1. Title and authors:

Predictive Analysis on the Success of Bank Telemarketing

Presented by Lakshmi Chaitanya Kakarla, Ravali Patlolla

2. Summary of questions:

1. **What type of job a particular education background person is doing?**
   So, here we are trying to analyze if there is any relation between the educational background of a person and to the type of job they are into.

2. **Does the duration of the call have any effect on term deposit subscription?**
   Here in this question we are trying to answer if duration of the call length has any effect on subscription like increase/decrease in call length lead to term deposit subscription.

3. **Has personal loan dependent on marital status and age?**
   In this question we are trying to analyze if marital status and age has led to take personal loans or not.

4. **Which is the customer most preferable communication type?**
   There are two types of communication modes used in this campaign. So we want to find out which one is most used.

5. **What factors are leading a customer/client to subscribe (yes) a term deposit (prediction variable)**
   Here we are planning to predict what all major variables are influencing to the success of the term deposit subscription.

3. Motivation and Background:

All most many of us might have received telemarketing calls quite often regarding health insurance, car insurance etc., so we took that idea, started searching for the data and found this data on bank telemarketing. The motivation for doing this project and selecting this particular data is Focus on maximizing customer lifetime value through the evaluation of available information and customer metrics, allows us to build longer and tighter relations in alignment with business demand. Also, the task of selecting the best set of clients, i.e., that are more likely to subscribe a product, is considered very important for the banking institution. The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y) based on the analysis of the input variables like age, type of job, educational status, contact communication type etc.,

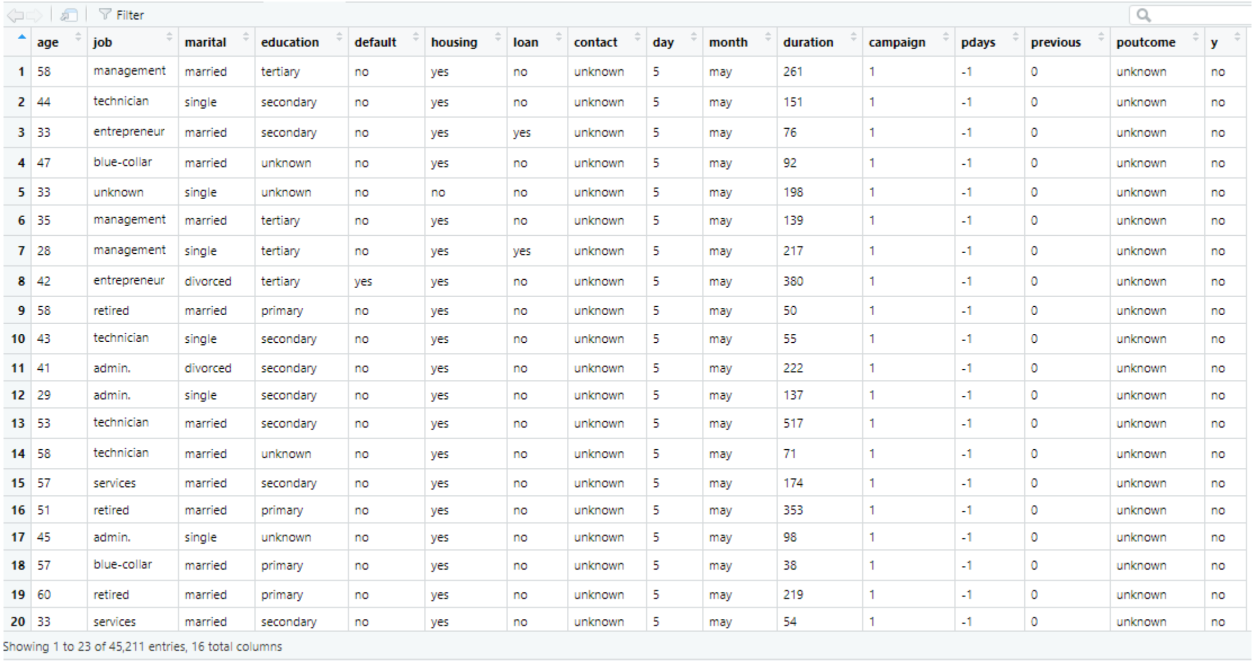
4. Dataset: The dataset which we are using is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. Below is the link to the URL from which we took the data. http://archive.ics.uci.edu/ml/datasets/Bank+Marketing

[Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

Attribute description:

Variable	**Description**
age	Age of the client
job	type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
marital	marital status (categorical: 'divorced', 'married', 'single'; note: 'divorced' means divorced or widowed)
education	(categorical: ‘primary', ‘secondary’, ‘tertiary’, ‘unknown')
default	Has credit in default? (categorical: 'no’, ‘yes')
housing	Has housing loan? (categorical: 'no’, ‘yes')
loan	Has personal loan? (categorical: 'no’, ‘yes')
contact	contact communication type (categorical: 'cellular', 'telephone', ’unknown’)
month	last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
day	last contact day of the month (“1-31 days of a month”)
duration	last contact duration, in seconds (numeric)
campaign	number of contacts performed during this campaign and for this client (numeric, includes last contact)
pdays	number of days that passed by after the client was last contacted from a previous campaign (numeric; -1 means client was not previously contacted)
previous	number of contacts performed before this campaign and for this client (numeric)
poutcome	outcome of the previous marketing campaign (categorical: 'failure’, ‘success', ‘other’, ‘unknown’)
Output variable (desired target): y	Has the client subscribed a **term deposit**? (binary: 'yes’, ‘no')

Sample Dataset:



	age	job	marital	education	default	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
1	58	management	married	tertiary	no	yes	no	unknown	5	may	261	1	-1	0	unknown	no
2	44	technician	single	secondary	no	yes	no	unknown	5	may	151	1	-1	0	unknown	no
3	33	entrepreneur	married	secondary	no	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
4	47	blue-collar	married	unknown	no	yes	no	unknown	5	may	92	1	-1	0	unknown	no
5	33	unknown	single	unknown	no	no	no	unknown	5	may	198	1	-1	0	unknown	no
6	35	management	married	tertiary	no	yes	no	unknown	5	may	139	1	-1	0	unknown	no
7	28	management	single	tertiary	no	yes	yes	unknown	5	may	217	1	-1	0	unknown	no
8	42	entrepreneur	divorced	tertiary	yes	yes	no	unknown	5	may	380	1	-1	0	unknown	no
9	58	retired	married	primary	no	yes	no	unknown	5	may	50	1	-1	0	unknown	no
10	43	technician	single	secondary	no	yes	no	unknown	5	may	55	1	-1	0	unknown	no
11	41	admin.	divorced	secondary	no	yes	no	unknown	5	may	222	1	-1	0	unknown	no
12	29	admin.	single	secondary	no	yes	no	unknown	5	may	137	1	-1	0	unknown	no
13	53	technician	married	secondary	no	yes	no	unknown	5	may	517	1	-1	0	unknown	no
14	58	technician	married	unknown	no	yes	no	unknown	5	may	71	1	-1	0	unknown	no
15	57	services	married	secondary	no	yes	no	unknown	5	may	174	1	-1	0	unknown	no
16	51	retired	married	primary	no	yes	no	unknown	5	may	353	1	-1	0	unknown	no
17	45	admin.	single	unknown	no	yes	no	unknown	5	may	98	1	-1	0	unknown	no
18	57	blue-collar	married	primary	no	yes	no	unknown	5	may	38	1	-1	0	unknown	no
19	60	retired	married	primary	no	yes	no	unknown	5	may	219	1	-1	0	unknown	no
20	33	services	married	secondary	no	yes	no	unknown	5	may	54	1	-1	0	unknown	no

