: df.head()
```

Out[5]:

	CustomerID	PlanTaken	Age	TypeofContact	CityTier	DurationOfPitch	Occupation	Gender	NumberOfPersons	NumberOfFollowups	PlanPitched	PreferredSe
0	200000	1	41.0	Self Enquiry	3	6.0	Salaried	Female	3	3.0	Deluxe	
1	200001	0	49.0	Company Invited	1	14.0	Salaried	Male	3	4.0	Deluxe	
2	200002	1	37.0	Self Enquiry	1	8.0	Free Lancer	Male	3	4.0	Basic	
3	200003	0	33.0	Company Invited	1	9.0	Salaried	Female	2	3.0	Basic	
4	200004	0	NaN	Self Enquiry	1	8.0	Small Business	Male	2	3.0	Basic	

3. What is the data type of columns ?

```
In [6]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4888 entries, 0 to 4887
Data columns (total 20 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   CustomerID               4888 non-null   int64  
 1   PlanTaken                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   NumberOfPersons          4888 non-null   int64  
 9   NumberOfFollowups        4843 non-null   float64
 10  PlanPitched              4888 non-null   object 
 11  PreferredServiceStar     4862 non-null   float64
 12  MaritalStatus            4888 non-null   object 
 13  NumberOfUpgrades         4748 non-null   float64
 14  iPhone                   4888 non-null   int64  
 15  PitchSatisfactionScore   4888 non-null   int64  
 16  PhoneContract            4888 non-null   int64  
 17  NumberOfChildren         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
```

4. Are there any missing values ?

```
In [7]: df.isnull().mean()*100
```

```
Out[7]: CustomerID               0.000000
PlanTaken                0.000000
Age                      4.623568
TypeofContact            0.511457
CityTier                 0.000000
DurationOfPitch          5.135025
Occupation               0.000000
Gender                   0.000000
NumberOfPersons          0.000000
NumberOfFollowups        0.920622
PlanPitched              0.000000
PreferredServiceStar     0.531915
MaritalStatus            0.000000
NumberOfUpgrades         2.864157
iPhone                   0.000000
PitchSatisfactionScore   0.000000
PhoneContract            0.000000
NumberOfChildren         1.350245
Designation              0.000000
MonthlyIncome            4.766776
dtype: float64
```

5. How does the data look mathematically?

In [8]:
```
df.describe()
```

Out[8]:

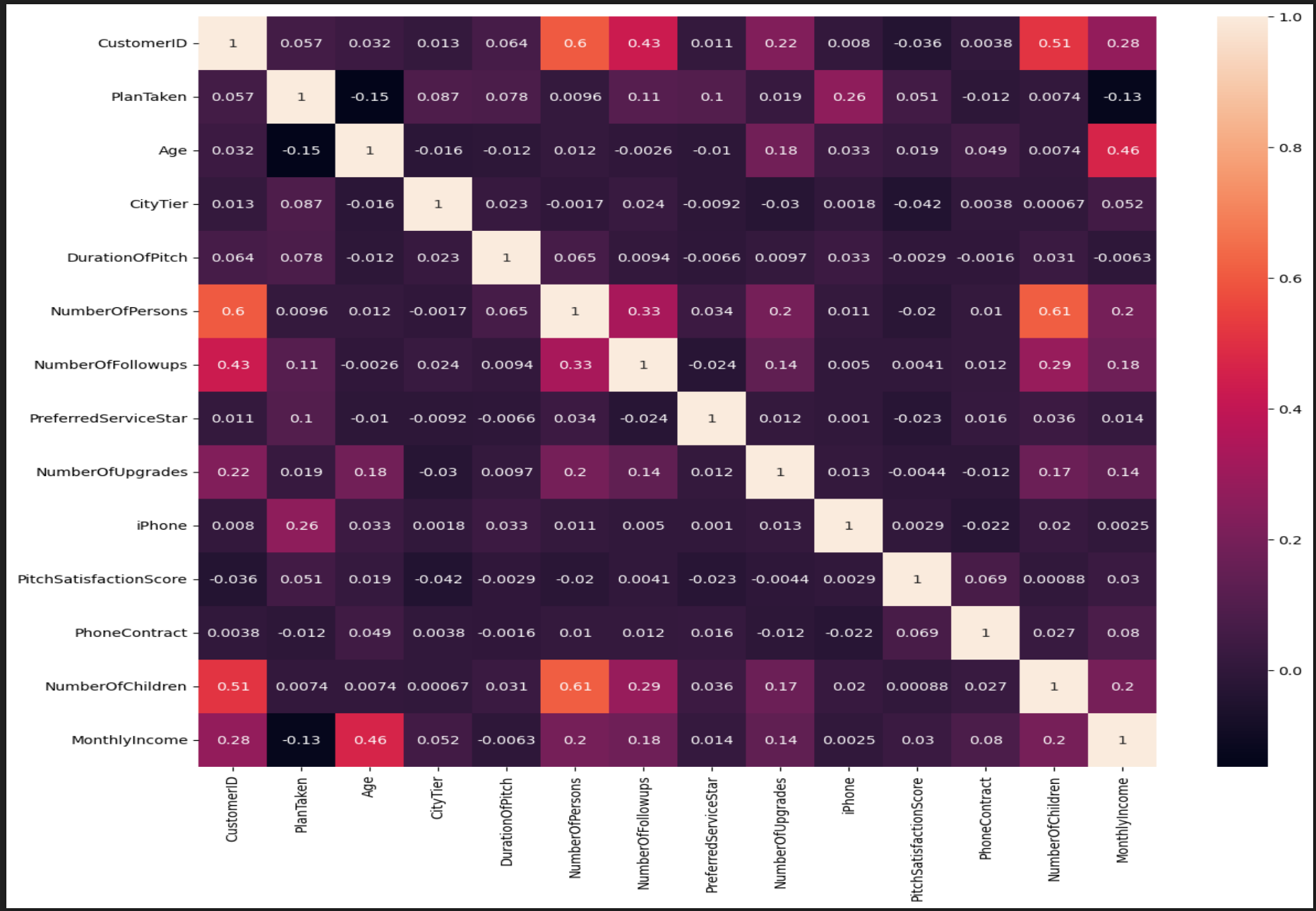
	CustomerID	PlanTaken	Age	CityTier	DurationOfPitch	NumberOfPersons	NumberOfFollowups	PreferredServiceStar	NumberOfUpgrades
count	4888.000000	4888.000000	4662.000000	4888.000000	4637.000000	4888.000000	4843.000000	4862.000000	4748.000000
mean	202443.500000	0.188216	37.622265	1.654255	15.490835	2.905074	3.708445	3.581037	3.236521
std	1411.188388	0.390925	9.316387	0.916583	8.519643	0.724891	1.002509	0.798009	1.849019
min	200000.000000	0.000000	18.000000	1.000000	5.000000	1.000000	1.000000	3.000000	1.000000
25%	201221.750000	0.000000	31.000000	1.000000	9.000000	2.000000	3.000000	3.000000	2.000000
50%	202443.500000	0.000000	36.000000	1.000000	13.000000	3.000000	4.000000	3.000000	3.000000
75%	203665.250000	0.000000	44.000000	3.000000	20.000000	3.000000	4.000000	4.000000	4.000000
max	204887.000000	1.000000	61.000000	3.000000	127.000000	5.000000	6.000000	5.000000	22.000000

