Bank Marketing Data Analysis

Understanding behaviors and trends of people who subscribe to term deposits

Overview

- What is a term deposit?
  - A deposit in a financial institution with a specific maturity date
- The infamous Wells Fargo cross-selling scandal
  - Many banks engage in marketing activities to prompt their financial products
- How data science comes into play
  - Leveraging marketing analytics in financial institutions to predict likelihood of deposit subscription
  - Indicator attributes on personal data

Research Scope

- Understand patterns in the dataset about bank user base, relationships between factors and how they contribute to whether people subscribe to term deposits
- In Scope
  - Exploratory data analysis on categorical variables
  - Exploratory data analysis on numerical variables
  - In-depth bi-variate analysis
- Out of Scope:
  - Machine learning
  - Statistical analysis

Dataset Description

Bank Marketing Data Set

Download: Data Folder, Data Set Description

Abstract: The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (variable y).

Data Set Characteristics:	Multivariate	**Number of Instances:**	45211	**Area:**	Business
Attribute Characteristics:	Real	**Number of Attributes:**	17	**Date Donated**	2012-02-14
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	1805119

- The data is related to direct marketing campaigns of a Portuguese bank institution from 2008 to 2010
- Campaigns were based on phone calls made to the clients to assess if term deposit would be subscribed
- 40k entries, 20 input variables (numerical and categorical), 1 output binary variable

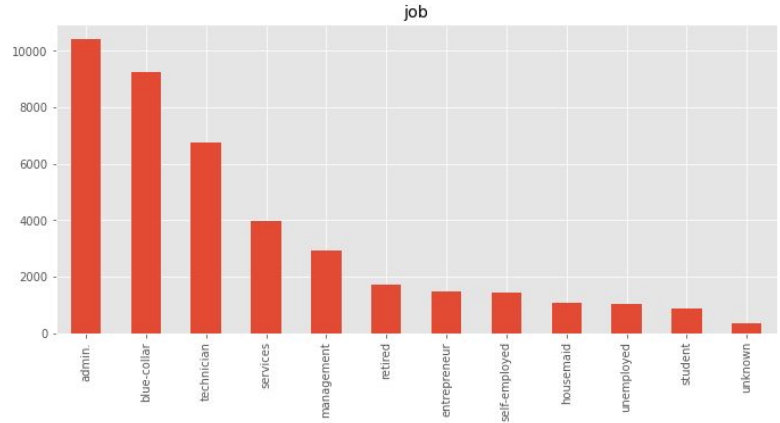
Sanity-Check & Cleaning Data

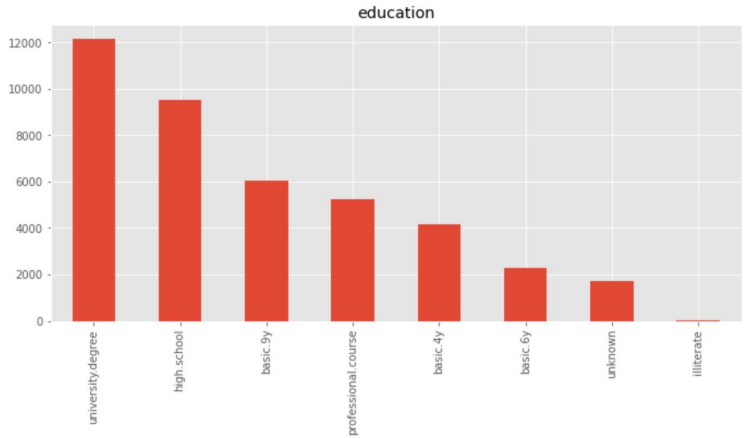
- Dataset has 41188 rows and 21 columns
- Cleaning column names- Replacing '.' with '_'
- 'unknown' data in categorical columns replaced with mode
- Analysing null values
- Analysing data types
- Dummies for target column with 'yes', 'no'

	age	job	marital	education	default	housing	loan	contact	month	day_of_
0	56	housemaid	married	basic.4y	no	no	no	telephone	may	
1	57	services	married	high.scho	unknown	no	no	telephone	may	
2	37	services	married	high.school	no	yes	no	telephone	may	
3	40	admin.	married	basic.6y	no	no	no	telephone	may	
4	56	services	married	high.school	no	no	yes	telephone	may	

