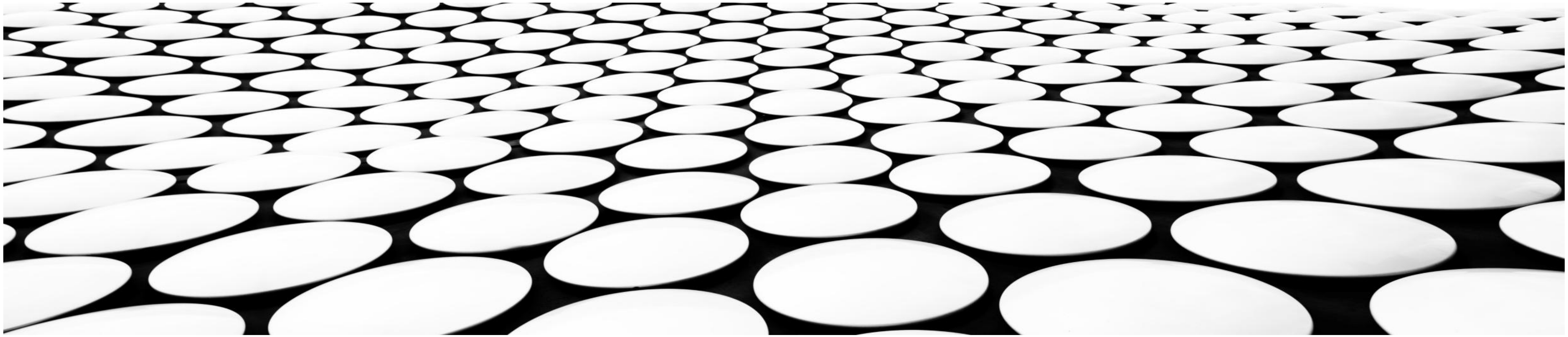
# DATA MINING AND PREDICTIVE DATA ANALYTICS

## CHAPTER-3

## EXPLORATORY DATA ANALYSIS



# INTRODUCTION TO EDA

- EDA is the Foundation of All Data Mining
  - EDA is the **first step** in any data mining task, helping analysts understand data before modeling.
- EDA Converts Raw Data into Insightful Understanding
  - It transforms raw, unorganized data into **interpretable information** by summarizing distributions, detecting anomalies, and revealing hidden trends.
- EDA Combines Statistics with visualization
  - EDA blends **quantitative summaries** (mean, variance, correlation) with **graphical methods** (histograms, boxplots, scatterplots) to uncover relationships that numbers alone might miss.

# INTRODUCTION TO EDA

- **EDA Guides Data Cleaning, Transformation, and Feature Selection**
  - Through EDA, we identify **missing values, outliers, redundancies, and variable correlations**.
  - It helps decide which features to **keep, discard, or transform**, ensuring that the subsequent data mining model is both **efficient and meaningful**.
- **EDA Bridges Business Context and Analytical Modeling**
  - In data mining, EDA serves as the **bridge between domain understanding and algorithmic modeling**.
  - It allows analysts to **align statistical findings with business logic**, ensuring that the models not only perform well but also **make practical, actionable sense**.

# HYPOTHESIS TESTING VS EXPLORATORY DATA ANALYSIS

- Two distinct approaches to data analysis
  - **Hypothesis Testing**: A confirmatory, formal procedure that tests a pre-specified idea or assumption.
  - **Exploratory Data Analysis (EDA)**: An open-ended, discovery-oriented process where the goal is to learn what the data suggest without a fixed hypothesis.
- Both approaches play a complementary role in data mining, statistics, and machine learning.

# HYPOTHESIS TESTING VS EXPLORATORY DATA ANALYSIS

- ## Hypothesis Testing
  - Hypothesis Testing is a **formal statistical procedure** used to evaluate whether a statement (hypothesis) about a population parameter is supported by sample data.
  - **Key Features**
    - Starts with an **a priori hypothesis** (before examining the data in detail).
    - Involves **null hypothesis ($H_0$)** and **alternative hypothesis ($H_1$)**.
    - Provides a **yes/no decision** (reject or fail to reject $H_0$).
  - **Example**
    - Mobile phone operators may hypothesize:
      - $H_0$: Market share has **not decreased** after fee hike.
      - $H_1$: Market share has **decreased** after fee hike.
      - Hypothesis testing procedures would be applied to evaluate this claim

# HYPOTHESIS TESTING VS EXPLORATORY DATA ANALYSIS

- **Exploratory Data Analysis (EDA)**
  - Exploratory Data Analysis (EDA) is an approach to analyzing datasets that emphasizes **visual exploration and descriptive statistics** to uncover patterns, anomalies, and relationships without relying on predetermined assumptions.
- **Primary reasons** for performing EDA is to:
  - Investigate the variables in the dataset.
  - Examine the distributions of **categorical variables** (e.g., frequency counts, bar charts).
  - Look at the **histograms of numeric variables** to understand their spread and shape.
  - Explore the **relationships among sets of variables**, both predictors and target variables.
  - Detect outliers, missing values, and data quality issues.
  - Develop initial hypotheses and guide subsequent modeling.

