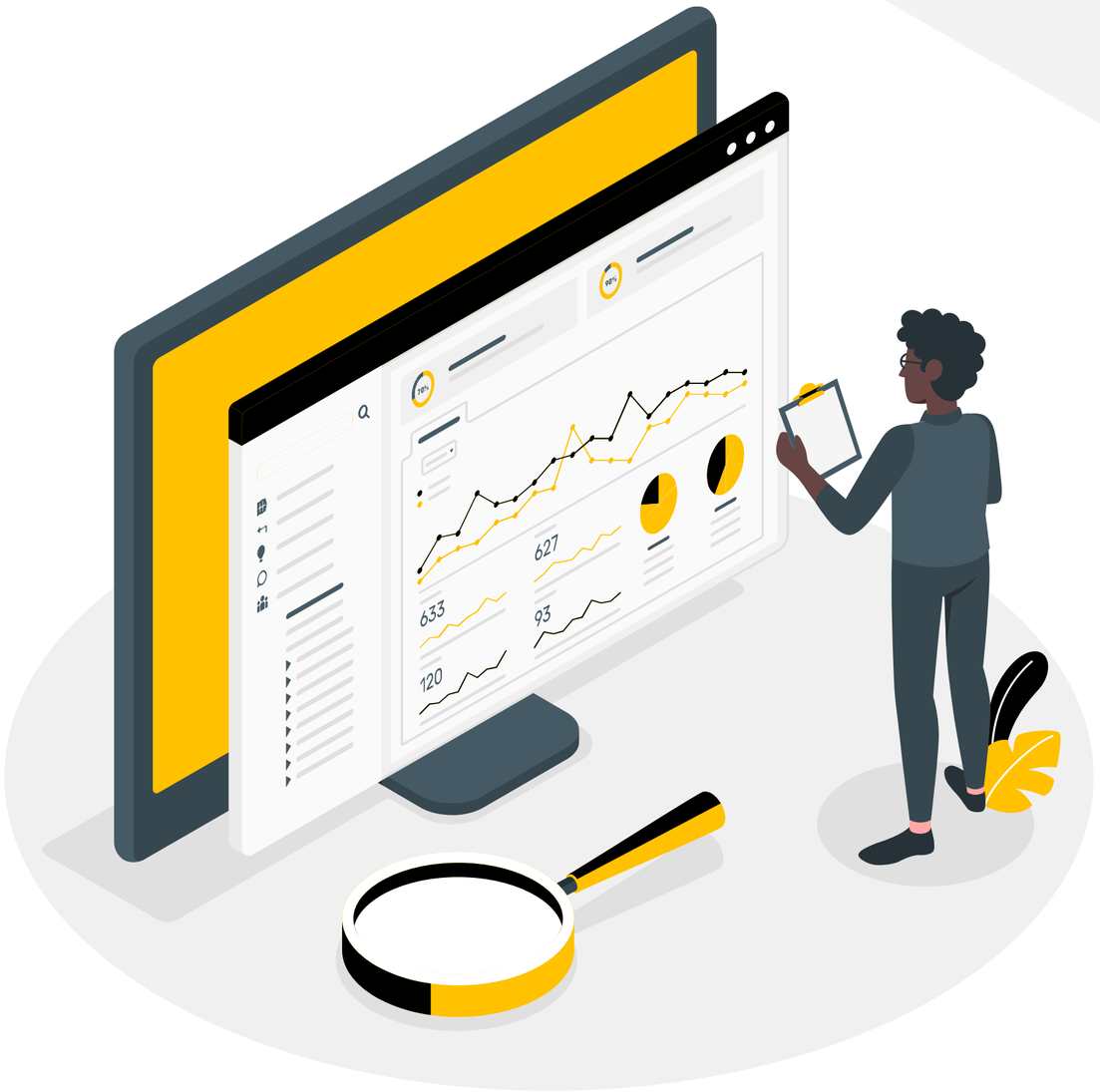
Analyzing Marketing Campaign Data



Prepared by:

Farida Mohamed	202000860
Hagar Alaaeldien	202000691
Sama Yousef	202000819
Mohamed Zayed	201800760
Nada Tayel	202001944

Table of Contents

Overview

01

Overview

- When was the last time you received a call marketing for their product?
- Was it relevant?
- How many marketing campaigns can you recall?

Don the marketer’s attire

- Revisiting the last campaign data
- Analysis related to new financial products.
- By Modeling the relationships between features and outcome
- By Applying Linear and Logistic Regression models

“Marketing is no longer about the stuff you make, but about the stories you tell.“

