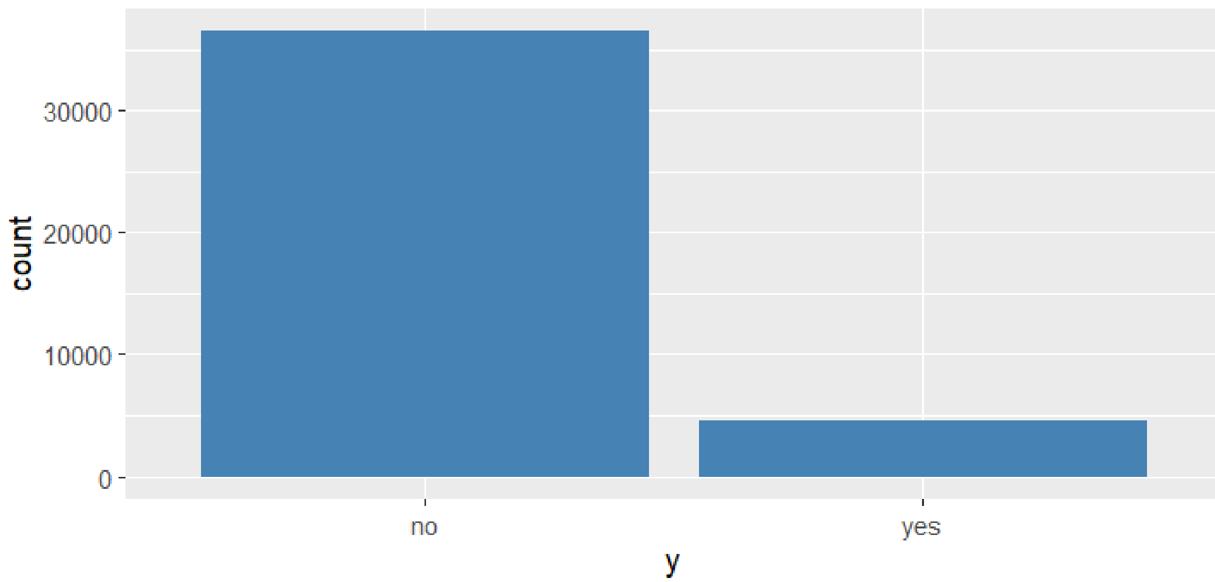
Matthew Voynovich
Data Analytics Assignment 6
Dr Eleish
5 December 2025

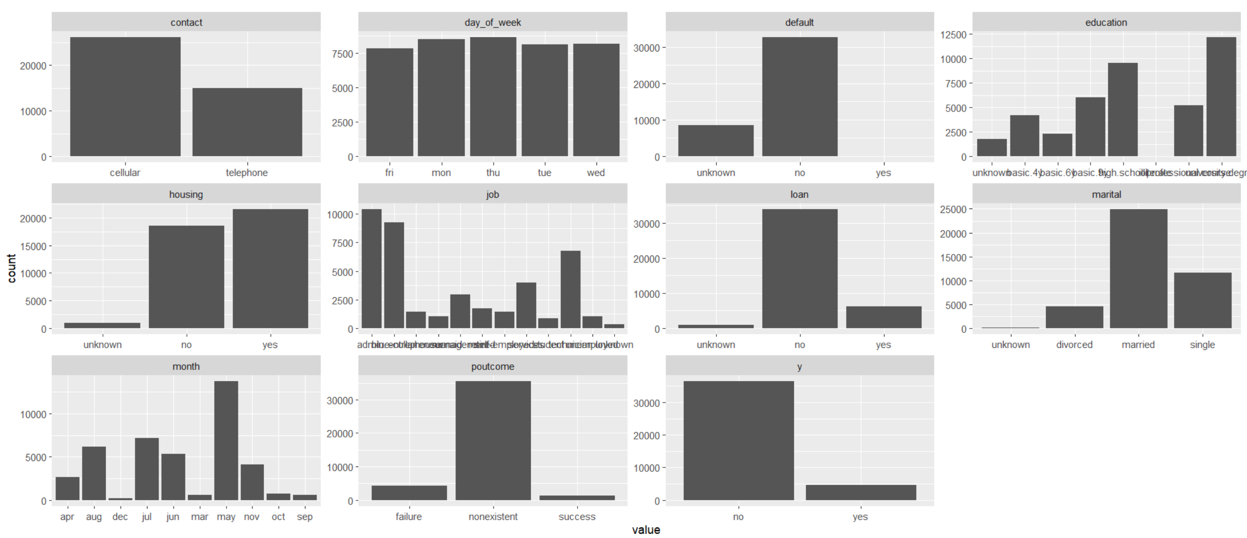
Dataset from:
https://archive.ics.uci.edu/dataset/222/bank+marketing