Categorical Value Analysis

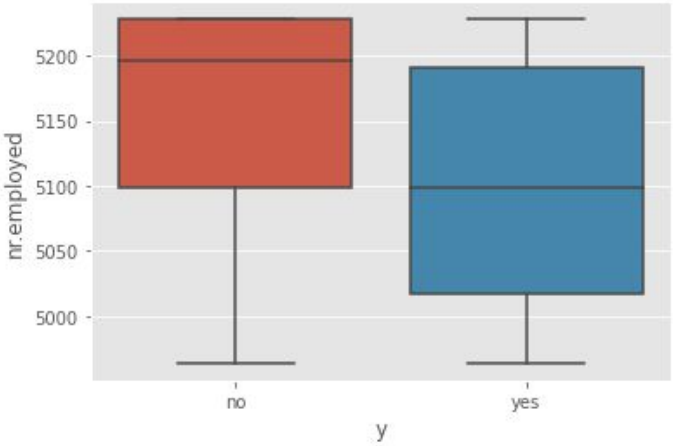
- Selected all categorical values and plotted for the frequency distribution using ggplot
- Key Findings:
  - Majority of customers married
  - Top professions were blue-collar jobs, technicians, administration, etc.
  - Majority of contacts occurred in May
  - Majority of customers did not subscribe (had ‘no’ as the output Y variable)





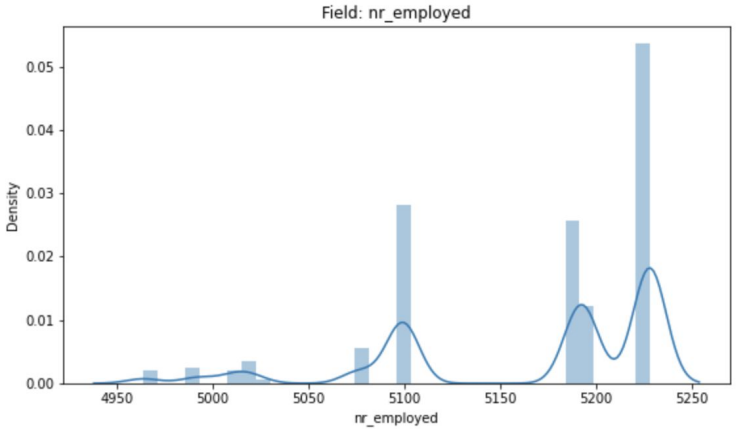
Categorical Value Analysis

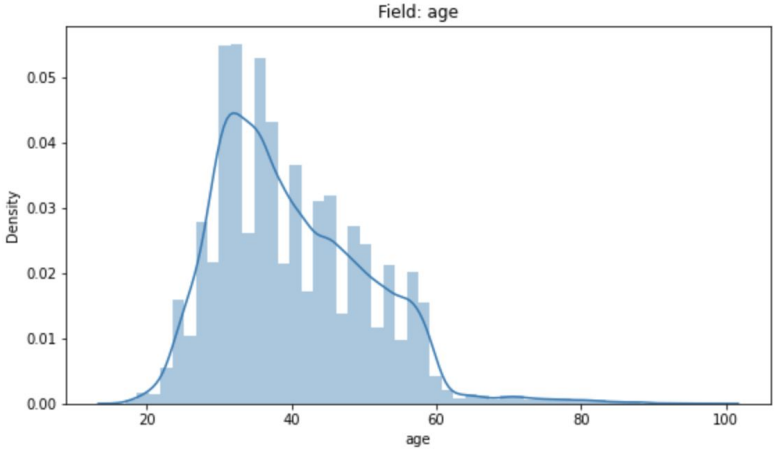
- Analyzing the distribution of categorical variables in relation to output variable
  - Created array of input categorical variables
  - Iterated and plotted via SNS Seaborn
- Inverse relationship between term subscription and education



