Predicting if client subscribes to Term Deposit(TD)

MATH2319 Machine Learning Applied Project Phase I

Ahmad Hasnain (s3712538)

28th April 2019

Contents

1 Introduction

The goal of this project was to build classifiers to predict whether an individual will subscribe to a term deposit from the direct marketing campaigns (phone calls) of a Portuguese banking institution run between 2008 to 2010.The data sets were sourced from the UCI Machine LearningRepository. The project has two phases. Phase I focuses on data preprocessing and exploration, as covered in this report. We shall present model building in Phase II. The rest of this report is organised as follow. Section 2 describes the data sets and their attributes. Section 3 covers data pre-processing. In Section 4, we explore each attribute and their inter-relationships. The last section ends with a summary.

2 Data Set

The UCI Machine Learning Repository provided four data sets, but only `bank-full.csv` were useful in this project. Data description is available in `bank-names.txt`.The data data set has 45,211 observations and consist of 15 descriptives features and one target feature i.e. Term Deposit. In Phase I we will do the necessary data preprocessing.In Phase II and we will build the classifiers from the the data set and evaluate their performance using cross-validation.

2.1 Target Feature

The response feature is term deposit(TD) which is given as:

$$\text{Term Deposit} = \begin{cases} Yes & \text{if customer subscribed for Term Deposit} \\ No & \text{if customer didnt subscribe for Term Deposit} \end{cases} \tag{1}$$

The target feature has two classes and hence it is a binary classification problem. To reiterate, The goal is to predict **if the client will subscribe a term deposit ?**.

2.2 Descriptive Features

The variable description is produced here from `bank-names.txt` file:

- age :numeric.
- job : categorical:(“admin.”,“unknown”,“unemployed”,“management”,“housemaid”,“entrepreneur”,“student”, “blue-collar”,“self-employed”,“retired”,“technician”,“services”).
- marital : marital status (categorical: “married”,“divorced”,“single”; note: “divorced” means divorced or widowed).
- education (categorical: “unknown”,“secondary”,“primary”,“tertiary”).
- default: has credit in default? (binary: “yes”,“no”).
- balance: average yearly balance, in euros (numeric).
- housing: has housing loan? (binary: “yes”,“no”).
- loan: has personal loan? (binary: “yes”,“no”).
- contact: contact communication type (categorical: “unknown”,“telephone”,“cellular”).
- day: last contact day of the month (numeric).
- month: last contact month of year (categorical: “jan”, “feb”, “mar”, . . . , “nov”, “dec”).
- 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:“unknown”,“other”,“failure”,“success”).

Most of the descriptive features are self-explanatory, contact,day,month,duration are related with the last contact made for the current marketing campaign. For more details, see RepositóriUM.

3 Data Pre-processing

3.1 Preliminaries

In this project, we used the following `R` packages.

```
library(readr)
library(magrittr)
library(knitr)
library(devtools)
library(ggplot2)
library(cowplot)
library(tidyr)
library(dplyr)
library(ggplot2)
library(mlr)
```

We read the datasets by treating the string values as characters, and trimming the extra white space in data. We would later convert the string columns to factor (categorical) after the data processing. To do away with ambiguity and help understand the variables,I have manually renamed the columns into more meaningful names.

```
bank <- read_delim("~/Dropbox/MC242/ML/R/bank/bank-full.csv",
                           ";", escape_double = FALSE, trim_ws = TRUE)

names(bank) <- c("age","job_category","marital_status","education_level",
                 "credit_default","avg_balance","housing_loan","personal_loan",
                 "contact_mode","contact_day","contact_month","last_contact_duration",
                 "no_of_timescontacted","previous_contact_days","previous_no_of_contacts",
                 "previous_outcome","Term_deposit")
```

3.2 Data Cleaning and Transformation

With `summarizeColumns` (see Table 1), we noticed the following anomalies:

- The min value of `capital_gain` is -1 , representing client *not contacted* .
- The min value of `avg_balance` was -8019 It could be a valid as many banks allow -ve balance for a certain period.
- On surface, each feature had no missing value.

```
summarizeColumns(bank) %>% knitr::kable( caption =  'Feature Summary before Data Preprocessing')
```

Table 1: Feature Summary before Data Preprocessing

name	type	na	mean	disp	median	mad	min	max	nlevs
age	integer	0	40.9362102	10.6187620	39	10.3782	18	95	0
job_category	character	0	NA	0.7847427	NA	NA	288	9732	12
marital_status	character	0	NA	0.3980668	NA	NA	5207	27214	3
education_level	character	0	NA	0.4868063	NA	NA	1857	23202	4
credit_default	character	0	NA	0.0180266	NA	NA	815	44396	2
avg_balance	integer	0	1362.2720577	3044.7658292	448	664.2048	-8019	102127	0
housing_loan	character	0	NA	0.4441618	NA	NA	20081	25130	2
personal_loan	character	0	NA	0.1602265	NA	NA	7244	37967	2
contact_mode	character	0	NA	0.3522594	NA	NA	2906	29285	3
contact_day	integer	0	15.8064188	8.3224762	16	10.3782	1	31	0
contact_month	character	0	NA	0.6955166	NA	NA	214	13766	12
last_contact_duration	integer	0	258.1630798	257.5278123	180	137.8818	0	4918	0
no_of_timescontacted	integer	0	2.7638407	3.0980209	2	1.4826	1	63	0
previous_contact_days	integer	0	40.1978280	100.1287460	-1	0.0000	-1	871	0
previous_no_of_contacts	integer	0	0.5803234	2.3034410	0	0.0000	0	275	0
previous_outcome	character	0	NA	0.1825220	NA	NA	1511	36959	4
Term_deposit	character	0	NA	0.1169848	NA	NA	5289	39922	2

Explore each non-numeric feature in the dataset:

```
sapply(bank [,sapply(bank, is.character)], table)
```

```
## $job_category
##
##        admin.   blue-collar  entrepreneur     housemaid    management
##          5171          9732          1487          1240          9458
##       retired self-employed      services       student    technician
##          2264          1579          4154           938          7597
##    unemployed       unknown
##          1303           288
##
## $marital_status
##
## divorced  married   single
##     5207    27214    12790
##
## $education_level
##
##   primary secondary  tertiary   unknown
##      6851     23202     13301      1857
##
## $credit_default
##
##    no   yes
## 44396   815
##
## $housing_loan
##
```

```
##    no   yes
## 20081 25130
##
## $personal_loan
##
##    no   yes
## 37967  7244
##
## $contact_mode
##
##  cellular telephone   unknown
##     29285      2906     13020
##
## $contact_month
##
##   apr   aug   dec   feb   jan   jul   jun   mar   may   nov   oct   sep
##  2932  6247   214  2649  1403  6895  5341   477 13766  3970   738   579
##
## $previous_outcome
##
## failure   other success unknown
##    4901    1840    1511   36959
##
## $Term_deposit
##
##    no   yes
## 39922  5289
```

We computed the level table for each character feature.The tables revealed:

- The value "unknown" for `job_category` is just 288 which is less not even 1% of the dataset hence, we have not omitted it but merged it with the new category unemployed in the next step.

- More than 50% of the clients had only secondary education and there are 1857 unknowns in the `education_level` category.At this stage we will not omit the unknowns as they are marginal in number.
- 60% of the individuals are married.
- Since 81% of the outcome is unknown for the previous campaign `previous_outcome`, therefore its not a predictive feature and we will do away with it.
- We reclassified the job-category to reduce the cardinality by merging unemployed, retired , student and unknown into unemployed. Entrepreneurs can be include in self-employed category. We created new feature to keep the original features intact as this helps in clear distinction and increases the granularity during the model building phase.

```
bank <- bank %>% mutate(
  job_category_new = ifelse( grepl('entrepreneur',job_category),"self-employed",
                      ifelse( grepl("housemaid",job_category), 'services',
ifelse(job_category %in% c("unemployed","retired","student","unknown"),"unemployed",
       job_category))))
```

- In the original dataset, 81% of the times customer was not contacted earlier (numerical representation as -1) therefore we will simplify the approach and create a categorical variable if customer was contacted earlier or not after the previous campaign.

```
bank <- bank %>% mutate(previous_contacted_new =
                          ifelse(previous_contact_days == "-1", "no", "yes"))

bank$previous_contact_days <- NULL
```

- Convert months into categorical variable with ordered levels .

```
bank$contact_month <- factor(bank$contact_month, levels = c("jan","feb","mar","apr","may",
                                                            "jun","jul","aug","sep","oct",
                                                            "nov","dec"),ordered = TRUE)
```

- Lastly, we converted all character features into factor .

```
bank[ ,sapply(bank,is.character)] <- lapply(bank[,sapply(bank,is.character)], factor)
```

- Table 2 presents the summary statistics after data-preprocessing.

```
summarizeColumns(bank) %>% kable( caption = 'Feature Summary after Data Preprocessing' )
```

