'data.frame': 31649 obs. of 18 variables:

$ serial\_number : int 1 2 3 4 5 6 7 8 9 10 ...

$ age\_in\_years : int 58 44 33 47 33 35 28 42 58 43 ...

$ job\_description : Factor w/ 12 levels "admin.","blue-collar",..: 5 10 3 2 12 5 5 3 6 10 ...

$ marital\_status : Factor w/ 3 levels "divorced","married",..: 2 3 2 2 3 2 3 1 2 3 ...

$ education\_details : Factor w/ 4 levels "primary","secondary",..: 3 2 2 4 4 3 3 3 1 2 ...

$ has\_default : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 2 1 1 ...

$ balance\_in\_account : int 2143 29 2 1506 1 231 447 2 121 593 ...

$ housing\_status : Factor w/ 2 levels "no","yes": 2 2 2 2 1 2 2 2 2 2 ...

$ previous\_loan : Factor w/ 2 levels "no","yes": 1 1 2 1 1 1 2 1 1 1 ...

$ phone\_type : Factor w/ 3 levels "cellular","telephone",..: 3 3 3 3 3 3 3 3 3 3 ...

$ date : int 5 5 5 5 5 5 5 5 5 5 ...

$ month\_of\_year : Factor w/ 11 levels "apr","aug","dec",..: 9 9 9 9 9 9 9 9 9 9 ...

$ call\_duration : int 261 151 76 92 198 139 217 380 50 55 ...

$ campaign\_contacts : int 1 1 1 1 1 1 1 1 1 1 ...

$ days\_passed : int -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...

$ previous\_contact : int 0 0 0 0 0 0 0 0 0 0 ...

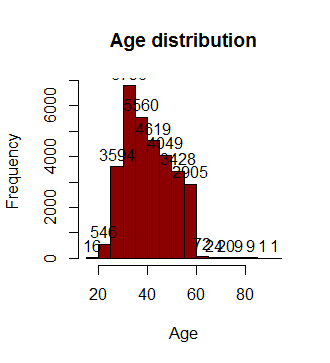
$ poutcome\_of\_campaign: Factor w/ 4 levels "failure","other",..: 4 4 4 4 4 4 4 4 4 4 ...

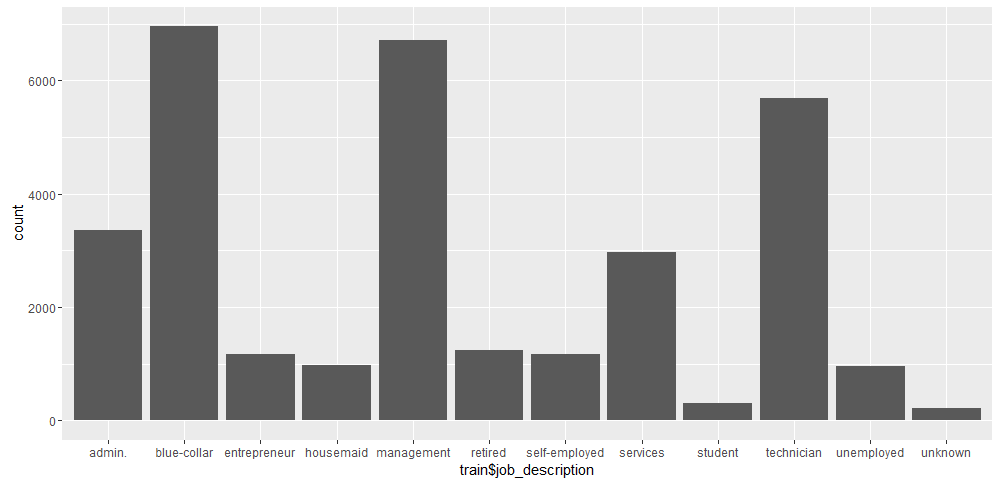
$ outcome : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...

