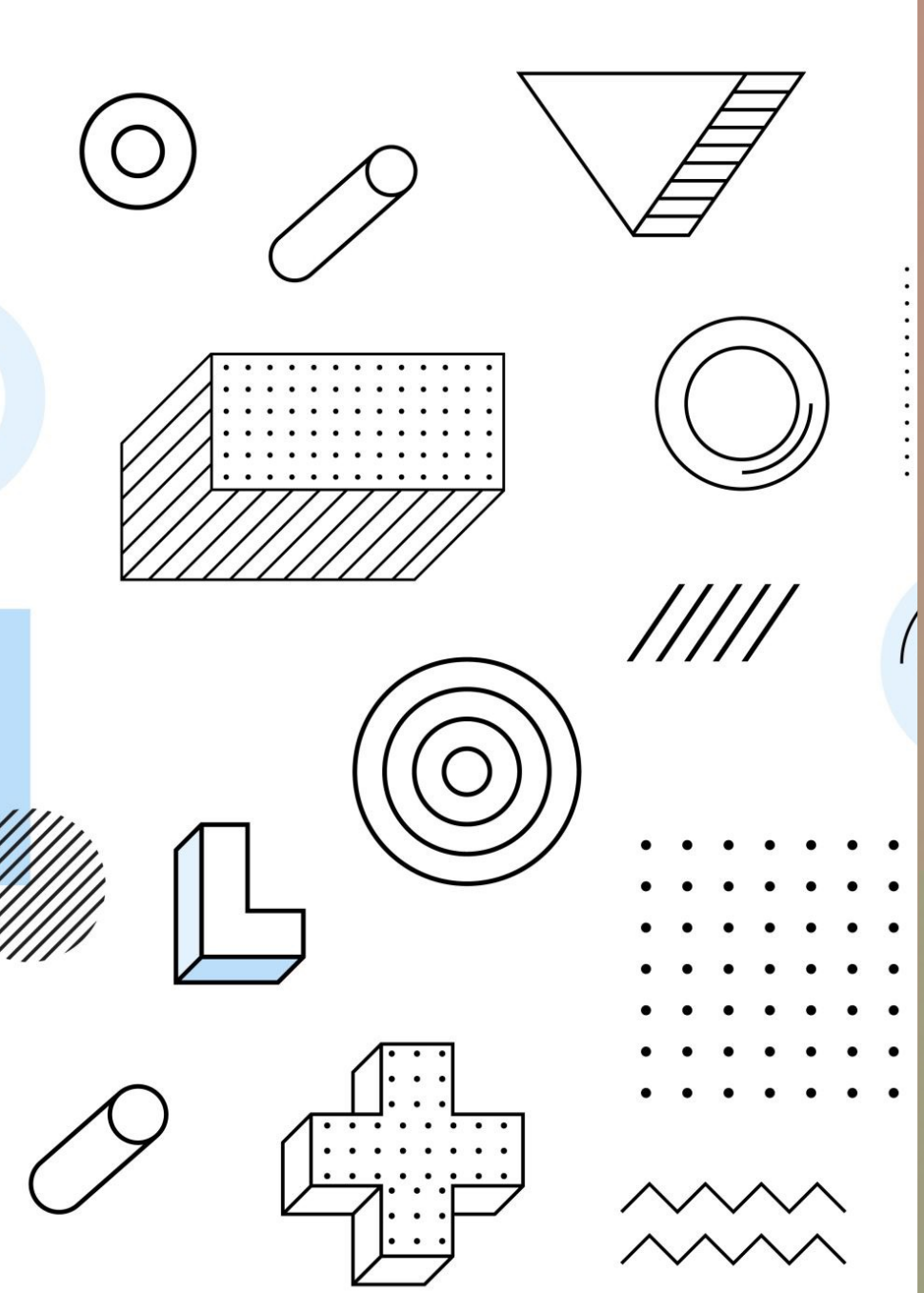


ANALYSIS OF BANK TELEMARKETING CAMPAIGN

STATISTICAL LEARNING MOD.B

BRENDA TELLEZ

DASARAJU ABHISHEK VARMA

MOHAMMAD HUZAIFA FAZAL

CONTENTS

INTRODUCTION

DATA COLLECTION & DESCRIPTION

DATA MANIPULATION

DATA CLEANING

EXPLORATORY DATA ANALYSIS

MODELS

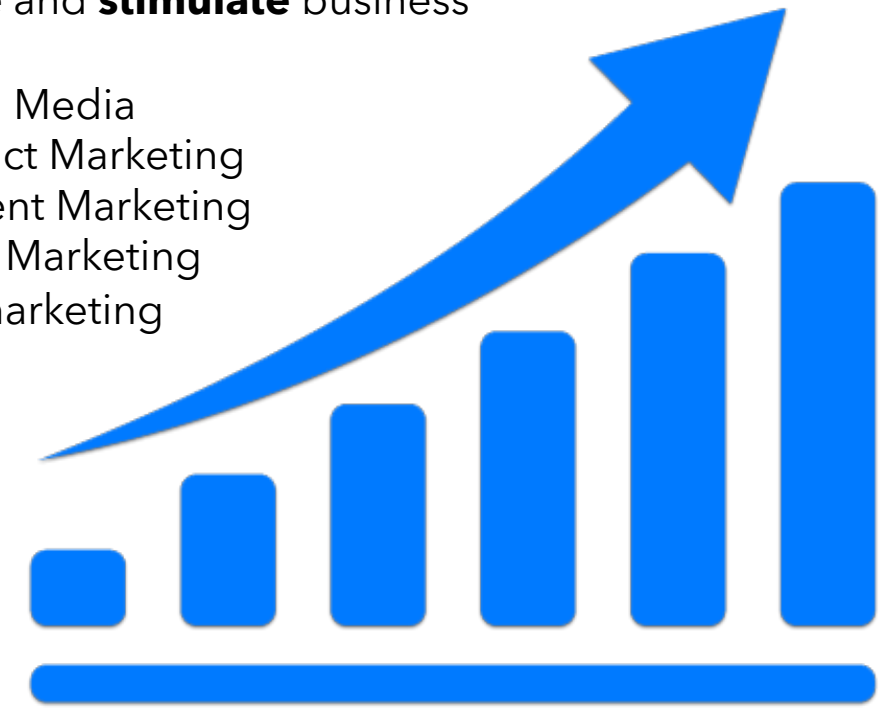
RESULTS

CONCLUSION

INTRODUCTION

Marketing campaigns are used to **enhance** and **stimulate** business growth.

- Social Media
- Product Marketing
- Content Marketing
- Email Marketing
- Telemarketing



- Call duration that yields the most positive results.
- Which day week and month increases odds of success?
- Call on cellphone or telephone line?
- Does the job or education significantly affect their decision?

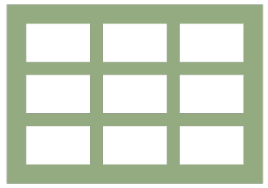
DATA COLLECTION AND DESCRIPTION

- Sourced from a Portuguese retail bank
- The dataset contains features related to direct marketing campaigns for the purpose of selling bank long-term deposits.
- We obtained the dataset from the UC Irvine Machine Learning Repository

Input variables:
bank client data:
1 - age (numeric)
2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
5 - default: has credit in default? (categorical: 'no','yes','unknown')
6 - housing: has housing loan? (categorical: 'no','yes','unknown')
7 - loan: has personal loan? (categorical: 'no','yes','unknown')
related with the last contact of the current campaign:
8 - contact: contact communication type (categorical: 'cellular','telephone')
9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
other attributes:
12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
14 - previous: number of contacts performed before this campaign and for this client (numeric)
15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
social and economic context attributes
16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)
17 - cons.price.idx: consumer price index - monthly indicator (numeric)
18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)
19 - euribor3m: euribor 3 month rate - daily indicator (numeric)
20 - nr.employed: number of employees - quarterly indicator (numeric)

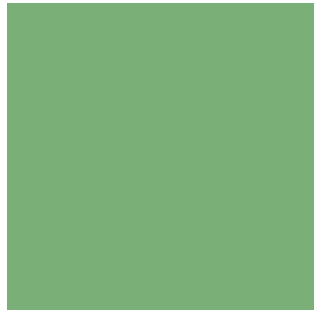
Output variable (desired target):
21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

DATASET DESCRIPTION



Contains 41,188 instances (rows)
with 21 features (columns)

Out of the 21 features, 20 are used
as potential predicting factors



10 Numerical columns

11 Categorical:

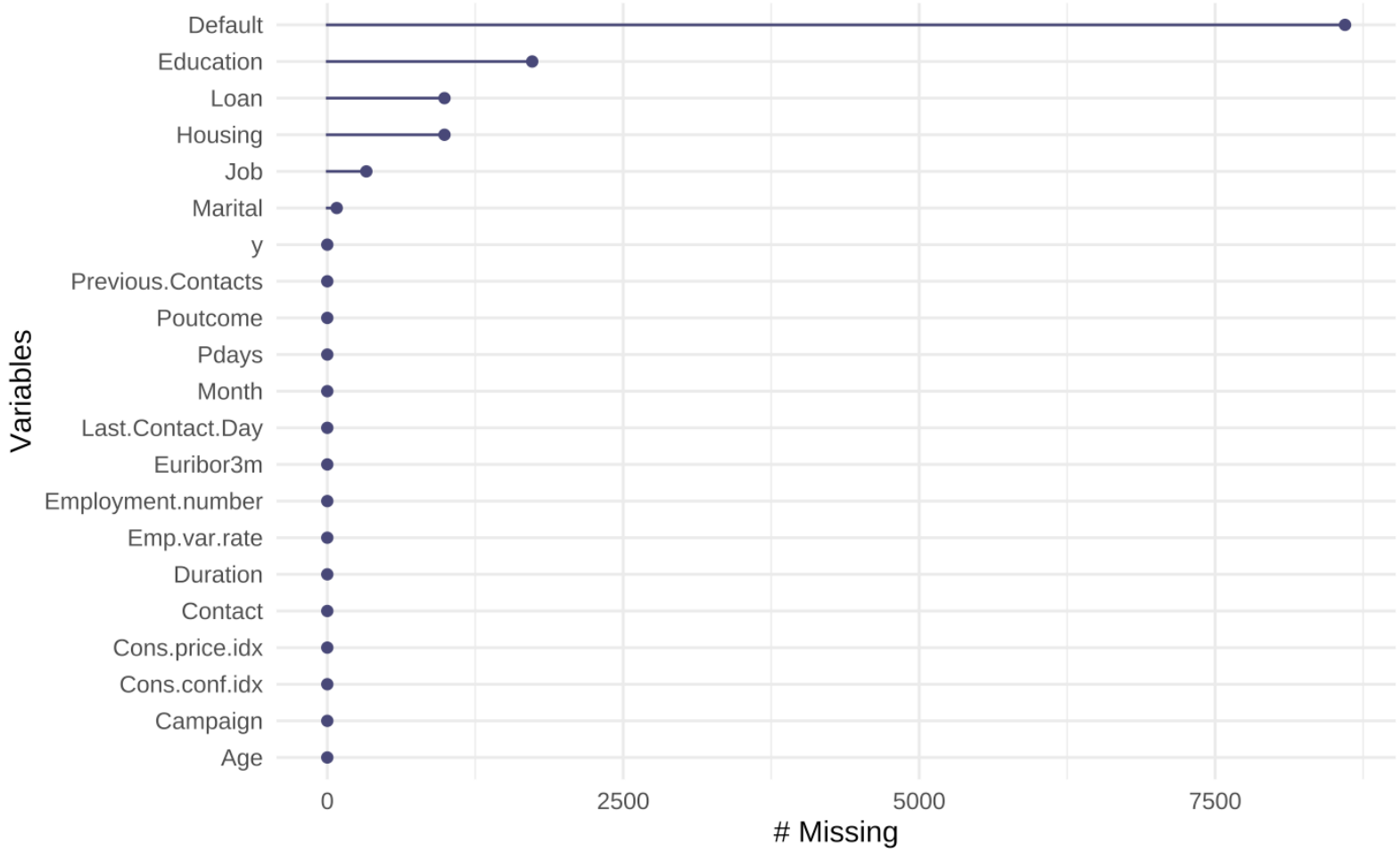
DATA MANIPULATION

```
'data.frame':   41188 obs. of  21 variables:
 $ age           : int  56 57 37 40 56 45 59 41 24 25 ...
 $ job           : chr  "housemaid" "services" "services" "admin." ...
 $ marital       : chr  "married" "married" "married" "married" ...
 $ education     : chr  "basic.4y" "high.school" "high.school" "basic.6y" ...
 $ default       : chr  "no" NA "no" "no" ...
 $ housing       : chr  "no" "no" "yes" "no" ...
 $ loan          : chr  "no" "no" "no" "no" ...
 $ contact       : chr  "telephone" "telephone" "telephone" "telephone" ...
 $ month         : chr  "may" "may" "may" "may" ...
 $ day_of_week   : chr  "mon" "mon" "mon" "mon" ...
 $ duration      : int  261 149 226 151 307 198 139 217 380 50 ...
 $ campaign      : int  1 1 1 1 1 1 1 1 1 1 ...
 $ pdays         : int  999 999 999 999 999 999 999 999 999 999 ...
 $ previous      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ poutcome      : chr  "nonexistent" "nonexistent" "nonexistent" "nonexistent" ...
 $ emp.var.rate  : num  1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 ...
 $ cons.price.idx: num  94 94 94 94 94 ...
 $ cons.conf.idx : num  -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 ...
 $ euribor3m     : num  4.86 4.86 4.86 4.86 4.86 ...
 $ nr.employed   : num  5191 5191 5191 5191 5191 ...
 $ y             : chr  "no" "no" "no" "no" ...
```

- Renamed Column names to standardize
- Removed the dot in ”admin.” in the job column

DATA CLEANING

- Check for NA values
  - Default: 20.87% of NA values, highly imbalanced(Dropped)
- 2 methods of handling NA values
  - Considered as 3rd category “unknown”
  - Deletion or Imputation



CONTINGENCY TABLE(JOB & EDUCATION)

	basic.4y	basic.6y	basic.9y	high.school	illiterate	professional.course	university.degree
administration	77	151	499	3329	1	363	5753
blue-collar	2318	1426	3623	878	8	453	94
entrepreneur	137	71	210	234	2	135	610
housemaid	474	77	94	174	1	59	139
management	100	85	166	298	0	89	2063
retired	597	75	145	276	3	241	285
self-employed	93	25	220	118	3	168	765
services	132	226	388	2682	0	218	173
student	26	13	99	357	0	43	170
technician	58	87	384	873	0	3320	1809
unemployed	112	34	186	259	0	142	262
<NA>	52	22	31	37	0	12	45

	<NA>
administration	249
blue-collar	454
entrepreneur	57
housemaid	42
management	123
retired	98
self-employed	29
services	150
student	167
technician	212
unemployed	19
<NA>	131

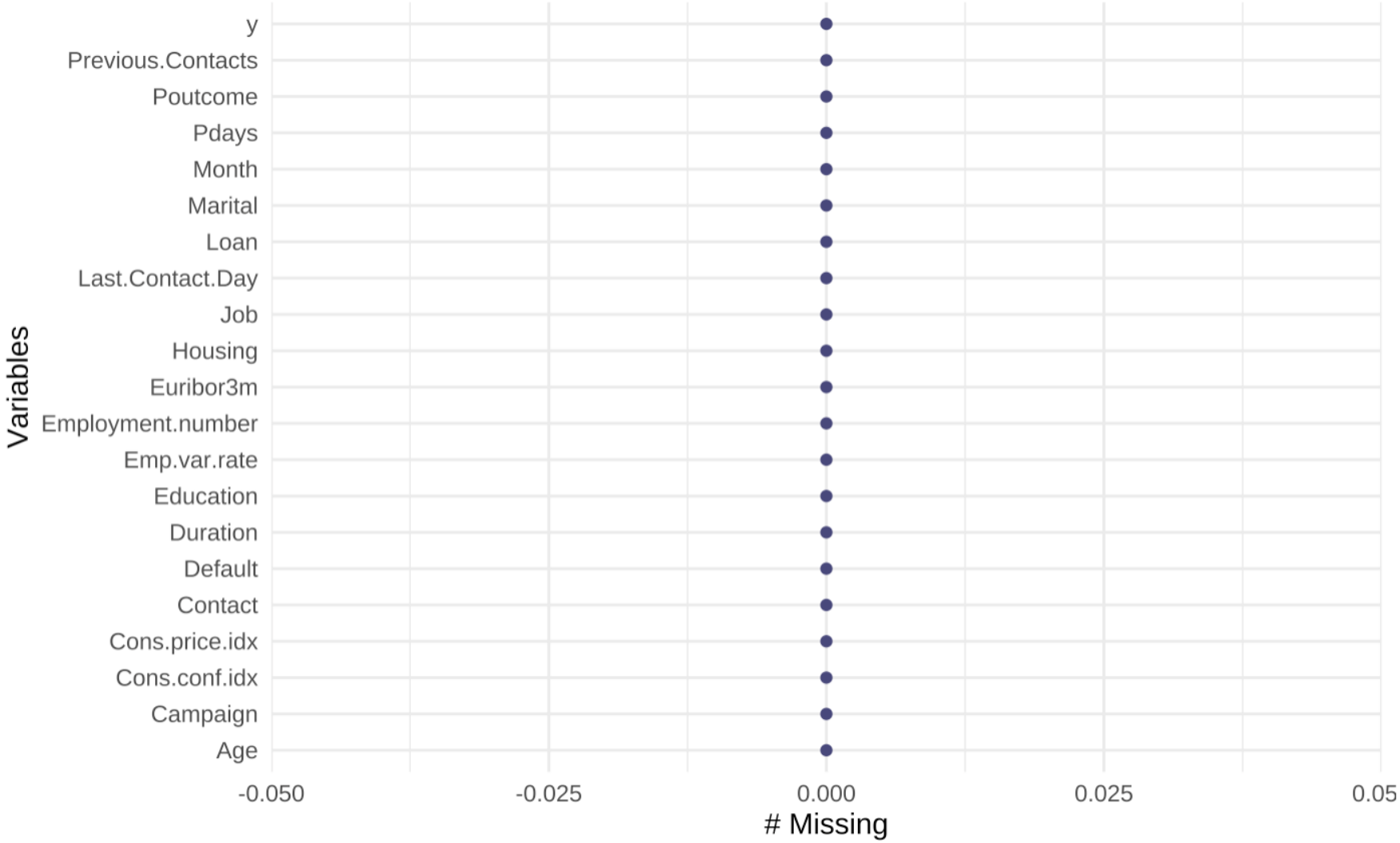
NA VALUES FOR HOUSING, LOAN & DEFAULT

2nd method of dealing with NA values

- For **Housing, Loan, and Default** columns the NA values were replaced back to "unknown" values as they will be considered as a category within their respective columns, as discussed before.

