Stat 627 Project - Exploratory Data Analysis

William Shields and Esha Teware

2025-10-17

Data Loading and Summary

Overall we have four different csv files with data. For the purpose of EDA we are using the file with the most observations and variables.

```
bank <- read_delim("../data/bank-additional-full.csv", delim = ";",
                   show_col_types = FALSE) %>%
  clean_names() %>%
  mutate_if(is.character, as.factor)

head(bank)
```

```
# A tibble: 6 x 21
    age job    marital education default housing loan  contact month day_of_week
  <dbl> <fct>  <fct>   <fct>     <fct>   <fct>   <fct> <fct>   <fct> <fct>
1    56 house~ married basic.4y  no      no      no    teleph~ may   mon
2    57 servi~ married high.sch~ unknown no      no    teleph~ may   mon
3    37 servi~ married high.sch~ no      yes     no    teleph~ may   mon
4    40 admin. married basic.6y  no      no      no    teleph~ may   mon
5    56 servi~ married high.sch~ no      no      yes   teleph~ may   mon
6    45 servi~ married basic.9y  unknown no      no    teleph~ may   mon
# i 11 more variables: duration <dbl>, campaign <dbl>, pdays <dbl>,
#   previous <dbl>, poutcome <fct>, emp_var_rate <dbl>, cons_price_idx <dbl>,
#   cons_conf_idx <dbl>, euribor3m <dbl>, nr_employed <dbl>, y <fct>
```

Our dataset consists of 41188 observations of 21 variables. Ten of these variables are numeric, and the remaining 11 are categorical(this 11 includes the response). The response variable, y is whether or not the client of the bank will subscribe a term deposit. This data comes from a Portuguese banking institution.

Variable Overview

age: Age of contacted person

job: Job of contacted person

marital: Marital status of contacted person

education: Highest education of contacted person

$default$: If contacted has credit in default

housing: If contacted has a housing loan

loan: If contacted has personal loan

contact: Method of contact(telephone = landline)

month: Month of contact

day_of_week: Day of week contacted

duration: Length of contact in seconds (Note: highly effects value of y, should not be included in a model for realistic predictions)

campaign: number of contacts performed during the campaign for this client(includes last contact)

pdays: days passed since last contacted

previous: number of contacts performed before this campaign

poutcome: outcome of previous marketing campaign

emp_var_rate: employment variation rate, updated quarterly

$cons\_price\_index$: monthly average of consumer price index

$cons\_conf\_index$: monthly average of consumer confidence index

$euribor3m$: daily 3 month euro rate

$nr\_employed$: quarterly average of the total number of employed citizens

y: has the client subscribed a term deposit?

Missing values check

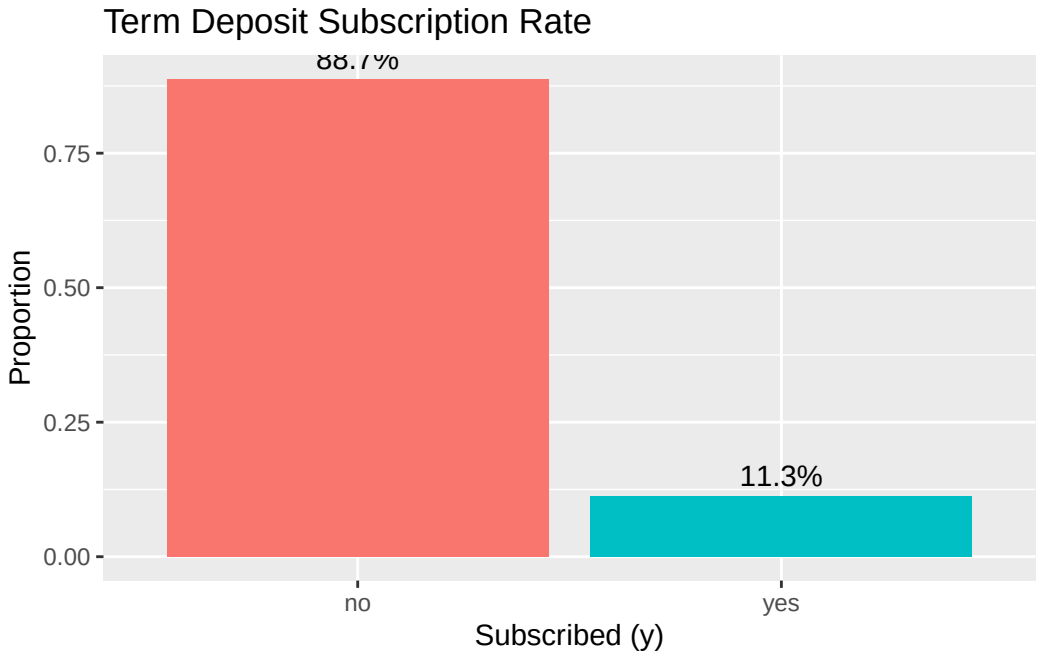
```r
bank %>%
  summarise(across(everything(), ~ sum(is.na(.)))) %>%
  pivot_longer(everything(), names_to = "variable", values_to = "missing_count") %>%
  arrange(desc(missing_count))
```

```
# A tibble: 21 x 2
   variable    missing_count
   <chr>               <int>
 1 age                     0
 2 job                     0
 3 marital                 0
 4 education               0
 5 default                 0
 6 housing                 0
 7 loan                    0
 8 contact                 0
 9 month                   0
10 day_of_week             0
# i 11 more rows
```

No missing values are present in the dataset.

Overview of the response

```r
bank %>%
  count(y) %>%
  mutate(Percent = n / sum(n)) %>%
  ggplot(aes(x = y, y = Percent, fill = y)) +
  geom_col(show.legend = FALSE) +
  geom_text(aes(label = scales::percent(Percent, accuracy = 0.1)), vjust = -0.5) +
  labs(title = "Term Deposit Subscription Rate", x = "Subscribed (y)", y = "Proportion")
```



A significant majority of the observations in the data have the value of *no* for the response variable.

Overview of Numeric Variables

```
numeric_vars <- bank %>% select(where(is.numeric))
skim(numeric_vars)
```

Table 1: Data summary

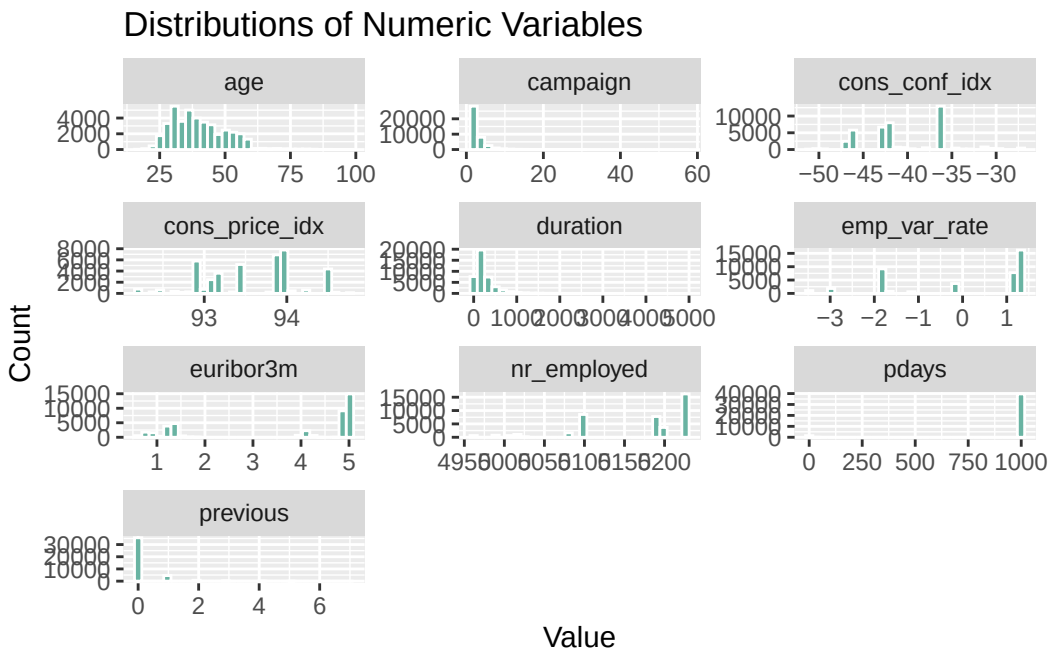
Name	numeric_vars
Number of rows	41188
Number of columns	10
_____	
Column type frequency:	
numeric	10
_____	
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1	40.02	10.42	17.00	32.00	38.00	47.00	98.00	
duration	0	1	258.29	259.28	0.00	102.00	180.00	319.00	4918.00	
campaign	0	1	2.57	2.77	1.00	1.00	2.00	3.00	56.00	
pdays	0	1	962.48	186.91	0.00	999.00	999.00	999.00	999.00	
previous	0	1	0.17	0.49	0.00	0.00	0.00	0.00	7.00	
emp_var_rate	0	1	0.08	1.57	-3.40	-1.80	1.10	1.40	1.40	
cons_price_idx	0	1	93.58	0.58	92.20	93.08	93.75	93.99	94.77	
cons_conf_idx	0	1	-	4.63	-	-	-	-	-	
			40.50		50.80	42.70	41.80	36.40	26.90	
euribor3m	0	1	3.62	1.73	0.63	1.34	4.86	4.96	5.04	
nr_employed	0	1	5167.04	72.25	4963.60	5099.10	5191.00	5228.10	5228.10	

