**Bank Client Analysis for Marketing Targets**

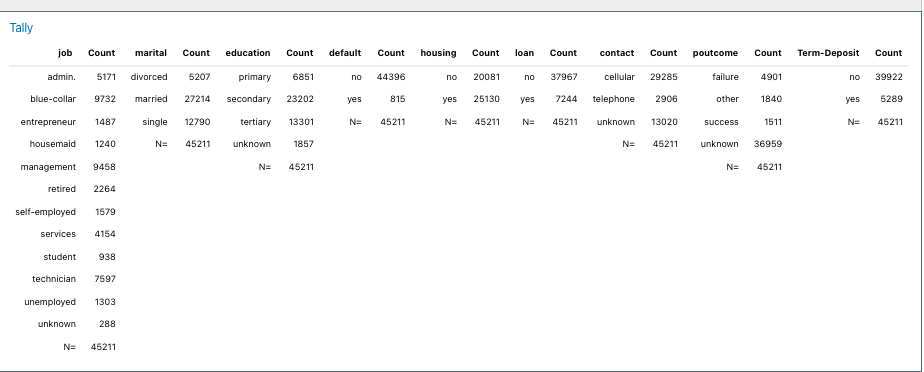
**Introduction**

In this analysis, we have taken a Bank client data of a Bank based out of Portugal. The data set is comprising of 45,211 clients and key variable includes age, job, marital status, education, balance, housing loan, personal loan, phone call details, previous calls from last campaign.

Based on the analysis, we aim to prioritize clients based on the different factors to invest in term deposits, Create personalized marketing campaigns to engage with the identified target clients effectively.

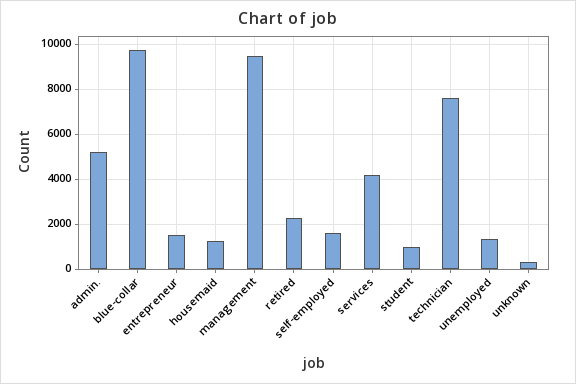
**Statistics**

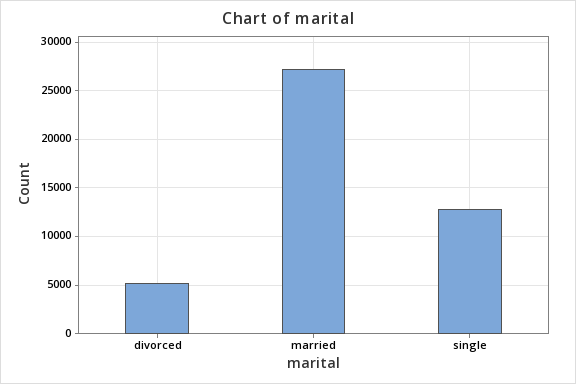
| **Variable** | **Mean** | **SE Mean** | **StDev** | **Minimum** | **Q1** | **Median** | **Q3** | **Maximum** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| age | 40.9362 | 0.0499404 | 10.6188 | 18 | 33 | 39 | 48 | 95 |
| balance | 1362.27 | 14.3196 | 3044.77 | -8019 | 72 | 448 | 1428 | 102127 |
| duration | 258.163 | 1.21116 | 257.528 | 0 | 103 | 180 | 319 | 4918 |
| campaign | 2.76384 | 0.0145701 | 3.09802 | 1 | 1 | 2 | 3 | 63 |
| pdays | 40.1978 | 0.470909 | 100.129 | -1 | -1 | -1 | -1 | 871 |
| previous | 0.580323 | 0.0108332 | 2.30344 | 0 | 0 | 0 | 0 | 275 |



