Automate Exploratory Data Analysis

- pandas profiling
- sweetviz
- AutoViz

In [2]:
```
import pandas as pd
import numpy as np
```

In [3]:
```
data = pd.read_csv(r"C:\Users\shail\Google Drive\Self\Learning\iNuerons\Project\Dataset\data1
data.head()
```

Out[3]:

	ProdTaken	Age	TypeofContact	CityTier	DurationOfPitch	Occupation	Gender	NumberOfPersonVisiting
0	1	41.0	Self Enquiry	3	6.0	Salaried	Female	3
1	0	49.0	Company Invited	1	14.0	Salaried	Male	3
2	1	37.0	Self Enquiry	1	8.0	Free Lancer	Male	3
3	0	33.0	Company Invited	1	9.0	Salaried	Female	2
4	0	36.0	Self Enquiry	1	8.0	Small Business	Male	2

In [4]:
```
data.shape
```

Out[4]:
```
(4888, 19)
```

Pandas Profiling

In [5]:
```
import pandas_profiling
```

In [6]:
```
from pandas_profiling import profile_report
```

In [7]:
```
data.profile_report()
```

Overview

Dataset statistics

Number of variables	19
Number of observations	4888
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	141
Duplicate rows (%)	2.9%
Total size in memory	725.7 KiB
Average record size in memory	152.0 B

Variable types

Categorical	14
Numeric	5

Alerts

Dataset has 141 (2.9%) duplicate rows	**Duplicates**
`NumberOfPersonVisiting` is highly correlated with `NumberOfChildrenVisiting`	**High correlation**
`NumberOfChildrenVisiting` is highly correlated with	**High correlation**

Out[7]:

sweetviz

In [ ]:
```
pip install sweetviz
```

In [9]:
```
import sweetviz as sv
analyze_report = sv.analyze(data)
analyze_report.show_notebook()
```

In [ ]:
```
pip install autoviz
```

In [11]:
```
from autoviz.AutoViz_Class import AutoViz_Class
AV = AutoViz_Class()
df = AV.AutoViz(r"C:\Users\shail\Google Drive\Self\Learning\iNuerons\Project\Dataset\data1\Tr
```

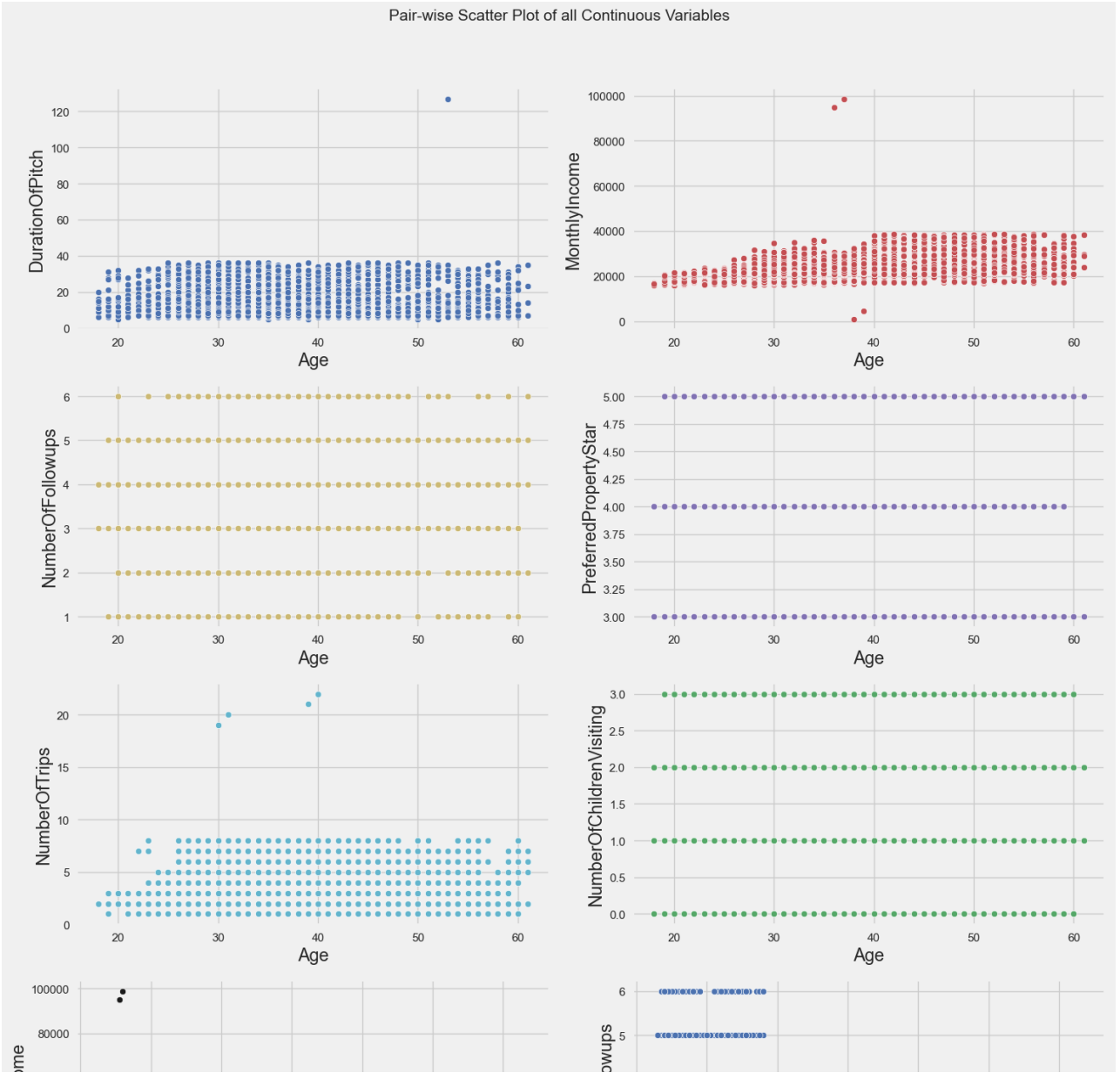
```
Imported v0.1.55. After importing, execute '%matplotlib inline' to display charts in Jupyter.
    AV = AutoViz_Class()
    dfte = AV.AutoViz(filename, sep=',', depVar='', dfte=None, header=0, verbose=1, lowess=Fa
lse,
                chart_format='svg',max_rows_analyzed=150000,max_cols_analyzed=30, save_plot_di
```

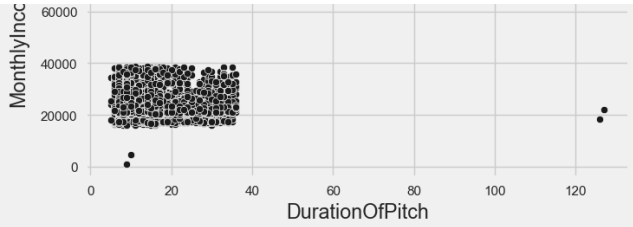
```
r=None)
Update: verbose=0 displays charts in your local Jupyter notebook.
        verbose=1 additionally provides EDA data cleaning suggestions. It also displays chart
s.
        verbose=2 does not display charts but saves them in AutoViz_Plots folder in local mac
hine.
        chart_format='bokeh' displays charts in your local Jupyter notebook.
        chart_format='server' displays charts in your browser: one tab for each chart type
        chart_format='html' silently saves interactive HTML files in your local machine

Shape of your Data Set loaded: (4888, 20)
############################################################################################
######################## C L A S S I F Y I N G  V A R I A B L E S  ##########################
############################################################################################
Classifying variables in data set...
```

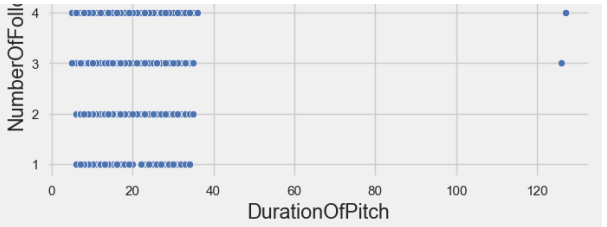
	Nuniques	dtype	Nulls	Nullpercent	NuniquePercent	Value counts Min	Data cleaning improvement suggestions
CustomerID	4888	int64	0	0.000000	100.000000	0	possible ID column: drop
MonthlyIncome	2475	float64	233	4.766776	50.634206	0	fill missing values, skewed column: cap or drop possible outliers
Age	44	float64	226	4.623568	0.900164	0	fill missing values
DurationOfPitch	34	float64	251	5.135025	0.695581	0	fill missing values, skewed column: cap or drop possible outliers
NumberOfTrips	12	float64	140	2.864157	0.245499	0	fill missing values, skewed column: cap or drop possible outliers
NumberOfFollowups	6	float64	45	0.920622	0.122750	0	fill missing values
Designation	5	object	0	0.000000	0.102291	230	
PitchSatisfactionScore	5	int64	0	0.000000	0.102291	0	
ProductPitched	5	object	0	0.000000	0.102291	230	
NumberOfPersonVisiting	5	int64	0	0.000000	0.102291	0	
MaritalStatus	4	object	0	0.000000	0.081833	682	
Occupation	4	object	0	0.000000	0.081833	2	
NumberOfChildrenVisiting	4	float64	66	1.350245	0.081833	0	fill missing values
Gender	3	object	0	0.000000	0.061375	155	
CityTier	3	int64	0	0.000000	0.061375	0	

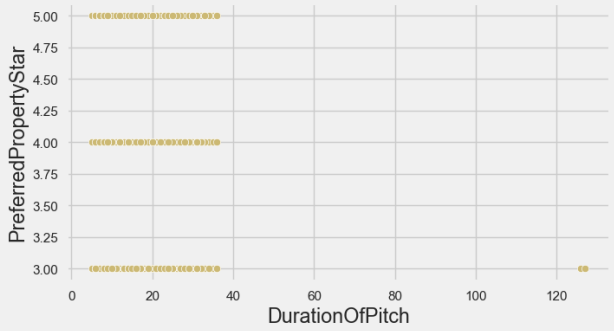
	Nuniques	dtype	Nulls	Nullpercent	NuniquePercent	Value counts Min	Data cleaning improvement suggestions
PreferredPropertyStar	3	float64	26	0.531915	0.061375	0	fill missing values
ProdTaken	2	int64	0	0.000000	0.040917	0	
TypeofContact	2	object	25	0.511457	0.040917	1419	fill missing values, fix mixed data types
Passport	2	int64	0	0.000000	0.040917	0	
OwnCar	2	int64	0	0.000000	0.040917	0	