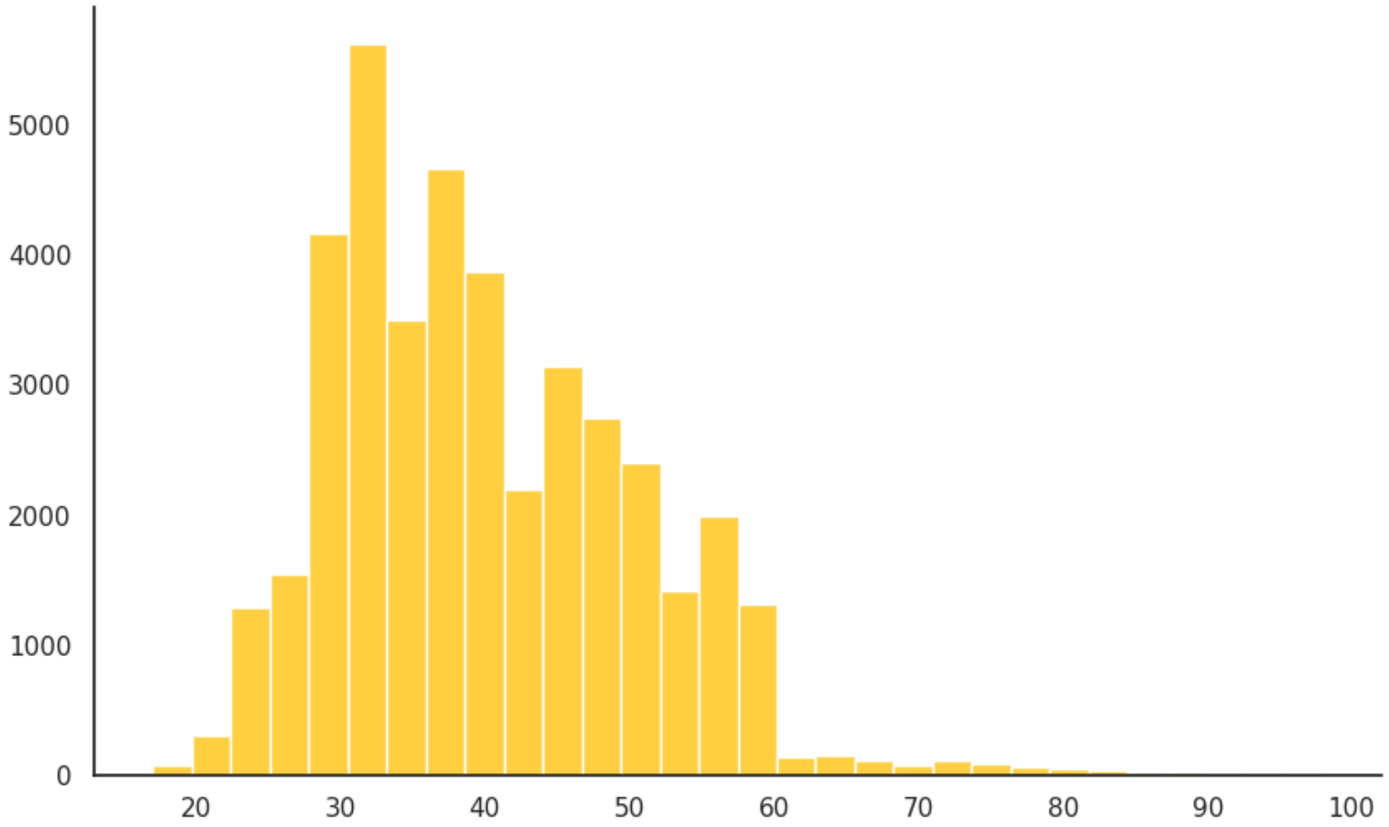
- Seth Godin

Exploring the data

Source: marketing campaign of a Portuguese banking institution based on phone calls performed between May 2008 - November 2010

Age	Age of the contacted client
Job	Type of job
Marital	Marital status (note: "divorced" means divorced or widowed)
Education	Education level of the contacted client
Default	Does the client have credit in default?
Housing	Does the client have a housing loan?
Loan	Does the client have a personal loan?
Contact	Type of communication with the client
Month	Last contact, month of the year
day_of_week	Last contact, day of the week
Duration	Last contact duration (in seconds)
Campaign	Number of contacts performed during this campaign for this client
Pdays	Number of days passed by after the client was contacted from a previous campaign
Previous	Number of contacts performed before this campaign and for this client
Poutcome	Outcome of the previous marketing campaign
emp.var.rate	Employment variation rate (quarterly indicator)
cons.price.idx	Consumer price index (monthly indicator)
cons.conf.idx	Consumer confidence index (monthly indicator)
euribor3m	Euribor 3-month rate
nr.employed	Number of employees (quarterly indicator)
Y	Has the client subscribed to a term deposit (outcome of the marketing campaign)?