Numerical Value Analysis

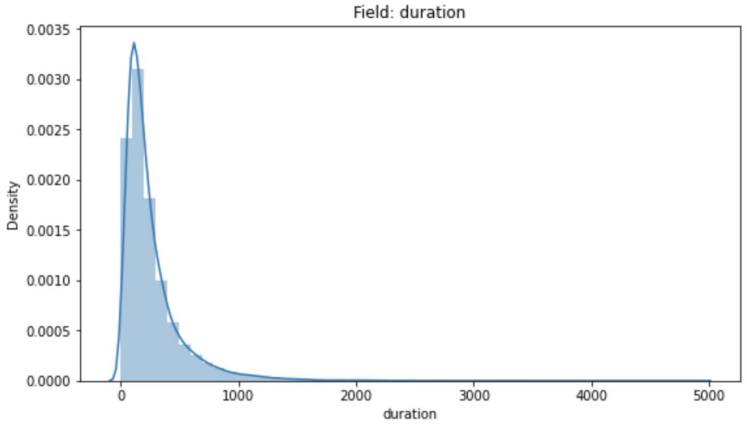
- Selected all numeric (continuous) values and analyzed them one column at a time
- Key Findings:
  - Median age is about 30, so many are on the younger side
  - Remaining fields, including indices, have much more discrete distributions, especially nr_employed

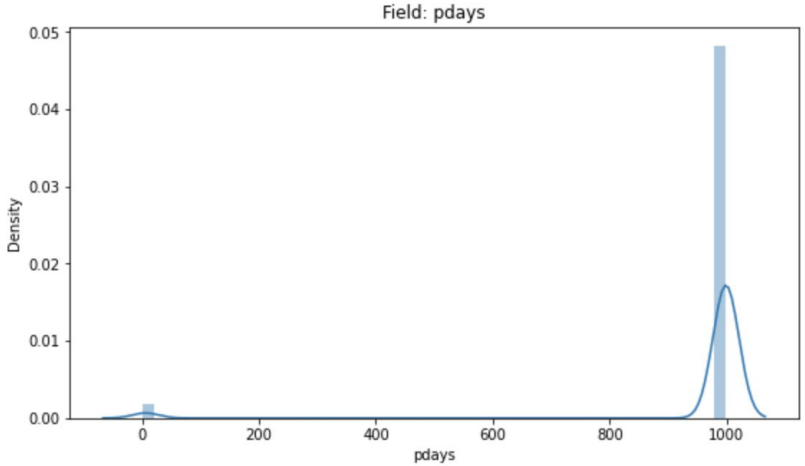




Numerical Value Analysis

- Key Findings
  - Durations of many transactions is shorter, with much of the data not taking more than ~200
  - ‘Pdays’ field only has two values. Majority are 1000



