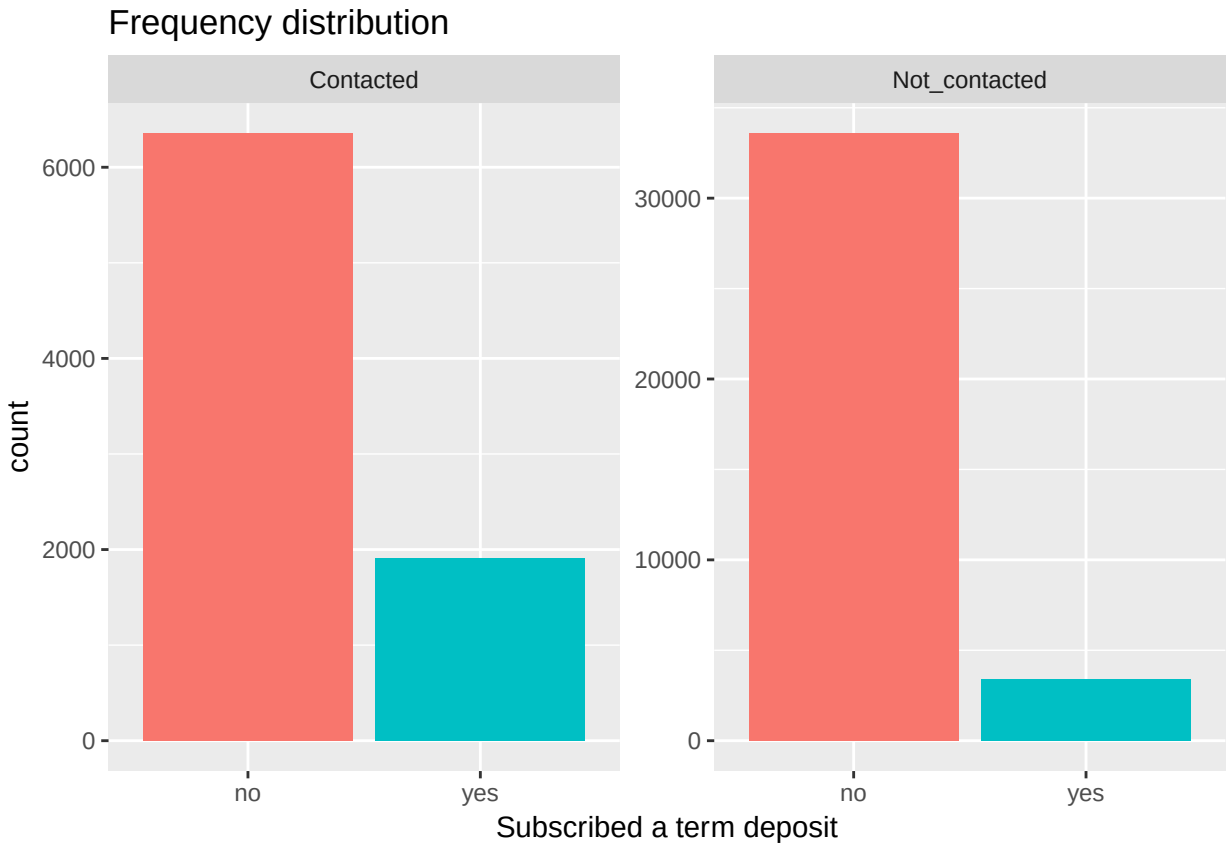
2nd & 3rd tasks

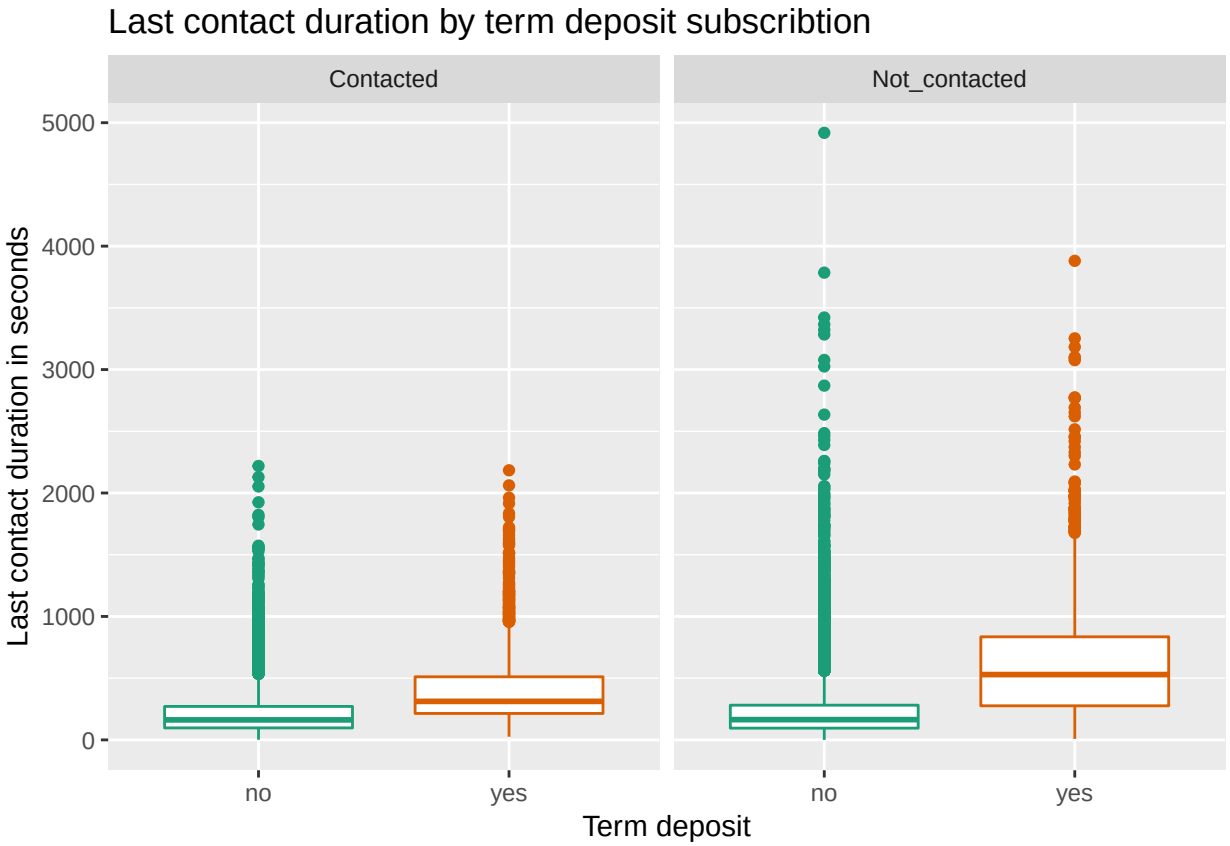
Manvydas Sokolovas