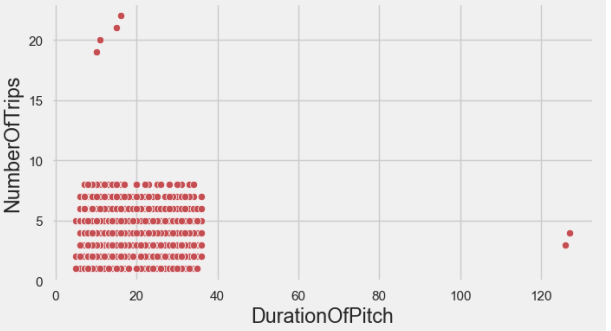
```
    20 Predictors classified...
        1 variables removed since they were ID or low-information variables
        List of variables removed: ['CustomerID']
Number of All Scatter Plots = 28
All Plots done
Time to run AutoViz = 6 seconds

 ###################### AUTO VISUALIZATION Completed ########################
```

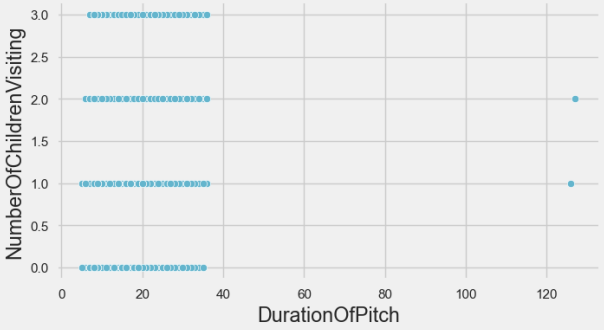



MonthlyInco
60000
40000
20000
0
0
20
40
60
80
100
120
DurationOfPitch

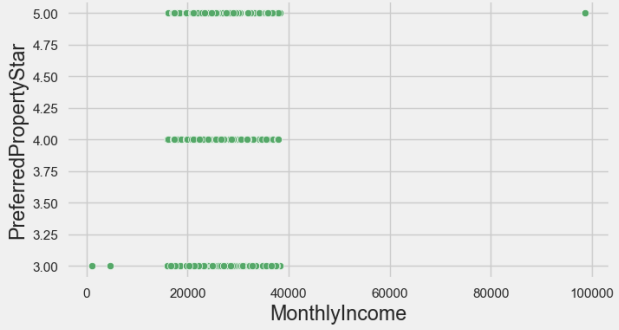

NumberOfFoll
4
3
2
1
0
20
40
60
80
100
120
DurationOfPitch

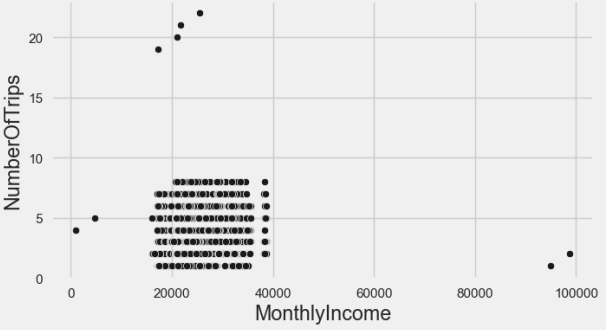

PreferredPropertyStar
5.00
4.75
4.50
4.25
4.00
3.75
3.50
3.25
3.00
0
20
40
60
80
100
120
DurationOfPitch

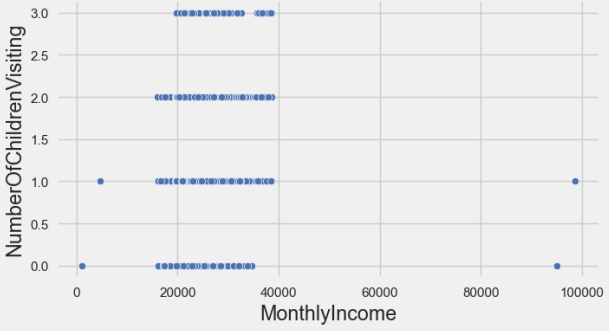

NumberOfTrips
20
15
10
5
0
0
20
40
60
80
100
120
DurationOfPitch

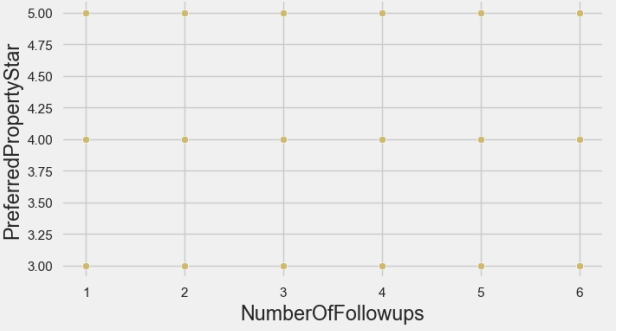

NumberOfChildrenVisiting
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
20
40
60
80
100
120
DurationOfPitch

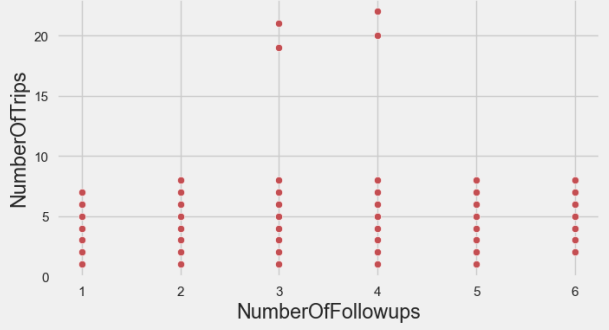

NumberOfFollowups
6
5
4
3
2
1
0
20000
40000
60000
80000
100000
MonthlyIncome

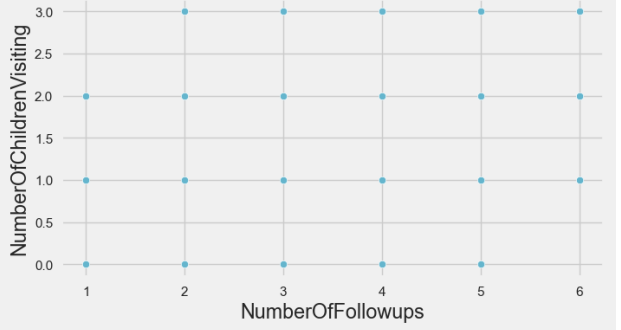

PreferredPropertyStar
5.00
4.75
4.50
4.25
4.00
3.75
3.50
3.25
3.00
0
20000
40000
60000
80000
100000
MonthlyIncome

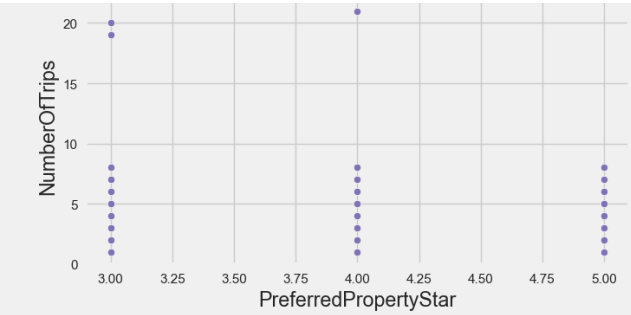

NumberOfTrips
20
15
10
5
0
0
20000
40000
60000
80000
100000
MonthlyIncome

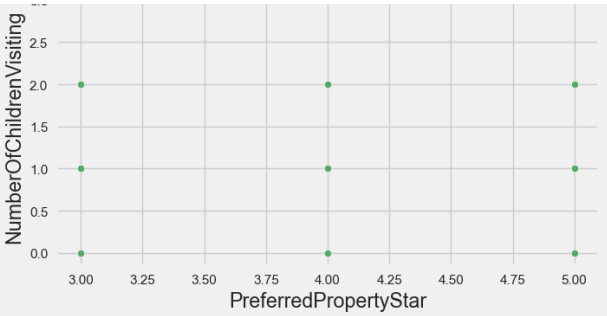

NumberOfChildrenVisiting
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
20000
40000
60000
80000
100000
MonthlyIncome

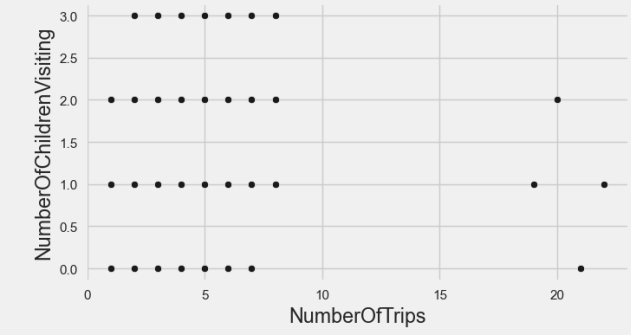

PreferredPropertyStar
5.00
4.75
4.50
4.25
4.00
3.75
3.50
3.25
3.00
1
2
3
4
5
6
NumberOfFollowups