- **Marital** column NA values – No logical inference possible, all rows containing NA values removed
- Rows where both Education & Job columns had NA values – entire rows removed

NO NA VALUES


y
Previous.Contacts
Poutcome
Pdays
Month
Marital
Loan
Last.Contact.Day
Job
Housing
Euribor3m
Employment.number
Emp.var.rate
Education
Duration
Default
Contact
Cons.price.idx
Cons.conf.idx
Campaign
Age
Variables
-0.050
-0.025
0.000
0.025
0.05
Missing

DATA TRANSFORMATION

- Categorical columns were converted to factors:
  - Ordinal: Education, Month, and day_of_week
  - Nominal: Job, marital, housing, loan, etc
  - Later on realized, not much benefit in ordinal columns so considered all categorical columns to nominal (One-hot encoding)
- Feature Scaling: MinMax scaling to normalize range

Categorical columns changed to factors

```
'data.frame':   40990 obs. of  21 variables:
 $ Age              : int  56 57 37 40 56 45 59 41 24 25 ...
 $ Job              : Factor w/ 11 levels "administration",..: 4 8 8 1 8 8 1 2 10 8 ...
 $ Marital          : Factor w/ 3 levels "divorced","married",..: 2 2 2 2 2 2 2 2 3 3 ...
 $ Education        : Factor w/ 7 levels "basic.4y","basic.6y",..: 1 4 4 2 4 3 6 3 6 4 ...
 $ Default          : Factor w/ 3 levels "no","unknown",..: 1 2 1 1 1 2 1 2 1 1 ...
 $ Housing          : Factor w/ 3 levels "no","unknown",..: 1 1 3 1 1 1 1 1 3 3 ...
 $ Loan             : Factor w/ 3 levels "no","unknown",..: 1 1 1 1 3 1 1 1 1 1 ...
 $ Contact          : Factor w/ 2 levels "cellular","telephone": 2 2 2 2 2 2 2 2 2 2 ...
 $ Month            : Factor w/ 10 levels "apr","aug","dec",..: 7 7 7 7 7 7 7 7 7 7 ...
 $ Last.Contact.Day : Factor w/ 5 levels "fri","mon","thu",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ Duration         : int  261 149 226 151 307 198 139 217 380 50 ...
 $ Campaign         : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Pdays            : int  999 999 999 999 999 999 999 999 999 999 ...
 $ Previous.Contacts: int  0 0 0 0 0 0 0 0 0 0 ...
 $ Poutcome         : Factor w/ 3 levels "failure","nonexistent",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ Emp.var.rate     : num  1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 ...
 $ Cons.price.idx   : num  94 94 94 94 94 ...
 $ Cons.conf.idx    : num  -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 ...
 $ Euribor3m        : num  4.86 4.86 4.86 4.86 4.86 ...
 $ Employment.number: num  5191 5191 5191 5191 5191 ...
 $ y                : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
 - attr(*, "na.action")= 'omit' Named int [1:198] 41 74 92 300 304 344 389 391 414 429 ...
  ..- attr(*, "names")= chr [1:198] "41" "74" "92" "300" ...
```

EXPLORATORY DATA ANALYSIS (EDA)

- Detailed statistics of Numerical columns

A tibble: 10 × 26

described_variables <chr>	n <int>	na <int>	mean <dbl>	sd <dbl>	se_mean <dbl>	IQR <dbl>	skewness <dbl>
Age	40990	0	39.99958526	10.4173059	0.051453682	15.000	0.7895061
Duration	40990	0	258.32688461	259.3483954	1.280986657	217.000	3.2660855
Campaign	40990	0	2.56528422	2.7661321	0.013662619	2.000	4.7796101
Pdays	40990	0	962.56525982	186.6903937	0.922110596	0.000	-4.9290023
Previous.Contacts	40990	0	0.17294462	0.4948439	0.002444158	0.000	3.8347775
Emp.var.rate	40990	0	0.08045621	1.5710700	0.007759908	3.200	-0.7219900
Cons.price.idx	40990	0	93.57523681	0.5787226	0.002858456	0.919	-0.2293914
Cons.conf.idx	40990	0	-40.50761893	4.6280751	0.022859222	6.300	0.3046367
Euribor3m	40990	0	3.61968524	1.7347790	0.008568508	3.617	-0.7072457
Employment.number	40990	0	5166.98231032	72.2687571	0.356953485	129.000	-1.0428245