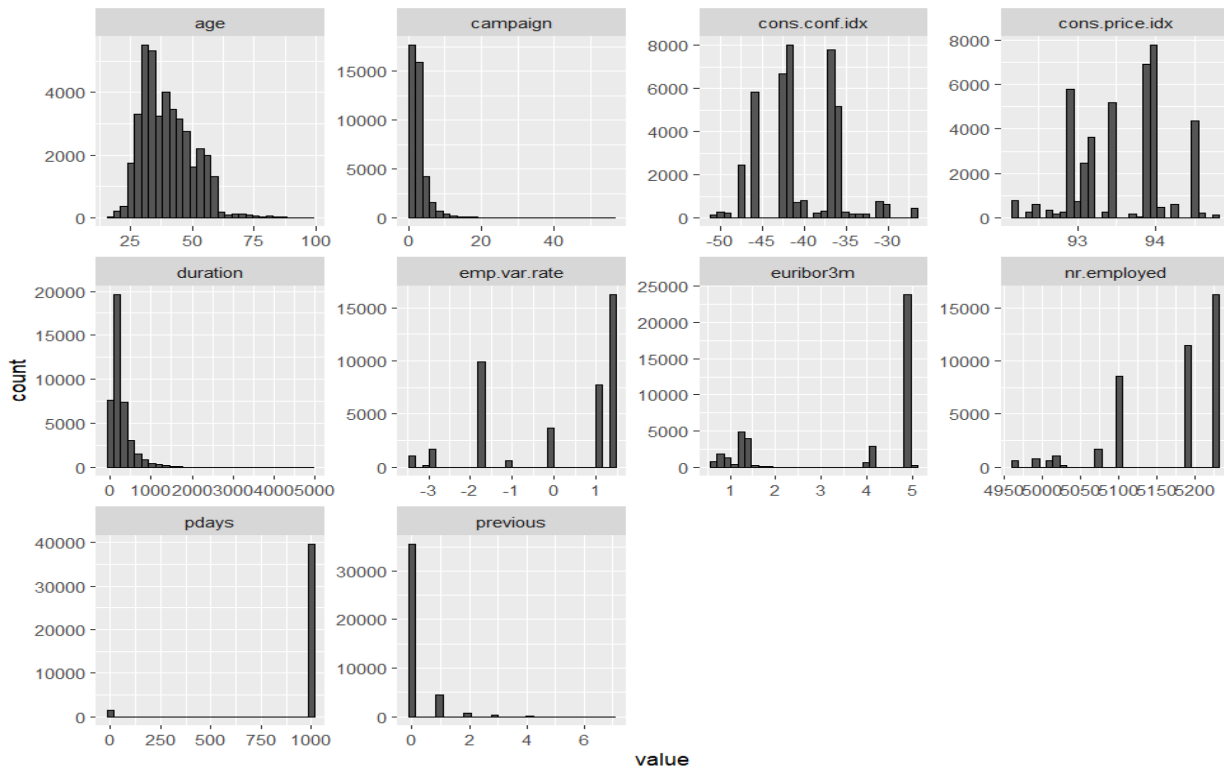
1. Exploratory Data Analysis (3%) Explore the statistical aspects of the dataset. Analyze the distributions and provide summaries of the relevant statistics. Perform any cleaning, transformations, interpolations, smoothing, outlier detection/ removal, etc. required on the data. Include figures and descriptions of this exploration and a short description of what you concluded (e.g. nature of distribution, indication of suitable model approaches you would try, etc.) Min.1 page text + graphics (required)