Table 2: Feature Summary after Data Preprocessing

name	type	na	mean	disp	median	mad	min	max	nlevs
age	integer	0	40.9362102	10.6187620	39	10.3782	18	95	0
job_category	factor	0	NA	0.7847427	NA	NA	288	9732	12
marital_status	factor	0	NA	0.3980668	NA	NA	5207	27214	3
education_level	factor	0	NA	0.4868063	NA	NA	1857	23202	4
credit_default	factor	0	NA	0.0180266	NA	NA	815	44396	2
avg_balance	integer	0	1362.2720577	3044.7658292	448	664.2048	-8019	102127	0
housing_loan	factor	0	NA	0.4441618	NA	NA	20081	25130	2
personal_loan	factor	0	NA	0.1602265	NA	NA	7244	37967	2
contact_mode	factor	0	NA	0.3522594	NA	NA	2906	29285	3
contact_day	integer	0	15.8064188	8.3224762	16	10.3782	1	31	0
contact_month	ordered	0	NA	0.6955166	NA	NA	214	13766	12
last_contact_duration	integer	0	258.1630798	257.5278123	180	137.8818	0	4918	0
no_of_timescontacted	integer	0	2.7638407	3.0980209	2	1.4826	1	63	0
previous_no_of_contacts	integer	0	0.5803234	2.3034410	0	0.0000	0	275	0
previous_outcome	factor	0	NA	0.1825220	NA	NA	1511	36959	4
Term_deposit	factor	0	NA	0.1169848	NA	NA	5289	39922	2
job_category_new	factor	0	NA	0.7847427	NA	NA	3066	9732	7
previous_contacted_new	factor	0	NA	0.1826325	NA	NA	8257	36954	2

4 Data Exploration

We explored the data for each feature individually and split them by the classes of target features. Then we proceeded to multivariate visualisation.

4.1 Univariate and Bivariate Visualisation

4.1.1 Numerical Features

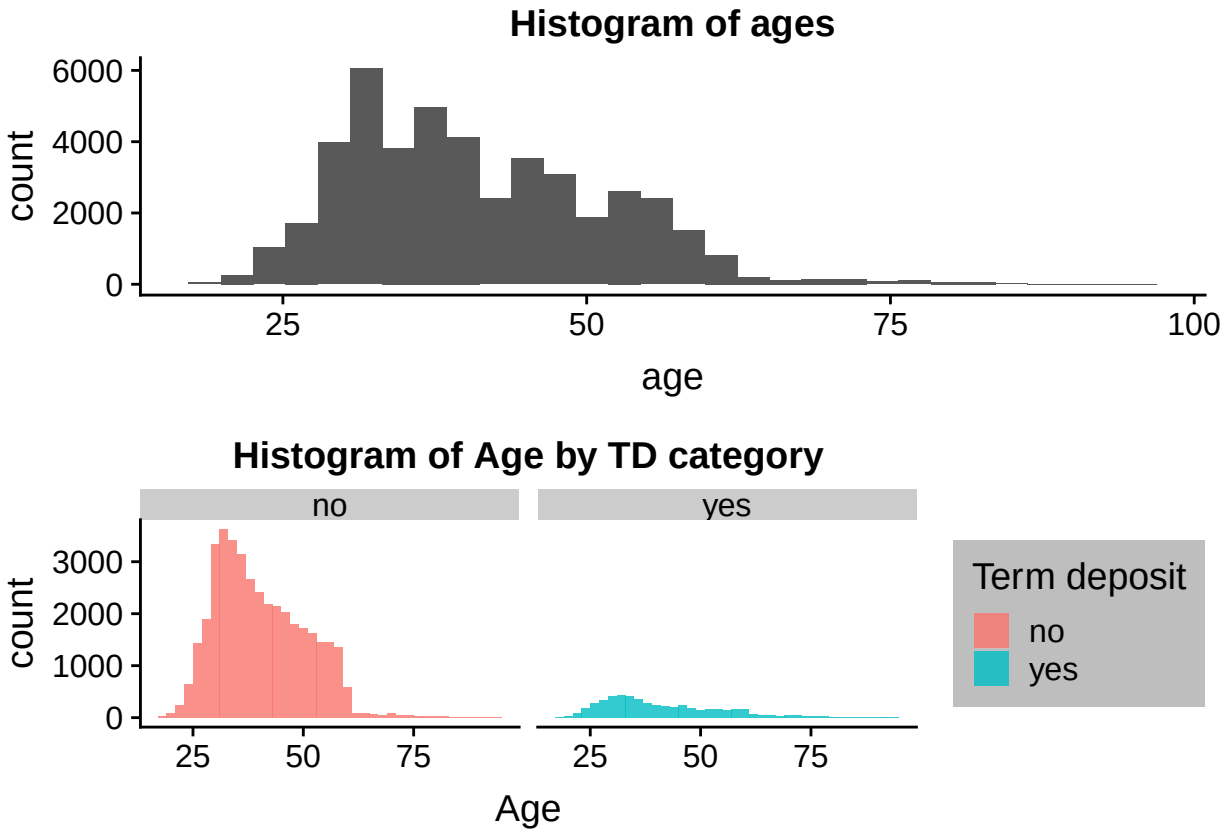
4.1.1.1 Age

Most of individuals aged between 30 and 45 with younger generation most likely subscribe for Term Deposit(TD). Clients who subscribed for TD appears to be following a normal distribution whereas those who didn't have a skewed distribution of age. Therefore, age can be a predictive feature.

```
p1 <- ggplot(data = bank,aes(x=age)) + geom_histogram(bins=30) + labs(title="Histogram of ages")

p2 <- ggplot(data = bank, aes(x=age, fill=Term_deposit))+
  facet_grid(~Term_deposit) +geom_histogram(alpha=0.8,binwidth = 2) +
  labs(title="Histogram of Age by TD category", x="Age", y="count")+
  theme(legend.background = element_rect(fill="grey", size=0.5, linetype="solid"),
        legend.text = element_text())+ guides(fill=guide_legend(title="Term deposit"))

plot_grid(p1, p2, ncol = 1)
```



4.1.1.2 Average yearly balance

As shown by the stacked histogram below, the distribution of the both clients who subscribed or not appears rightly skewed. Few enteries in box plot might appear like an outlier - but we can't omit them as its possible for few individuals to have very high or even negative yearly average balance as banks in Portugal do allow -ve balance.Also the median of yearly balance is almost same for both the categories.

```
summary(bank$avg_balance)
```

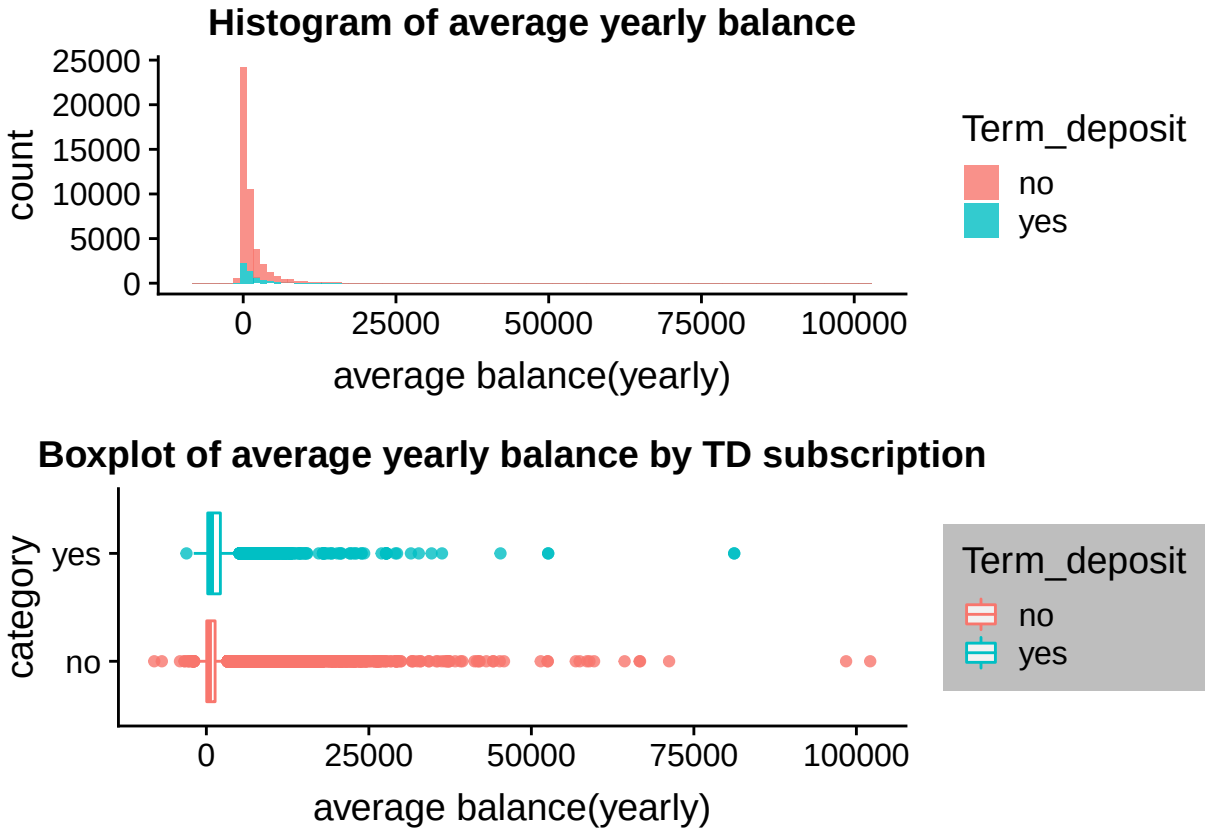
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -8019      72     448    1362    1428  102127
```

