Individual Assignment 4

Logistic Regression

Sina Bahrami

Part 1: Data pre-processing

The data are imported from the Kaggle website [1]. It represents marketing campaign of a financial institution.

The dataset is of shape 11162, 17 and the variables are age, job, marital, education, default, balance, housing, loan, contact, day, month, duration, campaign, pdays, previous, poutcome, deposit. The definitions of the variables are obtained from [2] and are provided in the appendix.

Exploratory data analysis

The results of the exploratory data analysis, as provided in the appendix, are explained below.

- The data is tidy and in good format, and no cleaning is required.
- Based on the summary of the data, there are no missing values.
- Box plots suggest presence of outliers, but the values are in a reasonable range so they are kept.
- The variables “poutcome” and “education” have “unknown” values. The observations that contain “unknown” should be dropped if the variables are used as predictors in the model because they are not basically a separate category.
- Number of contacts made during the campaign with the client (“campaign”) is intuitively expected to have impact on subscription to the term deposit, but the histogram implies no strong correlation.
- Balance doesn’t seems to have strong impact but intuitively it may have so it is kept.
- Also based on the intuition and the bar plot:
  - Clients without housing are more likely to subscribe to the term deposit.
  - Married clients seem to have less tendency to subscribe to the term deposit.
  - Clients with tertiary education are more likely to subscribe to the term deposit than people with lower levels of education.
  - Interestingly people at age range of 30-60 seem to be more likely to subscribe to the term deposit.
  - Certain job categories like retired, management, student and unemployed seem to be more likely to make the subscription.
  - Clients with loan seems to be less likely to subscribe.
  - The majority of the clients have no credit in default so it cannot be a very useful predictors.
  - The contacts have many “unknown”, so it is not considered among the predictors.

Part 2: Planning

Problem statement

Following the data pre-processing and some exploratory data analysis, in the rest of this study, we would like to adopt a model so that we can predict whether the customer is going to subscribe to a term deposit which is specified by the “deposit” variable.The prediction model can be used to decide on the best campaigning strategy to increase the number of subscriptions.

Variable selection

- Because the outcome variable (“deposit”) is of a binary nature, a logistic regression model can be adopted to predict the outcome variable.
- According to [2], the variable “duration” is a very deciding factor in the outcome, but cannot be used because it is not known until the customer is contacted, hence not useful for adopting a campaigning strategy.
- As explained in the EDA section, there seems to be an effect from the variable “age” as to whether it is in the range of 30-60 or not. By converting “age” to a categorical variable instead of a continuous one, we can capture this effect.
- Based on the bar plots, tertiary education has stronger effect compared to other levels, so the variable is converted to a binary or “tertiary”.
- According to the explanation in the exploratory data analysis section, the following predictors are chosen to predict subscription to the term deposit (“deposit”): marital, housing, loan, education, job, age, balance, campaign, deposit.

- The outcome is binary variable (yes/no) and the predictors are a combination of continuous and categorical variables.

Assumptions of the model

The assumptions of the logistic regression model include: independence of errors, multicollinearity, incomplete information and complete separation. The last two assumptions are fulfilled based on the plots of the exploratory data analysis. The first three assumptions, however, need to be checked after creating the model.

Part 3: Analysis

Base model

A logistic regression model is adopted using **marital, housing, loan, tertiary, job, mid_age, balance and campaign** as the predictors. The summary is provided in the appendix. It shows that all of the predictors except "jobunemployed" and "maritaldivorced" are statistically significant. Also by obtaining exponent of the 95% confidence interval of each coefficient, the odds ratio e^{β} corresponding to each coefficient, which represents the effect size, is presented in the appendix. The predictors with e^{β} close to 1 are removed from the alternative model.

- The only continuous variable in the model is "campaign". The linearity for this predictor needs to be checked. The p-value is 1.1999×10^{-4}, so the linearity is rejected at 5% level of significance. The predictor is removed from the model because its effect is not considerable too (e^{β} close to 0.9).
- Multicollinearity is checked by evaluating VIF criteria. For all of the predictors the VIF is close to 1 with the average of 1.0806. So there is no multicollinearity, hence no issue in this regard.
- For the independence of the errors, similar to linear regression, Durbin-Watson test is performed. The statistics value is 0.1821 with a p-value of near 0, so the independence of errors is rejected.

The alternative model

In the alternative model, the predictors with less statistical significance ("jobunemployed" and "maritaldivorced") are removed, also the predictors with a small effect size are removed, so "balance" is removed and the category of marital is changed to a binary with the two states of "married" or not. As a result the new predictors are **housing, loan, tertiary and mid_age**. The summary of the model is provided in the appendix.

- The new model has no continuous variables, so there is no need for linearity check.
- There is no multicollinearity because the base model, which includes all of the predictors in this model, is not multicolinear.
- For independence of errors, the p-value of the Durbin-Watson test is 0, so the independence is again rejected, so neither model 2 can be considered a valid model.

Model comparison

Despite the rejection of the independence assumption, for the sake of practice the two models are compared using the AIC criteria. AIC of the base model is 1.3573×10^{4} and for the model 2 (the simpler model) AIC is 1.3846×10^{4}. The model 2 has a higher AIC, so if both models were valid, model 1 would be preferred. Despite the fewer parameters of model 2, it has a higher deviance of 1.3836×10^{4} compared to the deviance of 1.3535×10^{4} of the model 1, thus resulting in a higher AIC.

Part 4: Conclusion

This study aims to predict whether a client decides to subscribe to a term deposit based on available data (predictors). To select suitable predictors, a exploratory data analysis is performed. Some predictors are removed throughout the analysis and some are modified to binary predictors. Two logistic regression models are created. The model assumptions were checked for both models and independence of errors is rejected for both models, so the models cannot be considered valid. The AIC of both models were compared. If the models were valid, the model 1 would be preferred.

References

1. Bachmann, J. (2017, November 12). Bank Marketing Dataset. Retrieved April 8, 2023, from https://www.kaggle.com/datasets/janiobachmann/bank-marketing-dataset?resource=download
2. Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22-31. doi:10.1016/j.dss.2014.03.001

Appendix

Variables definition

Candidate predictor variables

1. age: (numeric)
2. job: type of job (categorical: ‘admin.’,‘blue-collar’,‘entrepreneur’,‘housemaid’,‘management’,‘retired’,‘self-employed’,‘services’,‘stu
3. marital: marital status (categorical: ‘divorced’,‘married’,‘single’,‘unknown’; note: ‘divorced’ means divorced or widowed)
4. education: (categorical: primary, secondary, tertiary and 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’)
8. balance: Balance of the individual.
9. contact: contact communication type (categorical: ‘cellular’,‘telephone’)
10. month: last contact month of year (categorical: ‘jan’, ‘feb’, ‘mar’, …, ‘nov’, ‘dec’)
11. day: last contact day of the week (categorical: ‘mon’,‘tue’,‘wed’,‘thu’,‘fri’)
12. 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.
13. campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
14. 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)
15. previous: number of contacts performed before this campaign and for this client (numeric)
16. poutcome: outcome of the previous marketing campaign (categorical: ‘failure’,‘nonexistent’,‘success’)

Outcome variable

17. y - has the client subscribed a term deposit? (binary: ‘yes’,‘no’)

