Market Entry Analysis Project

-Submitted By
Dr. V . Swaroopa Rani

PROBLEM STATEMENT :

- This Project is based on a mobile company ‘XYZ Mobiles’ from China.
- XYZ Mobiles wants to enter Indian Market as it believes that the Indian market is very similar to China, in which the company is successfully running.
- Before entering the new market, the company wants to be sure that the whole process will be profitable for them.

  My task is to check for the following conditions that must be fulfilled in the Indian market for the company to enter:
- **Sale of a minimum of 12,000 phones** over the sample data in one year .
- **Collection of at least Rs. 20 crores** over the sample data in one year .

Segmentation and Model Development

- Gender was classified into binary data as male (1) and female(0). The annual income was converted into INR for matching the situation of Indian currency.
- From the Chinese customer data it is clear that purchase decision depends on 4 factors. Customer Age, Gender, Phone Age, and Annual Income.
- The Phone Age was classified into 4 segments.

Days	Segment
<200	1
200-360	2
360-500	3
>500	4

- The data was divided into training and test set with 70:30 Rule and then Logistic Regression was applied on training and testing datasets.
- And then ROC Curve, Beta Values(Coefficients) and Confusion Matrix (Including Accuracy, Sensitivity, Precision etc.)were computed from train and test data through K-means Clustering.

Task1: PIVOT ANALYSIS AFTER CLEANING OF DATASET:

Table: 1

Row Labels	Sum of PURCHASE	Count of Lead	Conversion rate
0	9836	17715	55.52
1	13195	22285	59.21
Grand Total	23031	40000	57.58

Table: 2

Row Labels	Sum of PURCHASE2	Count of PURCHASE	Conversion rate	
1	2351	6459	36.40	
2	7023	16545	42.45	
3	9208	11697	78.72	
4	4449	5299	83.96	
Grand Total	23031	40000	57.58	

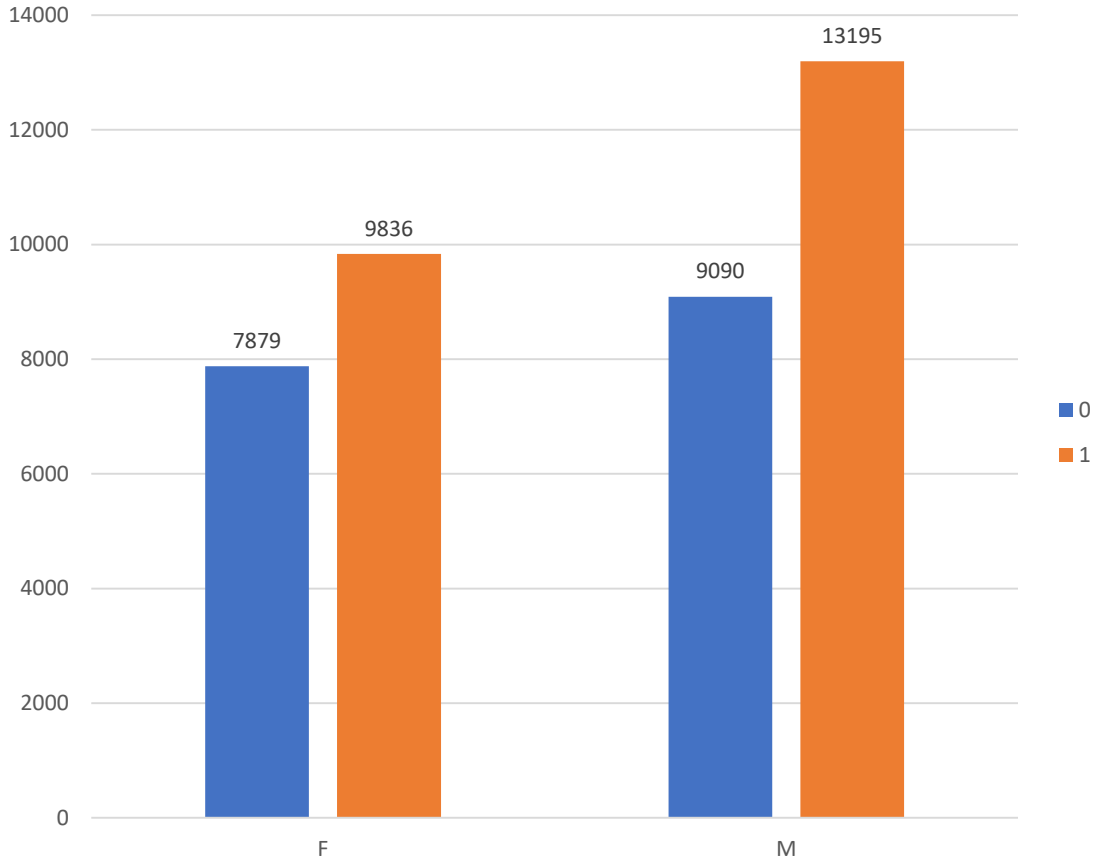
Note :

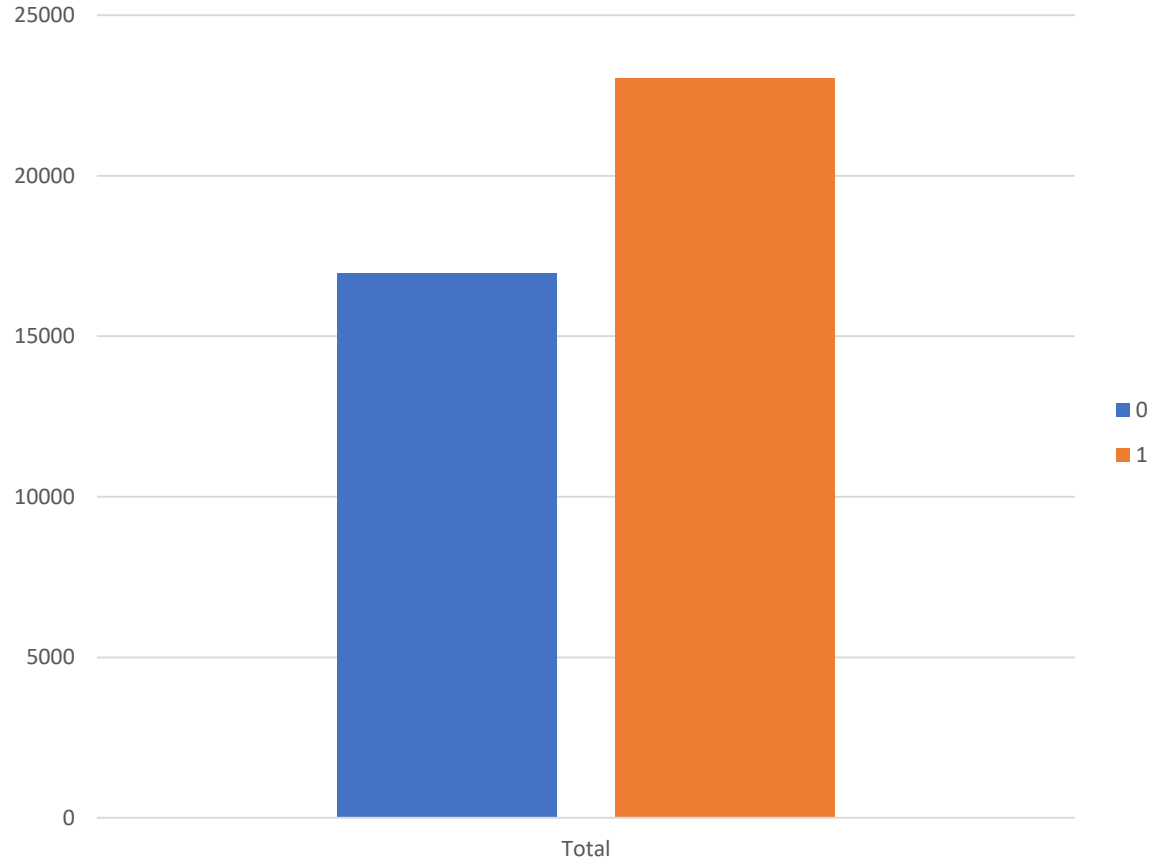
1. From Table 1 it is clear that conversion rate for Males is high As compared to Females that is, **59.21 %**.
2. From Table 2 it is clear that categories 3 and 4 (i.e., Phone Age > 365 and >500) are having highest Conversion Rates, **78.72% 83.96 %.**