The data is related to direct marketing campaigns of a Portuguese banking institution. The variable y is for whether the client subscribed a term deposit. The data did not require any transformations or cleaning. I got the summary statistics for all features alongside plotting their distributions. This was done by transforming everything from wide format to long format (2 cols, var name and var value). I then did a ggplot with a facet wrap of the variables to get the distributions.

```
      age                 job           marital                   education        default     
 Min.   :17.00   admin.      :10422   divorced: 4612   university.degree  :12168   no     :32588  
 1st Qu.:32.00   blue-collar: 9254   married :24928   high.school        : 9515   unknown: 8597  
 Median :38.00   technician : 6743   single  :11568   basic.9y           : 6045   yes    :    3  
 Mean   :40.02   services   : 3969   unknown :   80   professional.course: 5243                  
 3rd Qu.:47.00   management : 2924                    basic.4y           : 4176                  
 Max.   :98.00   retired    : 1720                    basic.6y           : 2292                  
                 (Other)    : 6156                    (Other)            : 1749                  
    housing           loan            contact          month        day_of_week    duration     
 no     :18622   no     :33950   cellular :26144   may    :13769   fri:7827    Min.   :   0.0  
 unknown:  990   unknown:  990   telephone:15044   jul    : 7174   mon:8514    1st Qu.: 102.0  
 yes    :21576   yes    : 6248                     aug    : 6178   thu:8623    Median : 180.0  
                                                   jun    : 5318   tue:8090    Mean   : 258.3  
                                                   nov    : 4101   wed:8134    3rd Qu.: 319.0  
                                                   apr    : 2632               Max.   :4918.0  
                                                   (Other): 2016                                 
    campaign          pdays          previous            poutcome      emp.var.rate      cons.price.idx 
 Min.   : 1.000   Min.   :  0.0   Min.   :0.000   failure    : 4252   Min.   :-3.40000   Min.   :92.20  
 1st Qu.: 1.000   1st Qu.:999.0   1st Qu.:0.000   nonexistent:35563   1st Qu.:-1.80000   1st Qu.:93.08  
 Median : 2.000   Median :999.0   Median :0.000   success    : 1373   Median : 1.10000   Median :93.75  
 Mean   : 2.568   Mean   :962.5   Mean   :0.173                       Mean   : 0.08189   Mean   :93.58  
 3rd Qu.: 3.000   3rd Qu.:999.0   3rd Qu.:0.000                       3rd Qu.: 1.40000   3rd Qu.:93.99  
 Max.   :56.000   Max.   :999.0   Max.   :7.000                       Max.   : 1.40000   Max.   :94.77  
                                                                                                        
 cons.conf.idx     euribor3m      nr.employed       y        
 Min.   :-50.8   Min.   :0.634   Min.   :4964   no :36548  
 1st Qu.:-42.7   1st Qu.:1.344   1st Qu.:5099   yes: 4640  
 Median :-41.8   Median :4.857   Median :5191              
 Mean   :-40.5   Mean   :3.621   Mean   :5167              
 3rd Qu.:-36.4   3rd Qu.:4.961   3rd Qu.:5228              
 Max.   :-26.9   Max.   :5.045   Max.   :5228              
```


count
30000
20000
10000
0
no
yes
y


contact
cellular
telephone
day_of_week
fri
mon
thu
tue
wed
default
unknown
no
yes
education
housing
unknown
no
yes
job
loan
unknown
no
yes
marital
unknown
divorced
married
single
month
apr
aug
dec
jul
jun
mar
may
nov
oct
sep
poutcome
failure
nonexistent
success
y
no
yes
count
value


age
campaign
cons.conf.idx
cons.price.idx
duration
emp.var.rate
euribor3m
nr.employed
pdays
previous
count
value

2. Model Development, Validation and Optimization (10% 4000-level / 7% 6000-level) Develop and evaluate three (4000-level) or four (6000-level) or more models. If possible, these models should cover more than one objective, i.e. regression, classification, clustering. Consider the efect of dimension reduction of the dataset on model performance. Diferent models means diferent combinations of an algorithm and a formula (input and output features). The choice of independent and response variables is up to you.

Explain why you chose them. Construct the models, test/ validate them. Briefly explain the validation approach. You can use any method(s) covered in the course. Include your code in your submission. Compare model results if applicable. Report the results of the model (fits, coeficients, sample trees, other measures of fit/ importance, etc., predictors and summary statistics). Min. 2 pages of text + graphics (required).

I did two different objectives. The first objective was classification of the target variable y. I used two separate models to evaluate this objective. The first model was a random forest classifier with 500 trees. I chose this model as I wanted to see if there was a complex non linear relationship across the features with whether a client made a bank term deposit. The other model I used was logistic regression. This model was accurate still but worse than the random forest classifier. These were evaluated using a train and test set (70% train) and used accuracy, precision, recall, and f1 score alongside a confusion matrix. The next objective was regression for the campaign variable (number of times client was contacted). This model was evaluated using mean squared error and mean absolute error. Finally weights and residual plot was generated.