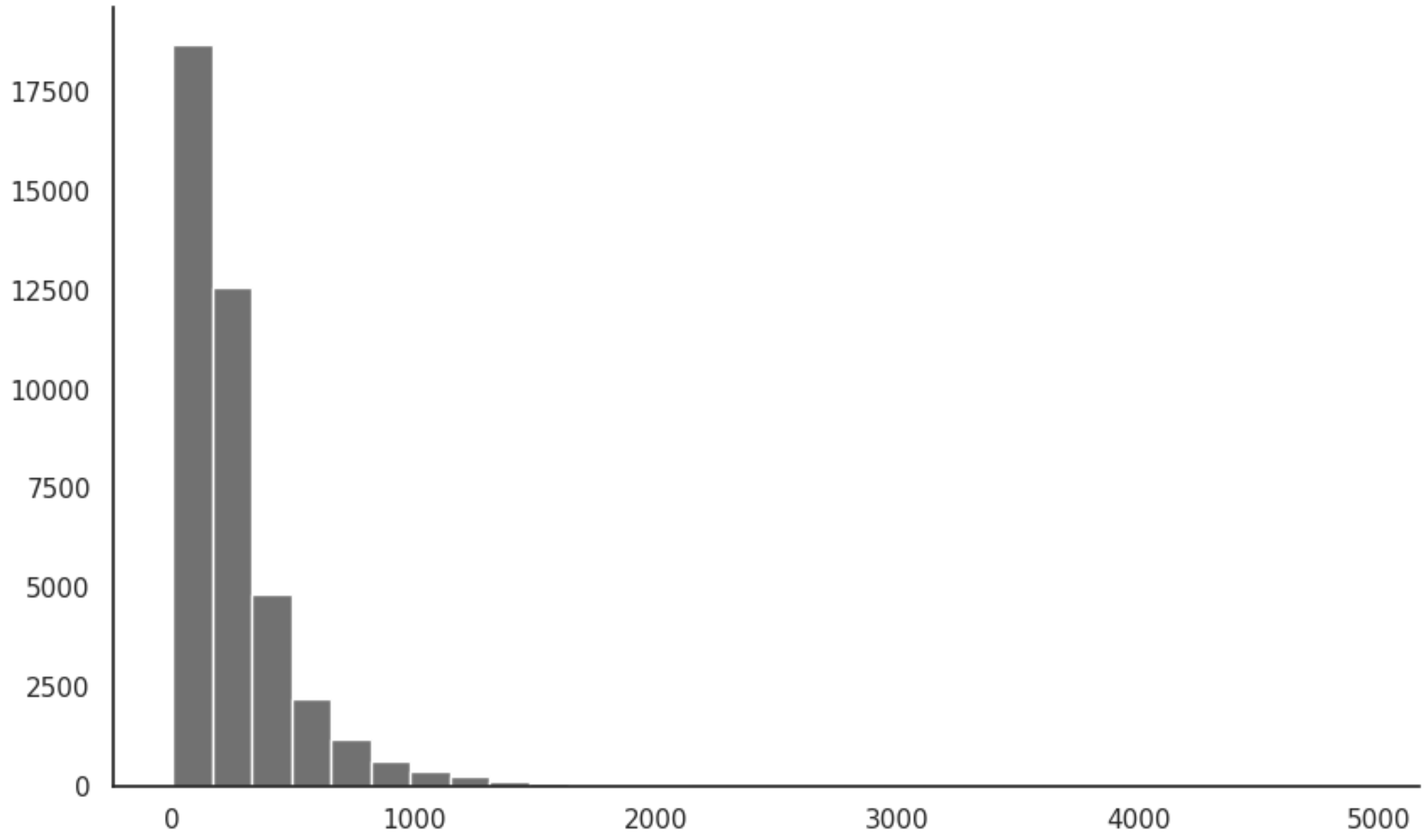
Ages Of Contacted Clients:



- **41,188** contacted Client.
- Ages Range between **17** and **98**.

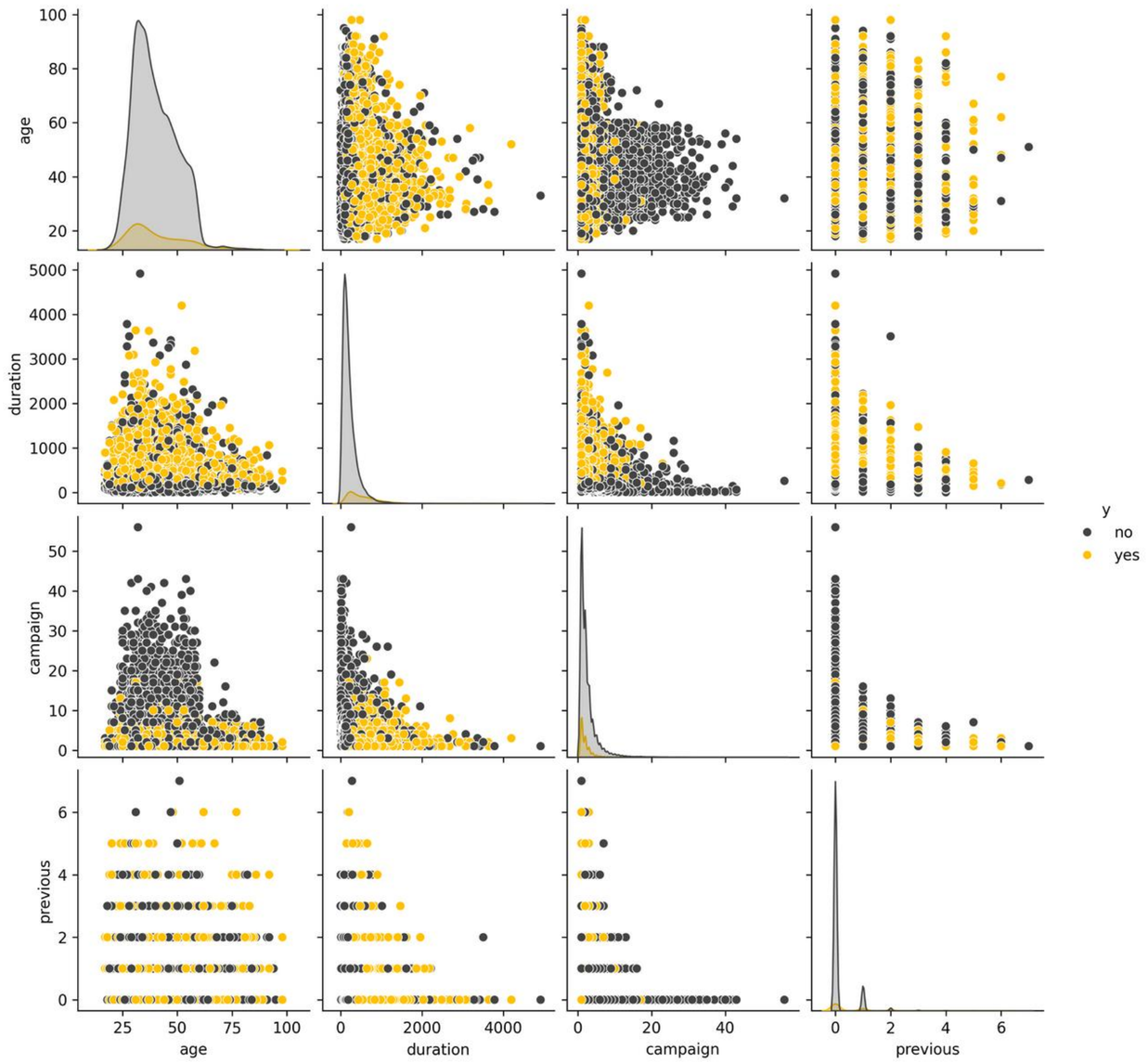
- 1st Quartile: **32** years
- 2nd Quartile: **38** years
- 3rd Quartile: **47** years