```
p3 <- ggplot(data = bank, aes(x=avg_balance, fill=Term_deposit))+
  geom_histogram(bins = 100,alpha=0.8)+labs(title="Histogram of average yearly balance",
                                            x = "average balance(yearly)")

p4 <- ggplot(data = bank, aes(y= avg_balance, x=Term_deposit,
                              col=Term_deposit)) +geom_boxplot(alpha=0.8)+
  labs(title="Boxplot of average yearly balance by TD subscription",x="category",
       y = "average balance(yearly)")+
  theme(legend.background = element_rect(fill="grey", size=0.5, linetype="solid"),
        legend.text = element_text())+ guides(fill=guide_legend(title="Term deposit"))

p4 <- p4 + coord_flip()

plot_grid(p3, p4, ncol = 1)
```



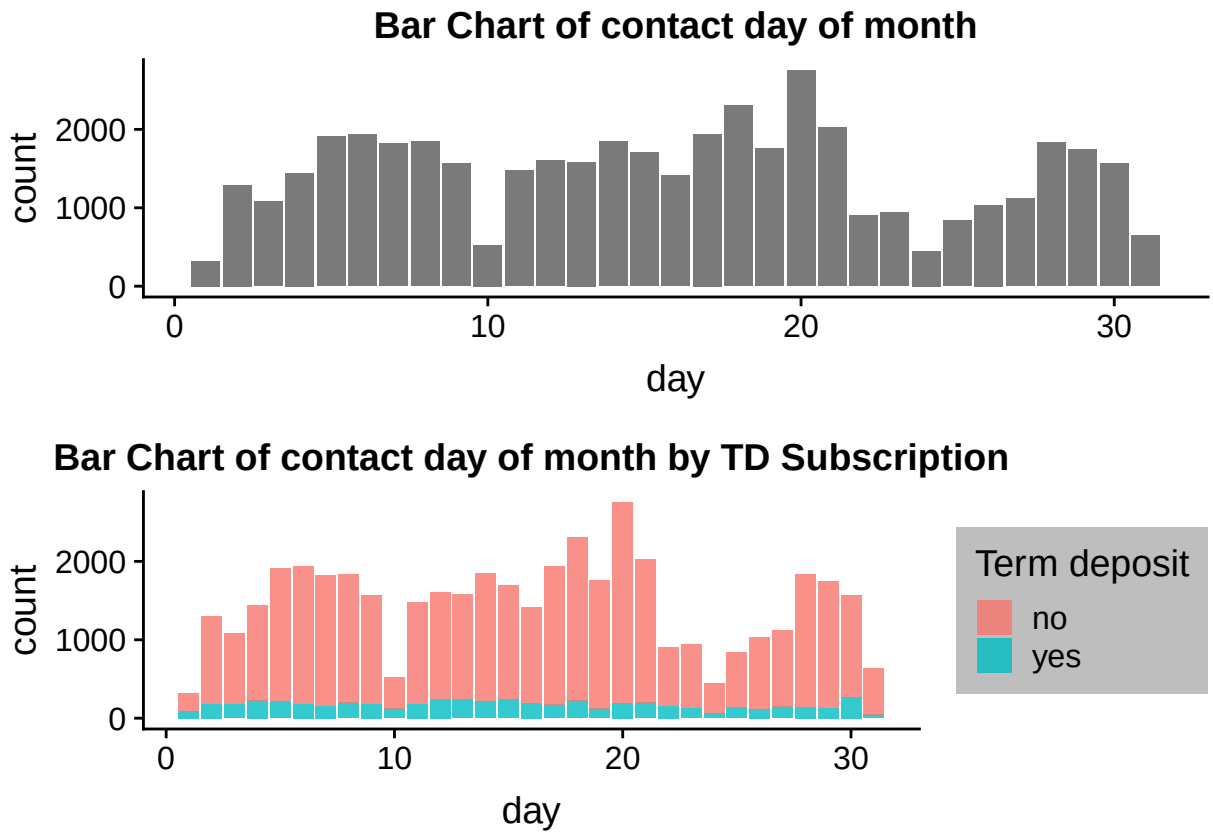
4.1.1.3 Contact day of the month

It appears that on 20th of each month maximum number of contacts are made. But for the clients who subscribe to the Term Deposit, distribution appears pretty even throughout the month. 30th day of the month has the maximum conversation rate (around 5%) which might be owing to fact that customers sign up wanting the interest to start immediately from next month.

```
p5 <- ggplot(data = bank, aes(x = contact_day)) + geom_bar(alpha=0.8)+
  labs(title="Bar Chart of contact day of month", x="day")
p6 <- ggplot(data = bank, aes(x = contact_day, fill=Term_deposit)) + geom_bar(alpha=0.8)+
  labs(title="Bar Chart of contact day of month by TD Subscription", x="day") +
  theme(legend.background = element_rect(fill="grey", size=0.5, linetype="solid"),
```

```
        legend.text = element_text())+ guides(fill=guide_legend(title="Term deposit"))

plot_grid(p5, p6, ncol = 1)
```



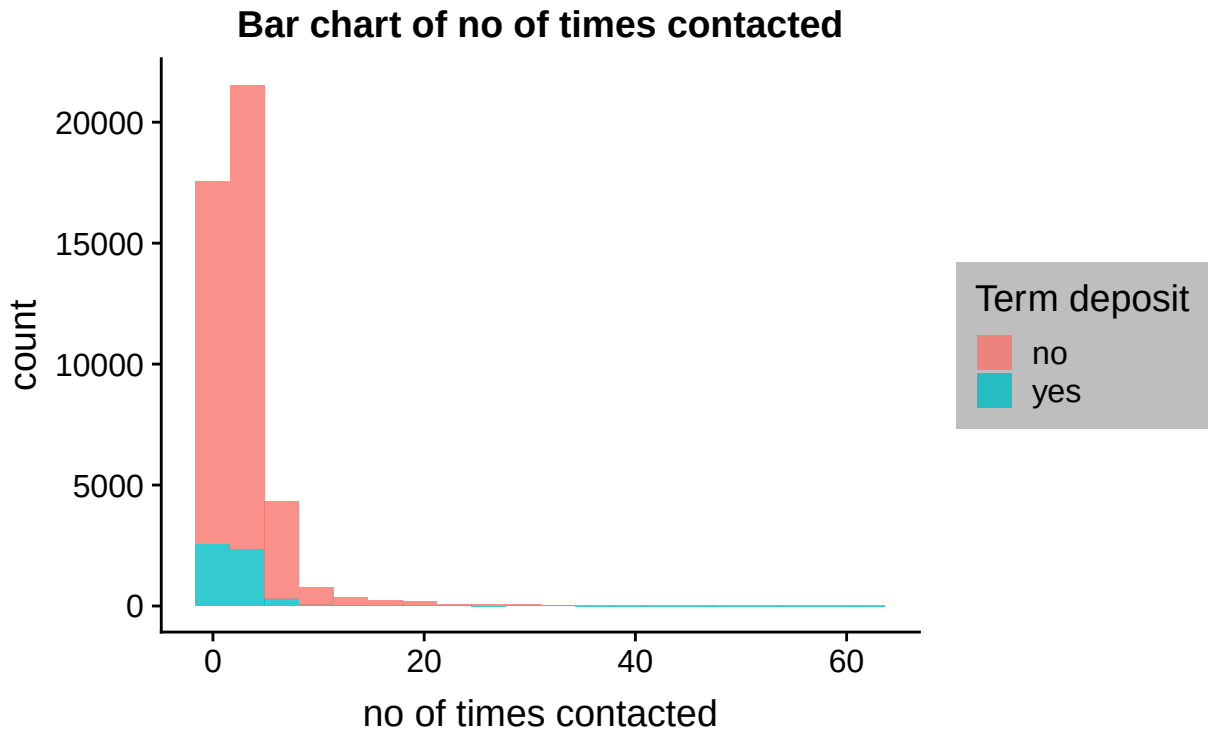
4.1.1.4 Number of contacts performed during this campaign

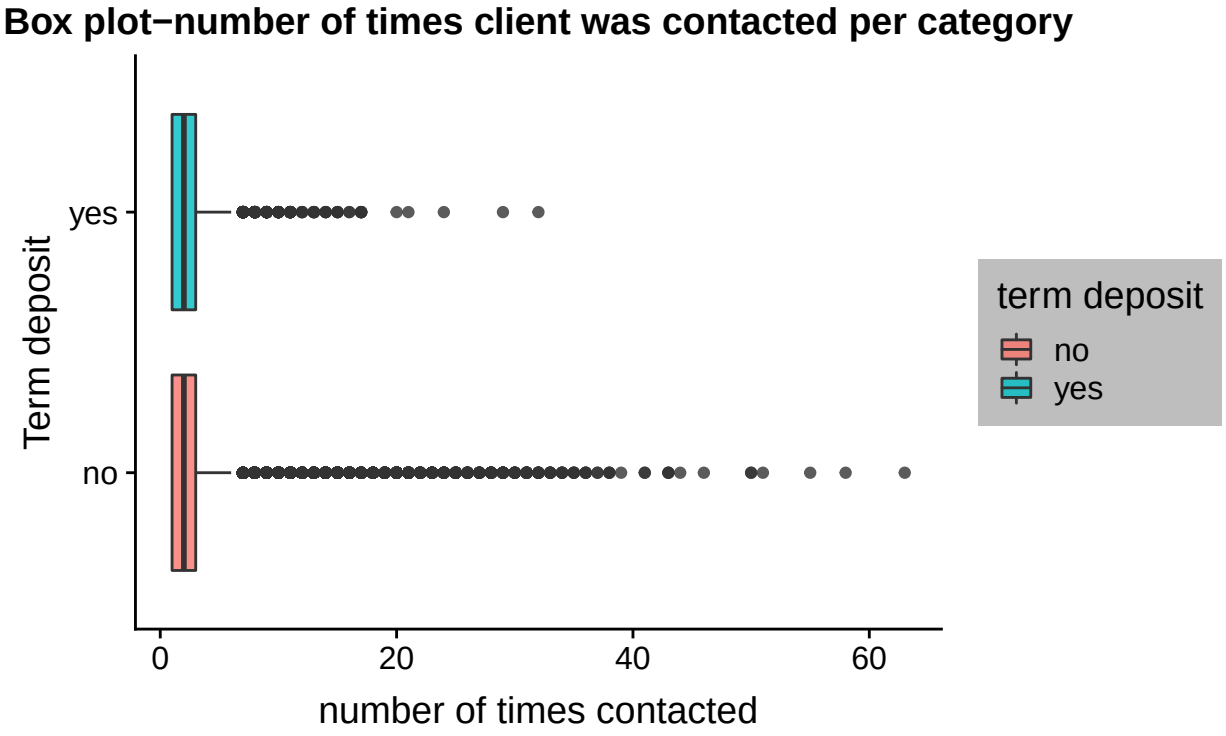
The distribution for both categories of client was rightly skewed as seen from the histogram. From the boxplot we can observe that median for customers who subscribed and those who didnt is same at 2 calls.