<u>Random Forest Metrics</u>

```
Classification Performance - Training Set
> print(cm_train$table)  # print confusion matrix
          Reference
Prediction    no   yes
       no  25582    69
       yes     2  3179
> cat("Accuracy:", accuracy_train, "\n")
Accuracy: 0.9975375
> cat("Precision:", precision_train, "\n")
Precision: 0.99731
> cat("Recall:", recall_train, "\n")
Recall: 0.9999218
> cat("F1 Score:", f1_train, "\n\n")
F1 Score: 0.9986142

Classification Performance - Test Set
> print(cm_test$table)  # print confusion matrix
          Reference
Prediction    no   yes
       no  10526   653
       yes   438   739
> cat("Accuracy:", accuracy_test, "\n")
Accuracy: 0.9117028
> cat("Precision:", precision_test, "\n")
Precision: 0.9415869
> cat("Recall:", recall_test, "\n")
Recall: 0.9600511
> cat("F1 Score:", f1_test, "\n")
F1 Score: 0.9507294
```

```
> print(feat_imp)
                       no         yes MeanDecreaseAccuracy MeanDecreaseGini
age            19.6650740   4.7436180           20.5811930        413.88257
job            32.0727993  -7.3652800           23.0859326        363.45145
marital         4.9204858  -2.6373341            2.6202843        111.71491
education      15.1726919   1.1402318           13.8423139        261.33107
default        -2.2531166  15.0625663            7.8372250         38.57692
housing         1.2614527  -4.0089573           -1.4740232         92.28155
loan           -0.5479091  -0.8936436           -0.9980942         71.79782
contact         6.0188655  25.7417107            8.8774408         49.34576
month          26.5105021   3.9107433           27.3358709        169.64899
day_of_week    25.3673207   7.2532979           26.2970920        242.60685
duration      138.9187175 204.7995333          206.8564872       1674.69706
campaign        9.2206486  14.3955436           17.0592957        199.61642
pdays           3.5310901  36.7512657           25.8658459        185.95023
previous        6.1672447   3.9371623            7.4743762         68.79238
poutcome       11.1940764  16.4667746           17.7744583        167.75779
emp.var.rate   17.9311442   8.7481248           18.8524654        119.01992
cons.price.idx 18.5639164  -0.5806048           18.7024000        116.53409
cons.conf.idx  15.8907996   1.9608324           16.5755580        144.46950
euribor3m      30.1352154  18.2382904           33.4174957        574.35427
nr.employed    19.7878526  22.5756729           23.6177023        357.49675
```

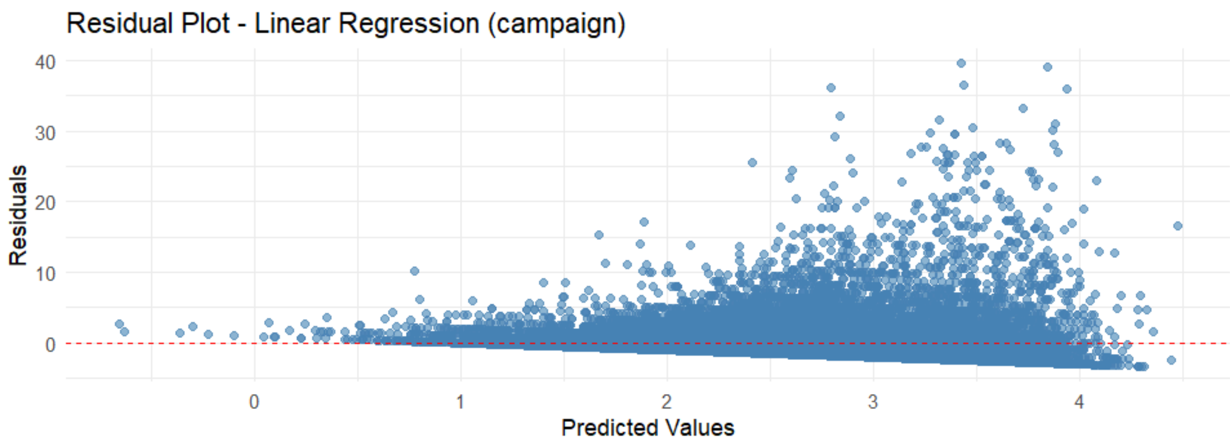
Logistic Regression

```
Logistic Regression - Training Set
> print(cm_train_logit$table)
          Reference
Prediction    no   yes
       no  24910  1849
       yes   674  1399
> cat("Accuracy:", accuracy_train_logit, "\n")
Accuracy: 0.9124931
> cat("Precision:", precision_train_logit, "\n")
Precision: 0.9309018
> cat("Recall:", recall_train_logit, "\n")
Recall: 0.9736554
> cat("F1 Score:", f1_train_logit, "\n\n")
F1 Score: 0.9517987
```

```
Logistic Regression - Test Set
> print(cm_test_logit$table)
          Reference
Prediction    no   yes
       no  10652   798
       yes   312   594
> cat("Accuracy:", accuracy_test_logit, "\n")
Accuracy: 0.9101651
> cat("Precision:", precision_test_logit, "\n")
Precision: 0.9303057
> cat("Recall:", recall_test_logit, "\n")
Recall: 0.9715432
> cat("F1 Score:", f1_test_logit, "\n\n")
F1 Score: 0.9504774
```