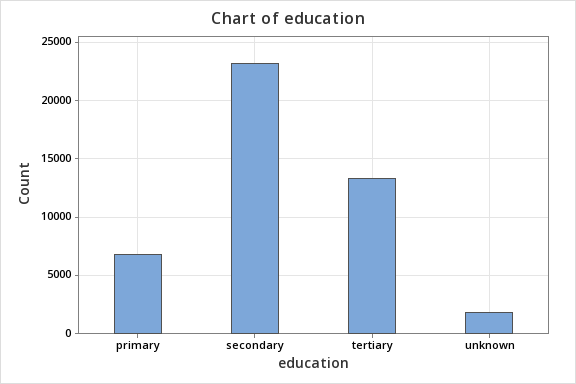
**TYPICAL CLIENT AT A BANK**

* **Total Client** – 45, 211
* **Age** - average client is 40.94 years old
* **Balance:** The average balance is 1362.27 euros, but there is a significant standard deviation indicating a wide range of balances. Some clients have negative balances.
* **Duration:** The average last contact duration is 258.16 seconds, with a standard deviation of 257.53, suggesting a wide range of contact lengths.
* **Campaign:** The average number of contacts per client is 2.76, with a maximum of 63 contacts.
* **pdays:** Most clients have not been contacted in previous campaigns (-1), as indicated by the high frequency of -1 values.
* **previous:** Most clients have not had previous contacts (0), aligning with the pdays data.