Statistical Analysis

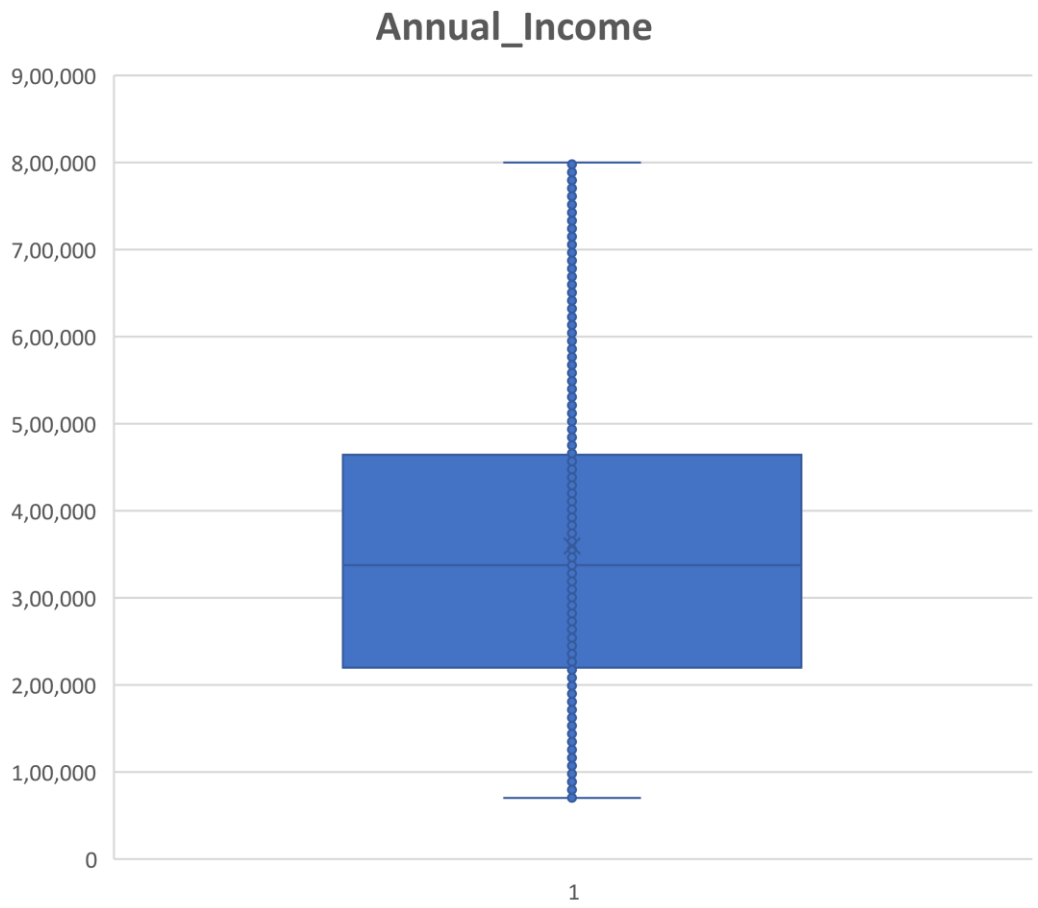
CURR_AGE		ANN_INCOME		AGE_PHN	
Mean	44.99745	Mean	359398.9	Mean	359.0803
Standard Error	0.0591	Standard Error	875.5463	Standard Error	1.015319
Median	45	Median	337656.8	Median	331
Mode	46	Mode	108632	Mode	326
Standard Deviation	11.82008	Standard Deviation	175109.3	Standard Deviation	203.0637
Sample Variance	139.7143	Sample Variance	3.07E+10	Sample Variance	41234.88
Kurtosis	-1.20187	Kurtosis	-0.3465	Kurtosis	1.437027
Skewness	0.00607	Skewness	0.590983	Skewness	1.060834
Range	40	Range	729881.7	Range	1019
Minimum	25	Minimum	70089	Minimum	1
Maximum	65	Maximum	799970.7	Maximum	1020
Sum	1799898	Sum	1.44E+10	Sum	14363210
Count	40000	Count	40000	Count	40000