10/29/2018

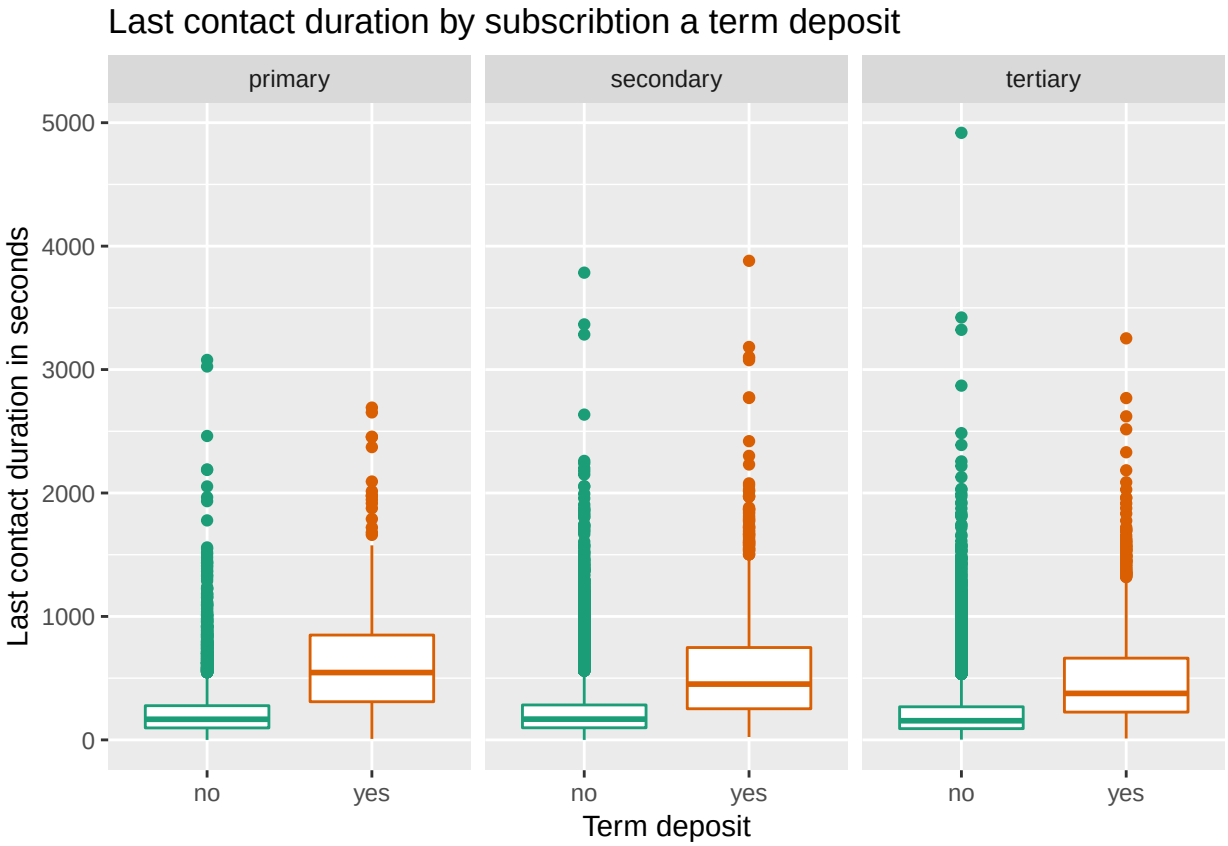
2) DATA VISUALIZATION TASK