Showing 1 to 23 of 45,211 entries, 16 total columns

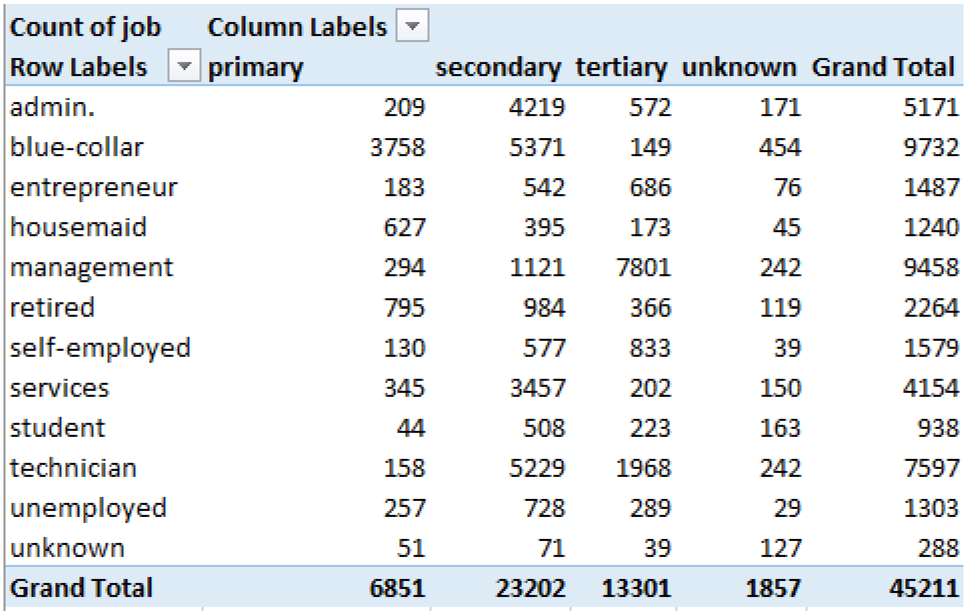
5. Methodology (analysis):

Over time, the increasing number of marketing campaigns has reduced their impact on the general public. In addition, economic pressures and competition have driven marketing executives to invest on directed campaigns with a stringent and thorough set of contacts. Using Business Intelligence (BI) and Data Mining (DM) techniques, these direct campaigns can be improved. Real-world data were collected from a Portuguese marketing campaign related with bank deposit subscription (from May 2008 to November 2010), in a total of 45,211 phone contacts. The business goal is to find a model that can explain success of a contact, i.e. if the client subscribes the deposit. Such a model can increase campaign efficiency by identifying key features that affect success, helping to better manage available resources (e.g. human effort, telephone calls, and time) and selecting a high quality and affordable set of potential buyers.

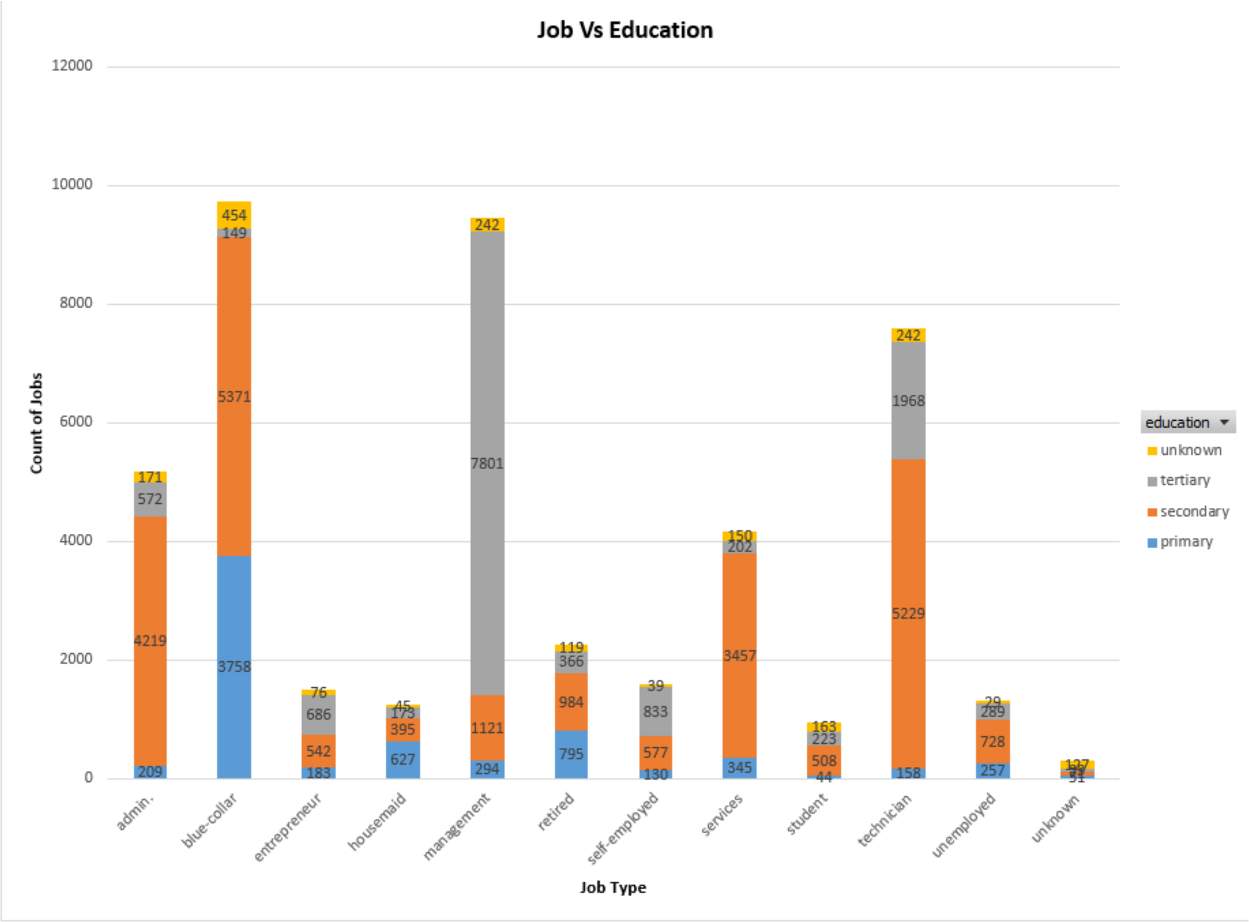
In this project, we first performed descriptive analysis of the bank telemarketing data to see how an attribute is related to other attributes and tried to answer the questions we got after looking into the data. Then we built a model using logistic regression algorithm to predict what factors are contributing to the subscription of a term deposit. With the help of this model, we trained our 75% of the data and then performed predictions over the rest 25% of the data. Finally, we calculated the accuracy of our predictions. With the predictions we made, we tried to give advices on possible outcomes which will help the business to focus on their targeted clients which helps them to turn this campaign into more successful one.