6. Are there duplicate values ?

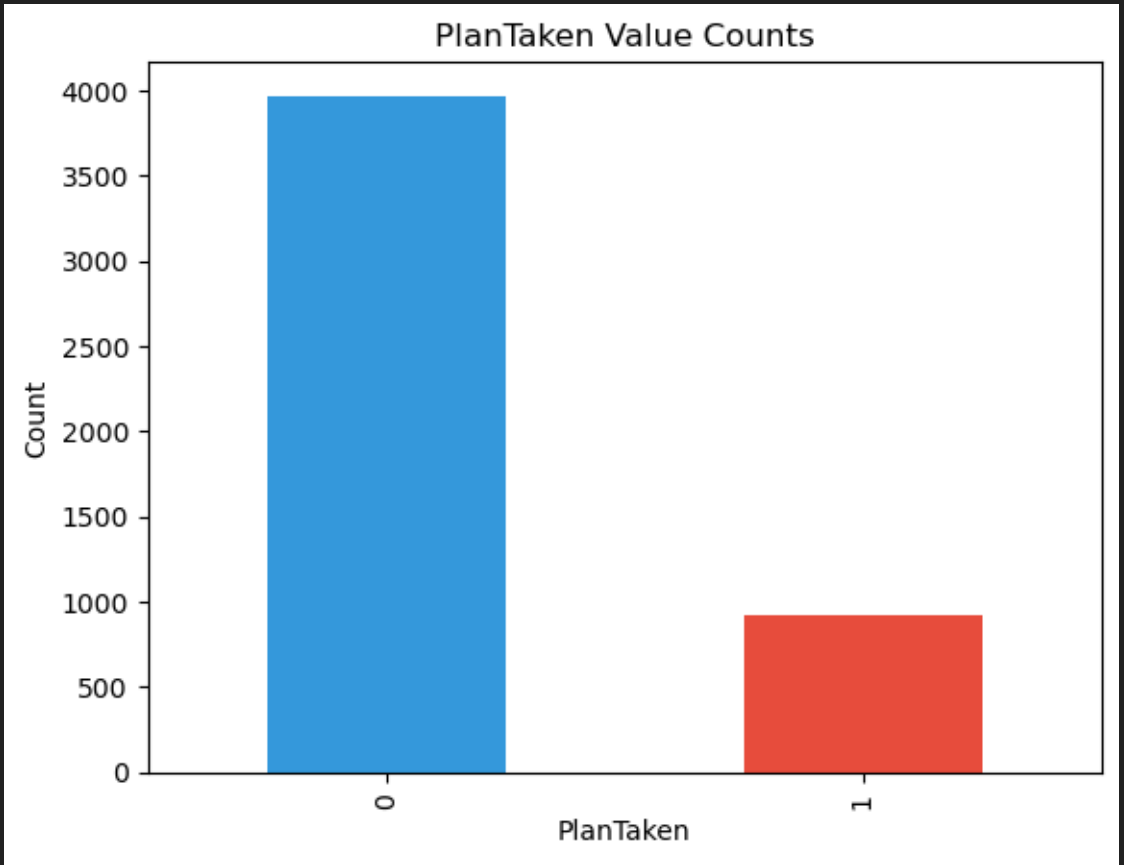
```
In [9]: df.duplicated().sum()
Out[9]: 0
```

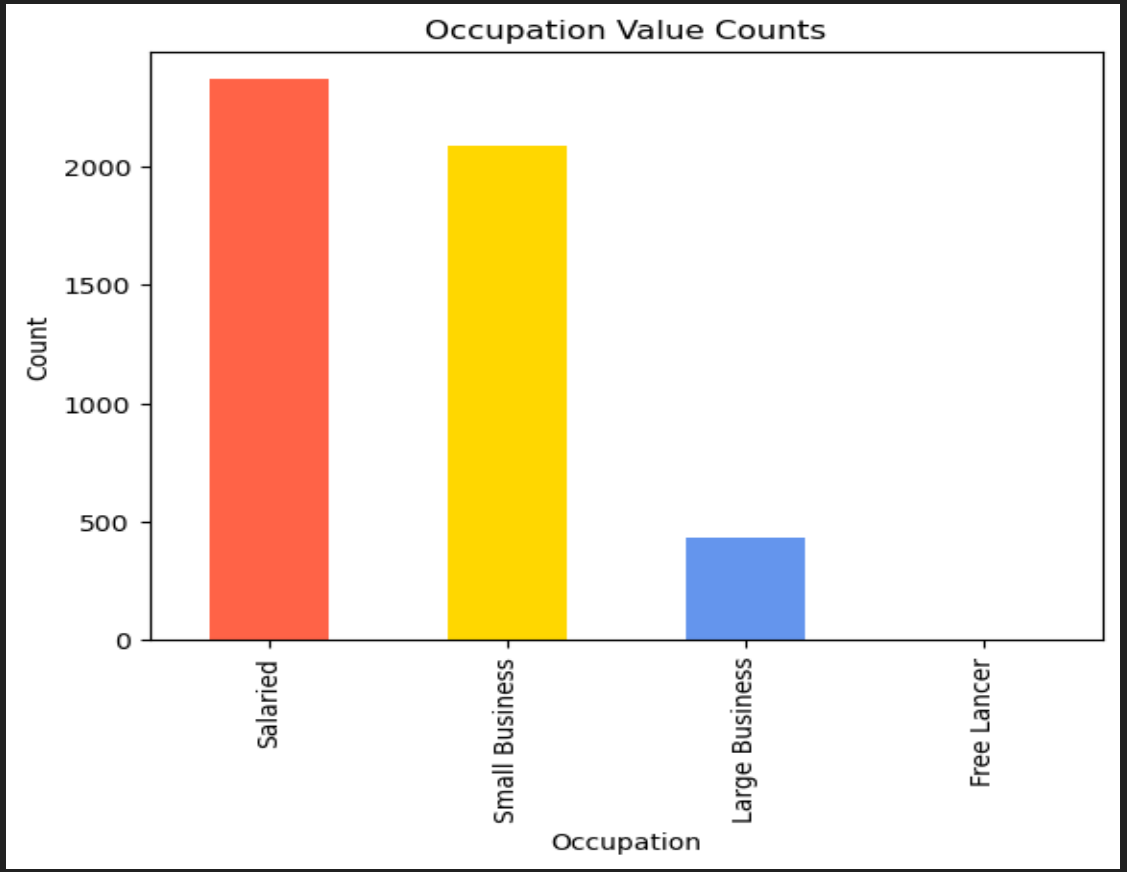
7. How is the correlation between columns ?

```
In [15]: tf = df.corr(numeric_only=True)
         plt.figure(figsize=(14, 12))
         sns.heatmap(tf, annot=True)
Out[15]: <Axes: >
```

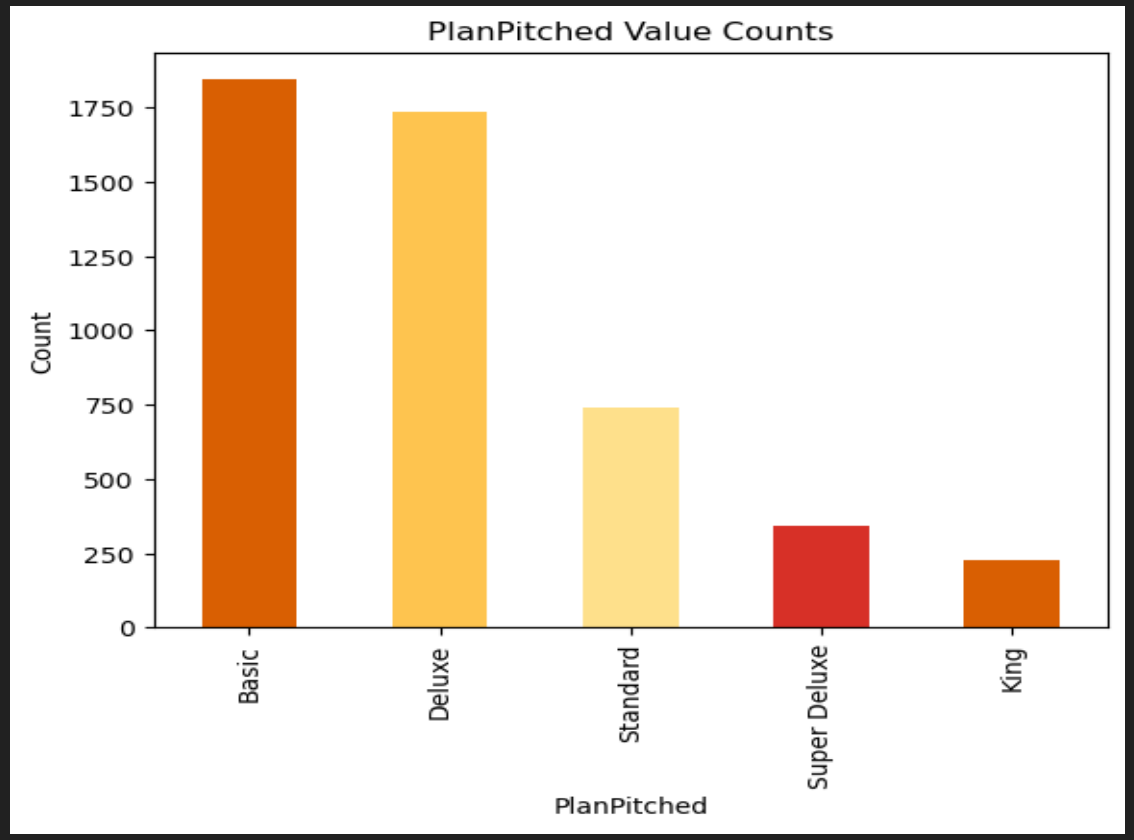

CustomerID 1 0.057 0.032 0.013 0.064 0.6 0.43 0.011 0.22 0.008 -0.036 0.0038 0.51 0.28
PlanTaken 0.057 1 -0.15 0.087 0.078 0.0096 0.11 0.1 0.019 0.26 0.051 -0.012 0.0074 -0.13
Age 0.032 -0.15 1 -0.016 -0.012 0.012 -0.0026 -0.01 0.18 0.033 0.019 0.049 0.0074 0.46
CityTier 0.013 0.087 -0.016 1 0.023 -0.0017 0.024 -0.0092 -0.03 0.0018 -0.042 0.0038 0.00067 0.052
DurationOfPitch 0.064 0.078 -0.012 0.023 1 0.065 0.0094 -0.0066 0.0097 0.033 -0.0029 -0.0016 0.031 -0.0063
NumberOfPersons 0.6 0.0096 0.012 -0.0017 0.065 1 0.33 0.034 0.2 0.011 -0.02 0.01 0.61 0.2
NumberOfFollowups 0.43 0.11 -0.0026 0.024 0.0094 0.33 1 -0.024 0.14 0.005 0.0041 0.012 0.29 0.18
PreferredServiceStar 0.011 0.1 -0.01 -0.0092 -0.0066 0.034 -0.024 1 0.012 0.001 -0.023 0.016 0.036 0.014