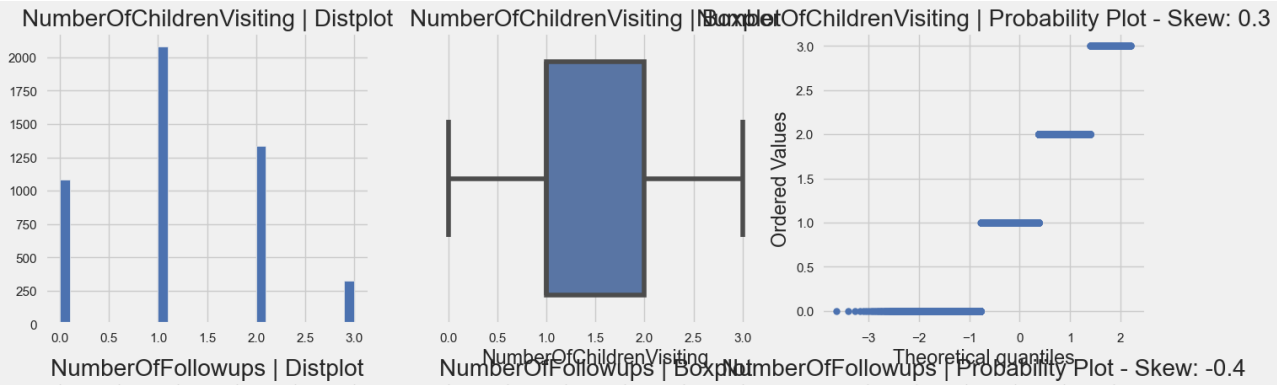

NumberOfTrips
20
15
10
5
0
1
2
3
4
5
6
NumberOfFollowups

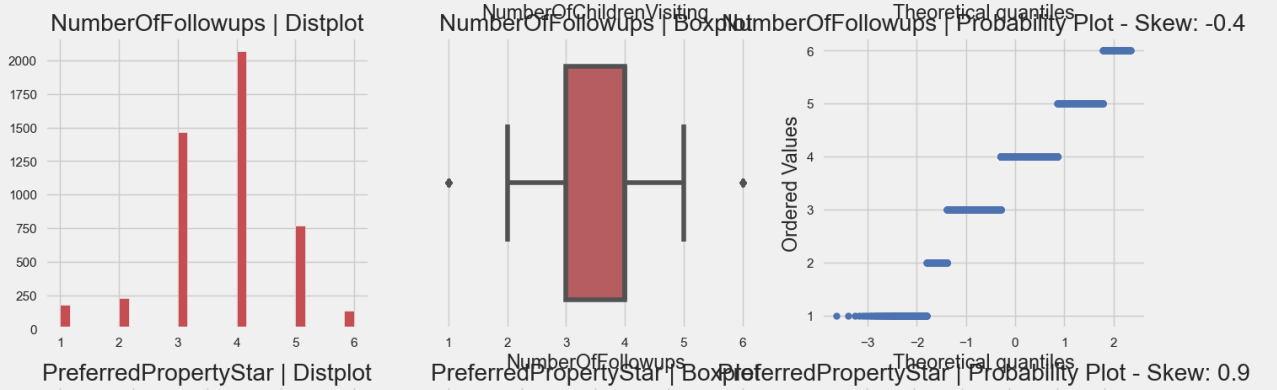

NumberOfChildrenVisiting
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
6
NumberOfFollowups

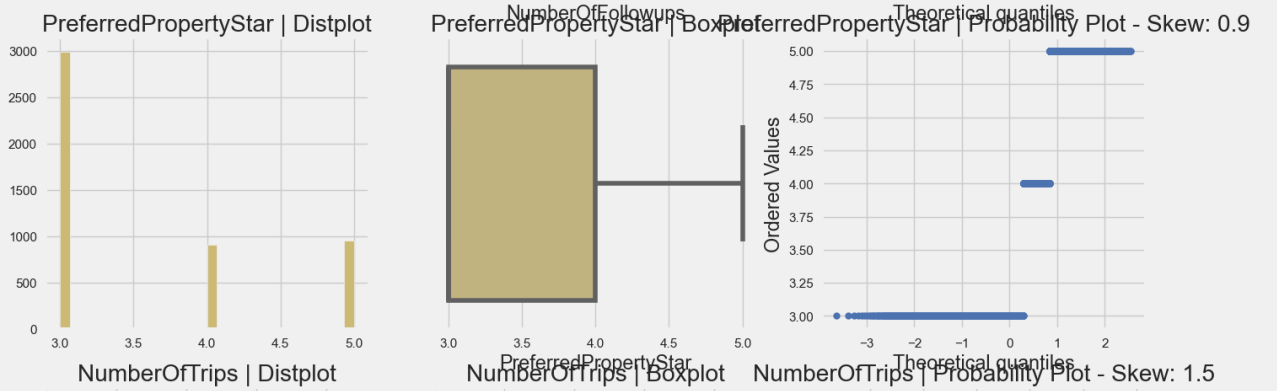

NumberOfTrips
PreferredPropertyStar


NumberOfChildrenVisiting
PreferredPropertyStar

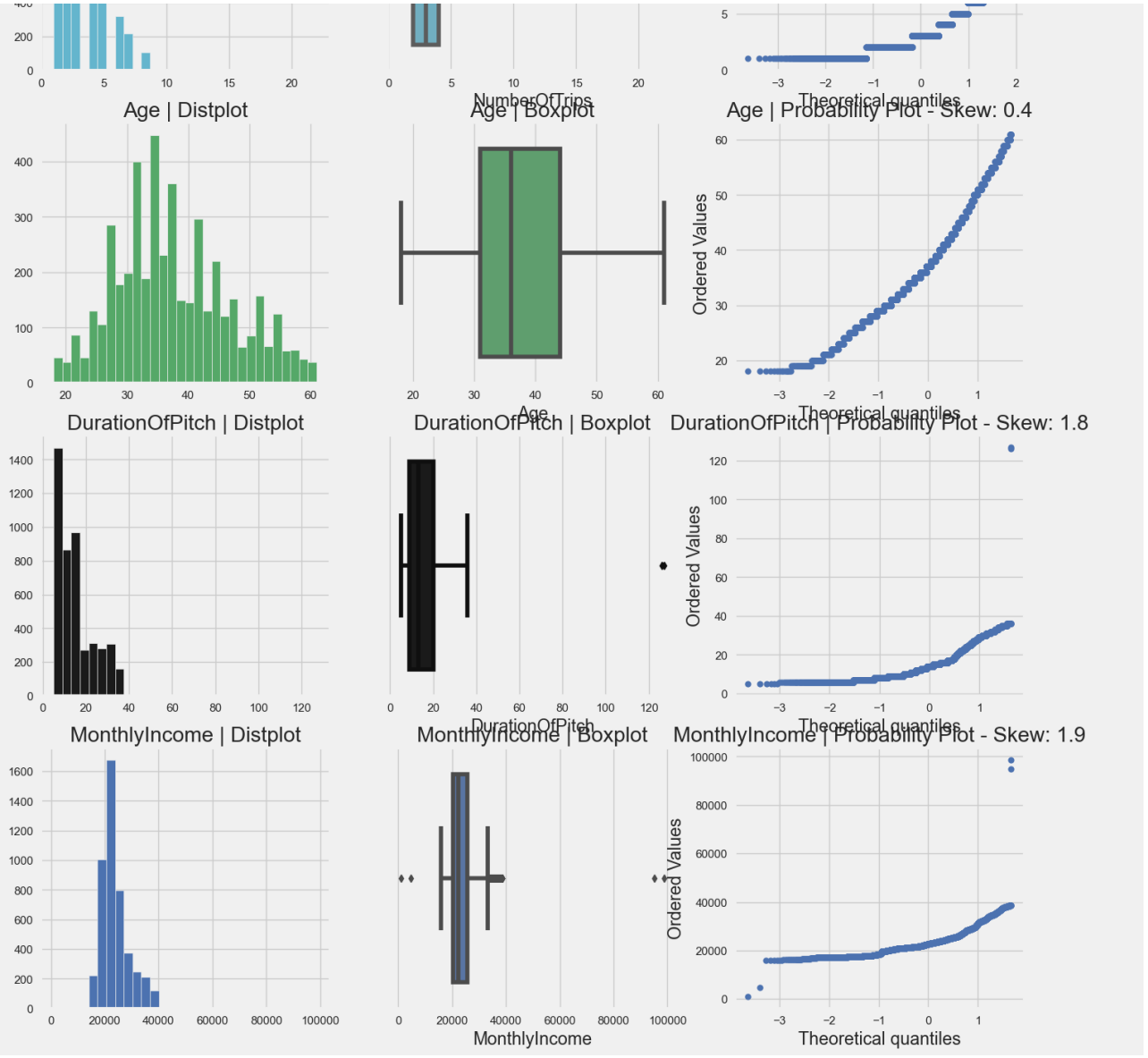

NumberOfChildrenVisiting
NumberOfTrips


NumberOfChildrenVisiting | Distplot
NumberOfChildrenVisiting | Boxplot
NumberOfChildrenVisiting | Probability Plot - Skew: 0.3
Ordered Values
Theoretical quantiles

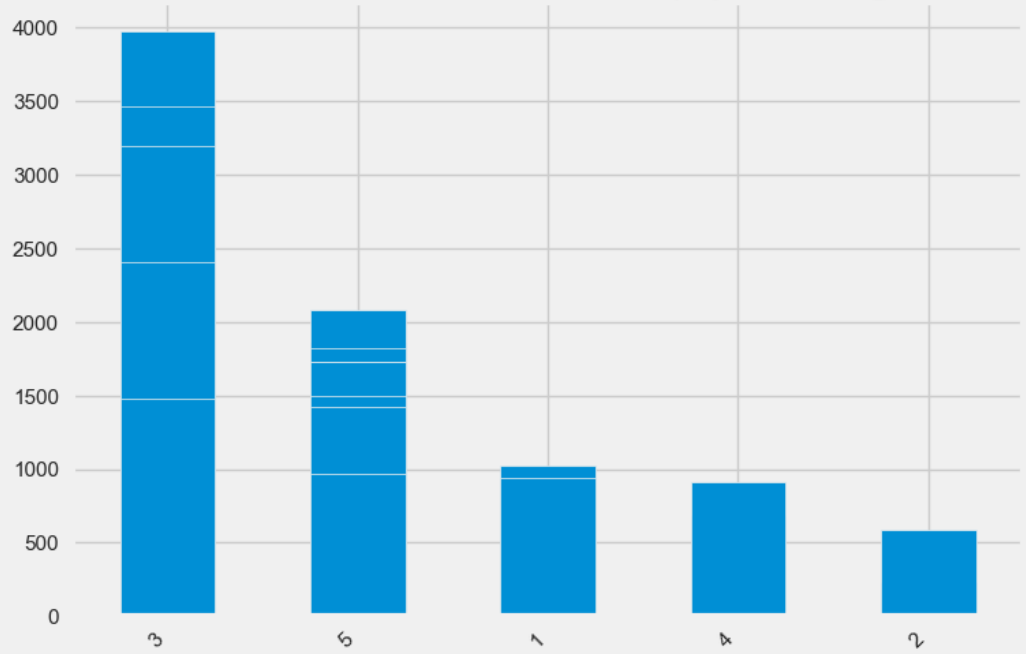

NumberOfFollowups | Distplot
NumberOfFollowups | Boxplot
NumberOfFollowups | Probability Plot - Skew: -0.4
Ordered Values
Theoretical quantiles