# HYPOTHESIS TESTING VS EXPLORATORY DATA ANALYSIS

- **Common EDA Techniques**
  - **Graphical**: Histograms, scatter plots, box plots, correlation heatmaps,
  - **Numerical**: Summary statistics (mean, median, variance, skewness,), correlation coefficients.
  - **Subset/Group analysis**: Identifying clusters, trends, or interesting subsets.
- EDA acts as the **foundation of data analysis**, shaping the direction of further investigation and hypothesis testing.

# HYPOTHESIS TESTING VS EXPLORATORY DATA ANALYSIS

- ## Complementary Roles
  - **EDA often comes first (**discovery stage) → helps analysts understand the dataset, distributions, and uncover important relationships and patterns that could indicate important areas for further investigation.
  - **Hypothesis testing follows (**confirmation stage) → validates the patterns or suspicions suggested by EDA with statistical rigor, i.e, testing assumptions with formal procedures.
  - Together, they form a **powerful cycle of discovery and confirmation** in data mining and statistical analysis.

# HYPOTHESIS TESTING VS EXPLORATORY DATA ANALYSIS

| Aspect | Hypothesis Testing | Exploratory Data Analysis (EDA) |
|---|---|---|
| **Purpose** | To confirm or reject a pre-specified idea. | To discover patterns, understand distributions, and generate new ideas. |
| **Approach** | Deductive, confirmatory. | Inductive, discovery-oriented. |
| **When Used** | When clear, theory-driven questions exist. | When data are unfamiliar, large, or complex. |
| **Focus** | Formal decision-making. | Investigation of variables, distributions, and relationships. |
| **Tools** | Statistical tests (t-test, chi-square, ANOVA, regression). | Graphical (histograms, scatter plots, box plots) and descriptive statistics. |
| **Outcome** | Binary decision (reject/fail to reject $H_0$). | Insights, hypotheses, directions for further study. |
| **Flexibility** | Rigid, structured. | Flexible, iterative. |

# EDA On The Churn Dataset – A Case Study

- In this case study
  - The Churn Dataset (UCI M/L Repository) is used to demonstrate EDA methods applied in a real-world business scenario.
- EDA helps in:
  - Detecting anomalies or missing data
  - Identifying patterns and relationships among variables
  - Suggesting potential predictors for the target variable
  - Gaining domain insights through visualizations and summary statistics before any formal modeling

# CHURN EXAMPLE- GETTING TO KNOW THE DATASET

- **Overview of the Dataset:**
  - **Number of Observations** (Rows): 3,333 customers
  - **Number of Predictors** (Features): 20
  - **Target Variable:** Churn – indicates whether a customer has left the company (True or False).
  - The dataset contains a mix of categorical, integer-valued, and continuous features describing customer demographics, account information, service usage, and interactions with customer service.

# CHURN EXAMPLE- GETTING TO KNOW THE DATASET

- **Variables in the Dataset:**
  - **(a) Customer Identification**
    1. **State** – Categorical; 50 U.S. states and the District of Columbia.
    2. **Account length** – Integer; duration (in days) the account has been active.
    3. **Area code** – Categorical; geographical area code.
    4. **Phone number** – Unique identifier (effectively a surrogate for customer ID).
  - **(b) Service Plans**
    5. **International plan** – Dichotomous categorical (Yes/No).
    6. **Voice mail plan** – Dichotomous categorical (Yes/No).
    7. **Number of voice mail messages** – Integer; count of saved messages.

# CHURN EXAMPLE- GETTING TO KNOW THE DATASET

- (C) Usage Metrics

8. **Total day minutes** – Continuous; daytime minutes used.
9. **Total day calls** – Integer; number of calls made during the day.
10. **Total day charge** – Continuous; charges (linked to day usage).
11. **Total eve minutes** – Continuous; evening minutes used.
12. **Total eve calls** – Integer; number of evening calls.
13. **Total eve charge** – Continuous; charges (linked to evening usage).
14. **Total night minutes** – Continuous; night-time minutes used.
15. **Total night calls** – Integer; number of night-time calls.
16. **Total night charge** – Continuous; charges (linked to night usage).
17. **Total international minutes** – Continuous; international call duration.
18. **Total international calls** – Integer; count of international calls.
19. **Total international charge** – Continuous; charges (linked to international usage).