6. Results:

Q1) what type of job a particular education background person is doing?

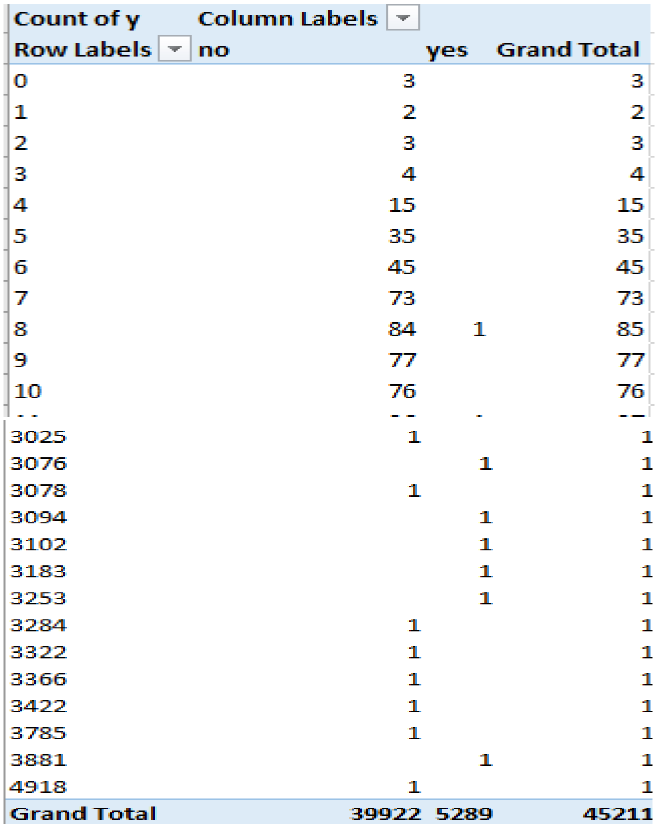


Count of job	Column Labels ▾				
Row Labels ▾	primary	secondary	tertiary	unknown	Grand Total
admin.	209	4219	572	171	5171
blue-collar	3758	5371	149	454	9732
entrepreneur	183	542	686	76	1487
housemaid	627	395	173	45	1240
management	294	1121	7801	242	9458
retired	795	984	366	119	2264
self-employed	130	577	833	39	1579
services	345	3457	202	150	4154
student	44	508	223	163	938
technician	158	5229	1968	242	7597
unemployed	257	728	289	29	1303
unknown	51	71	39	127	288
Grand Total	**6851**	**23202**	**13301**	**1857**	**45211**

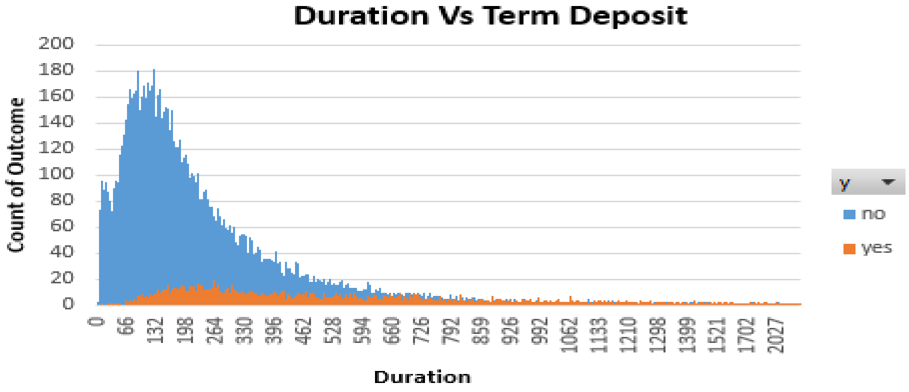


In this we tried to analyze what different kind of jobs persons with each educational level are performing. The highest number of jobs are seen in Management, blue-collar and technician categories across all education levels combined. From the above charts, we can see that most primary level education individuals are blue-collars (3758) and rest are scattered in small amounts to various other jobs. Secondary level education individuals are highly seen in technician (5229) and blue-collar (5371) jobs, with a very close difference in the number of persons into each job. A very large number of Management jobs (7801) are performed by Tertiary education background individuals.

Q2) does the duration of the call have any effect on term deposit subscription?

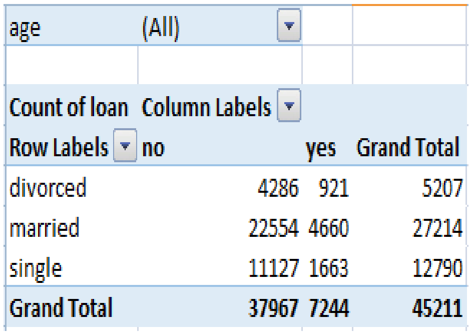


Count of y	Column Labels ▾		
Row Labels ▾	no	yes	Grand Total
0	3		3
1	2		2
2	3		3
3	4		4
4	15		15
5	35		35
6	45		45
7	73		73
8	84	1	85
9	77		77
10	76		76
3025	1		1
3076		1	1
3078	1		1
3094		1	1
3102		1	1
3183		1	1
3253		1	1
3284	1		1
3322	1		1
3366	1		1
3422	1		1
3785	1		1
3881		1	1
4918	1		1
Grand Total	39922	5289	45211



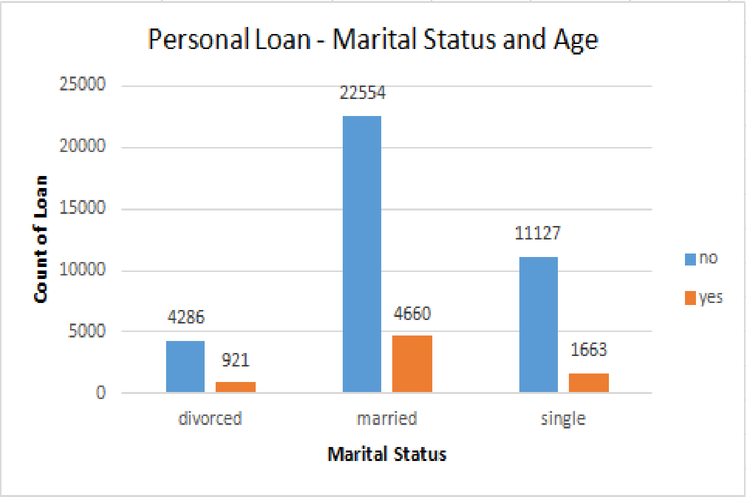
Here we tried to find if the duration of the call has made any difference in the subscription of the term deposit. We found that duration of call does had an effect on customers taking the term deposit. We have seen that increase in duration of the call has led the customers to subscribe the term deposit. From the graph above, though the number of customers who didnt subscribe to term deposit can be seen in large number, we can also see that the number of customers saying no to subscription has significantly decreased as length of duration increased.

Q3) has personal loan dependent on marital status and age?