EDA - AVERAGES

- Pdays mean = 962.6, median = 999, max = 999
  - Due to the way data was recorded
  - In the Pdays column, if a client was not contacted after a previous campaign then it is recorded as 999.
  - So, why not record it as “0”
  - 0 signifies zero days since the previous contact (on the day of recording the data)
  - Mean & Median tells us vast majority of clients were not contacted since the previous campaign.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   3.000   6.000   6.008   7.000  27.000
```

Averages without the 999 entries

EDA - AVERAGES

- **Age** and **Campaign** have reasonable mean and median values
  - Age: mean = 39.9 years, median = 38 years
  - Campaign: mean = 2.56 days, median = 2 days

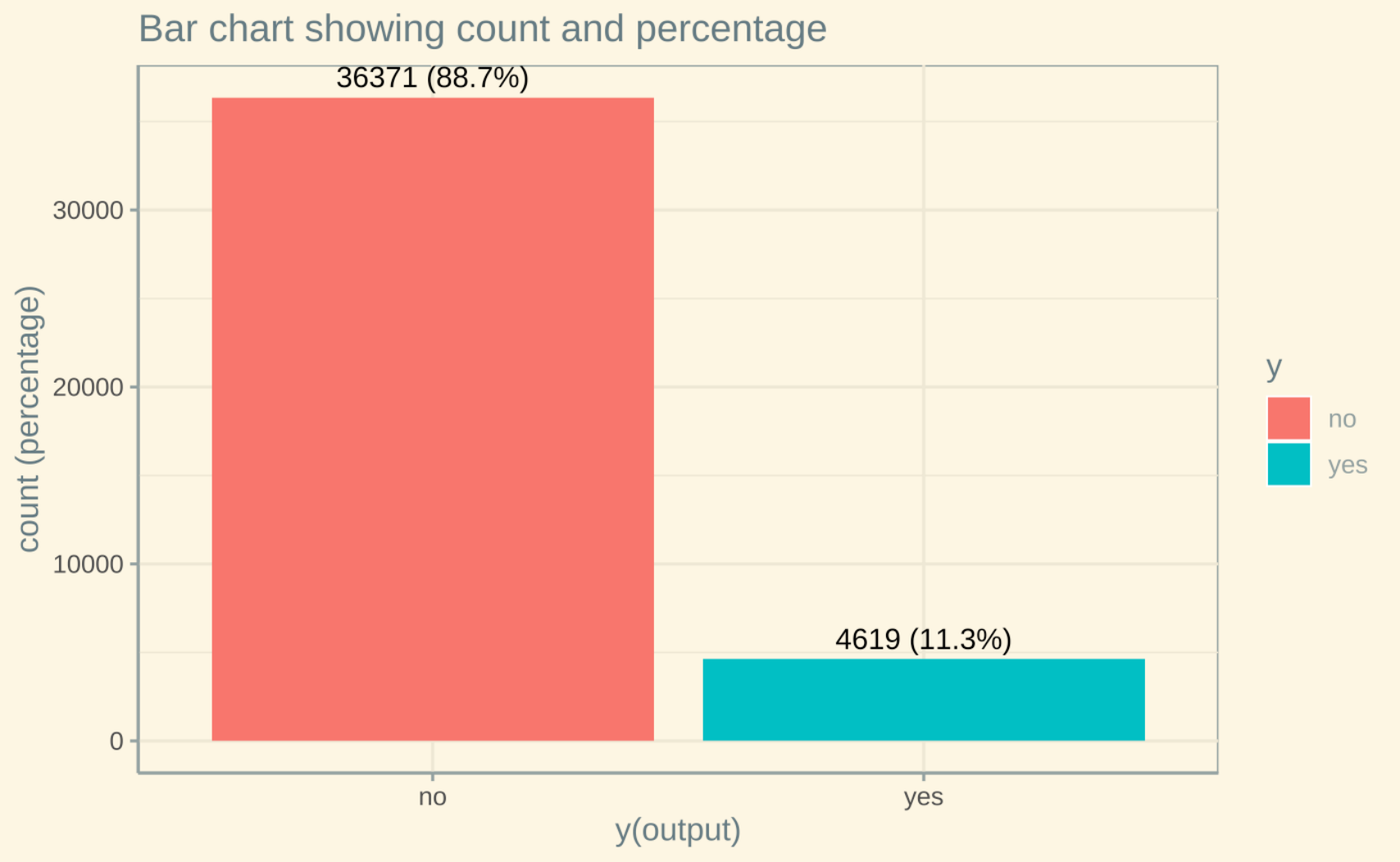
- '**Previous**: Mean = 0.1729, median = 0
  - suggests majority of clients were not contacted previously.
  - indicates bank is mostly focused on targeting new customers with their campaigns (as no previous contacts have been made)
  - or only recently started contacting customers for telemarketing purposes.

EDA

A tibble: 11 × 5

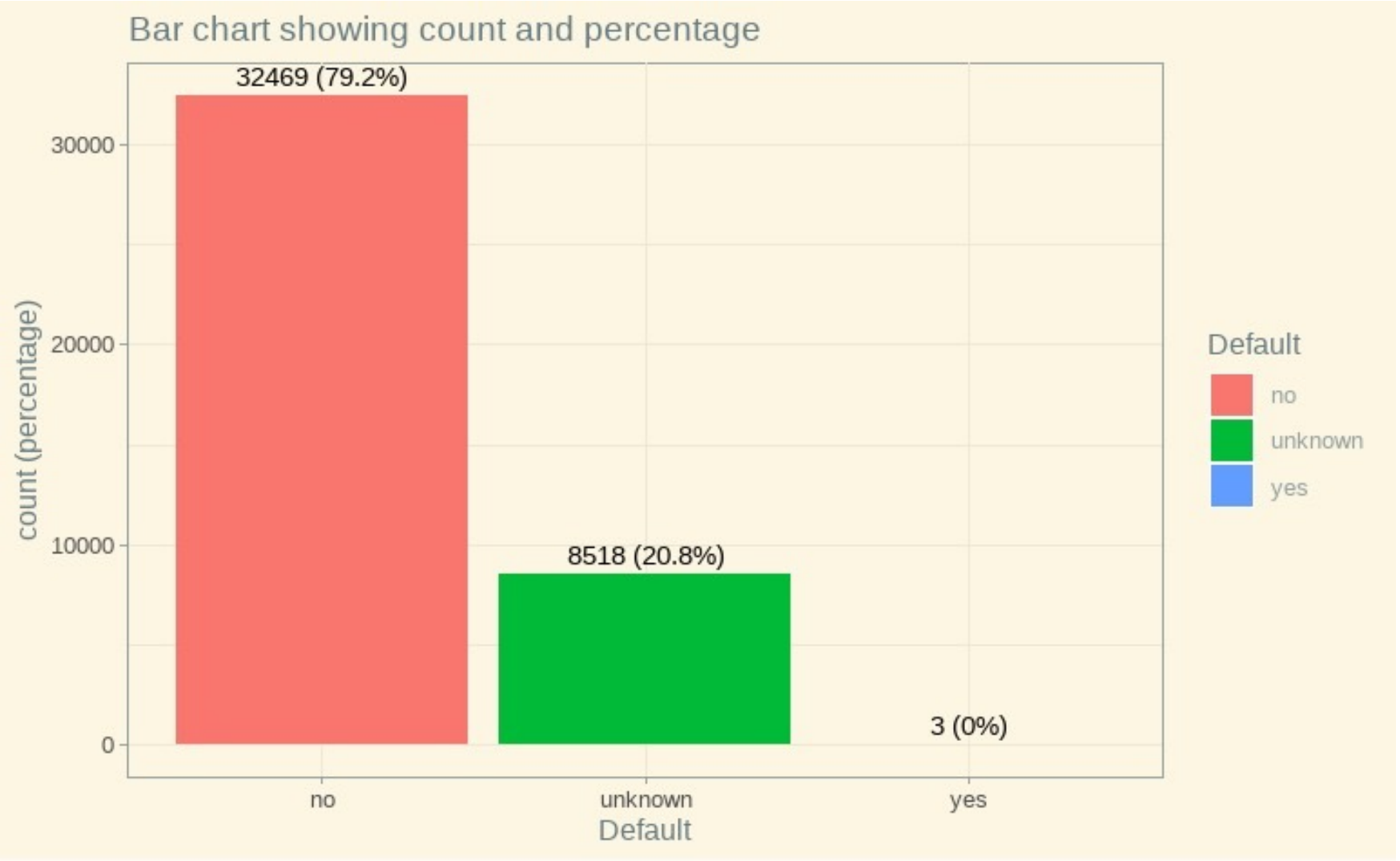
col_name <chr>	cnt <int>	common <chr>	common_pcnt <dbl>
Contact	2	cellular	63.50817
Default	3	no	79.21200
Education	7	university.degree	30.77336
Housing	3	yes	52.39571
Job	11	administration	25.57941