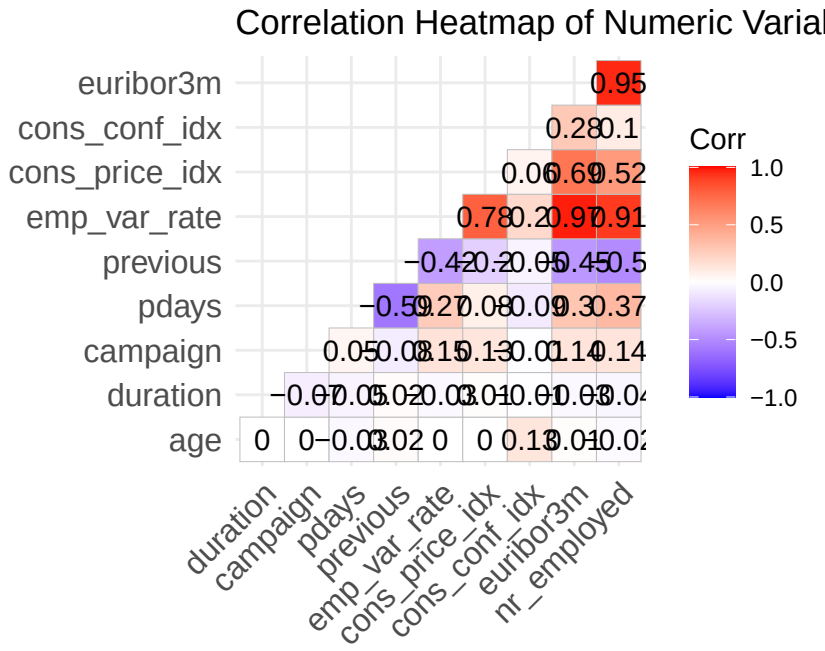
Histograms of Numeric Variables

```
numeric_vars %>%
  pivot_longer(cols = everything(), names_to = "variable", values_to = "value") %>%
  ggplot(aes(x = value)) +
  geom_histogram(bins = 30, fill = "#69b3a2", color = "white") +
  facet_wrap(~variable, scales = "free", ncol = 3) +
  labs(title = "Distributions of Numeric Variables", x = "Value", y = "Count")
```



Correlation of Numeric Variables

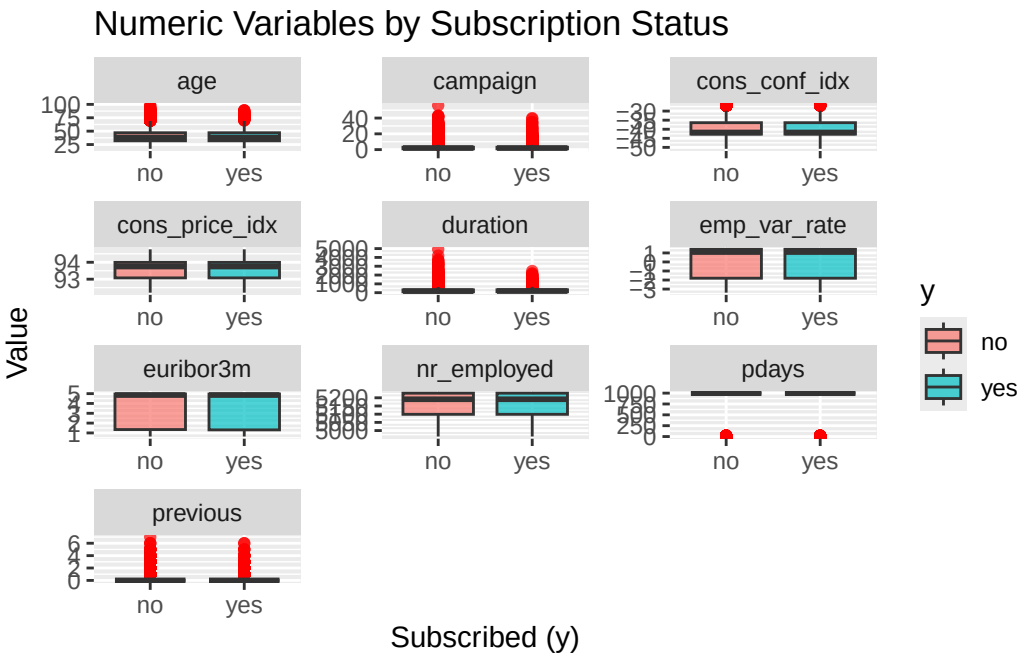
```
cor_matrix <- cor(numeric_vars)
ggcorrplot(cor_matrix, type = "lower", lab = TRUE,
           title = "Correlation Heatmap of Numeric Variables")
```



The pairs of, *euribor3m* and *nr_employed*, *euribor3m* and *emp_var_rate*, *nr_employed* and *emp_var_rate* are highly correlated. The values for these variables are the same across large groups of observations as they are economic indicators that change on either a quarterly(for *nr_employed* and *emp_var_rate*) or daily(for *euribor3m*) frequency.

Numeric Variables vs Response

```
numeric_vars %>%
  pivot_longer(cols = everything(), names_to = "variable", values_to = "value") %>%
  mutate(y = rep(bank$y, times = length(numeric_vars))) %>%
  ggplot(aes(x = y, y = value, fill = y)) +
  geom_boxplot(alpha = 0.7, outlier.color = "red") +
  facet_wrap(~variable, scales = "free", ncol = 3) +
  labs(title = "Numeric Variables by Subscription Status", x = "Subscribed (y)", y = "Value")
```