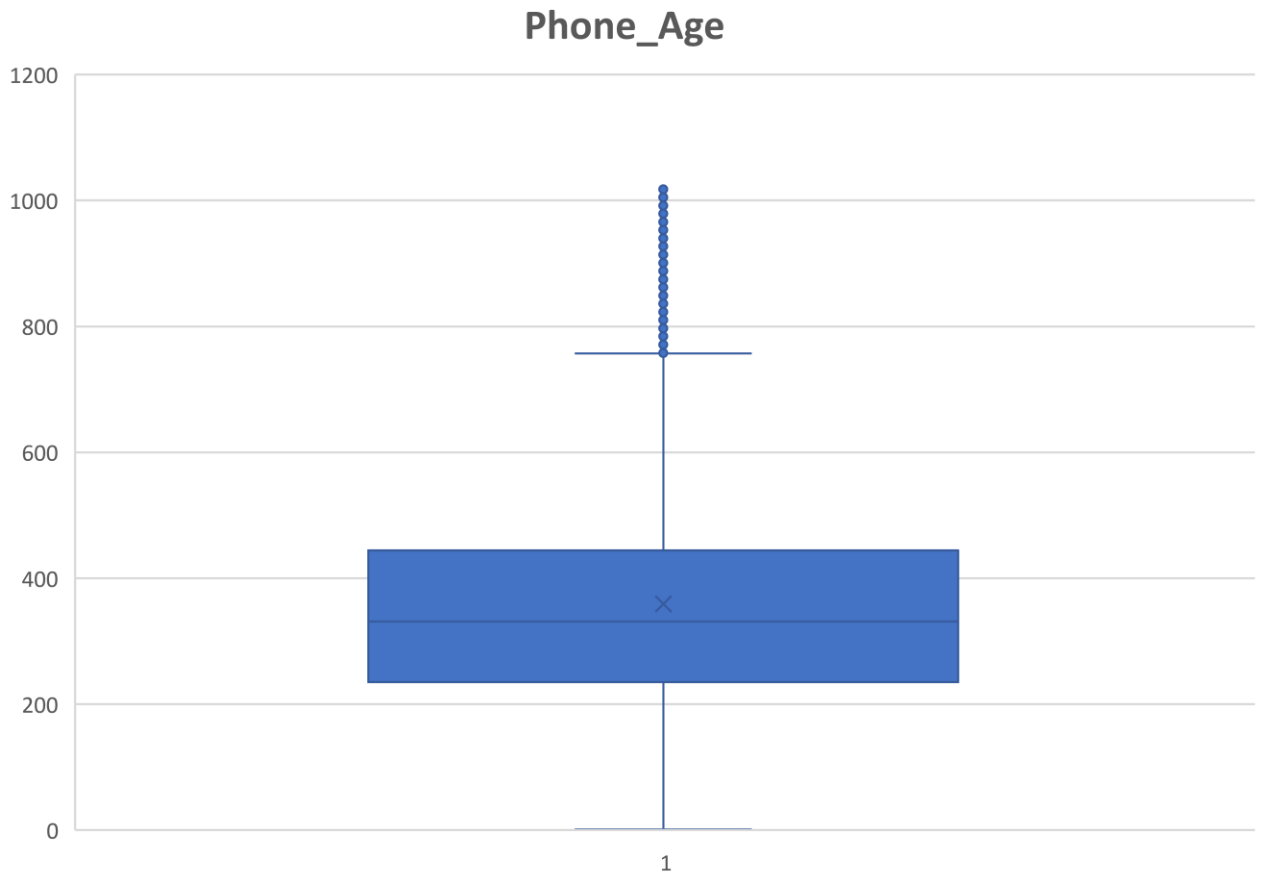
Data Visualisation and EDA





Box-plots Annual_income and Phone_Age



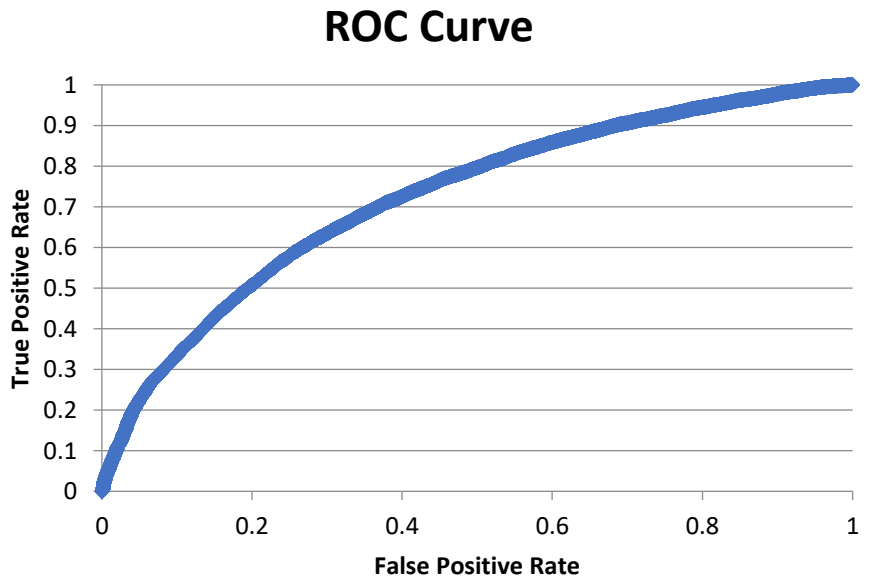


Task1: CLASSIFICATION MODEL BASED ON CHINESE DATASET:

- The logistic regression was applied on the train data by considering the factors like age, gender, income, phone life and purchase.
- The coefficients B0 to B4 and ROC Curve of training Data are computed as follows:

Coefficients	
B0	-1.53387
B1	-0.01181
B2	0.211286
B3	2.29E-06
B4	0.004219

	Confusion_matrix	
	predicted	
actual	no(0)	yes(1)
no(0)	7201	6412
yes(1)	4080	14307
	Performance Metrics	
Accuracy	0.672125	(TP + TN) / TP+TN+FP+FN
Precision	0.690526	TP / (TP + FP)
Recall	0.778104	TP / (TP + FN)
Specifivity	0.52898	TN / (TN + FP)
Sensitivity	0.778104	TP / (TP + FN)
F1_Score	0.731704	2 * (Precision * Recall) / (Precision + Recall)
TPR	0.778104	True Positives / (True Positives + False Negatives)
FPR	0.611132	False Positives / (False Positives + True Negatives)

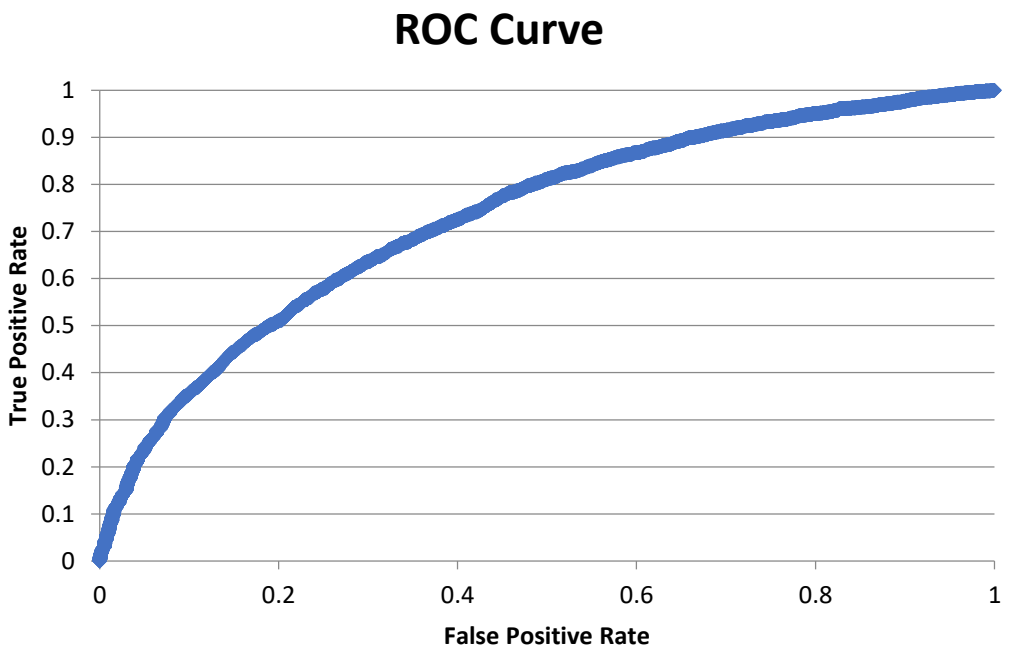


Task1: CLASSIFICATION MODEL ON CHINESE TEST DATASET AND PERFORMANCE METRICS ASSOCIATED:

- The coefficients B0 to B4 and ROC Curve of test Data are computed as follows:

Coefficients	
B0	-1.66007
B1	-0.01285
B2	0.293556
B3	2.81E-06
B4	0.00412

	Confusion_matrix	
	predicted	
actual	no(0)	yes(1)
no(0)	1799	1557
yes(1)	999	3645
	Performance Metrics	
Accuracy	0.6805	(TP + TN) / TP+TN+FP+FN
Precision	0.700692	TP / (TP + FP)
Recall	0.784884	TP / (TP + FN)
Specifivity	0.536055	TN / (TN + FP)
Sensitivity	0.784884	TP / (TP + FN)
F1_Score	0.740402	2 * (Precision * Recall) / (Precision + Recall)
TPR	0.784884	True Positives / (True Positives + False Negatives)



- From the coefficient it is clear that age has a negative impact. Old people are less likely to buy a phone.
- Gender plays a significant role, which is evident in the pivot analysis.
- The annual income plays an insignificant role as the coefficient is very small.
- The phone life also plays an important role and the categorial variation has an impact of its own.

5 Number Summary (Indian Dataset)

CURR_AGE		*GENDER*		*ANN_INCOME*		*AGE_PHN*	
Mean	45.08229	Mean	0.500934	Mean	1146029	Mean	576.1679
Standard Error	0.066686	Standard Error	0.002814	Standard Error	2249.446	Standard Error	1.077481
Median	45	Median	1	Median	1123316	Median	486
Mode	56	Mode	1	Mode	1171669	Mode	388
Standard Deviation	11.84938	Standard Deviation	0.500007	Standard Deviation	399699.3	Standard Deviation	191.4554
Sample Variance	140.4077	Sample Variance	0.250007	Sample Variance	1.6E+11	Sample Variance	36655.19
Kurtosis	-1.21016	Kurtosis	-2.00011	Kurtosis	-0.68412	Kurtosis	-0.60677
Skewness	-0.00116	Skewness	-0.00374	Skewness	0.136363	Skewness	0.87612
Range	40	Range	1	Range	1699791	Range	652
Minimum	25	Minimum	0	Minimum	300054	Minimum	368
Maximum	65	Maximum	1	Maximum	1999845	Maximum	1020
Sum	1423383	Sum	15816	Sum	3.62E+10	Sum	18191348
Count	31573	Count	31573	Count	31573	Count	31573
Confidence Level(95.0%)	0.130708	Confidence Level(95.0%)	0.005515	Confidence Level(95.0%)	4409.001	Confidence Level(95.0%)	2.111906

Task1: Data Modelling Considering Indian Dataset

- The data set is formatted such that gender is converted into a binomial model and the phone age is calculated by considering the purchase date as 1st July 2019.
- The phone life, age, income are segmented accordingly for further studies

DAYS	SEGMENT
<200	1
200-360	2
360-500	3
>500	4

Male	1
Female	0

Purchase	
1	Yes
0	No

	Segments	Age Criteria
Young Age	1	25-35
Mid Age	2	35-55
Old Age	3	55-65

1. Low Income:

   Minimum income: ₹3,00,054
   Maximum income: ₹7,67,967
2. Medium Income:

   Minimum income: ₹7,67,968
   Maximum income: ₹12,34,880
3. High Income:

   Minimum income: ₹12,34,881
   Maximum income: ₹19,99,845

- The probability is computed based on coefficients (B0 to B4) obtained from the Chinese dataset and the no of potential customers in India based on a cut-off 0.5 is 31573 with a conversion ratio of 45.10%.
- Applied k-means clustering on the potential customers in India dataset.
- 3 scenarios were considered in the study. One with 3 clusters, 4-clusters and standardized dataset.