Last.Contact.Day	5	thu	20.90754
Loan	3	no	82.41522
Marital	3	married	60.56111
Month	10	may	33.42035
Poutcome	3	nonexistent	86.34057

EDA – UNIVARIATE ANALYSIS

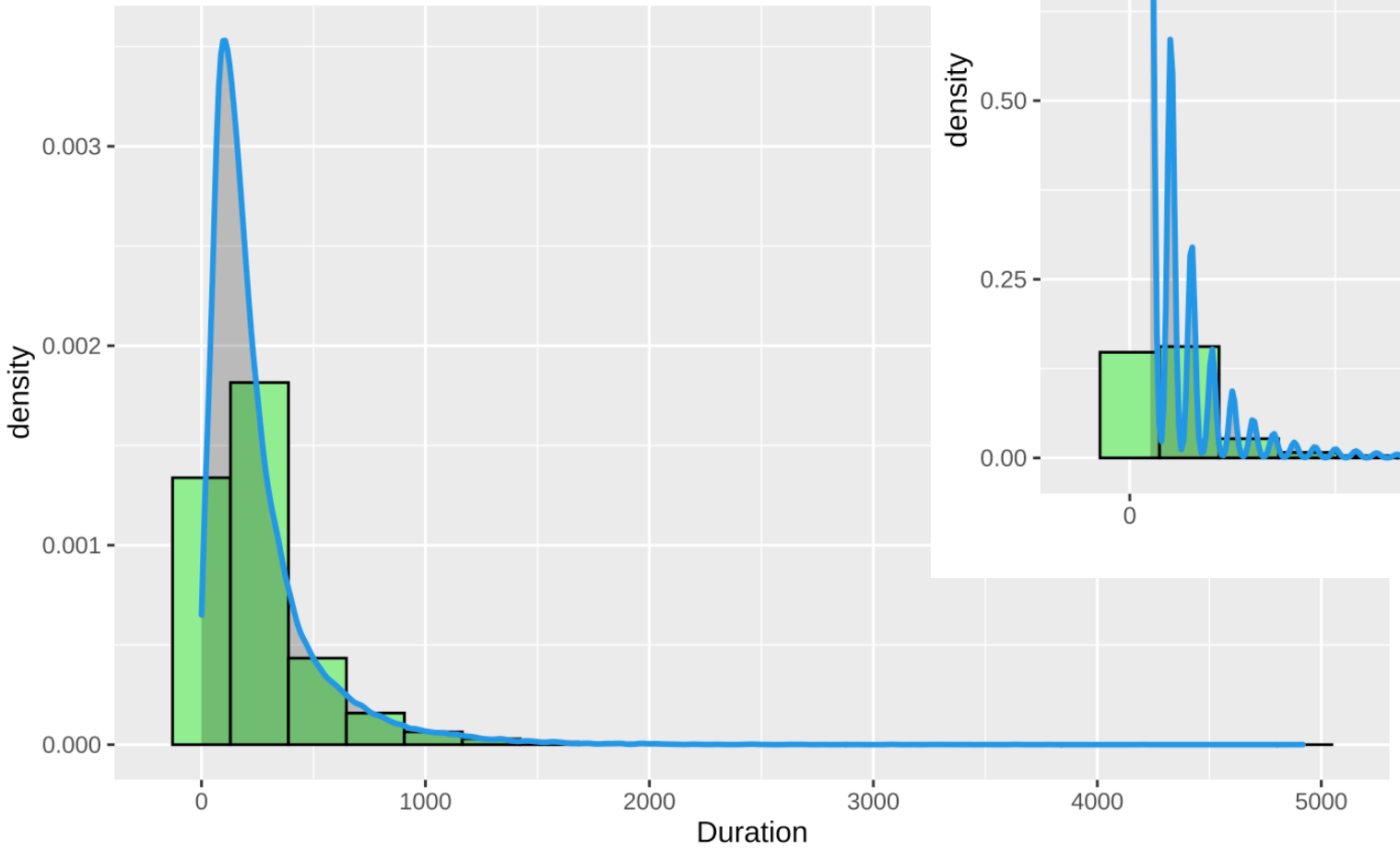


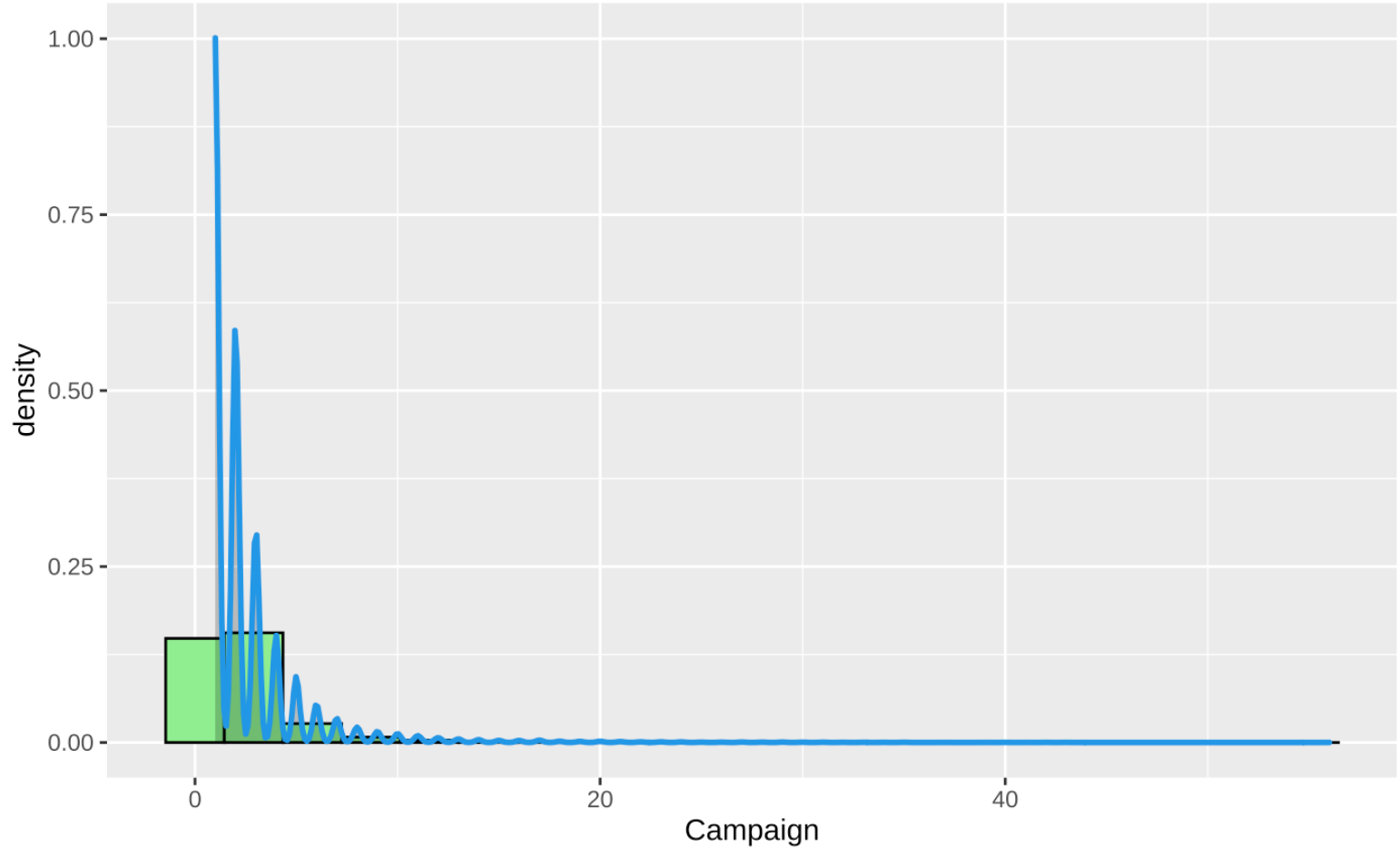
- Y-output highly **imbalanced**
- **36371 'No'** values and **4619 'Yes'**
- Imbalance ratio of 8:1 with 88.7% of the responses refusing the long-term deposit.

- Highly Imbalanced
- Removed Default column

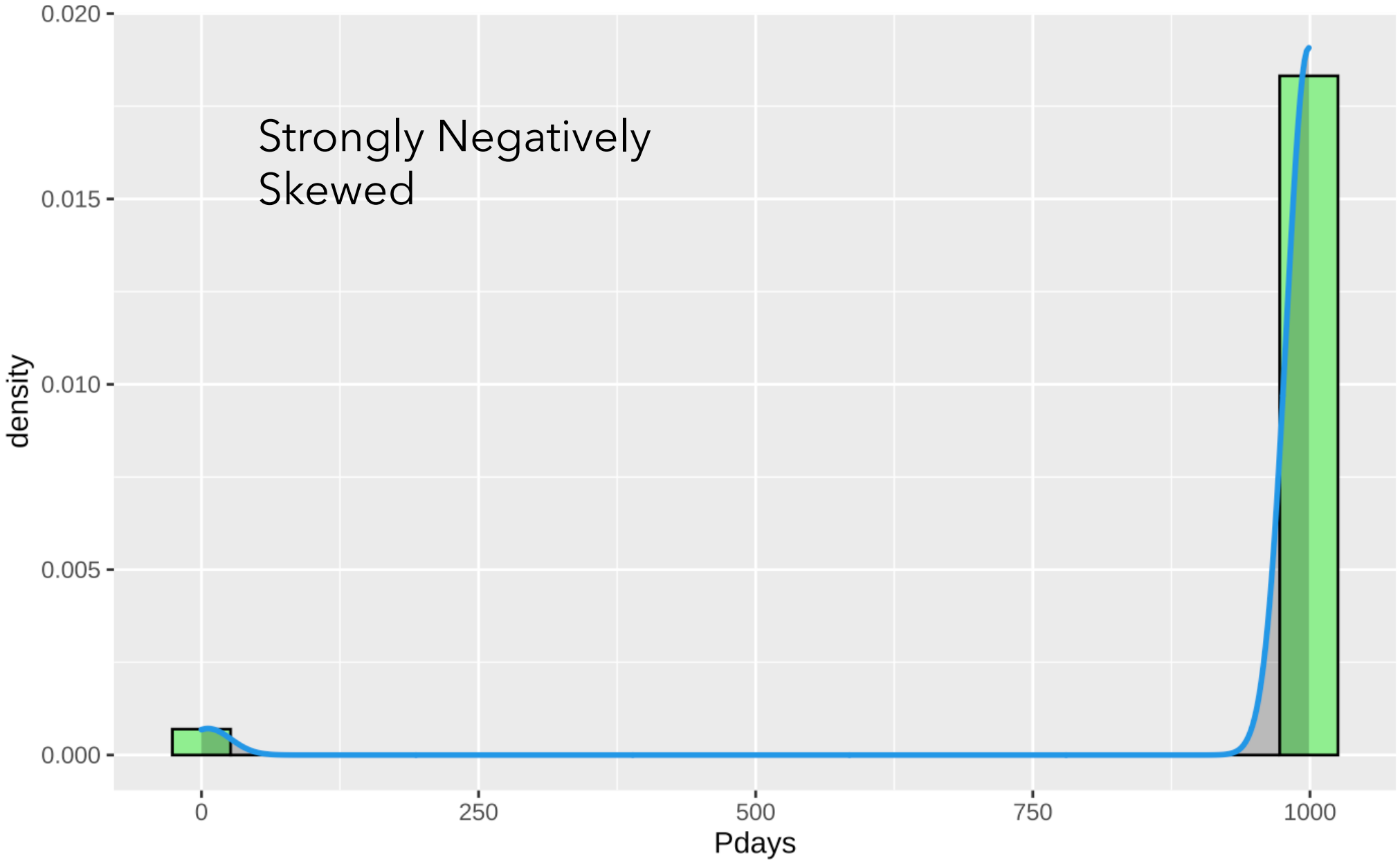


EDA – HISTOGRAMS WITH DENSITY PLOTS





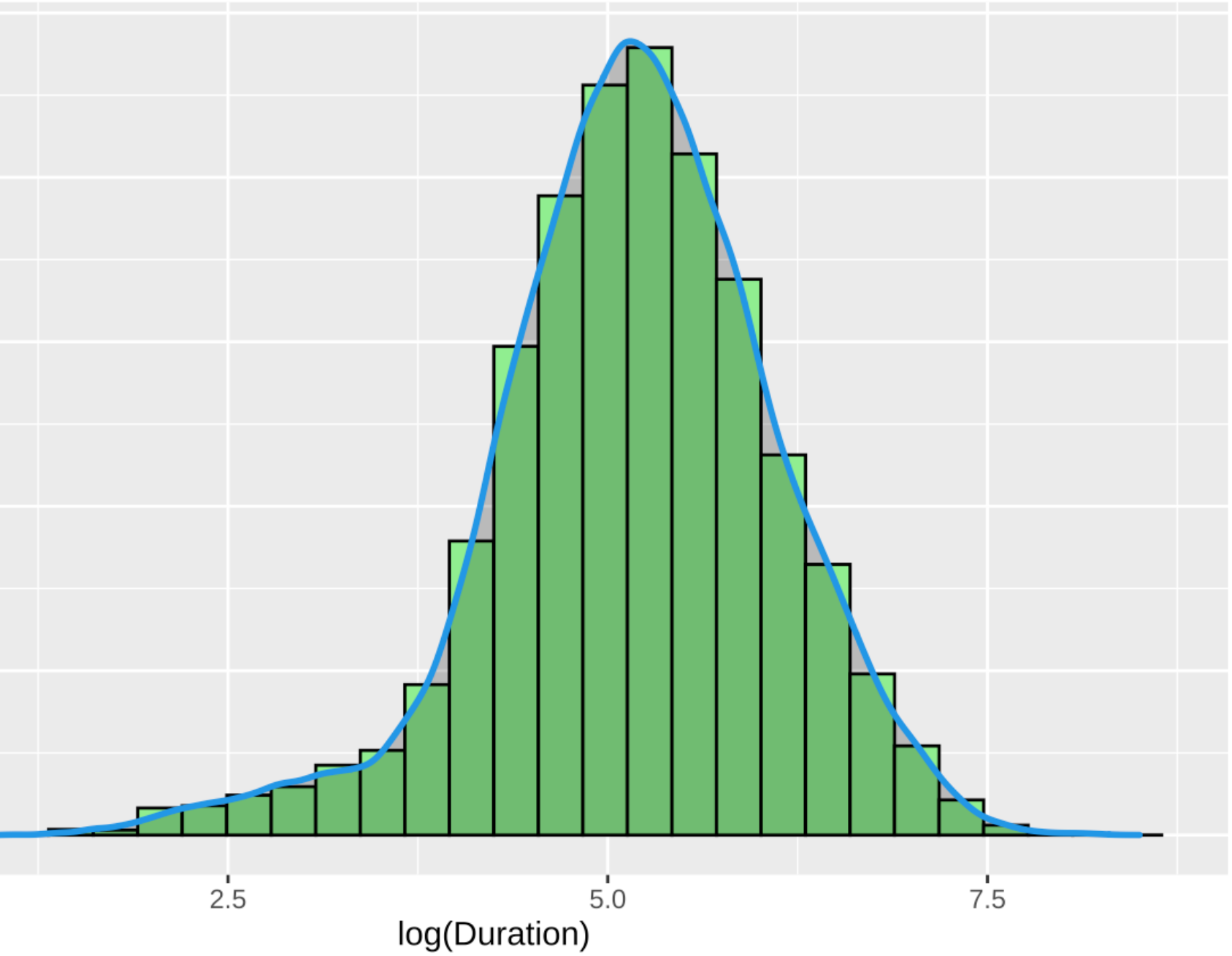
Strongly Positively skewed


Strongly Negatively
Skewed
0.020
0.015
0.010
0.005
0.000
density
0
250
500
750
1000
Pdays

Description: df [50 × 21]

Month <fctr>	Last.Contact.Day <fctr>	Duration <int>	Campaign <int>	Pdays <int>	Previous.Contacts <int>	Poutcome <fctr>
nov	mon	4918	1	999	0	nonexistent
aug	thu	4199	3	999	0	nonexistent
aug	fri	3785	1	999	0	nonexistent
jul	thu	3643	1	999	0	nonexistent
may	fri	3631	2	999	0	nonexistent
may	tue	3509	2	3	2	success
aug	thu	3422	1	999	0	nonexistent
may	tue	3366	3	999	0	nonexistent
aug	thu	3322	1	999	0	nonexistent
oct	mon	3284	1	999	0	nonexistent

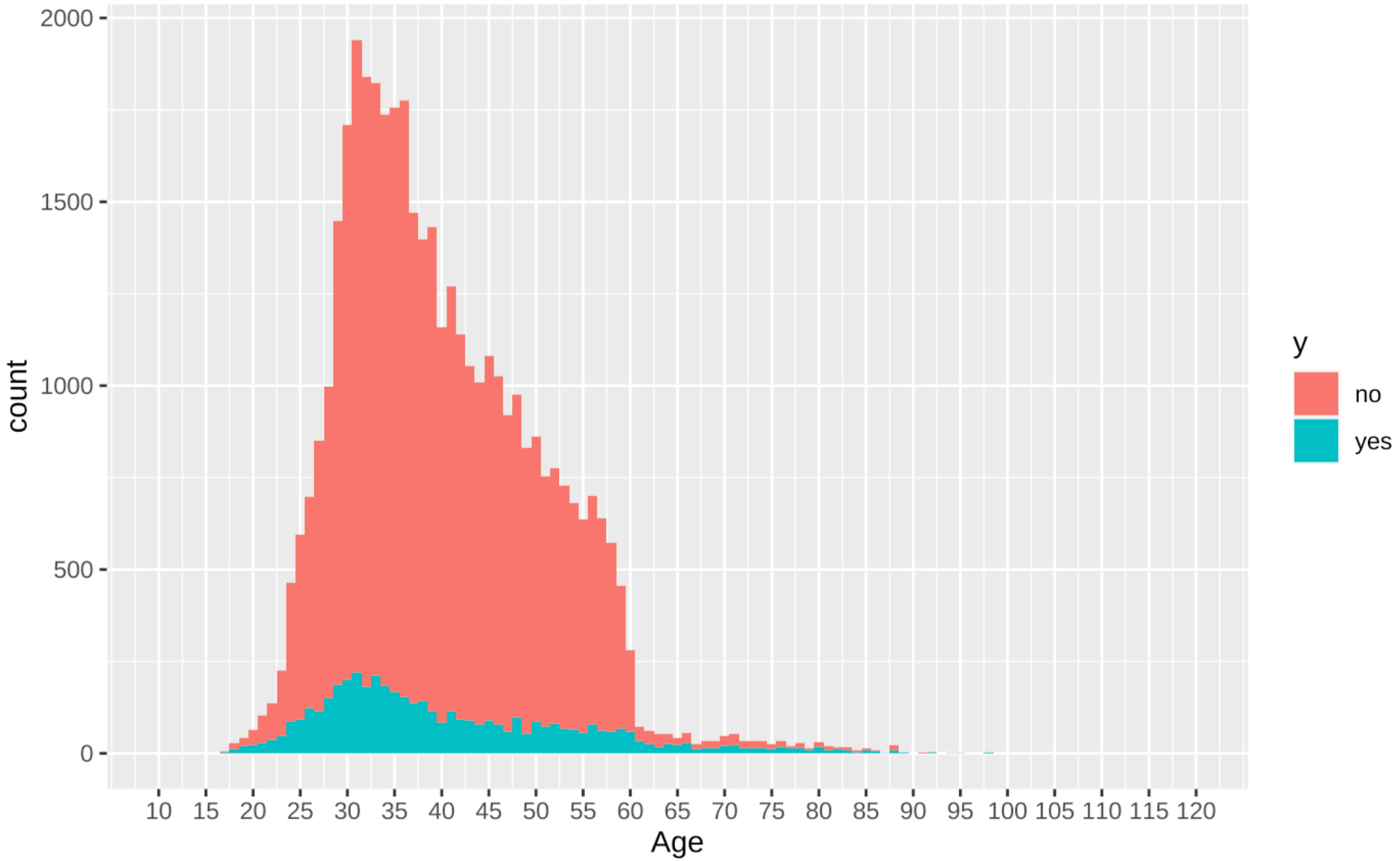
- Whenever duration is on the longer end (higher value), then the client has not been contacted prior to this campaign as well (pdays=999) and thus poutcome is "nonexistent".’