Code

```
knitr::opts_chunk$set(echo = TRUE)
remove(list=ls())
library(dplyr)
library(tidyr)
library(readr)
library(magrittr)
library(ggplot2)
# library(knitr)
library(pastecs)
library(GGally)
library(lmtest)
# library(desc)
library(car)
# library(leaps) # All-subsets method: regsubsets
# library(boot)
# library(ggm)
# library(corrplot)

options(digits=4)
detect_outlier <- function(x) {

  Quantile1 <- quantile(x, probs=.25, na.rm = TRUE)
  Quantile3 <- quantile(x, probs=.75, na.rm = TRUE)
  IQR = Quantile3-Quantile1
  x > Quantile3 + (IQR*1.5) | x < Quantile1 - (IQR*1.5)
}

# Takes a data frame and a column and removes the outliers from the
# original data frame
```

```r
remove_outlier <- function(dataframe, columns=names(dataframe)) {
  # for loop to traverse in columns vector
  for (colmn in columns) {
    # remove observation if it satisfies outlier function
    dataframe <- dataframe[!detect_outlier(dataframe[[colmn]]), ]
  }
  return(dataframe)
}
df_main <- read.csv(file="bank.csv",
                    na.strings = c("unknown", "Unknown", "NA", "na"),
                    stringsAsFactors = TRUE
)
columns <- names(df_main)
head(df_main)
```

```
##   age        job marital education default balance housing loan contact day
## 1  59     admin. married secondary      no    2343     yes   no    <NA>   5
## 2  56     admin. married secondary      no      45      no   no    <NA>   5
## 3  41 technician married secondary      no    1270     yes   no    <NA>   5
## 4  55   services married secondary      no    2476     yes   no    <NA>   5
## 5  54     admin. married  tertiary      no     184      no   no    <NA>   5
## 6  42 management  single  tertiary      no       0     yes  yes    <NA>   5
##   month duration campaign pdays previous poutcome deposit
## 1   may     1042        1    -1        0     <NA>     yes
## 2   may     1467        1    -1        0     <NA>     yes
## 3   may     1389        1    -1        0     <NA>     yes
## 4   may      579        1    -1        0     <NA>     yes
## 5   may      673        2    -1        0     <NA>     yes
## 6   may      562        2    -1        0     <NA>     yes
```

```r
str(df)
```

```
## function (x, df1, df2, ncp, log = FALSE)
```

```r
stat.desc(df_main)
```

```
##                    age job marital education default    balance housing loan
## nbr.val      1.116e+04  NA      NA        NA      NA  1.116e+04      NA   NA
## nbr.null     0.000e+00  NA      NA        NA      NA  7.740e+02      NA   NA
## nbr.na       0.000e+00  NA      NA        NA      NA  0.000e+00      NA   NA
## min          1.800e+01  NA      NA        NA      NA -6.847e+03      NA   NA
## max          9.500e+01  NA      NA        NA      NA  8.120e+04      NA   NA
## range        7.700e+01  NA      NA        NA      NA  8.805e+04      NA   NA
## sum          4.602e+05  NA      NA        NA      NA  1.706e+07      NA   NA
## median       3.900e+01  NA      NA        NA      NA  5.500e+02      NA   NA
## mean         4.123e+01  NA      NA        NA      NA  1.529e+03      NA   NA
## SE.mean      1.128e-01  NA      NA        NA      NA  3.053e+01      NA   NA
## CI.mean.0.95 2.210e-01  NA      NA        NA      NA  5.984e+01      NA   NA
## var          1.419e+02  NA      NA        NA      NA  1.040e+07      NA   NA
## std.dev      1.191e+01  NA      NA        NA      NA  3.225e+03      NA   NA
## coef.var     2.889e-01  NA      NA        NA      NA  2.110e+00      NA   NA
##              contact       day month  duration  campaign      pdays  previous
## nbr.val           NA 1.116e+04    NA 1.116e+04 1.116e+04  11162.000 1.116e+04
## nbr.null          NA 0.000e+00    NA 0.000e+00 0.000e+00      0.000 8.324e+03
## nbr.na            NA 0.000e+00    NA 0.000e+00 0.000e+00      0.000 0.000e+00
## min               NA 1.000e+00    NA 2.000e+00 1.000e+00     -1.000 0.000e+00
## max               NA 3.100e+01    NA 3.881e+03 6.300e+01    854.000 5.800e+01
## range             NA 3.000e+01    NA 3.879e+03 6.200e+01    855.000 5.800e+01
## sum               NA 1.748e+05    NA 4.152e+06 2.800e+04 572950.000 9.293e+03
## median            NA 1.500e+01    NA 2.550e+02 2.000e+00     -1.000 0.000e+00
## mean              NA 1.566e+01    NA 3.720e+02 2.508e+00     51.330 8.326e-01
## SE.mean           NA 7.970e-02    NA 3.286e+00 2.576e-02      1.029 2.169e-02
## CI.mean.0.95      NA 1.562e-01    NA 6.440e+00 5.050e-02      2.018 4.252e-02
## var               NA 7.091e+01    NA 1.205e+05 7.410e+00  11828.364 5.253e+00
## std.dev           NA 8.421e+00    NA 3.471e+02 2.722e+00    108.758 2.292e+00
## coef.var          NA 5.378e-01    NA 9.332e-01 1.085e+00      2.119 2.753e+00
##              poutcome deposit
## nbr.val            NA      NA
## nbr.null           NA      NA
## nbr.na             NA      NA
## min                NA      NA
## max                NA      NA
## range              NA      NA
## sum                NA      NA
## median             NA      NA
## mean               NA      NA
## SE.mean            NA      NA
## CI.mean.0.95       NA      NA
## var                NA      NA
## std.dev            NA      NA
## coef.var           NA      NA
```

```r
factor_cols <- df_main %>%
  select(where(is.factor)) %>%
  names()

print("The levels of the factor variables are provided below:")
```

```
## [1] "The levels of the factor variables are provided below:"
```

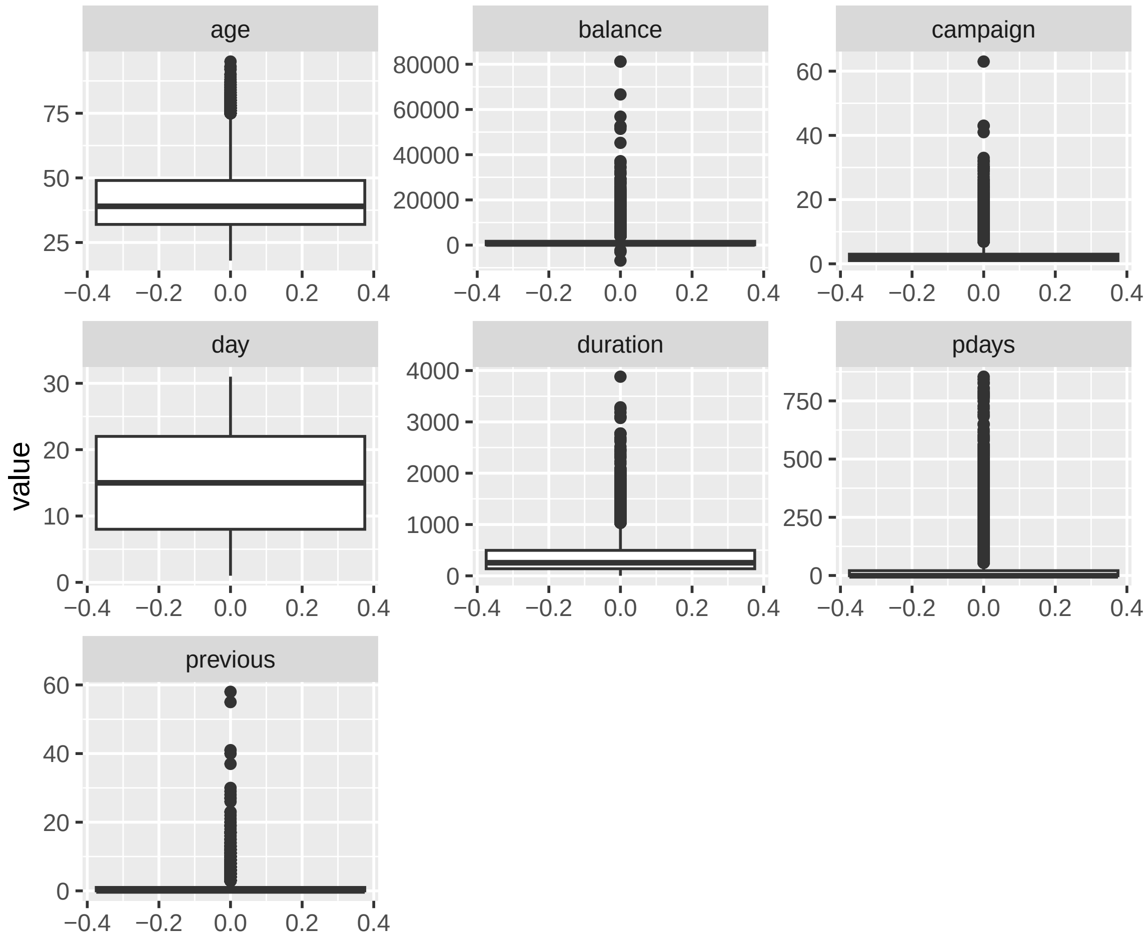
```r
sapply(df_main[factor_cols], levels)
```

```
## $job
##  [1] "admin."        "blue-collar"   "entrepreneur"  "housemaid"
##  [5] "management"    "retired"       "self-employed" "services"
##  [9] "student"       "technician"    "unemployed"
##
## $marital
## [1] "divorced" "married"  "single"
##
## $education
## [1] "primary"   "secondary" "tertiary"
##
## $default
## [1] "no"  "yes"
##
## $housing
## [1] "no"  "yes"
##
## $loan
## [1] "no"  "yes"
##
## $contact
## [1] "cellular"  "telephone"
##
## $month
##  [1] "apr" "aug" "dec" "feb" "jan" "jul" "jun" "mar" "may" "nov" "oct" "sep"
##
## $poutcome
## [1] "failure" "other"   "success"
##
## $deposit
## [1] "no"  "yes"
```