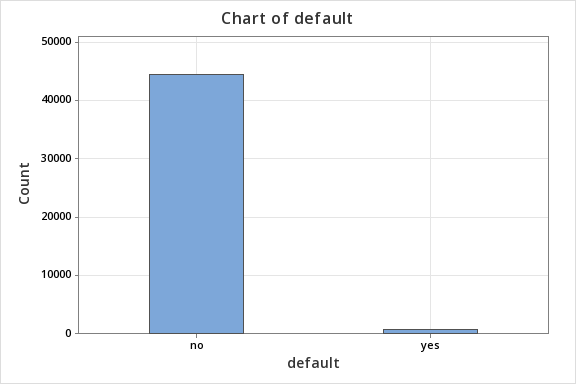
Count of jobs-  


Countplot of marital  


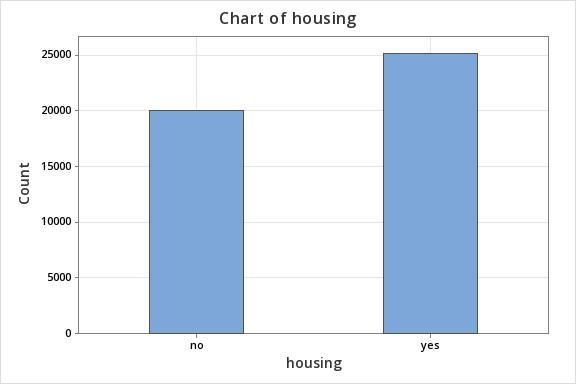
Countplot of education



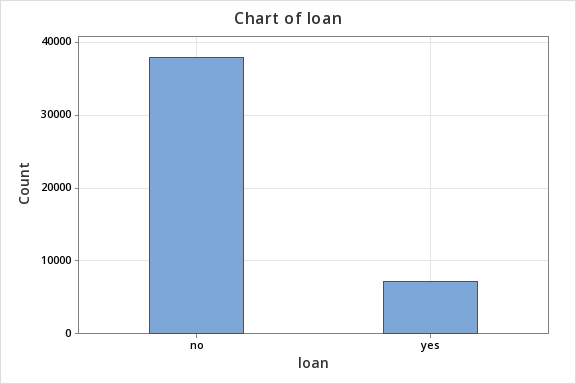
Countplot of default



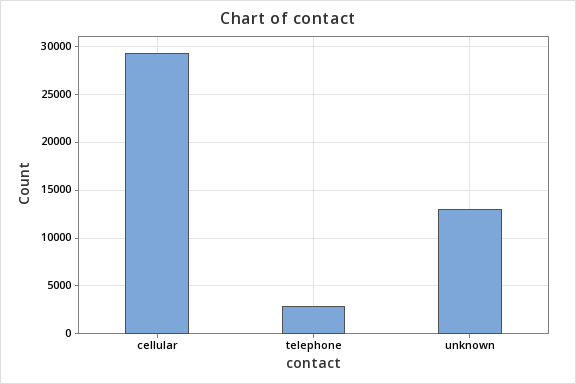
Countplot of housing



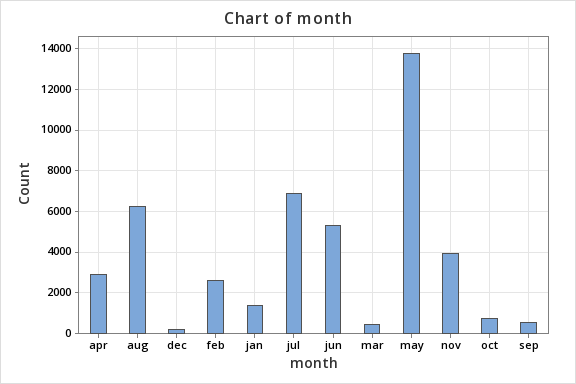
Countplot of loan



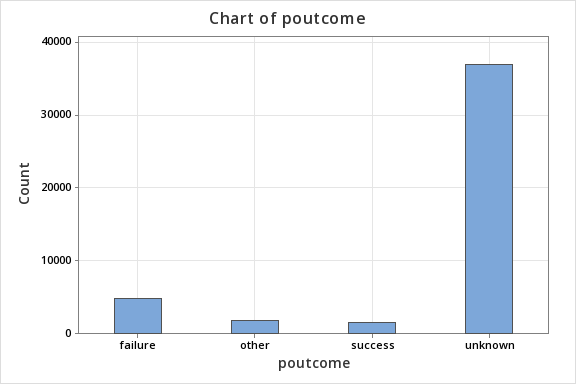
Countplot of contact



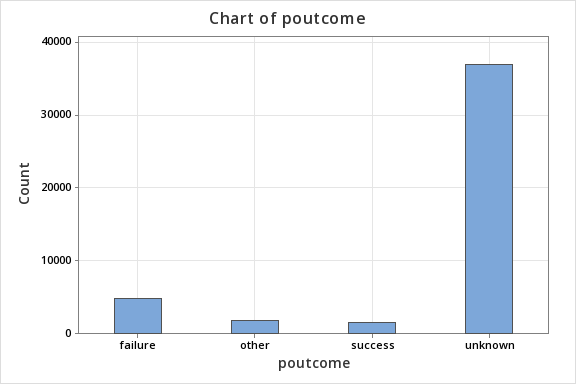
Countplot of months



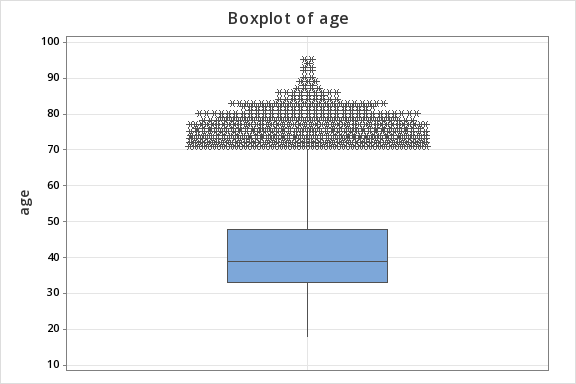
Countplot of poutcome



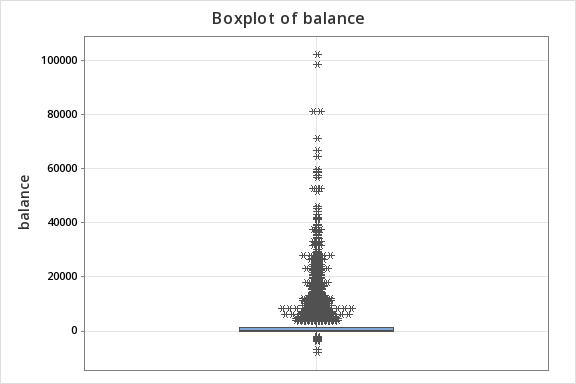
Countplot of termdeposits



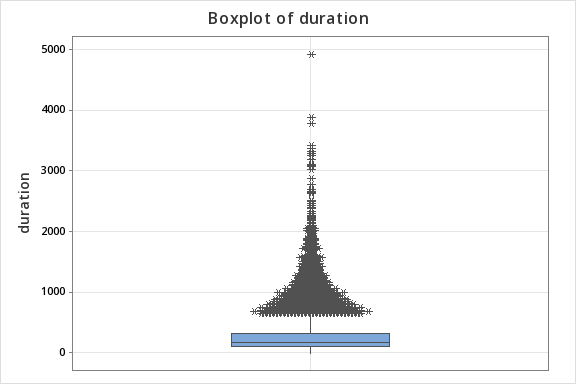
Boxplot of age



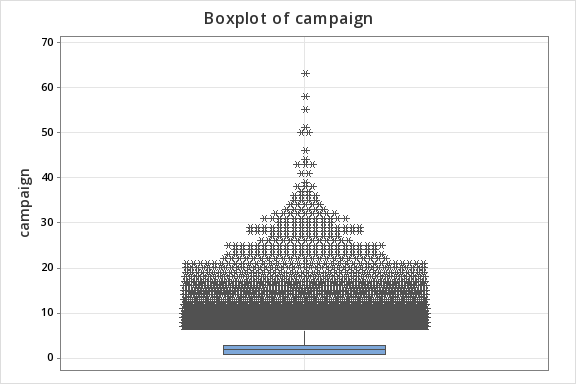