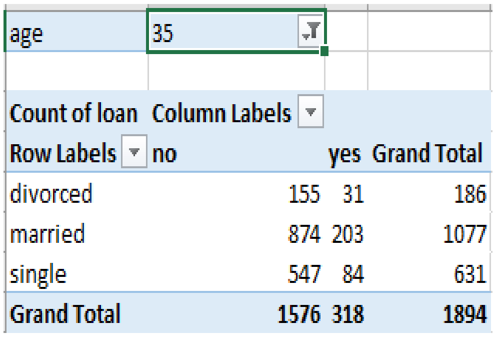


age (All)

Count of loan	Column Labels		
Row Labels	no	yes	Grand Total
divorced	4286	921	5207
married	22554	4660	27214
single	11127	1663	12790
Grand Total	37967	7244	45211

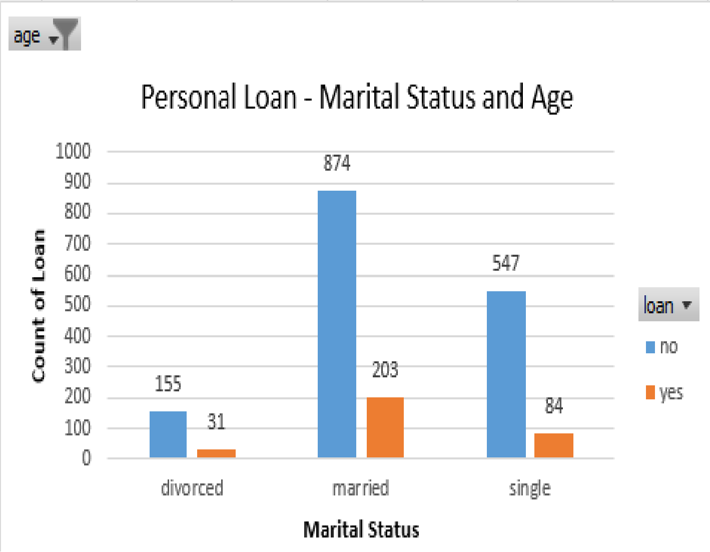


Data showing for all ages



age 35

Count of loan	Column Labels		
Row Labels	no	yes	Grand Total
divorced	155	31	186
married	874	203	1077
single	547	84	631
Grand Total	1576	318	1894



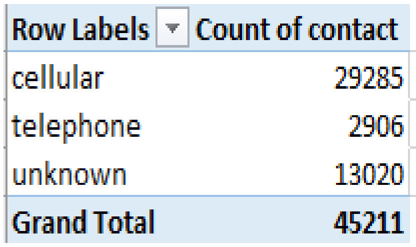
Data showing for age 35 where we highest number of people taking personal loans

In this question we wanted to analyze factors that are causing a person to go for a personal loan. So we took two factors: Marital status and age to analyze the dependency on personal loan based on these two. We found that, overall combining all ages married persons are the ones who have the most personal loans taken and divorced individuals are at lowest number to take personal loans. Out of all people of different ages, we found that married people between ages 30-40 years have taken the higher number of loans, with married people around 35years being highest. At the age 35, highest number of loans are taken by married people. With increase in the age after 40years, the persons with personal loans have significantly decreased.

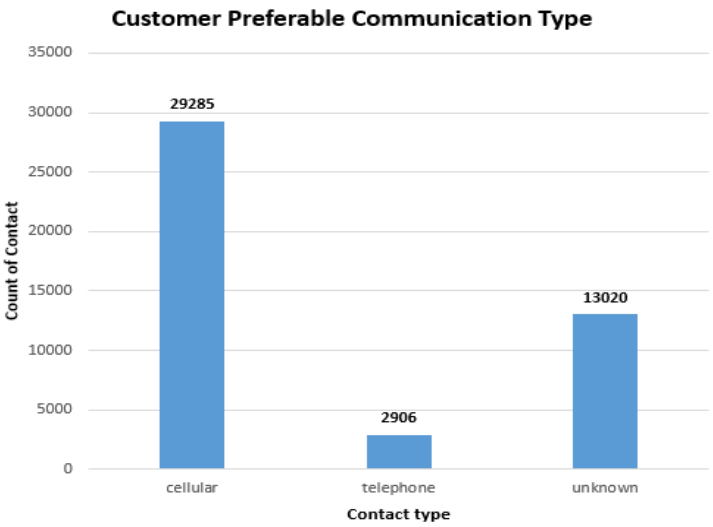
Age	No of loans
25	20
26	48
27	56
28	78
29	97
30	148
31	168
32	175
33	196
34	195
35	203
36	173
37	182
38	135

Age	No of loans
39	140
40	163
41	155
42	124
43	132
44	143
45	148
46	136
47	156
48	132
60	61
62	3
72	1
66	1

Q4) which is the customer/client most preferable communication type?



Row Labels	Count of contact
cellular	29285
telephone	2906
unknown	13020
Grand Total	**45211**

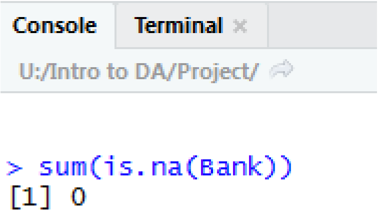


Here in this we tried to analyze which mode of communication is the most customers preferred one so that we can use this one for the next telemarketing campaign. From the charts above we found Cellular is the most preferred mode of communication over telephone, which numbers of customers using cellular as communication type to be 29285people and telephone customers at 2906. Here we have to take into account that a large part of customers preferred communication type is unknown (13020).

Q5) what factors are leading a customer/client to subscribe (yes) a term deposit (prediction variable)

PREPARATION AND CONVERSION OF THE DATA:

Before converting the data into numerical form we have checked for any missing values in the entire data frame. The image below shows that the data set has 0 missing values. After that the data which was in categorical form is mapped into numerical form to perform logistic regression and prediction.



U:/Intro to DA/Project/

```
> sum(is.na(Bank))
[1] 0
```

The mapping of categorical data is as follows:

Job: 'admin.-1','blue-collar-2','entrepreneur-3','housemaid-4','management-5','retired-6','self-employed-7','services-8','student-9','technician-10','unemployed-11','unknown-0'

Marital status: (categorical: 'divorced-0', 'married-1', and ‘single-2’)

Education: (categorical: ‘primary-1', ‘secondary-2’, ‘tertiary-3’, and ‘unknown-0')

Default: Has credit in default? (Categorical: 'no-0’, ‘yes-1')

Housing Loan: Has housing loan? (Categorical: 'no-0’, ‘yes-1')