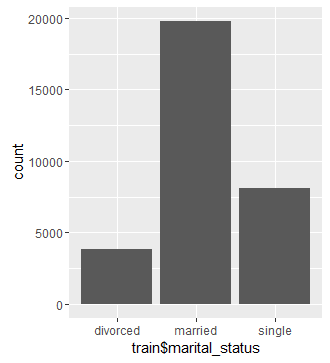
* Imbalanced data set – SMOTE needs to be done to balance.
* > # Check for data set imbalance
* > table(train$outcome)
* no yes
* 29809 1840

## Column Distribution & Missing values

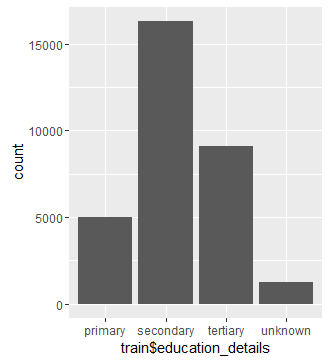
* Age
  + No Missing Values



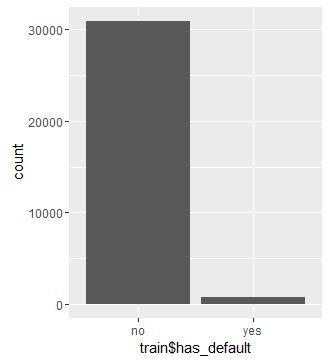
* + Job Description
    - 221 rows (~ 0.6%) with unknown values
* Marital Status
  + 0 missing/unknown values



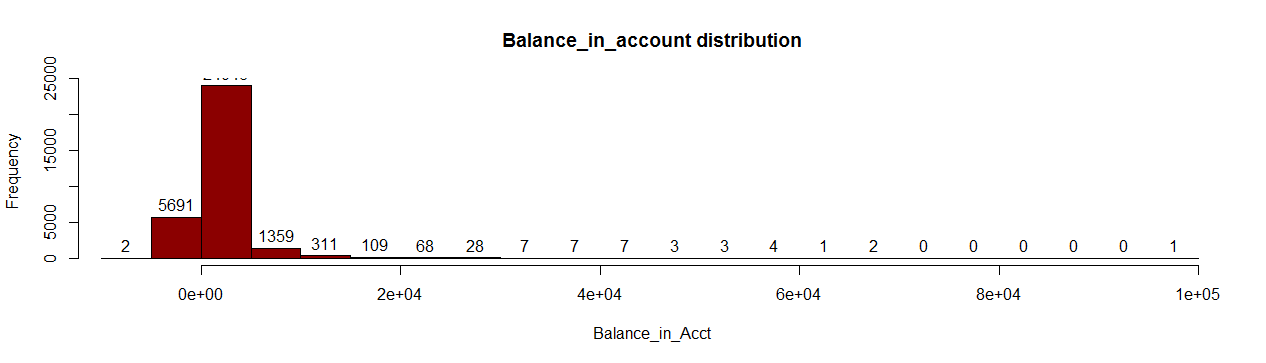
* Education Details
  + 1272 (~4%) unknown values



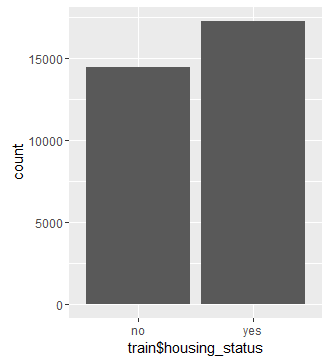
* Has Default
  + 0 missing/unknown values



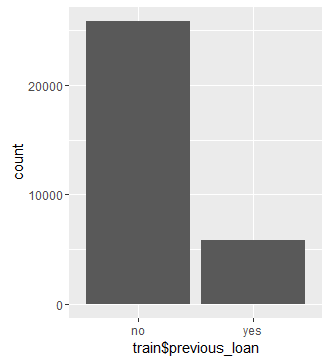
* Balance in account
  + 0 missing values, but with 2976 negative balance accounts/
  + Left Skewed



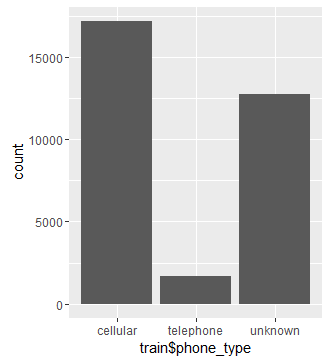
* Housing Status
  + 0 missing/unknown values