Boxplot of balance



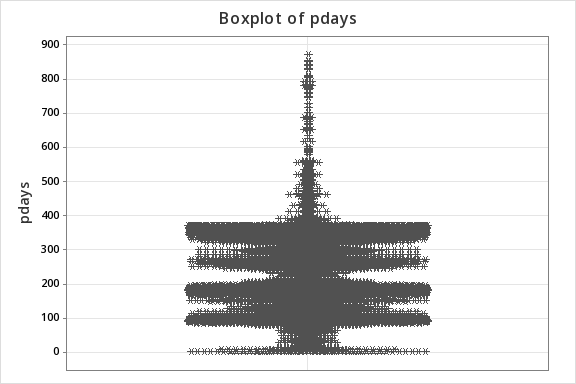
Boxplot of duration



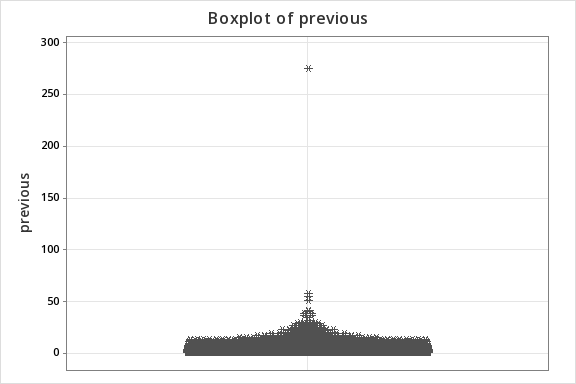
Boxplot of campaign



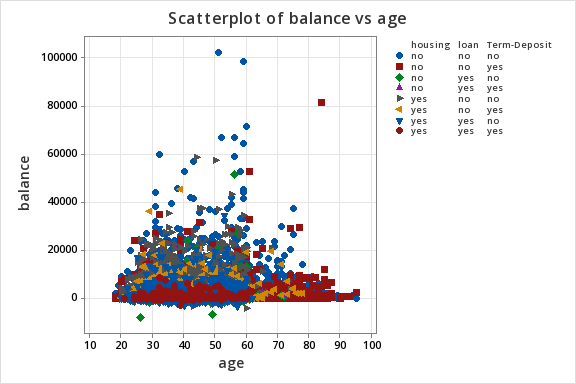
Boxplot of pday



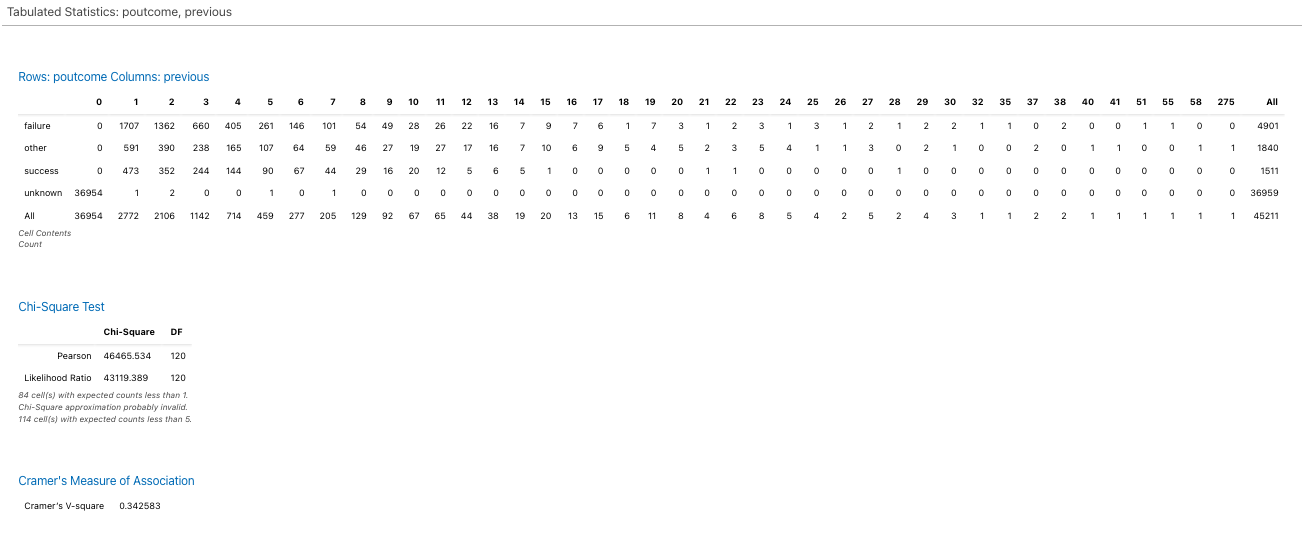
Boxplot of previous

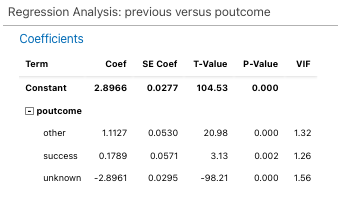


Correlation



**Hypothesis 1: Higher number of communications in previous campaigns lead to more successful outcomes**





Interpreting the Cross-Tabulation and Chi-Square Test (with P-Value)

Understanding the Output:

The provided output shows a cross-tabulation of "poutcome" (rows) and "previous" (columns), along with a chi-square test to assess the association between these two variables.