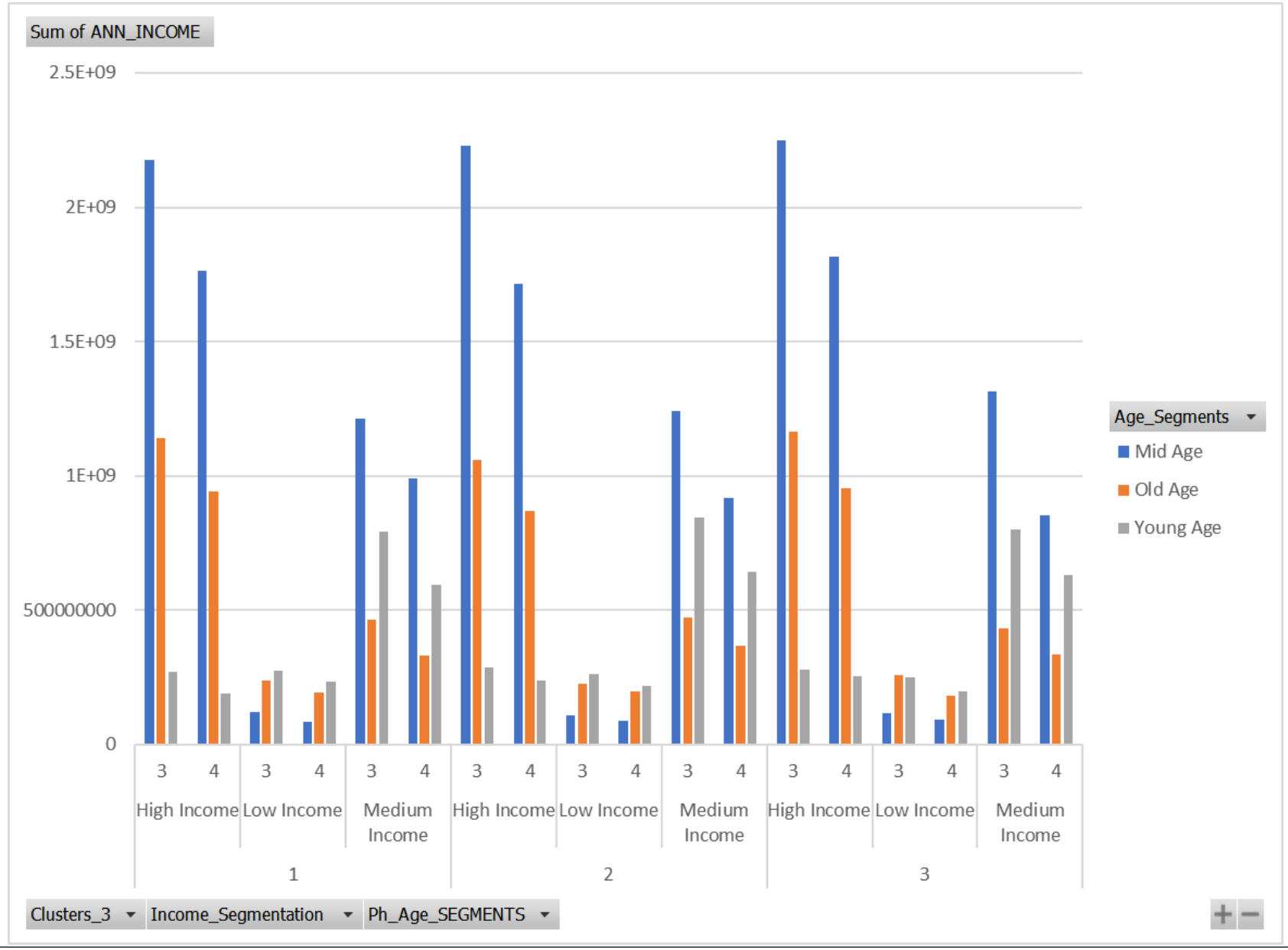
T2: JUSTIFICATION DURING CLUSTERING :

- Clustering is performed on 3 & 4 Clusters and their error terms (For more scaled and standardized data) is also found and with that centroid values were generated.
- In our analysis 3 clusters results were taken for further analysis and predicting results and then EDA is performed on each cluster & Centroid values and following results were obtained:

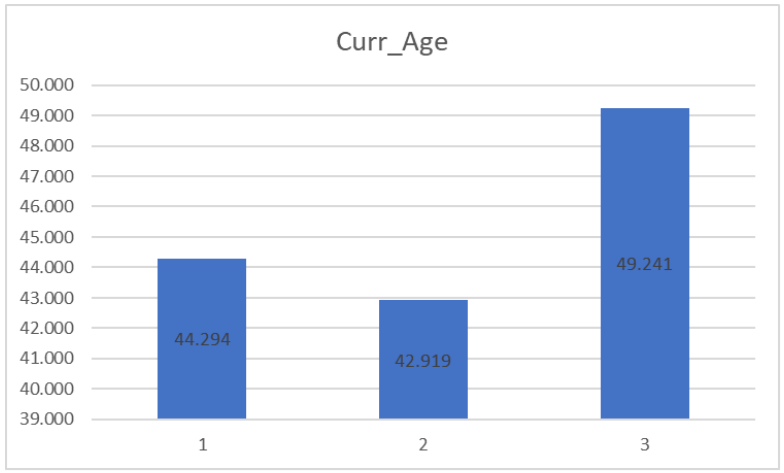
Centroid values (For Potential Customers)			
	1	2	3
CURR_AGE	44.294	42.918	49.241
GENDER	0.436	0.477	0.609
ANN_INCOME	696419.45	1151544.67	1655736.147
AGE_PHN	574.39	575.75	578.84
Centroid values (Error Terms)			
	1	2	3
CURR_AGE	0.460	0.525	-1.212
GENDER	0.998	-1.002	-0.072
ANN_INCOME	0.343	0.086	-0.539
AGE_PHN	-0.002	0.005	-0.004

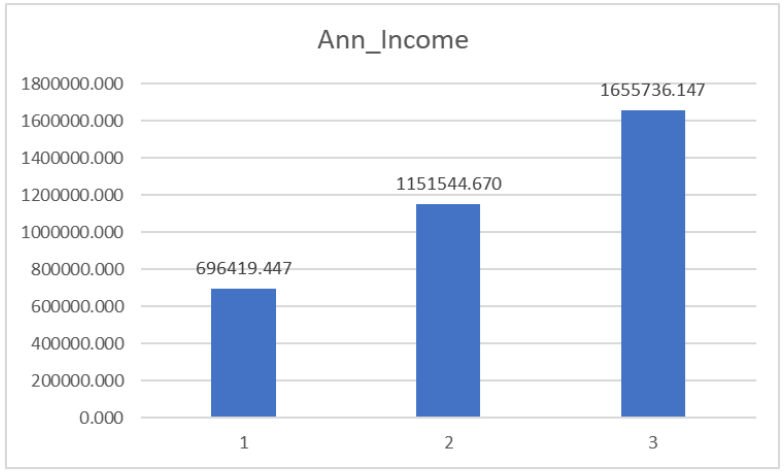
From the table above, **centroid value 3** is to be taken for business decision -0.539 and and It has error terms in negative as well which is good **so most peoples were clustered (i.e., their centroid) around with current age of 49, Annual income around 1655736 and age of phone around 579 days is to target For** through our clustering analysis and EDA analysis of that is also done in further slides.