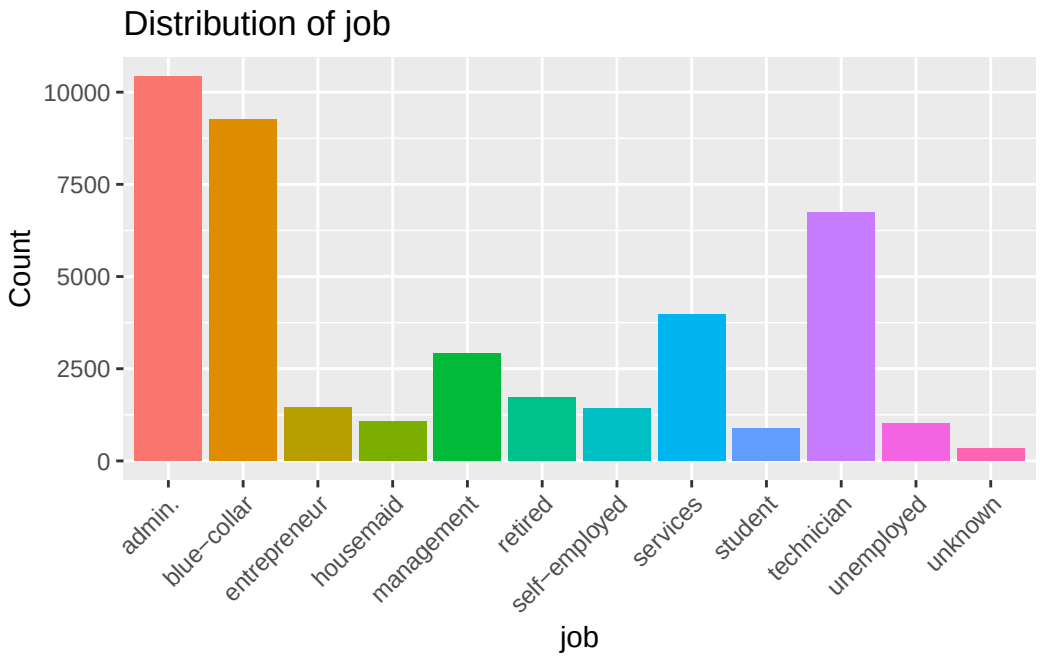


Overview of Categorical Variables

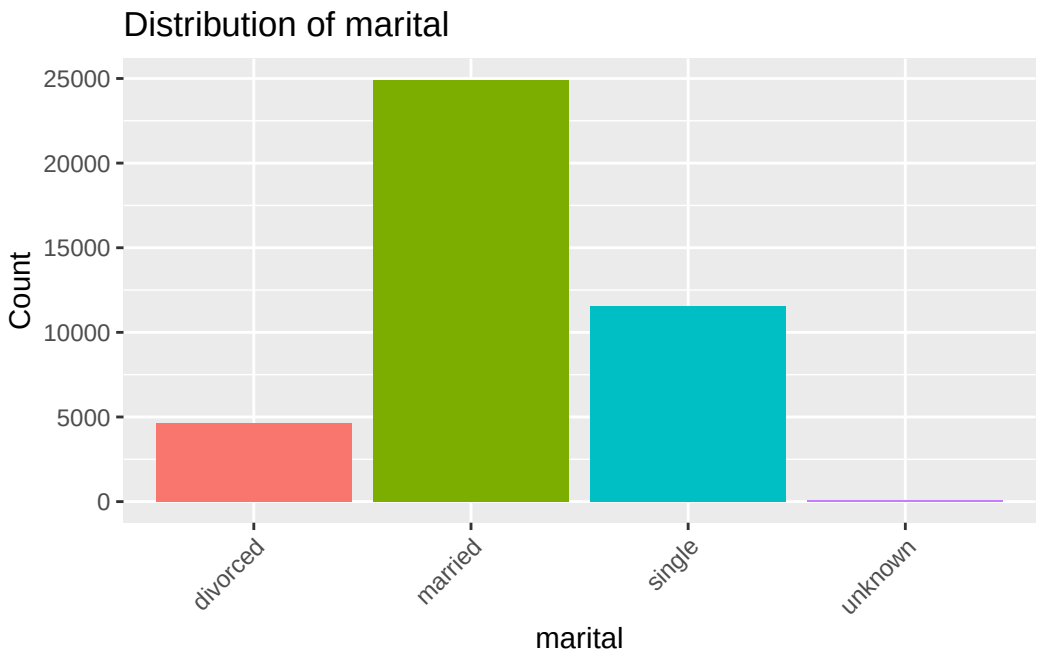
Frequency Plots of Categorical Variables

```
cat_vars <- bank %>% select(where(is.factor)) %>% select(-y)

cat_plot <- function(var) {
  ggplot(bank, aes(x = .data[[var]], fill = .data[[var]])) +
    geom_bar(show.legend = FALSE) +
    labs(title = paste("Distribution of", var), x = var, y = "Count") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
}

map(cat_vars %>% names(), cat_plot)
```

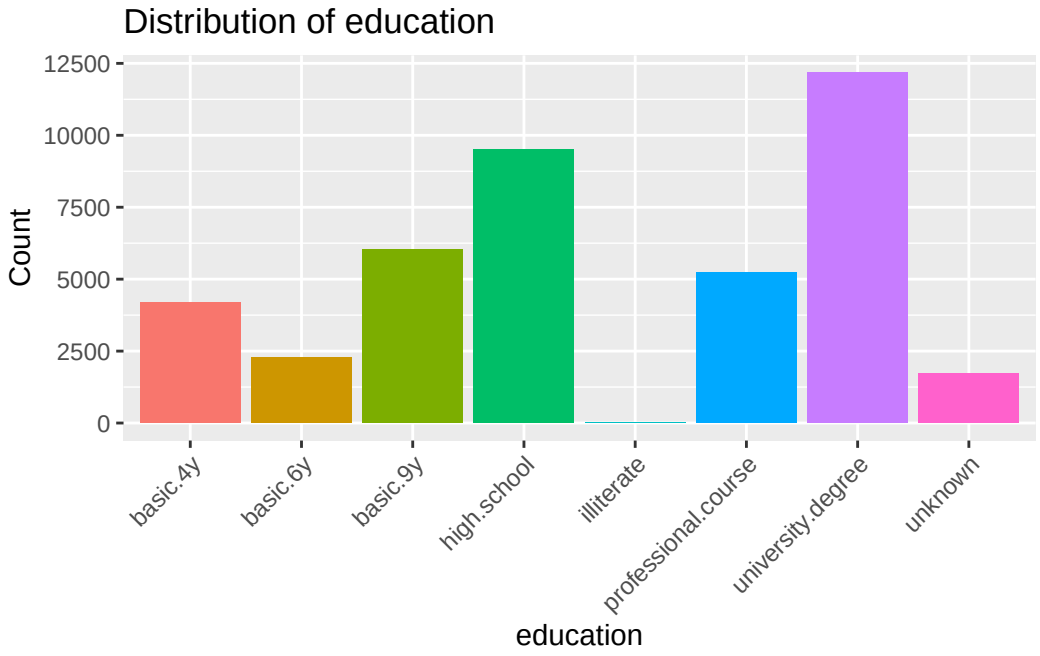
```
[[1]]
```


Distribution of job
Count
10000
7500
5000
2500
0
admin.
blue−collar
entrepreneur
housemaid
management
retired
self−employed
services
student
technician
unemployed
unknown
job

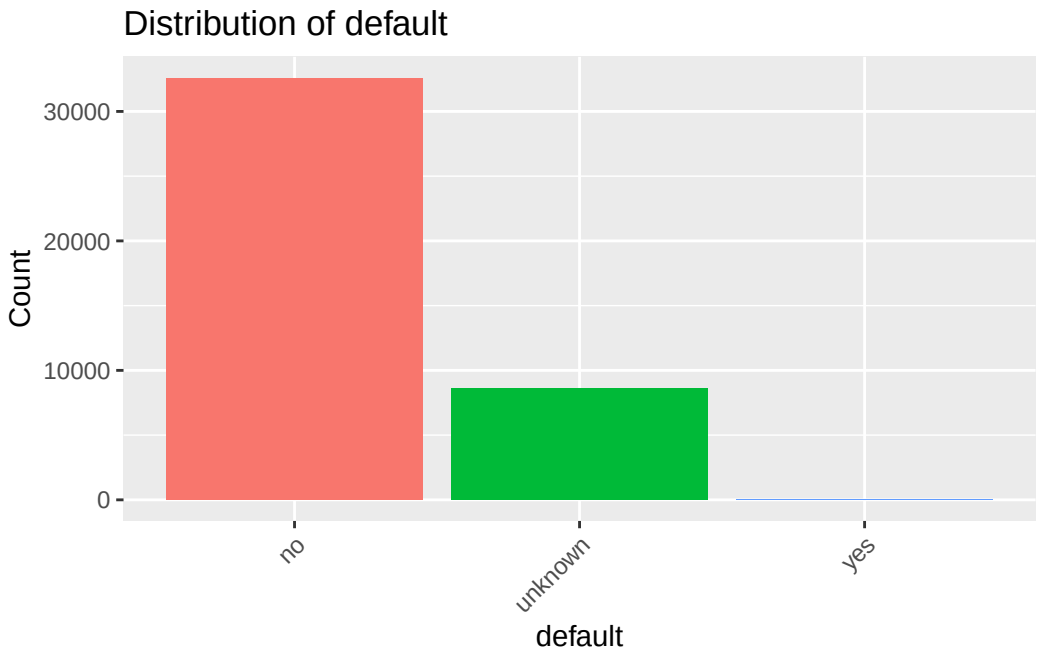
[[2]]


Distribution of marital
Count
25000
20000
15000
10000
5000
0
divorced
married
single
unknown
marital

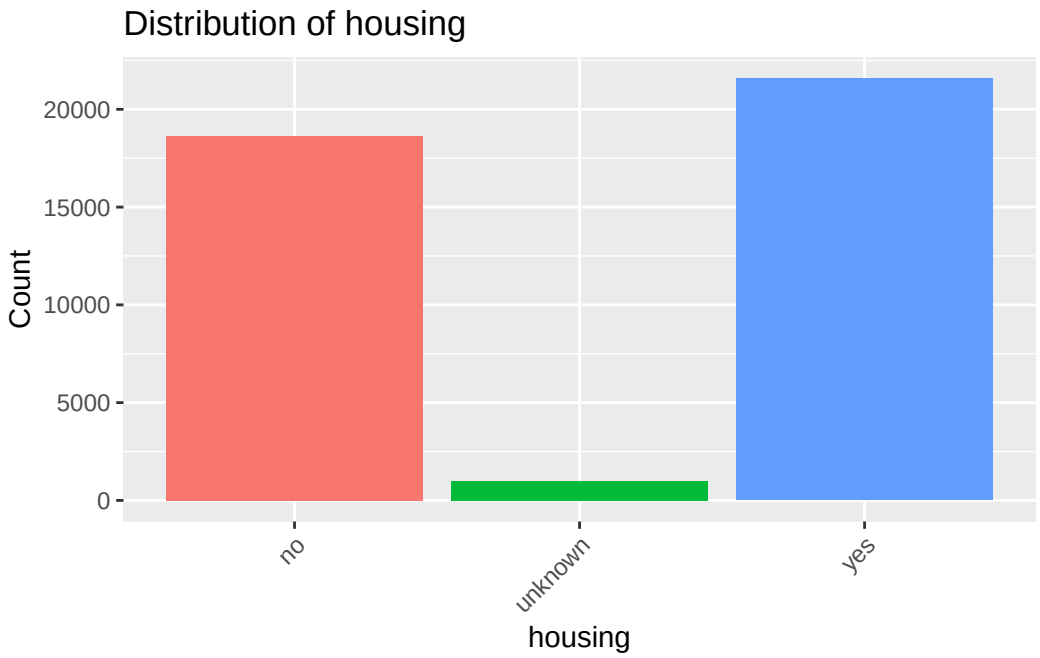
[[3]]