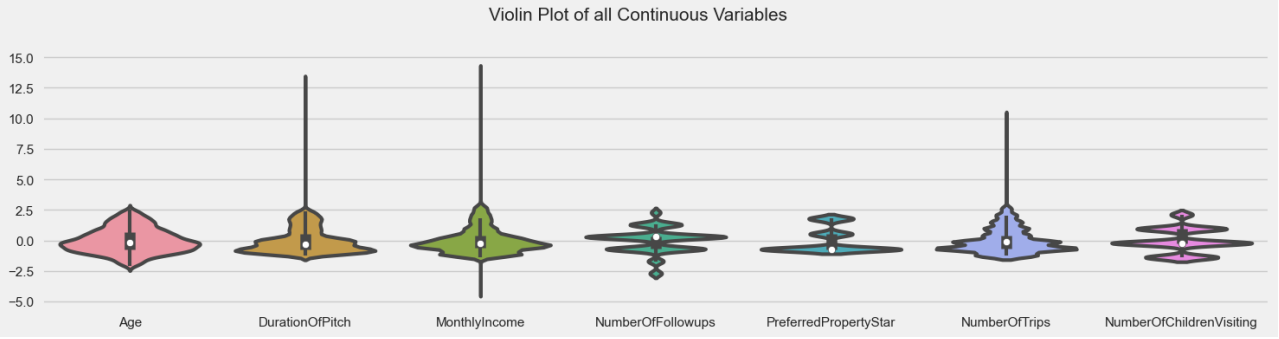

PreferredPropertyStar | Distplot
PreferredPropertyStar | Boxplot
PreferredPropertyStar | Probability Plot - Skew: 0.9
Ordered Values
Theoretical quantiles


NumberOfTrips | Distplot
NumberOfTrips | Boxplot
NumberOfTrips | Probability Plot - Skew: 1.5
Ordered Values

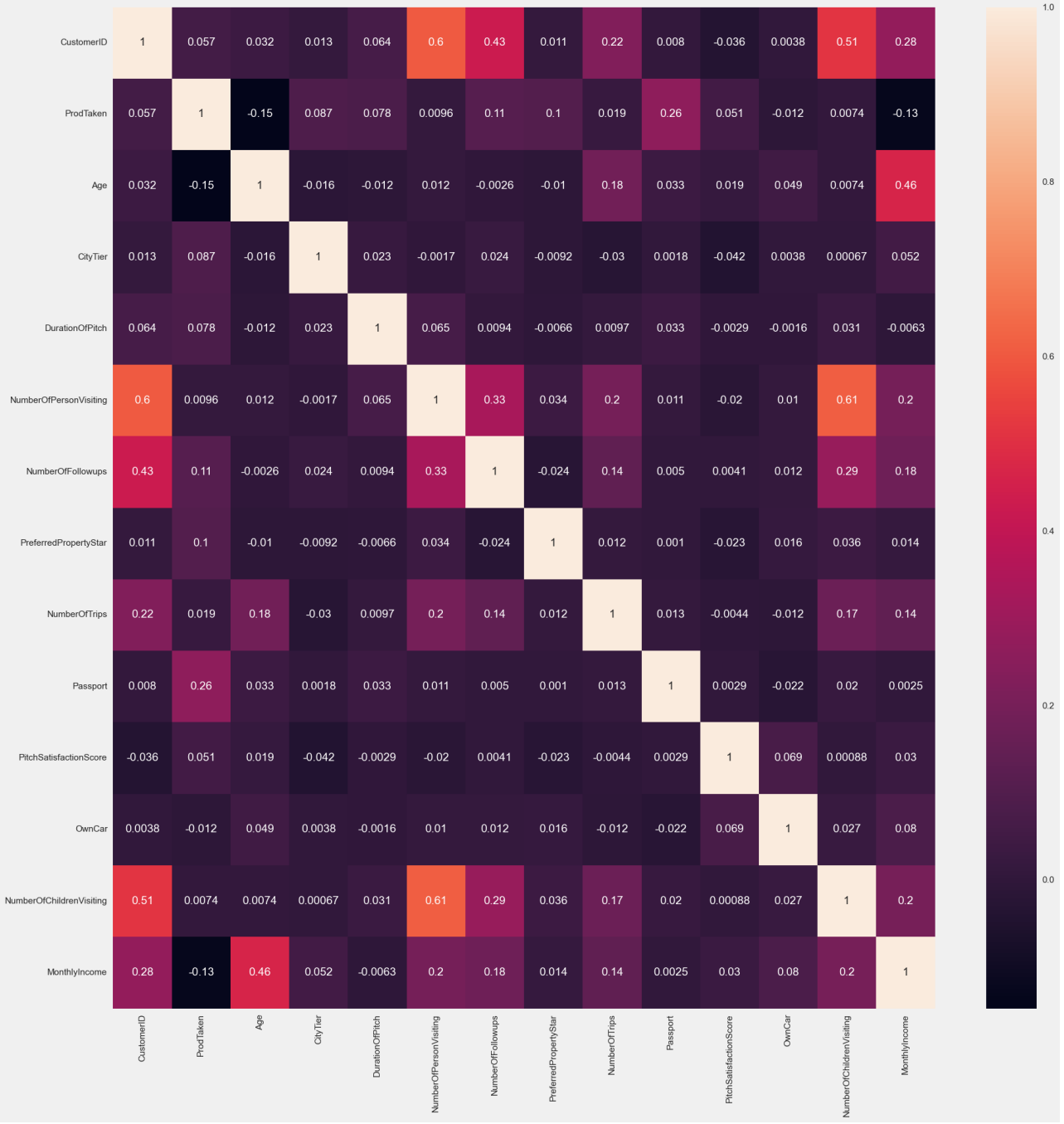

NumberOfTrips
Theoretical quantiles
Age | Distplot
Age | Boxplot
Age | Probability Plot - Skew: 0.4
Ordered Values
Age
Theoretical quantiles
DurationOfPitch | Distplot
DurationOfPitch | Boxplot
DurationOfPitch | Probability Plot - Skew: 1.8
Ordered Values
DurationOfPitch
Theoretical quantiles
MonthlyIncome | Distplot
MonthlyIncome | Boxplot
MonthlyIncome | Probability Plot - Skew: 1.9
Ordered Values
MonthlyIncome
Theoretical quantiles

Histograms (KDE plots) of all Continuous Variables

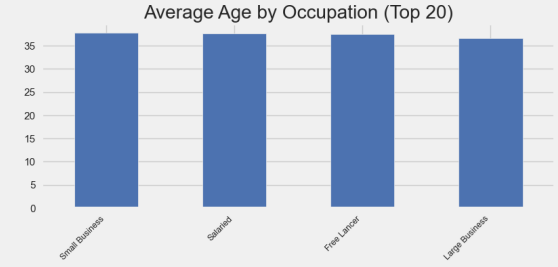


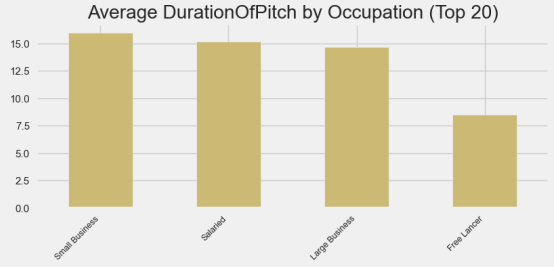


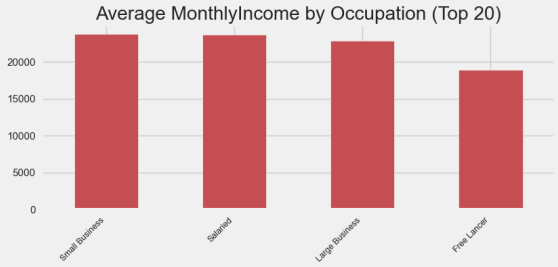
Heatmap of all Continuous Variables including target =

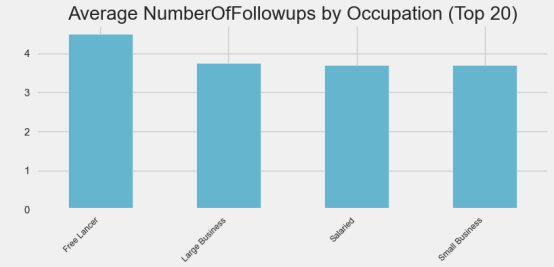


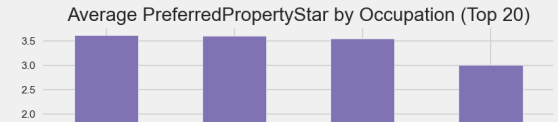
Bar plots for each Continuous by each Categorical variable


Average Age by Occupation (Top 20)
35
30
25
20
15
10
5
0
Small Business
Salaried
Free Lancer
Large Business

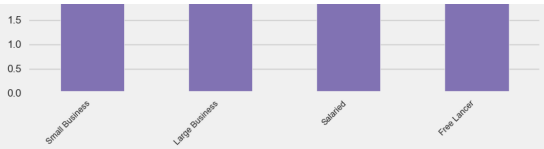

Average DurationOfPitch by Occupation (Top 20)
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Small Business
Salaried
Large Business
Free Lancer

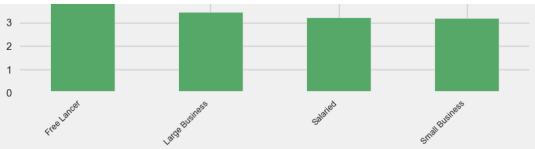

Average MonthlyIncome by Occupation (Top 20)
20000
15000
10000
5000
0
Small Business
Salaried
Large Business
Free Lancer