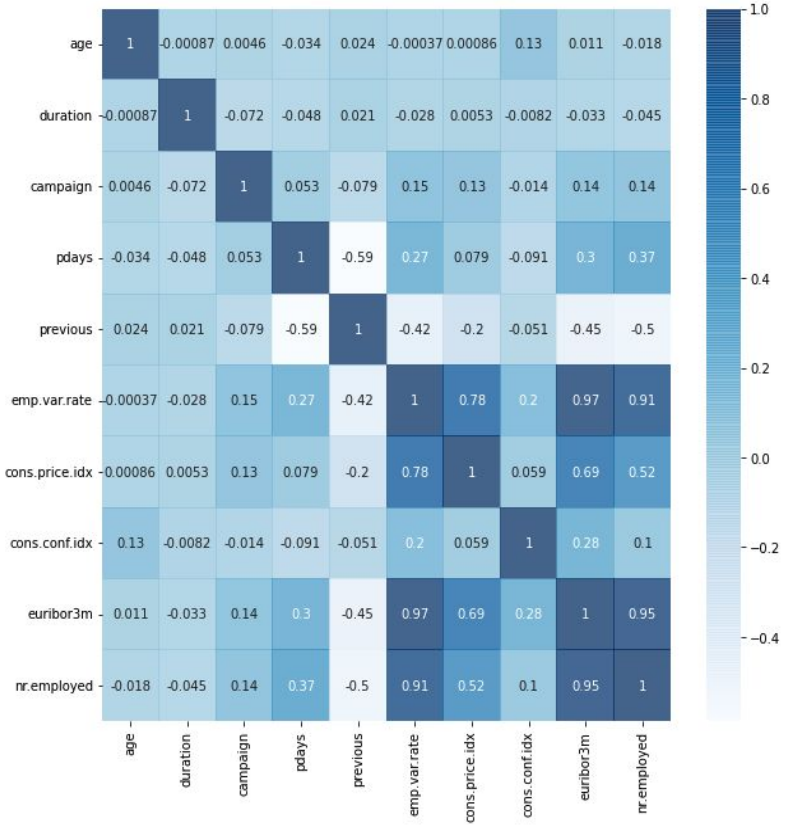


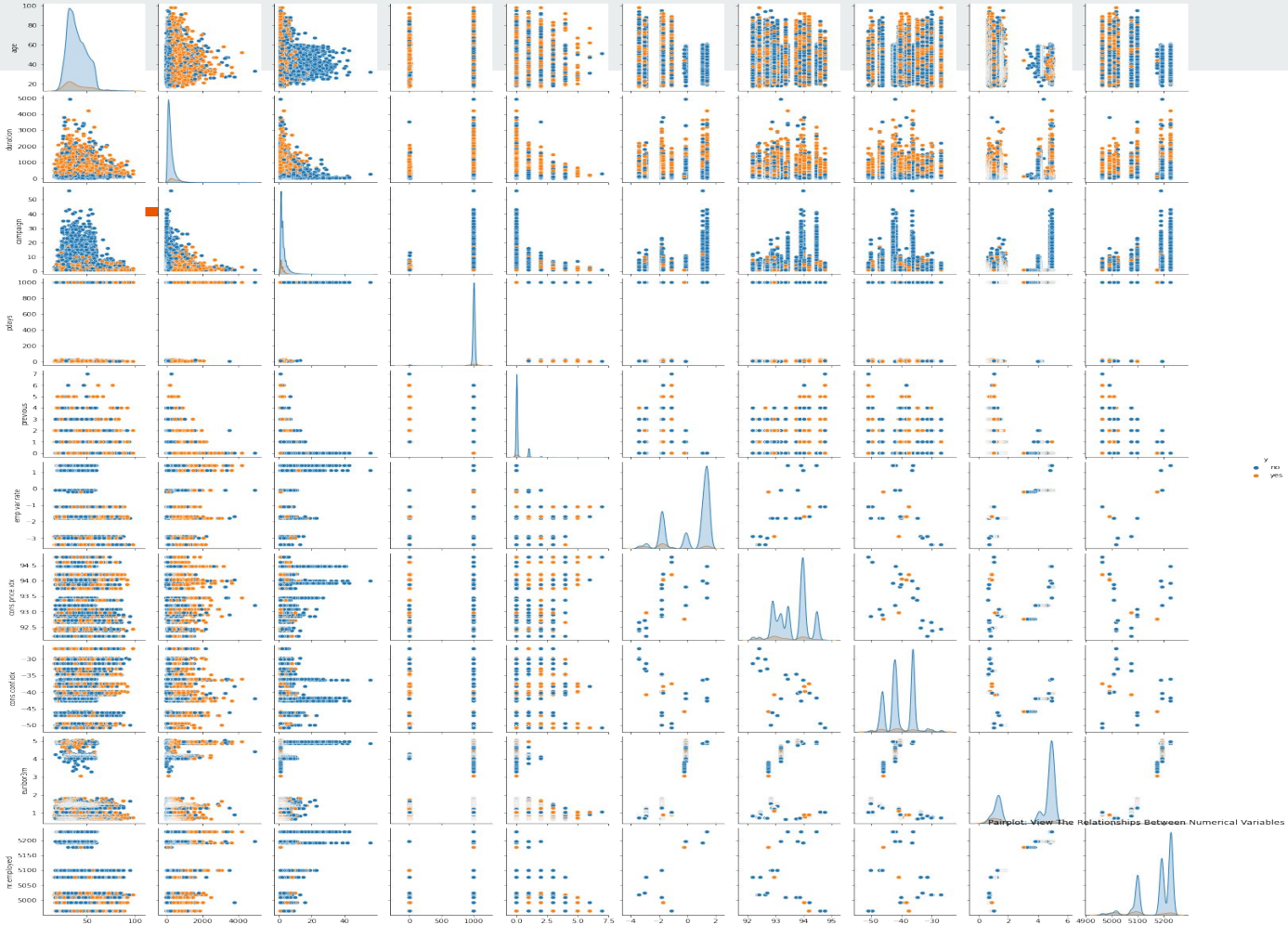
Bivariate Analysis

Bivariate Analysis of continuous variables

- Considered correlation >=0.5 or <=-0.5
- Ignored 'euribor3m' and 'cons_price_idx'
- Explore more on 'pdays' vs 'previous' , 'nr.employed' vs 'previous', emp.var.rate' vs 'nr.employed'

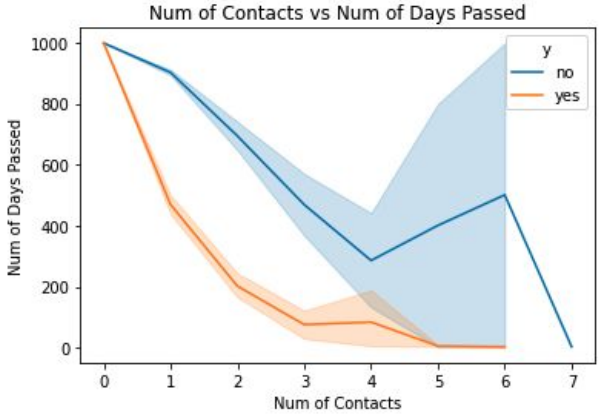
* previous-No of contacts made in previous in campaign