Distribution of education
Count
12500
10000
7500
5000
2500
0
basic.4y
basic.6y
basic.9y
high.school
illiterate
professional.course
university.degree
unknown
education

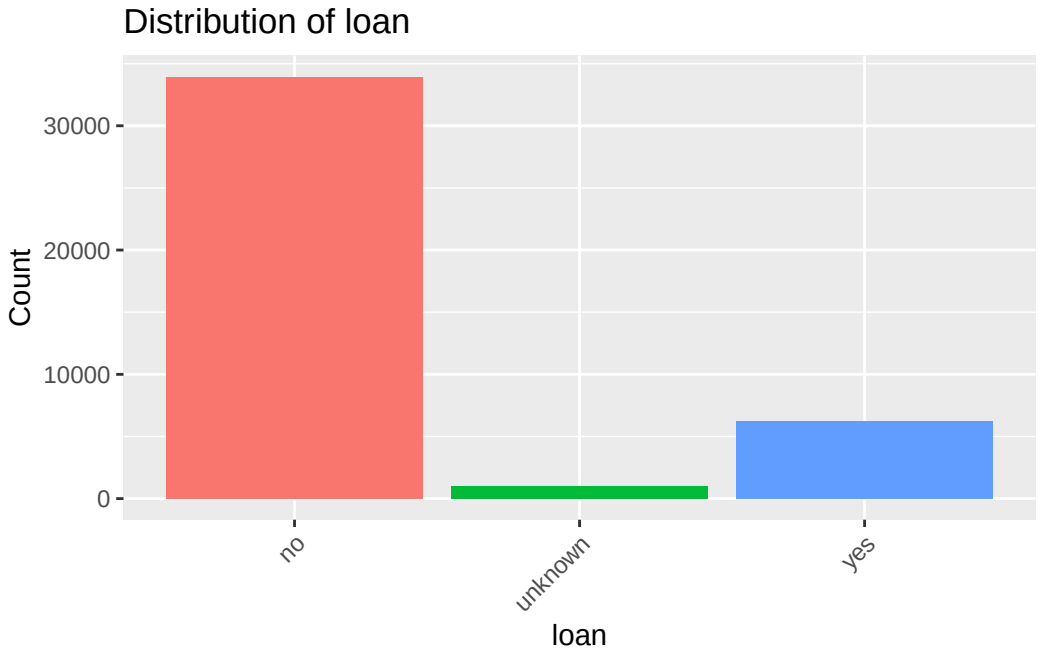
```
[[4]]
```


Distribution of default
Count
30000
20000
10000
0
no
unknown
yes
default

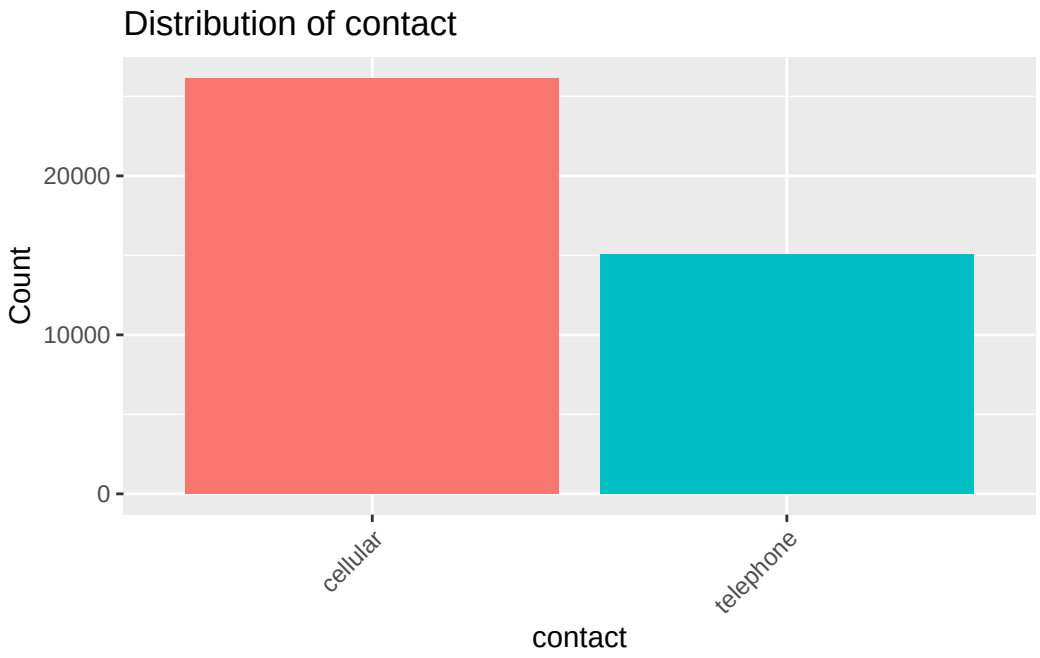
```
[[5]]
```


Distribution of housing
Count
20000
15000
10000
5000
0
no
unknown
yes
housing

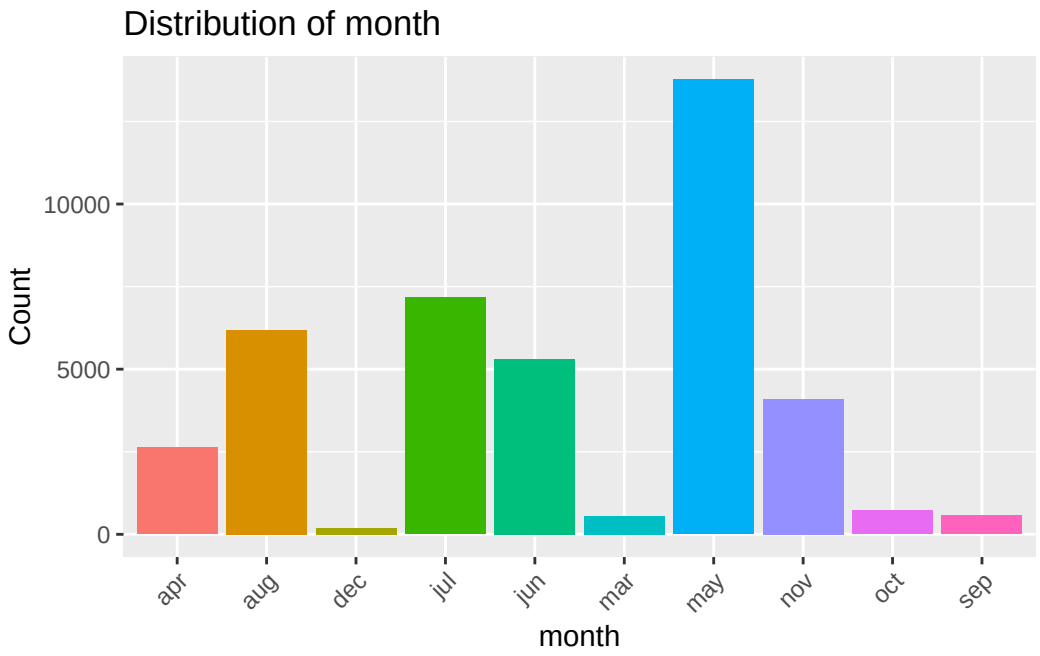
```
[[6]]
```


Distribution of loan
Count
30000
20000
10000
0
no
unknown
yes
loan

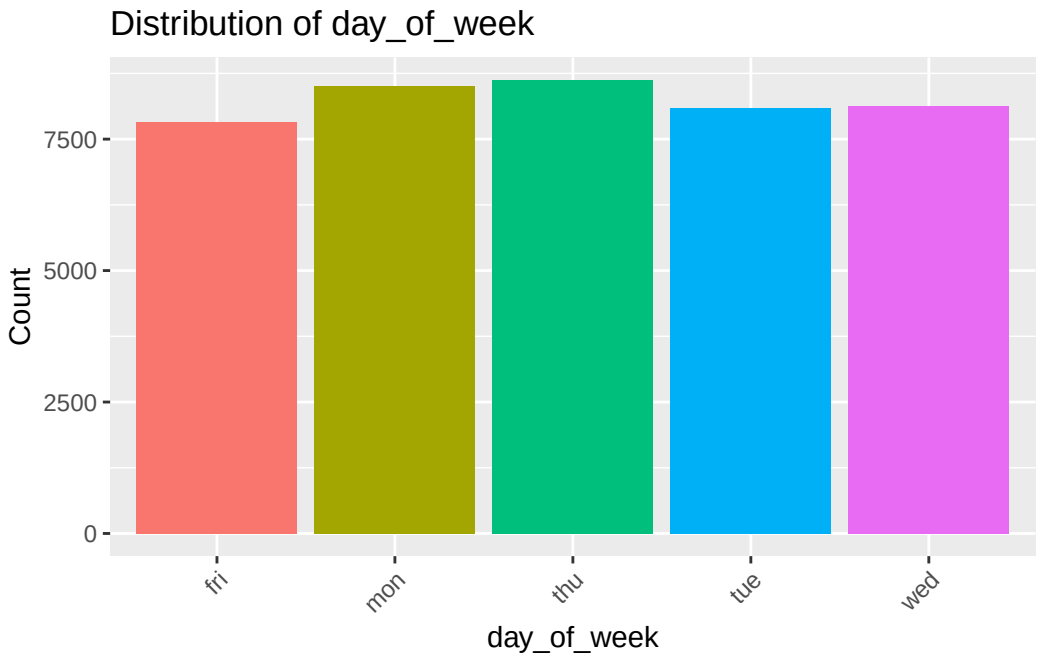
```
[[7]]
```


Distribution of contact
Count
20000
10000
0
cellular
telephone
contact

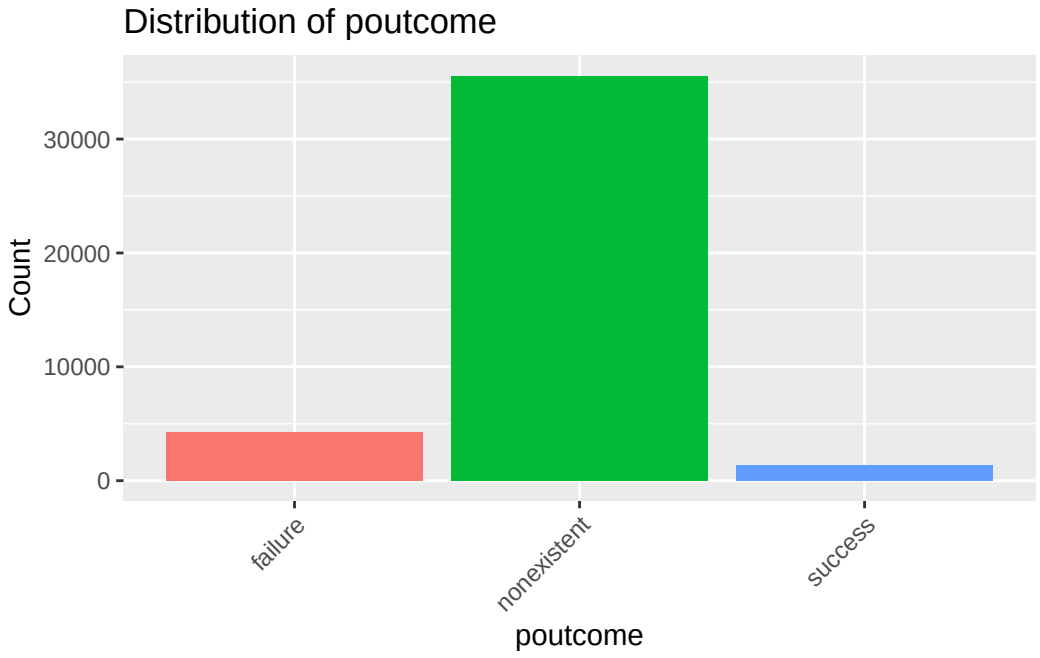
```
[[8]]
```


Distribution of month
Count
10000
5000
0
apr
aug
dec
jul
jun
mar
may
nov
oct
sep
month

```