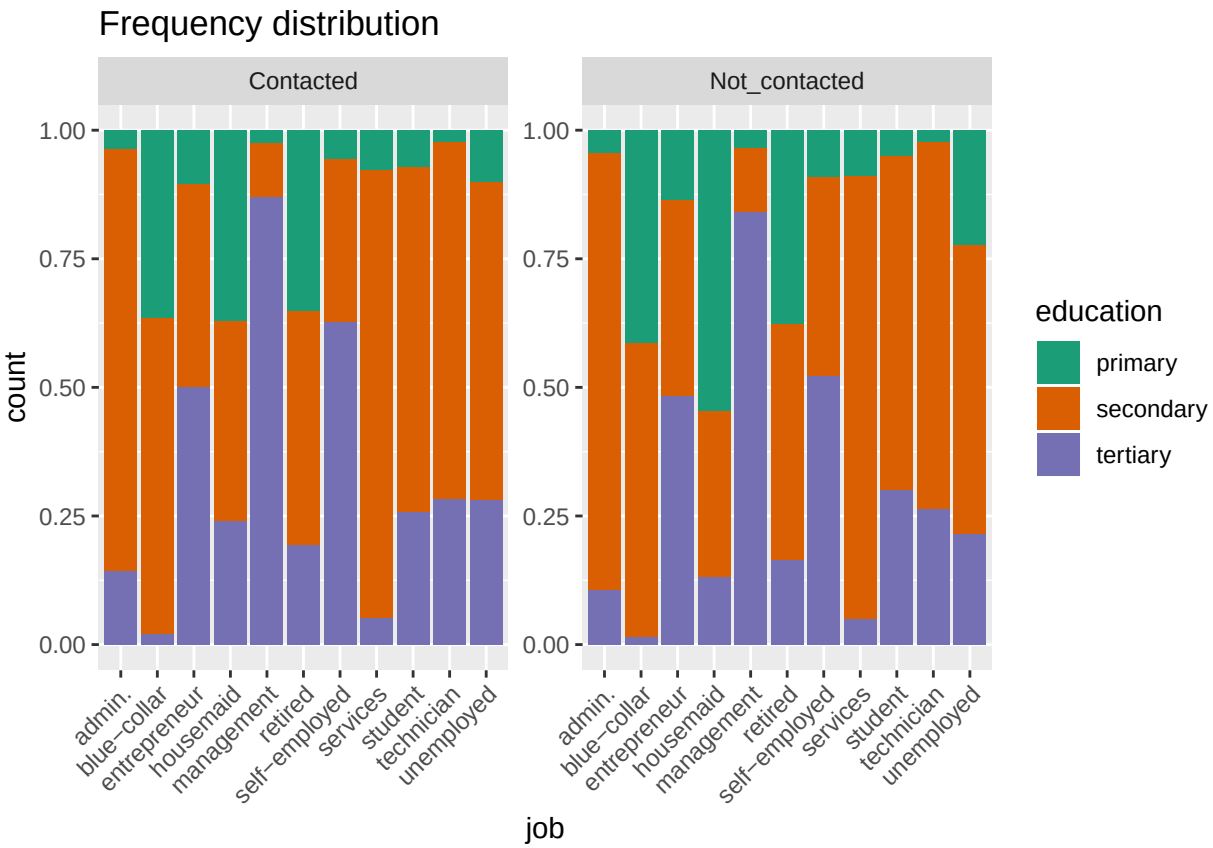
- Clients who were contacted before have way better subscribtion a term deposit ratio.