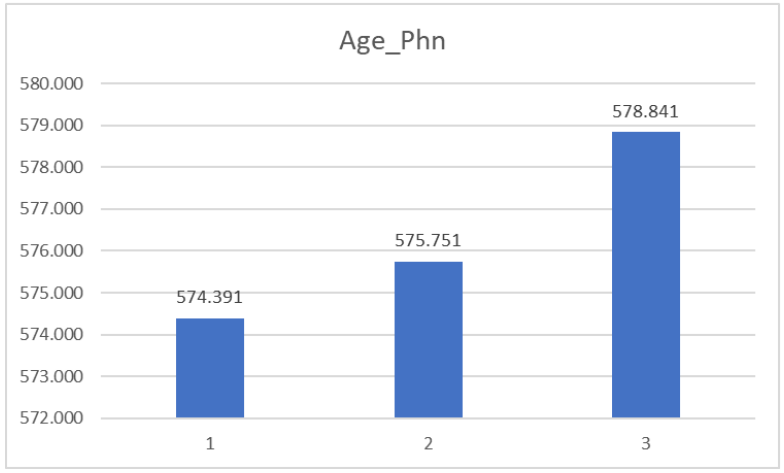
Potential Customers for all 3 clusters



T2 : EDA ANALYSIS :



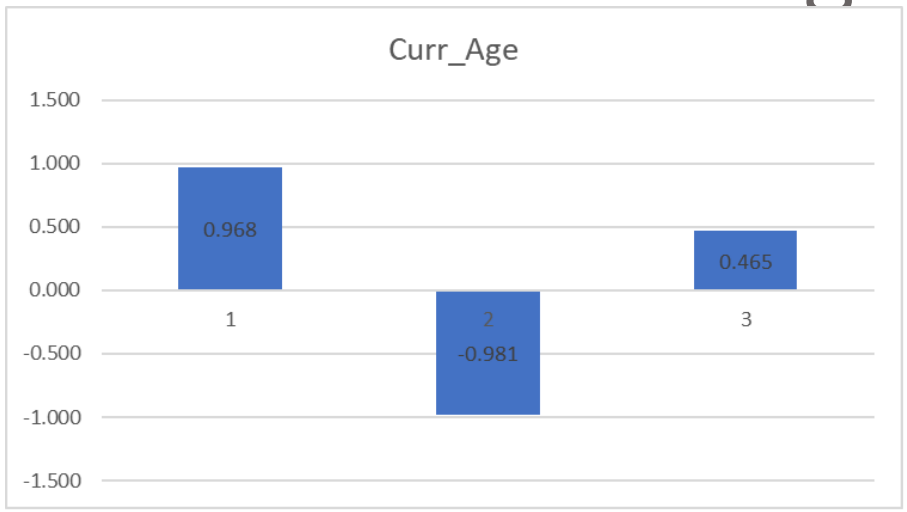


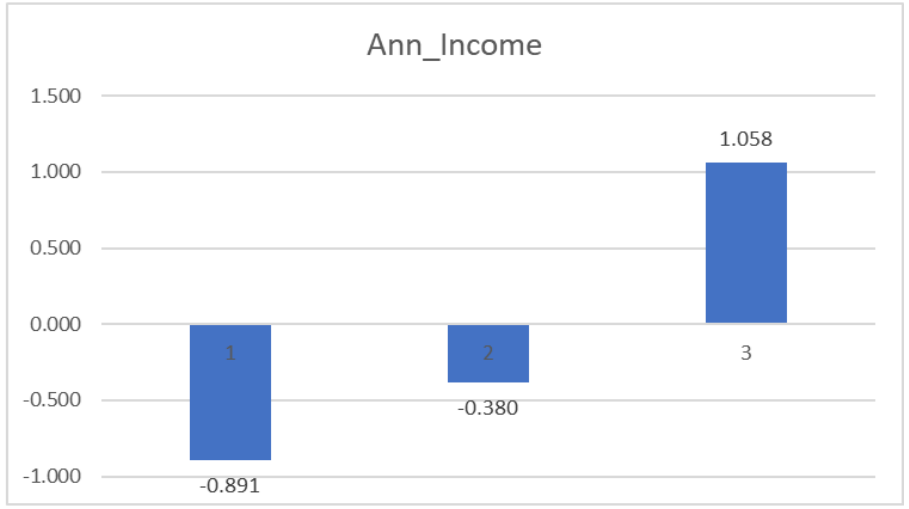


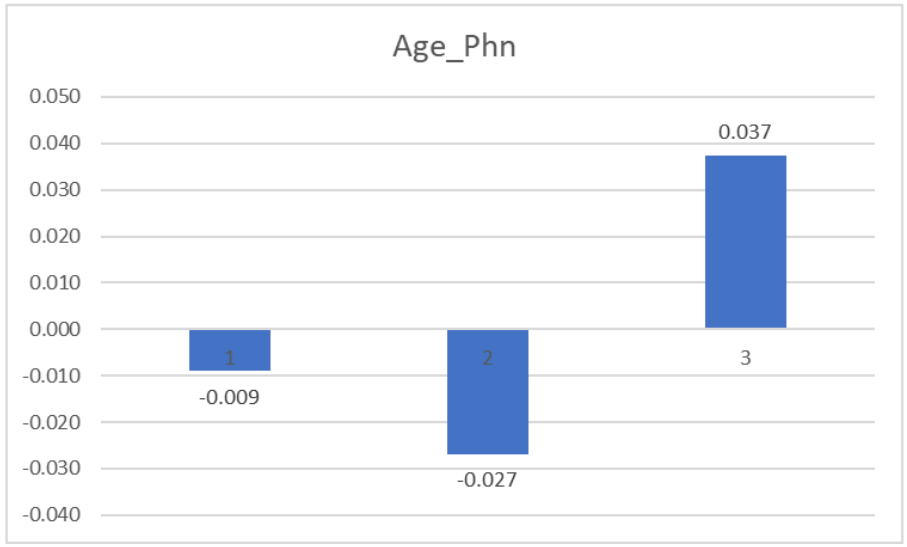
	Centroid Values		
	1	2	3
Curr_Age	44.294	42.919	49.241
Gender	0.437	0.477	0.610
Ann_Income	696419.447	1151544.670	1655736.147
Age_Phn	574.391	575.751	578.841

From the table above,
centroid value 3 is to be taken for business decision, with current age of 49,
Annual income around 1655736 and age of phone around 579 days is to target.

K-means Clustering on Standardized dataset(Clusters-3)







Centroid Values			
	1	2	3
Curr_Age	0.968	-0.981	0.465
Ann_Income	-0.891	-0.380	1.058
Age_Phn	-0.009	-0.027	0.037

T3: JUSTIFICATION FOR FINAL RESULTS:

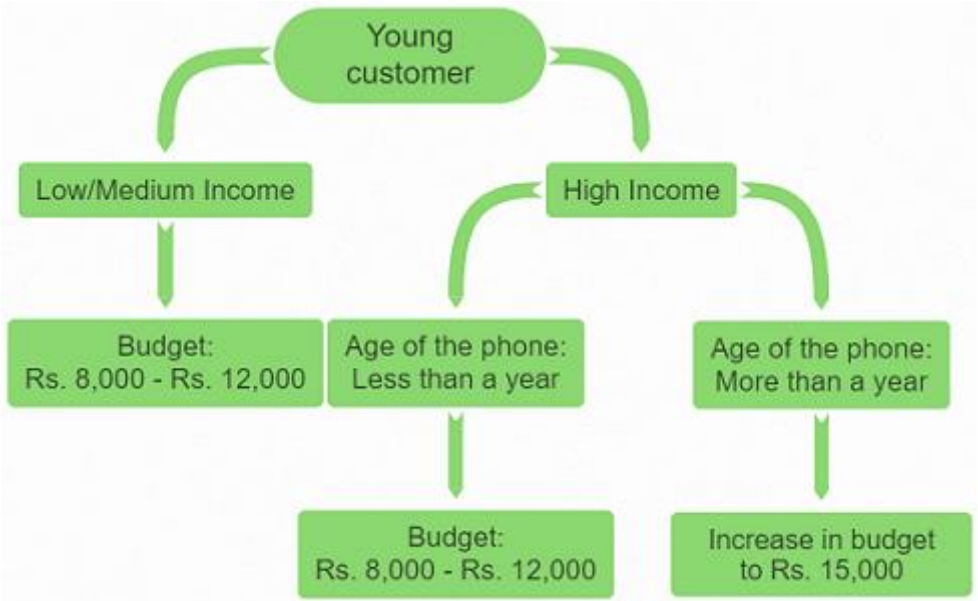
Now we need to segment our customers through their ages to check for our final results i.e.,

	Segments	Age Criteria
Young Age	1	25-35
Mid Age	2	35-55
Old Age	3	55-65

- **Sale of a minimum of 12,000 phones** over the sample data in one year .
- **Collection of at least Rs. 20 crores** over the sample data in one year .