[[9]]
```




```
[[10]]
```



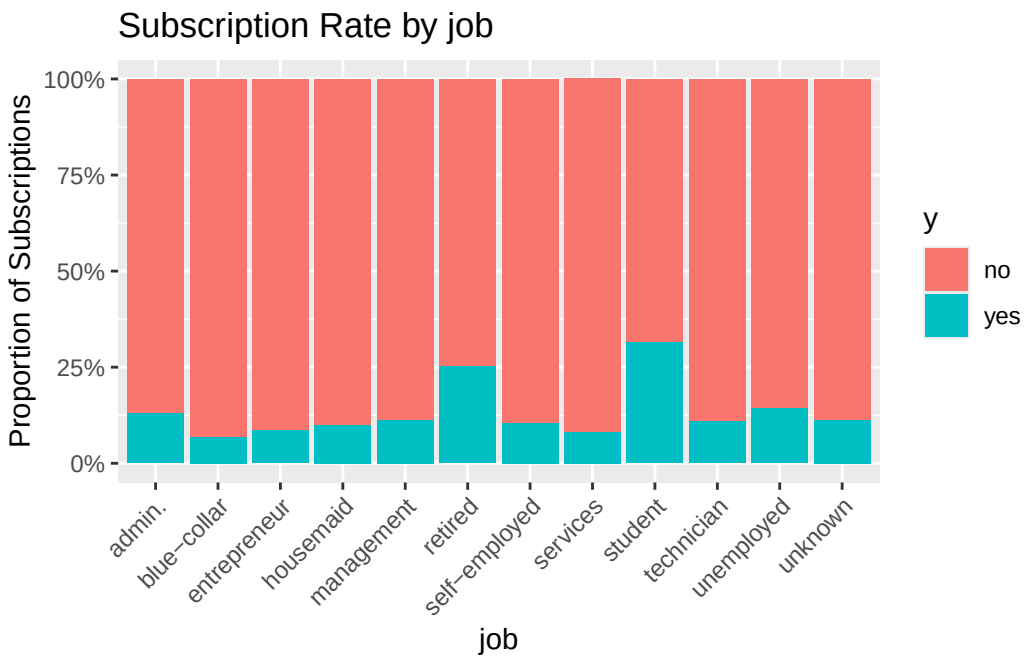
Most categorical variable distributions appear to dominated by one or two levels. Note: Some levels appear to have zero recorded observations. In actuality they have so few they do not

appear in this scale. For example, there are only three recorded observations of yes for $default$ in the entire dataset.

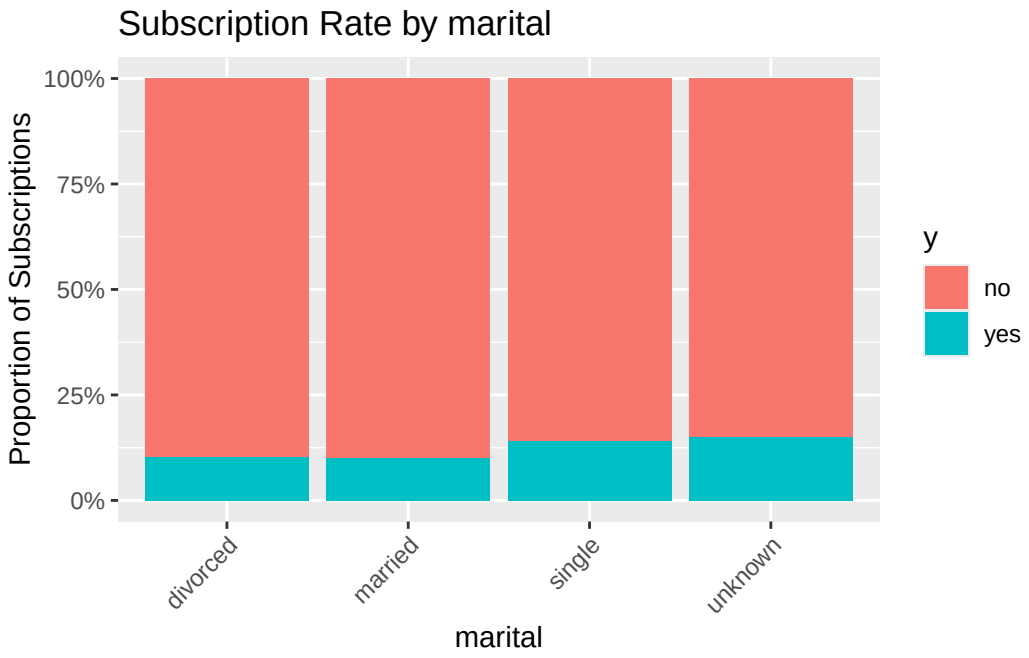
Categorical Variables vs Response

```
cat_vs_target <- function(var) {
  ggplot(bank, aes(x = .data[[var]], fill = y)) +
    geom_bar(position = "fill") +
    scale_y_continuous(labels = scales::percent) +
    labs(title = paste("Subscription Rate by", var),
         x = var, y = "Proportion of Subscriptions") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
}

map(cat_vars %>% names(), cat_vs_target)
```