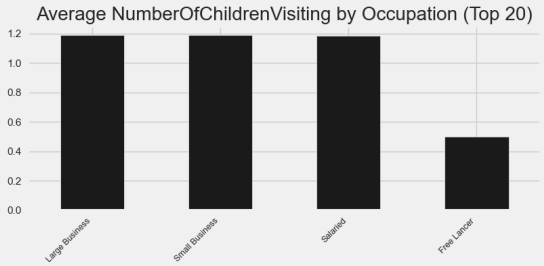

Average NumberOfFollowups by Occupation (Top 20)
4
3
2
1
0
Free Lancer
Large Business
Salaried
Small Business

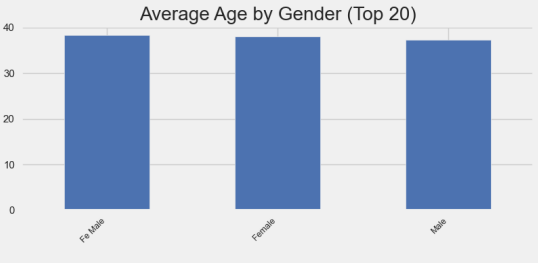

Average PreferredPropertyStar by Occupation (Top 20)
3.5
3.0
2.5
2.0

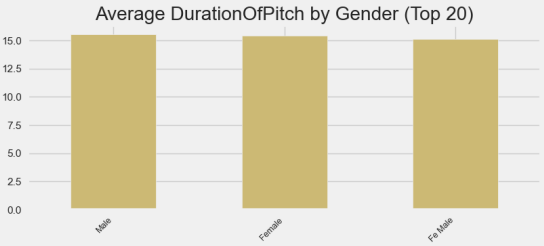

Average NumberOfTrips by Occupation (Top 20)
7
6
5
4

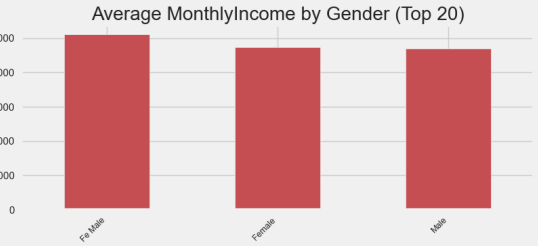

Small Business
Large Business
Salaried
Free Lancer

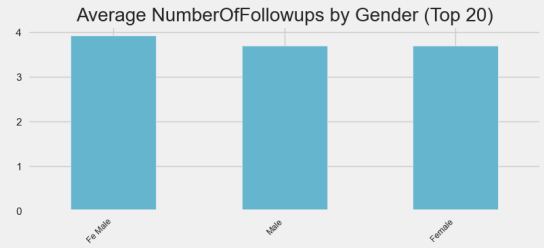

Free Lancer
Large Business
Salaried
Small Business


Average NumberOfChildrenVisiting by Occupation (Top 20)
Large Business
Small Business
Salaried
Free Lancer

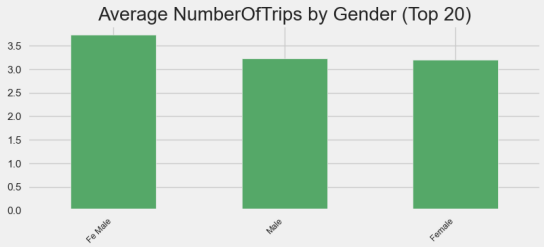

Average Age by Gender (Top 20)
Fe Male
Female
Male

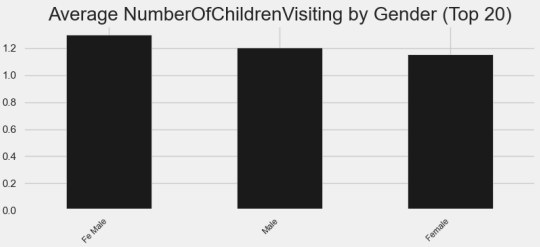

Average DurationOfPitch by Gender (Top 20)
Male
Female
Fe Male

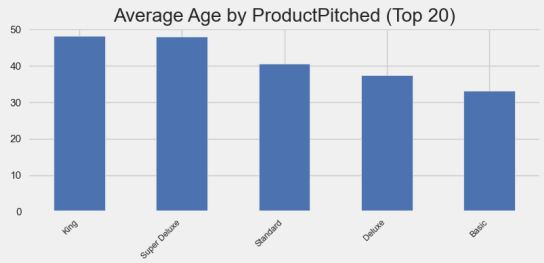

Average MonthlyIncome by Gender (Top 20)
Fe Male
Female
Male

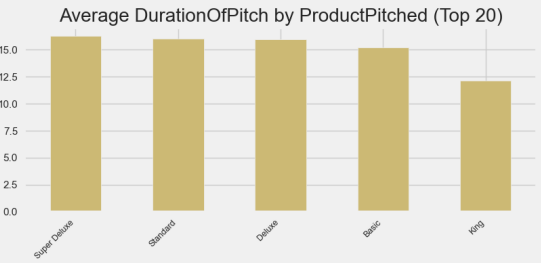

Average NumberOfFollowups by Gender (Top 20)
Fe Male
Male
Female

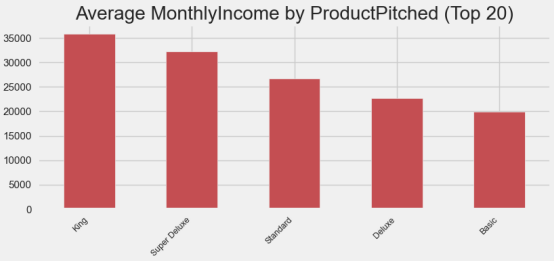

Average PreferredPropertyStar by Gender (Top 20)
Female
Male
Fe Male

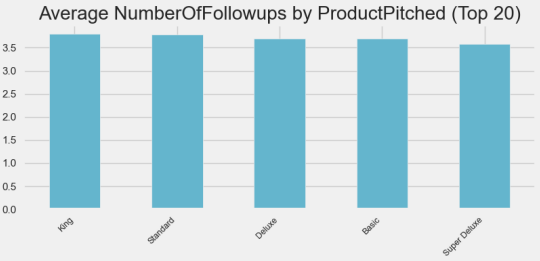

Average NumberOfTrips by Gender (Top 20)
Fe Male
Male
Female

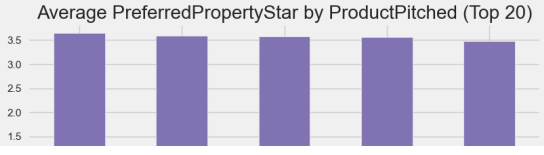

Average NumberOfChildrenVisiting by Gender (Top 20)
Fe Male
Male
Female

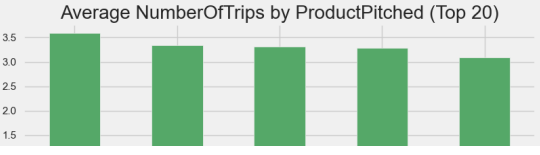

Average Age by ProductPitched (Top 20)
King
Super Deluxe
Standard
Deluxe
Basic

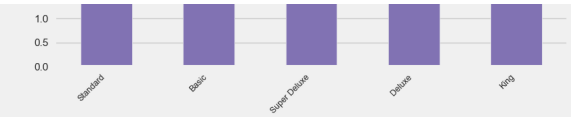

Average DurationOfPitch by ProductPitched (Top 20)
Super Deluxe
Standard
Deluxe
Basic
King

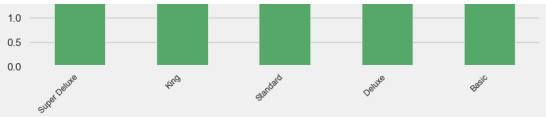

Average MonthlyIncome by ProductPitched (Top 20)
King
Super Deluxe
Standard
Deluxe
Basic


Average NumberOfFollowups by ProductPitched (Top 20)
King
Standard
Deluxe
Basic
Super Deluxe

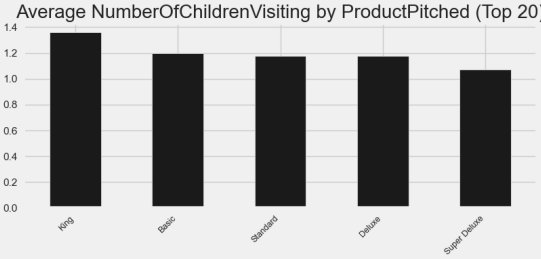

Average PreferredPropertyStar by ProductPitched (Top 20)


Average NumberOfTrips by ProductPitched (Top 20)

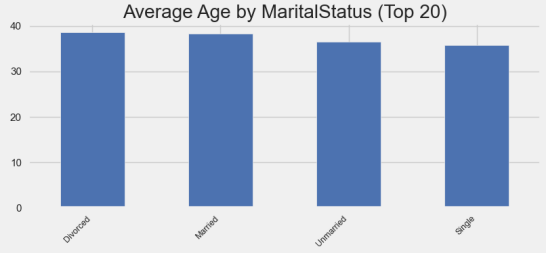




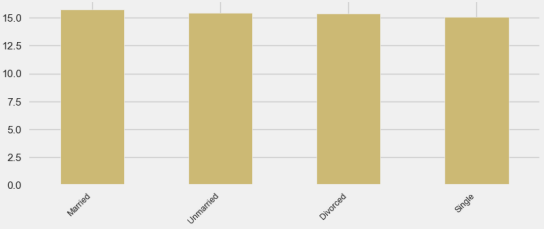
Average NumberOfChildrenVisiting by ProductPitched (Top 20)



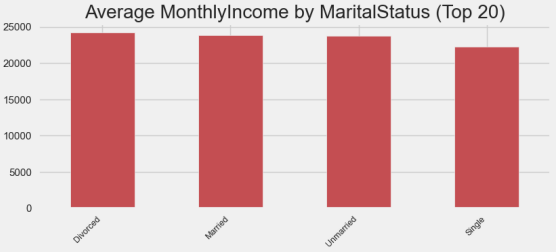
Average Age by MaritalStatus (Top 20)