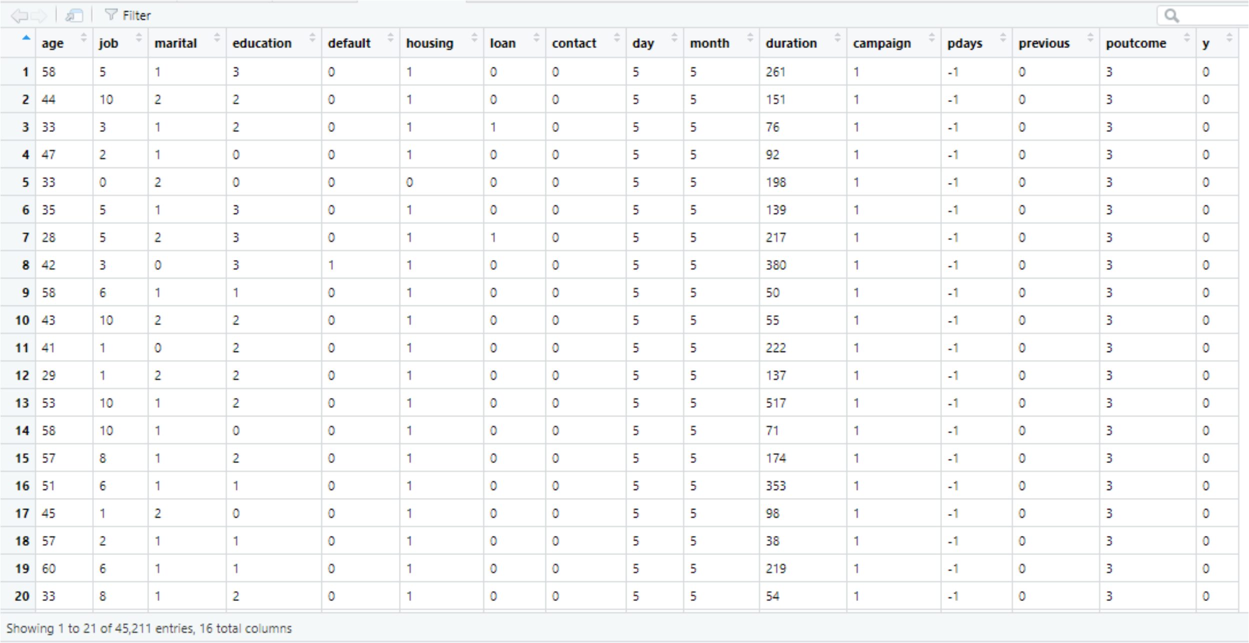
Loan: Has personal loan? (Categorical: 'no-0’, ‘yes-1')

Contact: Contact communication type (categorical: 'cellular-1', 'telephone-2', and ’unknown-0’)

Month: Last contact month of year (categorical: 'jan-1', 'feb-2', 'mar-3’... 'Nov-11', 'dec-12')
Poutcome: Outcome of the previous marketing campaign (categorical: 'failure-0’, ‘success-1', ‘other-2’, ‘unknown-3’)

Target Variable: Has the client subscribed a **term deposit**? (‘yes-1’, ‘no-0')

The image below shows the data after conversion into numerical values.

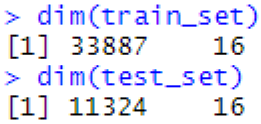


	age	job	marital	education	default	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
1	58	5	1	3	0	1	0	0	5	5	261	1	-1	0	3	0
2	44	10	2	2	0	1	0	0	5	5	151	1	-1	0	3	0
3	33	3	1	2	0	1	1	0	5	5	76	1	-1	0	3	0
4	47	2	1	0	0	1	0	0	5	5	92	1	-1	0	3	0
5	33	0	2	0	0	0	0	0	5	5	198	1	-1	0	3	0
6	35	5	1	3	0	1	0	0	5	5	139	1	-1	0	3	0
7	28	5	2	3	0	1	1	0	5	5	217	1	-1	0	3	0
8	42	3	0	3	1	1	0	0	5	5	380	1	-1	0	3	0
9	58	6	1	1	0	1	0	0	5	5	50	1	-1	0	3	0
10	43	10	2	2	0	1	0	0	5	5	55	1	-1	0	3	0
11	41	1	0	2	0	1	0	0	5	5	222	1	-1	0	3	0
12	29	1	2	2	0	1	0	0	5	5	137	1	-1	0	3	0
13	53	10	1	2	0	1	0	0	5	5	517	1	-1	0	3	0
14	58	10	1	0	0	1	0	0	5	5	71	1	-1	0	3	0
15	57	8	1	2	0	1	0	0	5	5	174	1	-1	0	3	0
16	51	6	1	1	0	1	0	0	5	5	353	1	-1	0	3	0
17	45	1	2	0	0	1	0	0	5	5	98	1	-1	0	3	0
18	57	2	1	1	0	1	0	0	5	5	38	1	-1	0	3	0
19	60	6	1	1	0	1	0	0	5	5	219	1	-1	0	3	0
20	33	8	1	2	0	1	0	0	5	5	54	1	-1	0	3	0

Showing 1 to 21 of 45,211 entries, 16 total columns

SPLITTING INTO TRAIN AND TEST DATA:

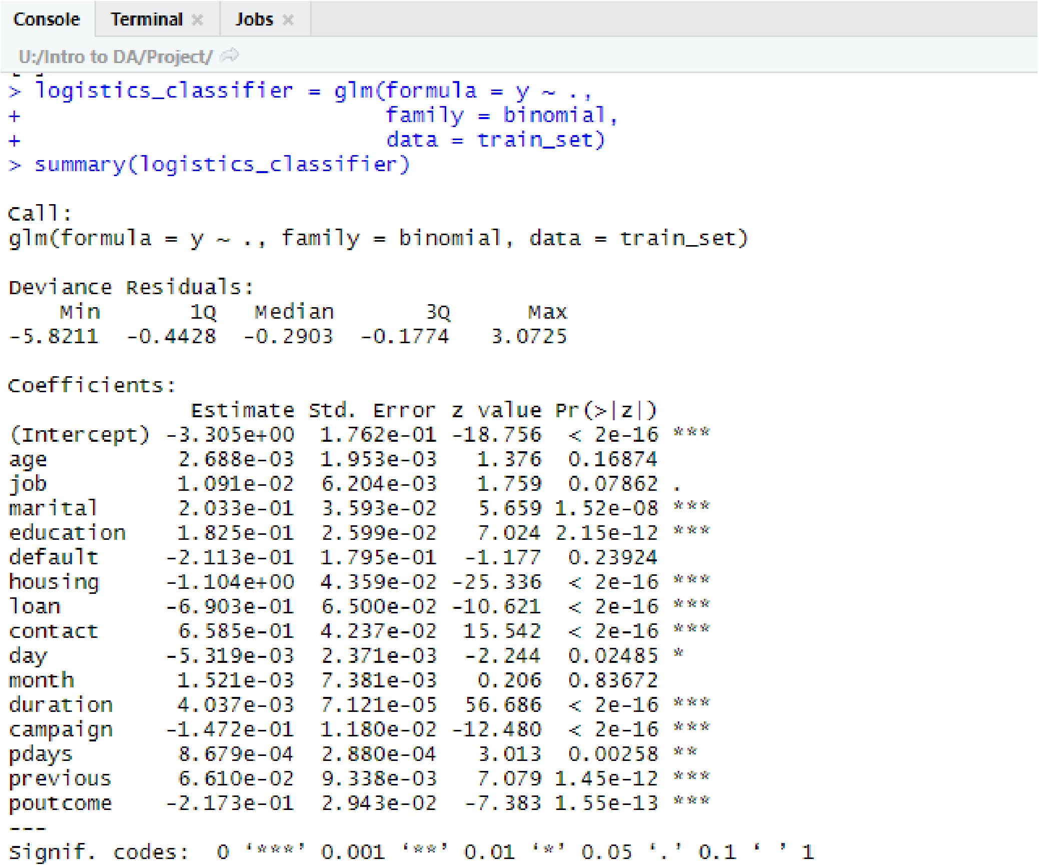
It is the usual practice in Machine Learning field to divide the data set into train and test set. The model will be built on the train set and the performance of the model will be tested on the test. Here we divided 75% of the data into train_set and 25% of the data into test_set. The image below is to show the dimensions of train and test data sets.



```
> dim(train_set)
[1] 33887    16
> dim(test_set)
[1] 11324    16
```