NumberOfUpgrades 0.22 0.019 0.18 -0.03 0.0097 0.2 0.14 0.012 1 0.013 -0.0044 -0.012 0.17 0.14
iPhone 0.008 0.26 0.033 0.0018 0.033 0.011 0.005 0.001 0.013 1 0.0029 -0.022 0.02 0.0025
PitchSatisfactionScore -0.036 0.051 0.019 -0.042 -0.0029 -0.02 0.0041 -0.023 -0.0044 0.0029 1 0.069 0.00088 0.03
PhoneContract 0.0038 -0.012 0.049 0.0038 -0.0016 0.01 0.012 0.016 -0.012 -0.022 0.069 1 0.027 0.08
NumberOfChildren 0.51 0.0074 0.0074 0.00067 0.031 0.61 0.29 0.036 0.17 0.02 0.00088 0.027 1 0.2
MonthlyIncome 0.28 -0.13 0.46 0.052 -0.0063 0.2 0.18 0.014 0.14 0.0025 0.03 0.08 0.2 1
CustomerID
PlanTaken
Age
CityTier
DurationOfPitch
NumberOfPersons
NumberOfFollowups
PreferredServiceStar
NumberOfUpgrades
iPhone
PitchSatisfactionScore
PhoneContract
NumberOfChildren
MonthlyIncome
1.0
0.8
0.6
0.4
0.2
0.0

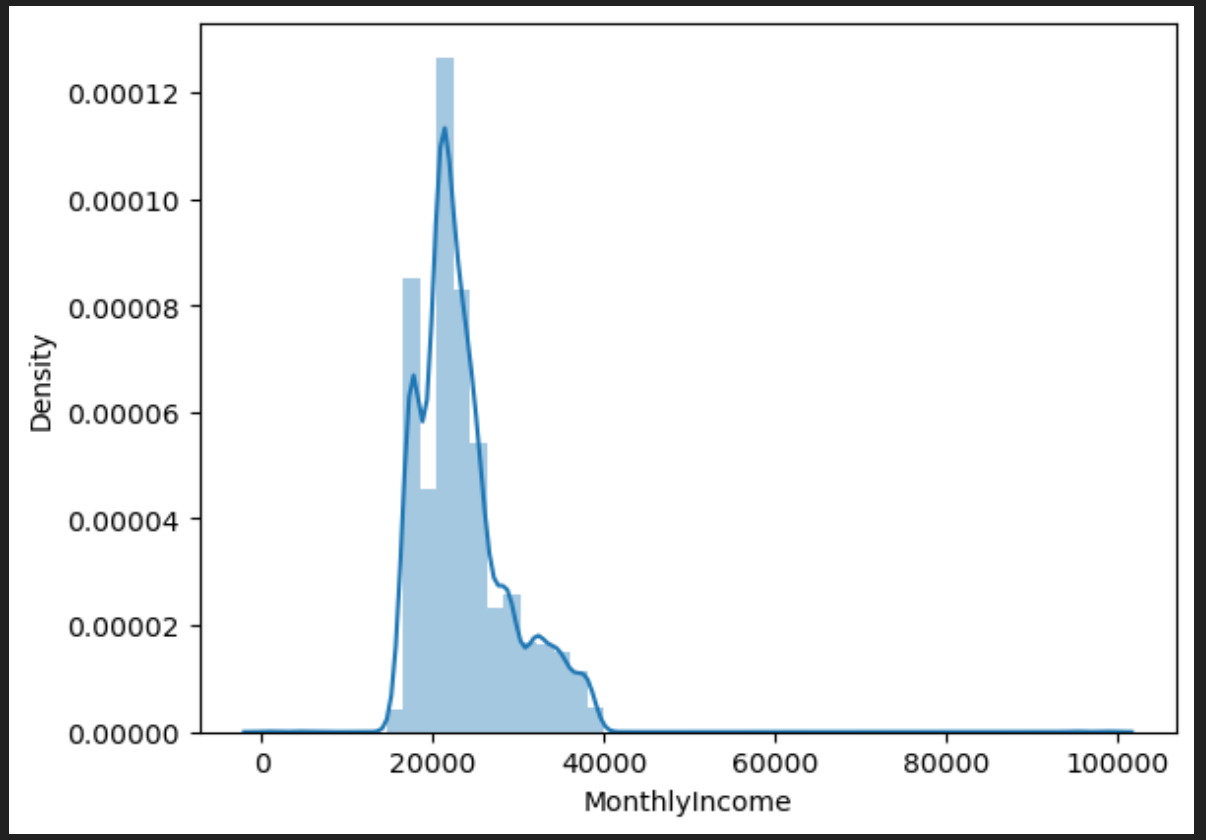
❑ Exploratory Data Analysis :



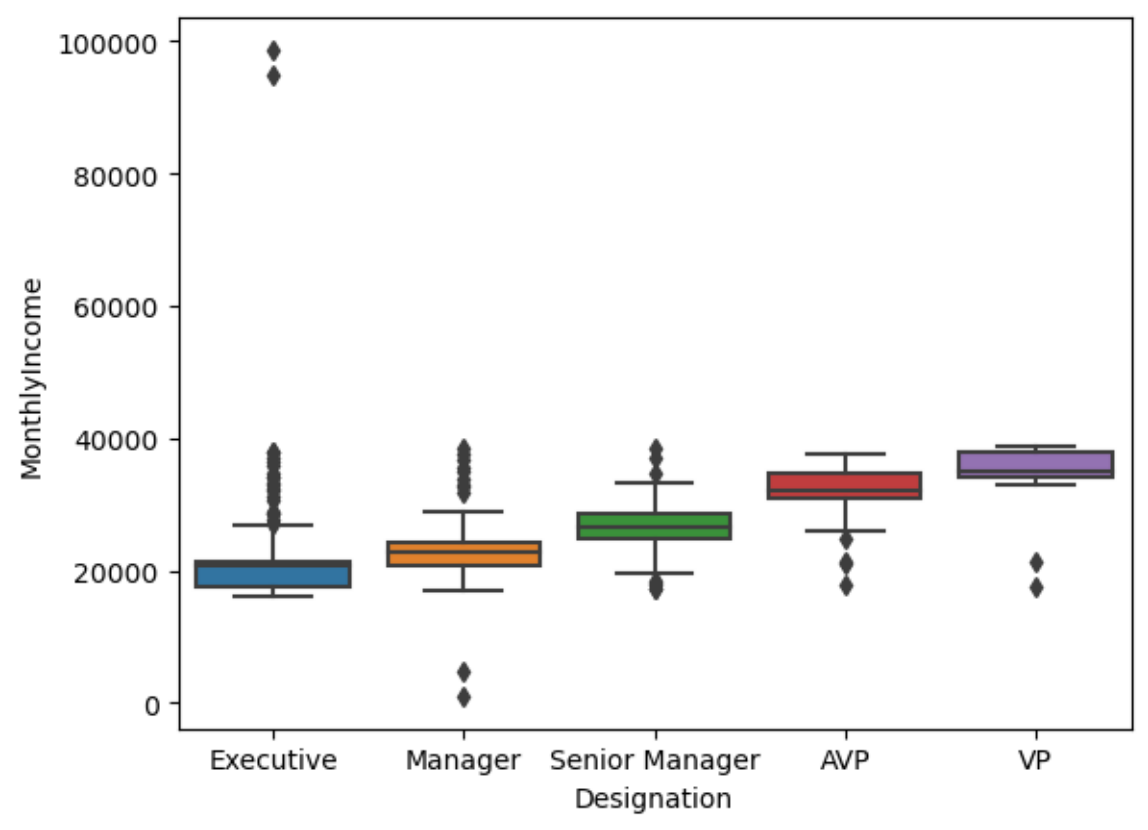


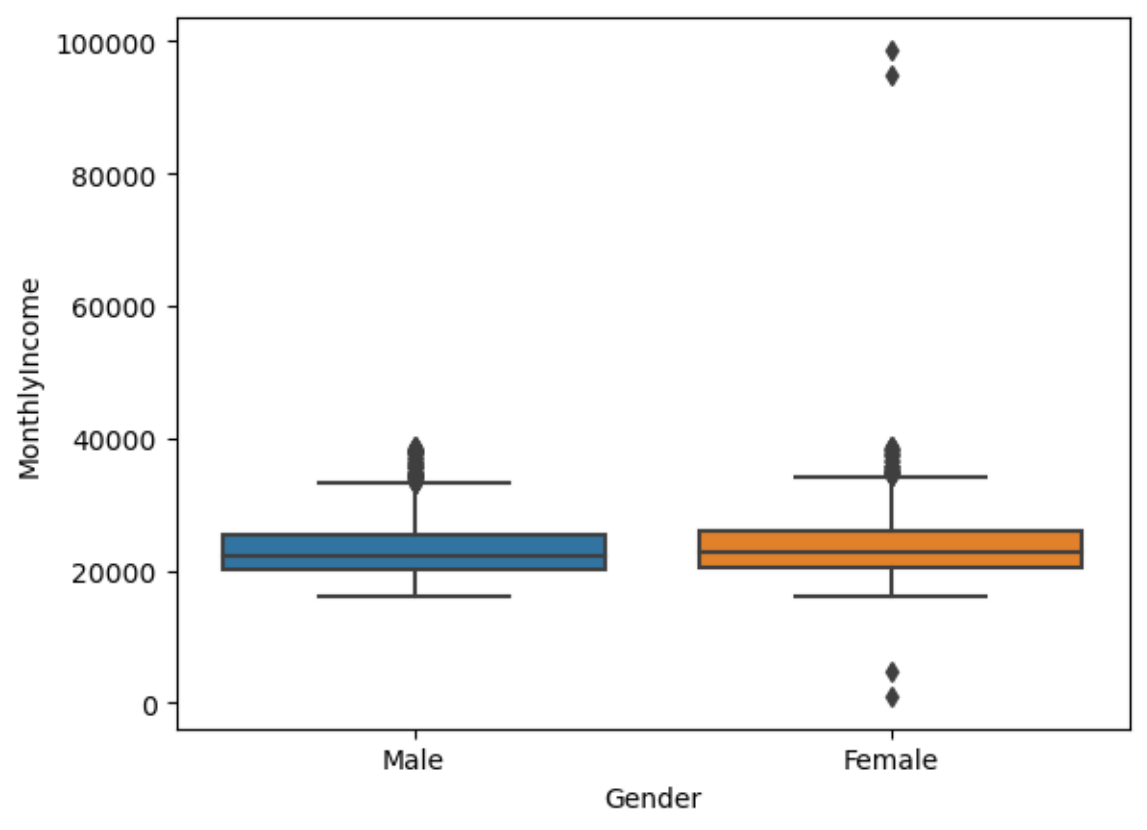
- 1000 people took wellness plan for healthy life style. But nearby 4000 people are not worried about their health.