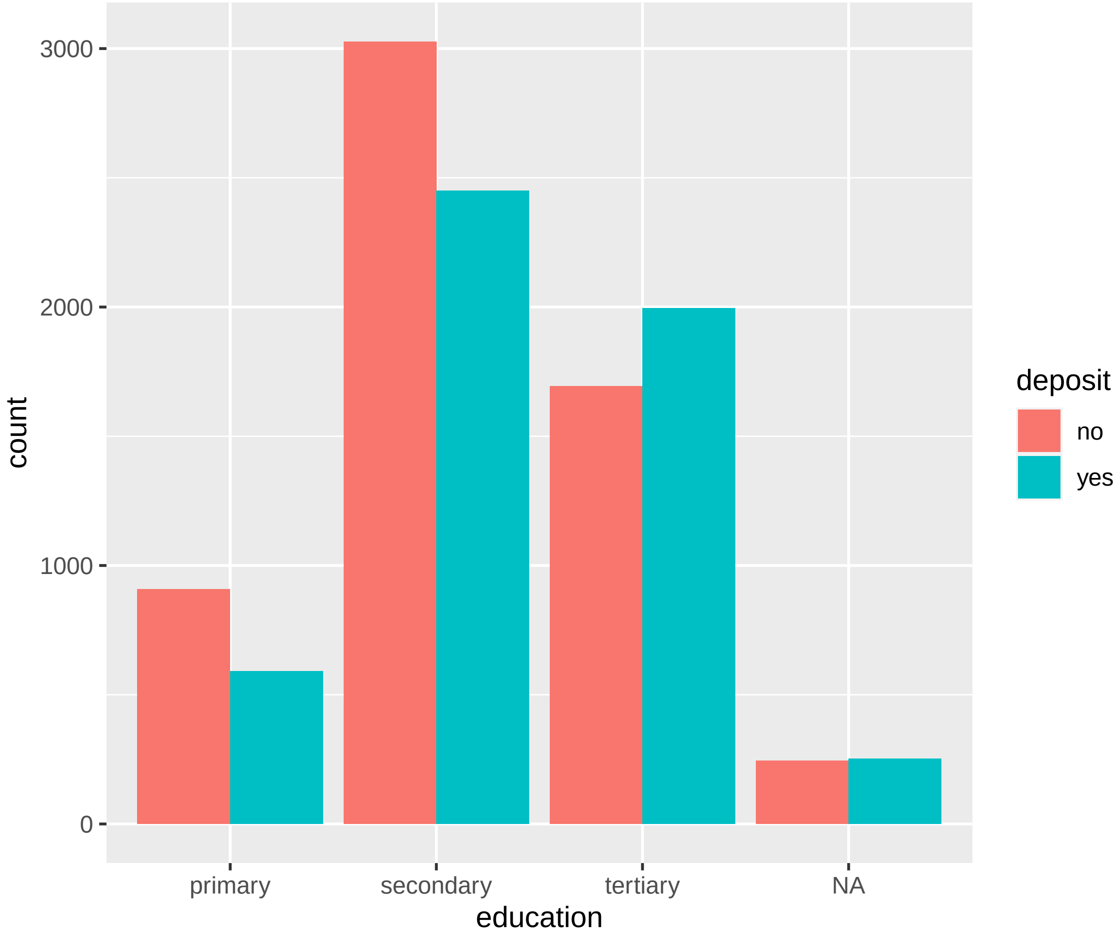
```
df_main %<>%
  mutate(
    education=factor(education, c("unknown", "primary", "secondary", "tertiary")),
    marital = factor(marital, c("single", "married", "divorced")),
    deposit = relevel(deposit, "no"),
    poutcome = factor(poutcome, c("failure", "success", "other", "unknown")),
    housing = relevel(housing, "no"),
  )


continuous_cols <- columns[!(columns %in% factor_cols)]
df_long <- df_main %>%
  pivot_longer(cols=continuous_cols, names_to="variable", values_to="value")
```

```
## Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.
## i Please use ‘all_of()‘ or ‘any_of()‘ instead.
##   # Was:
##   data %>% select(continuous_cols)
##
##   # Now:
##   data %>% select(all_of(continuous_cols))
##
## See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
```

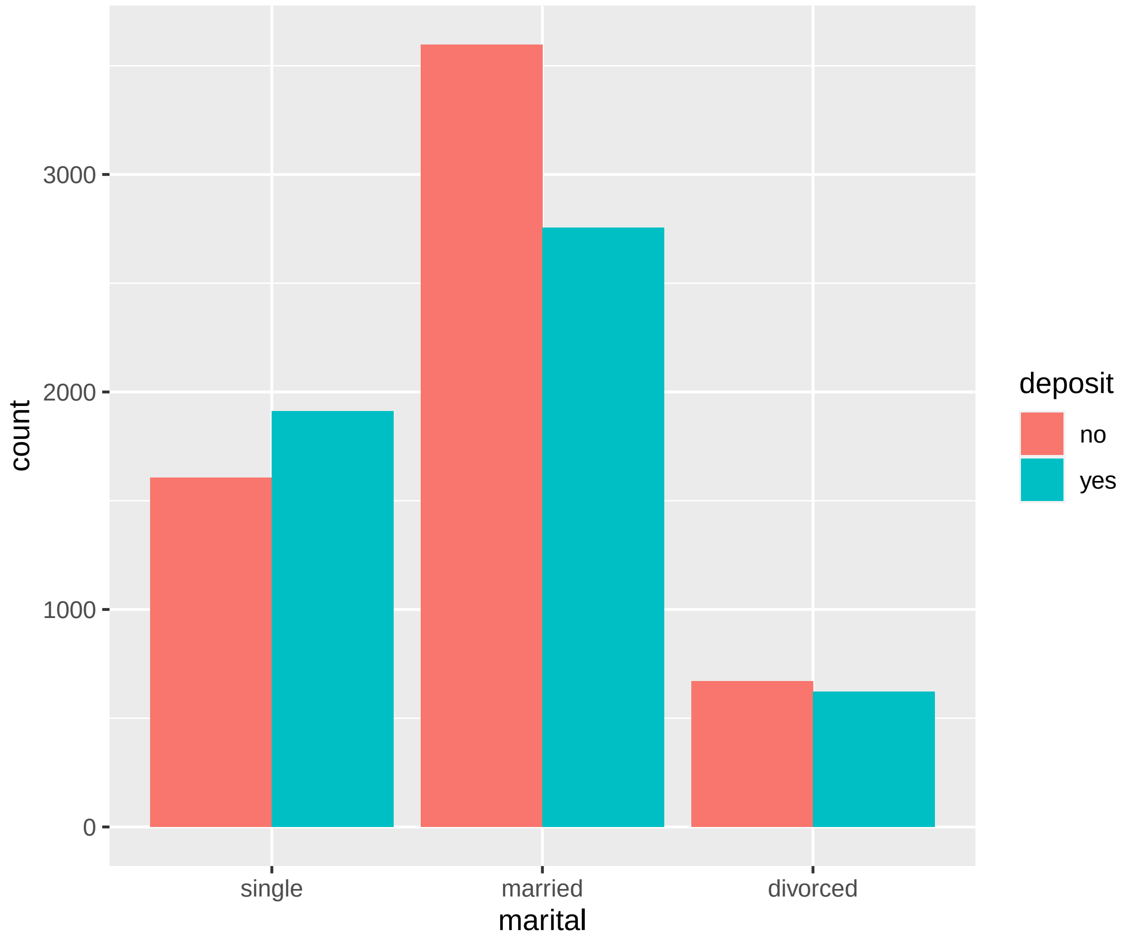
```
ggplot(df_long) +
  geom_boxplot(aes(y = value)) +
  facet_wrap(~variable, scales="free" )
```