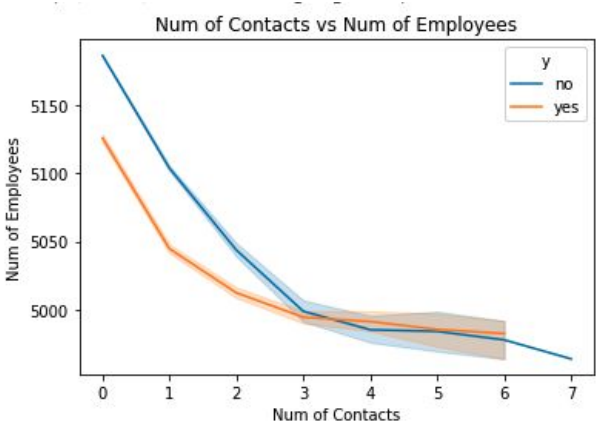


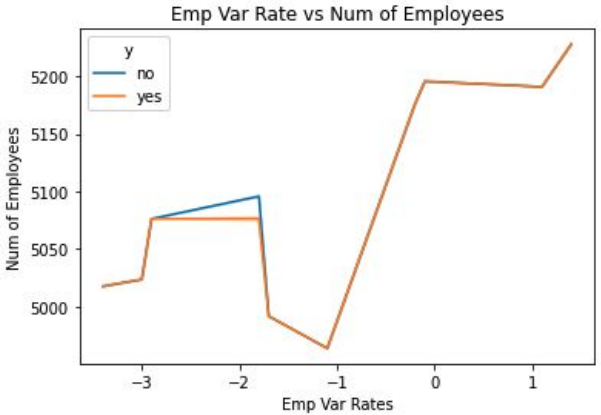


Pairplot: View The Relationships Between Numerical Variables

Bivariate Analysis





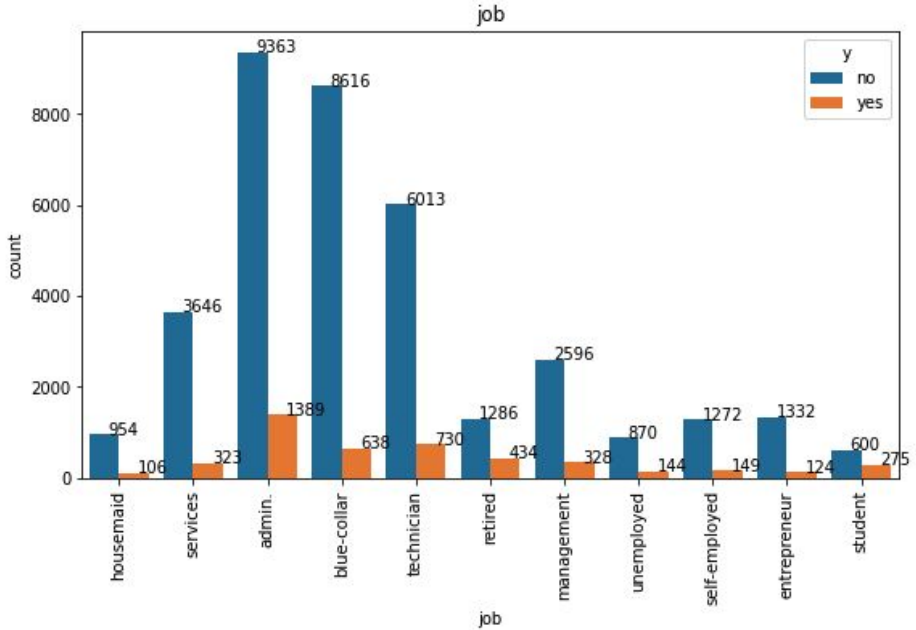


Keep a time gap between campaigns as 100-200 days, 3-5 times to contact and bigger the company size lesser employee job changes

Bivariate Analysis

Categorical Variables to Target

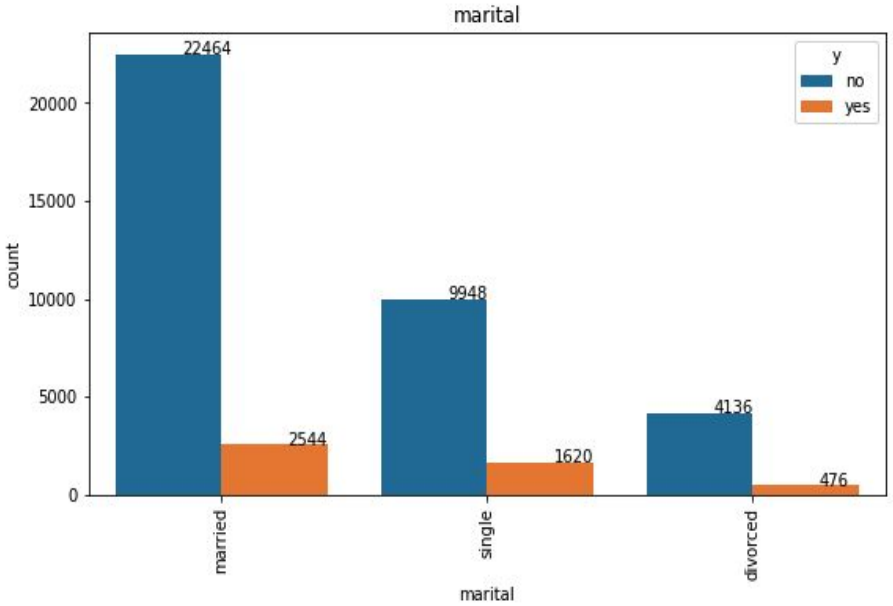
y	no	yes	accep_ratio
job			
student	600	275	0.31
retired	1286	434	0.25
unemployed	870	144	0.14
admin.	9363	1389	0.13
management	2596	328	0.11
technician	6013	730	0.11
housemaid	954	106	0.10
self-employed	1272	149	0.10
entrepreneur	1332	124	0.09
services	3646	323	0.08
blue-collar	8616	638	0.07