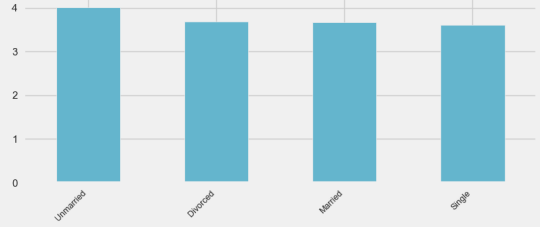
Average DurationOfPitch by MaritalStatus (Top 20)



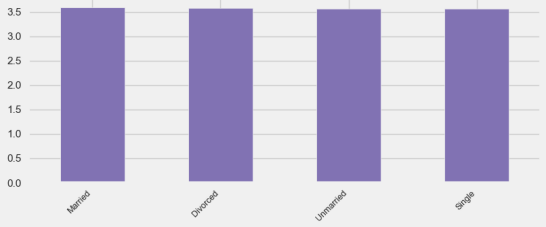
Average MonthlyIncome by MaritalStatus (Top 20)



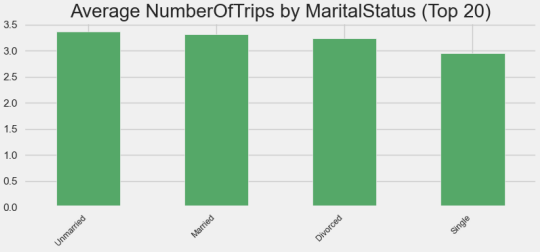
Average NumberOfFollowups by MaritalStatus (Top 20)



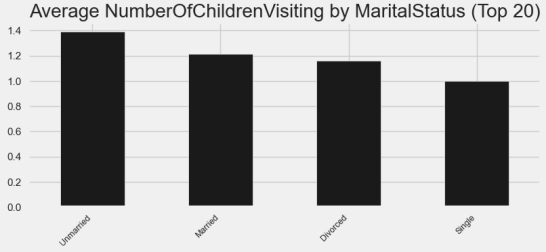
Average PreferredPropertyStar by MaritalStatus (Top 20)



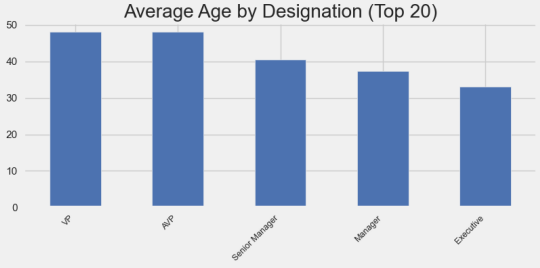
Average NumberOfTrips by MaritalStatus (Top 20)



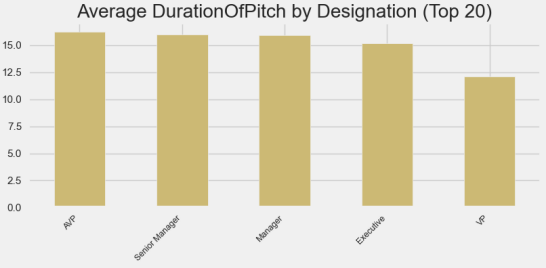
Average NumberOfChildrenVisiting by MaritalStatus (Top 20)



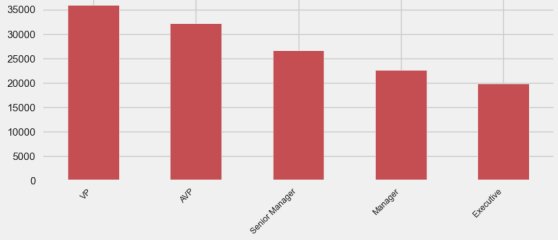
Average Age by Designation (Top 20)



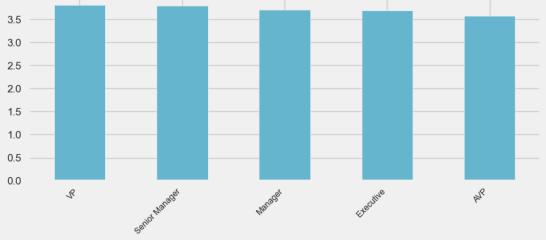
Average DurationOfPitch by Designation (Top 20)



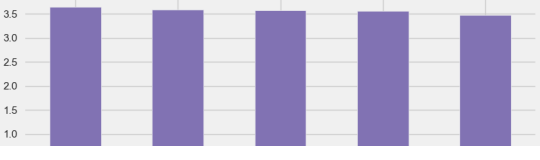
Average MonthlyIncome by Designation (Top 20)



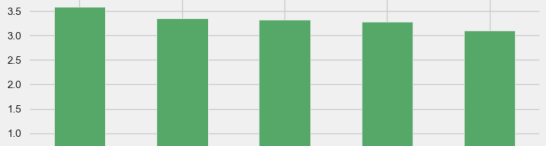
Average NumberOfFollowups by Designation (Top 20)



Average PreferredPropertyStar by Designation (Top 20)

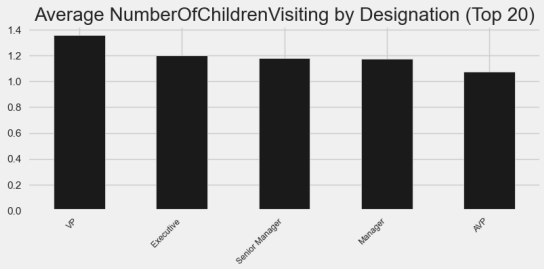


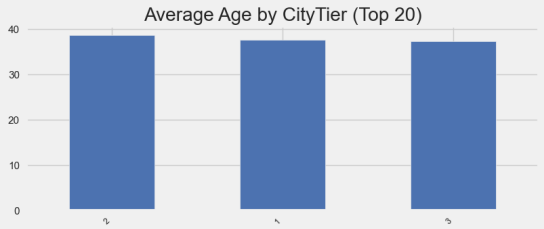
Average NumberOfTrips by Designation (Top 20)

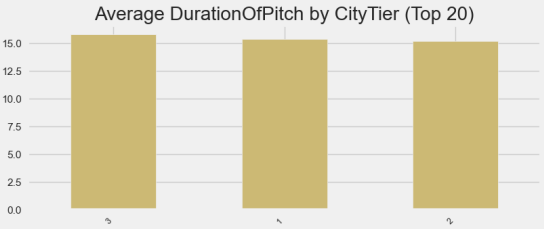


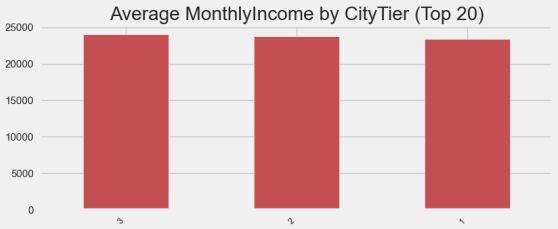

0.5
0.0
Senior Manager
Executive
AVP
Manager
VP

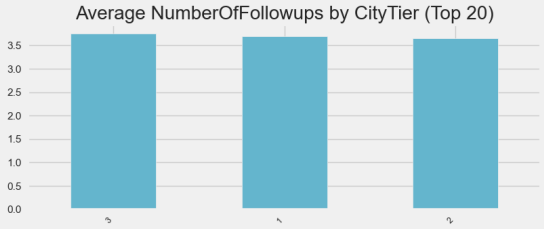

0.5
0.0
AVP
VP
Senior Manager
Manager
Executive

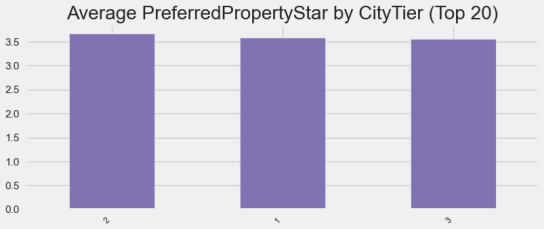

Average NumberOfChildrenVisiting by Designation (Top 20)
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
VP
Executive
Senior Manager
Manager
AVP

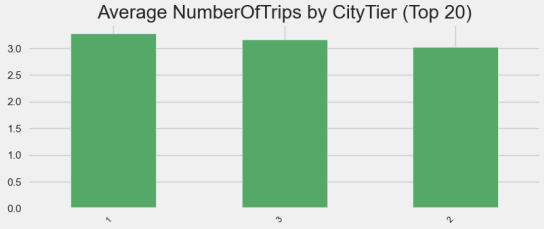

Average Age by CityTier (Top 20)
40
30
20
10
0
2
1
3

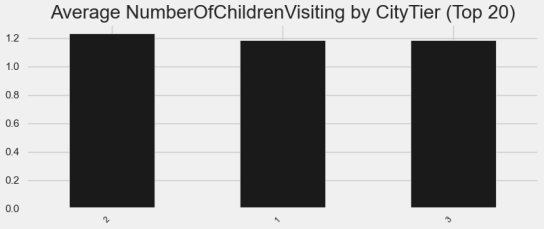

Average DurationOfPitch by CityTier (Top 20)
15.0
12.5
10.0
7.5
5.0
2.5
0.0
3
1
2

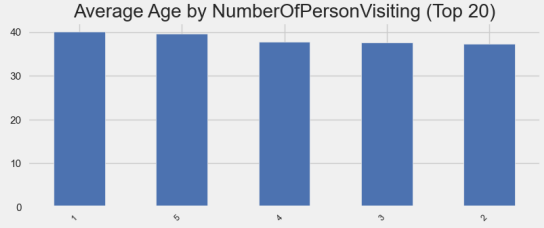

Average MonthlyIncome by CityTier (Top 20)
25000
20000
15000
10000
5000
0
3
2
1

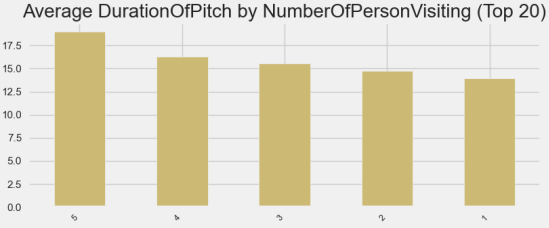

Average NumberOfFollowups by CityTier (Top 20)
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
3
1
2