- Last contact duration with clients who were not contacted before looks more spreaded. It could be because they needed more information or just because there are 5 times more data points. In both cases clients who subscribed a term deposit had longer contacts.



- Boxplots covers each other. That means they have some similar points. But we can see that clients who subscribed a term deposit had longer last contact duration and more educated people had a slightly shorter contact.



- It's easy to see that people with better education have more qualified jobs. Very low percentage of blue-collar workers have tertiary education and more than 80% of management workers have tertiary education. Target groups did not change after previous campaign.

3) MODELLING TASK

Model summary for previously contacted customers

```
##
## Call:
## glm(formula = y ~ job + education + housing + loan + contact +
##     day_of_week + season + duration + campaign + pdays, family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3084  -0.5985  -0.3708  -0.1758   2.6194
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.4165498  0.1801282  -2.313 0.020749 *
## job1               -0.3303450  0.0994338  -3.322 0.000893 ***
## educationsecondary  0.2209023  0.1285879   1.718 0.085813 .
## educationtertiary   0.4357563  0.1380511   3.156 0.001597 **
## housingyes         -1.0644448  0.0822775 -12.937  < 2e-16 ***
## loanyes            -0.7088590  0.1328199  -5.337 9.45e-08 ***
```

```
## contacttelephone   -0.3600966  0.1471784  -2.447 0.014418 *
## day_of_week1       -1.6096249  0.4255266  -3.783 0.000155 ***
## seasonspring       -1.3528544  0.1231766 -10.983  < 2e-16 ***
## seasonsummer       -0.0939160  0.1255765  -0.748 0.454533
## seasonwinter       -1.1877472  0.1191914  -9.965  < 2e-16 ***
## duration            0.0036176  0.0001646  21.979  < 2e-16 ***
## campaign           -0.1347812  0.0290885  -4.633 3.60e-06 ***
## pdays              -0.0008656  0.0003461  -2.501 0.012392 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6341.4  on 5880  degrees of freedom
## Residual deviance: 4722.5  on 5867  degrees of freedom
## AIC: 4750.5
##
## Number of Fisher Scoring iterations: 5
```

- We can see that majority of chosen variables are significant. There is a strong connection between y and housing loan (if person have housing loan then there is lower chances that he will subscribe a term deposit), there are less subscriptions in spring and winter. Negative estimate values lowers the probability that person will subscribe a term deposit.

Anova

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: y
##
## Terms added sequentially (first to last)
##
##
##             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                         5880     6341.4
## job          1   146.60      5879     6194.8 < 2.2e-16 ***
## education    2    17.84      5877     6177.0 0.0001335 ***
## housing      1   480.11      5876     5696.9 < 2.2e-16 ***
## loan         1    60.96      5875     5635.9 5.815e-15 ***
## contact      1     7.31      5874     5628.6 0.0068617 **
## day_of_week  1    29.02      5873     5599.6 7.175e-08 ***
## season       3   219.24      5870     5380.4 < 2.2e-16 ***
## duration     1   627.04      5869     4753.3 < 2.2e-16 ***
## campaign     1    24.51      5868     4728.8 7.409e-07 ***
## pdays        1     6.28      5867     4722.5 0.0121903 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- From chi square values we can figure out the significance of each variable.

Model accuracy

```
## [1] "Accuracy 0.8108"
```

- Model predicts pretty good. About 80% of trial data were predicted correctly.