- **(d) Customer Service Interaction**

20. **Number of calls to customer service** – Integer; reflects customer complaints or queries.

- **(e) Target Variable**

21. **Churn** – Boolean (True/False); indicates if the customer left the company.

# CHURN EXAMPLE- GETTING TO KNOW THE DATASET

| Variable | Type | Description |
|---|---|---|
| State | Categorical | 51 US states + DC |
| Account length | Integer | Duration of account in days |
| Area code | Categorical | Area classification |
| Phone number | Identifier | Surrogate for Customer ID |
| International plan | Dichotomous | Yes / No |
| Voice mail plan | Dichotomous | Yes / No |
| Number of voice mail messages | Integer | Number of messages |
| Total day minutes / calls / charge | Continuous / Integer | Usage during the day |
| Total evening minutes / calls / charge | Continuous / Integer | Usage during evening |
| Total night minutes / calls / charge | Continuous / Integer | Usage during night |
| Total international minutes / calls / charge | Continuous / Integer | International call activity |
| Number of calls to customer service | Integer | Frequency of customer support calls |
| **Churn (Target)** | Flag (True/False) | Customer left or stayed |

# CHURN EXAMPLE- GETTING TO KNOW THE DATASET

| Type | Variables | Description |
|---|---|---|
| **Categorical** | State, Area Code | Indicate geographic origin. |
| **Identification** | Phone number | Serves as a customer ID surrogate. |
| **Flag Variables** | International Plan, Voice Mail Plan | Dichotomous variables: Yes/No. |
| **Numerical (Continuous/Integer)** | Account length, number of voice mail messages, total day/eve/night/international minutes and calls, total charges, number of customer service calls | Capture usage statistics. |
| **Target** | Churn | Whether the customer left (True) or stayed (False). |

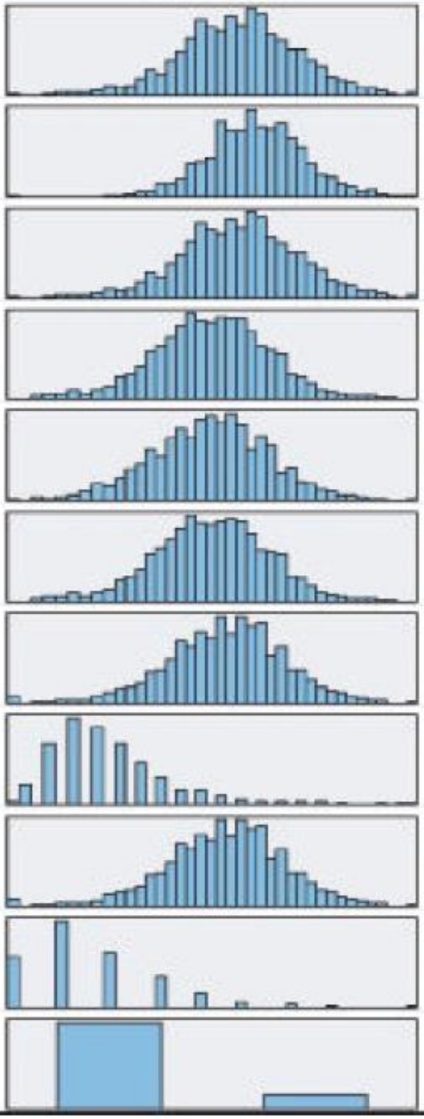
# CHURN EXAMPLE- GETTING TO KNOW THE DATASET

- **Preliminary Observations**
  - *Phone* is purely an identifier, not a predictor.
  - Two flag variables exist (*International Plan, Voice Mail Plan*).
  - Most variables are continuous.
  - Response variable *Churn* is binary.
- Visualization tools (histograms) and summary stats for each variable.
- Some variables (e.g., *Intl Calls, CustServ Calls*) are right-skewed;
- Most others appear near-normal.