- In occupation max of people are self employee which is 2000. Small business are nearby 2000 and Large business are below 500 While there are zero freelancer.



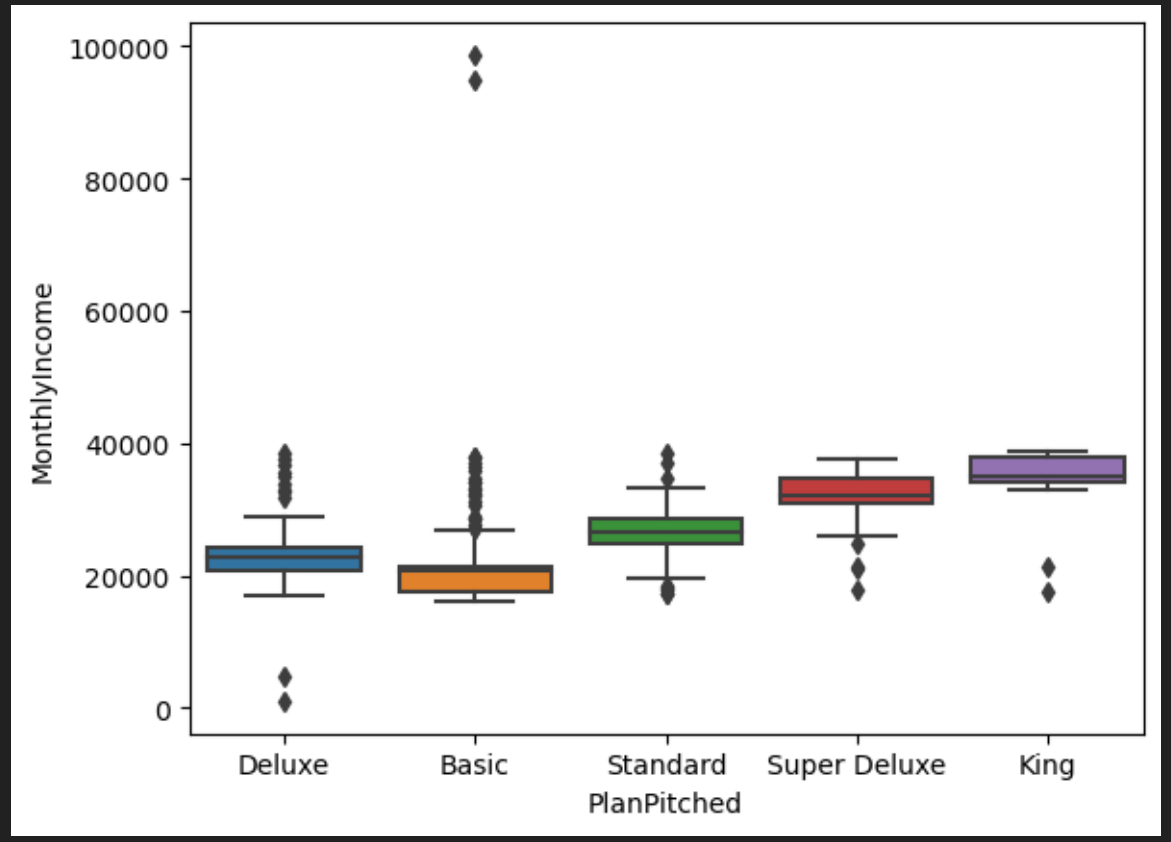


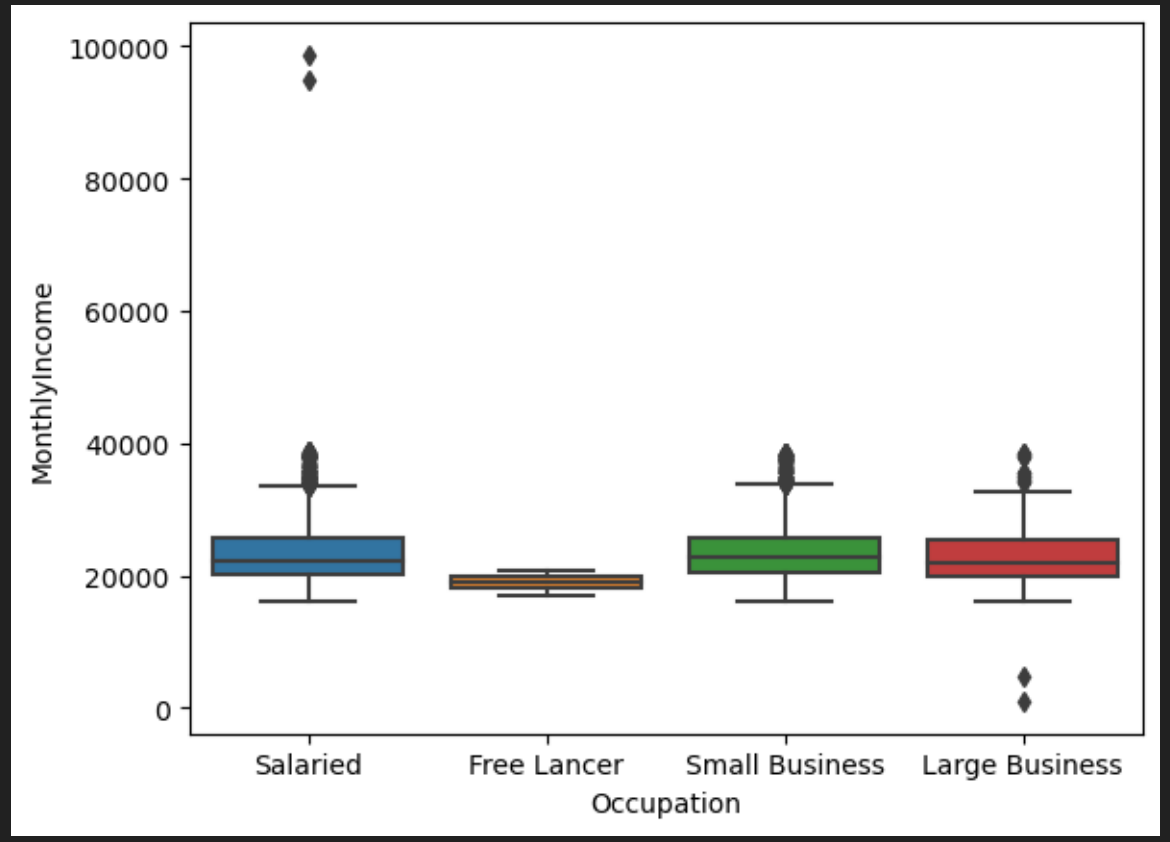
- Salesperson given max plan as basic plan to customers, which is 1750. Deluxe plans are slight less than basic plans, which is 1740. Standard plan are 3 highest given to customer those are 740. Customers aren't interested in super deluxe and king plan.
- density takes peak from monthly income 20000. Later monthly income density reduce at 0.00006 by minor fluctuation. Later onwards it goes to higher peak of 0.000012. After that reduce to monthly income of 40000.