```
p7 <- ggplot(data = bank, aes(x = no_of_timescontacted, fill= Term_deposit))+
  geom_histogram(bins=20,alpha=0.8)+
  labs(title="Bar chart of no of times contacted", x= "no of times contacted")+
  theme(legend.background = element_rect(fill="grey", size=0.5, linetype="solid"),
        legend.text = element_text())+ guides(fill=guide_legend(title="Term deposit"))


p8 <- ggplot(data = bank, aes( x =Term_deposit, y = no_of_timescontacted,
                          fill=Term_deposit)) + geom_boxplot(alpha=0.8)+
  labs(title="Box plot-number of times client was contacted per category",
       y="number of times contacted", x="Term deposit") +
  theme(legend.background = element_rect(fill="grey", size=0.5, linetype="solid"),
        legend.text = element_text())+ guides(fill=guide_legend(title="term deposit"))+
  coord_flip()

plot_grid(p7, p8, ncol = 1)
```


Bar chart of no of times contacted
20000
15000
10000
5000
0
count
0
20
40
60
no of times contacted
Term deposit
no
yes


Box plot−number of times client was contacted per category
yes
no
Term deposit
0
20
40
60
number of times contacted
term deposit
no
yes

We can also see that average number of calls for people who signed up is 2.14 against 2.84 for those who didnt signup, so making more calls doesn't equate to subscription.

```
bank %>% group_by(Term_deposit) %>% summarise(
  mean = mean(no_of_timescontacted),
  median = median(no_of_timescontacted))
```

```
## # A tibble: 2 x 3
##   Term_deposit  mean median
##   <fct>        <dbl>  <dbl>
## 1 no            2.85      2
## 2 yes           2.14      2
```

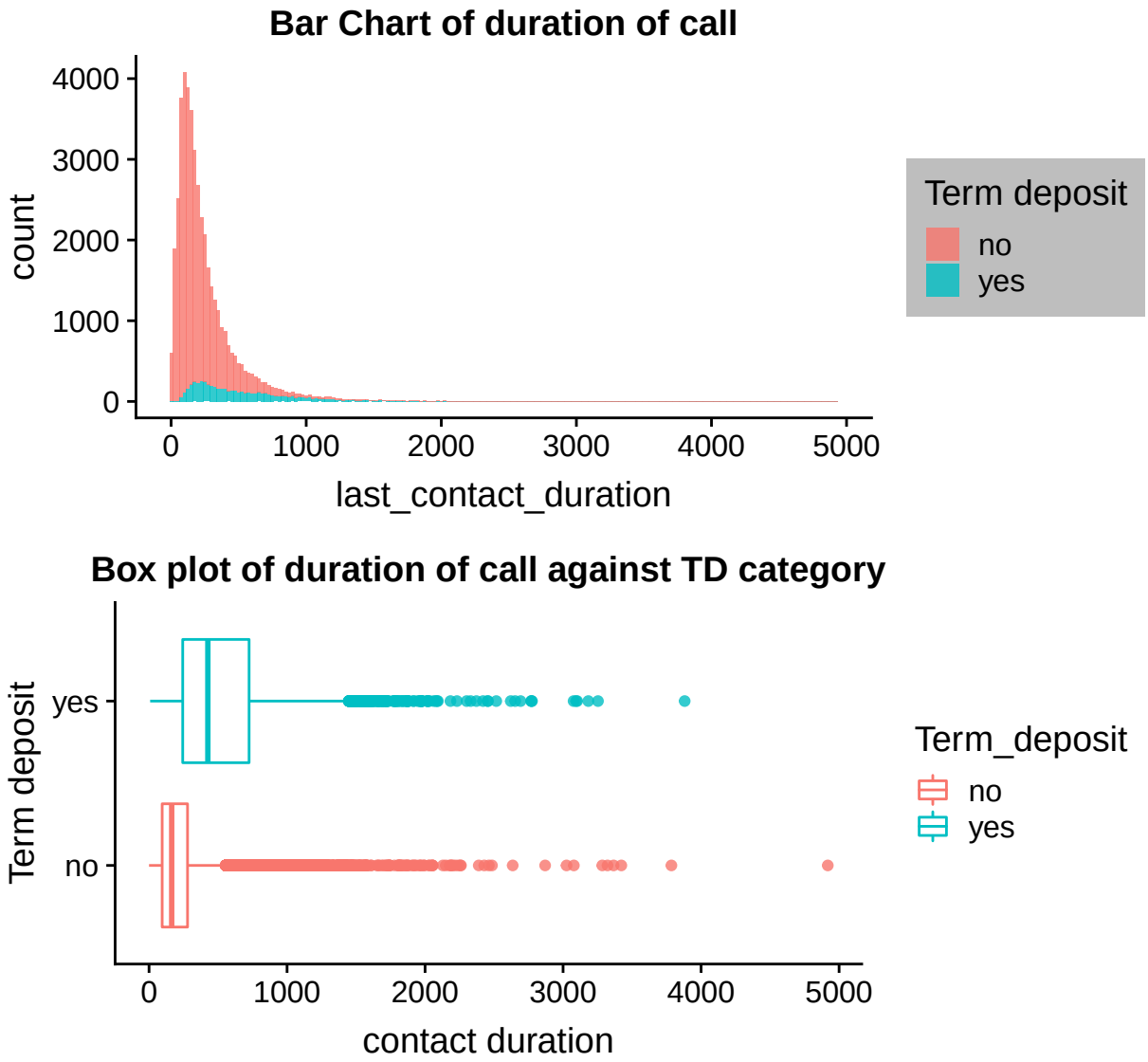
4.1.1.5 Last contact duration in secs :

The disctribution for both the categores is a bit skewed to right, but looking at the median in the boxplot we can clearly see that clients who subscribed to product had longer hours of discussion with the marketing agent.