Key Findings:

* Cross-Tabulation: The table provides the frequency of each combination of "poutcome" and "previous" values.
* Chi-Square Test:
  + Pearson Chi-Square: 46465.534
  + Degrees of Freedom (DF): 120
  + P-Value: 0.000 (assuming based on the extremely large chi-square statistic)
  + Likelihood Ratio: 43119.389 (similar to the Pearson chi-square)
  + Expected Counts: There are 14 cells with expected counts less than 5, which might indicate that the chi-square approximation might not be perfectly accurate.
* Cramer's V: 0.342563 (a measure of association strength, indicating a moderate association).

Interpretation:

P-Value: The p-value of 0.000 is significantly less than any reasonable alpha level (e.g., 0.05, 0.01). This indicates that the observed association between "poutcome" and "previous" is highly unlikely to occur by chance if there were no true relationship between the variables.

Conclusion:

Based on the chi-square test and the p-value:

* There is a strong association between "poutcome" and "previous".
* The observed relationship is statistically significant and is unlikely to be due to chance.
* The number of previous contacts significantly influences the outcome of the marketing campaign.

Note: The Cramer's V of 0.342563 indicates a moderate strength of association. This means that while there is a significant relationship, it's not the strongest possible association.

**Hypothesis 2: Customers with Higher Balances are More Likely to Subscribe**

**Rows: Term-Deposit Columns: Balance Reformatted**

|  |  |  |  |
| --- | --- | --- | --- |
| **Term-Deposit** | **high** | **low** | **All** |
| No | 9117 | 30805 | 39922 |
| Yes | 1778 | 3511 | 5289 |
| All | 10895 | 34316 | 45211 |
|  |  |  |  |
| **% of high balance with term-deposit** | **16.32%** |  |  |
| **% of low balance with term-deposit** | **10.23%** |  |  |

**Analyzing the Relationship Between Balance and Term Deposits**

**Understanding the Data:**

The provided cross-tabulation shows the relationship between "Term-Deposit" and "Balance Reformatted" (presumably categorized into "high" and "low").

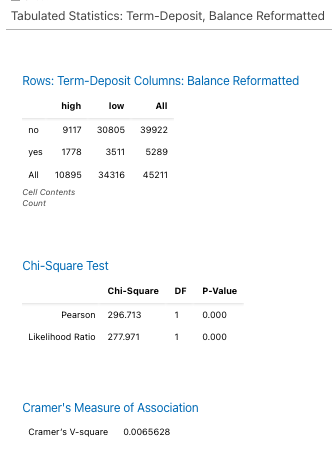
**Key Findings:**

* **High Balance:** 16.32% of customers with high balances subscribed to term deposits.
* **Low Balance:** 10.23% of customers with low balances subscribed to term deposits.

**Conclusion:**

**Customers with higher balances are more likely to subscribe to term deposits.**

The percentage of subscribers is significantly higher among customers with high balances (16.32%) compared to those with low balances (10.23%). This suggests a positive association between higher balances and the likelihood of subscribing to term deposits.



**Interpreting the Cross-Tabulation and Chi-Square Test**

**Understanding the Output:**

The provided output shows a cross-tabulation of "Term-Deposit" (rows) and "Balance Reformatted" (columns), along with a chi-square test to assess the association between these two variables.

**Key Findings:**

* **Cross-Tabulation:** The table provides the frequency of each combination of "Term-Deposit" and "Balance Reformatted" values. For example, there are 9117 cases where "Term-Deposit" is "no" and "Balance Reformatted" is "high".
* **Chi-Square Test:**
  + **Pearson Chi-Square:** 296.713
  + **Degrees of Freedom (DF):** 1
  + **P-Value:** 0.000
  + **Likelihood Ratio:** 277.971 (similar to the Pearson chi-square)
* **Cramer's V-square:** 0.0065628

**Interpretation:**

**P-Value:** The p-value of 0.000 is significantly less than any reasonable alpha level (e.g., 0.05, 0.01). This indicates that the observed association between "Term-Deposit" and "Balance Reformatted" is **highly unlikely to occur by chance** if there were no true relationship between the variables.