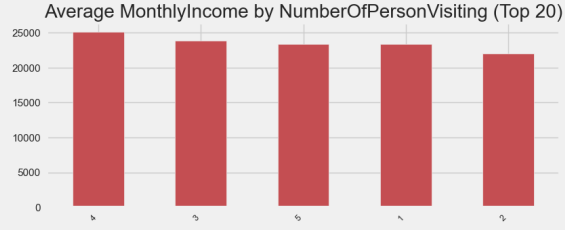

Average PreferredPropertyStar by CityTier (Top 20)
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
2
1
3

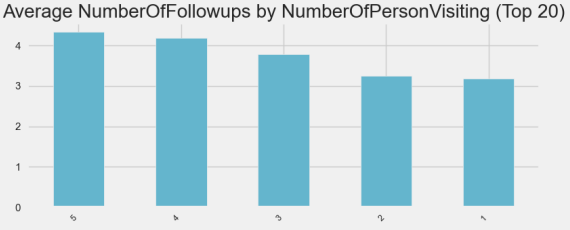

Average NumberOfTrips by CityTier (Top 20)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1
3
2

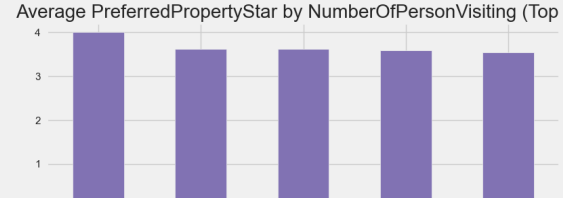

Average NumberOfChildrenVisiting by CityTier (Top 20)
1.2
1.0
0.8
0.6
0.4
0.2
0.0
2
1
3

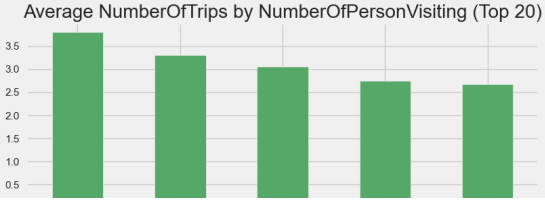

Average Age by NumberOfPersonVisiting (Top 20)
40
30
20
10
0
1
5
4
3
2

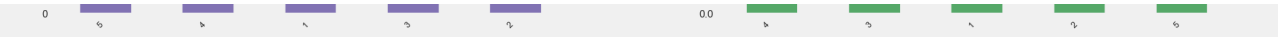

Average DurationOfPitch by NumberOfPersonVisiting (Top 20)
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
5
4
3
2
1

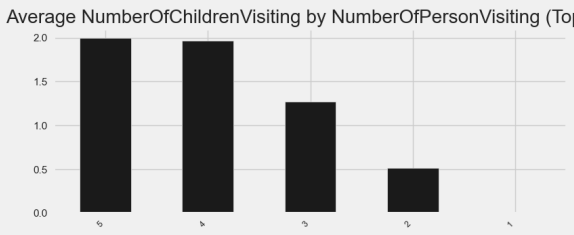

Average MonthlyIncome by NumberOfPersonVisiting (Top 20)
25000
20000
15000
10000
5000
0
4
3
5
1
2

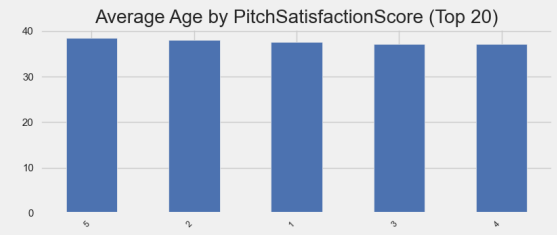

Average NumberOfFollowups by NumberOfPersonVisiting (Top 20)
4
3
2
1
0
5
4
3
2
1

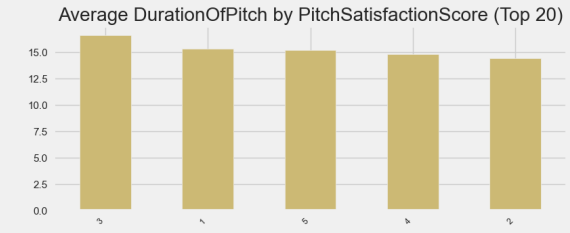

Average PreferredPropertyStar by NumberOfPersonVisiting (Top 20)
4
3
2
1

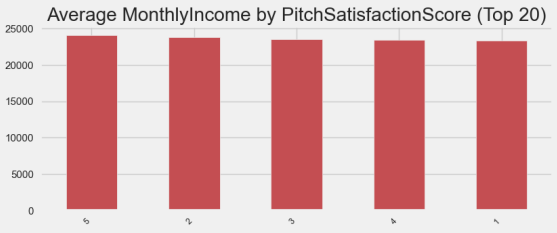

Average NumberOfTrips by NumberOfPersonVisiting (Top 20)
3.5
3.0
2.5
2.0
1.5
1.0
0.5

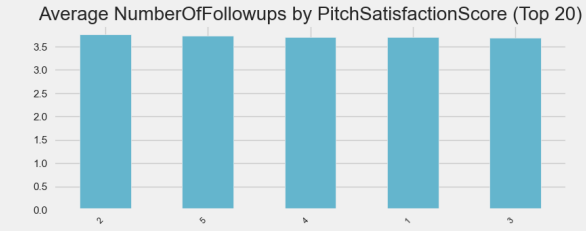


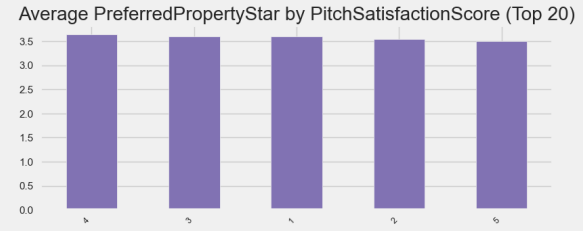

Average NumberOfChildrenVisiting by NumberOfPersonVisiting (Top 20)

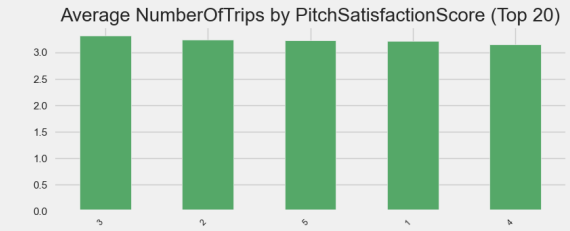

Average Age by PitchSatisfactionScore (Top 20)

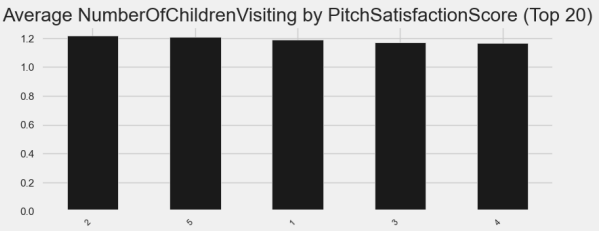

Average DurationOfPitch by PitchSatisfactionScore (Top 20)

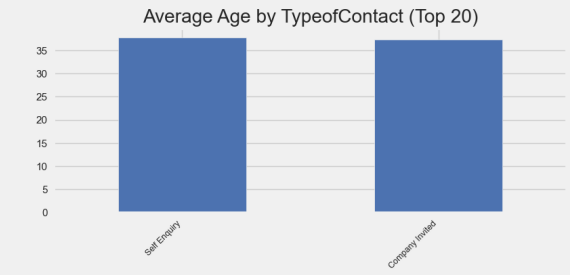

Average MonthlyIncome by PitchSatisfactionScore (Top 20)

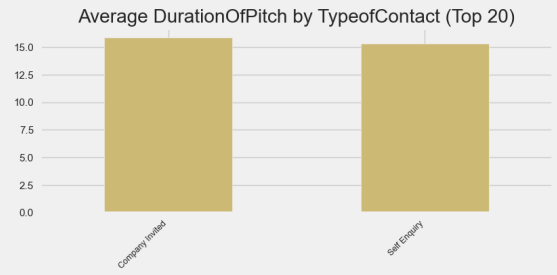

Average NumberOfFollowups by PitchSatisfactionScore (Top 20)

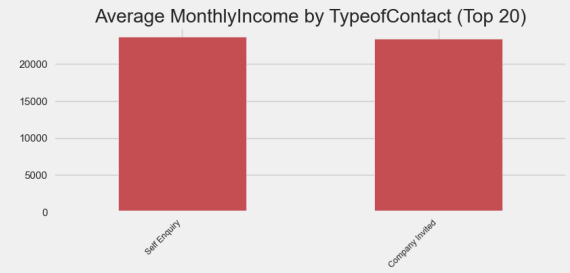

Average PreferredPropertyStar by PitchSatisfactionScore (Top 20)

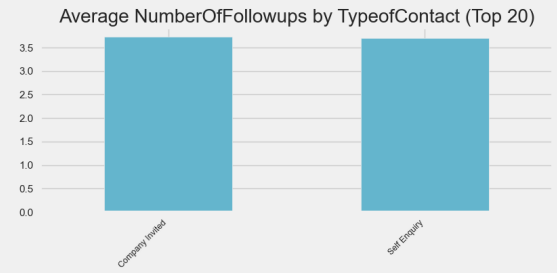

Average NumberOfTrips by PitchSatisfactionScore (Top 20)

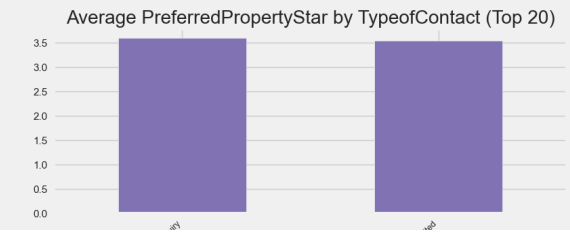

Average NumberOfChildrenVisiting by PitchSatisfactionScore (Top 20)