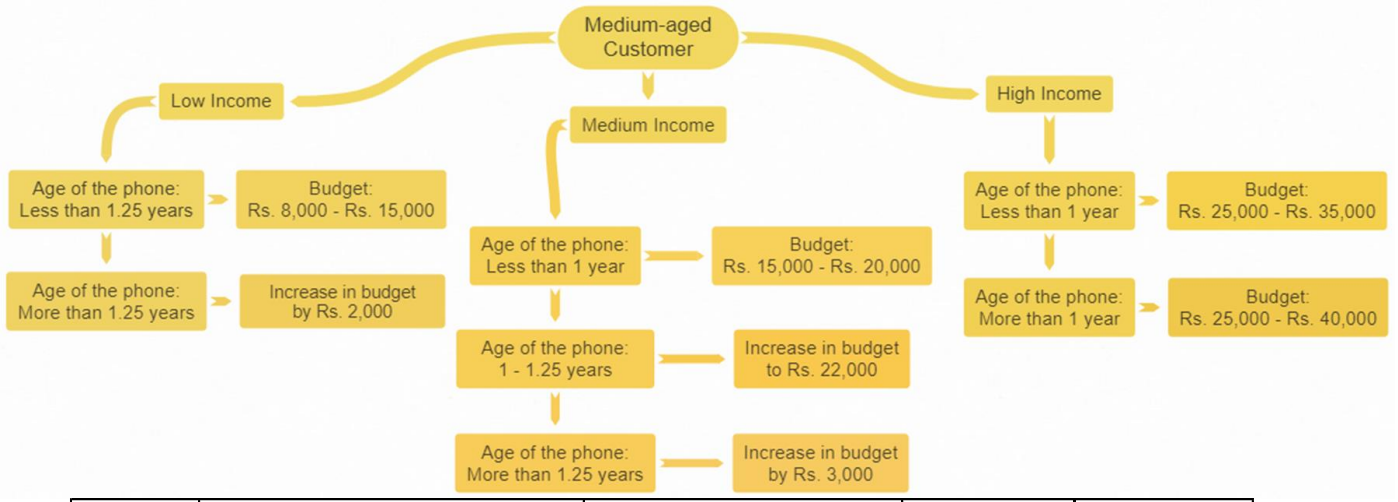
Task 2 : EDA ANALYSIS :

Segmentation(Young Age Customers) and EDA Analysis(Clusters-3)



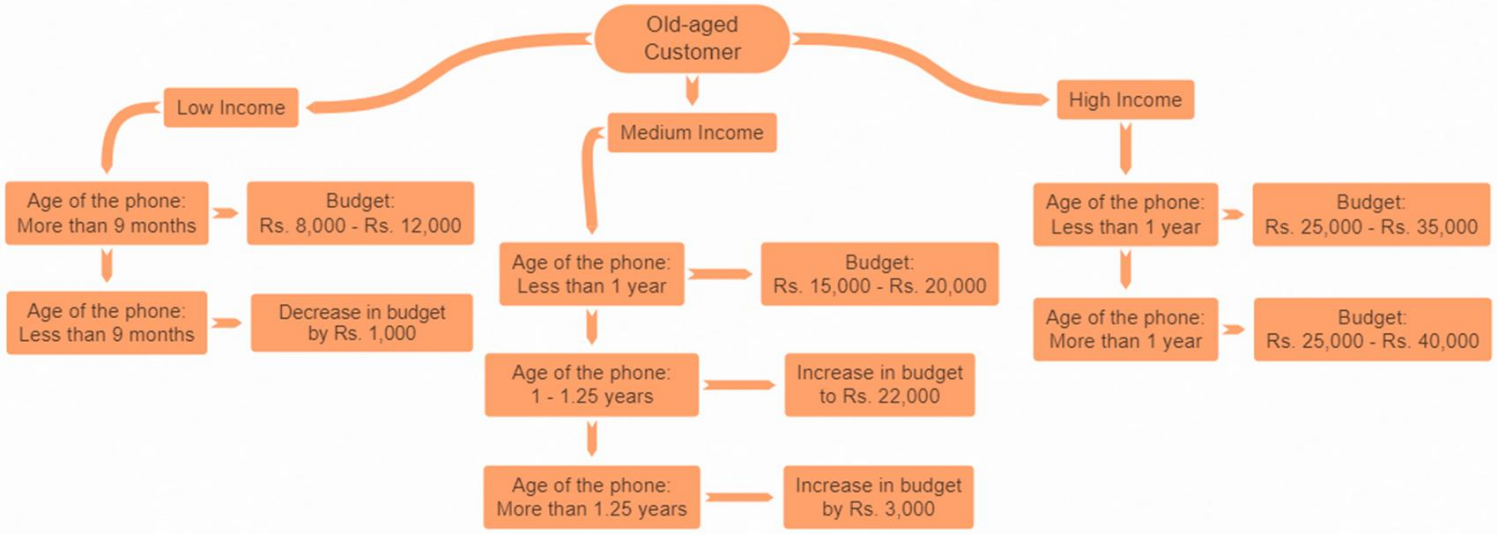
	Young Aged Customers				
	Budget	**Phn_age**	**count**	**Expected Revenue**	**IF >20 CRORE**
Low_Incor	8000	<1year	0	0	No
Medium_I	12000	>1 year	4339	52068000	No
High_Inco	12000	<1year	0	0	No
High_Inco	15000	>1year	1111	16665000	No
			Average	34366500	
			Segments	**Age Criteria**	
	Young Age		1	25-34	
	Mid Age		2	35-54	
	Old Age		3	55-65	

Segmentation(Medium Age Customers) and EDA Analysis (Clusters-3)



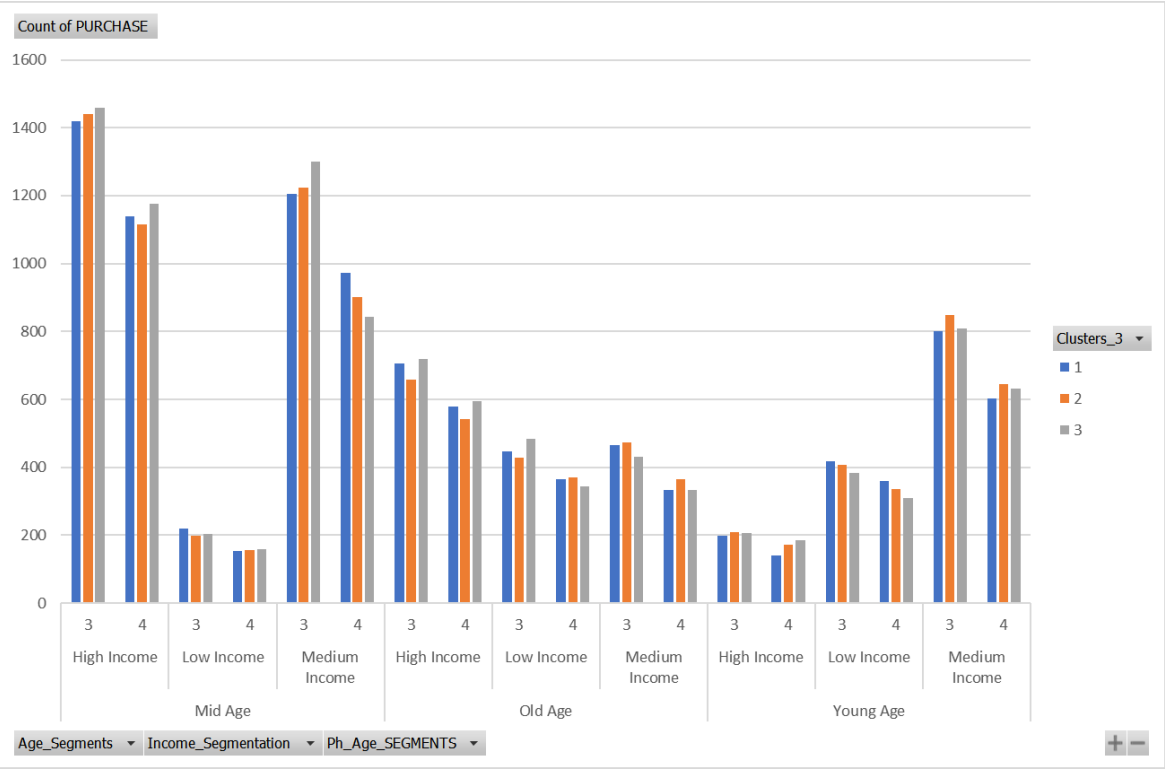
	Middle Aged Customers					
	Budget	Phn_age	count	Expected Revenue	sum	IF >20 CRORE
Low Incon	11500	<1.25years	442	5525000	14837500	No
Low Incon	13500	>1.25years	745	9312500		No
Medium I	17500	<1 year	0	0	137772000	No
Medium I	22000	ear-1.25ye	2396	51514000		
Medium I	25000	>1 .25year	4012	86258000		
High Incor	30000	<1year	0	0	242250000	Yes
High Incor	32500	>1year	7752	242250000		
Average					131619833.3	
	Segments				Age Criteria	
Young Age		1			25-34	
Mid Age		2			35-54	
Old Age		3			55-65	