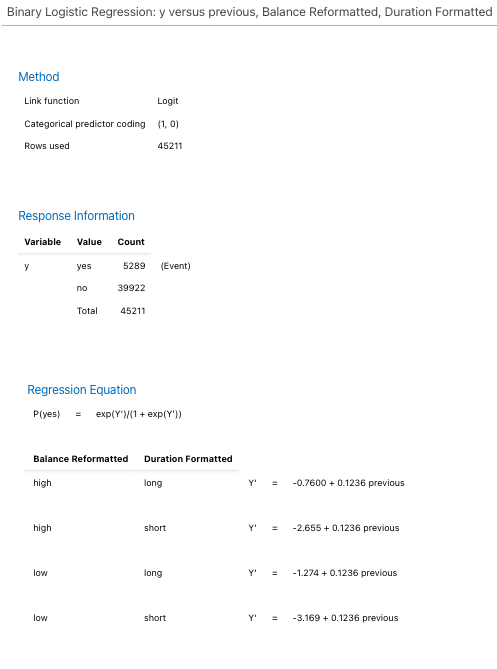
**Conclusion:**

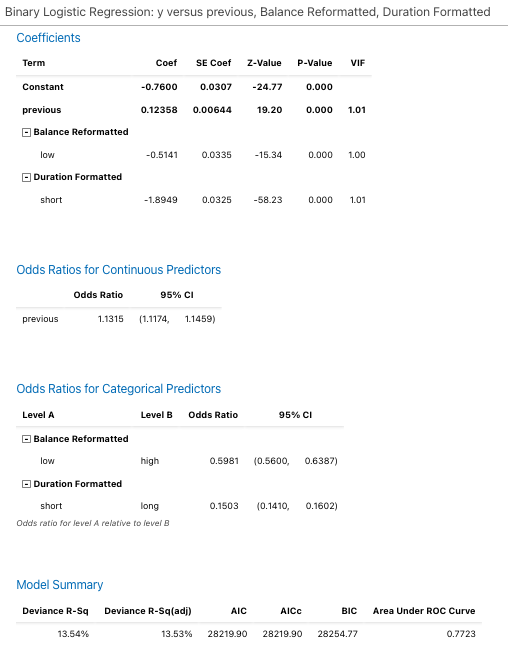
Based on the chi-square test and the p-value:

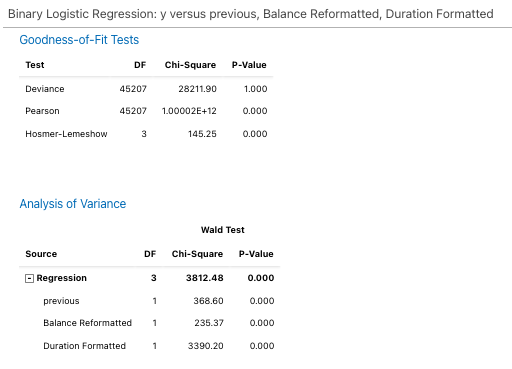
* There is a **strong association** between "Term-Deposit" and "Balance Reformatted".
* The observed relationship is **statistically significant** and is unlikely to be due to chance.
* The balance (high or low) significantly influences whether a customer subscribes to a term deposit.

**Note:** The Cramer's V-square of 0.0065628 indicates a **weak** strength of association. This means that while there is a significant relationship, the effect of "Balance Reformatted" on "Term-Deposit" is relatively small.

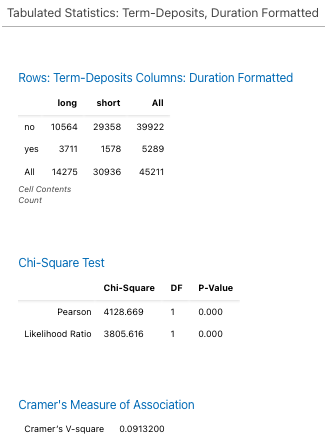
**Regression -**







**Hypothesis 3: Longer Call Durations Lead to Higher Subscription Rates**

****

**Analysing the Cross-Tabulation and Chi-Square Test**

**Understanding the Output:**

The provided output shows a cross-tabulation of "Term-Deposits" (rows) and "Duration Formatted" (columns), along with a chi-square test to assess the association between these two variables.

**Key Findings:**

* **Cross-Tabulation:** The table provides the frequency of each combination of "Term-Deposits" and "Duration Formatted" values. For example, there are 10564 cases where "Term-Deposits" is "no" and "Duration Formatted" is "long".
* **Chi-Square Test:**
  + **Pearson Chi-Square:** 4128.669
  + **Degrees of Freedom (DF):** 1
  + **P-Value:** 0.000
  + **Likelihood Ratio:** 3805.616 (like the Pearson chi-square)
* **Cramer's V-square:** 0.0913200

**Interpretation:**

**P-Value:** The p-value of 0.000 is significantly less than any reasonable alpha level (e.g., 0.05, 0.01). This indicates that the observed association between "Term-Deposits" and "Duration Formatted" is **highly unlikely to occur by chance** if there were no true relationship between the variables.