# CHURN EXAMPLE- GETTING TO KNOW THE DATASET

| Field | Sample Graph | Type | Min | Max | Mean | Std. Dev | Skewn... | Median | Mode | Unique | Valid |
|---|---|---|---|---|---|---|---|---|---|---|---|
| State | | Set | -- | -- | -- | -- | -- | -- | WV | 51 | 3333 |
| Account Length | | Range | 1 | 243 | 101.065 | 39.822 | 0.097 | 101 | 105 | -- | 3333 |
| Area Code | | Set | 408 | 510 | -- | -- | -- | -- | 415 | 3 | 3333 |
| Intl Plan | | Flag | -- | -- | -- | -- | -- | -- | no | 2 | 3333 |
| VMail Plan | | Flag | -- | -- | -- | -- | -- | -- | no | 2 | 3333 |
| VMail Message | | Range | 0 | 51 | 8.099 | 13.688 | 1.265 | 0 | 0 | -- | 3333 |
| Day Mins | | Range | 0.000 | 350.800 | 179.775 | 54.467 | -0.029 | 179.400 | 154.000' | -- | 3333 |
| Day Calls | | Range | 0 | 165 | 100.436 | 20.069 | -0.112 | 101 | 102 | -- | 3333 |
| Day Charge | | Range | 0.000 | 59.640 | 30.562 | 9.259 | -0.029 | 30.500 | 26.180' | -- | 3333 |

# CHURN EXAMPLE- GETTING TO KNOW THE DATASET



| Field | Sample Graph | Type | Min | Max | Mean | Std. Dev | Skewn... | Median | Mode | Unique | Valid |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Eve Mins | | Range | 0.000 | 363.700 | 200.980 | 50.714 | -0.024 | 201.400 | 169.900 | -- | 3333 |
| Eve Calls | | Range | 0 | 170 | 100.114 | 19.923 | -0.056 | 100 | 105 | -- | 3333 |
| Eve Charge | | Range | 0.000 | 30.910 | 17.084 | 4.311 | -0.024 | 17.120 | 14.250' | -- | 3333 |
| Night Mins | | Range | 23.200 | 395.000 | 200.872 | 50.574 | 0.009 | 201.200 | 188.200' | -- | 3333 |
| Night Calls | | Range | 33 | 175 | 100.108 | 19.569 | 0.032 | 100 | 105 | -- | 3333 |
| Night Charge | | Range | 1.040 | 17.770 | 9.039 | 2.276 | 0.009 | 9.050 | 9.450' | -- | 3333 |
| Intl Mins | | Range | 0.000 | 20.000 | 10.237 | 2.792 | -0.245 | 10.300 | 10.000 | -- | 3333 |
| Intl Calls | | Range | 0 | 20 | 4.479 | 2.461 | 1.321 | 4 | 3 | -- | 3333 |
| Intl Charge | | Range | 0.000 | 5.400 | 2.765 | 0.754 | -0.245 | 2.780 | 2.700 | -- | 3333 |
| CustServ Calls | | Range | 0 | 9 | 1.563 | 1.315 | 1.091 | 1 | 1 | -- | 3333 |
| Churn | | Flag | -- | -- | -- | -- | -- | -- | False | 2 | 3333 |

# CHURN EXAMPLE- GETTING TO KNOW THE DATASET

- **Objective of EDA- To see which variables are associated with *Churn***
- One of the primary reasons for performing EDA is to investigate the variables,
  - examine the distributions of the categorical variables,
  - look at the histograms of the numeric variables, and
  - explore the relationships among sets of variables.
- However, our overall objective for the data mining project as a whole (not just the EDA phase) is to develop a model of the type of customer likely to churn

# UNIVARIATE VS. MULTIVARIATE ANALYSIS

- **Univariate analysis** explores **a single variable** in isolation to understand its **distribution, central tendency, spread, and shape**.
- It does not deal with relationships or dependencies
- **Purpose**
  - Understand data range, outliers, and overall pattern.
  - Identify missing or extreme values.
  - Decide on data transformations (e.g., normalization, log-scaling).
  - Check assumptions for future modeling.

| Type of Variable | Common Techniques | Visualization |
|---|---|---|
| Categorical | Frequency counts, proportions, mode | Bar chart, pie chart |
| Numerical | Mean, median, standard deviation, skewness, | Histogram, box plot, density plot |

# UNIVARIATE VS. MULTIVARIATE ANALYSIS

- **Multivariate analysis** investigates **two or more variables simultaneously** to detect **patterns, relationships, correlations, and interactions** between them.
- **Purpose**
  - Find **dependencies** and **interaction effects** between variables.
  - Identify **predictors** for a target variable.
  - Support **feature selection** and **hypothesis formulation**.

| Relationship Type | Typical Analysis | Visualization |
|---|---|---|
| Two categorical | Contingency table, Chi-square test | Clustered bar chart |
| One categorical + one numeric | Group means, box plots | Side-by-side boxplots |
| Two numeric | Correlation, regression line | Scatter plot |
| Many numeric | PCA, heatmap | Matrix plots |