Linear Regression

```
> cat("Linear Regression MSE - Training Set:", mse_train, "\n")
Linear Regression MSE - Training Set: 7.334965
> cat("Linear Regression MSE - Test Set:", mse_test, "\n")
Linear Regression MSE - Test Set: 7.293563
> cat("Linear Regression Mean Absolute Error - Training Set:", mae_train, "\n")
Linear Regression Mean Absolute Error - Training Set: 1.600327
> cat("Linear Regression Mean Absolute Error - Test Set:", mae_test, "\n")
Linear Regression Mean Absolute Error - Test Set: 1.580199
```



```
> print(weights)
                  (Intercept)                           age 
                -4.068380e+01                  3.235802e-03 
              jobblue-collar               jobentrepreneur 
                -1.518917e-01                 -6.877704e-02 
                 jobhousemaid                 jobmanagement 
                -2.673365e-02                  4.528459e-03 
                  jobretired              jobself-employed 
                 5.457581e-02                  7.097699e-02 
                 jobservices                    jobstudent 
                -9.938007e-02                 -1.639930e-01 
               jobtechnician                 jobunemployed 
                -7.754265e-02                  2.811235e-02 
                  jobunknown                maritalmarried 
                -1.479850e-01                 -3.690300e-02 
               maritalsingle                maritalunknown 
                 6.582859e-03                  7.314091e-01 
           educationbasic.6y             educationbasic.9y 
                -7.240540e-02                  5.799961e-03 
       educationhigh.school          educationilliterate 
                 6.904560e-02                 -2.959848e-01 
educationprofessional.course   educationuniversity.degree 
                 3.659141e-03                  4.036213e-02 
            educationunknown                defaultunknown 
                 4.911269e-02                  4.454198e-02 
                  defaultyes                housingunknown 
                -1.061218e+00                 -7.121594e-02 
                  housingyes                   loanunknown 
                -3.291092e-02                            NA 
                     loanyes              contacttelephone 
                 5.309074e-02                  6.096281e-01 
                    monthaug                      monthdec 
                 4.040205e-01                  1.103982e+00 
                    monthjul                      monthjun 
                 1.002968e+00                  6.796677e-01 
                    monthmar                      monthmay 
                 3.496972e-01                  2.832586e-01 
                    monthnov                      monthoct 
                 4.059842e-01                  6.298680e-01 
                    monthsep               day_of_weekmon 
                 7.270800e-01                 -1.161568e-01 
              day_of_weekthu                day_of_weektue 
                -1.614403e-01                 -2.996407e-01 
              day_of_weekwed                      duration 
                -3.030210e-01                 -7.856850e-04 
                       pdays                      previous 
                -1.848679e-04                  7.209526e-02 
         poutcomenonexistent               poutcomesuccess 
                 2.125761e-01                 -2.768348e-01 
                emp.var.rate                cons.price.idx 
                 8.398994e-01                 -1.474739e-01 
               cons.conf.idx                     euribor3m 
                 1.283602e-02                 -1.098402e+00 
                 nr.employed                          yyes 
                 1.181778e-02                  1.065831e-01 
> 
```

3. Decisions (2% 4000-level / 5% 6000-level) Describe your conclusions from the model fits, predictions and how well (or not) it could be used for decisions and why. Min. 1/2 page of text + graphics

The models all fit fairly well overall. Among them, the most usable was the random forest classification model, which achieved the highest accuracy and consistently outperformed the logistic regression model. This makes sense because random forests can naturally capture non-linear relationships and interactions between features, while logistic regression assumes linear separability. Additionally, logistic regression struggled due to the presence of intentional null values in certain features, which the random forest handled more gracefully.

For regression, I initially attempted to use a random forest regressor, but each attempt caused R to freeze or take an excessively long time to complete. Because of this, I switched to lighter models, starting with linear regression and applying multiple non-linear feature transformations such as polynomial terms. After evaluating several models, the standard linear regression model without any transformations produced the best MSE and MAE.

When plotting the residuals, the pattern indicated a non-linear structure in the data. The residuals were not randomly scattered, suggesting that the linear model even though it performed best among models I tested, it is likely not the best overall.

Given this, I would recommend using a decision tree regressor or another tree-based method. These models tend to capture non-linear relationships more effectively and would likely yield better performance without requiring manual feature transformations.