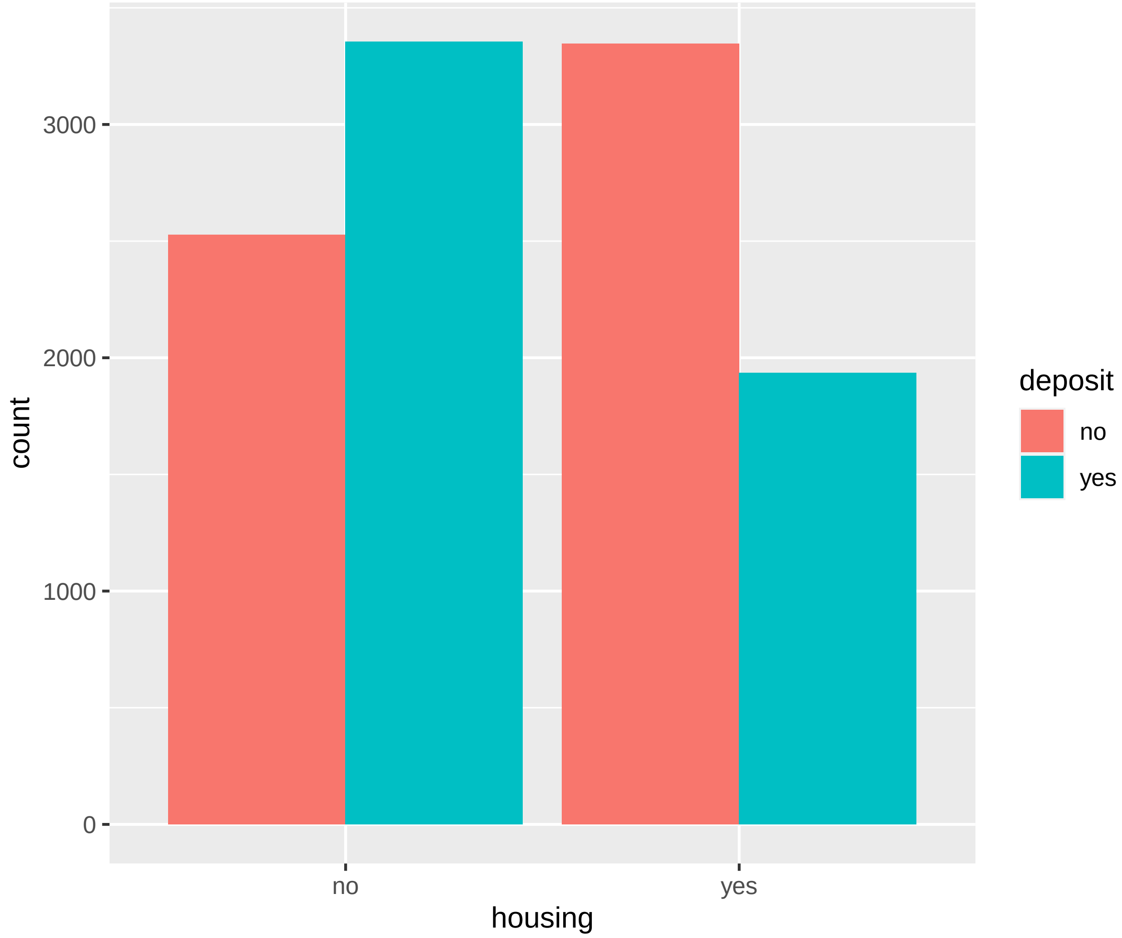
```
# Exploration of factor variables
ggplot(df_main, aes(fill=deposit)) +
  geom_bar(position="dodge", aes(x=education))
```



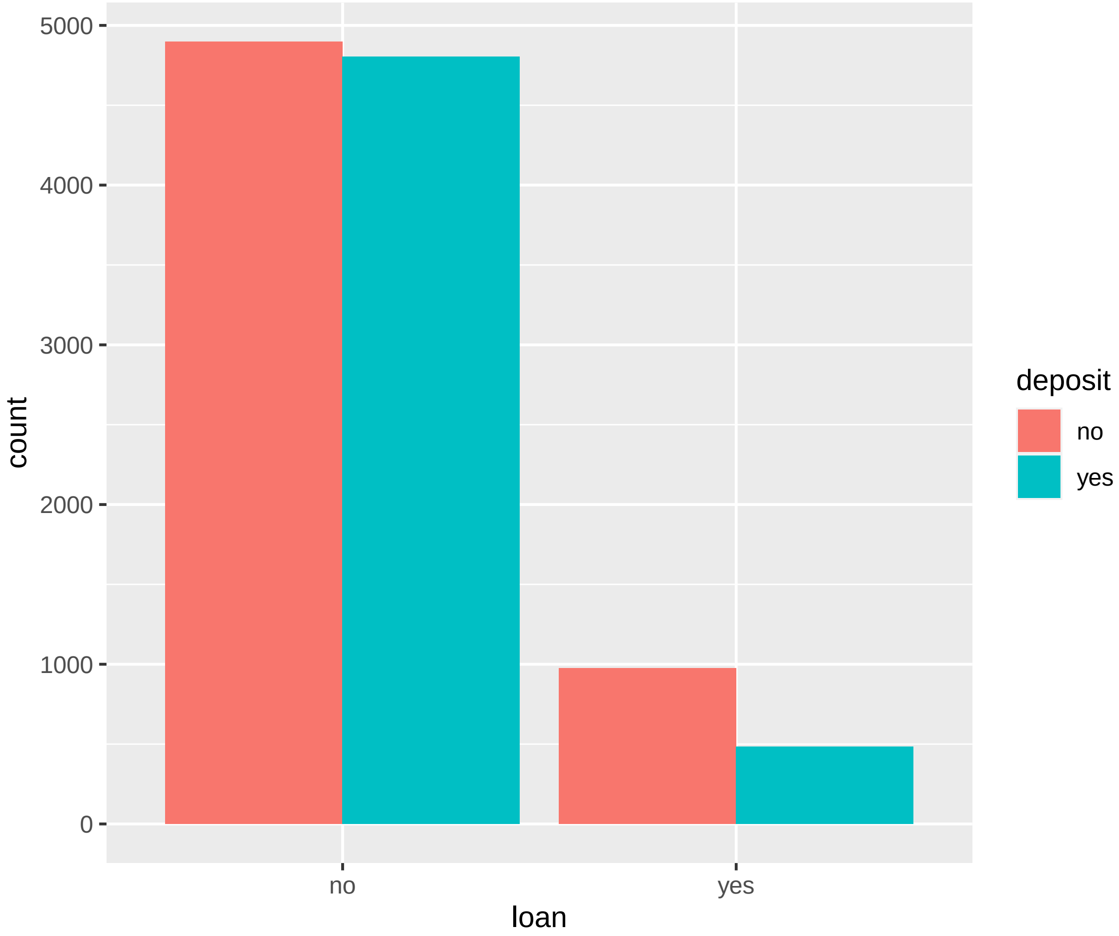
```
ggplot(df_main, aes(fill=deposit)) +
  geom_bar(position="dodge", aes(x=marital))
```