# EXPLORING CATEGORICAL VARIABLES

- **Understanding the Target Distribution**

| Value | Proportion | % | Count |
|---|---|---|---|
| False | | 85.51 | 2850 |
| True | | 14.49 | 483 |

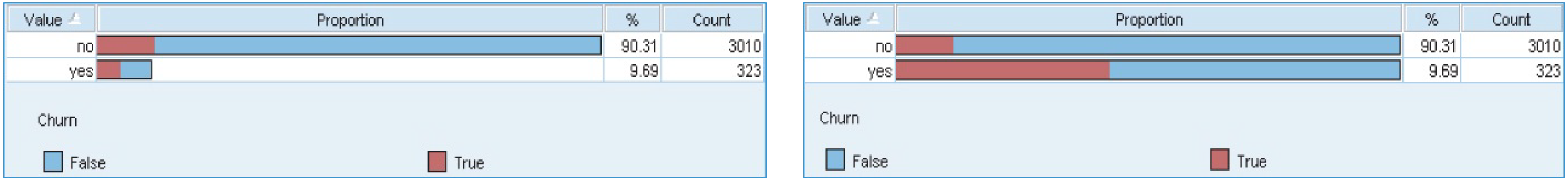
- Only **14.49 % of customers churned**.

- **Objective:** To identify the Categorical Variables variables influencing this minority class.
- We are to test TWO Categorical Variables:
  - ***International Plan,***
  - ***Voice Mail Plan***

# EXPLORING CATEGORICAL VARIABLES

- ***International Plan vs. Churn***

- A comparison of the proportion of churners and non-churners, with International Plan (yes, 9.69% of customers) or without (no, 90.31% of customers).




Bar chart of the *International Plan*, with an *overlay* of *churn*

- Clearly, those who have selected the International Plan have a greater chance of leaving the company's service than do those who do not have the International Plan

# EXPLORING CATEGORICAL VARIABLES

## CONTINGENCY TABLE

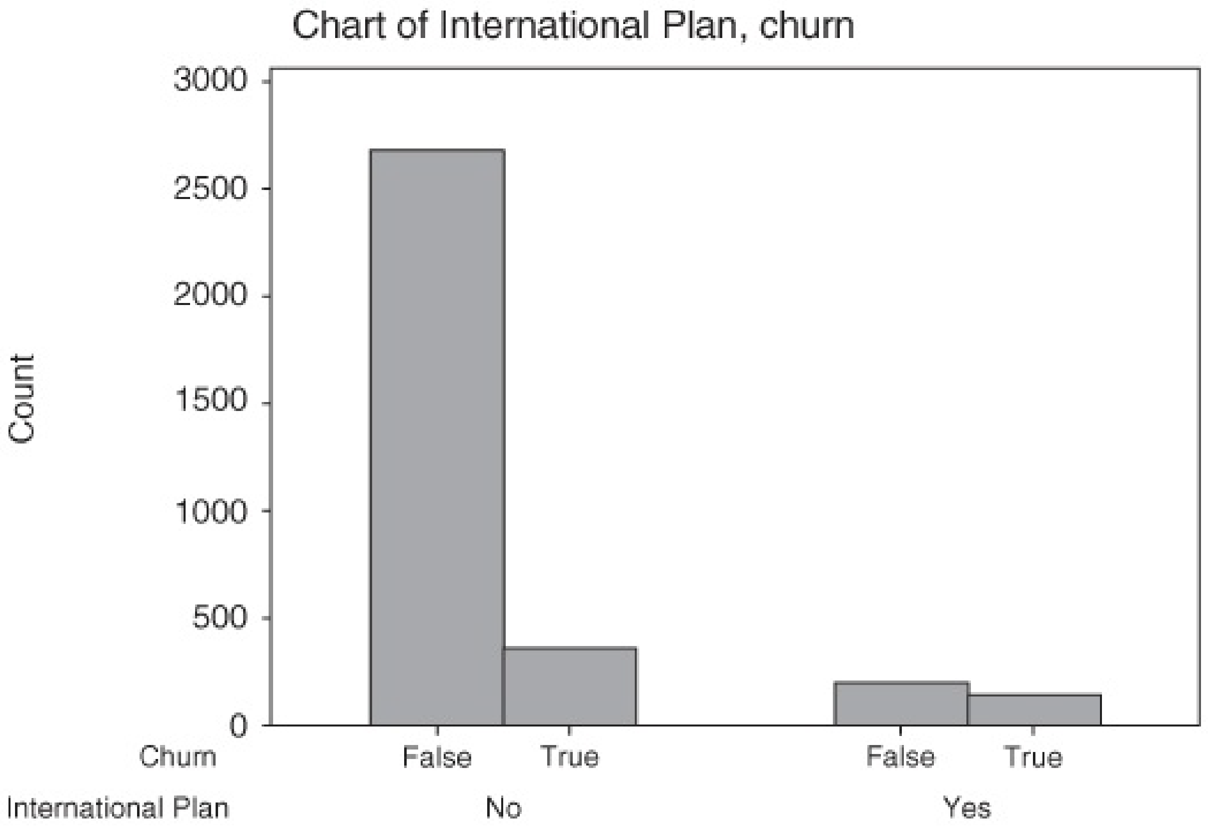
| | Intl Plan = No | Intl Plan = Yes |
|---|---|---|
| Churn = False | 2664 | 186 |
| Churn = True | 346 | 137 |