Segmentation(Old Age Customers) and EDA Analysis (Clusters-3)



Old Aged Customers					
Budget	**Phn_age**	**count**	**Expected Revenue**	**sum**	**IF >20 CRORE**
10000	>9months	2438	23161000	23161000	No
9000	<9months	0	0		
17500	<1 year	0	0	51277500	No
22000	ear-1.25ye	892	19178000		
25000	>1 .25year	1493	32099500		
30000	<1year	0	0	118593750	No
32500	>1year	3795	118593750		
			average	48258062.5	
	Segments				**Age Criteria**
Young Age		1			25-34
Mid Age		2			35-54

Business Decision



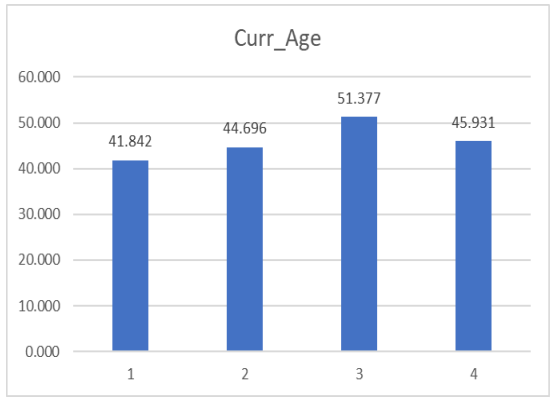
Looking at the table beside we can conclude that Customer Age Segment 2 ,count of purchase is 15283 which is way over our required result that is minimum 12000 Phones, sum of their ages is very high as 31573 and their Conversion rate is also high of 48.40 % in INDIA as compared to other segments so XYZ company must target Middle aged customers and so, yes they can enter in INDIAN market as there is no loss for XYZ Mobile Company.

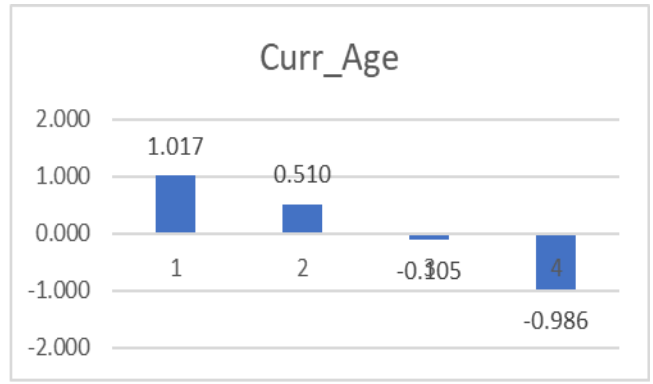
Count of PURCHASE	Column Labels			
Row Labels	1	2	3	Grand Total
Mid Age	**5111**	**5033**	**5139**	**15283**
High Income	**2559**	**2555**	**2635**	**7749**
3	1420	1441	1459	4320
4	1139	1114	1176	3429
Low Income	**374**	**353**	**363**	**1090**
3	220	197	204	621
4	154	156	159	469
Medium Income	**2178**	**2125**	**2141**	**6444**
3	1205	1223	1299	3727
4	973	902	842	2717
Old Age	**2896**	**2836**	**2902**	**8634**
High Income	**1285**	**1201**	**1312**	**3798**
3	705	659	718	2082
4	580	542	594	1716
Low Income	**813**	**799**	**826**	**2438**
3	447	429	483	1359
4	366	370	343	1079
Medium Income	**798**	**836**	**764**	**2398**
3	466	472	430	1368
4	332	364	334	1030
Young Age	**2517**	**2618**	**2521**	**7656**
High Income	**337**	**383**	**391**	**1111**
3	197	210	205	612
4	140	173	186	499
Low Income	**776**	**742**	**691**	**2209**
3	417	407	383	1207
4	359	335	308	1002
Medium Income	**1404**	**1493**	**1439**	**4336**
3	802	848	808	2458
4	602	645	631	1878
Grand Total	**10524**	**10487**	**10562**	**31573**

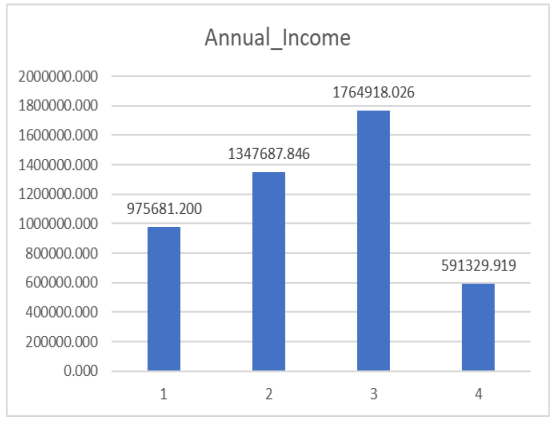
Cluster Centroids and error terms

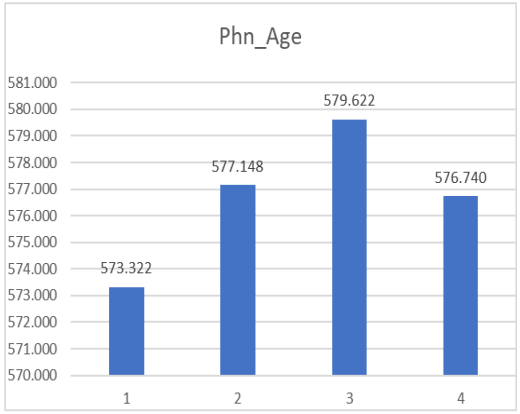
Centroid	1	2	3	4
Curr_Age	41.842	44.696	51.377	45.931
Annual_Income	975681.200	1347687.846	1764918.026	591329.919
Phn_Age	573.322	577.148	579.622	576.740

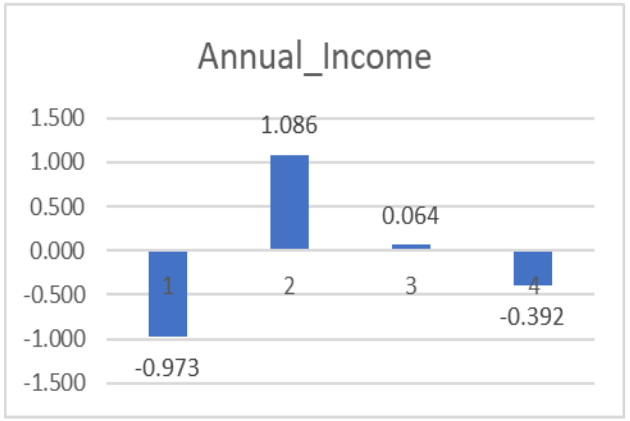
Centroid	1	2	3	4
Curr_Age	1.017	0.510	-0.105	-0.986
Annual_Income	-0.973	1.086	0.064	-0.392
Phn_Age	-0.368	-0.463	1.552	-0.517