Duration of Last Contact:



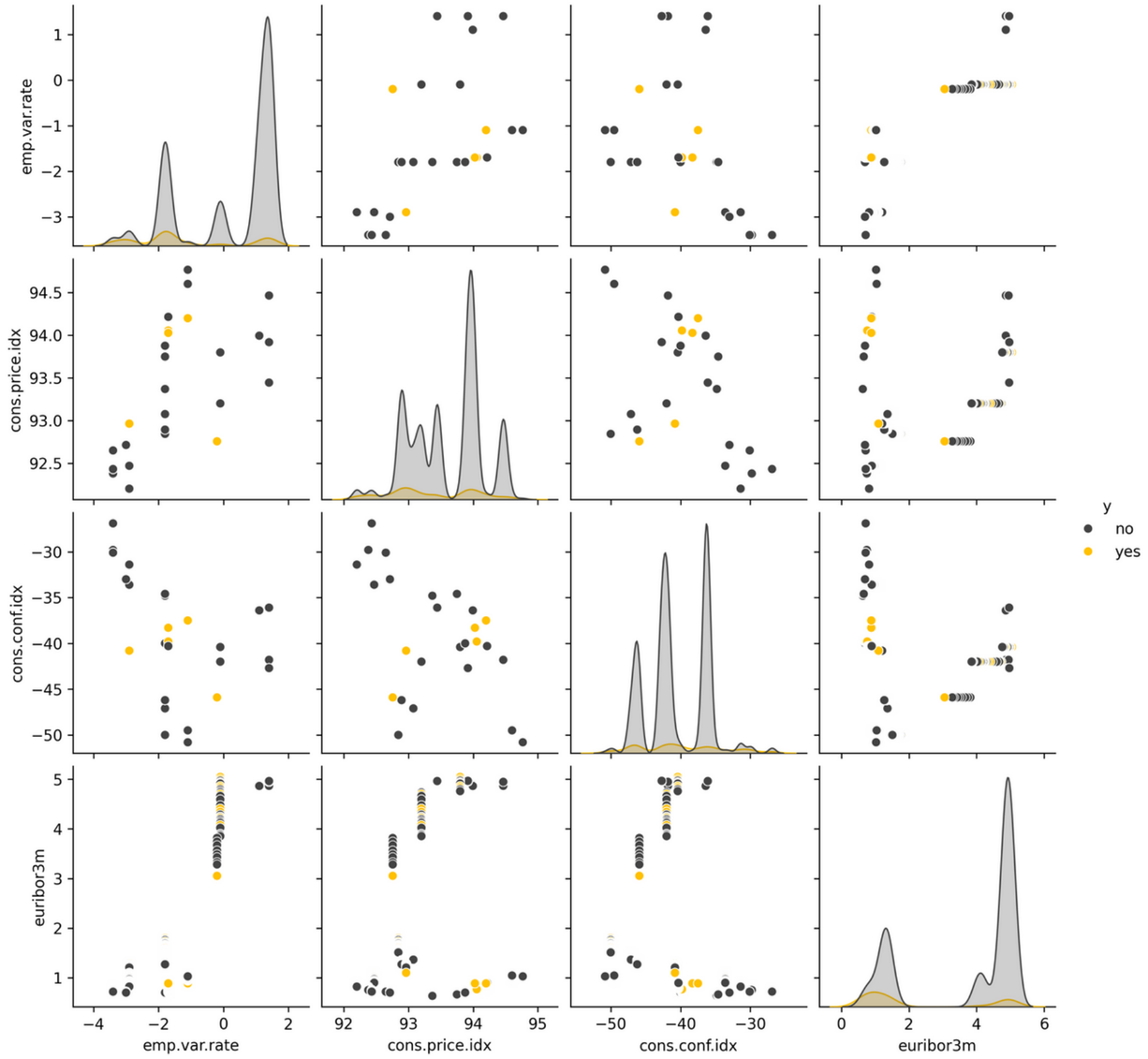
- Durations Range between **0** and **4918** Secs.