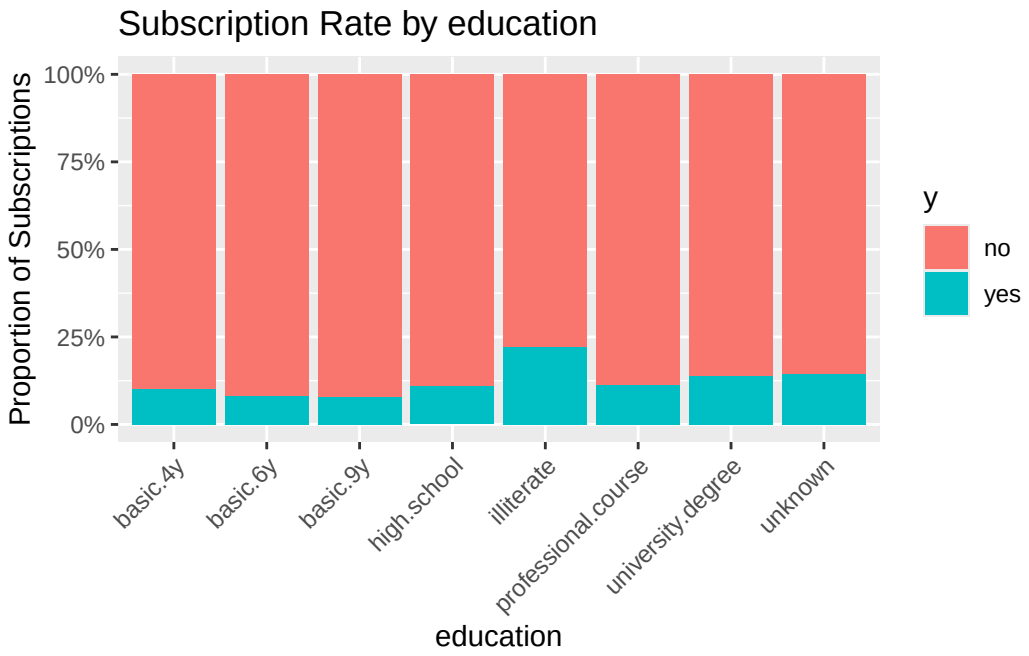
```
[[1]]
```



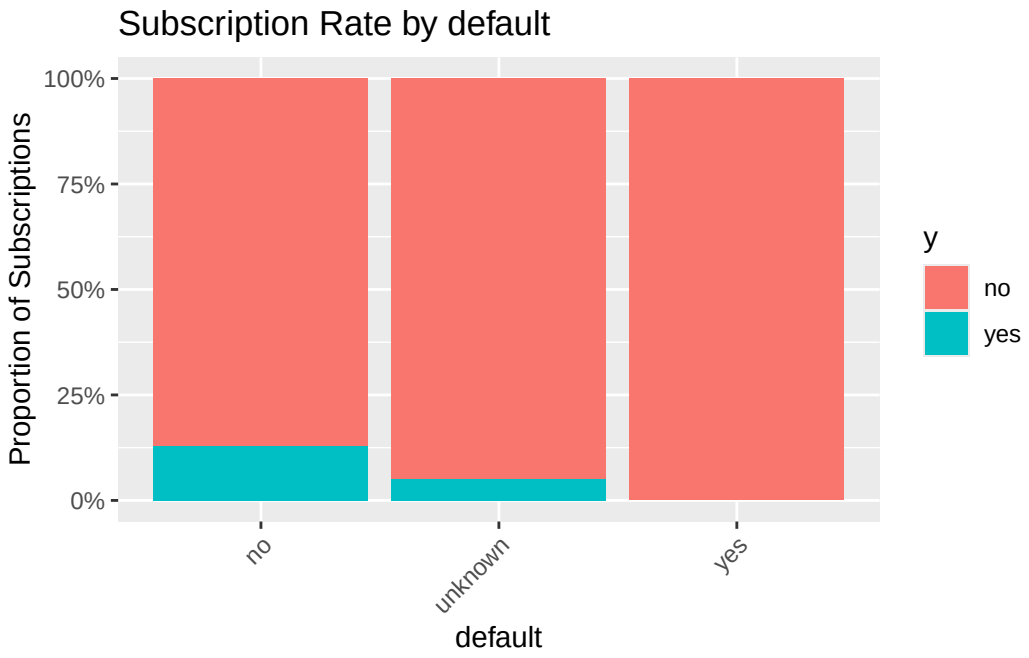

```
[[2]]
```


Subscription Rate by marital
Proportion of Subscriptions
100%
75%
50%
25%
0%
divorced
married
single
unknown
marital
y
no
yes

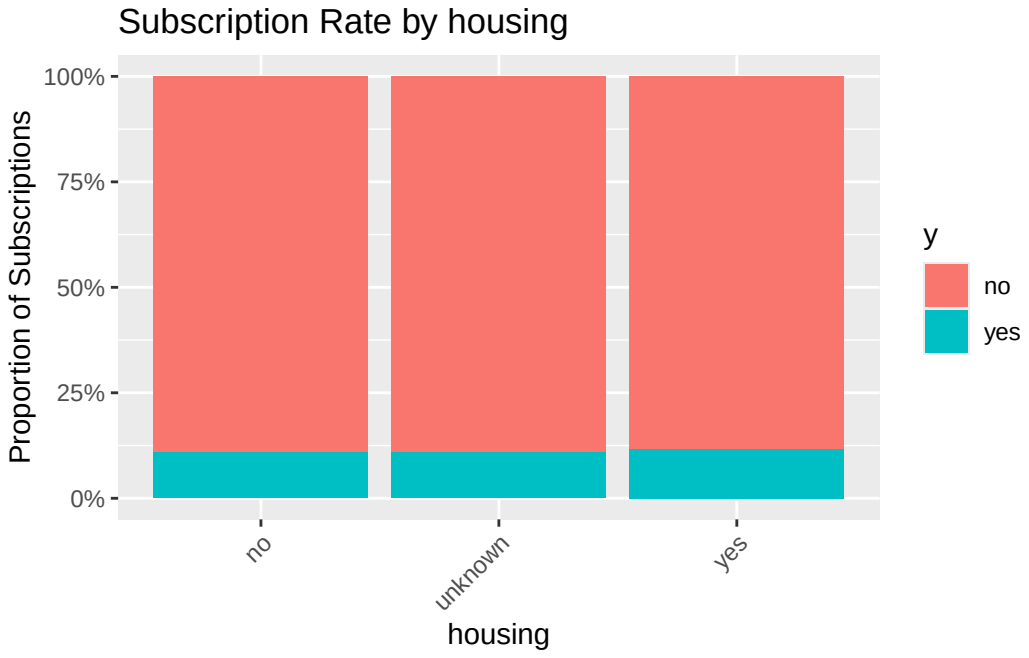
```
[[3]]
```


Subscription Rate by education
Proportion of Subscriptions
100%
75%
50%
25%
0%
basic.4y
basic.6y
basic.9y
high.school
illiterate
professional.course
university.degree
unknown
education
y
no
yes

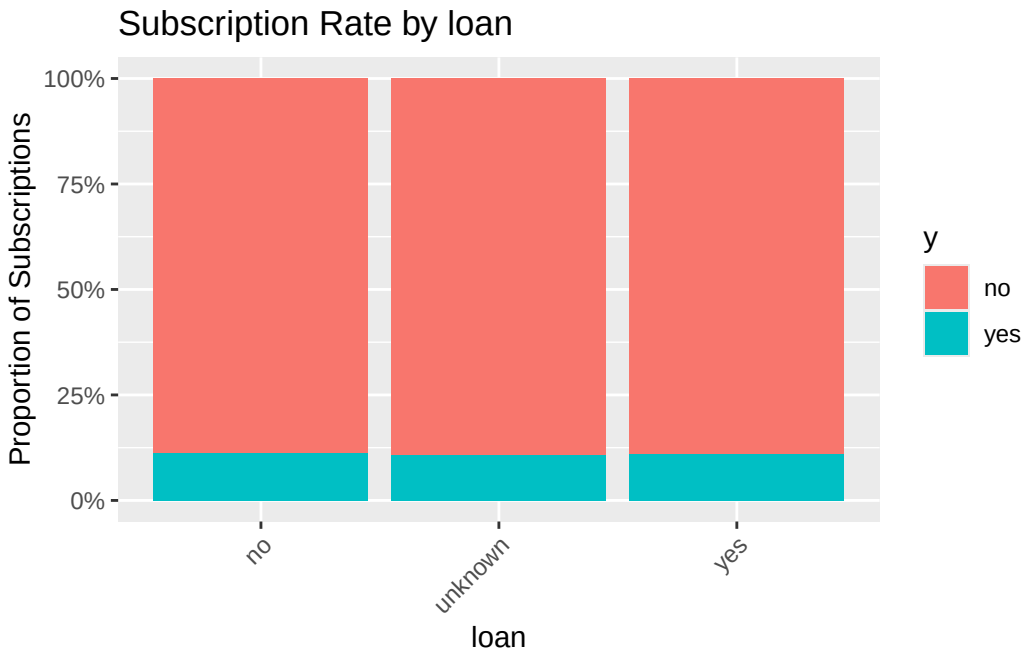
```
[[4]]
```


Subscription Rate by default
100%
75%
50%
25%
0%
Proportion of Subscriptions
no
unknown
yes
default
y
no
yes