- Clients not contacted before require more time (duration)

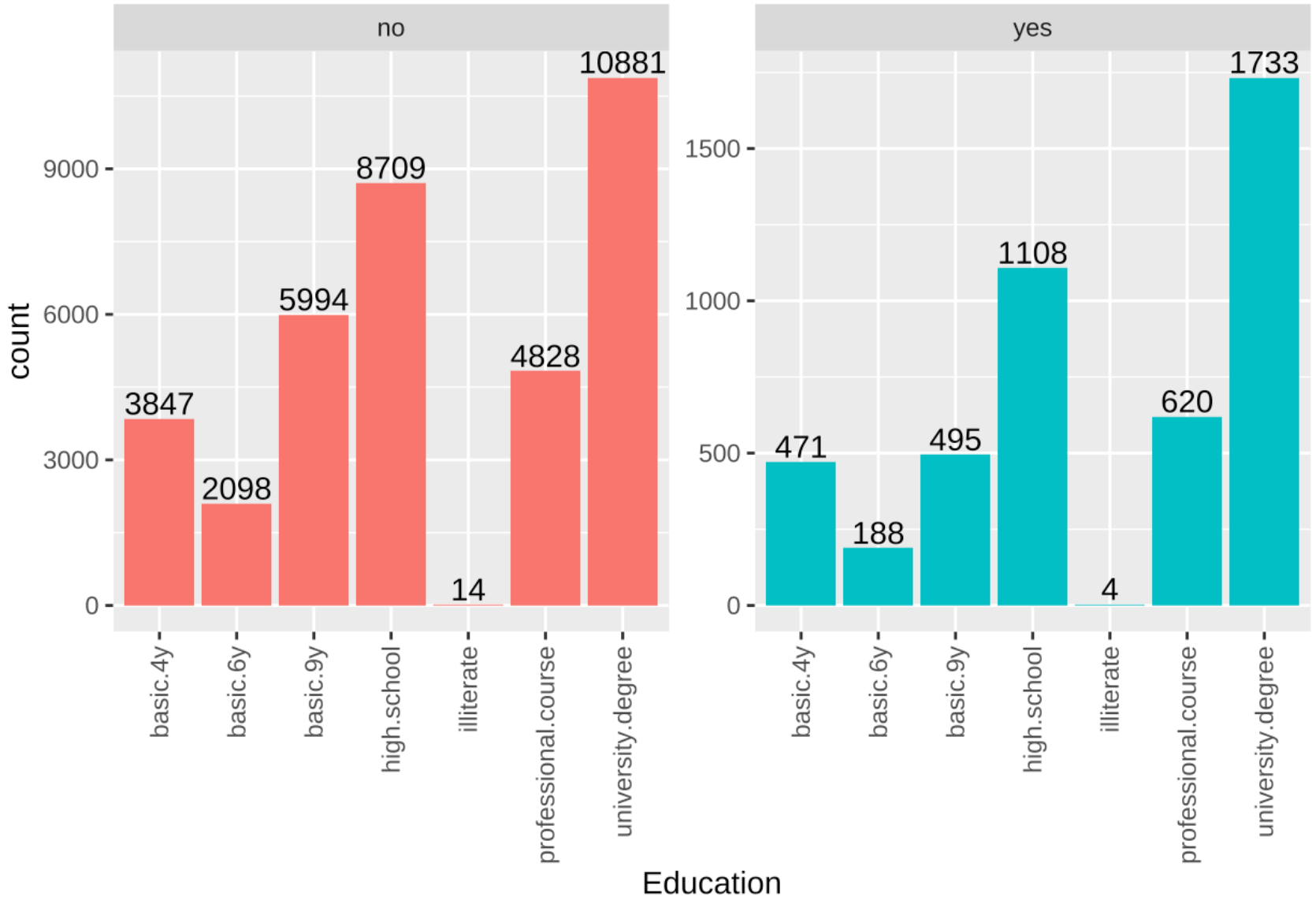


- Log transformations
- No significant improved obtained with log transformations

- In regression models, no assumptions made on the independent variables

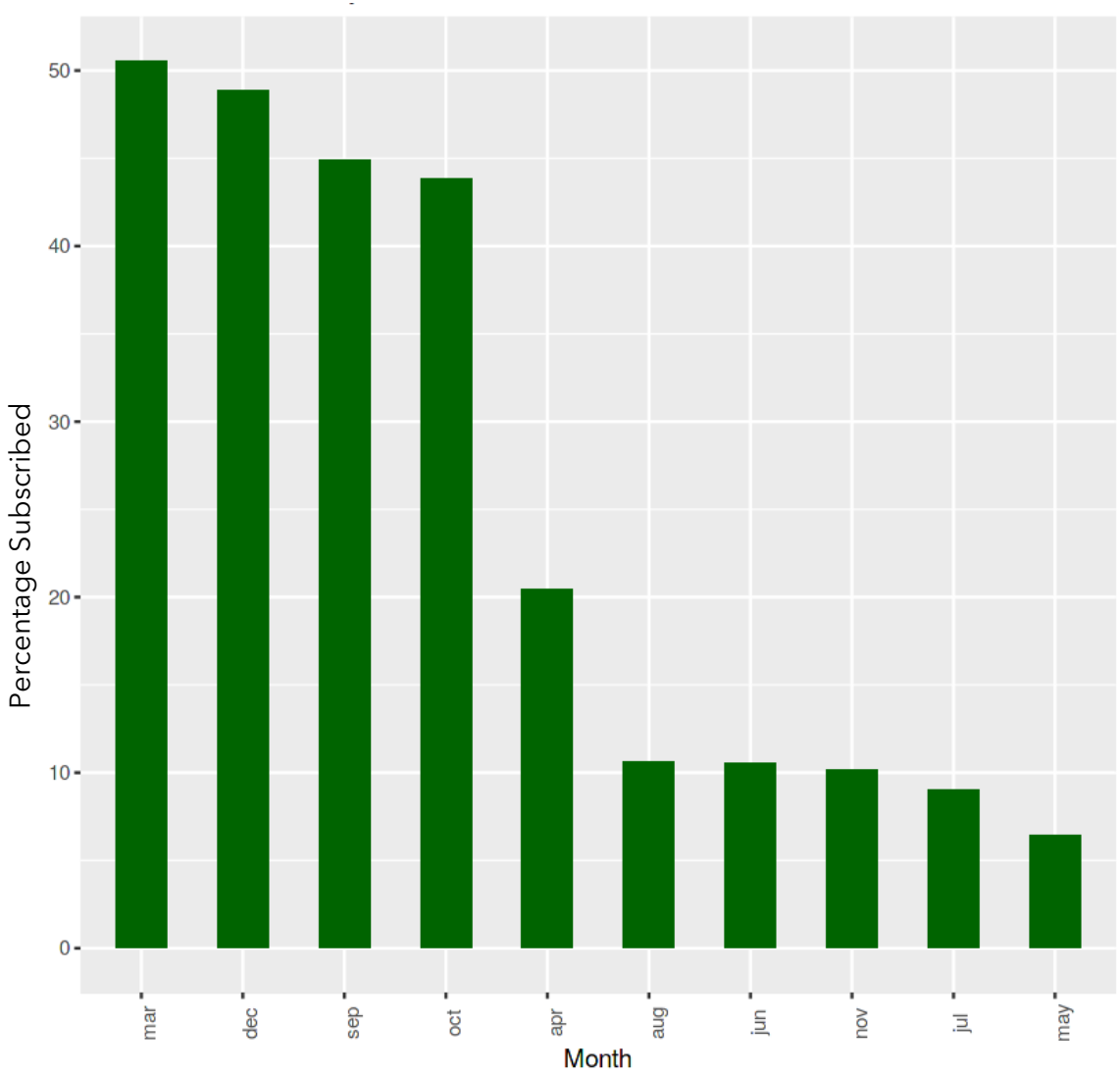
OVERLAY HISTOGRAM OF AGE VS Y-OUTPUT





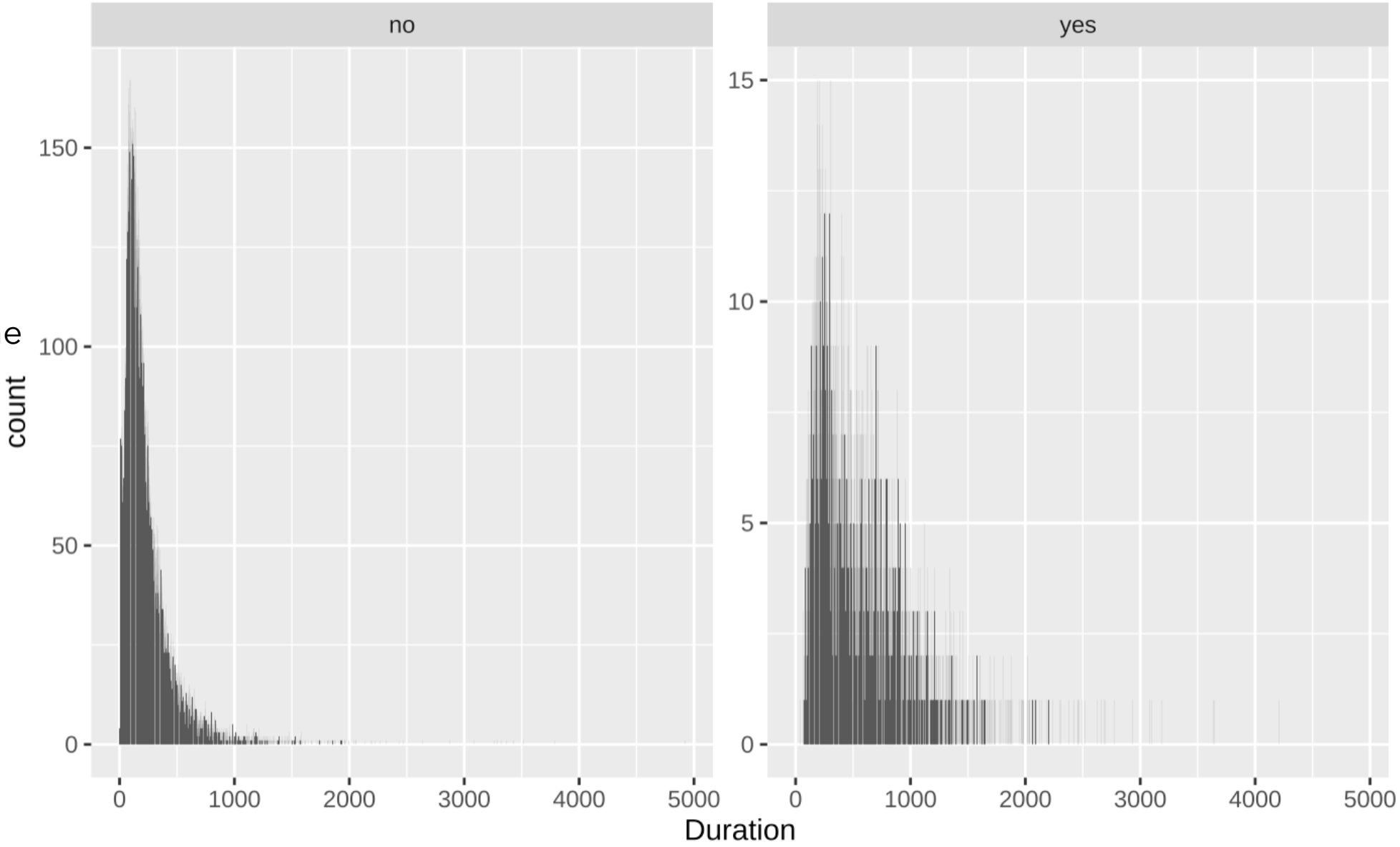
- Few illiterate, thus removed
- Similar trend for jobs

Month Analysis



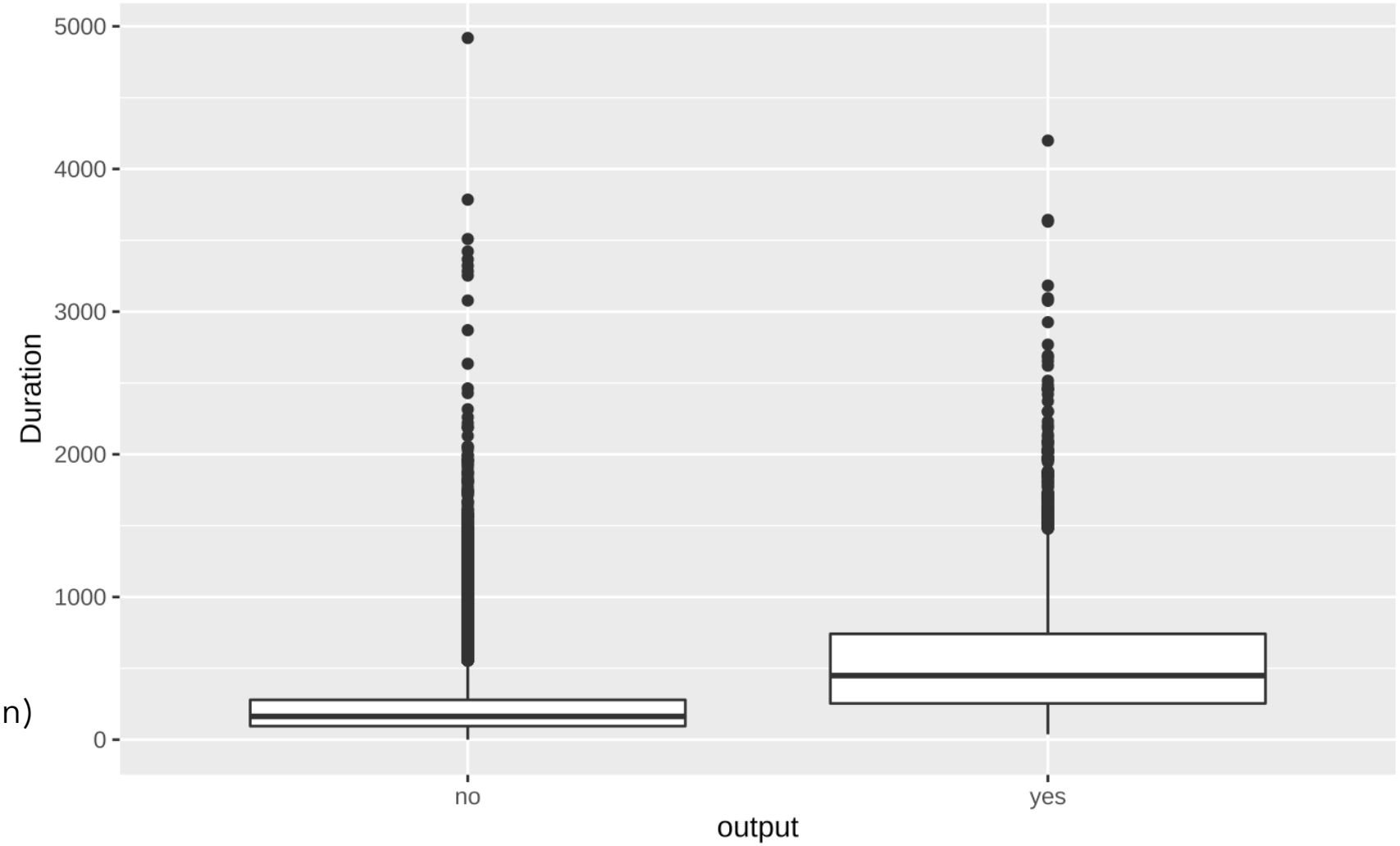
- Data shows March, December, and September are the months where clients subscribed the most.

- No: call durations mostly to towards the lower end
- Yes: calls **longer** on average



BOXPLOTS

DURATION

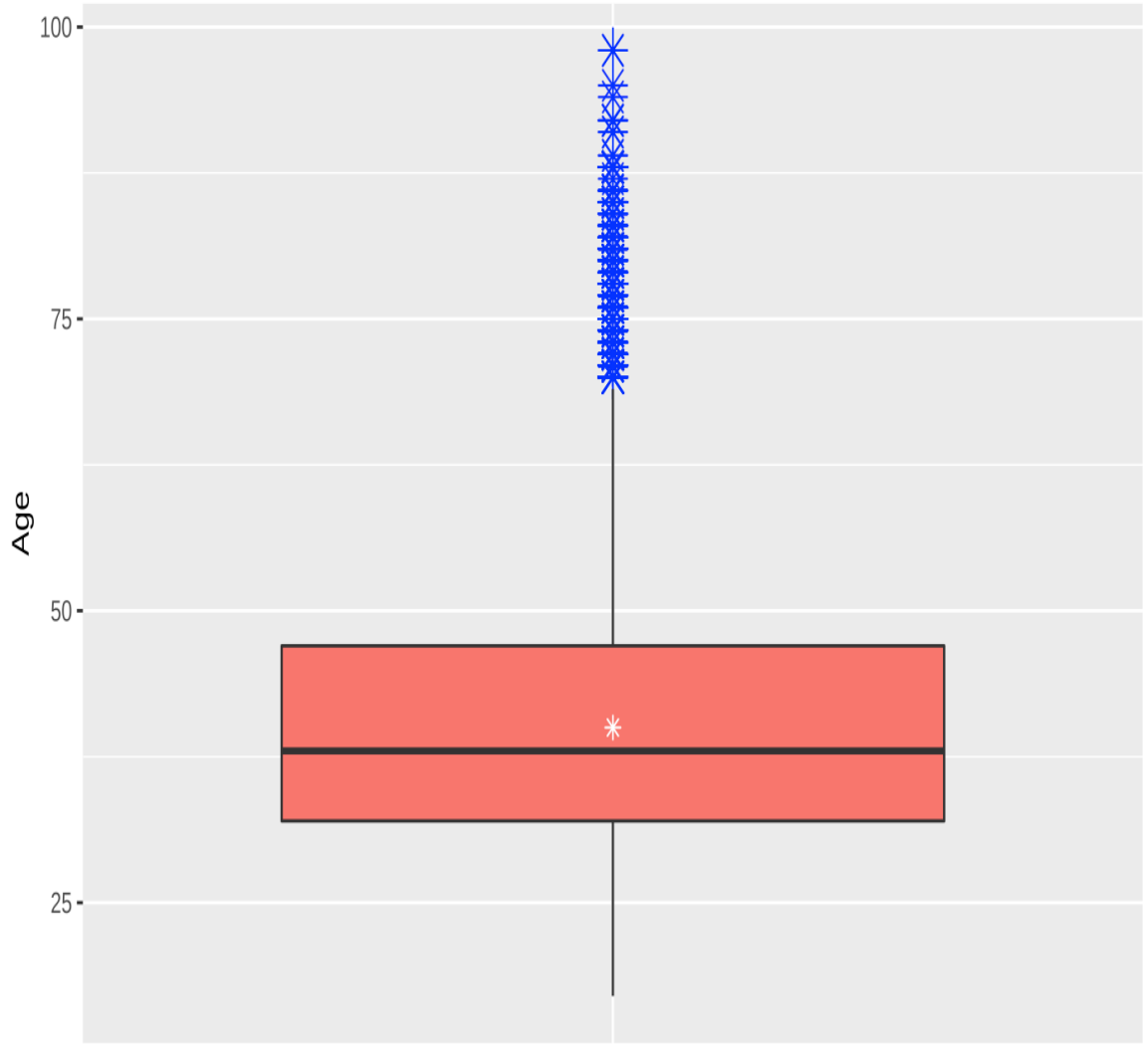


Interquartile range: 254s – 741s
(4.23 min – 12.35 min)

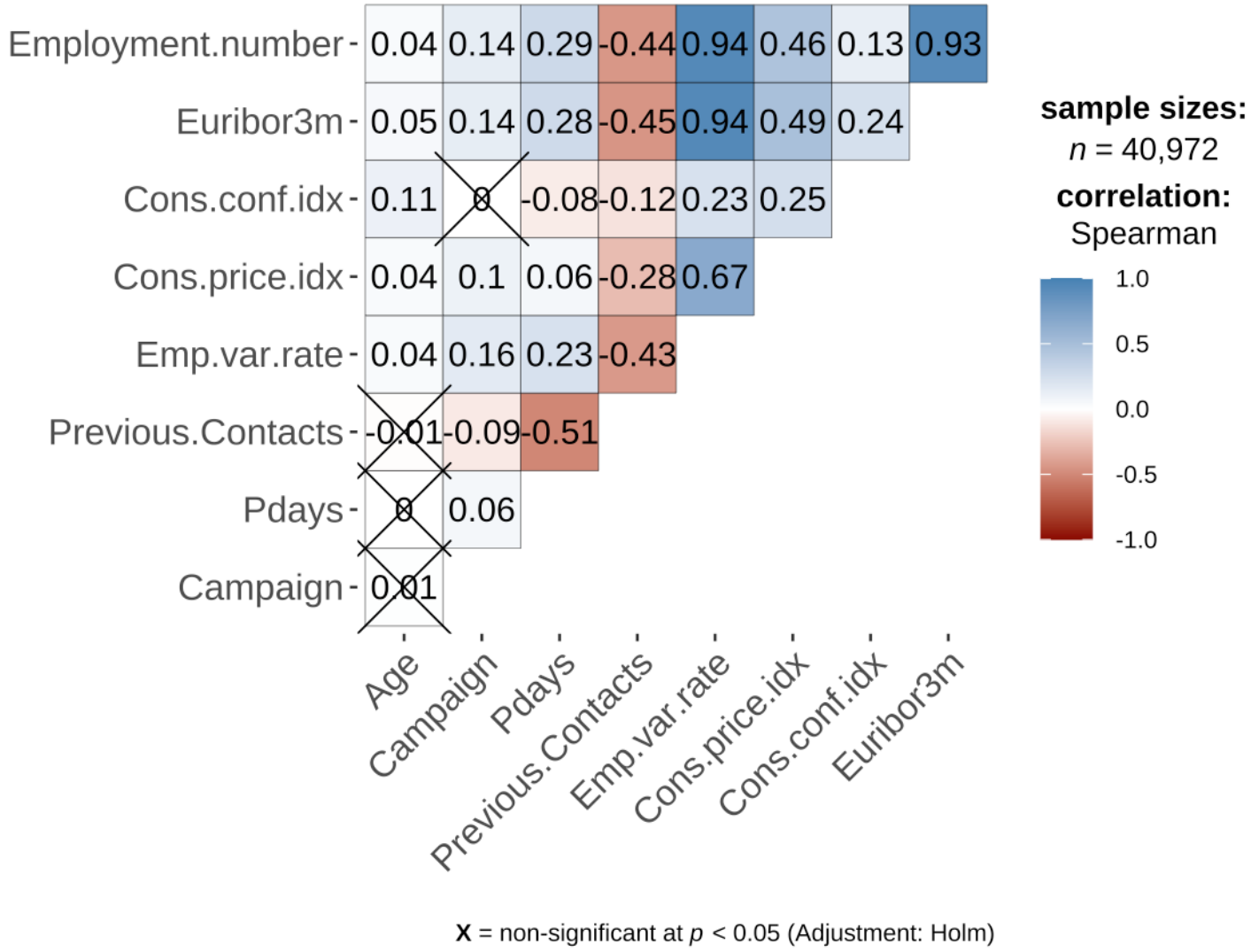
Mean: 553s (9.21min)

BOXPLOTS

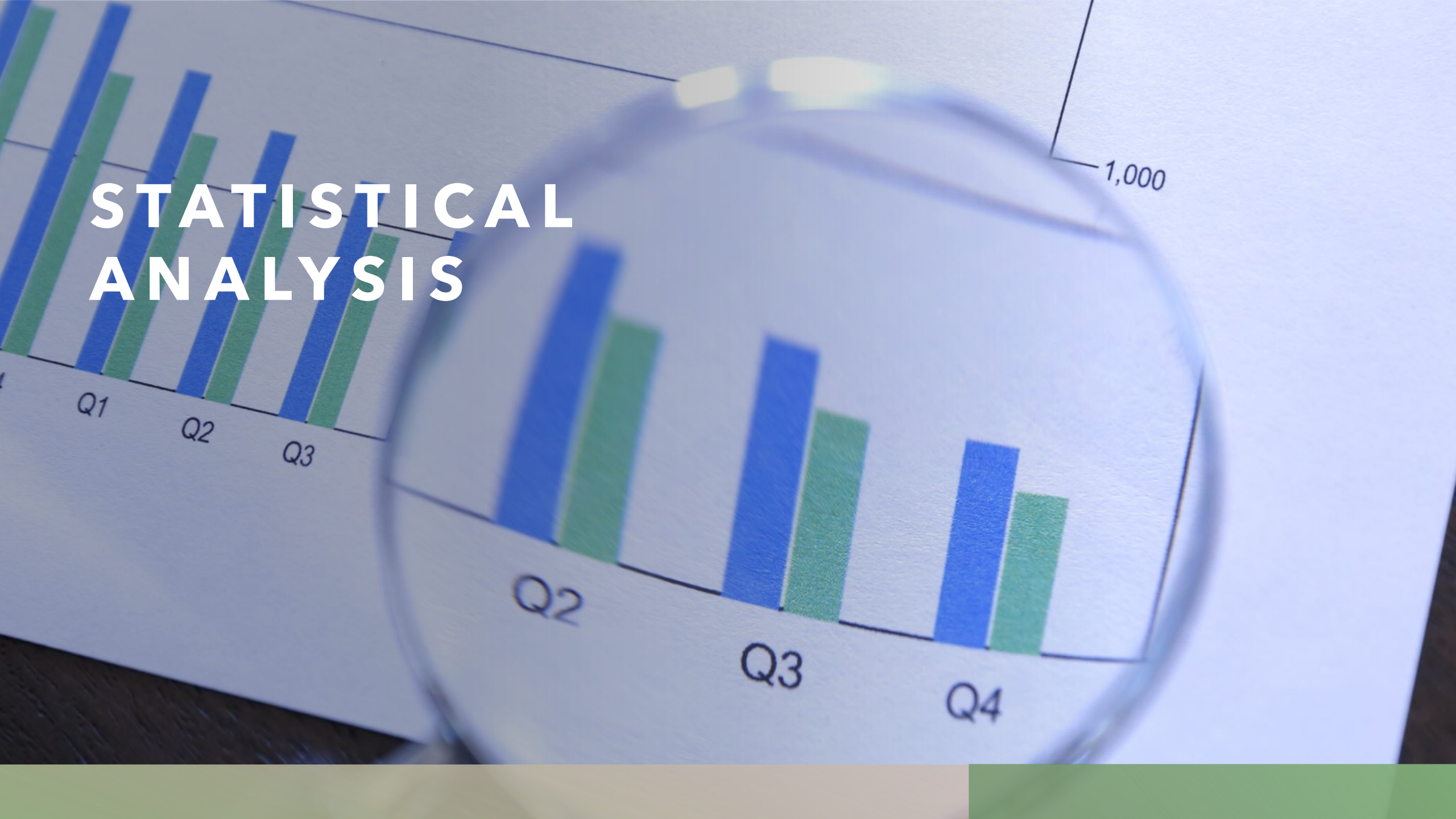
AGE



CORRELATION MATRIX



STATISTICAL ANALYSIS



CHI-SQUARE STATISTICAL TEST

CHECKING RELATIONSHIP BETWEEN CATEGORIES AND OUTPUT

Job

	no	yes
administration	9123	1362
blue-collar	8703	641
entrepreneur	1330	123
housemaid	951	106
management	2593	328
retired	1286	439
self-employed	1267	149
services	3640	323
student	599	275
technician	6014	729
unemployed	865	144