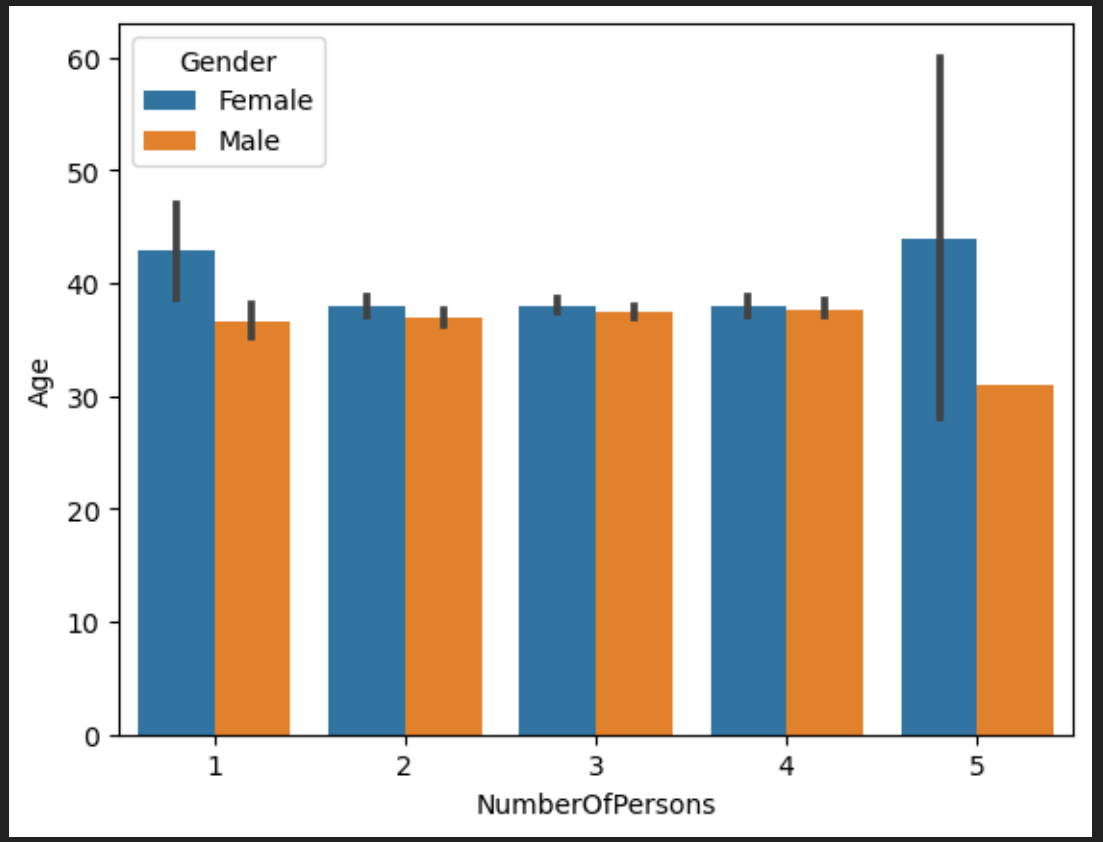


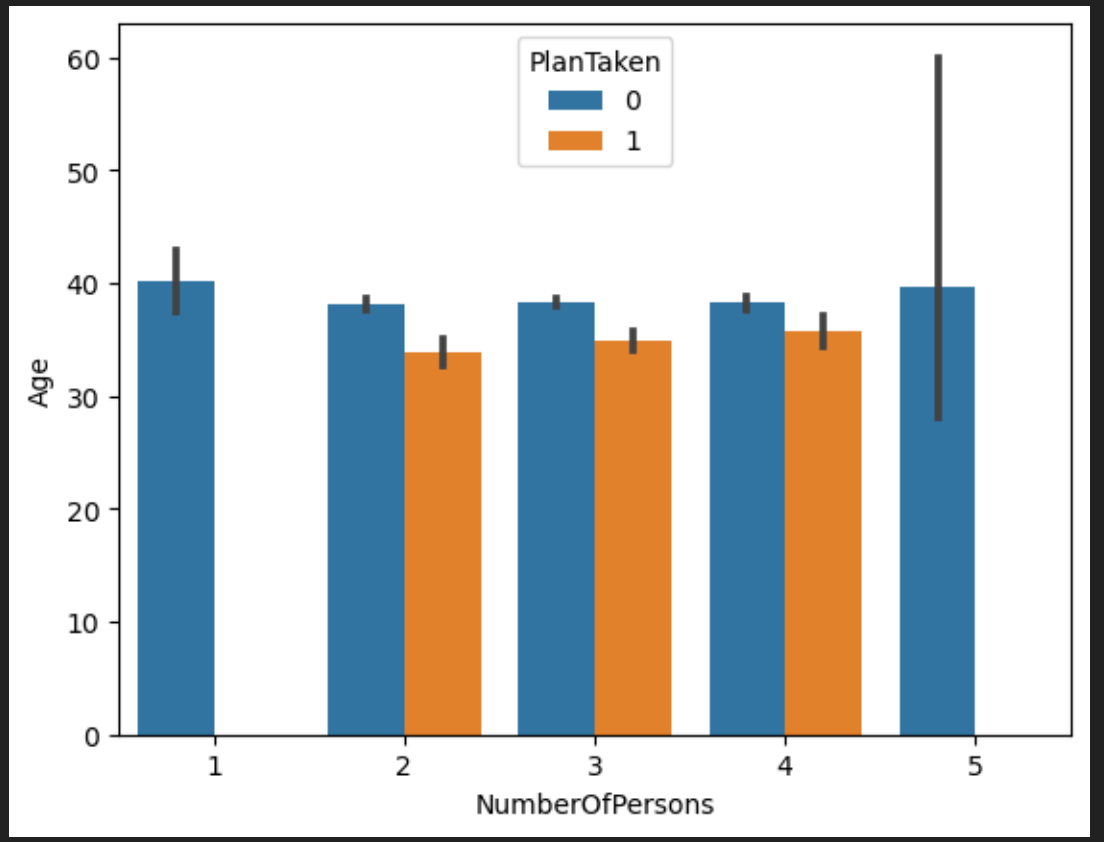
- We have currently different designation, such as executive, manager, senior manager, AVP and VP. In which we have highest income for executive. Also, few manager not satisfy with their income. AVP and VP income are nearby 2000.
- Feamles have a higher monthly income than male, but few of the feamles are not satisfied with their monthly income. Also, male have higher monthly income of 40,000.



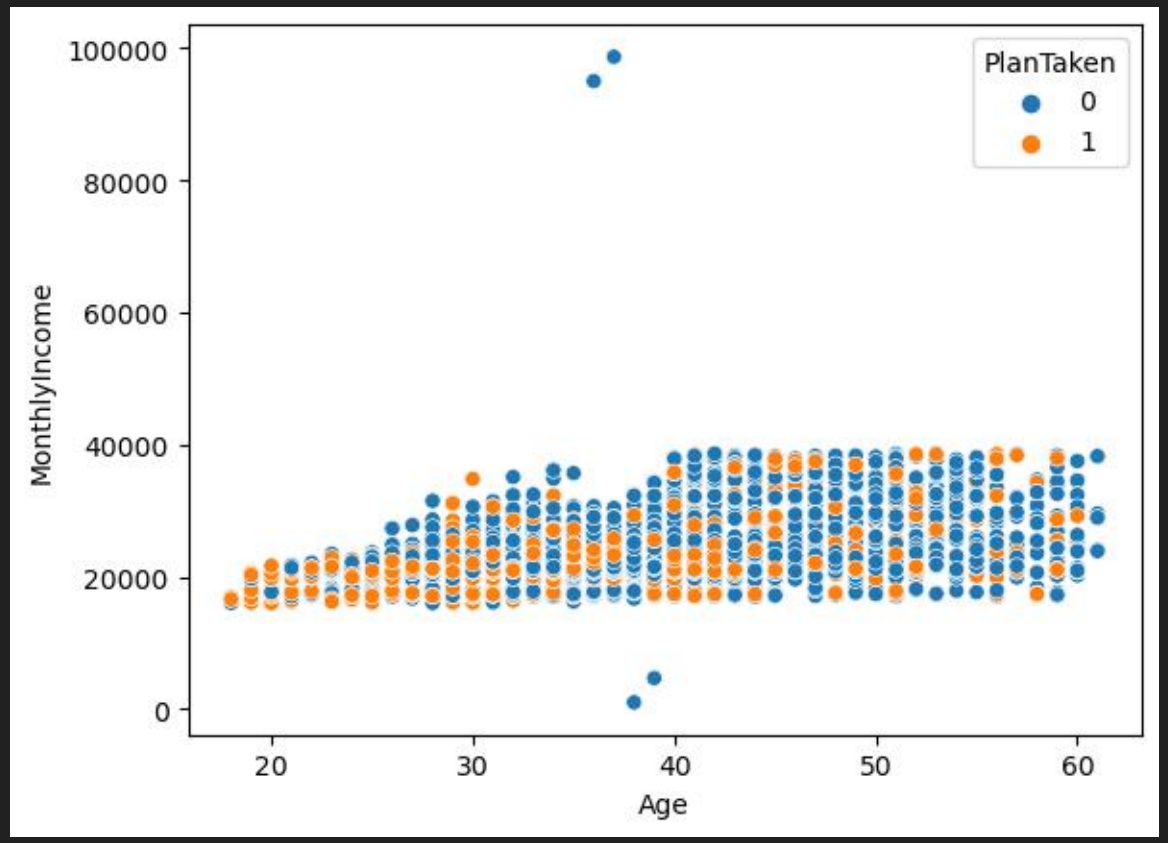


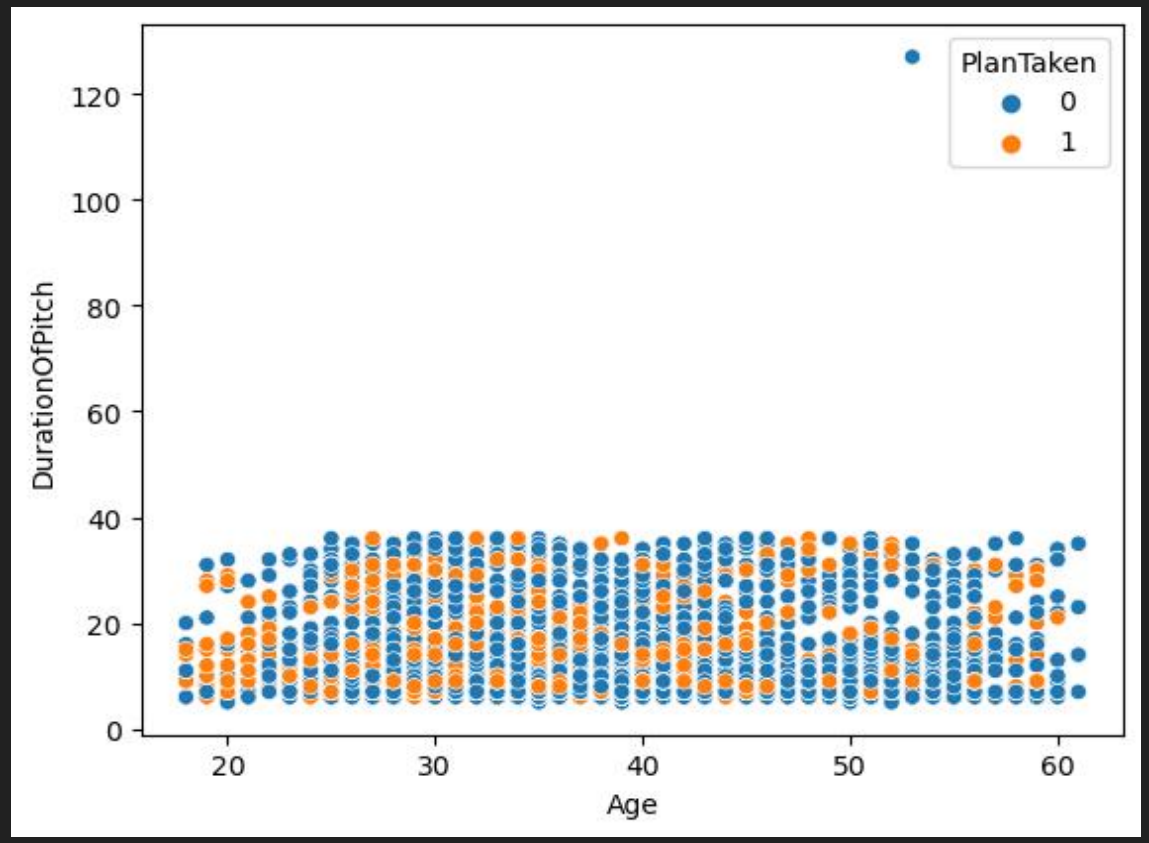
- Sales person have highest plan pitched of basic which is near by 1,00,000. sales person given minimum plan pitched of deluxe, which is near by zero.