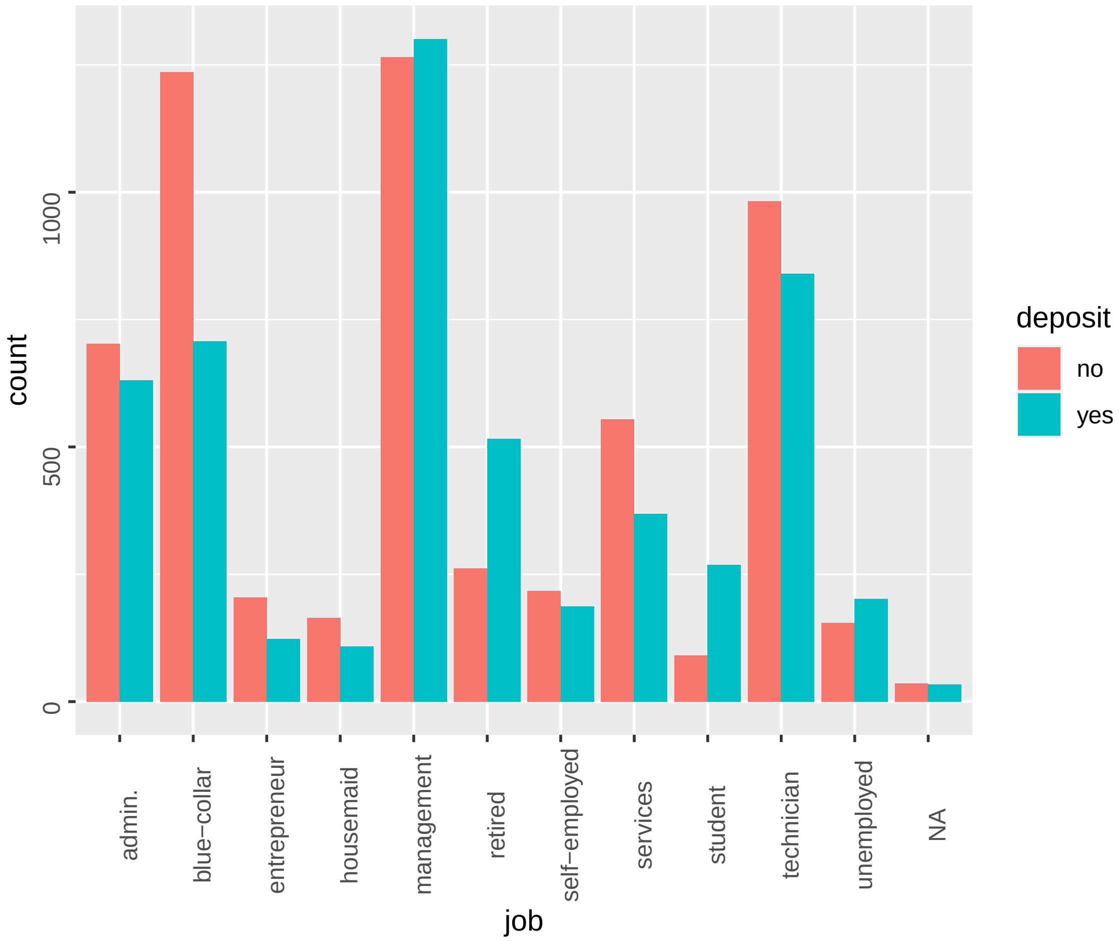
```
ggplot(df_main, aes(fill=deposit)) +
  geom_bar(position="dodge", aes(x=housing))
```



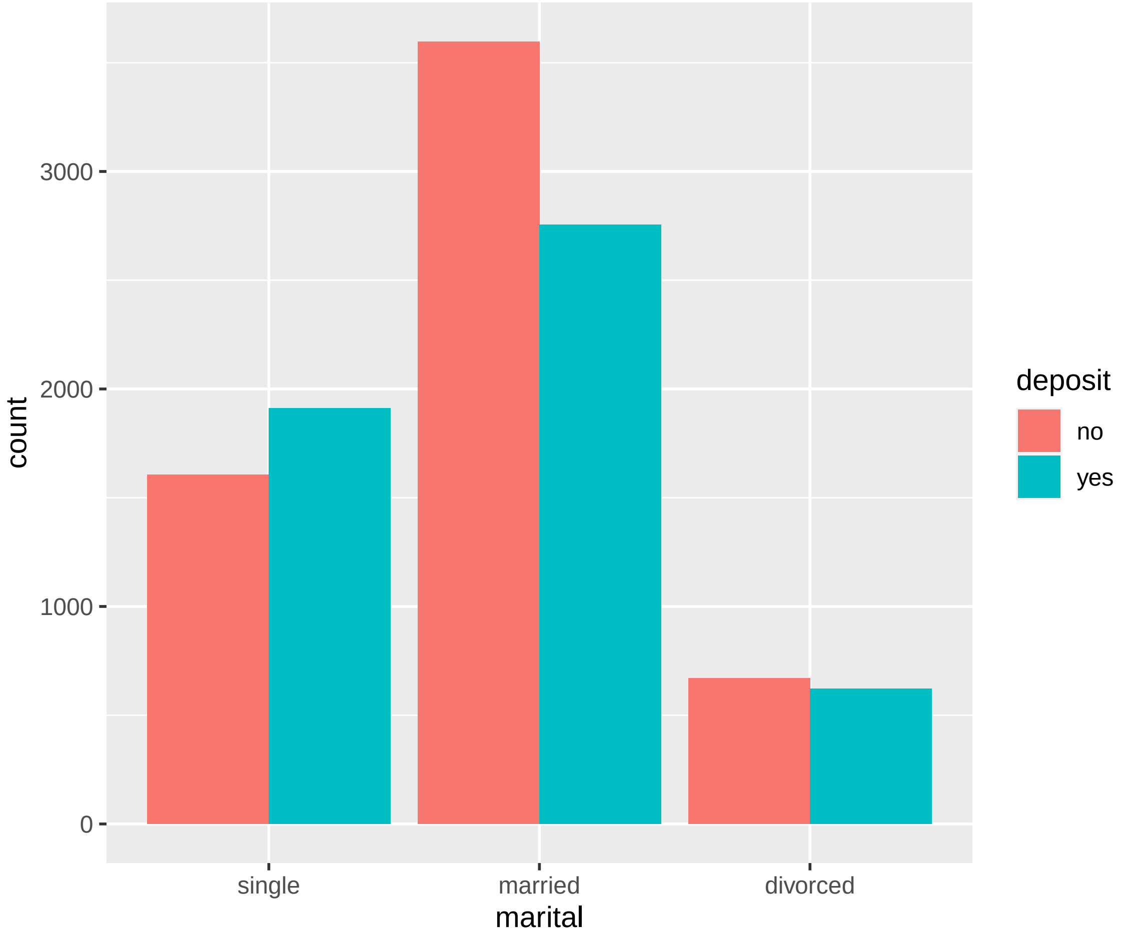
```
ggplot(df_main, aes(fill=deposit)) +
  geom_bar(position="dodge", aes(x=loan))
```



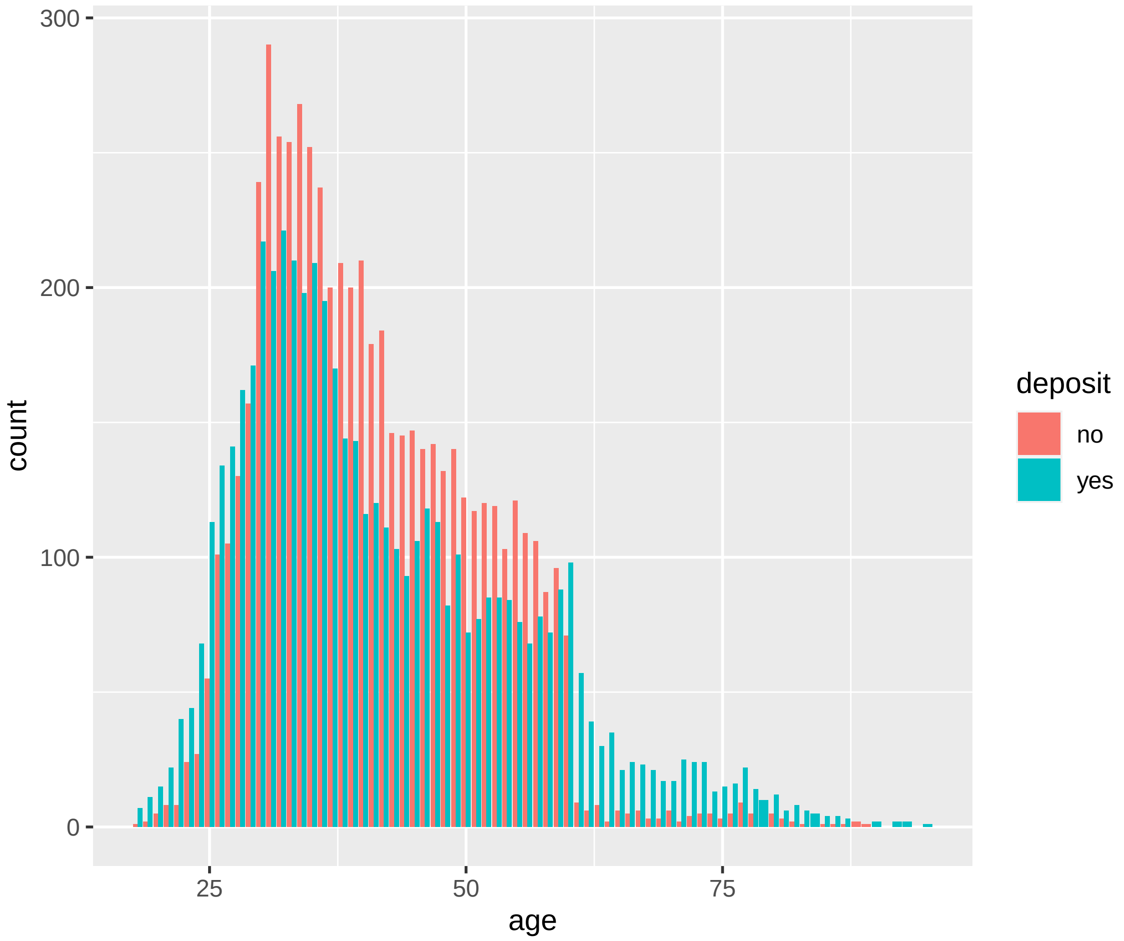
```
ggplot(df_main, aes(fill=deposit)) +
  geom_bar(position="dodge", aes(x=job)) +
  theme(axis.text=element_text(angle = 90))
```