```
        Pearson's Chi-squared test

data:  bank[[col]] and bank$y
X-squared = 979.04, df = 10, p-value < 2.2e-16

[1] "effect size: 0.0488721275382095"
```

Marital

```
X-squared = 120.38, df = 2, p-value < 2.2e-16

[1] "effect size: 0.038319780063254 2"
```

Education

```
X-squared = 186.96, df = 6, p-value < 2.2e-16

[1] "effect size: 0.0275714940810703"
```

H_0: There is no relationship with the variable and the output (Independent)

H_1: There is a relationship with the variable and the output (dependent)

CHI-SQUARE TEST RESULTS

Housing

X-squared = 6.0813, df = 2, p-value = 0.0478

[1] "effect size: 0.00861276577315766"

Low effect size (Cramer's V)

Loan

X-squared = 0.99884, df = 2, p-value = 0.6069

[1] "effect size: 0.0034905550364950 5"

Contact

X-squared = 860.48, df = 1, p-value < 2.2e-16

[1] "effect size: 0.144887806747072"

Month

X-squared = 3078.9, df = 9, p-value < 2.2e-16

[1] "effect size: 0.0913560419150834"

Previous Contact day

X-squared = 25.771, df = 4, p-value = 3.519e-05

[1] "effect size: 0.0125370732017601"

Previous Outcome

X-squared = 4214.1, df = 2, p-value < 2.2e-16

[1] "effect size: 0.226725131514327"

DATA TRANSFORMATION

HANDLING CATEGORIES

One-Hot Encoded Variables

```
$ Job_blue-collar              : int  0 0 0 0 0 0 0 1 0 0 ...
$ Job_entrepreneur             : int  0 0 0 0 0 0 0 0 0 0 ...
$ Job_housemaid                : int  1 0 0 0 0 0 0 0 0 0 ...
$ Job_management               : int  0 0 0 0 0 0 0 0 0 0 ...
$ Job_retired                  : int  0 0 0 0 0 0 0 0 0 0 ...
$ Job_self-employed            : int  0 0 0 0 0 0 0 0 0 0 ...
$ Job_services                 : int  0 1 1 0 1 1 0 0 0 1 ...
$ Job_student                  : int  0 0 0 0 0 0 0 0 0 0 ...
$ Job_technician               : int  0 0 0 0 0 0 0 0 1 0 ...
$ Job_unemployed               : int  0 0 0 0 0 0 0 0 0 0 ...
$ Marital_married              : int  1 1 1 1 1 1 1 1 0 0 ...
$ Marital_single               : int  0 0 0 0 0 0 0 0 1 1 ...
$ Education_basic.6y           : int  0 0 0 1 0 0 0 0 0 0 ...
$ Education_basic.9y           : int  0 0 0 0 0 1 0 1 0 0 ...
$ Education_high.school        : int  0 1 1 0 1 0 0 0 0 1 ...
$ Education_professional.course: int  0 0 0 0 0 0 1 0 1 0 ...
$ Education_university.degree  : int  0 0 0 0 0 0 0 0 0 0 ...
$ Housing_unknown              : int  0 0 0 0 0 0 0 0 0 0 ...
$ Housing_yes                  : int  0 0 1 0 0 0 0 0 1 1 ...
$ Loan_unknown                 : int  0 0 0 0 0 0 0 0 0 0 ...
$ Loan_yes                     : int  0 0 0 0 1 0 0 0 0 0 ...
$ Contact_telephone            : int  1 1 1 1 1 1 1 1 1 1 ...
```

- Used for LDA, QDA, KNN and Naïve Bayes Modelling

Factor Variables