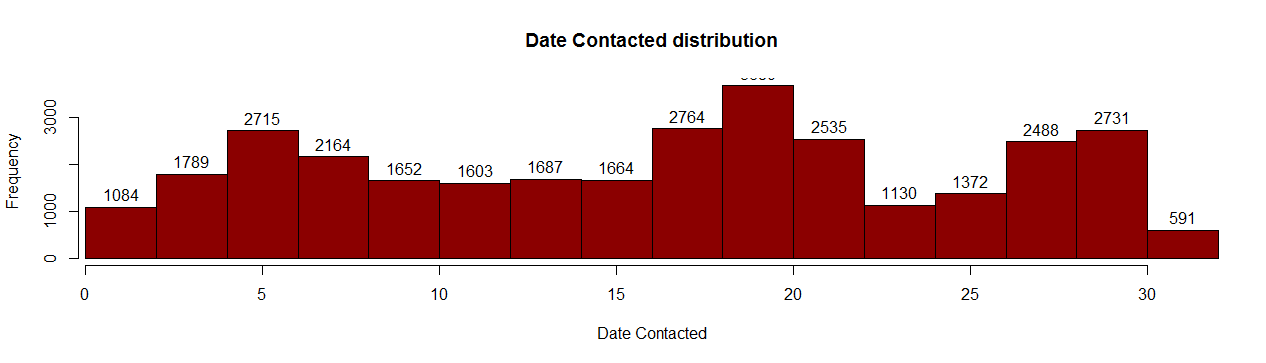
* Previous\_Loan
  + 0 missing/unknown values