- 1st Quartile: **102** Secs
- 2nd Quartile: **180** Secs
- 3rd Quartile: **319** Secs



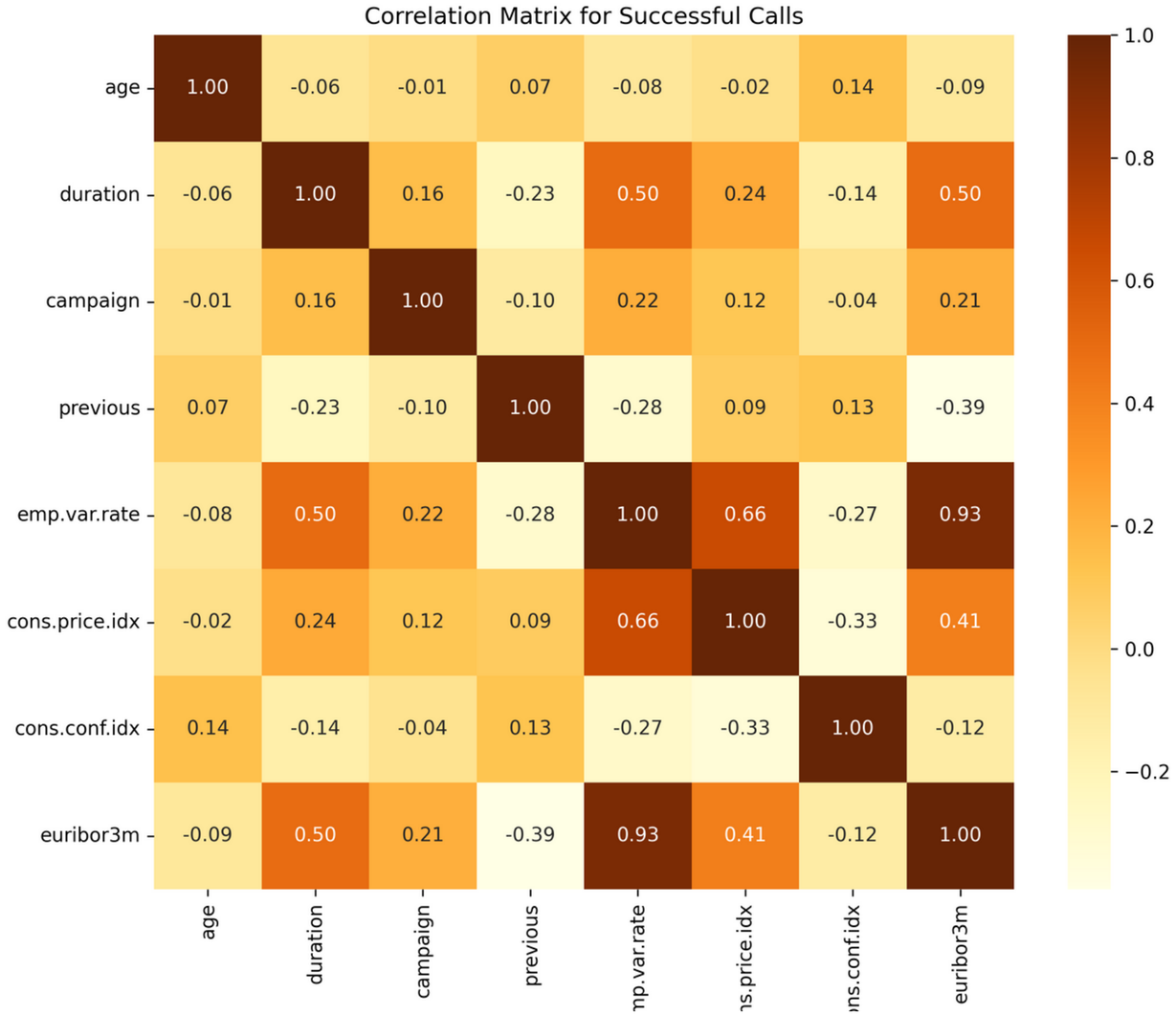
Pairplot of the campaign features, grouped by marketing outcome

- Fewer contacts result in more successful campaign.
- Reducing the number of contacts per customer might lead to higher success rates.
- Prioritizing quicker follow-ups after initial contact could improve outcomes.



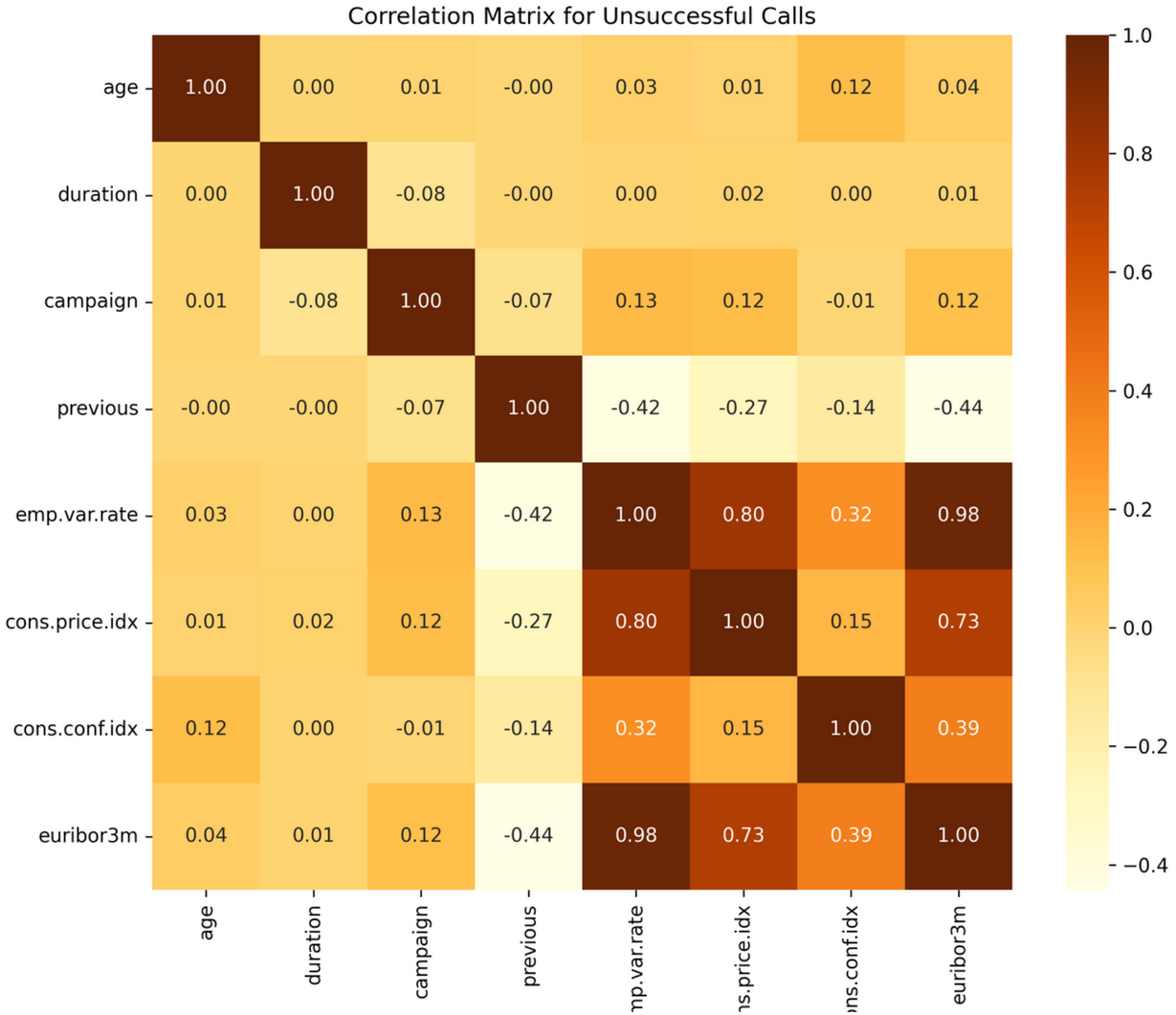
Pairplot of the financial features, grouped by marketing outcome