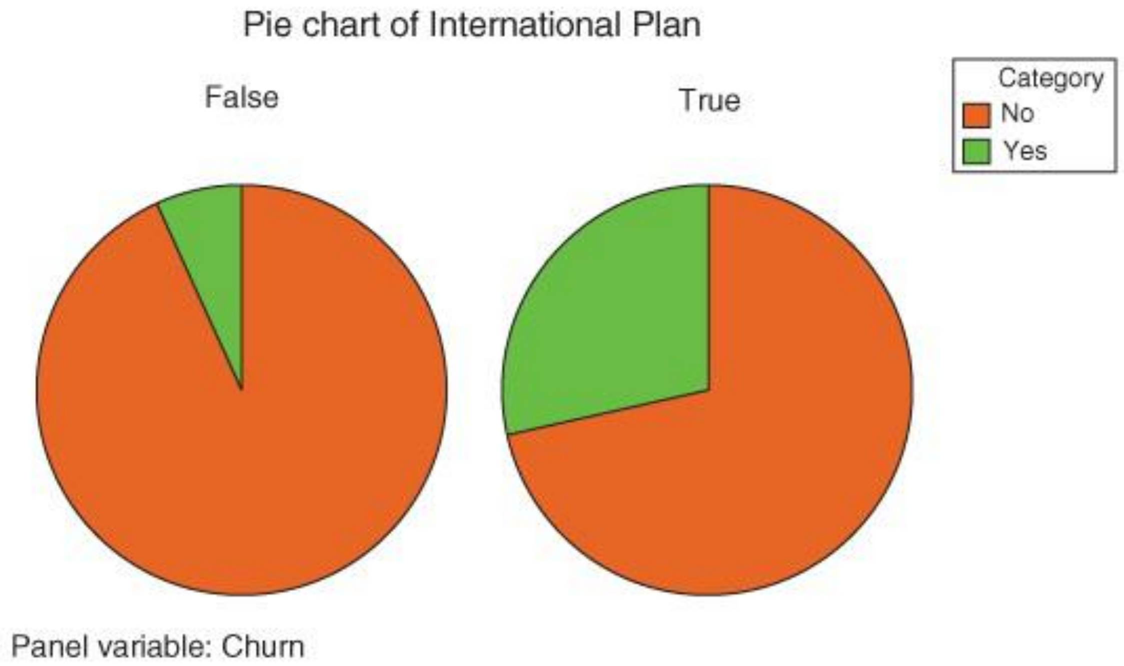
| | Intl Plan = No | Intl Plan = Yes |
|---|---|---|
| Churn = False | 88.5% | 57.6% |
| Churn = True | 11.5% | 42.4% |

- **Interpretation:**
  - Churn rate for “Yes” = 137 / (186+137) = 42.5%
  - Churn rate for “No” = 346 / (2664+346) = 11.5%
  - 42.4% of international plan holders churned, compared to only 11.5% of others.
  - Thus, customers with international plans are over 3× more likely to leave.
  - Possible business implication: Investigate dissatisfaction with international service.

# EXPLORING CATEGORICAL VARIABLES

- Clustered Bar Chart and Comparative Pie Chart
  - The graphical counterpart of the contingency table
  - Clustered Bar Chart conveys counts as well as proportions
  - Comparative pie chart conveys only proportions.