- Employees have maximum monthly income, while few of the large businesses are not satisfied with their monthly income.



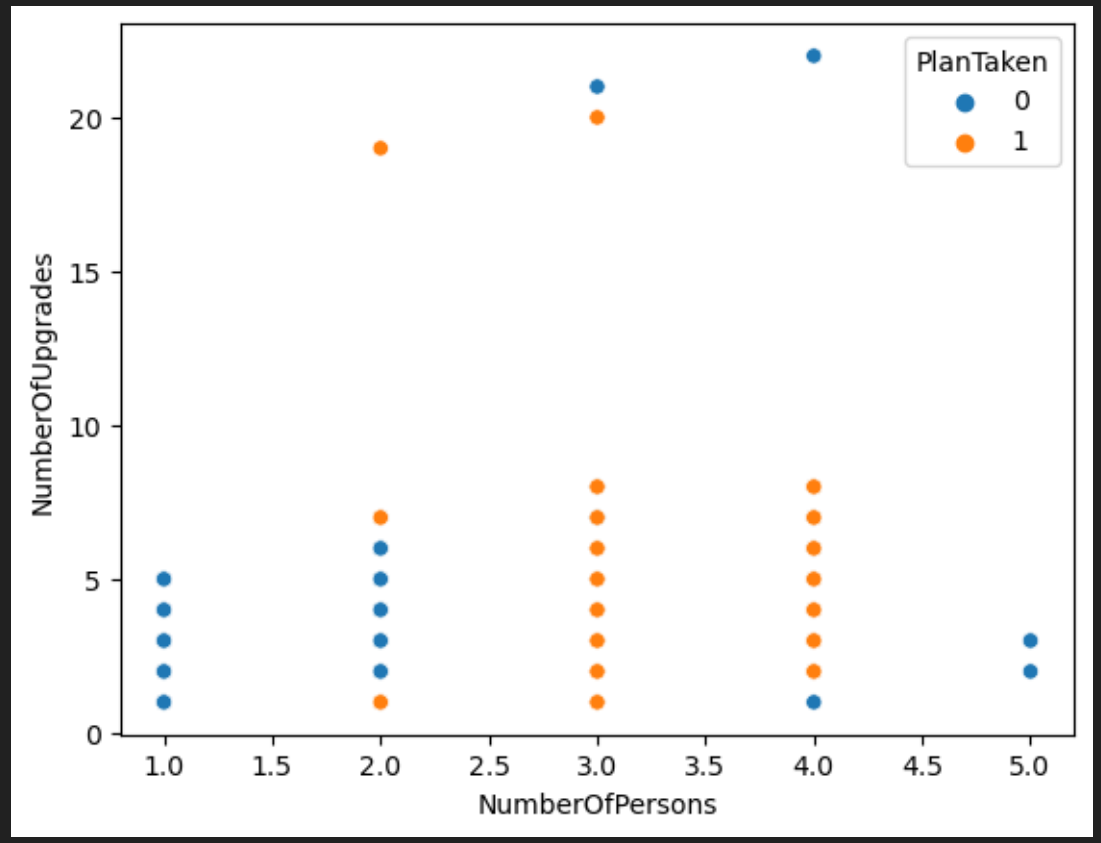


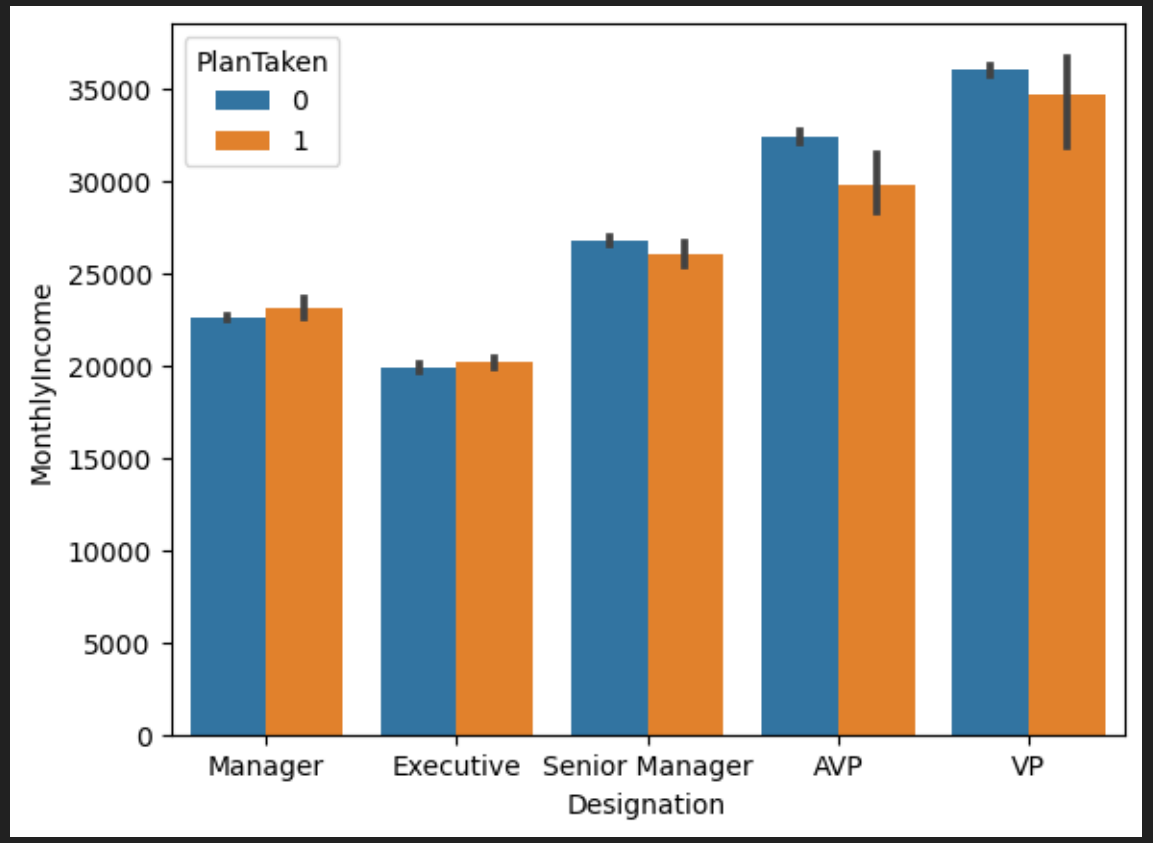
- Most of the customers having age near by 40.
- Most of the customers of age 40 took the plan. However, the customers less than age 40 mostly did not took the plan.