- Lower euro interest rates (euribor3m) generally correlate with higher success rates in campaigns.
- Campaigns conducted during periods of favorable economic conditions may be more effective.



Correlation Matrix for Successful Calls

- **Emp.Var.Rate and Euribor3m (0.93 Correlation):** indicates that an expanding economy results in more successful calls.
- **Emp.Var.Rate and Cons.Price.Idx (0.66 Correlation):** reinforces this hypothesis.



Correlation Matrix for Unsuccessful Calls

- Higher correlations among unsuccessful calls with factors like higher number of contacts and longer days since the last campaign.
- Shows inverse patterns compared to successful calls, particularly in the rate of contact and age distribution.

Modeling Data

02

Modeling via Regression

- **Regression:** A fundamental technique in data analysis and machine learning.
- **Estimating Relationships:** The process of estimating the relationship between 'm' different features (X1, ..., Xm) and a target variable (Y).
- **General Form**: Linear/Logistic regression equation:

$$Y \approx f(X_1, \ldots, X_m)$$

Figure 3.16: General form of linear/logistic regression

Linear Regression

- **Linear Regression:** The target variable, Y, is continuous.
- **General Form:** The linear regression equation.

$$Y \approx \alpha_0 + \sum_{j=1}^{m} \alpha_j X_j$$

Figure 3.17: Linear regression equation

- **Specific Form:** Applying the equation to 'i' samples with 'm' features each.

$$y_i \approx \hat{y}_i = \alpha_0 + \sum_{j=1}^{m} \alpha_j X_{i,j}$$

Figure 3.19: Linear regression equation in a specific form

Assumptions Least Squares Method

- **Assumptions:** The dependency of Y from the feature vectors {X1,..., Xm} is either linear or can be approximated with a linear equation.
- **Estimating Relationships:** Finding the parameters $\alpha 0, ..., \alpha m$ so that the "distance" between Y and $\hat{Y}$ is minimized.
- **Least Squares Method:** A common method used to minimize the distance between the two vectors Y and $\hat{Y}$.

Matrix Notations

- **Matrix Notations:** Introduction of the matrix notations of the target variable, Y, the feature matrix, X, and the parameter vector, α

$$Y = \begin{bmatrix} y_1 \\ . \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 x_{11} & \dots & x_{1m} \\ \dots & & \\ 1 x_{n1} & \dots & x_{nm} \end{bmatrix} \quad \alpha = \begin{bmatrix} \alpha_1 \\ . \\ \alpha_m \end{bmatrix}$$

Figure 3.21: Definitions of the matrix notation of Y, X, and α

- **Residual Sum of Squares (RSS):** Expression for RSS(α) using matrix notation.

$$RSS(\alpha) = (Y - X_\alpha)^T (Y - X_\alpha)$$

Figure 3.22: Expression for RSS(α)

Minimization Conditions

- **Conditions for Minimization:** The gradient of RSS is equal to zero and its Hessian matrix is positive definite.

$$\frac{\partial RSS}{\partial \alpha}(\alpha) = -2X^T(y - X\alpha) = 0$$

Figure 3.23: Condition of the gradient of RSS to be equal to zero

$$\frac{\partial^2 RSS}{\partial \alpha^2}(\alpha) = 2X^T X \text{ positive definite}$$

Figure 3.24: Condition of the hessian matrix of RSS to be positive definite

- **Analytical Solution:** The value of α, which minimizes the function on the right-hand side of the equation.

$$\hat{\alpha} = (X^T X)^{-1} X^T y$$

Figure 3.25: Analytical solution of equation in Figure 3.20

R-squared in Regression

- **R-squared:** Also known as the coefficient of determination, Defined as the proportion of variance in the dependent variable (`cons.conf.idx`: Consumer Confidence Index, in experiment data) Confidence Index, in experiment data) that is predicted by the model from Figure 3.26.