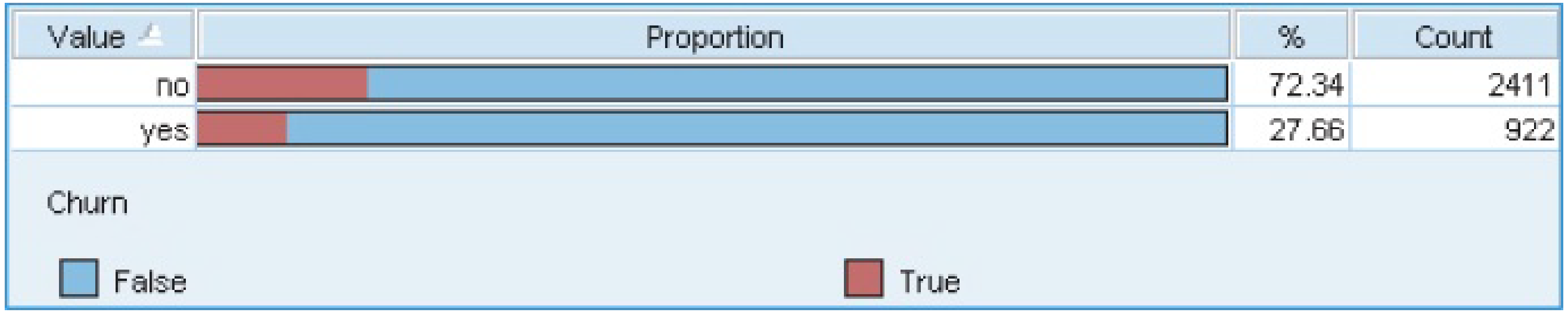



# EXPLORING CATEGORICAL VARIABLES

- To summarize, EDA on the *International Plan* has indicated that
  1. Customers selecting the *International Plan* are more than three times as likely to leave the company's service and those without the plan
  2. Perhaps we should investigate what is it about our *international plan* that is inducing our customers to leave;
  3. We should expect that, whatever data mining algorithms we use to predict *churn*, the model will probably include whether or not the customer selected the *International Plan*.

# EXPLORING CATEGORICAL VARIABLES

- ***Voice Mail Plan vs. Churn***




- Comparing using a bar graph with equalized lengths, it is observed that those who do not have the Voice Mail Plan are more likely to churn than those who do have the plan.
- The numbers in the graph indicate proportions and counts of those who do and do not have the Voice Mail Plan, without reference to churning.

# EXPLORING CATEGORICAL VARIABLES

**CONTINGENCY TABLE**

| | VMail Plan = No | VMail Plan = Yes |
|---|---|---|
| Churn = False | 83.3% | 91.3% |
| Churn = True | 16.7% | 8.7% |

| | VMail Plan = No | VMail Plan = Yes |
|---|---|---|
| Churn = False | 2008 | 842 |
| Churn = True | 403 | 80 |

- **Interpretation:**
  - Churn rate without plan = 403 / (2008+403) = 16.7%
  - Churn rate with plan = 80 / (842+80) = 8.7%
  - Those without the plan are **twice as likely to churn**.
  - Suggestion: Make voice mail plans more accessible or attractive to increase retention.

# EXPLORING CATEGORICAL VARIABLES

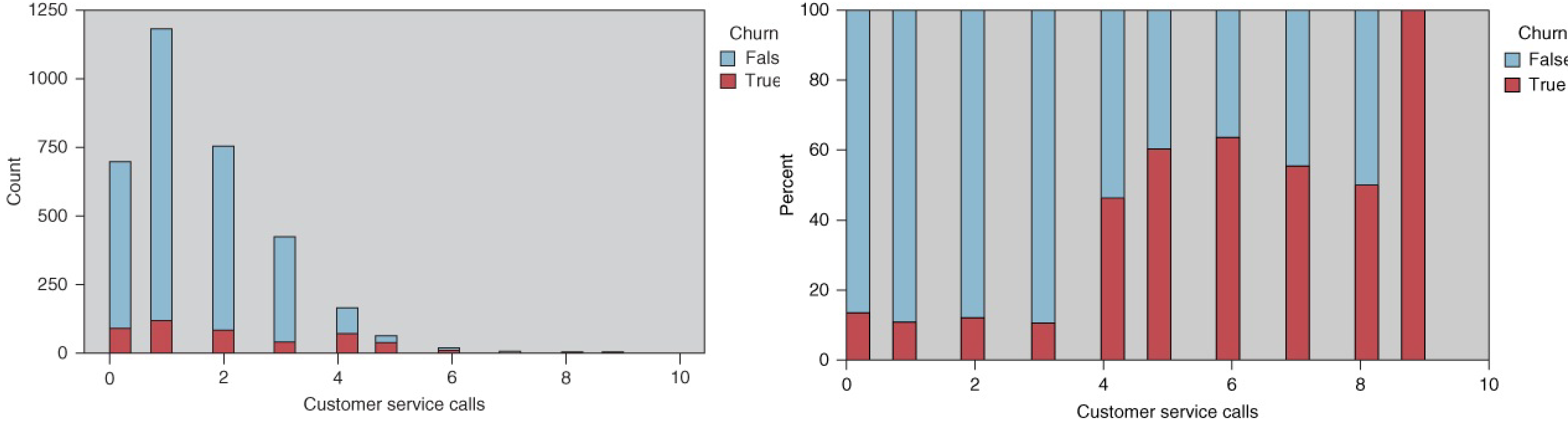
- **To summarize, this EDA on the Voice Mail Plan has indicated that**
  1. Customers without the *Voice Mail Plan* are more likely to churn.
  2. Perhaps *Voice Mail Plan* should be still enhanced further, or make it easier for customers to join it, as an instrument for increasing customer loyalty;
  3. Whatever data mining algorithms we use to predict churn, the model will probably include whether or not the customer selected the *Voice Mail Plan*.
  4. Our confidence in this expectation is perhaps not quite as high as for the *International Plan*.