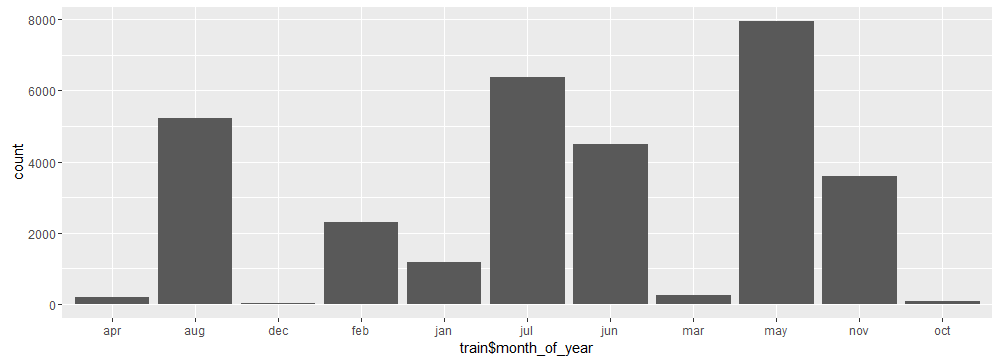
* Phone\_Type
  + 12765 unknown values



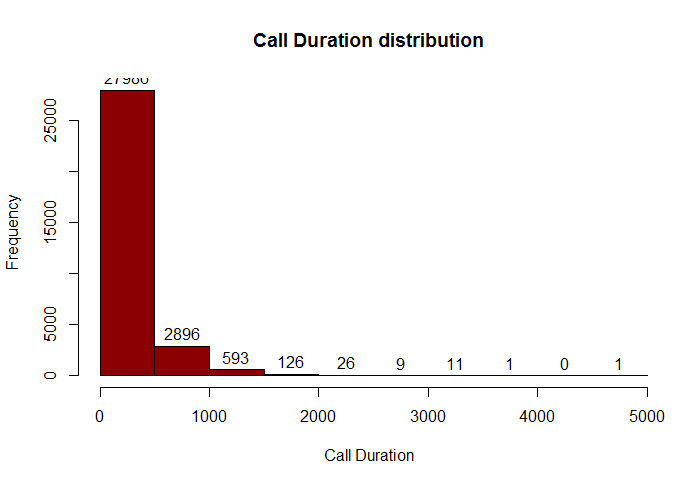
* Date Contacted
  + 0 missing values



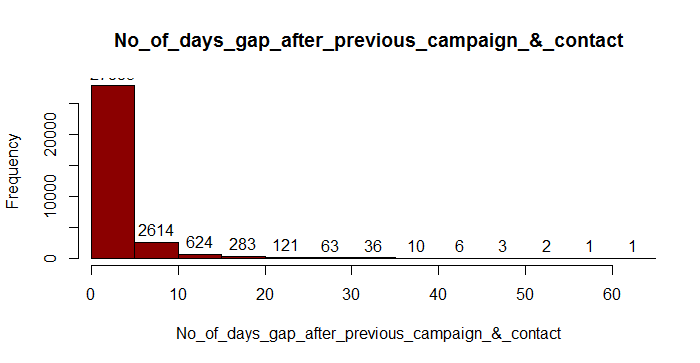
* Month\_of\_year
  + 0 missing values
  + More contacts during Jan, Feb, May, Jun, Jul, Aug, Nov
  + No contacts in Sep at all.



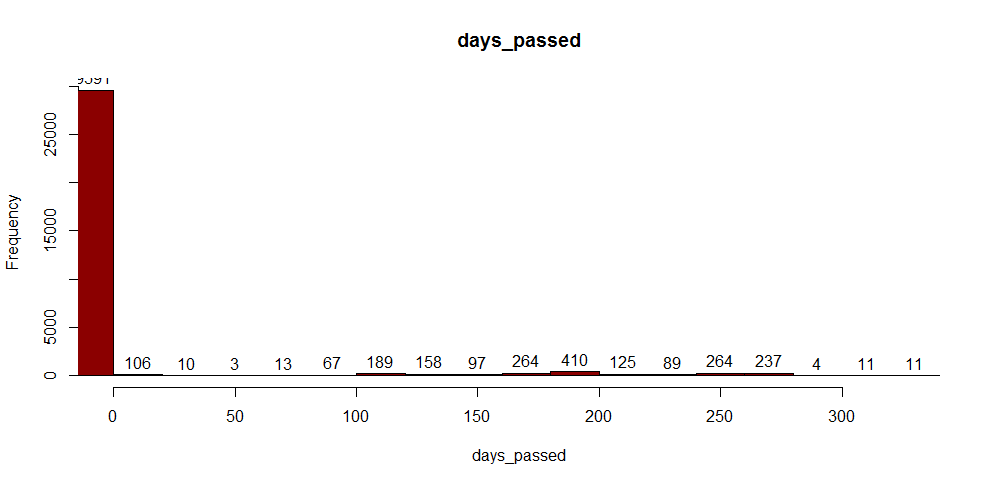
* Call\_Duration
  + 2 records with 0 call duration



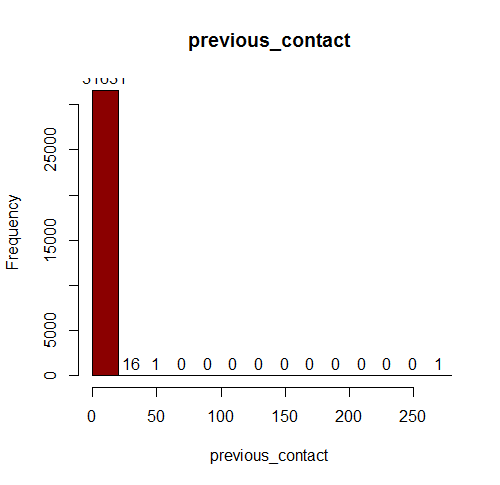
* Campaign contacts
  + 0 missing values

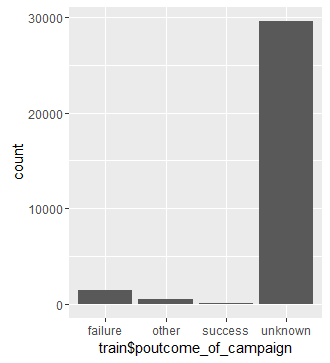


* Days\_Passed
  + Many records with -1



* Previous\_campaign\_contact\_count



* Previous\_campaign\_outcome
  + 29591(~93%) unknown values
  + 538 with other value. It should be either success or failure.

## Handling Missing values

* Job Descrition
  + Verify records with unknown values for corresponding bank balance, housing loan status.
  + Replacing with blue collar
* Education Details
  + 50% of records with unknown education details have previous housing loan.
  + We can assume that they had good education qualification
  + Hence replacing unknown with tertiary.