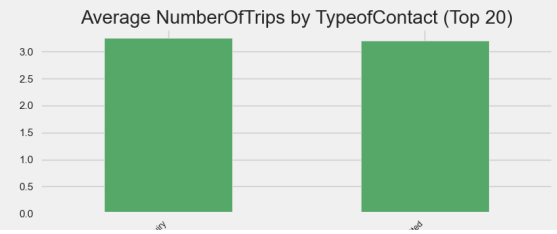

Average Age by TypeofContact (Top 20)
Self Enquiry
Company Invited

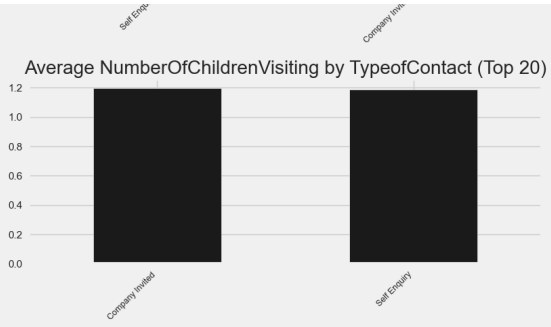

Average DurationOfPitch by TypeofContact (Top 20)
Company Invited
Self Enquiry

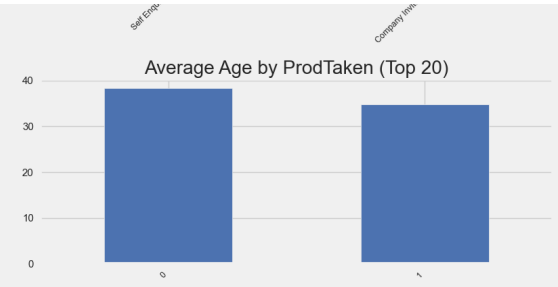

Average MonthlyIncome by TypeofContact (Top 20)
Self Enquiry
Company Invited

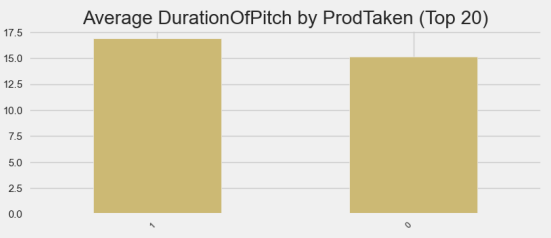

Average NumberOfFollowups by TypeofContact (Top 20)
Company Invited
Self Enquiry

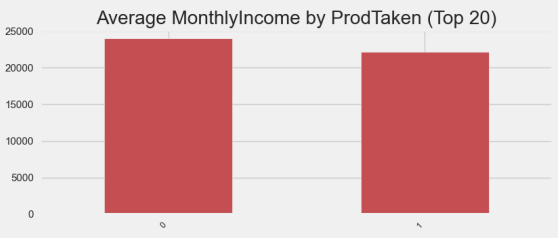

Average PreferredPropertyStar by TypeofContact (Top 20)

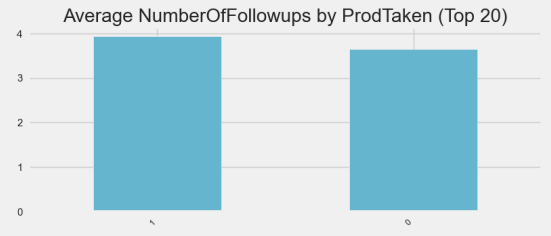

Average NumberOfTrips by TypeofContact (Top 20)


Average NumberOfChildrenVisiting by TypeofContact (Top 20)
Company Invited
Self Enquiry

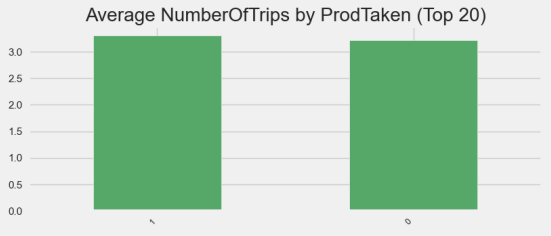

Average Age by ProdTaken (Top 20)

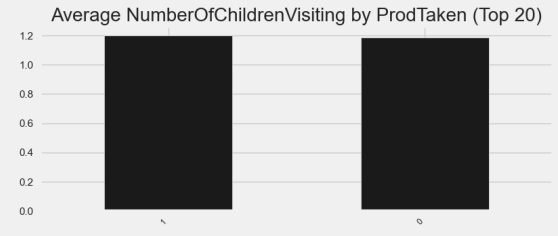

Average DurationOfPitch by ProdTaken (Top 20)

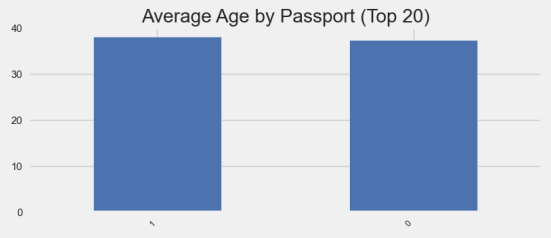

Average MonthlyIncome by ProdTaken (Top 20)

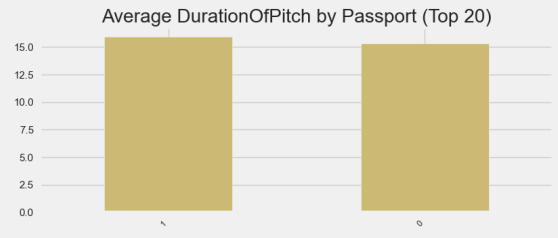

Average NumberOfFollowups by ProdTaken (Top 20)

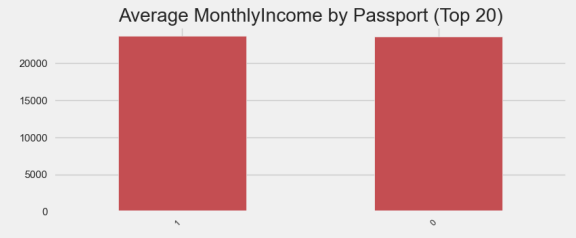

Average PreferredPropertyStar by ProdTaken (Top 20)

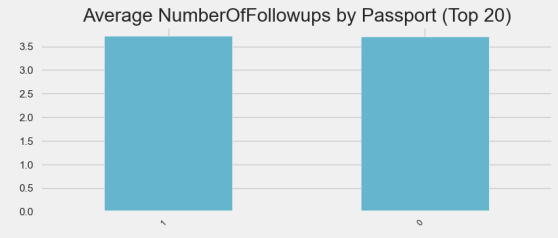

Average NumberOfTrips by ProdTaken (Top 20)


Average NumberOfChildrenVisiting by ProdTaken (Top 20)


Average Age by Passport (Top 20)

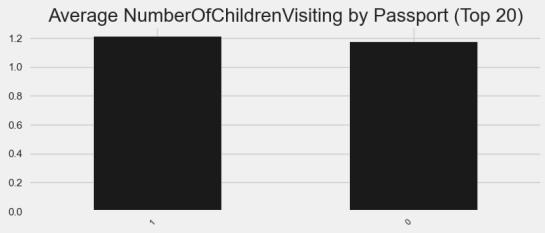

Average DurationOfPitch by Passport (Top 20)

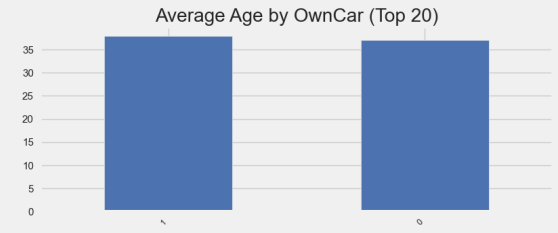

Average MonthlyIncome by Passport (Top 20)

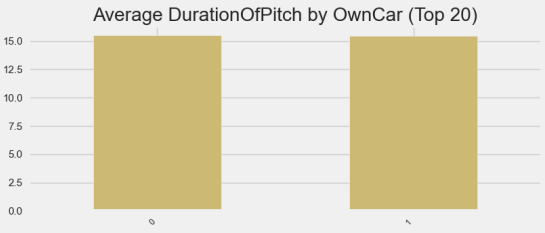

Average NumberOfFollowups by Passport (Top 20)

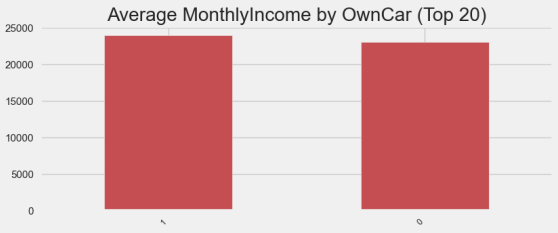

Average PreferredPropertyStar by Passport (Top 20)

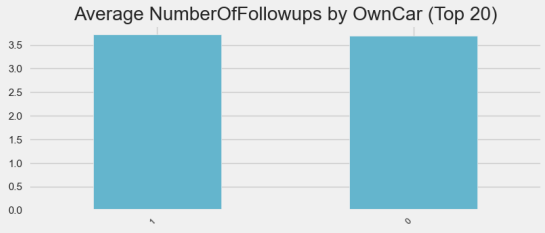

Average NumberOfTrips by Passport (Top 20)

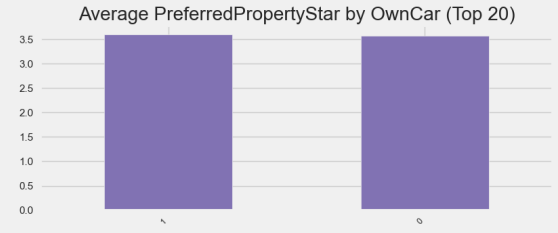


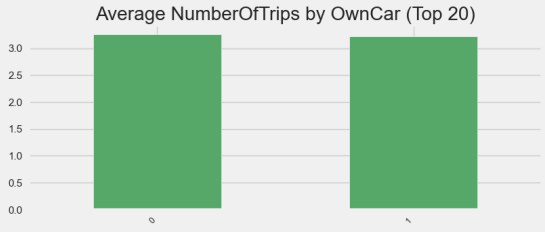


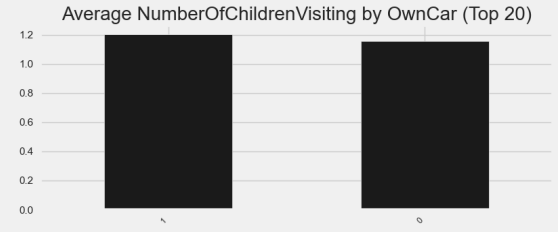












Thank a Lot for being with us!

In [ ]: