Time Series Analysis for Airplane Departures & Logistic Regression for Bank Marketing Campaign

Rian Dwi Putra
MSc Data Analytics
National College of Ireland
Dublin, Ireland
x22108637@student.ncirl.ie

Abstract— Trends, seasonal fluctuations, irregular cycles, and occasional shifts in level or variability are typical patterns that can be seen in time series data in a variety of scientific fields, including medicine, the environment, business, economics, and medicine. Analyzing such series often aims to estimate the effect of known exogenous interventions, find unknown interventions, and extrapolate the data's dynamic pattern to predict future observations.

One of the most widely used regression techniques for modeling binary dependent variables is logistic regression analysis. A mathematical modeling technique called logistic regression is used to define the relationship between independent variables like X1, X2, etc., and a binary dependent variable called Y, which can be coded as either 0 or 1 for either of two possible categories. Continuous, discrete, binary, or a combination of the three can serve as the independent variables. This paper examines logistic regression models.

Keywords—R, Statistics, Logistic Regression, Time Series Analysis

I. INTRODUCTION

- Airplane Departure

The expression "running a city is like running an airport" comes from a saying. Like a city, an airport has numerous facilities, systems, users, employees, and regulations. Like how cities thrive on trade and commerce with other cities, airports are successful in part because they are able to successfully serve as the location where passengers and cargo travel to and from other airports. Air terminals should be successful as a part of the country's organization of air terminals as urban communities find their place in the economy of their province, state, and nation.

The air travel industry will be liberated a long time from this point as we enter the second 10 years of the 21st hundred years. This significant step has forever changed the civil aviation industry. Globally, air travel has been privatized and liberalized. Today, a lot of the larger and smaller airports are heavily privatized; In parallel, there has been a significant decline in national governments' involvement in airport ownership and management.

Another significant development since the 1980s was the establishment of airline alliances. These all influence how airplane as of now use air terminals, requiring association arrangement for greatest business benefit. The development of massive long-range aircraft has also influenced airline equipment. Electronic help has come about because of the spread of the Web, influencing web-based booking, tagging, registration, flight following, traveler treatment of postponements and undoing, etc. The International Air Transport Association (IATA) has backed and encouraged this endeavor. E-documentation has been implemented to replace paper archives for cargo transportation.

The airport is an essential component of the air transport system because it is the physical location where a modal transfer is made from the air mode to the land mode or vice versa. As a result, it serves as the amalgamation of the three fundamental components of the structure of air transportation:

The airport, its commercial and operational concessionaires, tenants, and partners, as well as the airways control system; the airline; the user; and the planning and operation of airports all require consideration of the interactions between these three major components or system actors for the purposes of this discussion. For the system to function properly, each actor needs to reach equilibrium with the other two. If this isn't done, things won't be perfect, as several bad things that are signs of bad operation show. Any one of these things could eventually result in a decrease in the overall scale of operations at the airport facility or, at the very least, a decrease in the total share of traffic as traffic is drawn elsewhere in a situation where there is no restriction on competition.

- Bank Marketing Campaign

Marketing is now an essential part of any business's efforts to promote products and services that are geared toward customers. It is without a doubt also a significant event in the fields of economics and social science. In economics, marketing-related issues have been studied using multivariate statistical analysis. The legitimate application of reasonable strategies to a particular problem has been a constant challenge that necessitates the utilization of information regarding the potential outcomes of standard procedures.

Banks and other monetary organizations can zero in on clients who are probably going to buy into their items, offers, and different packs thanks to coordinate promoting. Monetary organizations find it hard to recognize this gathering of clients often.

Due to the increasing number of advertising campaigns, its impact on the general public has diminished over time. Due to competition and economic pressures, marketing managers have also invested in targeted campaigns with a strict and rigorous selection of contacts.

There are two main ways that businesses can advertise their products and services: through either coordinated promoting, which centers around a specific gathering of contacts, or mass missions, which are focused on a sweeping public. According to a similar report, positive responses to mass missions are typically extremely low—under 1%—in today's global competitive environment. In contrast, targeted marketing targets individuals who are more likely to be interested in a specific product or service due to its efficiency. However, there are some disadvantages to directed marketing, such as the risk of instilling a hostile attitude toward banks due to the invasion of privacy.

Businesses can gain a competitive advantage over their "peers" and improve their financial standing by outsourcing their work through marketing campaigns.

Businesses use direct marketing to reach specific customer groups for a specific objective. Utilizing remote communication centers to communicate with the client simplifies campaign administration. Customers can communicate with a business through a variety of channels, including mobile phones and fixed lines, thanks to these call centers. Because of its far-off nature, item promoting brought out through a contact community is alluded to as selling. The two primary methods by which businesses promote their products are mass crusades, which are directed at the entire population, and direct campaigns, which are directed only at a specific group of people.

The formal review shows that the mass campaign doesn't work very well. Commonly, short of what one percent of the populace will uphold the mass mission. Strangely, direct campaigns are much more effective at engaging people because they only target a small group of people who are thought to be more interested in the product being advertised. Matching characteristics of the customer, such as age, marital status, educational level, and so forth makes it extremely difficult for statistics to select these potential customers. and additional characteristics, like loan application and repayment, to a variety of outputs, like whether a customer will subscribe to a term deposit or not.

II. DATA SOURCE & RESEARCH QUESTION

- Dataset Description

The experiments presented in this study are based on these datasets:

1. The 'departure.csv' datafile, uploaded on Moodle, is a monthly time series of number of departures from Ireland via airports, commencing in 2010 to September 2022.
2. The bank file, uploaded on Moodle, contains details of a marketing campaign that aims to convince the customer to buy a bank product.
   -Age
   -Job
   -Marital status
   -Education
   -Credit
   -Housing
   -Has Mortgage
   -Loan: has a personal loan?
   -Contact communication type
   -Month
   -Last contact day of the week- Duration: duration of the last contact, in seconds (numeric).
   - Campaign: number of contacts made for this client and during this campaign
   - pdays: the number of days since the client was last contacted
   - previous: number of contacts made before and for this client as a result of this campaign
   - poutcome: outcome of the previous promotion
   - y: has the customer signed up for the bank product?

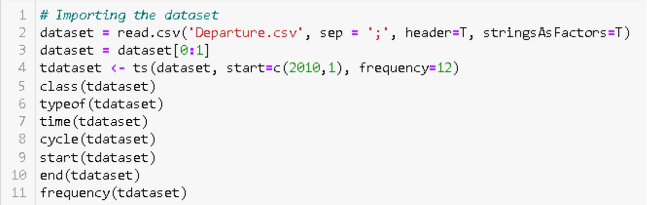
- Research Questions

Research Question in this study will be written on every section in Part III. Methodology.

III. METHODOLOGY

- **Airplane Departure**

First, this study needs to read the file and see the structure of dataset by using this code.



```
# Importing the dataset
dataset = read.csv('Departure.csv', sep = ';', header=T, stringsAsFactors=T)
dataset = dataset[0:1]
tdataset <- ts(dataset, start=c(2010,1), frequency=12)
class(tdataset)
typeof(tdataset)
time(tdataset)
cycle(tdataset)
start(tdataset)
end(tdataset)
frequency(tdataset)
```

1. A preliminary assessment of the nature and components of the raw time series, using visualizations as appropriate.
   a. Boxplots

      Boxplots show how evenly distributed a data set's data are. It creates three quartiles for the data set. The dataset's minimum, maximum, median, first quartile, and third quartile are all depicted on this graph. By creating boxplots for each of the data sets, it is also useful for comparing how the data are distributed across them.

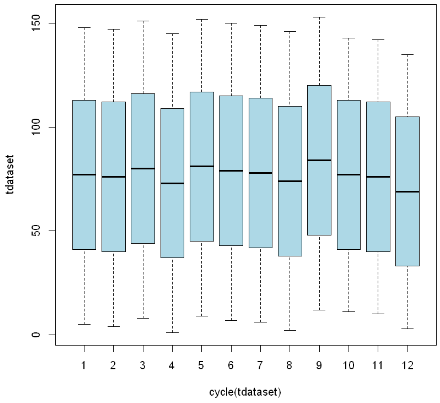


Figure. Boxplot

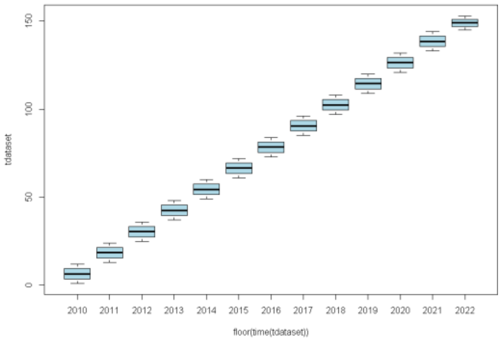


Figure. Boxplot

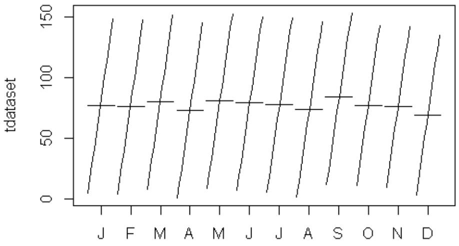


Figure. Monthplot

b. Seasonal Plot

The only difference between a seasonal plot and a standard time series plot is that the x-axis displays data from each season. This kind of plot makes it easier to see the underlying seasonal pattern and is especially useful for figuring out which years the pattern changes.

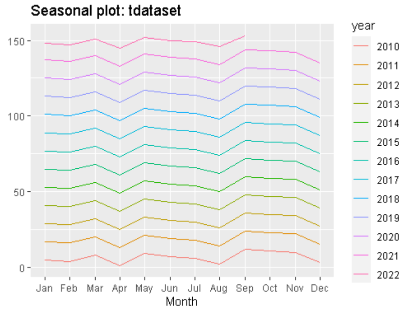


Figure. Seasonal Plot

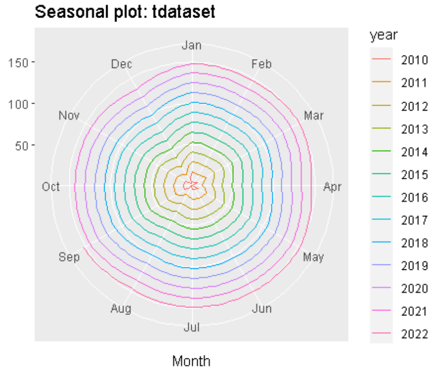
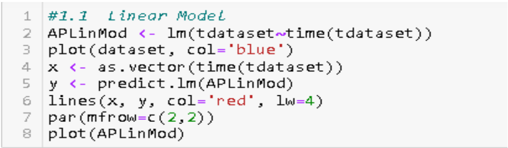


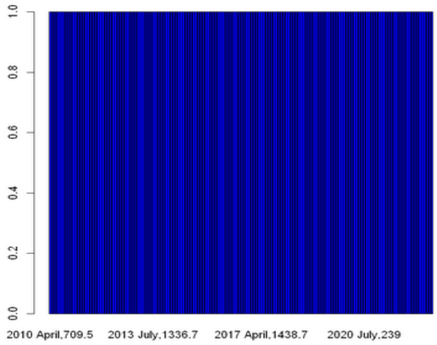
Figure. Seasonal Plot

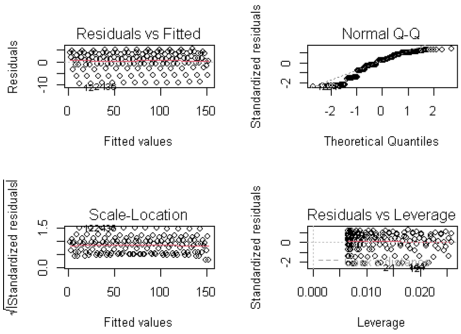
c. Linear Model

A continuous response variable is portrayed in linear models as a function of one or more predictor variables. The formula suggests a relationship between the left and right sides that is determined by "linear" or "constant" parameters for each term that are solved for to minimize the total deviation of the data from the model. Therefore the model is referred to as a linear model.



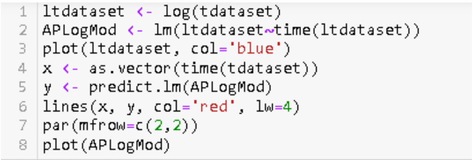
```
#1.1  Linear Model
APLinMod <- lm(tdataset~time(tdataset))
plot(dataset, col='blue')
x <- as.vector(time(tdataset))
y <- predict.lm(APLinMod)
lines(x, y, col='red', lw=4)
par(mfrow=c(2,2))
plot(APLinMod)
```



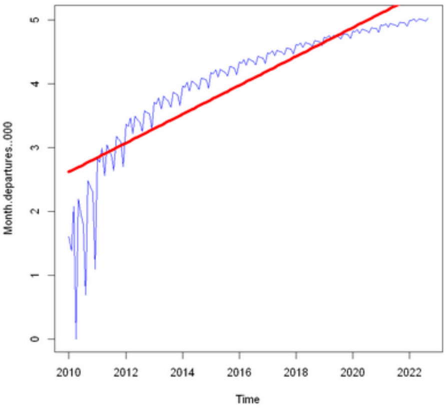


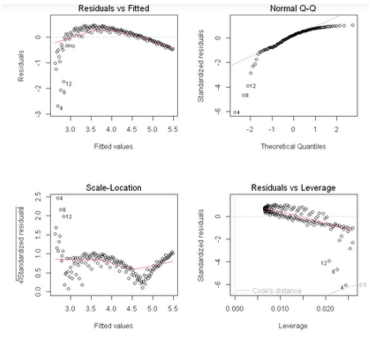
d. Logarithmic Model

A model that measures the magnitude of the thing it is measuring is called a logarithmic model. It can also be thought of as an exponential model's opposite. To investigate the properties of exponential functions and solve exponential equations, logarithms can be utilized.



```
ltdataset <- log(tdataset)
APLogMod <- lm(ltdataset~time(ltdataset))
plot(ltdataset, col='blue')
x <- as.vector(time(tdataset))
y <- predict.lm(APLogMod)
lines(x, y, col='red', lw=4)
par(mfrow=c(2,2))
plot(APLogMod)
```



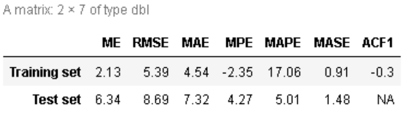


2. Estimation and discussion of suitable time series models.

a. Simple Exponential Smoothing

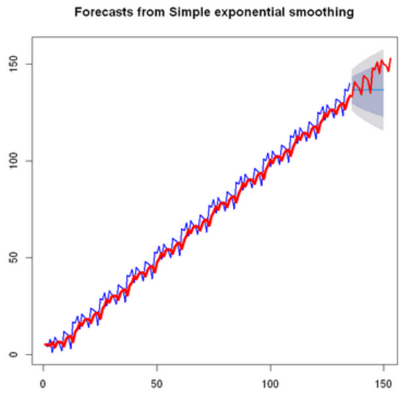
In an exponential smoothing calculation, Error, Trend, and Seasonal components are combined. Each term can be added together, multiplied together, or left out of the model entirely. These three terms (Error, Trend, and Season) are alluded to as ETS.

- Using Simple Exponential Smoothing - SES

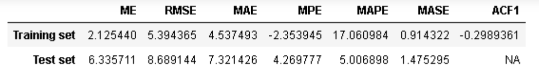


A matrix: 2 × 7 of type dbl

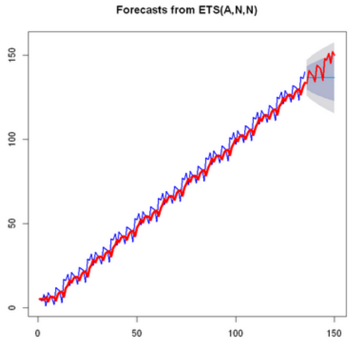
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	2.13	5.39	4.54	-2.35	17.06	0.91	-0.3
Test set	6.34	8.69	7.32	4.27	5.01	1.48	NA

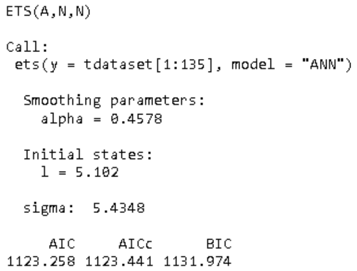


- Using ets with model='ANN'



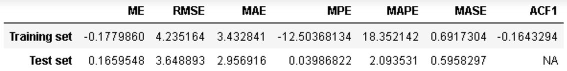
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	2.125440	5.394365	4.537493	-2.353945	17.060984	0.914322	-0.2989361
Test set	6.335711	8.689144	7.321426	4.269777	5.006898	1.475295	NA



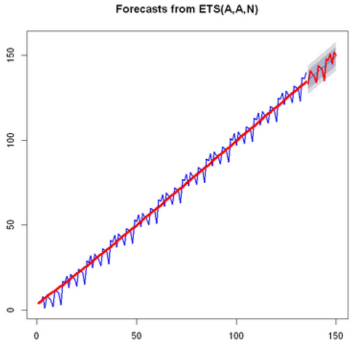


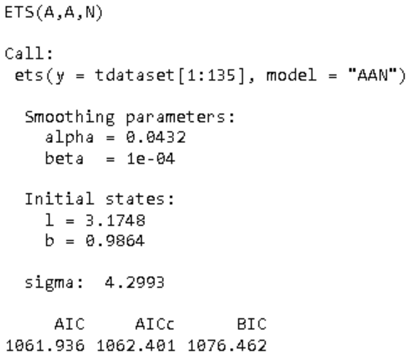
```
ETS(A,N,N)

Call:
 ets(y = tdataset[1:135], model = "ANN")

  Smoothing parameters:
    alpha = 0.4578

  Initial states:
    l = 5.102

  sigma:  5.4348

     AIC     AICc      BIC 
1123.258 1123.441 1131.974
```

- Using ets with model='AAN'



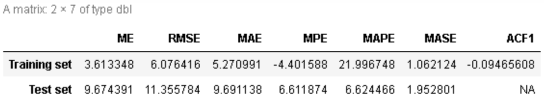
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-0.1779860	4.235164	3.432841	-12.50368134	18.352142	0.6917304	-0.1643294
Test set	0.1659548	3.648893	2.956916	0.03986822	2.093531	0.5958297	NA





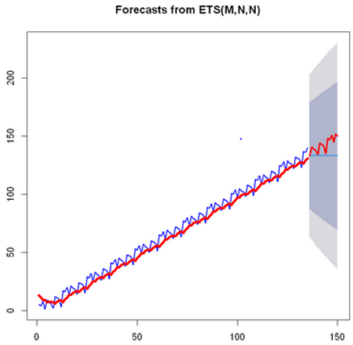
```
ETS(A,A,N)

Call:
 ets(y = tdataset[1:135], model = "AAN")

  Smoothing parameters:
    alpha = 0.0432
    beta  = 1e-04

  Initial states:
    l = 3.1748
    b = 0.9864

  sigma:  4.2993

     AIC     AICc      BIC 
1061.936 1062.401 1076.462
```

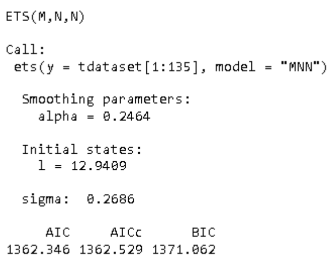
- Using ets with model='MNN'



A matrix: 2 × 7 of type dbl

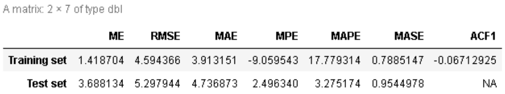
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	3.613348	6.076416	5.270991	-4.401588	21.996748	1.062124	-0.09465608
Test set	9.674391	11.355784	9.691138	6.611874	6.624466	1.952801	NA





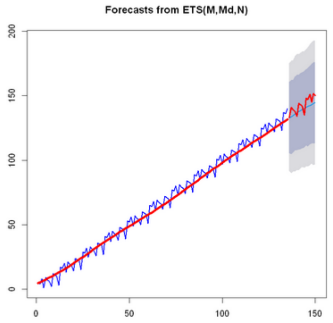
```
ETS(M,N,N)

Call:
 ets(y = tdataset[1:135], model = "MNN")

  Smoothing parameters:
    alpha = 0.2464

  Initial states:
    l = 12.9409

  sigma:  0.2686

     AIC     AICc      BIC 
1362.346 1362.529 1371.062 
```

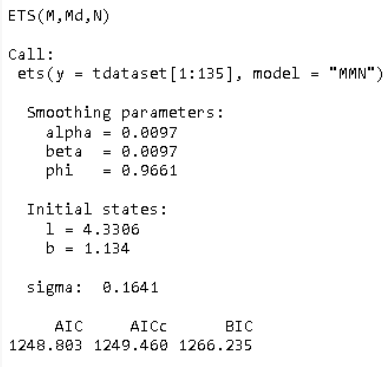
- Using ets with model='MMN'



A matrix: 2 × 7 of type dbl

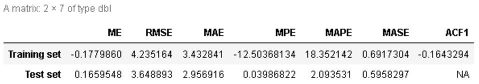
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	1.418704	4.594366	3.913151	-9.059543	17.779314	0.7885147	-0.06712925
Test set	3.688134	5.297944	4.736873	2.496340	3.275174	0.9544978	NA





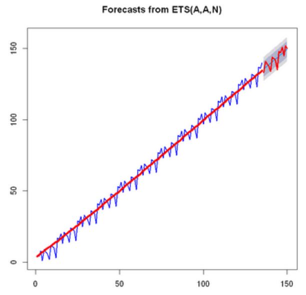
```
ETS(M,Md,N)

Call:
 ets(y = tdataset[1:135], model = "MMN")

  Smoothing parameters:
    alpha = 0.0097
    beta  = 0.0097
    phi   = 0.9661

  Initial states:
    l = 4.3306
    b = 1.134

  sigma:  0.1641

     AIC     AICc      BIC 
1248.803 1249.460 1266.235 
```

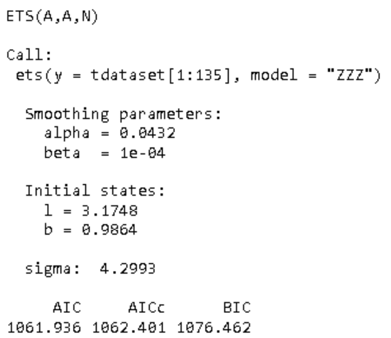
- Using ets with model='ZZZ'



A matrix: 2 × 7 of type dbl

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-0.1779860	4.235164	3.432841	-12.50368134	18.352142	0.6917304	-0.1643294
Test set	0.1659548	3.648893	2.956916	0.03986822	2.093531	0.5958297	NA



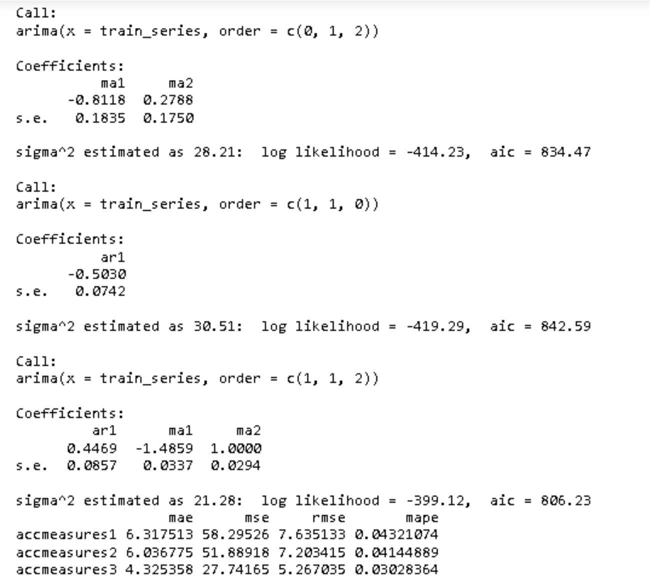


```
ETS(A,A,N)

Call:
 ets(y = tdataset[1:135], model = "ZZZ")

  Smoothing parameters:
    alpha = 0.0432
    beta  = 1e-04

  Initial states:
    l = 3.1748
    b = 0.9864

  sigma:  4.2993

     AIC     AICc      BIC 
1061.936 1062.401 1076.462 
```

b. ARIMA

The ARIMA models are all static. AutoRegressive Integrated Moving Average Model is abbreviated as ARIMA. The stationary nature of the data is the objective of ARIMA.

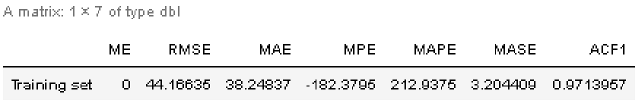
There are two kinds of ARIMA models: non-occasional and occasional. A moving average model and differencing with autoregression are combined in the non-seasonal model.



```
Call:
arima(x = train_series, order = c(0, 1, 2))

Coefficients:
          ma1     ma2
      -0.8118  0.2788
s.e.   0.1835  0.1750

sigma^2 estimated as 28.21:  log likelihood = -414.23,  aic = 834.47

Call:
arima(x = train_series, order = c(1, 1, 0))

Coefficients:
          ar1
      -0.5030
s.e.   0.0742

sigma^2 estimated as 30.51:  log likelihood = -419.29,  aic = 842.59

Call:
arima(x = train_series, order = c(1, 1, 2))

Coefficients:
         ar1      ma1     ma2
      0.4469  -1.4859  1.0000
s.e.  0.0857   0.0337  0.0294

sigma^2 estimated as 21.28:  log likelihood = -399.12,  aic = 806.23
                   mae      mse     rmse       mape
accmeasures1 6.317513 58.29526 7.635133 0.04321074
accmeasures2 6.036775 51.88918 7.203415 0.04144889
accmeasures3 4.325358 27.74165 5.267035 0.03028364
```

3. Forecast the number of departures in the first 6 months of 2021 and discuss the choice of an ‘optimum’ model for this series

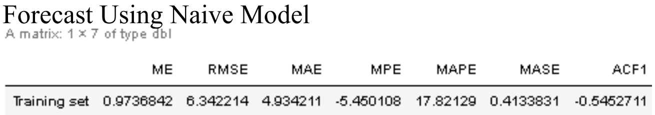
a. Forecast Using the Mean Model



A matrix: 1 × 7 of type dbl

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	0	44.16635	38.24837	-182.3795	212.9375	3.204409	0.9713957

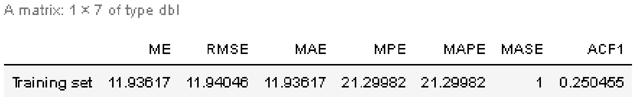
b. Forecast Using Naive Model



A matrix: 1 × 7 of type dbl

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	0.9736842	6.342214	4.934211	-5.450108	17.82129	0.4133831	-0.5452711

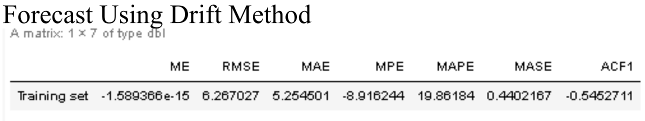
c. Forecast Using Seasonal Naive Model



A matrix: 1 × 7 of type dbl

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	11.93617	11.94046	11.93617	21.29982	21.29982	1	0.250455

d. Forecast Using Drift Method

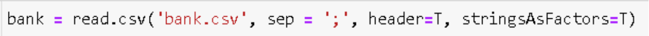


A matrix: 1 × 7 of type dbl

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-1.589366e-15	6.267027	5.254501	-8.916244	19.86184	0.4402167	-0.5452711

- **Bank Marketing Campaign**

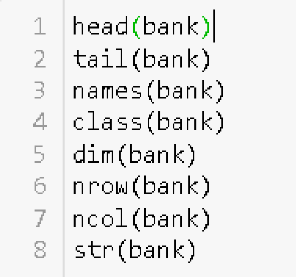
First, this study needs to read the file in Jupyter Notebook file.



```
bank = read.csv('bank.csv', sep = ';', header=T, stringsAsFactors=T)
```

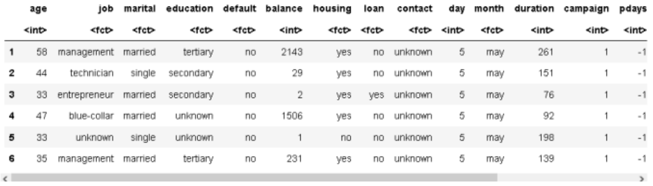
Figure. Reading the bank data file

After the data being read, we can see the structure and characteristics of dataset by using this code.



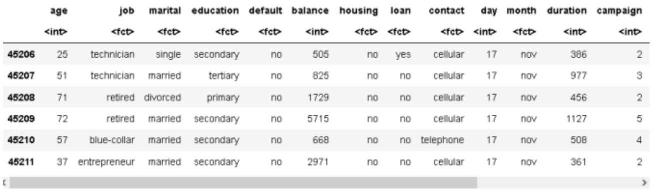
```
head(bank)
tail(bank)
names(bank)
class(bank)
dim(bank)
nrow(bank)
ncol(bank)
str(bank)
```

Figure. Code to see the structure of dataset



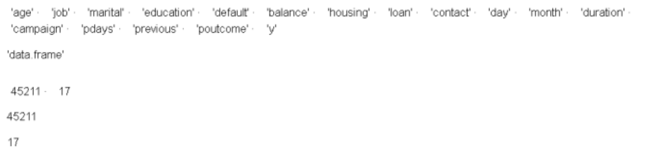
	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays
	<int>	<fct>	<fct>	<fct>	<fct>	<int>	<fct>	<fct>	<fct>	<int>	<fct>	<int>	<int>	<int>
1	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1
2	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1
3	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1
4	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1
5	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1
6	35	management	married	tertiary	no	231	yes	no	unknown	5	may	139	1	-1

Figure. Head of dataset



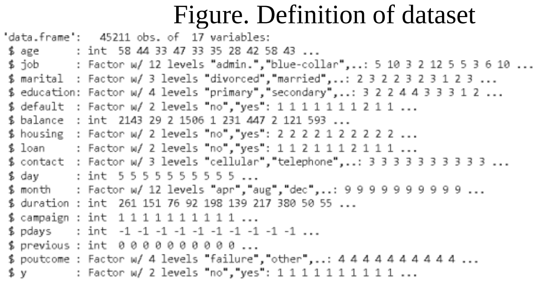
	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign
	<int>	<fct>	<fct>	<fct>	<fct>	<int>	<fct>	<fct>	<fct>	<int>	<fct>	<int>	<int>
45206	25	technician	single	secondary	no	505	no	yes	cellular	17	nov	386	2
45207	51	technician	married	tertiary	no	825	no	no	cellular	17	nov	977	3
45208	71	retired	divorced	primary	no	1729	no	no	cellular	17	nov	456	2
45209	72	retired	married	secondary	no	5715	no	no	cellular	17	nov	1127	5
45210	57	blue-collar	married	secondary	no	668	no	no	telephone	17	nov	508	4
45211	37	entrepreneur	married	secondary	no	2971	no	no	cellular	17	nov	361	2

Figure. Tail of dataset



```
'age' · 'job' · 'marital' · 'education' · 'default' · 'balance' · 'housing' · 'loan' · 'contact' · 'day' · 'month' · 'duration' ·
'campaign' · 'pdays' · 'previous' · 'poutcome' · 'y'

'data.frame'

45211 · 17

45211

17
```

Figure. Definition of dataset

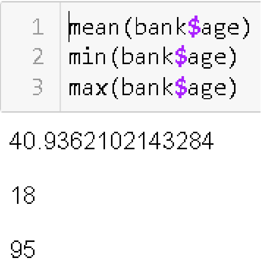


```
'data.frame':	45211 obs. of  17 variables:
 $ age      : int  58 44 33 47 33 35 28 42 58 43 ...
 $ job      : Factor w/ 12 levels "admin.","blue-collar",..: 5 10 3 2 12 5 5 3 6 10 ...
 $ marital  : Factor w/ 3 levels "divorced","married",..: 2 3 2 2 3 2 3 1 2 3 ...
 $ education: Factor w/ 4 levels "primary","secondary",..: 3 2 2 4 4 3 3 3 1 2 ...
 $ default  : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 2 1 1 ...
 $ balance  : int  2143 29 2 1506 1 231 447 2 121 593 ...
 $ housing  : Factor w/ 2 levels "no","yes": 2 2 2 2 1 2 2 2 2 2 ...
 $ loan     : Factor w/ 2 levels "no","yes": 1 1 2 1 1 1 2 1 1 1 ...
 $ contact  : Factor w/ 3 levels "cellular","telephone",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ day      : int  5 5 5 5 5 5 5 5 5 5 ...
 $ month    : Factor w/ 12 levels "apr","aug","dec",..: 9 9 9 9 9 9 9 9 9 9 ...
 $ duration : int  261 151 76 92 198 139 217 380 50 55 ...
 $ campaign : int  1 1 1 1 1 1 1 1 1 1 ...
 $ pdays    : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
 $ previous : int  0 0 0 0 0 0 0 0 0 0 ...
 $ poutcome : Factor w/ 4 levels "failure","other",..: 4 4 4 4 4 4 4 4 4 4 ...
 $ y        : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
```

Figure. Structure of dataset

Exploratory Data Analysis : Data analysis will be used in this section to determine the bank marketing campaign's demographic target segmentation. Among the demographic factors are: age, education, occupation, and marital status

a. *What is the age range of bank marketing target segment? What is the average age?*



```
mean(bank$age)
min(bank$age)
max(bank$age)
```

```
40.9362102143284

18

95
```

Figure. Age range

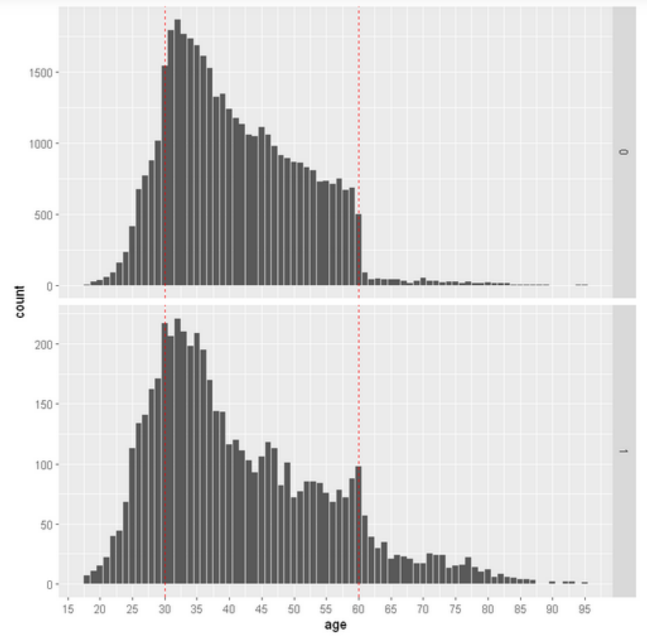
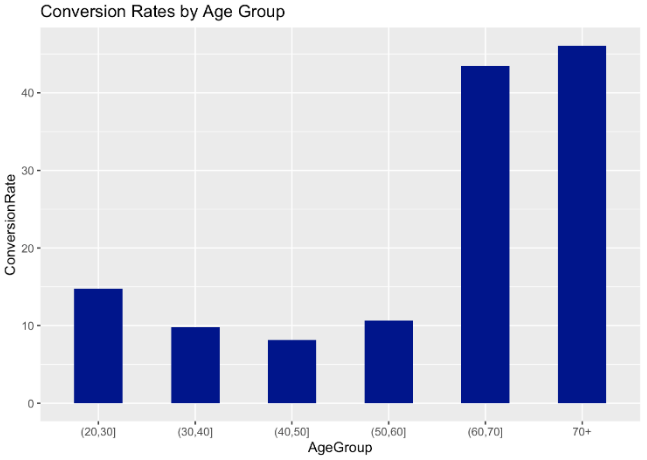


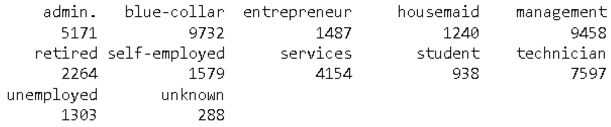
Figure. Chart of age range

It is evident from the graph that the bank is not interested in contacting the over-60 population.



The chart shows that people over 60 have the highest conversion rate for taking out a term deposit; however, the bank has not given this group as much attention or communication. It could be because older people are hard to reach with telemarketing because they aren't very familiar with technology.

b. *What kind of jobs does the customer pool represent?*



```
       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 
```

Figure. Total customer for every job

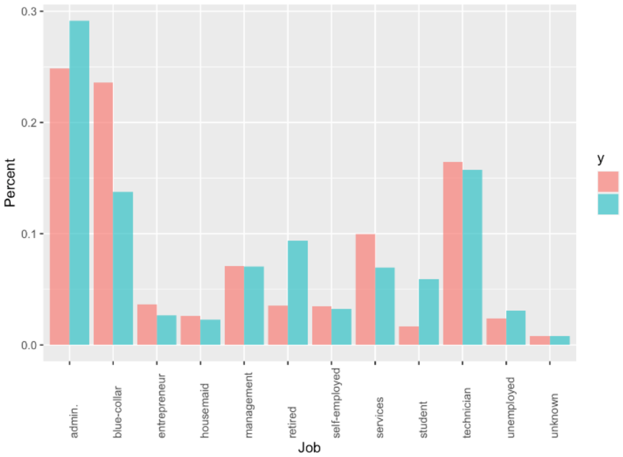


Figure. Chart of every customer’s job

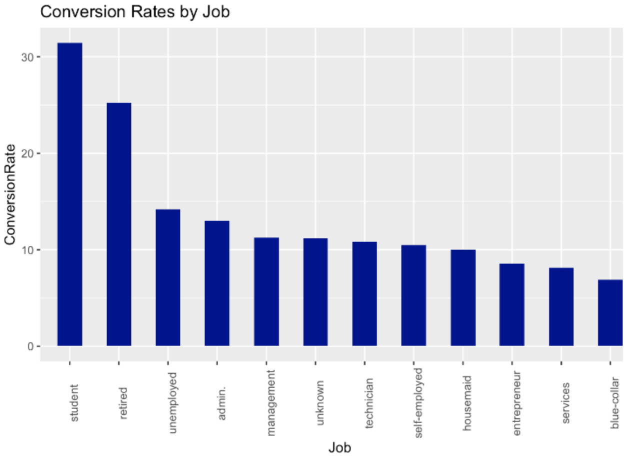
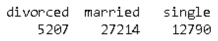


Figure. Conversion rates by Job

Surprisingly, the categories of students, retired people, administrators, and unemployed people have the highest relative frequencies of subscription to term deposits. In the blue-collar sector, there is surprisingly little demand for deposit subscriptions.

c. *How is the marital status of potential clients?*



```
divorced  married   single
    5207    27214    12790
```

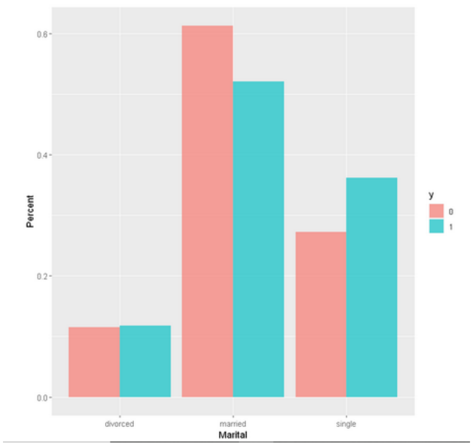
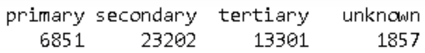


Figure. Chart of Customer based on Marital Status

The chart shows clearly that single clients are more likely than other groups (divorced and married) to subscribe to term deposits.

d. *What are the target customers' levels of education?*



```
primary secondary  tertiary   unknown
   6851     23202     13301      1857
```

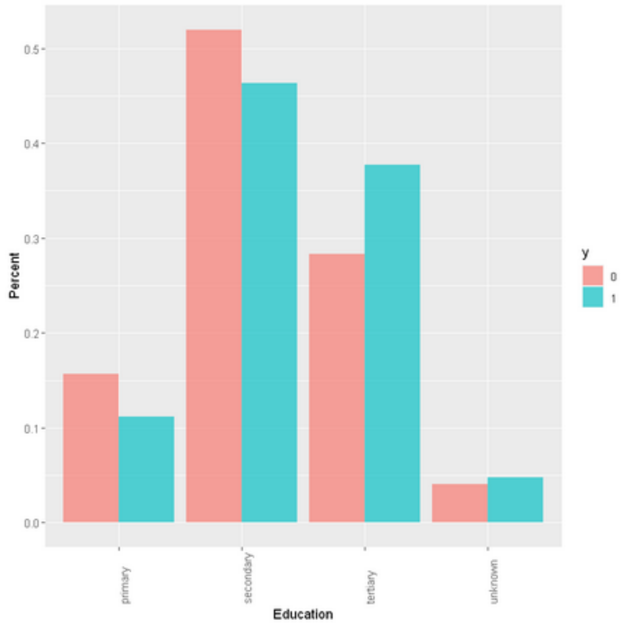


Figure. Chart of customer based on education

Apparently there is a positive connection between's the quantity of long stretches of training and the probability to buy into a term store. Individuals with college degree is the gathering that have the most noteworthy probability of taking up term store. Bank marketing ought to concentrate on high school, professional courses, and university degrees as the three groups.

e. *Which month has the highest term deposit subscription?*

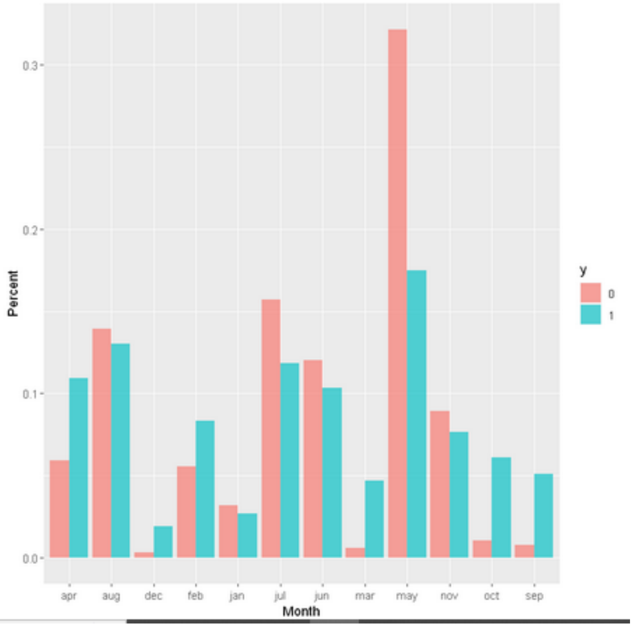


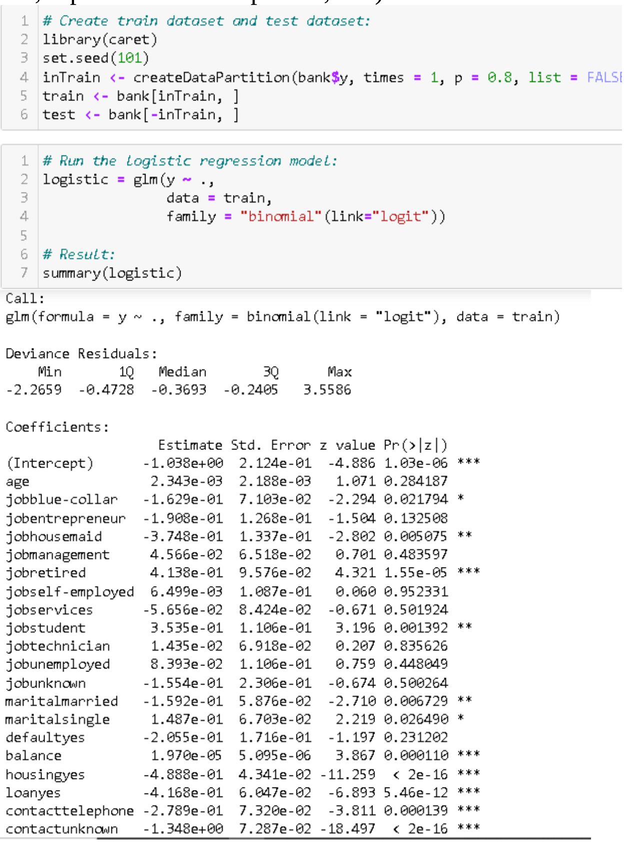
Figure. Percentage of subscription by month

f. Applying logistic regression

An order calculation is what Logistic Regression is. It is utilized to anticipate a double outcome in light of numerous unrestricted factors.

A paired result is one in which there are only two possible outcomes: either the event occurs (option 1) or it does not occur. Free factors are those that have the potential to affect the outcome (also known as the ward variable).

When working with paired data, the most appropriate type of analysis is logistic regression. When the result or ward variable is dichotomous or clear-cut, we know we are managing paired information; overall, if it falls into one of two categories (such as "yes" or "no," "pass" or "come up short, etc.).

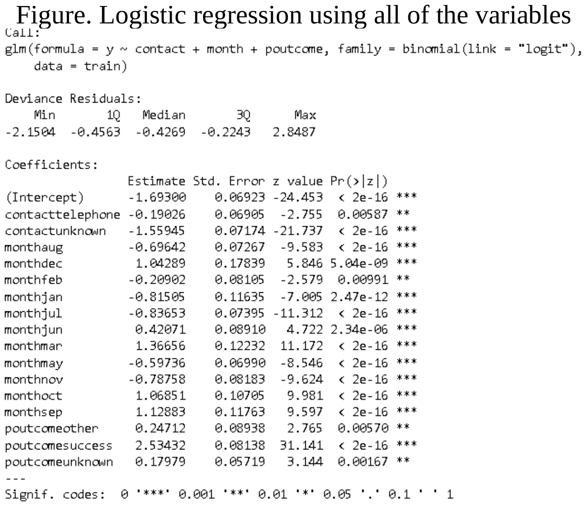


```r
# Create train dataset and test dataset:
library(caret)
set.seed(101)
inTrain <- createDataPartition(bank$y, times = 1, p = 0.8, list = FALSE
train <- bank[inTrain, ]
test <- bank[-inTrain, ]
```

```r
# Run the logistic regression model:
logistic = glm(y ~ .,
               data = train,
               family = "binomial"(link="logit"))

# Result:
summary(logistic)
```

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2659  -0.4728  -0.3693  -0.2405   3.5586

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.038e+00  2.124e-01  -4.886 1.03e-06 ***
age               2.343e-03  2.188e-03   1.071 0.284187
jobblue-collar   -1.629e-01  7.103e-02  -2.294 0.021794 *
jobentrepreneur  -1.908e-01  1.268e-01  -1.504 0.132508
jobhousemaid     -3.748e-01  1.337e-01  -2.802 0.005075 **
jobmanagement     4.566e-02  6.518e-02   0.701 0.483597
jobretired        4.138e-01  9.576e-02   4.321 1.55e-05 ***
jobself-employed  6.499e-03  1.087e-01   0.060 0.952331
jobservices      -5.656e-02  8.424e-02  -0.671 0.501924
jobstudent        3.535e-01  1.106e-01   3.196 0.001392 **
jobtechnician     1.435e-02  6.918e-02   0.207 0.835626
jobunemployed     8.393e-02  1.106e-01   0.759 0.448049
jobunknown       -1.554e-01  2.306e-01  -0.674 0.500264
maritalmarried   -1.592e-01  5.876e-02  -2.710 0.006729 **
maritalsingle     1.487e-01  6.703e-02   2.219 0.026490 *
defaultyes       -2.055e-01  1.716e-01  -1.197 0.231202
balance           1.970e-05  5.095e-06   3.867 0.000110 ***
housingyes       -4.888e-01  4.341e-02 -11.259  < 2e-16 ***
loanyes          -4.168e-01  6.047e-02  -6.893 5.46e-12 ***
contacttelephone -2.789e-01  7.320e-02  -3.811 0.000139 ***
contactunknown   -1.348e+00  7.287e-02 -18.497  < 2e-16 ***
```

Figure. Logistic regression using all of the variables



```
Call:
glm(formula = y ~ contact + month + poutcome, family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1504  -0.4563  -0.4269  -0.2243   2.8487

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.69300    0.06923 -24.453  < 2e-16 ***
contacttelephone -0.19026    0.06905  -2.755  0.00587 **
contactunknown   -1.55945    0.07174 -21.737  < 2e-16 ***
monthaug         -0.69642    0.07267  -9.583  < 2e-16 ***
monthdec          1.04289    0.17839   5.846 5.04e-09 ***
monthfeb         -0.20902    0.08105  -2.579  0.00991 **
monthjan         -0.81505    0.11635  -7.005 2.47e-12 ***
monthjul         -0.83653    0.07395 -11.312  < 2e-16 ***
monthjun          0.42071    0.08910   4.722 2.34e-06 ***
monthmar          1.36656    0.12232  11.172  < 2e-16 ***
monthmay         -0.59736    0.06990  -8.546  < 2e-16 ***
monthnov         -0.78758    0.08183  -9.624  < 2e-16 ***
monthoct          1.06851    0.10705   9.981  < 2e-16 ***
monthsep          1.12883    0.11763   9.597  < 2e-16 ***
poutcomeother     0.24712    0.08938   2.765  0.00570 **
poutcomesuccess   2.53432    0.08138  31.141  < 2e-16 ***
poutcomeunknown   0.17979    0.05719   3.144  0.00167 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure. Logistic regression using only significant variables

IV. RESULTS AND CONCLUSION

- **Airplane Departure**

  The Mean Absolute Error (MAE) is calculated by using the mean of the absolute differences between the actual and predicted values (y_hat).

  Outliers, or large errors, can be more effectively identified with the Mean Squared Error (MSE). As the name suggests, it employs the mean of the squared errors (differences between y and y_hat). Due to its squaring, it favors large errors over small ones, which can be a disadvantage in some situations. Consequently, the MSE is appropriate when the metric loses its unit when it squares and we really want to focus on significant errors.

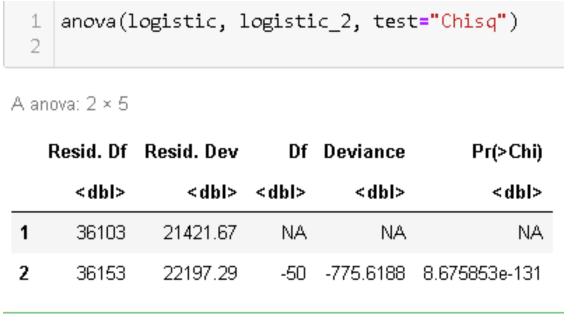
  The MSE won't lose its unit if we use its square root. The Root Mean Squared Error (RMSE) is the new mistake metric that outcomes from this. It has the same advantages as its siblings, MAE and MSE. Be that as it may, it is additionally delicate to anomalies, very much like MSE.

  Based on those descriptions, this study will focus on finding the lowest RMSE to get the best result. Answering question for every Time Series Model:

  - about Exponential Smoothing, the best model is Using ets with model='AAN'
  - about ARIMA, the best model is ARIMA (1, 1, 2)
  - about forecasting number of departures, the best model is forecast using drift method

- **Bank Marketing Campaign**

The statistical method known as analysis of variance (ANOVA) is used to determine whether two or more groups' means are significantly different from one another. By comparing the means of various samples, ANOVA determines whether one or more factors have an effect.



```r
anova(logistic, logistic_2, test="Chisq")
```

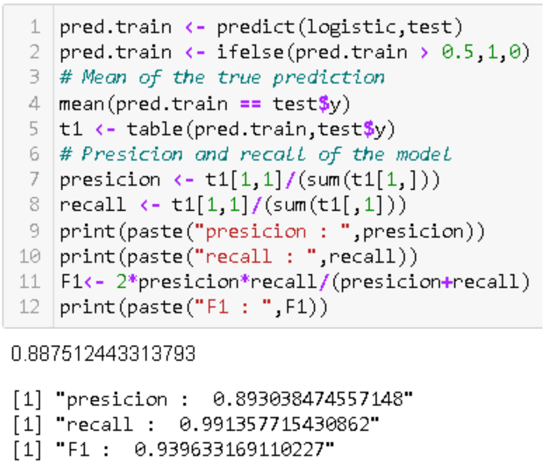
A anova: 2 × 5

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	36103	21421.67	NA	NA	NA
2	36153	22197.29	-50	-775.6188	8.675853e-131

Figure. ANOVA to the models

This ANOVA test also supports the claim that our first model predicts more accurately than our second model with p<0.05.

The percentage of positive predictions that are accurate (true positives) is known as precision. The F1-Score is a measure that combines precision and recall, and recall is a measure of how many of the positive cases the classifier correctly predicted.



```
pred.train <- predict(logistic,test)
pred.train <- ifelse(pred.train > 0.5,1,0)
# Mean of the true prediction
mean(pred.train == test$y)
t1 <- table(pred.train,test$y)
# Presicion and recall of the model
presicion <- t1[1,1]/(sum(t1[1,]))
recall <- t1[1,1]/(sum(t1[,1]))
print(paste("presicion : ",presicion))
print(paste("recall : ",recall))
F1<- 2*presicion*recall/(presicion+recall)
print(paste("F1 : ",F1))
```

```
0.887512443313793

[1] "presicion :  0.893038474557148"
[1] "recall :  0.991357715430862"
[1] "F1 :  0.939633169110227"
```

Figure. Precision, Recall and F1

V. REFERENCES

1. Best, H., & Wolf, C. (2014). The SAGE handbook of regression analysis and causal inference (H. Best & C. Wolf, Eds.). SAGE Publications.
2. Chatfield, C. (2016). The analysis of time series: An introduction, sixth edition (6th ed.). Chapman and Hall.
3. Hilbe, J. M. (2018). Practical guide to logistic regression. CRC Press.
4. Little, T. D. (Ed.). (2013). The oxford handbook of quantitative methods in psychology: Vol. 2: Statistical analysis. Oxford University Press.