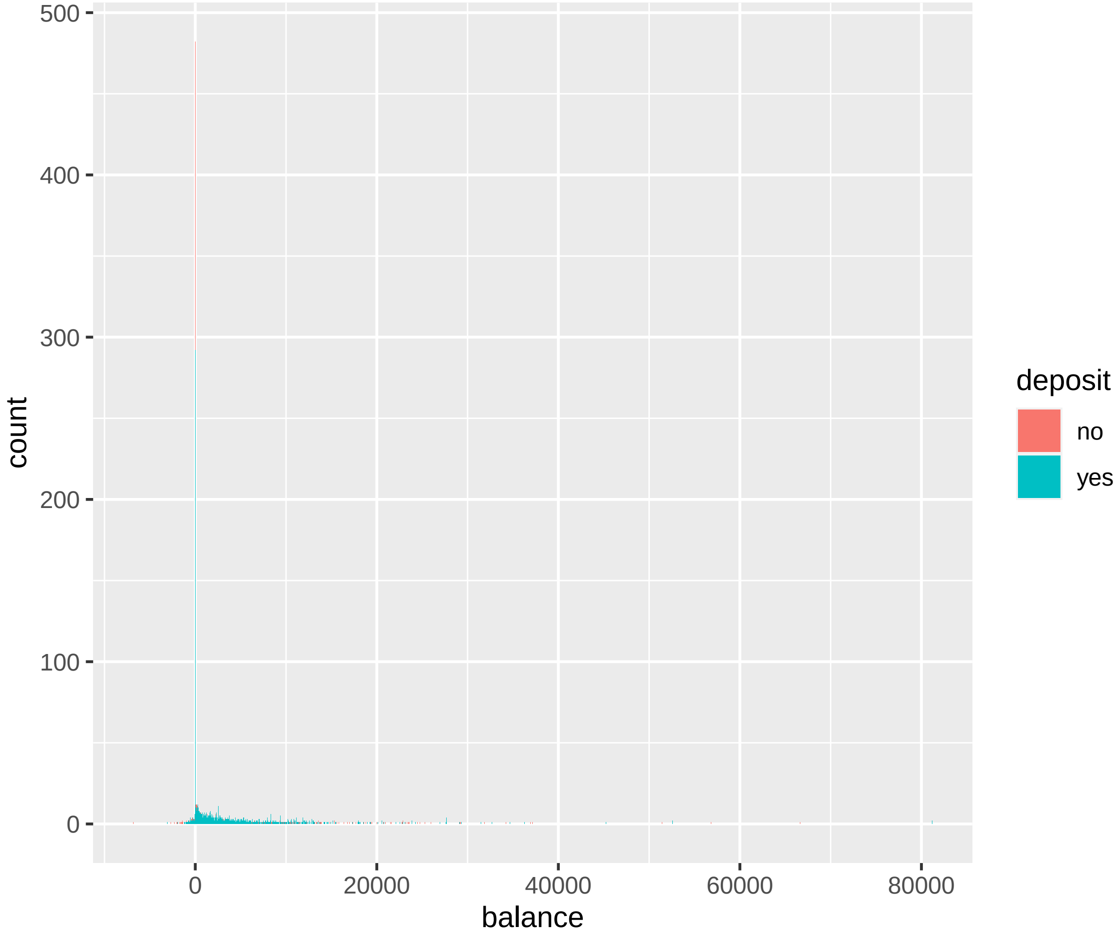
```
ggplot(df_main, aes(fill=deposit)) +
  geom_bar(position="dodge", aes(x=marital))
```



```
ggplot(df_main, aes(fill=deposit)) +
  geom_bar(position="dodge", aes(x=age))
```

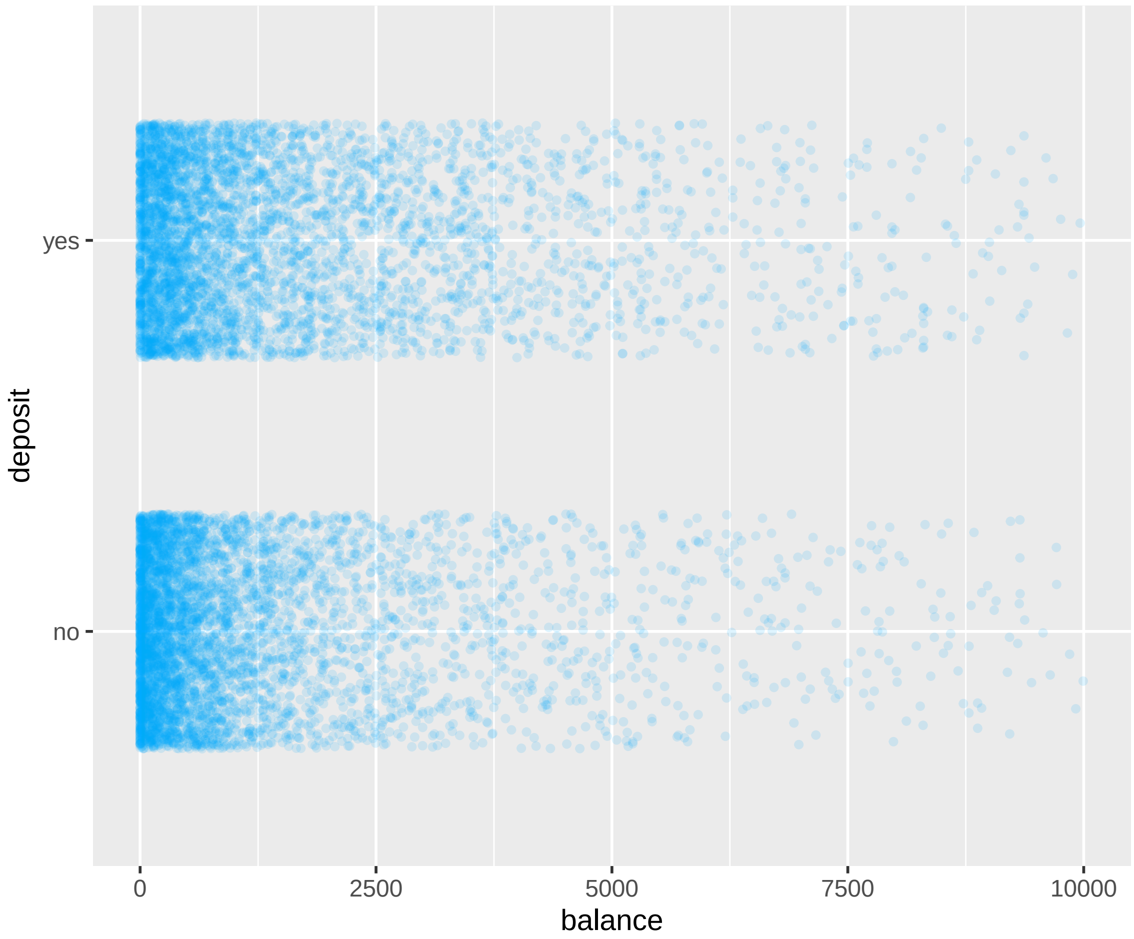


```
ggplot(df_main, aes(fill=deposit)) +
  geom_bar(position="dodge", aes(x=balance))
```