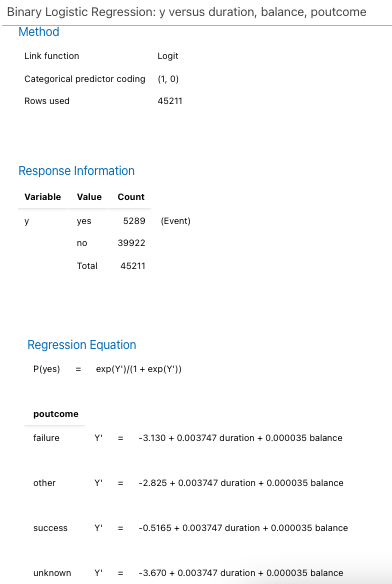
**Conclusion:**

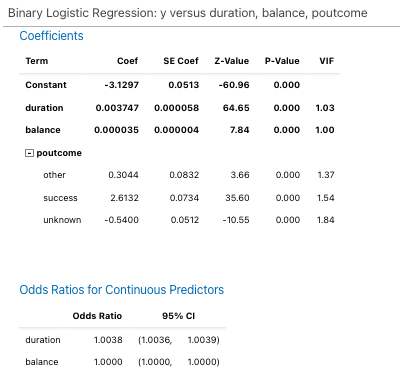
Based on the chi-square test and the p-value:

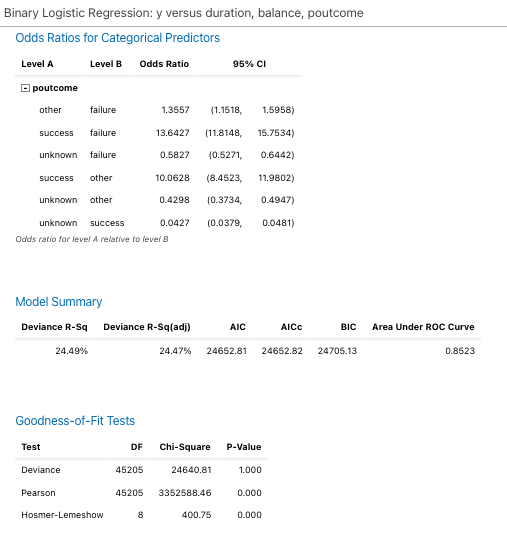
* There is a **strong association** between "Term-Deposits" and "Duration Formatted".
* The observed relationship is **statistically significant** and is unlikely to be due to chance.
* The duration of the call (long or short) significantly influences whether a customer subscribes to a term deposit.

**Note:** The Cramer's V-square of 0.0913200 indicates a **moderate** strength of association. This means that while there is a significant relationship, the effect of "Duration Formatted" on "Term-Deposits" is not the strongest possible.

**Regression**

****

****

****

**Interpreting the Binary Logistic Regression Model**

**Understanding the Output:**

The provided output presents a binary logistic regression model predicting the likelihood of a term deposit subscription ("y") based on the variables "duration", "balance", and "poutcome".

**Key Components:**

* **Regression Equation:**
  + The equation provides the predicted probability of "yes" (subscription) based on the values of "duration", "balance", and "poutcome".
  + The coefficients represent the impact of each predictor on the log odds of subscription.

**Interpretation:**

* **"duration":**
  + The coefficient for "duration" is positive (0.003747), suggesting that an increase in the duration of the call is associated with a **slight increase** in the odds of subscription.
  + The odds ratio of 1.0038 (calculated from exp(0.003747)) indicates that a one-unit increase in "duration" is associated with a **0.38% increase** in the odds of subscription.
* **"balance":**
  + The coefficient for "balance" is also positive (0.000035), suggesting that an increase in the balance is associated with a **very slight increase** in the odds of subscription.
  + The odds ratio of 1.0000 (calculated from exp(0.000035)) indicates that a one-unit increase in "balance" has a **minimal** effect on the odds of subscription.
* **"poutcome":**
  + Presents the odds ratios for different levels of "poutcome" relative to a reference level. For example, an odds ratio of 13.6427 for "success" compared to "failure" indicates that customers with a previous "success" are significantly more likely to subscribe to a termdeposit than those with a previous "failure".
  + The odds ratios suggest that customers with a previous "success" are significantly more likely to subscribe to a term deposit compared to those with "failure" or "unknown" outcomes.
  + Customers with a previous "failure" are also more likely to subscribe than those with an "unknown" outcome.