OLS Regression Results

Dep. Variable:	cons.conf.idx	R-squared:	0.177
Model:	OLS	Adj. R-squared:	0.177
Method:	Least Squares	F-statistic:	2960.
Date:	Mon, 10 Feb 2020	Prob (F-statistic):	0.00
Time:	23:28:51	Log-Likelihood:	-1.1753e+05
No. Observations:	41188	AIC:	2.351e+05
Df Residuals:	41184	BIC:	2.351e+05
Df Model:	3		
Covariance Type:	nonrobust		

- **Computation:** Can be computed using the formula shown in Figure 3.27.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Figure 3.27: Definition of R^2

Adjusted R-squared

- **Limitations of R-squared:** It tends to increase with the complexity of the model, which doesn't always mean that our model is becoming more accurate.

- **Adjusted R-squared:** A modification of R-squared that adjusts for the number of predictors in the model.

$$\overline{R^2} = 1 - (1 - R^2)\frac{n-1}{n-m-1}$$

Figure 3.30: Definition of $\overline{R^2}$

- **Benefits of adjusted R-squared:** Takes into account both the accuracy and the complexity of the model. It penalizes excessive complexity.

Interpreting Coefficients & P-values

- **Coefficients:** The 'coef' column provides the coefficients of the linear regression formula. These values tell us how much the dependent variable (`cons.conf.idx`) will change if we increase one of the variables by 1 while keeping the other constant.

- **Correlations:** `cons.conf.idx` is positively correlated with `cons.price.idx` and `euribor3m`, and negatively correlated with `emp.var.rate`.

- **P-values:** The P>|t| column returns the p-value of a hypothesis test, where the null hypothesis is that the relative coefficient is equal to zero. A p-value of 0 means that the coefficient is statistically significant.

	coef	std err	t	P>|t|	[0.025	0.975]
const	-82.4025	5.999	-13.736	0.000	-94.161	-70.644
emp.var.rate	-4.1814	0.072	-57.960	0.000	-4.323	-4.040
cons.price.idx	0.2828	0.063	4.478	0.000	0.159	0.407
euribor3m	4.3582	0.057	76.618	0.000	4.247	4.470

Omnibus:	3246.559	Durbin-Watson:	0.001
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4034.493
Skew:	0.761	Prob(JB):	0.00
Kurtosis:	2.811	Cond. No.	2.72e+04

Figure 3.26: Results from the OLS model

Linear Regression Model

- **Linear regression formula:** The model assumes the following form:

$$cons.conf\ .idx = -82.4025 - 4.1814\ .emp.var.rate + 0.2828\ .cons.price.idx + 4.3582.euribor3m$$

Figure 3.31: Linear regression model of cons.conf.idx as a function of emp.var.rate, cons. price.idx, and euribor3m

- Where cons.conf.idx, emp.var.rate, cons.price.idx, and euribor3m refer to the Consumer Confidence Index, Employment Variation Rate, Consumer Price index, and Euribor 3-month rate respectively.