```
$ Job              : Factor w/ 11 levels "administration",..: 4 8 8 1 8 8 1 2 10 8 ...
$ Marital          : Factor w/ 3 levels "divorced","married",..: 2 2 2 2 2 2 2 2 3 3 ...
$ Education        : Factor w/ 6 levels "basic.4y","basic.6y",..: 1 4 4 2 4 3 5 3 5 4 ...
$ Housing          : Factor w/ 3 levels "no","unknown",..: 1 1 3 1 1 1 1 1 3 3 ...
$ Loan             : Factor w/ 3 levels "no","unknown",..: 1 1 1 1 3 1 1 1 1 1 ...
$ Contact          : Factor w/ 2 levels "cellular","telephone": 2 2 2 2 2 2 2 2 2 2 ...
$ Month            : Factor w/ 10 levels "apr","aug","dec",..: 7 7 7 7 7 7 7 7 7 7 ...
$ Last.Contact.Day : Factor w/ 5 levels "fri","mon","thu",..: 2 2 2 2 2 2 2 2 2 2 ...
```

- Used for Logistic Regression

RESCALING CONTINUOUS COLUMNS

MINMAX SCALING – Numerical Columns

```
'data.frame':   40972 obs. of  19 variables:
 $ Age              : num  0.481 0.494 0.247 0.284 0.481 ...
 $ Job              : Factor w/ 11 levels "administration",..: 4 8 8 1 8 8 1 2 10 8 ...
 $ Marital          : Factor w/ 3 levels "divorced","married",..: 2 2 2 2 2 2 2 2 3 3 ...
 $ Education        : Factor w/ 6 levels "basic.4y","basic.6y",..: 1 4 4 2 4 3 5 3 5 4 ...
 $ Housing          : Factor w/ 3 levels "no","unknown",..: 1 1 3 1 1 1 1 1 3 3 ...
 $ Loan             : Factor w/ 3 levels "no","unknown",..: 1 1 1 1 3 1 1 1 1 1 ...
 $ Contact          : Factor w/ 2 levels "cellular","telephone": 2 2 2 2 2 2 2 2 2 2 ...
 $ Month            : Factor w/ 10 levels "apr","aug","dec",..: 7 7 7 7 7 7 7 7 7 7 ...
 $ Last.Contact.Day : Factor w/ 5 levels "fri","mon","thu",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ Campaign         : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Pdays            : num  1 1 1 1 1 1 1 1 1 1 ...
 $ Previous.Contacts: num  0 0 0 0 0 0 0 0 0 0 ...
 $ Poutcome         : Factor w/ 3 levels "failure","nonexistent",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ Emp.var.rate     : num  0.938 0.938 0.938 0.938 0.938 ...
 $ Cons.price.idx   : num  0.699 0.699 0.699 0.699 0.699 ...
 $ Cons.conf.idx    : num  0.603 0.603 0.603 0.603 0.603 ...
 $ Euribor3m        : num  0.957 0.957 0.957 0.957 0.957 ...
 $ Employment.number: num  0.86 0.86 0.86 0.86 0.86 ...
 $ y                : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

Scaled from a range of 0 to 1

SPLITTING DATASET INTO TRAINING AND TESTING SET

```
# Split the data into training and test set
```{r}
set.seed(115)
trainIndices = sample(1:dim(bank)[1],round(.8 * dim(bank)[1]))
```

Build bank test/train
```{r}
bank.train = bank[trainIndices,]
bank.test = bank[-trainIndices,]
```

```{r}
print(table(bank.train$y)/nrow(bank.train))
print(table(bank.test$y)/nrow(bank.test))
bank.test.class <- bank.test$y
bank.test <- subset(bank.test,select=-c(y))
```
```

```
        0         1
0.8876306 0.1123694

        0         1
0.8864464 0.1135536
```

Training split

Testing split

MODELLING

LOGISTIC REGRESSION

LR FULL MODEL

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23219  on 32777  degrees of freedom
Residual deviance: 18080  on 32732  degrees of freedom
AIC: 18172

Number of Fisher Scoring iterations: 6
```

- Loan.unknown and Housing.unknown are perfectly colinear
- Reason: Clients choose not to disclose both information together.

```
Coefficients: (1 not defined because of singularities)
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -1.2573025  0.5151832  -2.440 0.014667 *
Age                         -0.1460792  0.1920763  -0.761 0.446940
Jobblue-collar              -0.1173002  0.0788588  -1.487 0.136890
Jobentrepreneur             -0.0043824  0.1180007  -0.037 0.970374
Jobhousemaid                -0.0938806  0.1482344  -0.633 0.526522
Jobmanagement               -0.0253760  0.0834163  -0.304 0.760969
Jobretired                   0.3401969  0.1060906   3.207 0.001343 **
Jobself-employed            -0.0104956  0.1127120  -0.093 0.925810
Jobservices                 -0.1042535  0.0849642  -1.227 0.219812
Jobstudent                   0.2848847  0.1095732   2.600 0.009324 **
Jobtechnician                0.0002343  0.0712199   0.003 0.997375
Jobunemployed                0.0103556  0.1236843   0.084 0.933274
Maritalmarried               0.0329193  0.0674172   0.488 0.625343
Maritalsingle                0.0290947  0.0771153   0.377 0.705959
Educationbasic.6y            0.1950516  0.1131341   1.724 0.084694 .
Educationbasic.9y           -0.0366701  0.0902569  -0.406 0.684533
Educationhigh.school         0.0460973  0.0888163   0.519 0.603748
Educationprofessional.course 0.0956247  0.0980003   0.976 0.329184
Educationuniversity.degree   0.1700879  0.0888256   1.915 0.055511 .
Housingunknown              -0.0977340  0.1334515  -0.732 0.463951
Housingyes                  -0.0213076  0.0404615  -0.527 0.598461
Loanunknown                         NA         NA      NA       NA
Loanyes                     -0.0773423  0.0565385  -1.368 0.171325
Contacttelephone            -0.7887348  0.0758115 -10.404  < 2e-16 ***
Monthaug                     0.3743492  0.1206825   3.102 0.001923 **
Monthdec                     0.5381976  0.2118717   2.540 0.011079 *
Monthjul                     0.0489800  0.0934453   0.524 0.600169
Monthjun                    -0.6306972  0.1235239  -5.106 3.29e-07 ***
Monthmar                     1.4968538  0.1463015  10.231  < 2e-16 ***
Monthmay                    -0.4397211  0.0808314  -5.440 5.33e-08 ***
Monthnov                    -0.4596285  0.1175851  -3.909 9.27e-05 ***
Monthoct                    -0.0552917  0.1514693  -0.365 0.715085
Monthsep                     0.1331698  0.1777168   0.749 0.453654
Last.Contact.Daymon         -0.2422047  0.0644226  -3.760 0.000170 ***
Last.Contact.Daythu          0.0287649  0.0623119   0.462 0.644349
Last.Contact.Daytue          0.0337580  0.0640165   0.527 0.597962
Last.Contact.Daywed          0.1205612  0.0637077   1.892 0.058436 .
Campaign                    -2.0791475  0.5546419  -3.749 0.000178 ***
Pdays                       -1.2254246  0.2262510  -5.416 6.09e-08 ***
Previous.Contacts           -0.4122987  0.4357934  -0.946 0.344104
Poutcomenonexistent          0.4729014  0.0966416   4.893 9.91e-07 ***
Poutcomesuccess              0.6813410  0.2218922   3.071 0.002136 **
Emp.var.rate                -7.0228597  0.6689174 -10.499  < 2e-16 ***
Cons.price.idx               5.3380864  0.6315784   8.452  < 2e-16 ***
Cons.conf.idx                0.8801959  0.1872645   4.700 2.60e-06 ***
Euribor3m                    0.6511612  0.5666569   1.149 0.250503
Employment.number            1.8949782  0.7984677   2.373 0.017631 *
```

LOGISTIC REGRESSION

BEST MODEL SELECTION USING STEPWISE BACKWARD ELIMINATION USING AIC

```
Call:  glm(formula = y ~ Job + Education + Contact + Month + Last.Contact.Day +
    Campaign + Pdays + Poutcome + Emp.var.rate + Cons.price.idx +
    Cons.conf.idx + Employment.number, family = "binomial", data = bank.train)
```

- These are the variables that gives the best (lowest) AIC score
- Full model AIC: 18,172
- Best model AIC: 18,160

```
Degrees of Freedom: 32777 Total (i.e. Null);  32740 Residual
Null Deviance:      23220
Residual Deviance: 18090        AIC: 18160
```

- Big difference between Null deviance and Residual deviance

LOGISTIC REGRESSION

WITH BOOTSTRAP SAMPLING

Bootstrap samples: 50
Direction: backward
Penalty: 2 * df

Covariates selected

	(%)
Cons.conf.idx	100
Cons.price.idx	100
Contact	100
Emp.var.rate	100
Job	100
Last.Contact.Day	100
Month	100
Pdays	100
Poutcome	100
Campaign	98
Education	82
Employment.number	82
Euribor3m	44
Loan	36
Age	32
Previous.Contacts	26
Housing	18
Marital	16

Coefficient sign shows the variables that are stable

- Consumer Confidence index was included in 100% of the candidate models
- Marital was only included in 16% of the candidate models

Coefficients Sign

	+ (%)	- (%)
Cons.conf.idx	100.00	0.00
Cons.price.idx	100.00	0.00
Educationuniversity.degree	100.00	0.00
Employment.number	100.00	0.00
Euribor3m	100.00	0.00
Jobretired	100.00	0.00
Last.Contact.Daywed	100.00	0.00
Monthaug	100.00	0.00
Monthdec	100.00	0.00
Monthmar	100.00	0.00
Poutcomenonexistent	100.00	0.00
Poutcomesuccess	100.00	0.00
Educationbasic.6y	97.56	2.44
Jobstudent	96.00	4.00
Educationprofessional.course	87.80	12.20
Maritalmarried	87.50	12.50
Educationhigh.school	78.05	21.95
Maritalsingle	75.00	25.00
Last.Contact.Daytue	74.00	26.00
Monthjul	72.00	28.00
Monthsep	72.00	28.00
Last.Contact.Daythu	62.00	38.00
Jobentrepreneur	50.00	50.00
Jobself-employed	50.00	50.00
Jobunemployed	50.00	50.00
Jobtechnician	46.00	54.00
Monthoct	38.00	62.00
Jobmanagement	34.00	66.00
Educationbasic.9y	31.71	68.29
Jobhousemaid	28.00	72.00
Jobservices	14.00	86.00
Jobblue-collar	12.00	88.00
Housingunknown	11.11	88.89
Housingyes	11.11	88.89
Age	6.25	93.75

STATISTICAL SIGNIFICANCE - VARIABLES

Stat Significance

	(%)
Campaign	100.00
Cons.conf.idx	100.00
Cons.price.idx	100.00
Contacttelephone	100.00
Emp.var.rate	100.00
Monthmar	100.00
Monthmay	100.00
Pdays	100.00
Poutcomenonexistent	100.00
Last.Contact.Daymon	98.00
Monthjun	98.00
Monthnov	98.00
Employment.number	92.68
Jobretired	92.00
Monthaug	90.00
Poutcomesuccess	86.00
Jobstudent	82.00
Educationuniversity.degree	70.73
Previous.Contacts	69.23
Monthdec	68.00
Maritalmarried	62.50
Age	56.25
Euribor3m	54.55
Maritalsingle	50.00
Last.Contact.Daywed	48.00
Housingyes	44.44
Loanyes	44.44
Educationbasic.6y	34.15
Educationprofessional.course	31.71
Jobblue-collar	28.00
Jobservices	28.00
Monthsep	24.00
Housingunknown	22.22
Loanunknown	18.75
Educationhigh.school	14.63
Monthoct	14.00
Jobtechnician	10.00
Jobself-employed	8.00
Monthjul	8.00
Jobhousemaid	6.00
Jobmanagement	6.00
Jobunemployed	6.00
Last.Contact.Daythu	6.00
Jobentrepreneur	4.00
Last.Contact.Daytue	4.00
Educationbasic.9y	2.44

- Shows the statistical significance of the variables
- Campaign, etc have consistently high significance in the models

MODEL SELECTION PROCESS

```
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
y ~ Age + Job + Marital + Education + Housing + Loan + Contact +
    Month + Last.Contact.Day + Campaign + Pdays + Previous.Contacts +
    Poutcome + Emp.var.rate + Cons.price.idx + Cons.conf.idx +
    Euribor3m + Employment.number
```