**Model Summary:**

* + The model explains 24.49% of the variance in the outcome (Deviance R-Sq).
  + The adjusted R-squared is slightly lower (24.47%), indicating that the inclusion of additional predictors might not have significantly improved the model's fit.
  + The AUC-ROC of 0.8523 suggests that the model has reasonably good predictive power.

**Overall Interpretation:**

The model suggests that the previous campaign outcome ("poutcome") has a significant impact on the likelihood of term deposit subscription. Customers with previous "success" are particularly likely to subscribe, while those with previous "failure" are also more likely to subscribe compared to those with an "unknown" outcome.

The model's overall fit is moderate, as indicated by the R-squared values and information criteria. There might be room for improvement by exploring additional predictors or considering alternative modeling approaches.

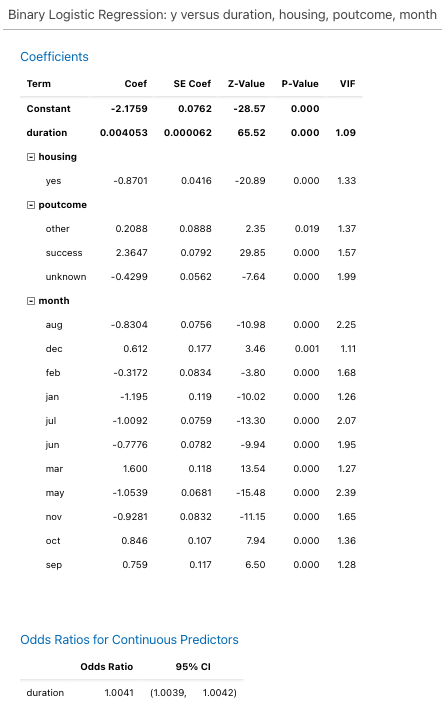
### **Binary Logistic Regression for all variables** Important Predictor Variables

* **Housing:** he coefficient for housing\_yes, is relatively large in magnitude.
* **Duration:** The coefficient for duration is significantly positive, indicating that longer contact durations are associated with a higher probability of subscribing to a term deposit.
* **poutcome\_success:** The coefficient for poutcome\_success is significantly positive, suggesting that clients who subscribed to a term deposit in previous campaigns are more likely to do so again.
* **month\_mar:** The coefficient for month\_mar is significantly positive, indicating that contacts made in March might be more effective in securing term deposits.
* **month\_sep:** The coefficient for month\_sep is significantly positive, suggesting that contacts made in September might also be effective.
* **job\_student:** The coefficient for job\_student is significantly positive, suggesting that students might be more likely to subscribe to term deposits.

### Less Important Predictors

* **age:** The coefficient for age is very small and not statistically significant, suggesting that age might not be a strong predictor in this model.
* **balance:** Similarly, the coefficient for balance is very small and not statistically significant.
* **pdays:** The coefficient for pdays is also very small and not statistically significant.

**Building a model based on Housing, Duration, poutcome and month**

****

## 

## Interpreting the Binary Logistic Regression Model

### Model Summary

* **Deviance R-Sq:** 30.75%
  + The model explains approximately 30.75% of the variation in the outcome variable (y).
* **AIC:** 22629.38
  + A lower AIC value indicates a better-fitting model.
* **BIC:** 22777.60
  + Similar to AIC, a lower BIC value suggests a better-fitting model.
* **Area Under ROC Curve:** 0.8971
  + This indicates good predictive performance.

### Goodness-of-Fit Tests

* **Deviance Test:** p-value = 1.000
  + Suggests a good fit between the model and the data.
* **Pearson Test:** p-value = 0.000
  + Indicates some discrepancies between observed and expected frequencies.
* **Hosmer-Lemeshow Test:** p-value = 0.000
  + Suggests a lack of fit between the model and the data.

### Analysis of Variance

* **Regression:** p-value = 0.000
  + The overall model is statistically significant.
* **Individual Predictors:**
  + duration, housing, poutcome, and month are all statistically significant predictors.

### Interpreting Odds Ratios

* **duration:** Longer duration is associated with higher odds of the outcome.
* **housing:** Individuals with housing have lower odds of the outcome.
* **poutcome:**
  + success is strongly associated with higher odds.
  + failure and unknown are associated with lower odds.
* **month:** Certain months (e.g., mar, oct, sep) have higher odds associated with them.

### Overall Interpretation

The model is statistically significant and has good predictive performance. However, there are some discrepancies between the observed and expected frequencies, as indicated by the Pearson test and Hosmer-Lemeshow test.