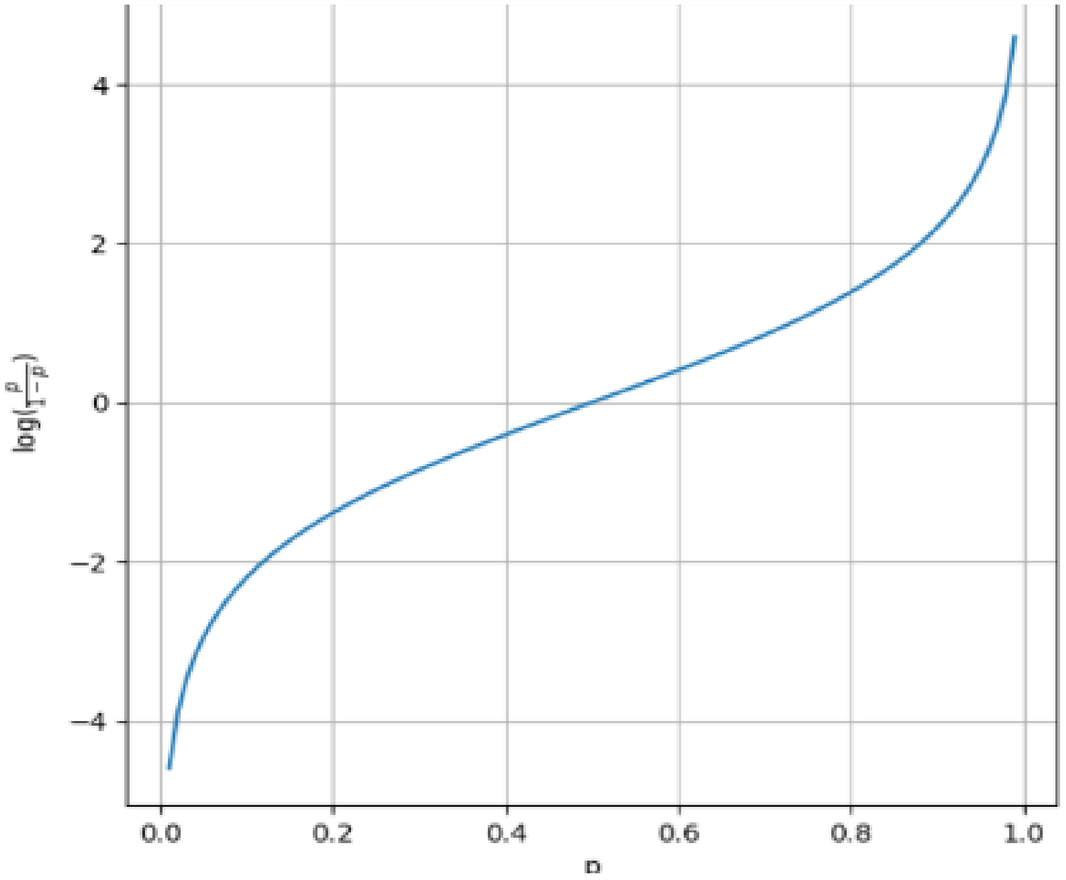
Logistic Regression

	Logistic Regression	Linear regression
Predicts	the probability of categorical outcomes	continuous numerical outcomes
Target Variable	binary {0,1}	continous [0, 1]
Deal with	Numeric data	Numeric data

Modeling Probability with Logistic Regression :

- Sigmoid function transforms the output of the linear equation into probability

- Logit (p)=log(p/(1-p)) ,in which it maps the interval (0,1) into (-∞,∞)



Applying on Our data

Dep. Variable:	y	No. Observations:	41188
Model:	Logit	Df Residuals:	41183
Method:	MLE	Df Model:	4
Date:	Thu, 09 May 2024	Pseudo R-squ.:	0.2331
Time:	19:58:44	Log-Likelihood:	-11119.
converged:	True	LL-Null:	-14499.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.7793	0.076	-49.435	0.000	-3.929	-3.629
age	0.0091	0.002	5.665	0.000	0.006	0.012
duration	0.0039	6.16e-05	62.973	0.000	0.004	0.004
campaign	-0.1163	0.011	-10.706	0.000	-0.138	-0.095
previous	1.0579	0.027	39.232	0.000	1.005	1.111

Probability Estimation in Logistic Regression

Probability estimation via logistic regression (Y = 1)

$$Pr(Y=1|X=x) = \frac{exp(\alpha_0 + \sum_{j=1}^{m} \alpha_j X_j)}{1 + exp(\alpha_0 + \sum_{j=1}^{m} \alpha_j X_j)}$$

Probability estimation via logistic regression (Y = 0)

$$Pr(Y=0|X=x) = \frac{1}{1 + exp(\alpha_0 + \sum_{j=1}^{m} \alpha_j X_j)}$$

Logistic regression model

Logistic regression formula: The model takes the following form:

$$Pr(\text{y= yes}) = \frac{exp(-3.7793 + 0.0091.age + 0.0039.duration - 0.1163.campaign + 1.0579.previous)}{1 + exp(-3.7793 + 0.0091.age + 0.0039.duration - 0.1163.campaign + 1.0579.previous)}$$

Figure 3.39: Probability of the y column being equal to "yes," according to the logistic regression model

$$Pr(\text{y=yes}) = \frac{1}{1 + exp(-3.7793 + 0.0091.age + 0.0039.duration - 0.1163.campaign + 1.0579.previous)}$$

Figure 3.40: Probability of the y column being equal to "no," according to the logistic regression model

where previous stands for the number of contacts before this campaign and for specific client

Logistic regression model

- High p-values (>1) indicate insignificance of features
- infinity p-values signify convergence issues

Logit Regression Results

Dep. Variable:	y	No. Observations:	41188
Model:	Logit	Df Residuals:	41183
Method:	MLE	Df Model:	4
Date:	Tue, 11 Feb 2020	Pseudo R-squ.:	0.2331
Time:	17:19:35	Log-Likelihood:	-11119.
converged:	True	LL-Null:	-14499.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.7793	0.076	-49.435	0.000	-3.929	-3.629
age	0.0091	0.002	5.665	0.000	0.006	0.012
duration	0.0039	6.16e-05	62.973	0.000	0.004	0.004
campaign	-0.1163	0.011	-10.706	0.000	-0.138	-0.095
previous	1.0579	0.027	39.232	0.000	1.005	1.111

Leaner Logistic Regression

Leaner Logistic Regression Model

- Leaner logistic regression model was created by selecting the important features improving the results of the model

```
Optimization terminated successfully.
         Current function value: 0.222140
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                41188
Model:                          Logit   Df Residuals:                    41181
Method:                           MLE   Df Model:                            6
Date:                Wed, 12 Feb 2020   Pseudo R-squ.:                  0.3690
Time:                        19:53:11   Log-Likelihood:                -9149.5
converged:                       True   LL-Null:                       -14499.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const        -43.1379      3.524    -12.240      0.000     -50.046     -36.230
duration       0.0045   7.11e-05     63.505      0.000       0.004       0.005
campaign      -0.0495      0.011     -4.331      0.000      -0.072      -0.027
pdays         -0.0016   6.81e-05    -22.928      0.000      -0.002      -0.001
cons.price.idx 0.4921      0.038     12.877      0.000       0.417       0.567
cons.conf.idx  0.0699      0.003     20.137      0.000       0.063       0.077
euribor3m     -0.7200      0.015    -47.462      0.000      -0.750      -0.690
==============================================================================
```

Figure 3.45: Results from the logistic regression model

Conclusion

- The mathematical analysis of marketing campaign data has revealed key insights that can significantly impact our marketing strategies.

- optimizing future campaigns, such as allocating resources to high-performing channels, refining targeting strategies, and tailoring messages to specific customer segments.

Thanks