```
p9 <- ggplot(data = bank, aes(x = last_contact_duration, fill= Term_deposit))+
  geom_histogram(binwidth = 25, alpha=0.8)+labs(title="Bar Chart of duration of call")+
  theme(legend.background = element_rect(fill="grey", size=0.5, linetype="solid"),
        legend.text = element_text())+ guides(fill=guide_legend(title="Term deposit"))

p10 <- ggplot(data = bank, aes(y = last_contact_duration,
      x=Term_deposit, col=Term_deposit)) + geom_boxplot(alpha=0.8) +
  labs(title="Box plot of duration of call against TD category",
  x="Term deposit", y= "contact duration") +coord_flip()

plot_grid(p9, p10, ncol = 1)
```



4.1.2 Categorical Features

Investigating the important categorical features along with a few newly created ones.

4.1.2.1 New Job Category

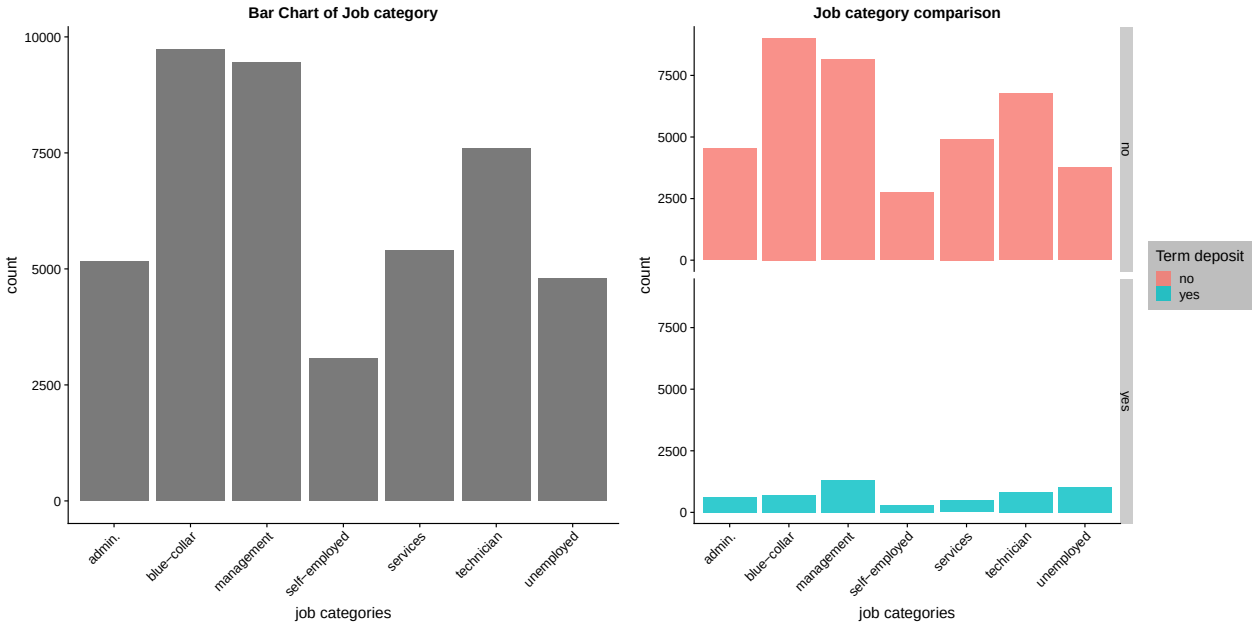
Most of the clients worked in management or were blue collar workers.Among the clients who subscribed from Term Deposit majority are from management (almost 25%)[Table 3], which makes it an important feature to predict.

```
p11 <- ggplot(data = bank, aes(x=job_category_new)) +geom_bar(alpha=0.8)+
  labs(title="Bar Chart of Job category", x="job categories")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1.0))

p12 <- ggplot(data = bank, aes(x=job_category_new, fill=Term_deposit))+
  geom_bar(alpha=0.8)+ facet_grid(Term_deposit~.)+
  labs(title="Job category comparison", x="job categories")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1.0))+
```

```
  theme(legend.background = element_rect(fill="grey", size=0.5, linetype="solid"),
        legend.text = element_text())+ guides(fill=guide_legend(title="Term deposit"))

plot_grid(p11, p12, ncol = 2)
```




```
T1 <- table(bank$Term_deposit, bank$job_category_new) %>% prop.table(margin = 1) *100
kable(T1, caption = "%agewise comparison for each job category")
```

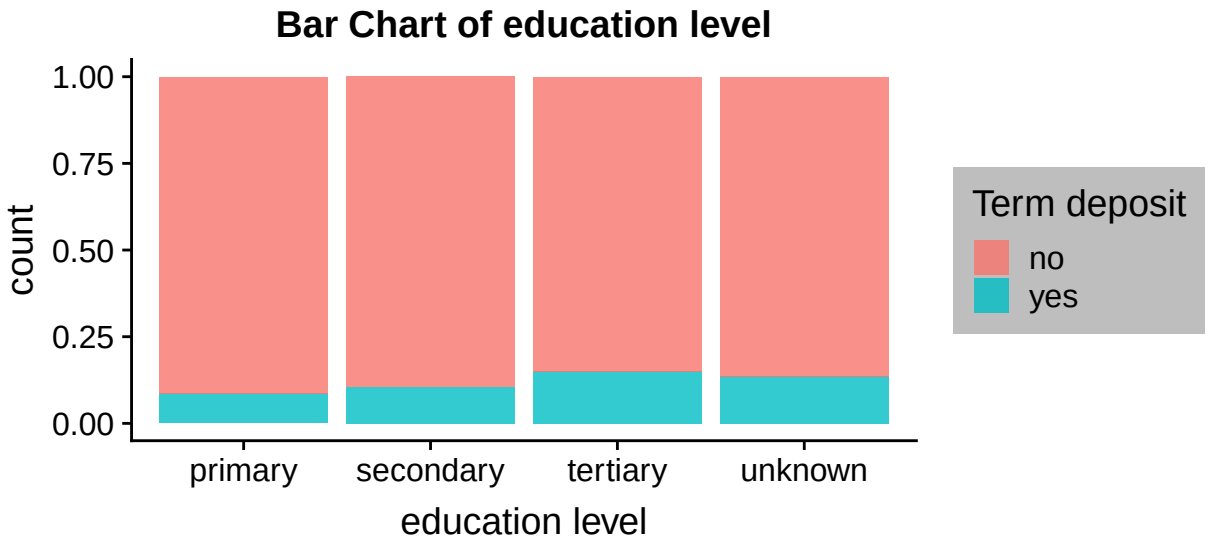
Table 3: %agewise comparison for each job category

	admin.	blue-collar	management	self-employed	services	technician	unemployed
no	11.37218	22.60408	20.43234	6.903462	12.314012	16.92550	9.448424
yes	11.93042	13.38627	24.59822	5.861221	9.037625	15.88202	19.304216

4.1.2.2 Education Level

Segregating education by Term deposit category showed that the clients who subscribed for TD were mostly from tertiary education level which might be owing to their high income level or being more informed, but we can't confirm as do not have sufficient data about income level of clients[Table 4].

```
ggplot(data = bank, aes(x= education_level, fill=Term_deposit))+
  geom_bar(alpha=0.8,position = "fill")+
  labs(title="Bar Chart of education level", x="education level")+
  theme(legend.background = element_rect(fill="grey", size=0.5, linetype="solid"),
        legend.text = element_text())+ guides(fill=guide_legend(title="Term deposit"))
```




```
T2 <- table(bank$Term_deposit, bank$education_level) %>% prop.table(margin = 2) *100
kable(T2, caption = "%agewise comparison for marital status")
```

Table 4: %agewise comparison for marital status

	primary	secondary	tertiary	unknown
no	91.373522	89.44057	84.99361	86.42973
yes	8.626478	10.55943	15.00639	13.57027

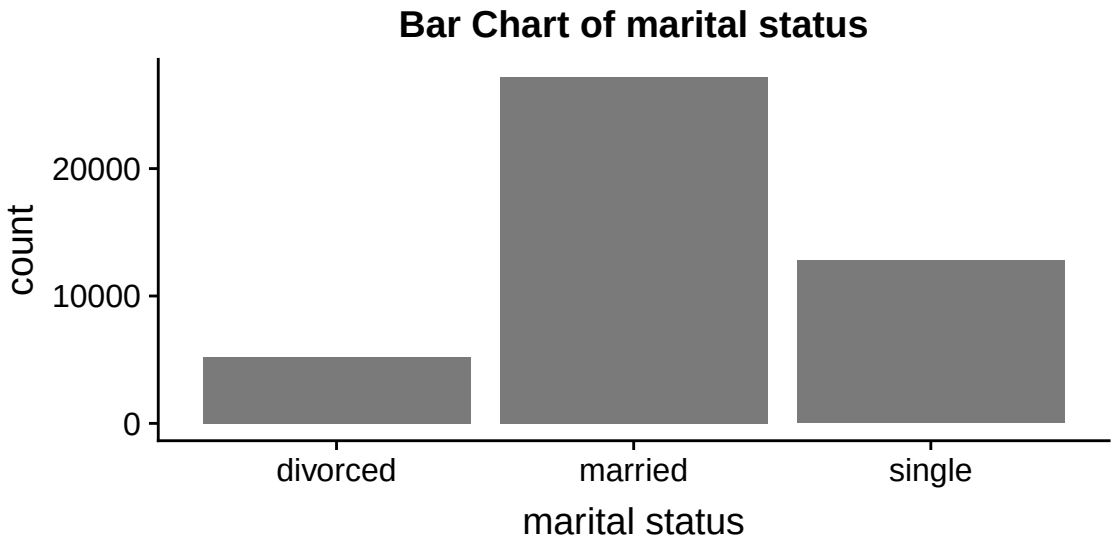
4.1.2.3 Marital Status

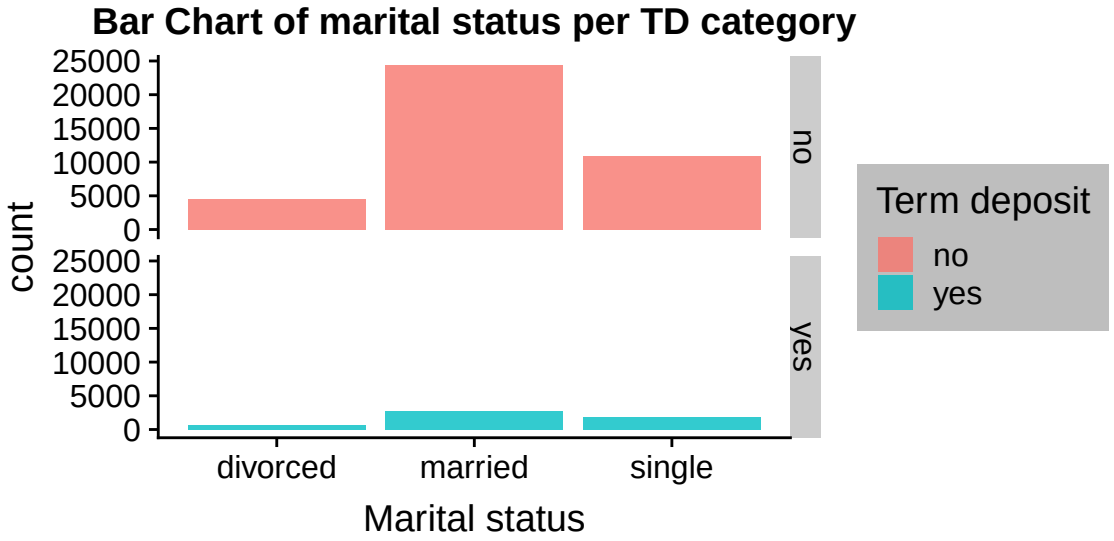
Most client were married. Also, there are more chances of a married person subscribing for term deposit compared to single to divorced which again is understandable.