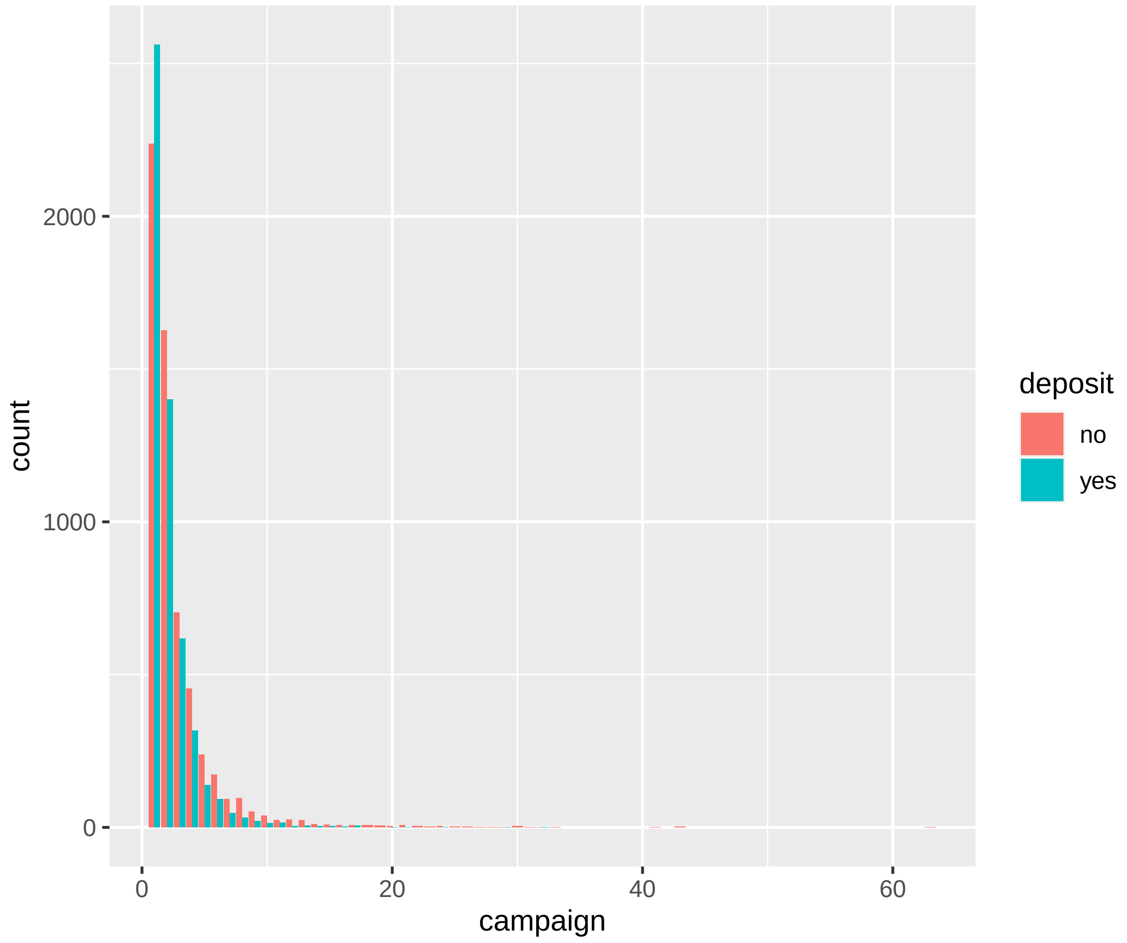


```
ggplot(df_main, aes(x=balance, y=deposit)) +
  geom_jitter(height=0.3, width=0.3, size=1, color="#00abff22") +
  scale_x_continuous(limits = c(0,10000))
```

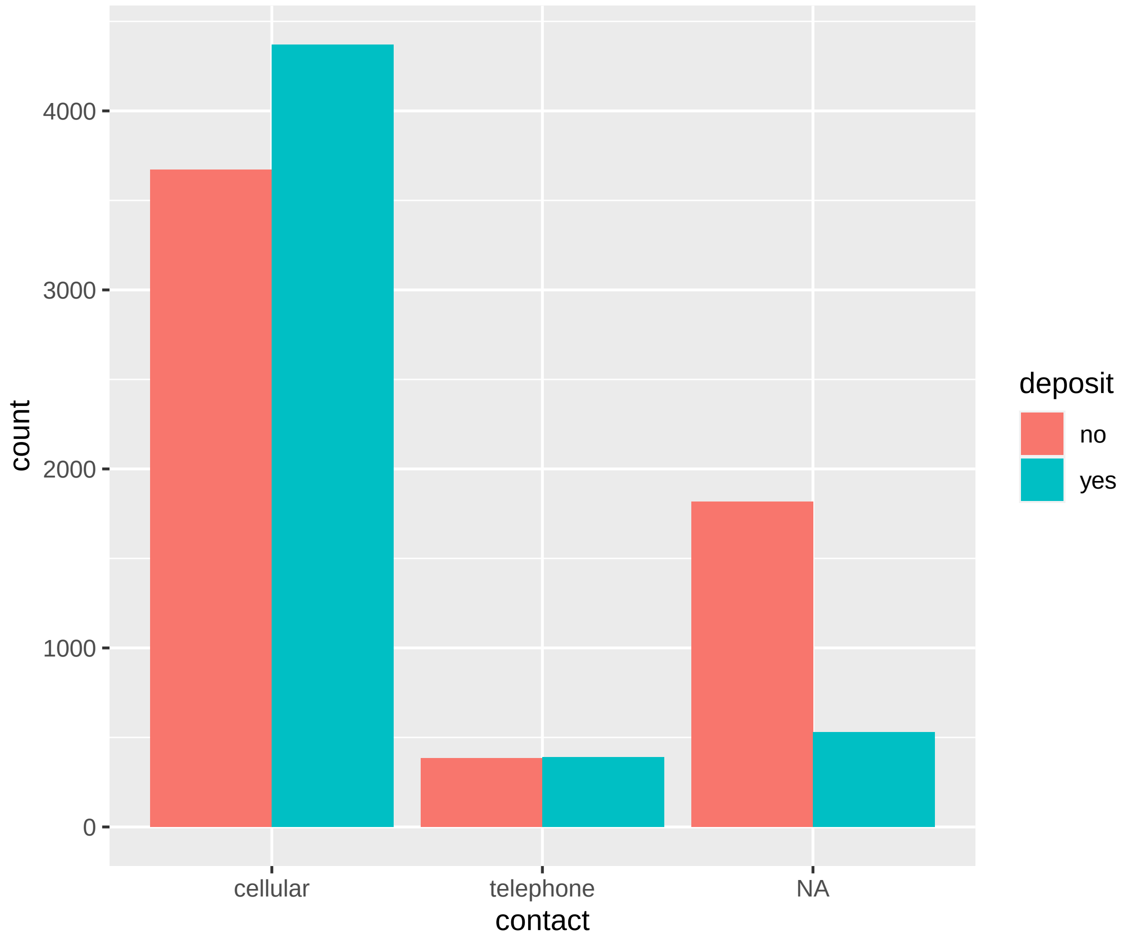
```
## Warning: Removed 1315 rows containing missing values (‘geom_point()‘).
```