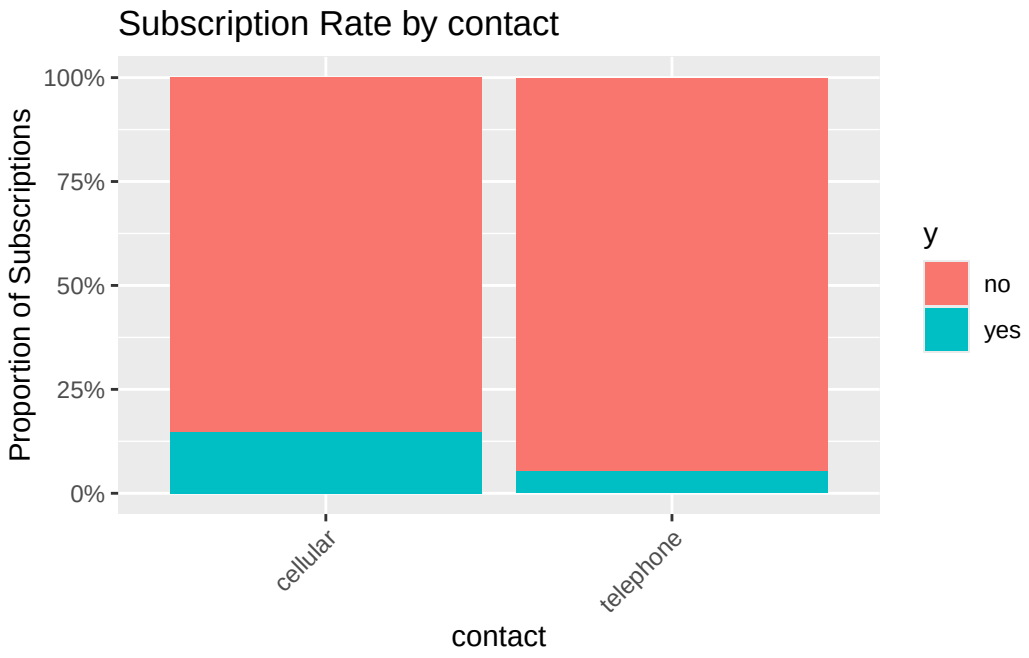
```
[[5]]
```


Subscription Rate by housing
100%
75%
50%
25%
0%
Proportion of Subscriptions
no
unknown
yes
housing
y
no
yes

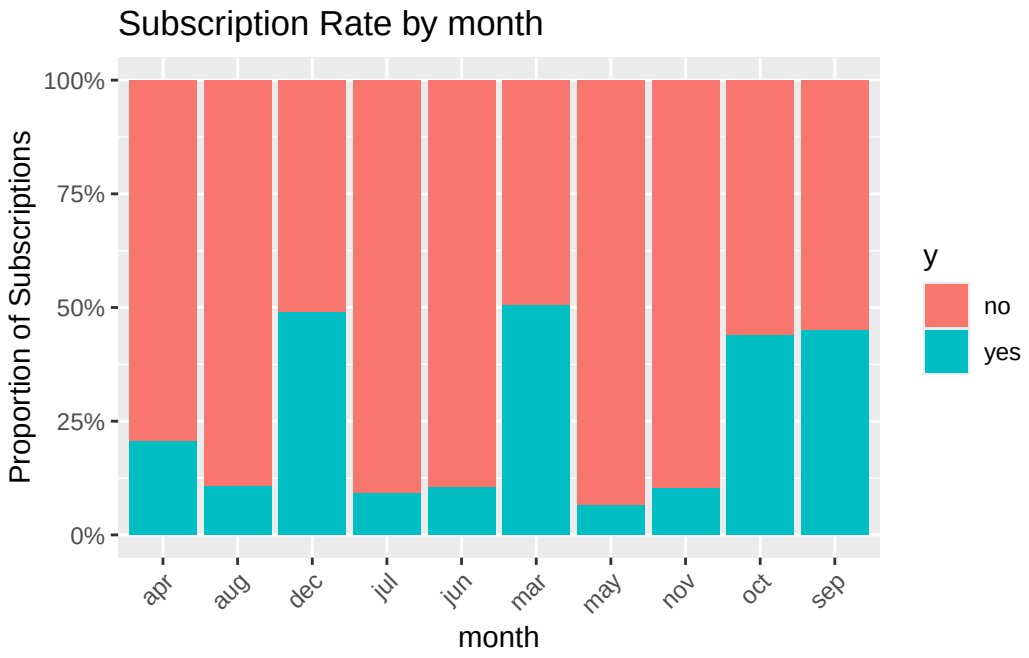
```
[[6]]
```


Subscription Rate by loan
Proportion of Subscriptions
100%
75%
50%
25%
0%
no
unknown
yes
loan
y
no
yes

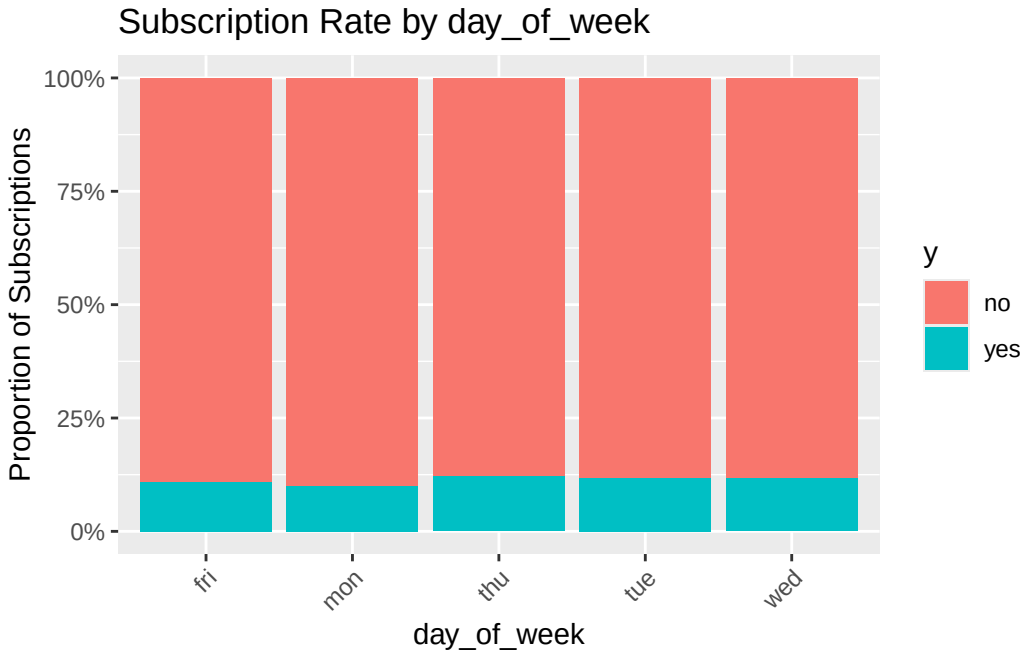
```
[[7]]
```


Subscription Rate by contact
Proportion of Subscriptions
100%
75%
50%
25%
0%
cellular
telephone
contact
y
no
yes

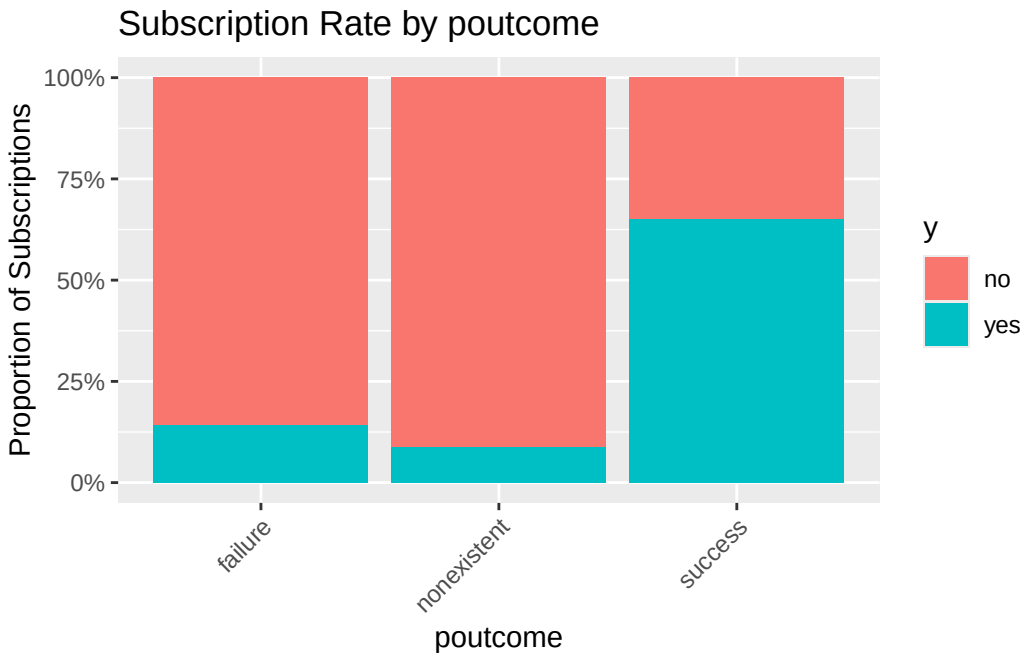
```
[[8]]
```


Subscription Rate by month
Proportion of Subscriptions
100%
75%
50%
25%
0%
apr
aug
dec
jul
jun
mar
may
nov
oct
sep
month
y
no
yes

```
[[9]]
```


Subscription Rate by day_of_week
Proportion of Subscriptions
100%
75%
50%
25%
0%
fri
mon
thu
tue
wed
day_of_week
y
no
yes

```
[[10]]
```



As expected from looking at the frequency of the response earlier, most levels of our categorical variables primarily result in *no* for the response. Notable differences are the increased amount of *yes* in the months of December, March, October, and September. As well as the large number of *yes* when *poutcome* is *success*. *poutcome* is the result of the previous marketing campaign with a customer, so it is not surprising that a customer who has subscribed for a term deposit before would be more likely to do so again.