```
p13 <- ggplot(data = bank, aes(x= marital_status)) +geom_bar(alpha=0.8)+
  labs(title="Bar Chart of marital status", x="marital status")

p14 <- ggplot(data = bank, aes(x=marital_status, fill=Term_deposit)) +geom_bar(alpha=0.8)+facet_grid(Te
  labs(title="Bar Chart of marital status per TD category", x="Marital status") +
theme(legend.background = element_rect(fill="grey", size=0.5, linetype="solid"),
legend.text = element_text())+ guides(fill=guide_legend(title="Term deposit"))

plot_grid(p13, p14, ncol = 1)
```





52% of married client subscribing to Term Deposit makes it an important predictive feature to consider.

```
T3 <- table(bank$Term_deposit, bank$marital_status) %>% prop.table(margin = 1) *100
kable(T3, caption = "%agewise comparison for marital status")
```

Table 5: %agewise comparison for marital status

	divorced	married	single
no	11.48490	61.26697	27.24813
yes	11.76026	52.08924	36.15050

4.1.2.4 Credit default

We did not display any visualtisation for credit default as 98% of the clients in dataset do not have credit default , so the feature doesnt help in prediction.

4.1.2.5 Contact mode

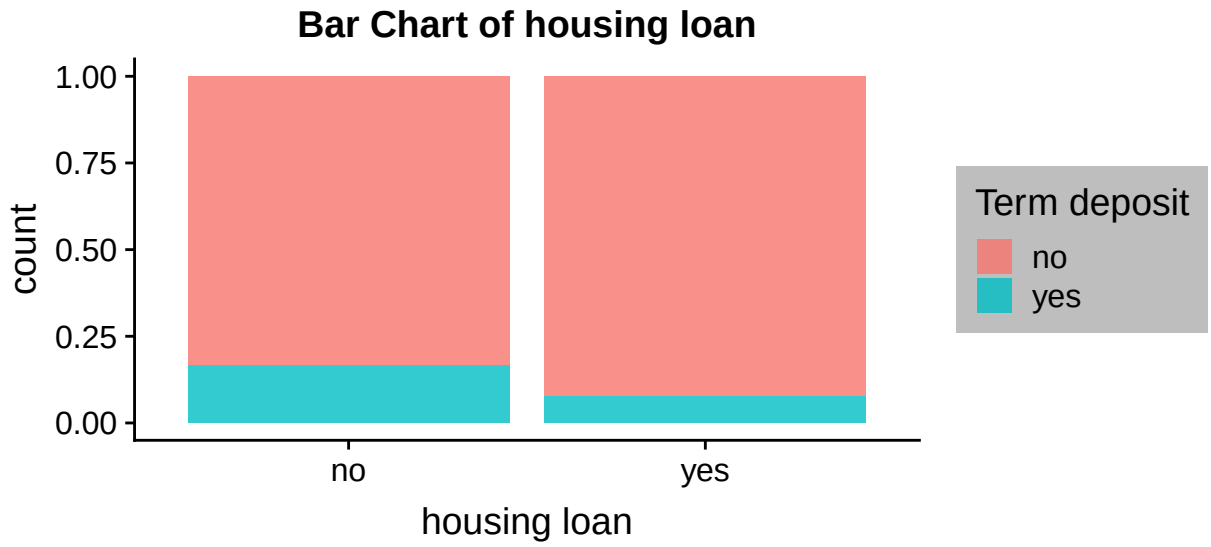
65% of the client is contacted on personal mobile and only 5% on telephone and rest is unknown which could be fax or email. Also the proportion of client subscribing to Term Deposit was found to be equally likely in both cell phones and telephone.Nothing can be deduced about the unknown mode. So we will ignore this feature.

```
bank$contact_mode <- NULL
```

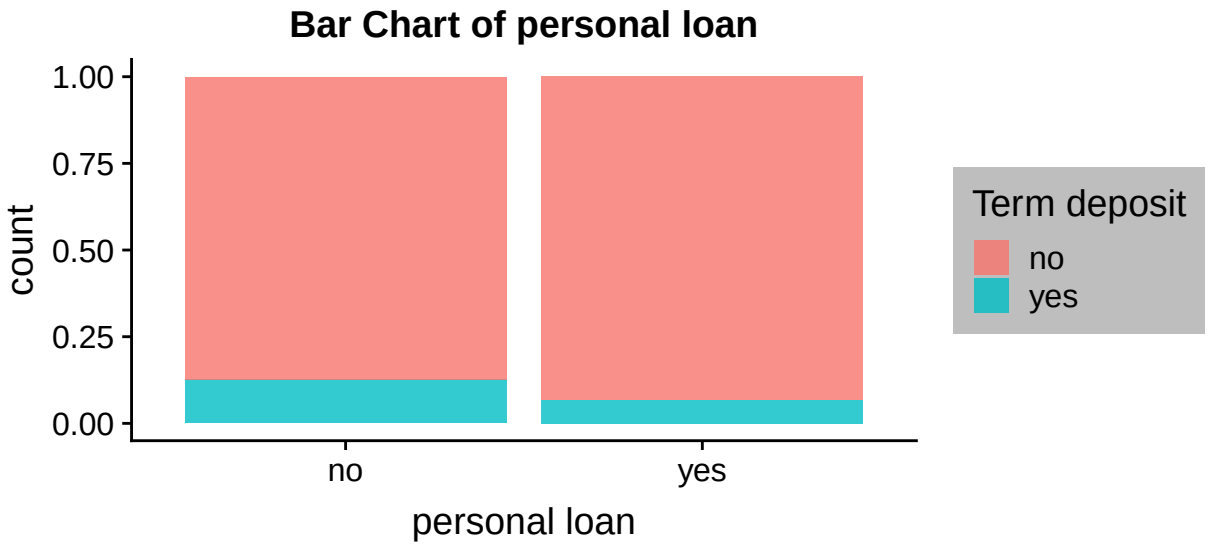
4.1.2.6 Housing Loan and Personal Loan

Client with no housing and personal loans are almost double the takers of term deposit than the ones with already a loan. We will explore more of this feature in multivariate analysis.

```
ggplot(data = bank, aes(x= housing_loan, fill=Term_deposit))+
geom_bar(alpha=0.8,position = "fill")+
  labs(title="Bar Chart of housing loan", x="housing loan") +
  theme(legend.background = element_rect(fill="grey", size=0.5, linetype="solid"),
        legend.text = element_text())+ guides(fill=guide_legend(title="Term deposit"))
```




```
ggplot(data = bank, aes(x= personal_loan, fill=Term_deposit))+
  geom_bar(alpha=0.8,position = "fill")+ labs(title="Bar Chart of personal loan", x="personal loan") +
  theme(legend.background = element_rect(fill="grey", size=0.5, linetype="solid"),
        legend.text = element_text())+ guides(fill=guide_legend(title="Term deposit"))
```



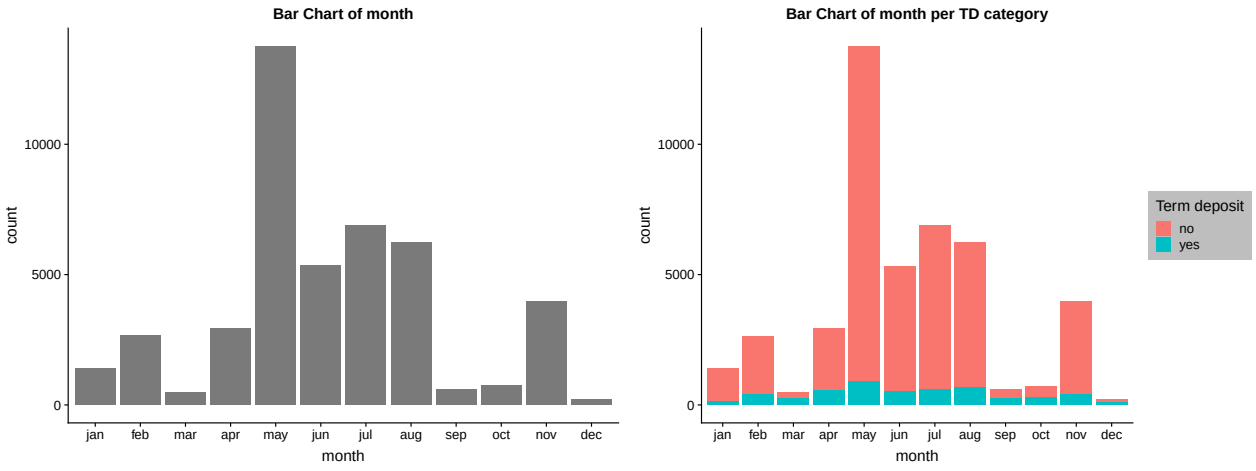
4.1.2.7 Month

May is the month when max contacts where made with the customers and we can clearly see the effect of it. 17.5% of the term deposits were sold in the month of May.