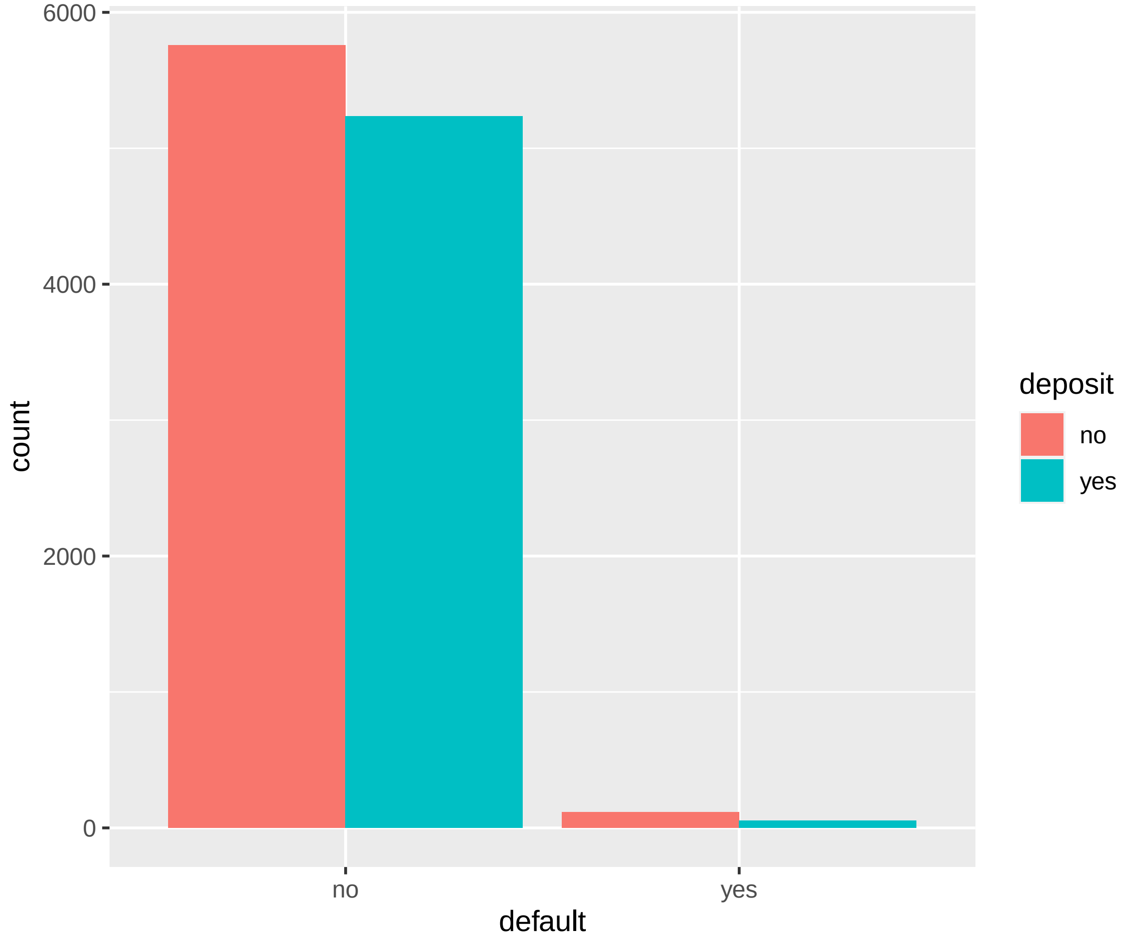
```
ggplot(df_main, aes(fill=deposit)) +
  geom_bar(position="dodge", aes(x=campaign))
```



```
ggplot(df_main, aes(fill=deposit)) +
  geom_bar(position="dodge", aes(x=contact))
```



```
ggplot(df_main, aes(fill=deposit)) +
  geom_bar(position="dodge", aes(x=default))
```



```
sel_variables <- c("marital", "housing", "loan", "education", "job", "age", "balance", "campaign", "deposit")
df_model <- df_main %>%
  mutate(
    mid_age=as.factor(if_else(age>=30 & age<=60, "yes", "no")),
    tertiary=as.factor(if_else(education=="tertiary", "yes", "no")))
sel_variables <- c("marital", "housing", "loan", "tertiary", "job", "mid_age", "balance", "campaign", "deposit"
df_model <- df_model[sel_variables]
deposit_model1 <- glm(deposit ~ marital + housing + loan + tertiary + job + mid_age + balance + campaign, famil
model1_summary <- summary(deposit_model1)
print(summary(deposit_model1))
```

```
##
## Call:
## glm(formula = deposit ~ marital + housing + loan + tertiary +
##     job + mid_age + balance + campaign, family = binomial, data = na.omit(df_model))
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.354  -1.046  -0.689   1.105   2.978
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.15e+00   8.49e-02   13.60  < 2e-16 ***
## maritalmarried     -1.78e-01   4.94e-02   -3.60  0.00032 ***
## maritaldivorced     8.86e-03   7.31e-02    0.12  0.90357
## housingyes         -6.56e-01   4.28e-02  -15.34  < 2e-16 ***
## loanyes            -4.84e-01   6.29e-02   -7.69  1.4e-14 ***
## tertiaryyes         3.70e-01   5.90e-02    6.27  3.7e-10 ***
## jobblue-collar     -3.16e-01   7.71e-02   -4.11  4.0e-05 ***
## jobentrepreneur    -4.77e-01   1.37e-01   -3.50  0.00047 ***
## jobhousemaid     -5.15e-01   1.44e-01   -3.58  0.00034 ***
## jobmanagement    -2.34e-01   8.43e-02   -2.77  0.00554 **
## jobretired        2.09e-01   1.06e-01    1.97  0.04910 *
## jobself-employed -3.52e-01   1.25e-01   -2.83  0.00472 **
## jobservices      -2.44e-01   9.20e-02   -2.65  0.00808 **
## jobstudent        4.32e-01   1.60e-01    2.71  0.00682 **
## jobtechnician    -1.51e-01   7.71e-02   -1.95  0.05076 .
## jobunemployed     1.31e-01   1.27e-01    1.03  0.30091
## mid_ageyes       -6.63e-01   5.99e-02  -11.06  < 2e-16 ***
## balance           3.90e-05   7.47e-06    5.22  1.8e-07 ***
## campaign         -1.13e-01   9.87e-03  -11.46  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14709  on 10633  degrees of freedom
## Residual deviance: 13535  on 10615  degrees of freedom
## AIC: 13573
##
## Number of Fisher Scoring iterations: 4
```

```
print(exp(confint(deposit_model1)))
```

```
## Waiting for profiling to be done...
```

```
##                   2.5 % 97.5 %
## (Intercept)      2.6877 3.7492
## maritalmarried   0.7599 0.9224
## maritaldivorced  0.8742 1.1644
## housingyes       0.4772 0.5643
## loanyes          0.5445 0.6968
## tertiaryyes      1.2896 1.6252
## jobblue-collar   0.6265 0.8475
## jobentrepreneur  0.4737 0.8095
## jobhousemaid     0.4500 0.7910
## jobmanagement    0.6709 0.9336
## jobretired       1.0012 1.5190
## jobself-employed 0.5506 0.8973
## jobservices      0.6541 0.9383
## jobstudent       1.1318 2.1186
## jobtechnician    0.7395 1.0005
## jobunemployed    0.8897 1.4635
## mid_ageyes       0.4581 0.5795
## balance          1.0000 1.0001
## campaign         0.8756 0.9102
```

```
df_model %<>%
  mutate(
    log_campaign=log(campaign)
  )
deposit_vs_campaign <- glm(deposit ~ campaign + log_campaign, family=binomial,
                     data=na.omit(df_model))
vif_res <- vif(deposit_model1)
avg_vif <- mean(vif_res[,3])
# Durbin-Watson test dwtest
dw_res <- dwtest(deposit_model1)
# Marital is changed to married or not
df_model2 <- df_model %>%
  mutate(married=as.factor(if_else(marital=="married", "yes", "no"))) %>%
  select(!c("marital","balance"))
deposit_model2 <- glm(deposit ~ housing + loan + tertiary + mid_age, family=binomial, data=na.omit(df_model2))
```

```
model2_summary <- summary(deposit_model2)
model2_summary
```

```
##
## Call:
## glm(formula = deposit ~ housing + loan + tertiary + mid_age,
##     family = binomial, data = na.omit(df_model2))
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.721  -1.038  -0.709   1.175   1.735
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.8681     0.0527   16.46   <2e-16 ***
## housingyes   -0.6989     0.0409  -17.08   <2e-16 ***
## loanyes      -0.5613     0.0619   -9.07   <2e-16 ***
## tertiaryyes   0.3550     0.0428    8.29   <2e-16 ***
## mid_ageyes   -0.8612     0.0531  -16.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14709  on 10633  degrees of freedom
## Residual deviance: 13836  on 10629  degrees of freedom
## AIC: 13846
##
## Number of Fisher Scoring iterations: 4
```

```
exp(confint(deposit_model2))
```

```
## Waiting for profiling to be done...
```

```
##              2.5 % 97.5 %
## (Intercept) 2.1496 2.6432
## housingyes  0.4588 0.5386
## loanyes     0.5051 0.6437
## tertiaryyes 1.3114 1.5512
## mid_ageyes  0.3807 0.4688
```

```
# Durbin-Watson test dwtest
dw_res2 <- dwtest(deposit_model2)
```