Logistic Model with all Variables

Example logistic regression model ran on all possible predictors.

```
# Split 80/20 for train/test
set.seed(123)
train_index <- createDataPartition(bank$y, p = 0.8, list = FALSE)
train <- bank[train_index, ]
test <- bank[-train_index, ]
```

```
# Basic logistic model using all predictors
logit_model <- glm(y ~ ., data = train, family = binomial)

# View summary
summary(logit_model)
```

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

Coefficients: (1 not defined because of singularities)
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -2.182e+02  4.297e+01  -5.078 3.82e-07 ***
age                          8.420e-04  2.718e-03   0.310 0.756749
jobblue-collar              -2.671e-01  8.899e-02  -3.002 0.002683 **
jobentrepreneur             -1.592e-01  1.381e-01  -1.153 0.249023
jobhousemaid                -2.131e-02  1.641e-01  -0.130 0.896699
jobmanagement               -3.507e-02  9.462e-02  -0.371 0.710925
jobretired                   2.857e-01  1.197e-01   2.387 0.016983 *
jobself-employed            -9.936e-02  1.282e-01  -0.775 0.438194
jobservices                 -2.115e-01  9.890e-02  -2.138 0.032484 *
jobstudent                   2.182e-01  1.245e-01   1.752 0.079822 .
jobtechnician               -3.832e-02  7.956e-02  -0.482 0.630099
jobunemployed                1.002e-01  1.409e-01   0.711 0.477119
jobunknown                  -2.366e-01  2.786e-01  -0.849 0.395699
maritalmarried              -2.214e-02  7.686e-02  -0.288 0.773341
maritalsingle                5.888e-02  8.769e-02   0.671 0.501944
maritalunknown              -1.053e-01  4.699e-01  -0.224 0.822734
educationbasic.6y            6.725e-02  1.364e-01   0.493 0.622119
educationbasic.9y           -2.870e-02  1.068e-01  -0.269 0.788117
educationhigh.school        -2.386e-02  1.031e-01  -0.231 0.817001
educationilliterate          1.335e+00  8.356e-01   1.598 0.110077
educationprofessional.course 1.216e-01  1.126e-01   1.081 0.279852
educationuniversity.degree   1.570e-01  1.028e-01   1.528 0.126551
educationunknown             5.633e-02  1.340e-01   0.420 0.674251
defaultunknown              -3.019e-01  7.514e-02  -4.018 5.86e-05 ***
defaultyes                  -7.316e+00  1.135e+02  -0.064 0.948597
housingunknown              -1.566e-01  1.544e-01  -1.014 0.310683
housingyes                  -1.407e-02  4.621e-02  -0.305 0.760735
loanunknown                         NA         NA      NA       NA
loanyes                     -3.360e-02  6.404e-02  -0.525 0.599783
contacttelephone            -5.905e-01  8.477e-02  -6.965 3.28e-12 ***
monthaug                     8.277e-01  1.344e-01   6.157 7.42e-10 ***
monthdec                     2.283e-01  2.334e-01   0.978 0.327897
monthjul                     1.290e-01  1.067e-01   1.210 0.226467
monthjun                    -5.544e-01  1.405e-01  -3.946 7.94e-05 ***
monthmar                     1.924e+00  1.598e-01  12.046  < 2e-16 ***
monthmay                    -4.946e-01  9.154e-02  -5.403 6.57e-08 ***
monthnov                    -4.640e-01  1.353e-01  -3.430 0.000603 ***
```

```
monthoct                     2.160e-01  1.718e-01   1.257 0.208611
monthsep                     3.128e-01  2.005e-01   1.560 0.118758
day_of_weekmon              -7.727e-02  7.374e-02  -1.048 0.294702
day_of_weekthu               9.468e-02  7.167e-02   1.321 0.186511
day_of_weektue               1.118e-01  7.396e-02   1.511 0.130751
day_of_weekwed               2.141e-01  7.333e-02   2.919 0.003507 **
duration                     4.655e-03  8.320e-05  55.948  < 2e-16 ***
campaign                    -3.956e-02  1.296e-02  -3.053 0.002264 **
pdays                       -7.899e-04  2.374e-04  -3.327 0.000878 ***
previous                    -2.950e-02  6.741e-02  -0.438 0.661661
poutcomenonexistent          4.674e-01  1.060e-01   4.409 1.04e-05 ***
poutcomesuccess              1.072e+00  2.308e-01   4.646 3.38e-06 ***
emp_var_rate                -1.744e+00  1.579e-01 -11.045  < 2e-16 ***
cons_price_idx               2.055e+00  2.825e-01   7.274 3.50e-13 ***
cons_conf_idx                1.492e-02  8.630e-03   1.729 0.083778 .
euribor3m                    3.854e-01  1.463e-01   2.635 0.008423 **
nr_employed                  4.197e-03  3.505e-03   1.197 0.231224
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23199  on 32950  degrees of freedom
Residual deviance: 13673  on 32898  degrees of freedom
AIC: 13779

Number of Fisher Scoring iterations: 10
```

Predictions and Confusion Matrix

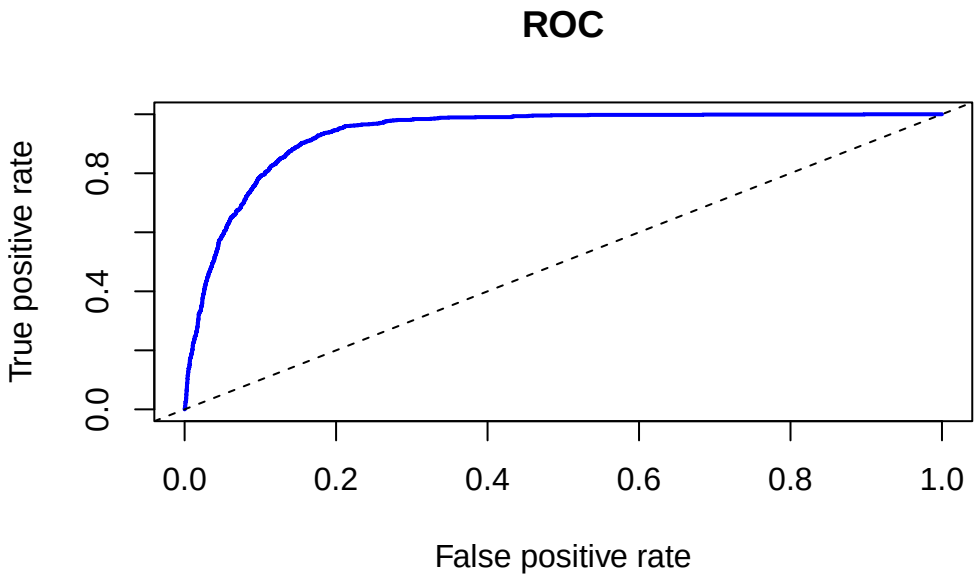
```r
# Predict probabilities and classes
test$pred_prob <- predict(logit_model, newdata = test, type = "response")
test$pred_class <- ifelse(test$pred_prob > 0.5, "yes", "no")

# Confusion matrix
confusionMatrix(factor(test$pred_class, levels = c("no", "yes")),
                test$y, positive = "yes")
```

```
Confusion Matrix and Statistics
```

```
          Reference
Prediction   no  yes
       no  7115  543
       yes  194  385

               Accuracy : 0.9105
                 95% CI : (0.9042, 0.9166)
    No Information Rate : 0.8873
    P-Value [Acc > NIR] : 3.635e-12

                  Kappa : 0.4646

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.41487
            Specificity : 0.97346
         Pos Pred Value : 0.66494
         Neg Pred Value : 0.92909
             Prevalence : 0.11266
         Detection Rate : 0.04674
   Detection Prevalence : 0.07029
      Balanced Accuracy : 0.69416

       'Positive' Class : yes
```

ROC Curve

```r
roc.predictions <- prediction(test$pred_prob, test$y)

roc.perf <- performance(roc.predictions, "tpr", "fpr")

plot(roc.perf, col = "blue", lwd = 2, main = "ROC", print.thres = TRUE)
abline(a = 0, b = 1, col = "black", lty = 2)
```



AUC:

```
auc <- performance(roc.predictions, "auc")
auc@y.values[[1]]
```

```
[1] 0.9353211
```