```
p15 <- ggplot(data = bank, aes(x = contact_month)) + geom_bar(alpha=0.8)+
  labs(title="Bar Chart of month", x="month")

p16 <- ggplot(data = bank, aes(x = contact_month, fill=Term_deposit)) + geom_bar()+
  labs(title="Bar Chart of month per TD category", x="month") +
  theme(legend.background = element_rect(fill="grey", size=0.5, linetype="solid"),
        legend.text = element_text())+ guides(fill=guide_legend(title="Term deposit"))

plot_grid(p15, p16, ncol = 2)
```



4.1.2.8 Previous contacted new

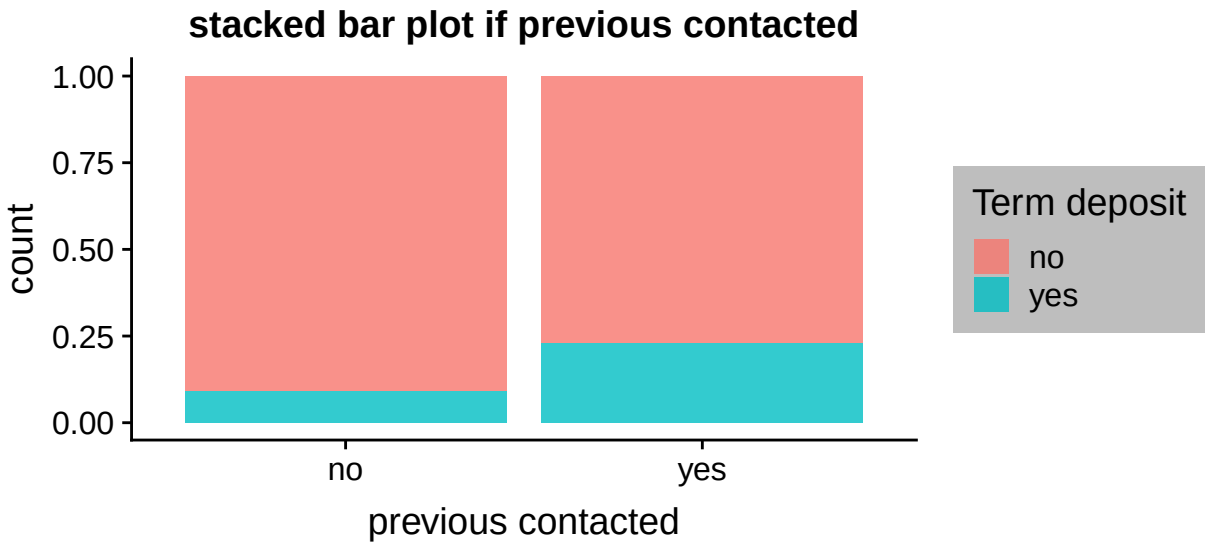
```
T10 <- table(bank$Term_deposit, bank$previous_contacted_new) %>% prop.table(margin = 1)*100
rownames(T10) <- c("TD_No:","TD_Yes:")
kable(T10, caption = "%agewise comparison for previous contact")
```

Table 6: %agewise comparison for previous contact

	no	yes
TD_No:	84.08897	15.91103
TD_Yes:	63.98185	36.01815

Around 64% of the times people signed up for Term Deposit if they had been contacted [Table6] which can also be seen in the stacked bar chart below.

```
ggplot(bank, aes(x=previous_contacted_new, fill = Term_deposit)) +
  geom_bar(position = "fill", alpha=0.8)+
  labs(title="stacked bar plot if previous contacted", x= "previous contacted")+
  theme(legend.background = element_rect(fill="grey", size=0.5, linetype="solid"),
        legend.text = element_text())+ guides(fill=guide_legend(title="Term deposit"))
```



4.1.2.9 outcome of the previous marketing campaign

On initial investigaion we found 81% of the outcome was unknown, therefore its better to remove this feature.

```
bank$previous_outcome <- NULL
```

4.2 Multivariate Visualisation

4.2.1 Proportional bar chart - Job category, Marital status and Education Level :

Across the marital status, *singles* with tertiary education were the ones who subscribed the most for Term Deposit. Across the job category *unemployed* clients mostly subscribed to TD which is a little unusual.Apart from that the distribution of TD[1] clients is pretty even across other job categories.

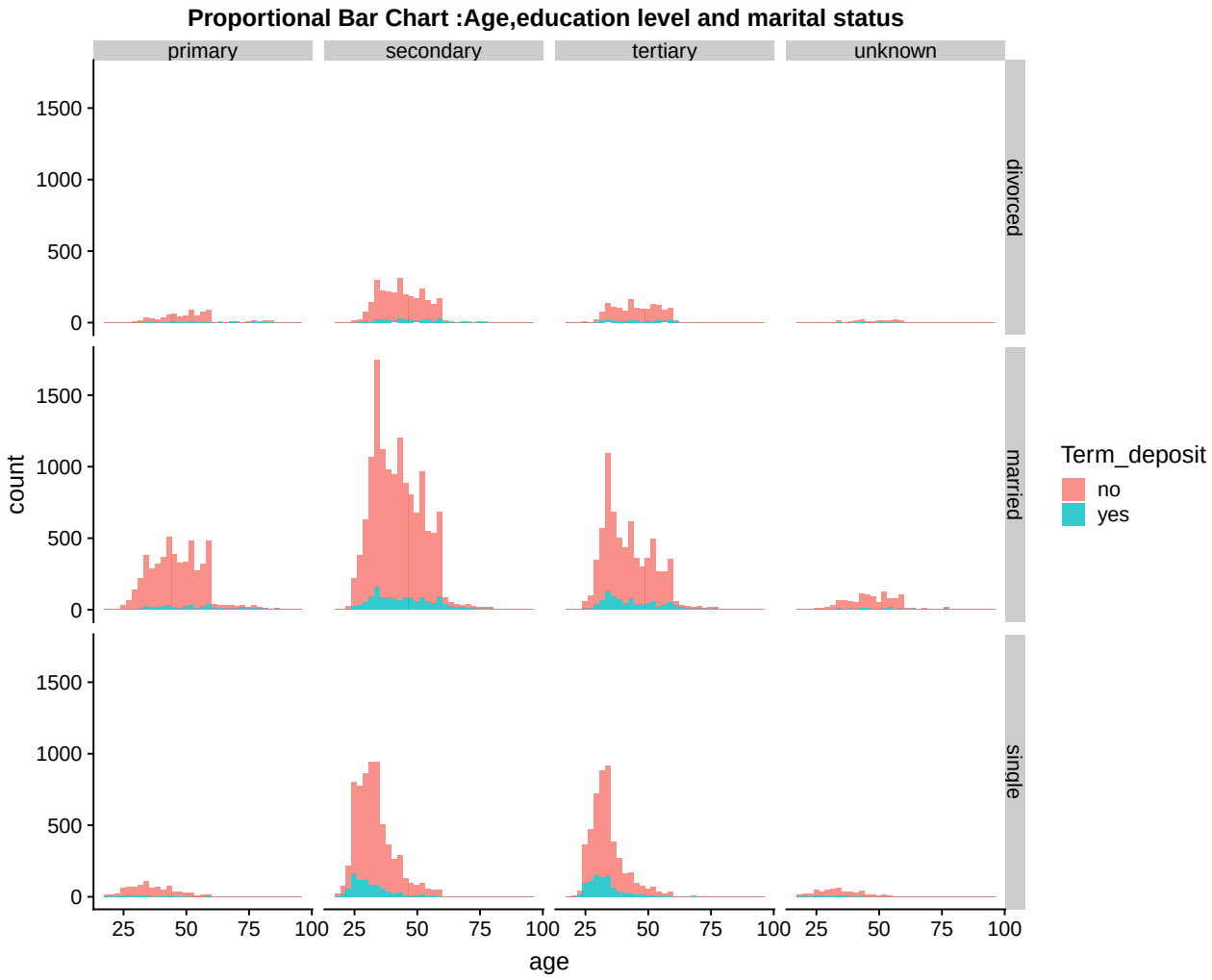
[1]TD stands for Term Deposit.