LOGISTIC REGRESSION ALGORITHM:

Logistic Regression is a classification algorithm. It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. To represent binary/categorical outcome, we use dummy variables. It uses sigmoid function to classify variables into classes and it’s basically applicable to classification problems. Fitting Logistic Regression to the train_set and finding which factors are leading a customer/client to subscribe (yes) for a term deposit (prediction variable)



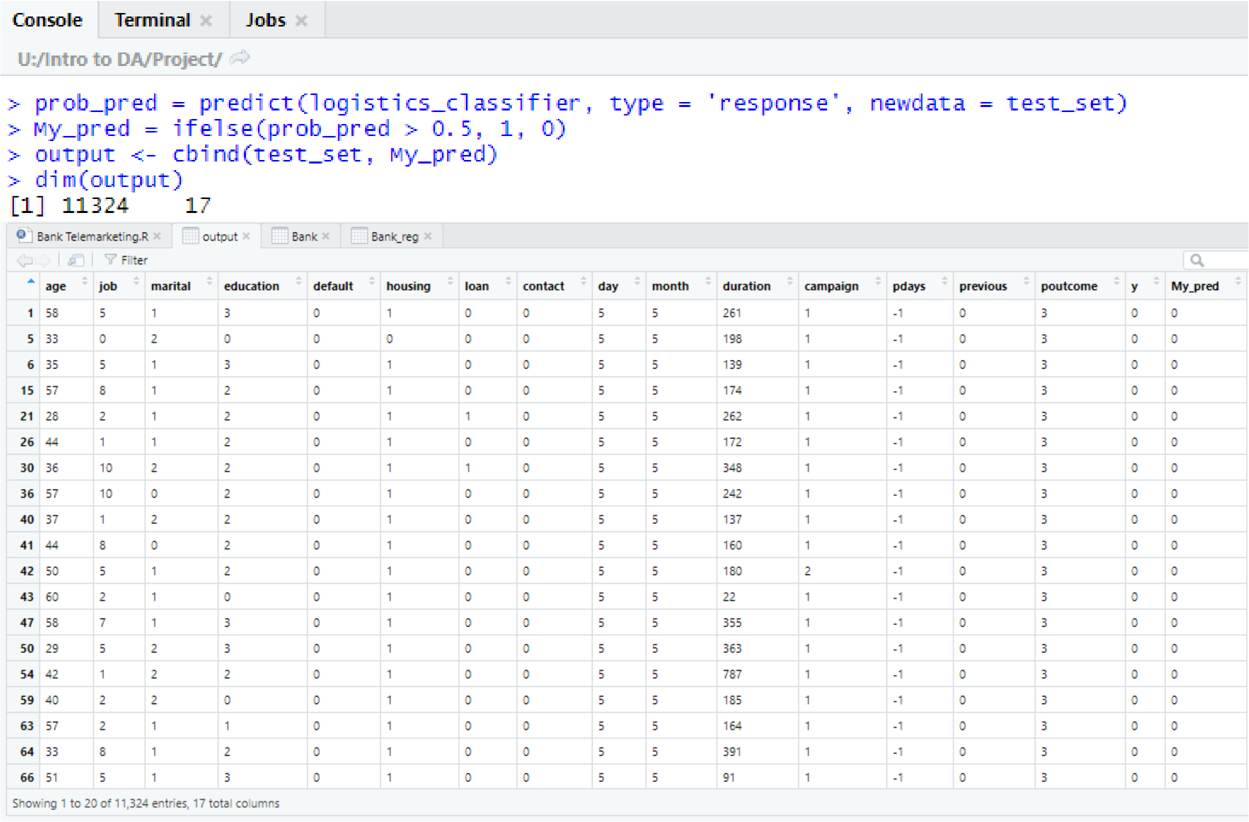
```
> logistics_classifier = glm(formula = y ~ .,
+                            family = binomial,
+                            data = train_set)
> summary(logistics_classifier)

Call:
glm(formula = y ~ ., family = binomial, data = train_set)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-5.8211  -0.4428  -0.2903  -0.1774   3.0725  

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -3.305e+00  1.762e-01 -18.756  < 2e-16 ***
age          2.688e-03  1.953e-03   1.376  0.16874    
job          1.091e-02  6.204e-03   1.759  0.07862 .  
marital      2.033e-01  3.593e-02   5.659 1.52e-08 ***
education    1.825e-01  2.599e-02   7.024 2.15e-12 ***
default     -2.113e-01  1.795e-01  -1.177  0.23924    
housing     -1.104e+00  4.359e-02 -25.336  < 2e-16 ***
loan        -6.903e-01  6.500e-02 -10.621  < 2e-16 ***
contact      6.585e-01  4.237e-02  15.542  < 2e-16 ***
day         -5.319e-03  2.371e-03  -2.244  0.02485 *  
month        1.521e-03  7.381e-03   0.206  0.83672    
duration     4.037e-03  7.121e-05  56.686  < 2e-16 ***
campaign    -1.472e-01  1.180e-02 -12.480  < 2e-16 ***
pdays        8.679e-04  2.880e-04   3.013  0.00258 ** 
previous     6.610e-02  9.338e-03   7.079 1.45e-12 ***
poutcome    -2.173e-01  2.943e-02  -7.383 1.55e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Based on the output of the Logistic regression, only 9 variables are significant while other are insignificant. The significant variables to be considered in our prediction model are marital, education, housing, loan, contact, duration, campaign, previous and poutcome.

PREDICTION USING THE MODEL

Using this logistic classifier model we have trained our dataset and now it’s the time for prediction. In logistic regression, we are only concerned about the probability of outcome dependent variable (success or failure). If the probability of the prediction is greater than 0.5 then we are predicting it as Yes to term deposit, otherwise No.

Whenever the log of odd ratio is found to be positive, the probability of success is always more than 50%. We can say that the probability never goes below 0 and above 1.