Bivariate Analysis

Categorical Variables to Target

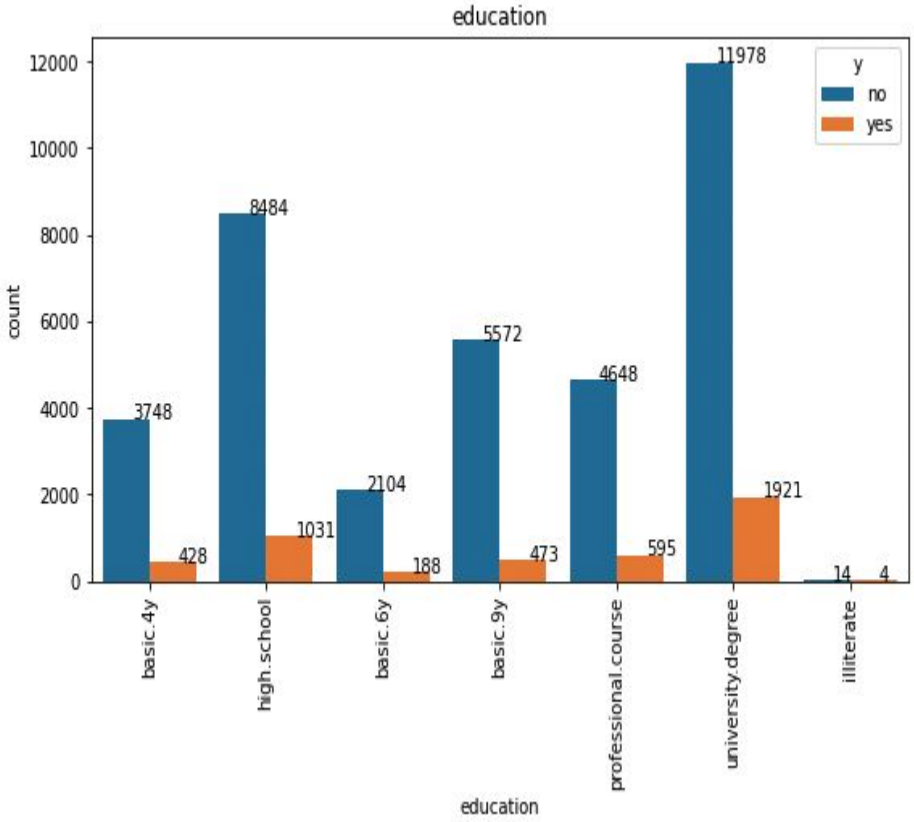
y	no	yes	accep_ratio
marital			
single	9948	1620	0.14
divorced	4136	476	0.10
married	22464	2544	0.10



Bivariate Analysis

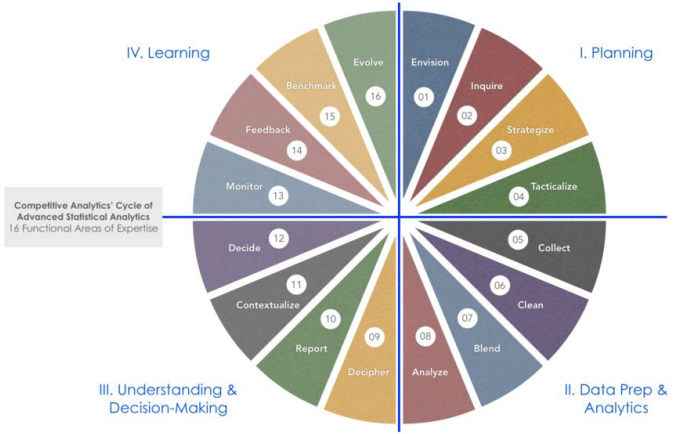
Categorical Variables to Target

y	no	yes	accep_ratio
education			
illiterate	14	4	0.22
university.degree	11978	1921	0.14
high.school	8484	1031	0.11
professional.course	4648	595	0.11
basic.4y	3748	428	0.10
basic.6y	2104	188	0.08
basic.9y	5572	473	0.08



Next Steps

- Advanced statistical methods
  - Regression on numerical values
- Machine learning prediction
  - Using supervised learning and unsupervised learning to predict whether people would subscribe



Thank you!