```
ggplot(bank, aes(x=education_level, fill=Term_deposit))+
  geom_bar(position = "fill", alpha=0.8)+labs(x="education level")+
  facet_grid(  job_category_new ~ marital_status )+
  labs(title="Proportional Bar Chart :Job category, Marital status and Education Level")+
  theme(legend.background = element_rect(fill="grey", size=0.5, linetype="solid"),
        legend.text = element_text())+ guides(fill=guide_legend(title="Term deposit"))
```

4.2.2 Age,education level and marital status

Because of the relatively few numbers, its hard to see the distribution in "unknown" category for singles , the distribution is rightly skewed across all education level and for married people with secondary and tertiary education its relatively symmetrical..

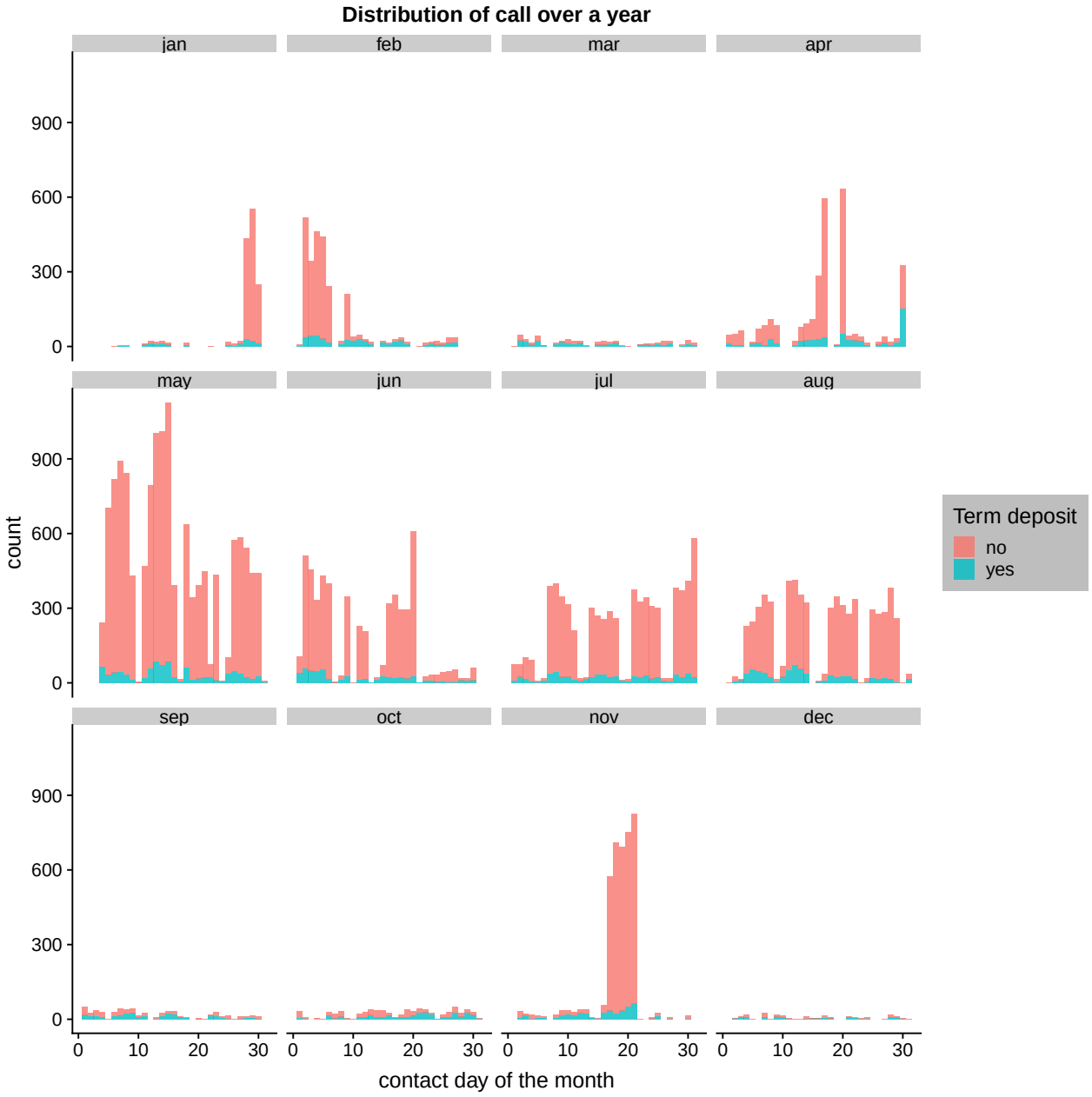
```
ggplot(bank, aes(x=age, fill=Term_deposit))+geom_histogram(bins= 35,alpha=0.8)+
  facet_grid( marital_status~ education_level) +
  labs(title="Proportional Bar Chart :Age,education level and marital status ")
```



4.2.3 Month, Days and Term Deposit subscription:

The stacked bars reveals that,april to aug is the time when most customers were contacted and most of the subscription happened. There is no repeated pattern in days of the months regarding subscription.

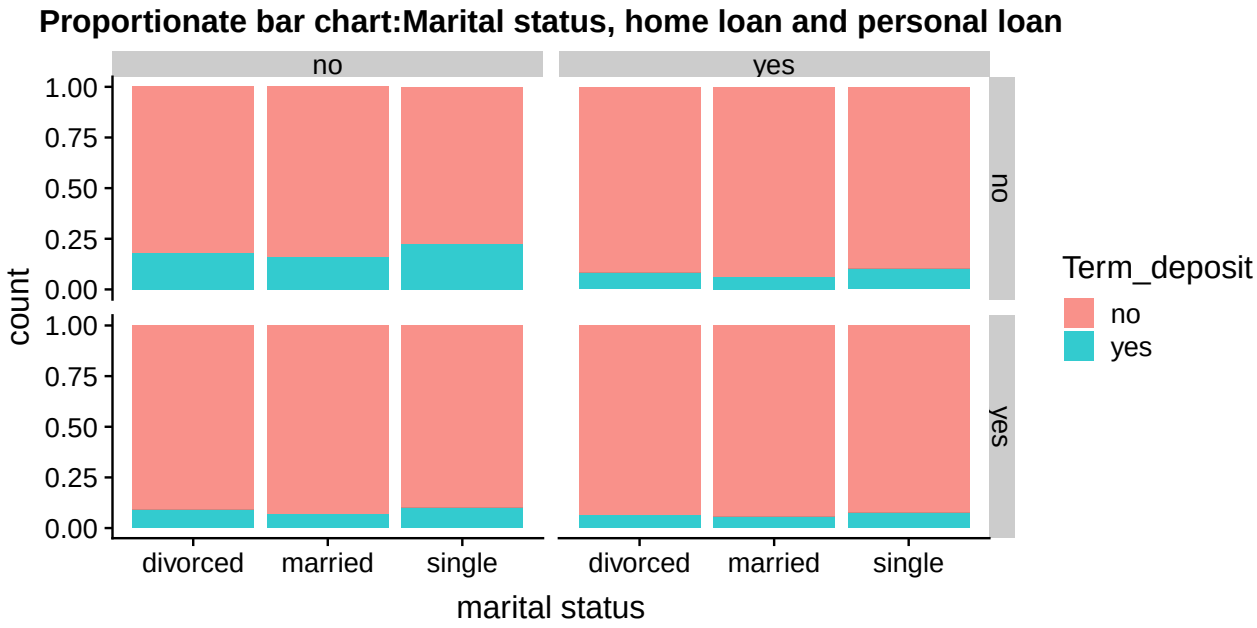
```
ggplot(bank,aes(x=contact_day, fill=Term_deposit)) + geom_bar(alpha=0.8)+
  facet_wrap(~contact_month)+ labs(title="Distribution of call over a year",
                                   x="contact day of the month")+
  theme(legend.background = element_rect(fill="grey", size=0.5, linetype="solid"),
        legend.text = element_text())+ guides(fill=guide_legend(title="Term deposit"))
```



4.2.4 Marital status, home loan and personal loan

As seen in the visualisation, singles without either a perosnal or home loan are the ones who subscribed the most to the term deposit.

```
ggplot(data = bank, aes(x= marital_status, fill=Term_deposit)) +
  labs(title="Proportionate bar chart:Marital status, home loan and personal loan",
       x="marital status")+
  geom_bar(position = "fill", alpha=0.8)+
  facet_grid(housing_loan ~ personal_loan)
```



5 Summary

We removed features which were either unknown or not predictive in nature like `previous_contact_days`, `contact_mode` and `previous_outcome`.We defined new features like`previous_contacted_new` and `job_category_new` which binned their corresponding original features into lower cardinalities but we did not remove the original features since we would like to experiment the model building by changing the granularity of the data. From the data exploration, we found that age, education levels, marital status, call duration and job category were potentially useful features in predicting the term deposit.