```
> prob_pred = predict(logistics_classifier, type = 'response', newdata = test_set)
> My_pred = ifelse(prob_pred > 0.5, 1, 0)
> output <- cbind(test_set, My_pred)
> dim(output)
[1] 11324    17
```

	age	job	marital	education	default	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y	My_pred
1	58	5	1	3	0	1	0	0	5	5	261	1	-1	0	3	0	0
5	33	0	2	0	0	0	0	0	5	5	198	1	-1	0	3	0	0
6	35	5	1	3	0	1	0	0	5	5	139	1	-1	0	3	0	0
15	57	8	1	2	0	1	0	0	5	5	174	1	-1	0	3	0	0
21	28	2	1	2	0	1	1	0	5	5	262	1	-1	0	3	0	0
26	44	1	1	2	0	1	0	0	5	5	172	1	-1	0	3	0	0
30	36	10	2	2	0	1	1	0	5	5	348	1	-1	0	3	0	0
36	57	10	0	2	0	1	0	0	5	5	242	1	-1	0	3	0	0
40	37	1	2	2	0	1	0	0	5	5	137	1	-1	0	3	0	0
41	44	8	0	2	0	1	0	0	5	5	160	1	-1	0	3	0	0
42	50	5	1	2	0	1	0	0	5	5	180	2	-1	0	3	0	0
43	60	2	1	0	0	1	0	0	5	5	22	1	-1	0	3	0	0
47	58	7	1	3	0	1	0	0	5	5	355	1	-1	0	3	0	0
50	29	5	2	3	0	1	0	0	5	5	363	1	-1	0	3	0	0
54	42	1	2	2	0	1	0	0	5	5	787	1	-1	0	3	0	0
59	40	2	2	0	0	1	0	0	5	5	185	1	-1	0	3	0	0
63	57	2	1	1	0	1	0	0	5	5	164	1	-1	0	3	0	0
64	33	8	1	2	0	1	0	0	5	5	391	1	-1	0	3	0	0
66	51	5	1	3	0	1	0	0	5	5	91	1	-1	0	3	0	0

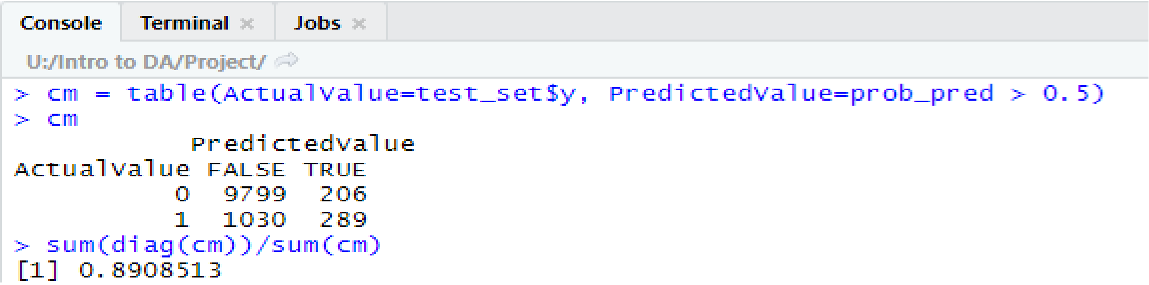
Showing 1 to 20 of 11,324 entries, 17 total columns

CONFUSION MATRIX

We can check by building a confusion matrix to display the success rate of our model's predictions on the test_set data we created earlier. The table function builds the confusion matrix. Going diagonally, (True Negative=9799, True Positive=289) represent the number of correct predictions.

Conversely, the going up diagonally, (False Negative=1030, False Positive=206) represent the number of incorrect predictions. We can calculate the accuracy of your model with the formula:

$$\frac{\text{True Positive} + \text{True Negatives}}{\text{True Positive} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}}$$

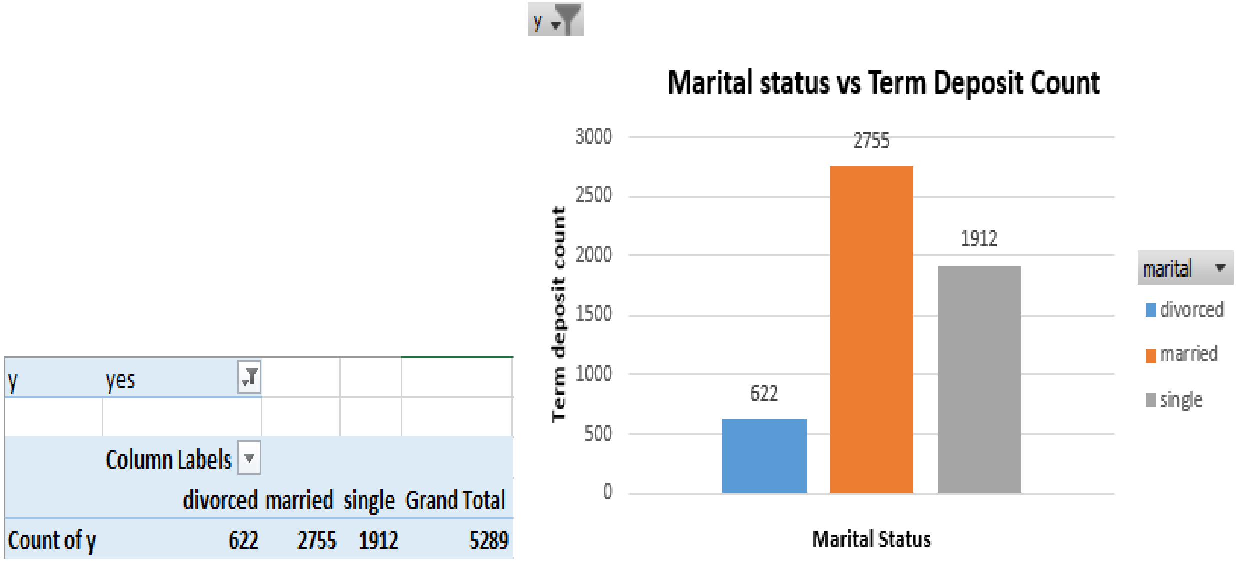