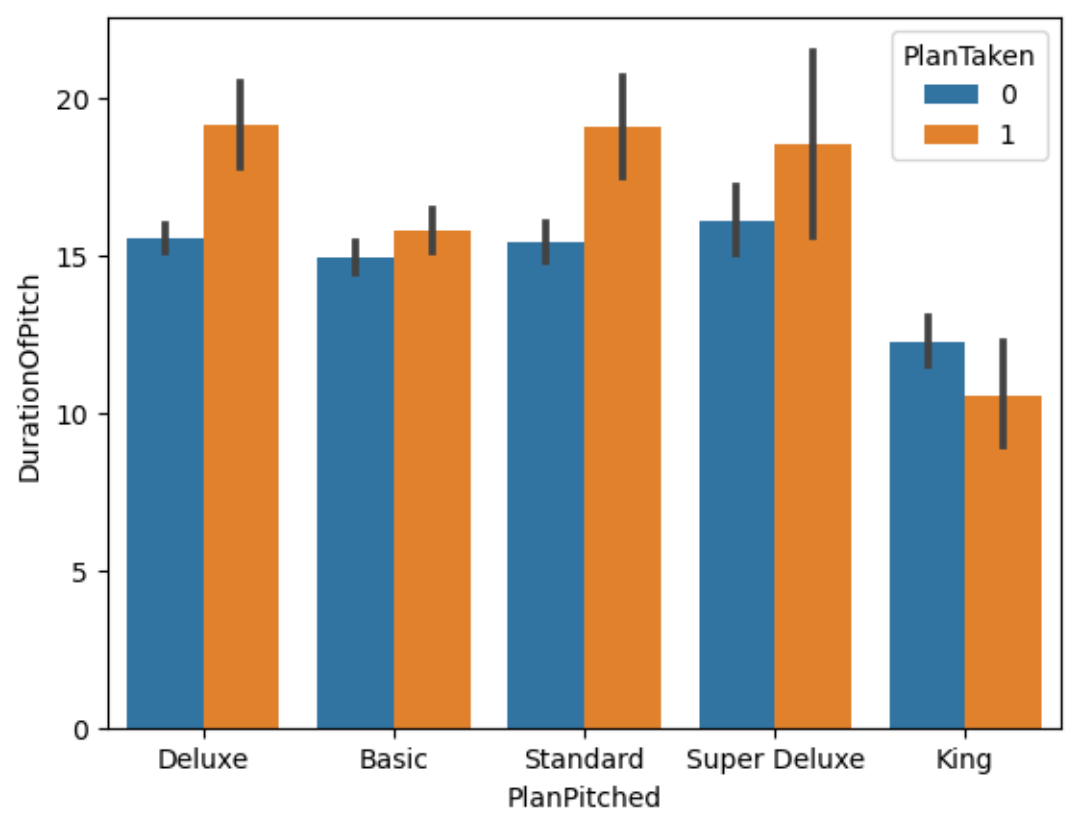


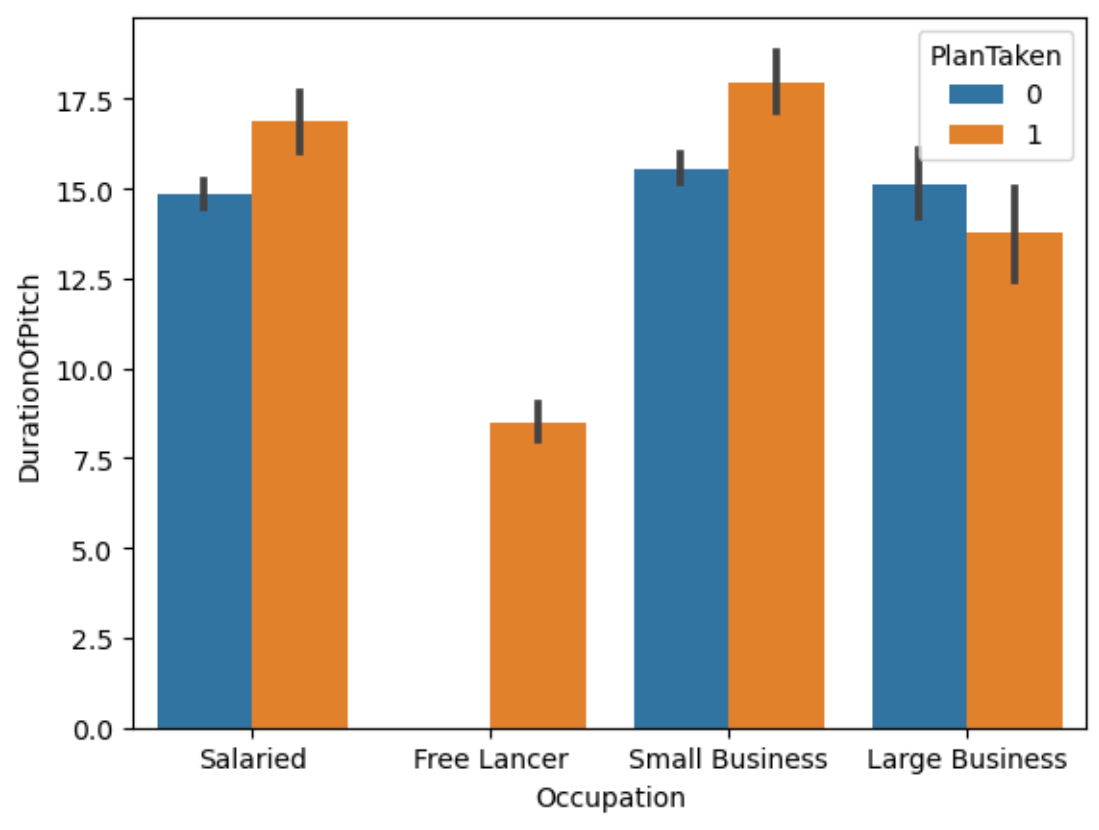
- Customers more than 45 age did not took the plan.
- Above age of 50 customers did not took the plan, even after duration pitch of 40.



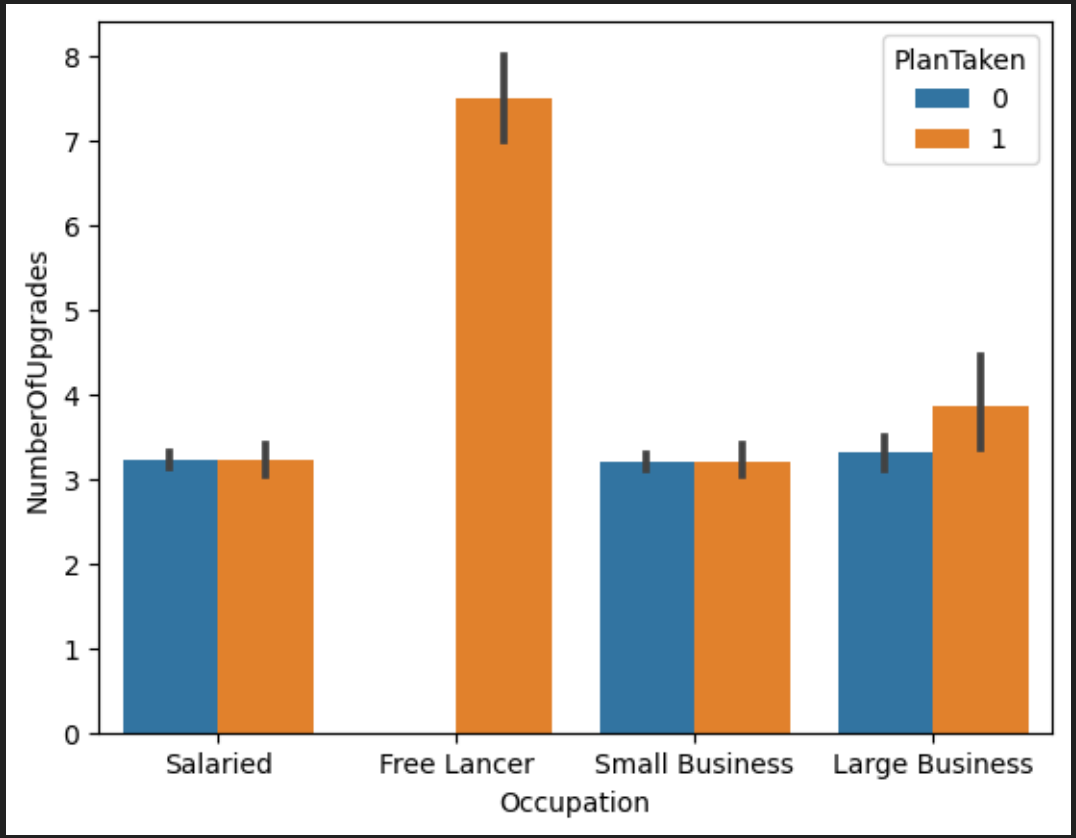


- 3 and 4 customers upgraded their plans for 10 times.
- VP Designation has highest monthly income.

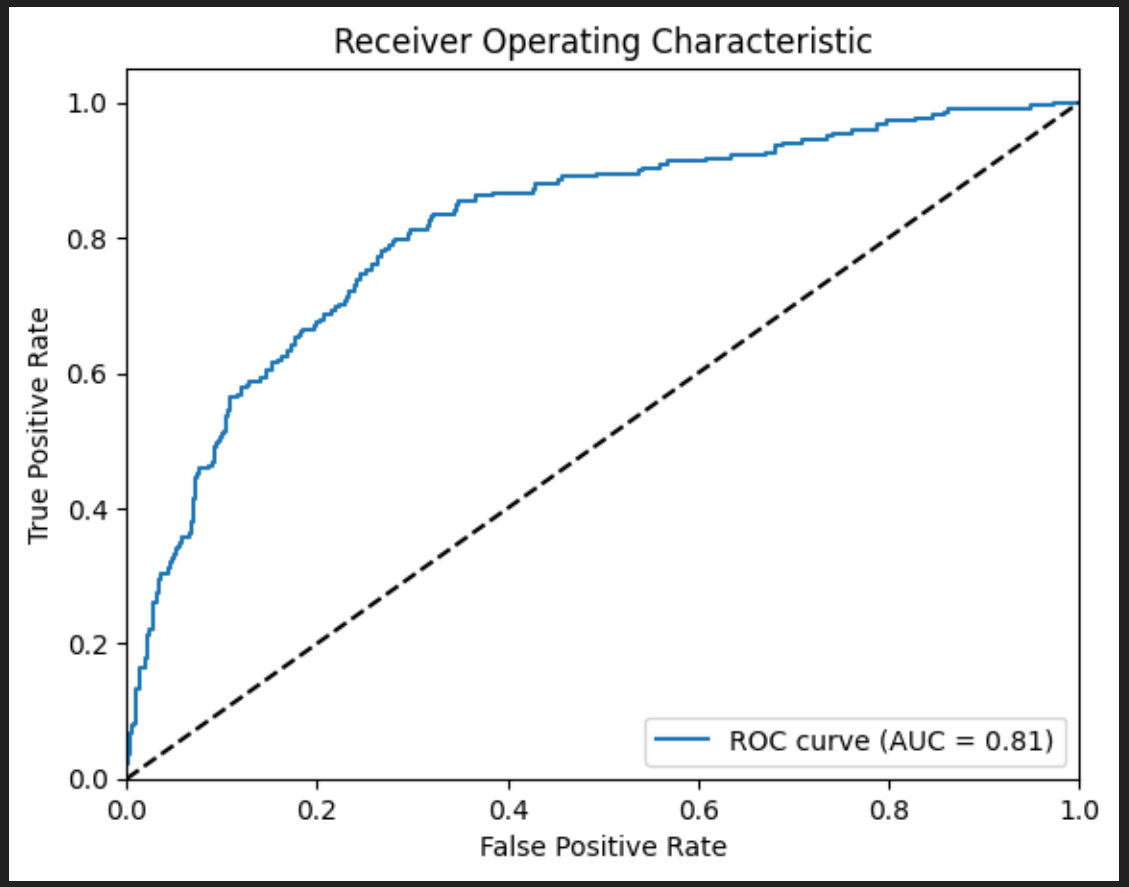


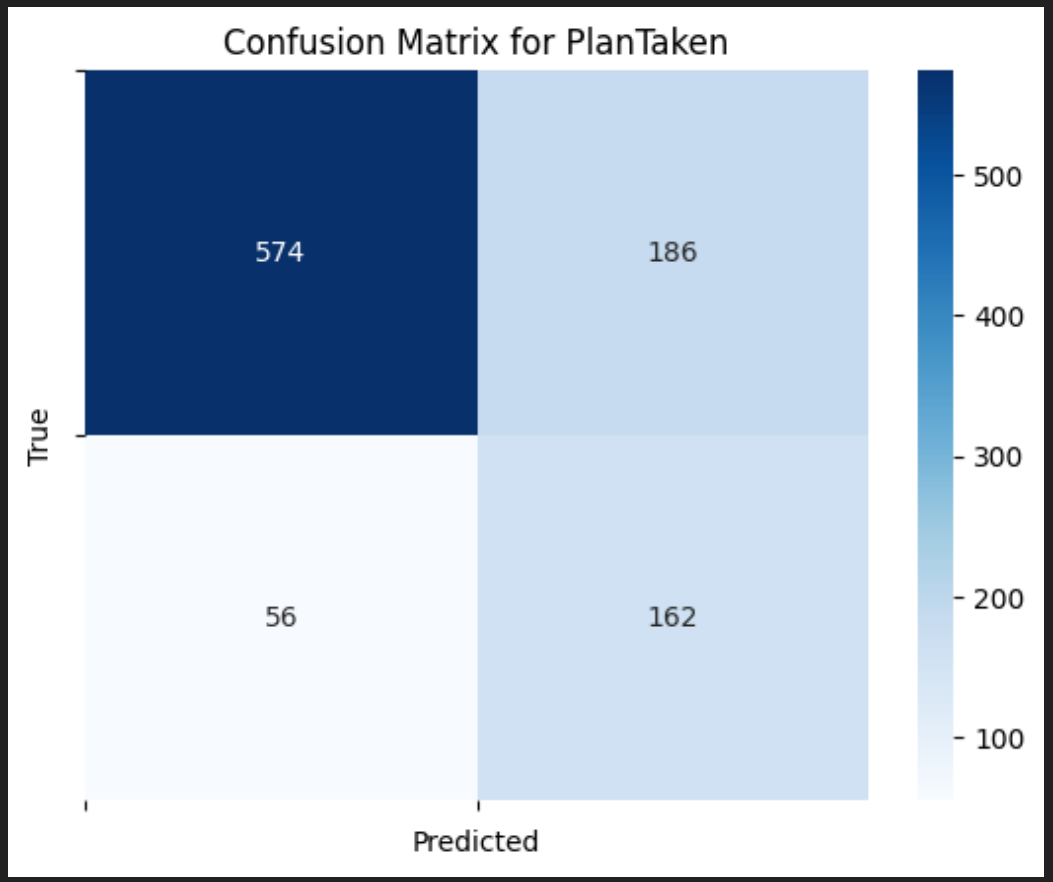