```
Final Model:
y ~ Job + Education + Contact + Month + Last.Contact.Day + Campaign +
    Pdays + Poutcome + Emp.var.rate + Cons.price.idx + Cons.conf.idx +
    Employment.number
```

Variables that were removed in each step of the model selection

	Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
1				32732	18080.41	18172.41
2	- Marital	2	0.2396035	32734	18080.65	18168.65
3	- Housing	1	0.2757271	32735	18080.92	18166.92
4	- Loan	2	2.2873374	32737	18083.21	18165.21
5	- Age	1	0.7662522	32738	18083.98	18163.98
6	- Previous.Contacts	1	0.9031918	32739	18084.88	18162.88
7	- Euribor3m	1	1.3187374	32740	18086.20	18162.20

Comparing Variance Inflation Factor

Our Best Model

High scores

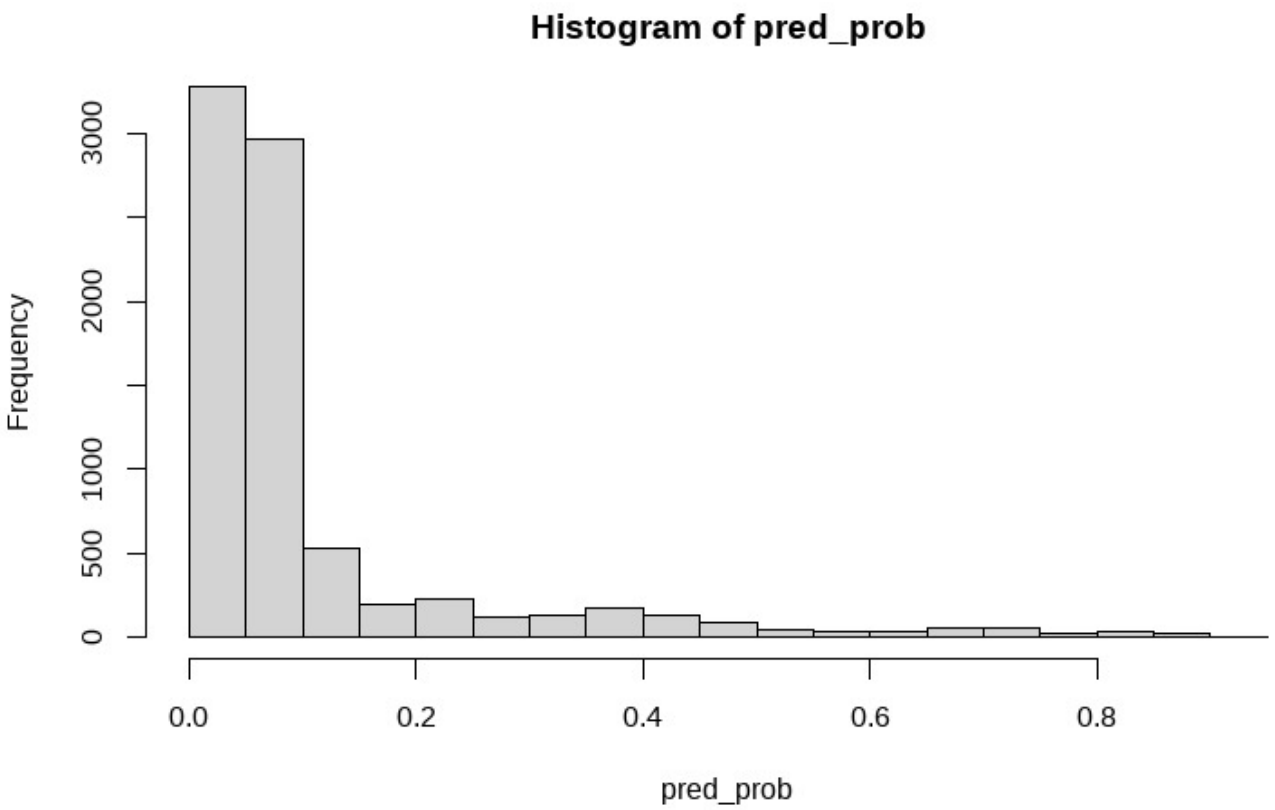
	GVIF	Df	GVIF^(1/(2*Df))
Job	3.771678	10	1.068628
Education	3.359351	5	1.128822
Contact	2.425169	1	1.557295
Month	27.190240	9	1.201405
Last.Contact.Day	1.047319	4	1.005796
Campaign	1.042488	1	1.021023
Pdays	9.448266	1	3.073803
Poutcome	10.594971	2	1.804160
Emp.var.rate	144.033959	1	12.001415
Cons.price.idx	53.853756	1	7.338512
Cons.conf.idx	2.602511	1	1.613230
Employment.number	75.155914	1	8.669251

Removed Employment Variation rate

Low scores

	GVIF	Df	GVIF^(1/(2*Df))
Job	3.767926	10	1.068575
Education	3.367212	5	1.129086
Contact	1.900338	1	1.378527
Month	5.112846	9	1.094889
Last.Contact.Day	1.044074	4	1.005406
Campaign	1.040604	1	1.020100
Pdays	9.436236	1	3.071846
Poutcome	10.562847	2	1.802790
Cons.price.idx	1.882988	1	1.372220
Cons.conf.idx	2.293264	1	1.514353
Employment.number	2.098835	1	1.448736

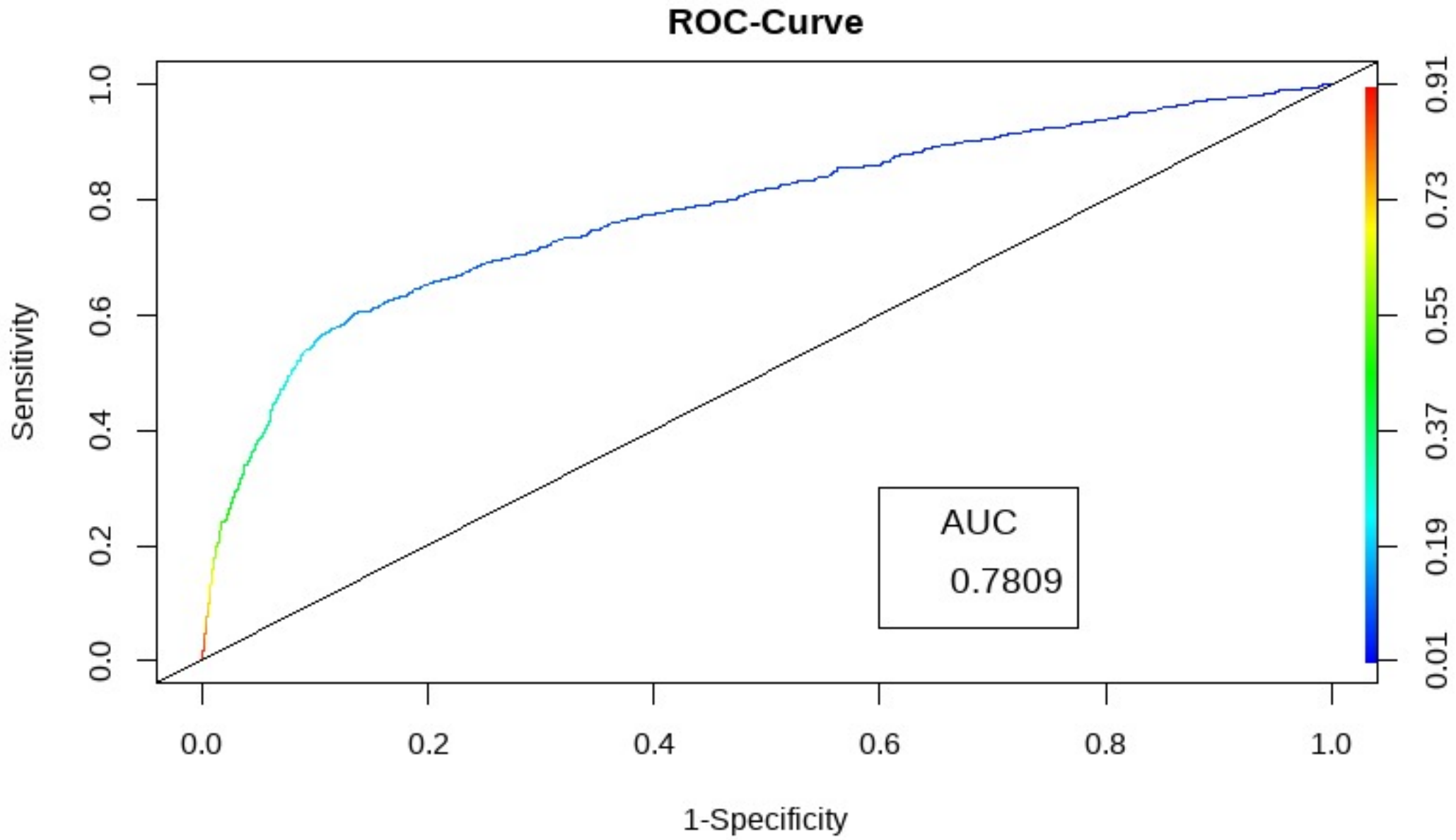
Eventhough this model has low scores, it performs worse than the other model



Confusion Matrix (threshold = 0.2)

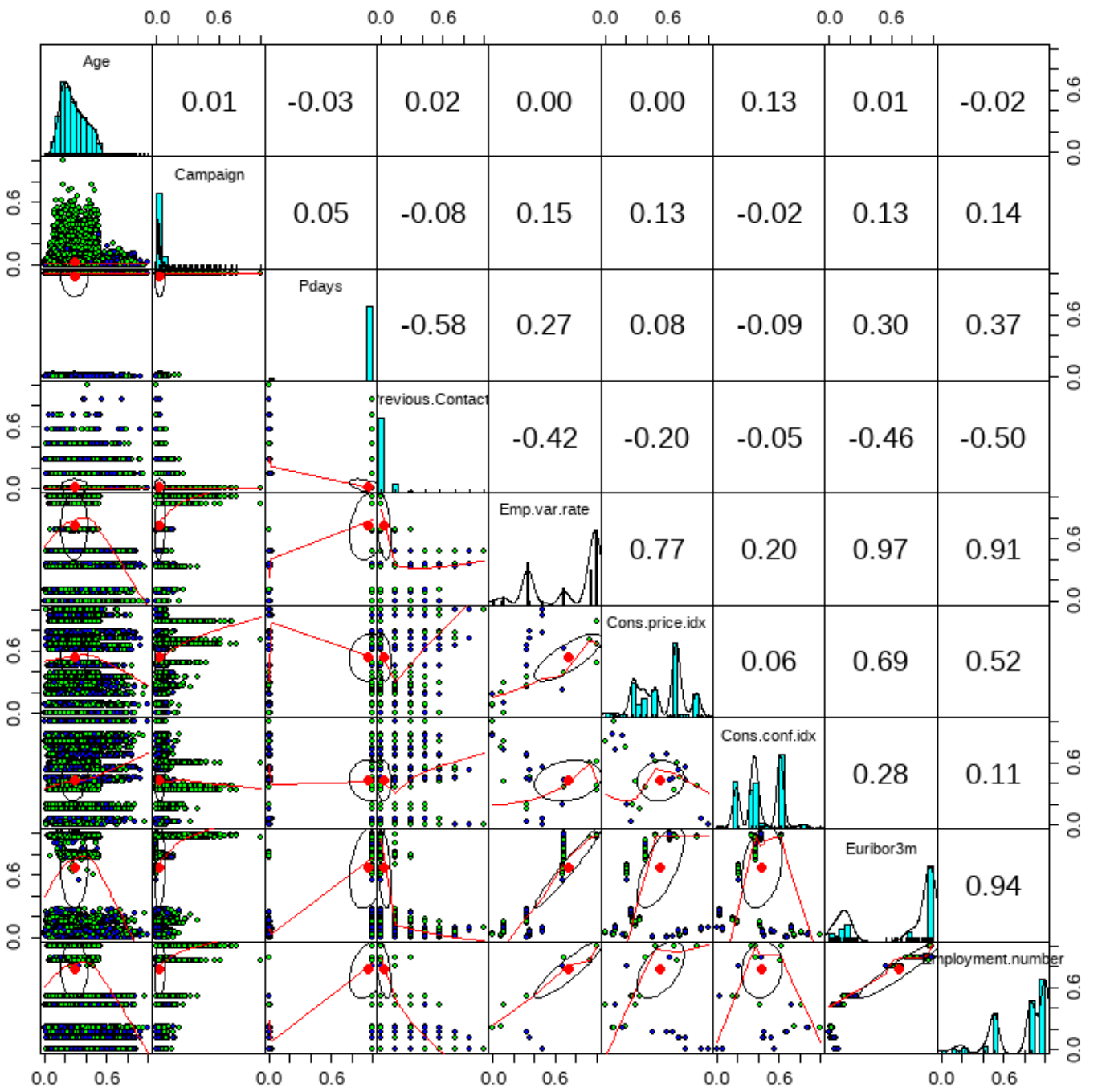
	Actual	
Predicted	0	1
0	6574	395
1	732	493

Accuracy: 86.25%
Sensitivity: 89.98%
Specificity: 55.52%


ROC-Curve
Sensitivity
1-Specificity
0.0
0.2
0.4
0.6
0.8
1.0
0.01
0.19
0.37
0.55
0.73
0.91
AUC
0.7809

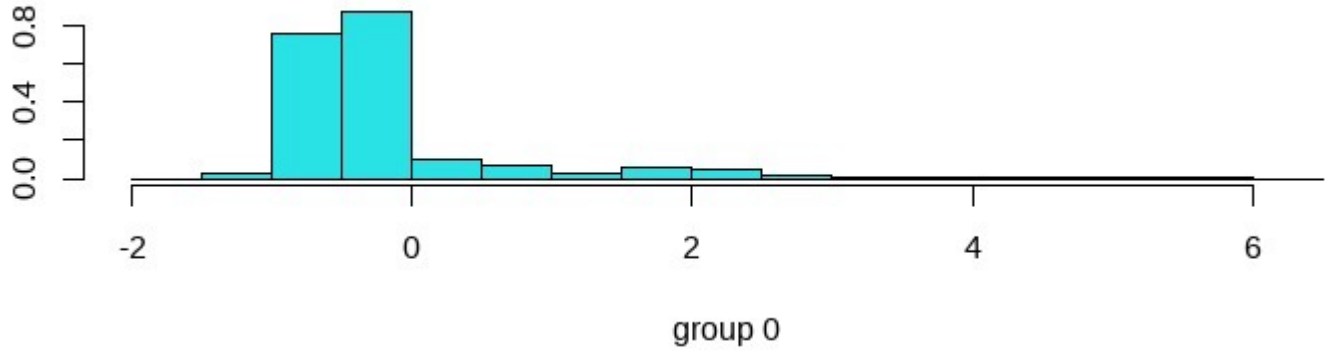
MODELS-LDA & QDA

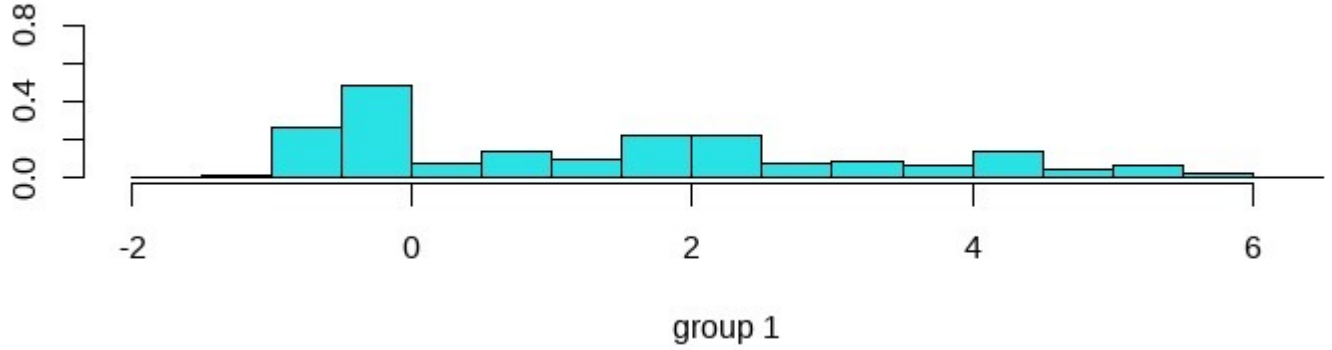
- No significant linear & quadratic relationships in the graph between variables and output
- No 2 variables are properly separating the output (green=1, blue=0) as seen in the pairplots .
- Due to Collinearity problem, Housing column was removed.



- Significant predictors for LDA

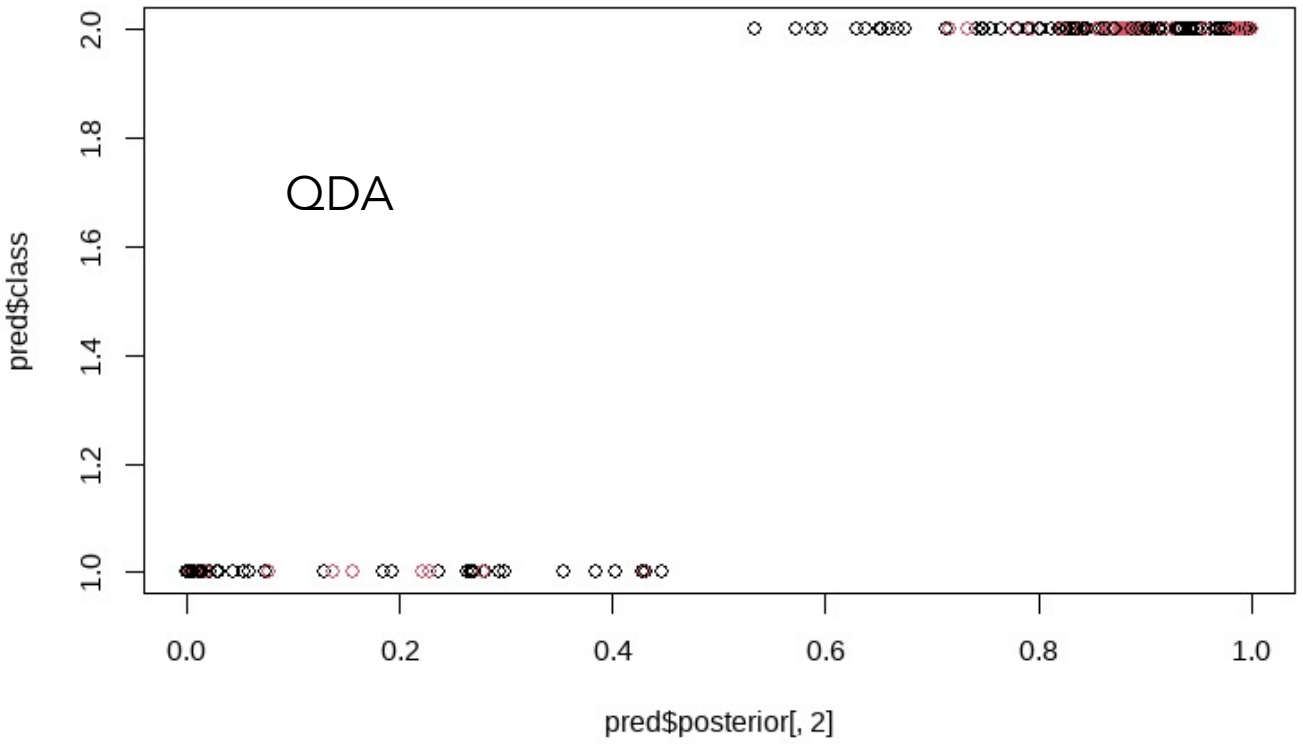
Emp.var.rate	-7.170172538
Cons.price.idx	5.024595071
Cons.conf.idx	1.024463201
Euribor3m	2.121984235

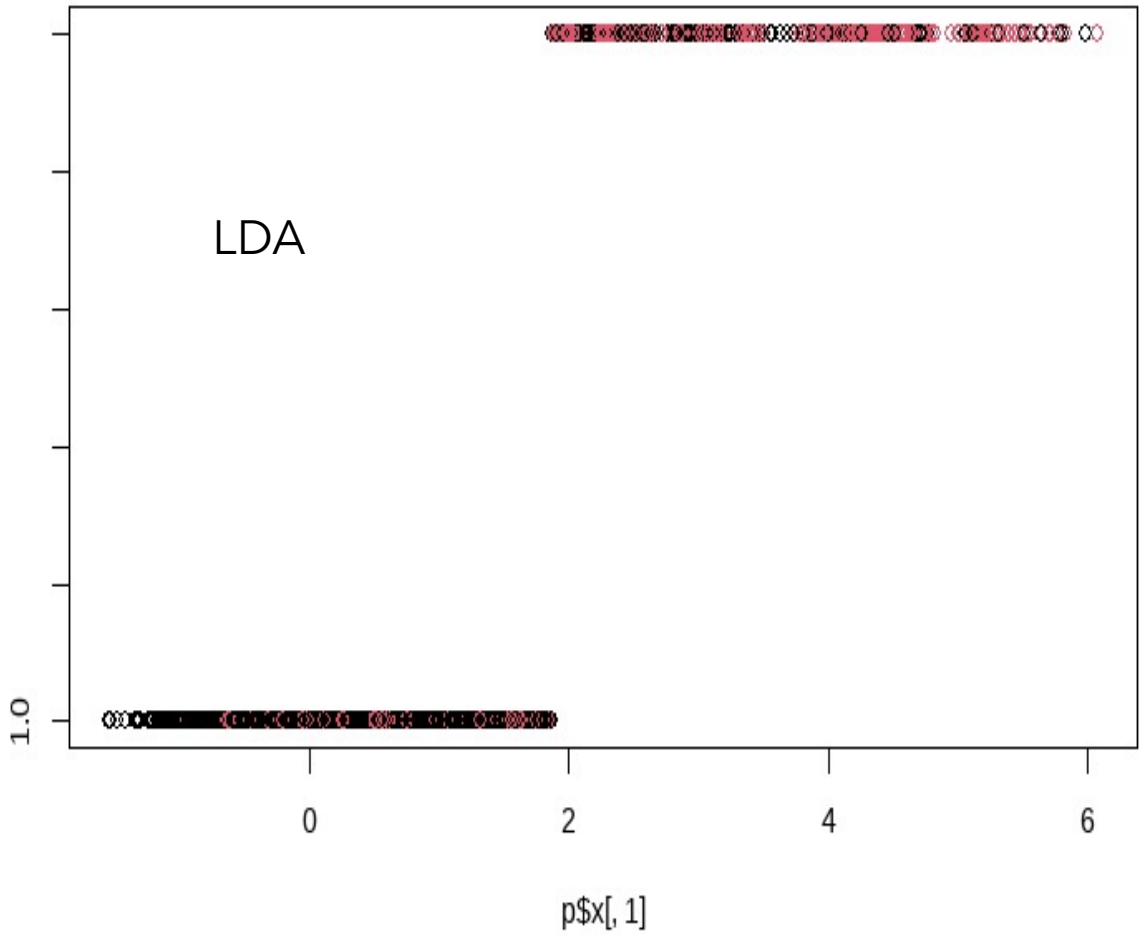




LDA shows poor separation between the classes (yes or no) due to the fact there are non-linear relationships between variables and output

Classification prediction Analysis





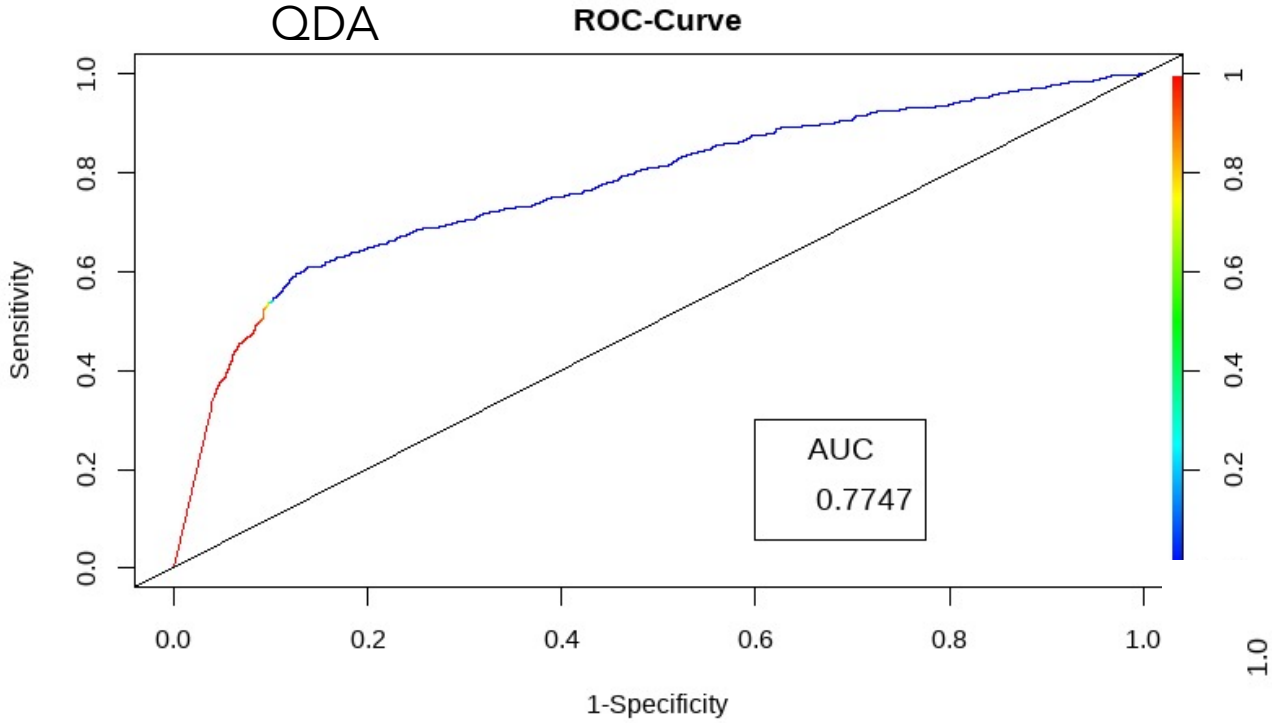
- The predicted points were coinciding

CONFUSION MATRIX – LDA & QDA

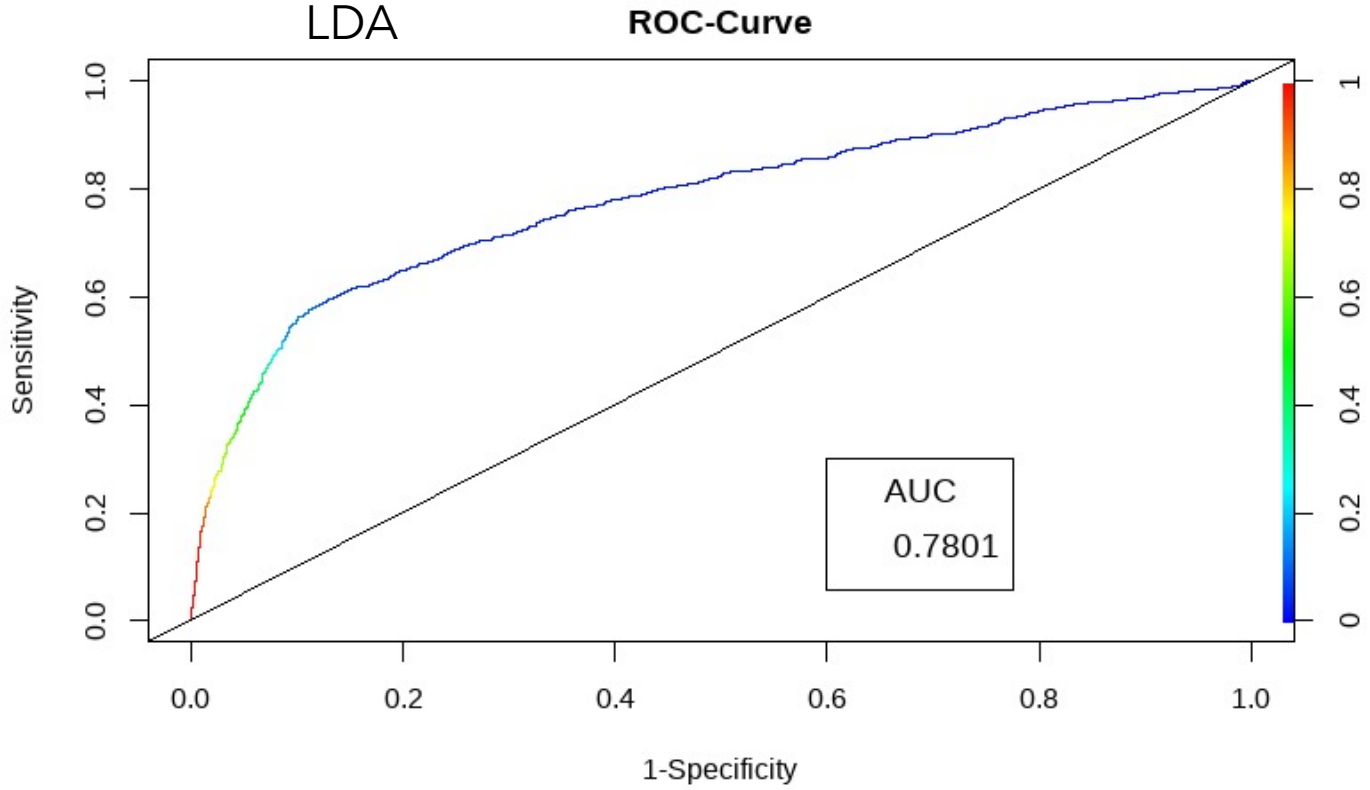
Predicted \ Actual	0	1
0	6936	542
1	370	346

Predicted \ bin.test.class	0	1
0	6579	411
1	727	477

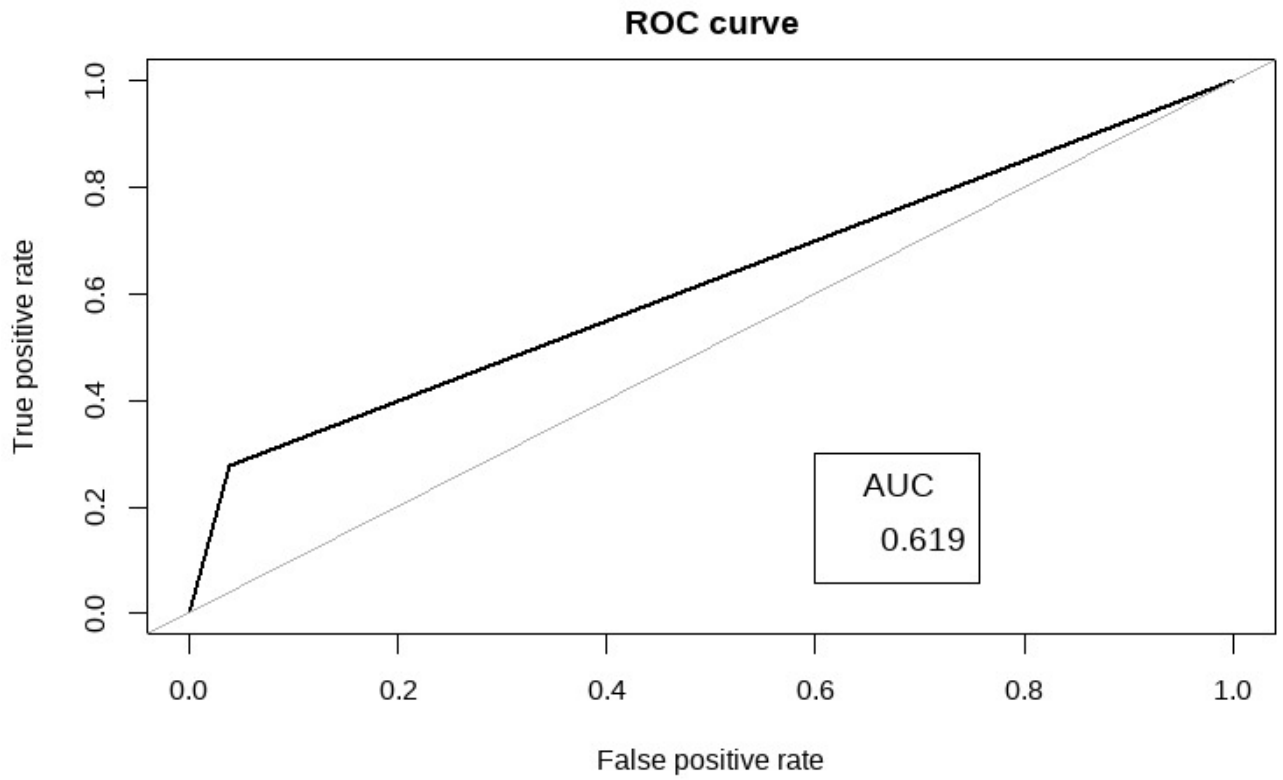
QDA


ROC-Curve
Sensitivity
1-Specificity
0.0
0.2
0.4
0.6
0.8
1.0
AUC
0.7747

LDA


ROC-Curve
Sensitivity
1-Specificity
0.0
0.2
0.4
0.6
0.8
1.0
AUC
0.7801

KNN

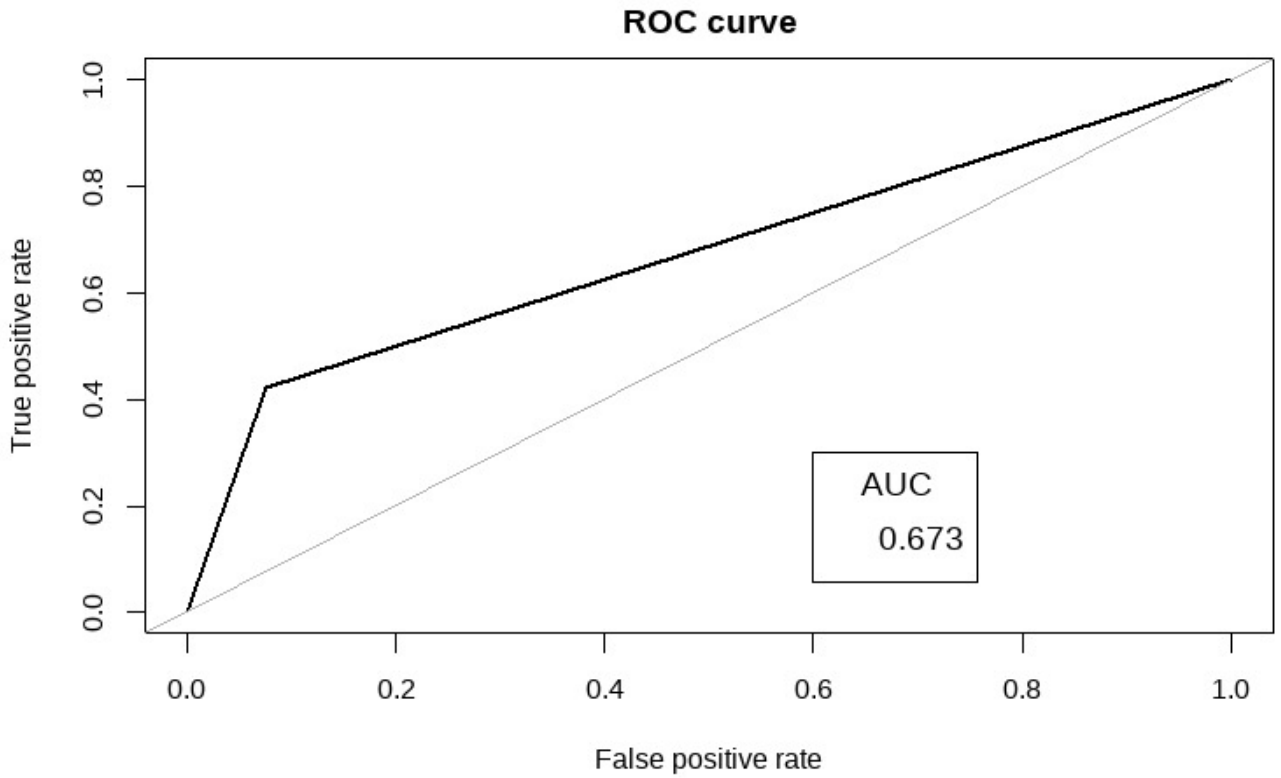


Confusion Matrix

	Actual	
Predicted	0	1
0	7024	642
1	282	246

Accuracy: 88.7%%
Sensitivity: 96.14%
Specificity: 27.70%

NAÏVE BAYES



Confusion Matrix

	Actual	
Predicted	0	1
0	6760	514
1	546	374

Accuracy: 87.06%
Sensitivity: 92.53%
Specificity: 42.12%

RESULTS

- Logistic Regression: AUC=78.09%, Best model AIC = 18,162, Full model AIC:18,172
- LDA: AUC=78.01%
- QDA: AUC=77.47%
- KNN: AUC= 62%
- Naïve Bayes: AUC=67.3%

CONCLUSION

- Among the models, the best performing was Logistic Regression
- Job + Education + Contact + Month + Last.Contact.Day + Campaign + Pdays + Poutcome + Emp.var.rate + Cons.price.idx+Cons.conf.idx + Employment.number decides whether client subscribe the term deposit or not

- It was also noted that calls were most successful during the months March, December, September, and October
- Clients should be contacted on their cellphones
- Call durations should be from 4.23 min to 12.35 min

THANK YOU