Model summary for previously not contacted customers

```
##
## Call:
## glm(formula = y ~ job + marital + education + balance + housing +
##     loan + contact + month + duration + campaign + day_of_week +
##     ageclass, family = binomial(link = "logit"), data = train2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.7852  -0.3842  -0.2644  -0.1782   3.2161
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -2.150e+00  1.824e-01 -11.785  < 2e-16 ***
## job1                 -2.091e-01  7.712e-02  -2.712  0.00670 **
## maritalmarried       -2.372e-01  9.277e-02  -2.557  0.01057 *
## maritalsingle         1.059e-02  1.049e-01   0.101  0.91964
## educationsecondary    3.190e-01  1.003e-01   3.182  0.00146 **
## educationtertiary     4.297e-01  1.094e-01   3.927 8.61e-05 ***
## balance               9.870e-06  7.895e-06   1.250  0.21123
## housingyes           -8.483e-01  6.969e-02 -12.173  < 2e-16 ***
## loanyes              -4.512e-01  9.277e-02  -4.864 1.15e-06 ***
## contacttelephone     -2.475e-01  1.116e-01  -2.218  0.02654 *
## monthaug             -1.415e+00  1.132e-01 -12.501  < 2e-16 ***
## monthdec              1.117e+00  2.849e-01   3.920 8.85e-05 ***
## monthfeb             -6.281e-01  1.240e-01  -5.063 4.12e-07 ***
## monthjan             -1.743e+00  1.946e-01  -8.953  < 2e-16 ***
## monthjul             -1.442e+00  1.105e-01 -13.055  < 2e-16 ***
## monthjun              6.898e-01  1.459e-01   4.728 2.27e-06 ***
## monthmar              1.451e+00  1.713e-01   8.470  < 2e-16 ***
## monthmay             -6.269e-01  1.107e-01  -5.661 1.50e-08 ***
## monthnov             -1.341e+00  1.287e-01 -10.422  < 2e-16 ***
## monthoct              5.606e-01  1.713e-01   3.273  0.00106 **
## monthsep              5.622e-01  2.113e-01   2.661  0.00780 **
## duration              4.278e-03  1.002e-04  42.690  < 2e-16 ***
## campaign             -6.217e-02  1.427e-02  -4.357 1.32e-05 ***
## day_of_weekMonday    -9.977e-02  1.288e-01  -0.775  0.43853
## day_of_weekSaturday  -3.446e-03  8.724e-02  -0.040  0.96849
## day_of_weekSunday    -2.728e-01  9.501e-02  -2.872  0.00408 **
## day_of_weekThursday   4.312e-02  9.121e-02   0.473  0.63636
## day_of_weekTuesday    6.550e-01  2.550e-01   2.569  0.01020 *
## day_of_weekWednesday -1.051e-01  1.048e-01  -1.003  0.31610
## ageclass(34,50]      -1.845e-01  7.149e-02  -2.580  0.00988 **
## ageclass(50,70]       4.611e-02  9.045e-02   0.510  0.61020
## ageclass(70,95]       8.385e-01  1.938e-01   4.326 1.52e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 12481.1  on 17292  degrees of freedom
## Residual deviance:  8684.9  on 17261  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 8748.9
##
## Number of Fisher Scoring iterations: 6
```

- Variables for this model were chosen by stepAIC. Most of them are statisticaly significant.

Anova

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: y
##
## Terms added sequentially (first to last)
##
##
##             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                        17292    12481.1
## job          1    59.45     17291    12421.7 1.256e-14 ***
## marital      2    80.38     17289    12341.3 < 2.2e-16 ***
## education    2     3.80     17287    12337.5 0.1492263
## balance      1    15.52     17286    12322.0 8.168e-05 ***
## housing      1   138.74     17285    12183.2 < 2.2e-16 ***
## loan         1    75.94     17284    12107.3 < 2.2e-16 ***
## contact      1     2.26     17283    12105.0 0.1328516
## month       11   820.50     17272    11284.5 < 2.2e-16 ***
## duration     1  2518.79     17271     8765.7 < 2.2e-16 ***
## campaign     1    24.13     17270     8741.6 8.989e-07 ***
## day_of_week  6    23.21     17264     8718.4 0.0007284 ***
## ageclass     3    33.47     17261     8684.9 2.564e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- From chi square values we can figure out the significance of each variable.

Model accuracy

```
## Warning in `!=.default`(fitted.results, test$y): longer object length is
## not a multiple of shorter object length
```

```
## Warning in is.na(e1) | is.na(e2): longer object length is not a multiple of
## shorter object length
```

```
## [1] "Accuracy NA"
```

- Model predicts pretty good. About 85% of trial data were predicted correctly.