```
> cm = table(Actualvalue=test_set$y, Predictedvalue=prob_pred > 0.5)
> cm
           Predictedvalue
Actualvalue FALSE TRUE
          0  9799  206
          1  1030  289
> sum(diag(cm))/sum(cm)
[1] 0.8908513
```

Logistics Regression was able to give us an accuracy of 89.08%, which means that we can expect our model to classify correct about 9 observations in every 10.

After the prediction, we performed correlation between each of the suggested variable and output variable, term deposit= yes individually to know more insights.

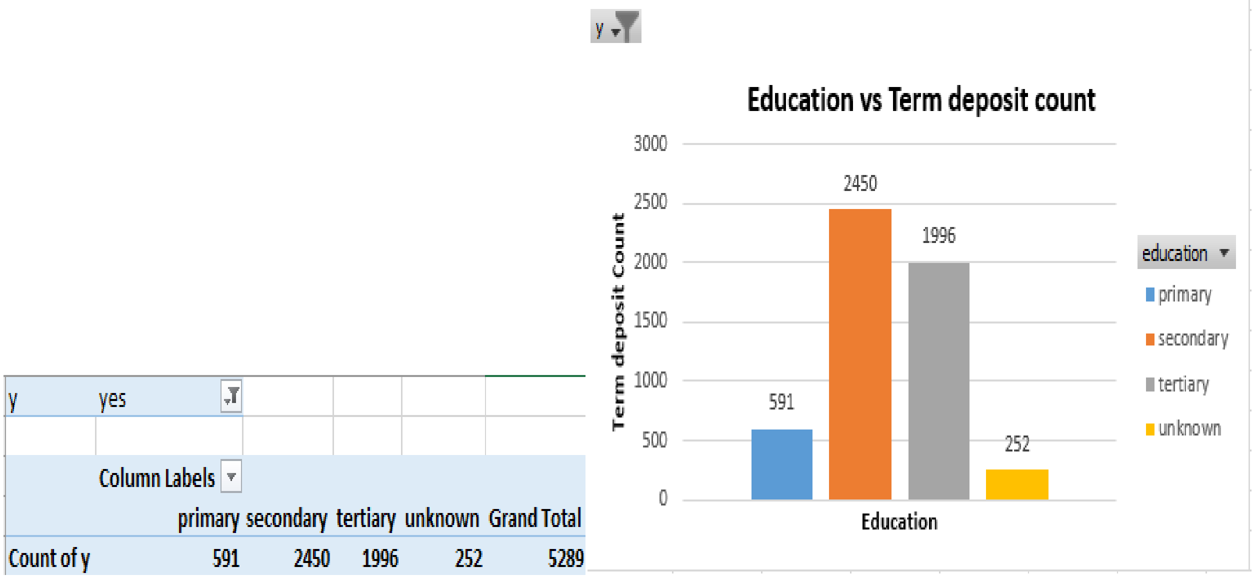
1) Marital vs Term deposit = yes



y	yes			
	Column Labels			
	divorced	married	single	Grand Total
Count of y	622	2755	1912	5289

From this graph, we can say that the married people are higher in number who said yes to a term deposit compared to divorced and single marital status.

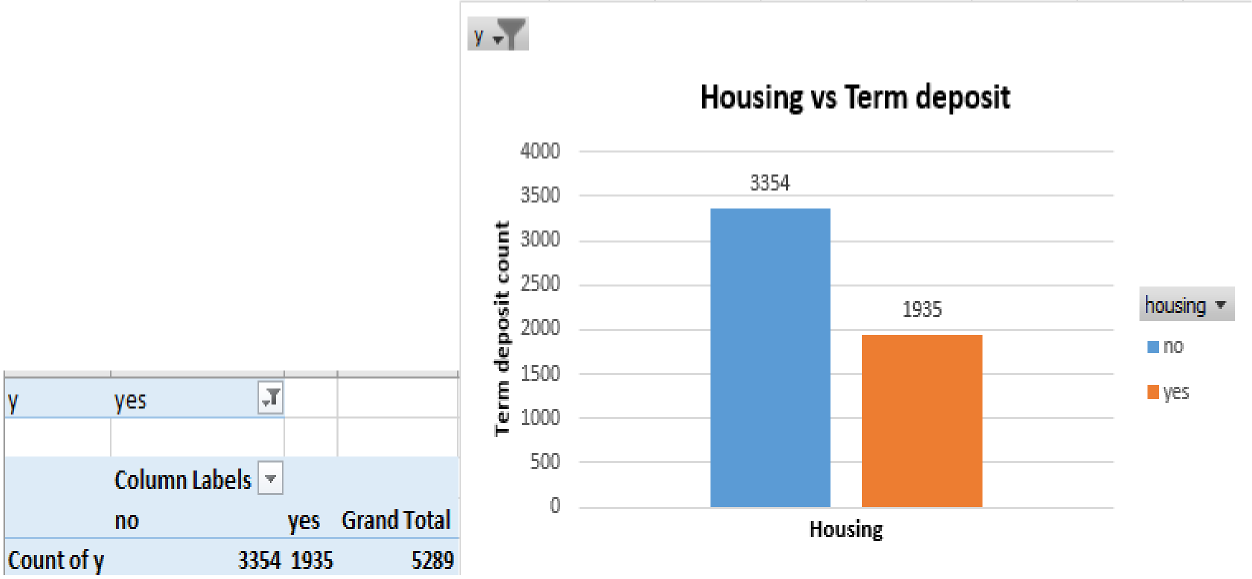
2) Education vs Term deposit = yes



y	yes				
	Column Labels				
	primary	secondary	tertiary	unknown	Grand Total
Count of y	591	2450	1996	252	5289

From this graph, we can say that the people who have done their secondary education are higher in number who said yes to a term deposit compared to other education status.

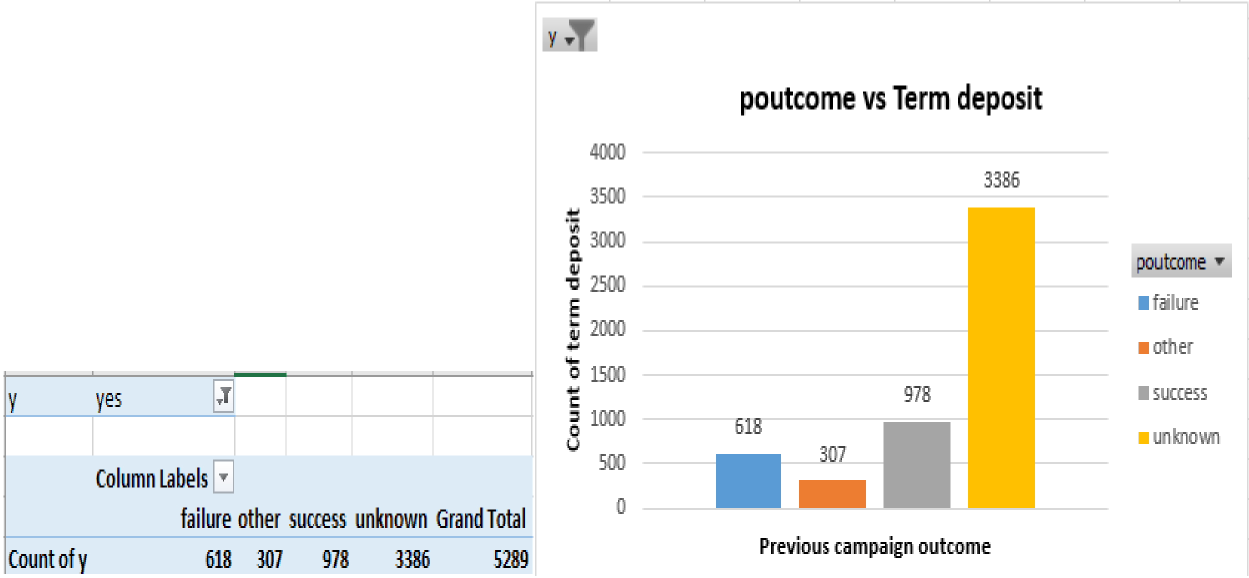
3) Housing vs Term deposit = yes



y	yes		
	Column Labels		
	no	yes	Grand Total
Count of y	3354	1935	5289

From this graph, we can say that the people who have no housing loan are higher in number who said yes to a term deposit compared to people who have a housing loan.

4) poutcome vs Term deposit = yes



y	yes				
	Column Labels				
	failure	other	success	unknown	Grand Total
Count of y	618	307	978	3386	5289

From this graph, we can say that the success of the previous marketing campaign has a positive influence on term deposit subscription.

7. Collaboration:

We both discussed, and proceeded with the project step by step.

Reflection: From this assignment, we learnt how to perform descriptive analysis of the data with which we were able to draw some insights and how to do further predictions with that data. As the dataset is unbalanced, as only 5289 (11.69%) records are related with successes. We wish we had more positive data but we understand that is not always the case to get positive results, so we wish our analysis will help the business to focus on their targeted clients which helps them to turn this campaign into more successful one. We wish we had explored more options with Tableau which would reduce our efforts. The piece of advice that we would like to offer to future students embarking on this project is to consider the socio-economic factors of the customer/client also into account while performing the predictions.