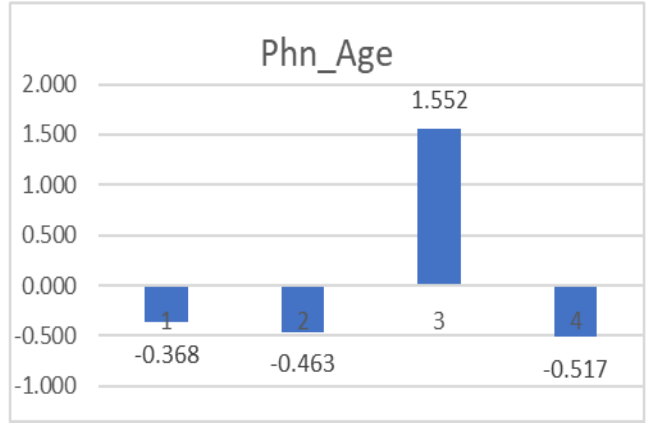












Task3: Model Evaluation (4 Clusters):

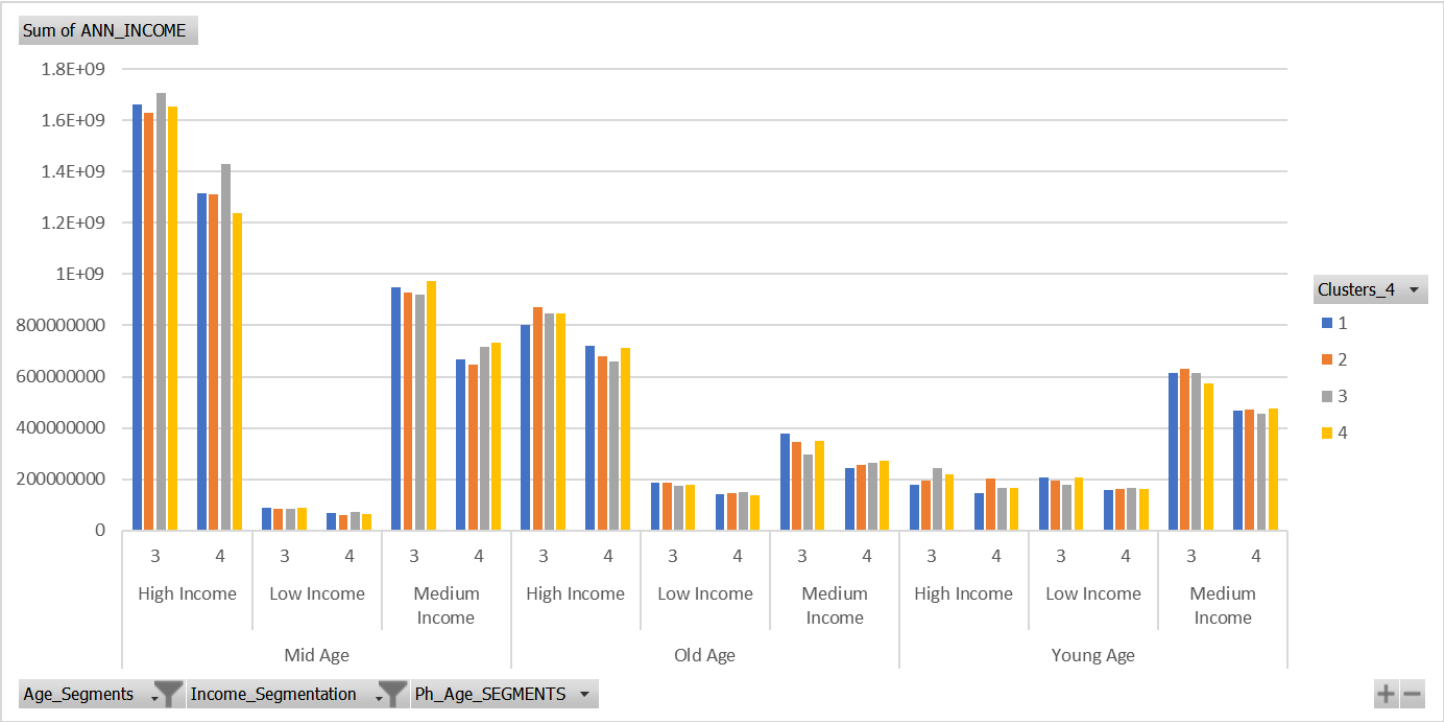
From the table below we can Conclude that **revenue collection is way over Rs. 20 Crores for segments i.e. Middle Age and Old Age with Medium Income and High income and Phone age >1year.**

	Young Aged Customers				
	Budget	**Phn_age**	**count**	**Expected Revenue**	**IF >20 CRORE**
Low_Income	8000	<1year	0	0	No
Medium_Income	12000	>1 year	13178	158136000	No
High_Income	12000	<1year	0	0	No
High_Income	15000	>1year	12658	189870000	No
			Average	174003000	
			Segments	**Age Criteria**	
	Young Age		1	25-34	
	Mid Age		2	35-54	
	Old Age		3	55-65	

	Middle Aged Customers					
	Budget	**Phn_age**	**count**	**Expected Revenue**	**sum**	**IF >20 CRORE**
Low Income	11500	<1.25years	2333	29162500	78750000	No
Low Income	13500	>1.25years	3967	49587500		No
Medium Inc	17500	<1 year	0	0	282015500	Yes
Medium Inc	22000	1 year-1.25years	4891	105156500		
Medium Inc	25000	>1 .25year	8226	176859000		
High Income	30000	<1year	0	0	395562500	Yes
High Income	32500	>1year	12658	395562500		
Average					252109333.3	
		Segments			**Age Criteria**	
Young Age		1			25-34	
Mid Age		2			35-54	
Old Age		3			55-65	

	Old Aged Customers					
	Budget	**Phn_age**	**count**	**Expected Revenue**	**sum**	**IF >20 CRORE**
Low_Income	10000	>9months	5737	54501500	54501500	No
Low_Income	9000	<9months	0	0		
Medium_Income	17500	<1 year	0	0	282015500	Yes
Medium_Income	22000	ear-1.25ye	4891	105156500		
Medium_Income	25000	>1 .25year	8226	176859000		
High_Income	30000	<1year	0	0	395562500	Yes
High_Income	32500	>1year	12658	395562500		
				average	244026500	
		Segments				**Age Criteria**
	Young Age		1			25-34
	Mid Age		2			35-54
	Old Age		3			55-65

Suggestion for Business Decision



Count of PURCHASE	Column Labels				
Row Labels	1	2	3	4	Grand Total
Mid Age	**3805**	**3719**	**3925**	**3834**	**15283**
High Income	**1932**	**1903**	**2035**	**1879**	**7749**
3	1077	1054	1111	1078	4320
4	855	849	924	801	3429
Low Income	**278**	**264**	**273**	**275**	**1090**
3	160	155	149	157	621
4	118	109	124	118	469
Medium Income	**1595**	**1552**	**1617**	**1680**	**6444**
3	935	918	910	964	3727
4	660	634	707	716	2717
Old Age	**2167**	**2189**	**2098**	**2180**	**8634**
High Income	**938**	**960**	**937**	**963**	**3798**
3	495	539	526	522	2082
4	443	421	411	441	1716
Low Income	**608**	**626**	**600**	**604**	**2438**
3	342	352	327	338	1359
4	266	274	273	266	1079
Medium Income	**621**	**603**	**561**	**613**	**2398**
3	378	347	298	345	1368
4	243	256	263	268	1030
Young Age	**1892**	**1943**	**1908**	**1913**	**7656**
High Income	**239**	**291**	**298**	**283**	**1111**
3	131	143	177	161	612
4	108	148	121	122	499
Low Income	**563**	**545**	**530**	**571**	**2209**
3	318	296	273	320	1207
4	245	249	257	251	1002
Medium Income	**1090**	**1107**	**1080**	**1059**	**4336**
3	619	635	622	582	2458
4	471	472	458	477	1878
Grand Total	**7864**	**7851**	**7931**	**7927**	**31573**

From the Table we can conclude that Customer Age Segment 2 ,count of purchase is 15283 which is way over our required result that is minimum 12000 Phones, sum of their ages is very high as 31573 and their Conversion rate is also high of 48.40 % in INDIA as compared to other segments so XYZ company must target Middle aged customers and so, yes they can enter in INDIAN market as there is no loss for XYZ Mobile Company.