- King plan pitch has the lowest duration pitch compare to other types.
- Free Lancer did not reject any plans.

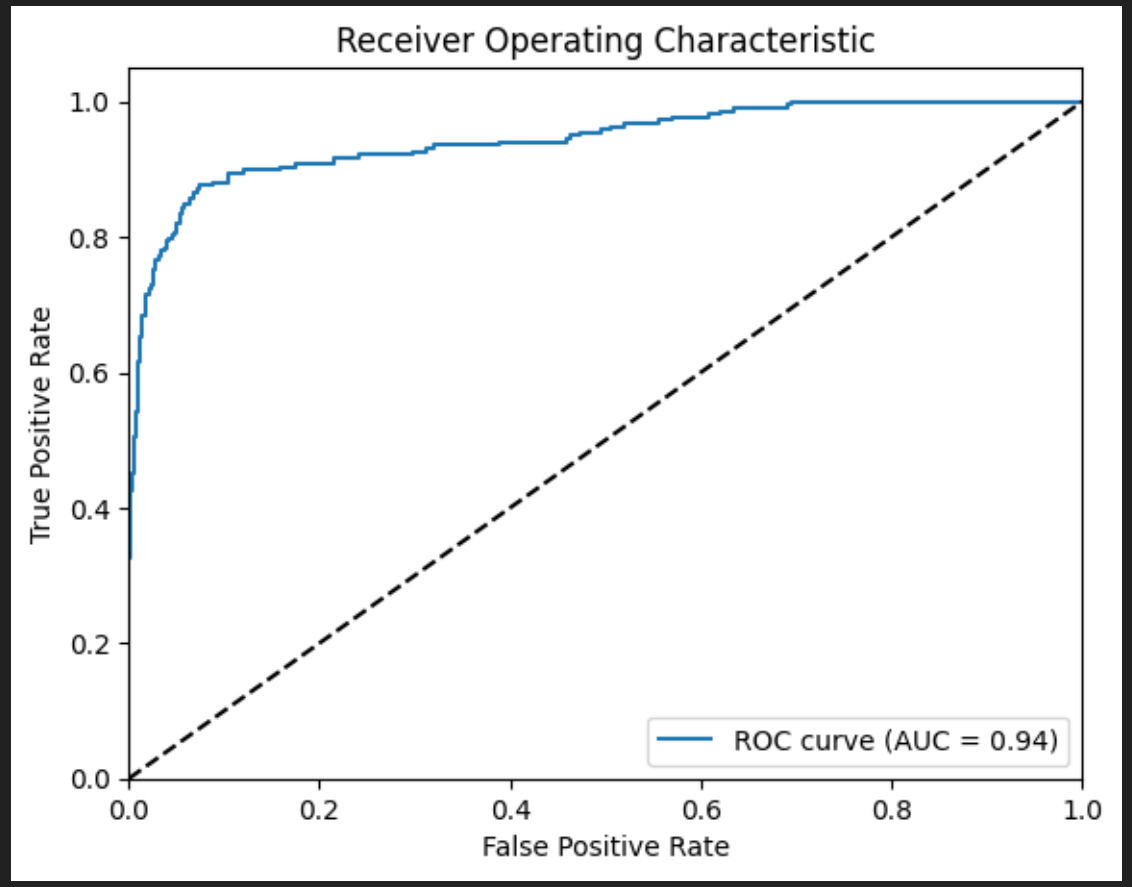


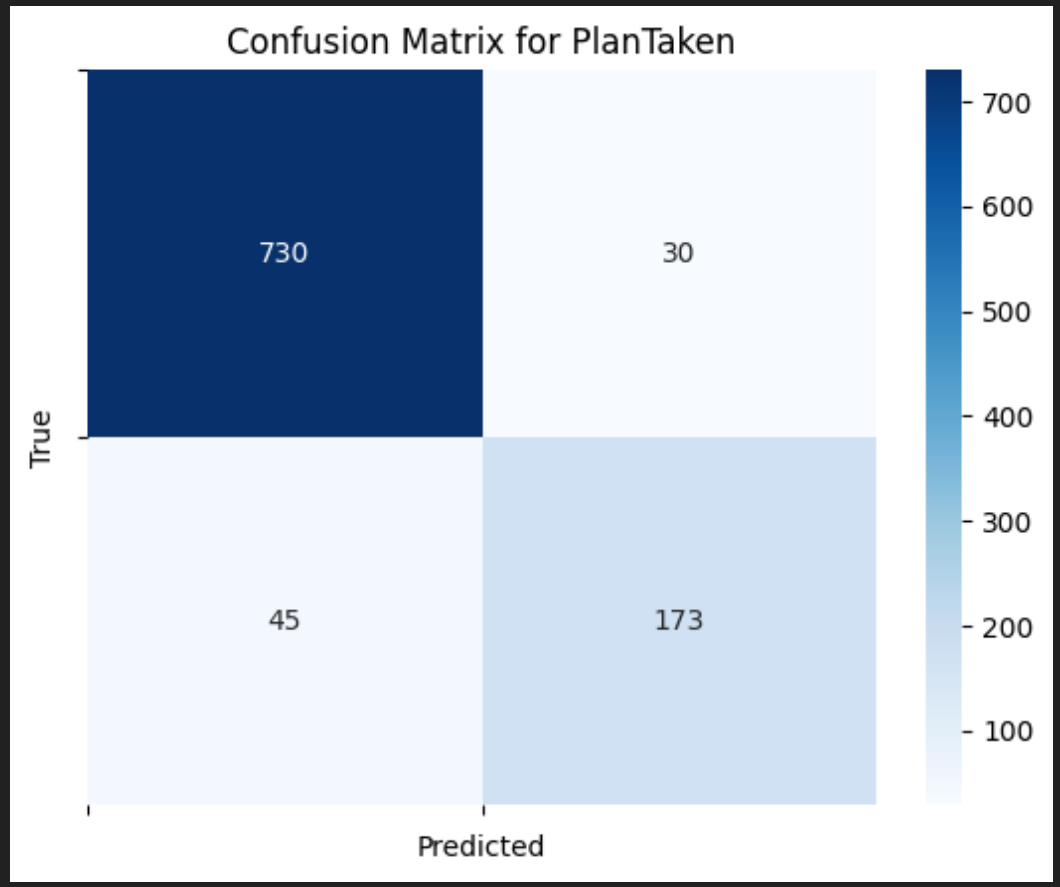
- Free Lancer having the highest number of upgrades and took the plan without any hesitation.





- Final ROC curve and confusion matrix for logistic regression.





- Final ROC curve and confusion matrix for SVM.

END