# EXPLORING NUMERIC VARIABLES

- EDA of numeric predictors focuses on data shape, symmetry, and relation to the target variable.
- **Univariate Patterns**
  - *Account length* and most usage fields are roughly symmetric.
  - *Voice mail messages* has median = 0 (half customers lack the service).
  - *Customer service calls* shows right skew (few customers make many calls).
- **Overlay and Normalized Histograms**
  - Plain histograms show frequency, but **overlay histograms** (color-coded by churn) reveal how predictors relate to the target.

# EXPLORING NUMERIC VARIABLES

- *Customer Service Calls* With *Churn* Overlay




- Customers who have called customer service three times or less have a markedly lower churn rate (red part of the rectangle) than customers who have called customer service four or more times.

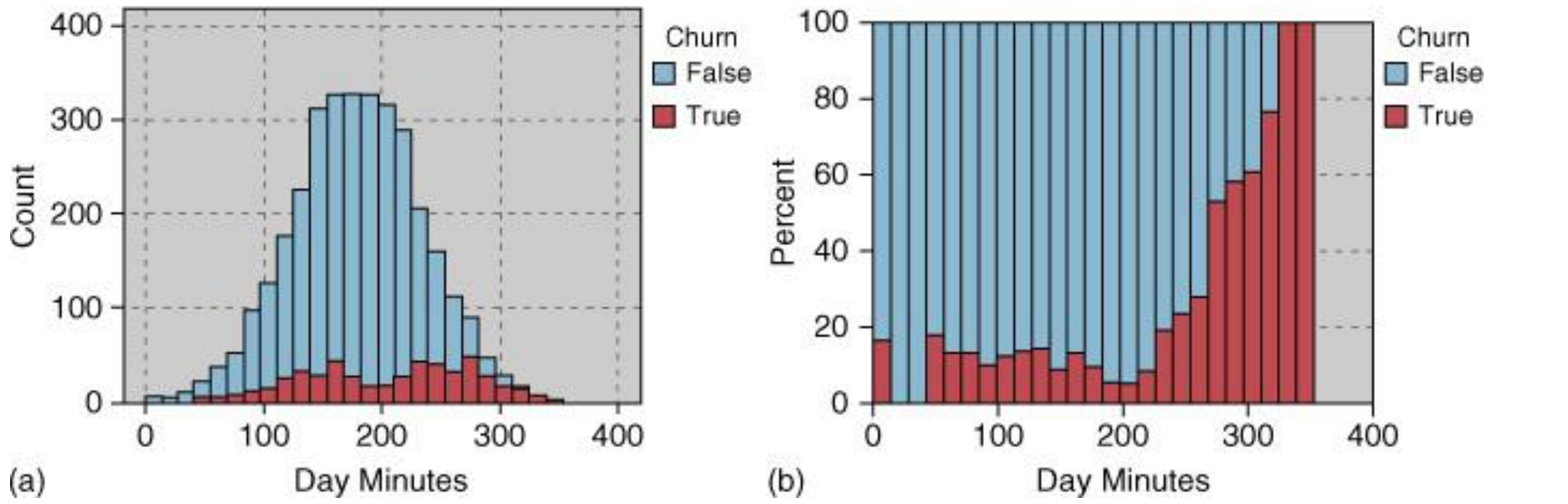
# EXPLORING NUMERIC VARIABLES

- This EDA on the *customer service calls* has indicated that

1. We should carefully track the number of *customer service calls* made by each customer.
   - By the third call, specialized incentives should be offered to retain customer loyalty, because, by the fourth call, the probability of *churn* increases greatly;
2. we should expect that, whatever data mining algorithms we use to predict churn, the model will probably include the number of *customer service calls* made by the customer.

# EXPLORING NUMERIC VARIABLES

- *Day Minutes Vs. Churn*



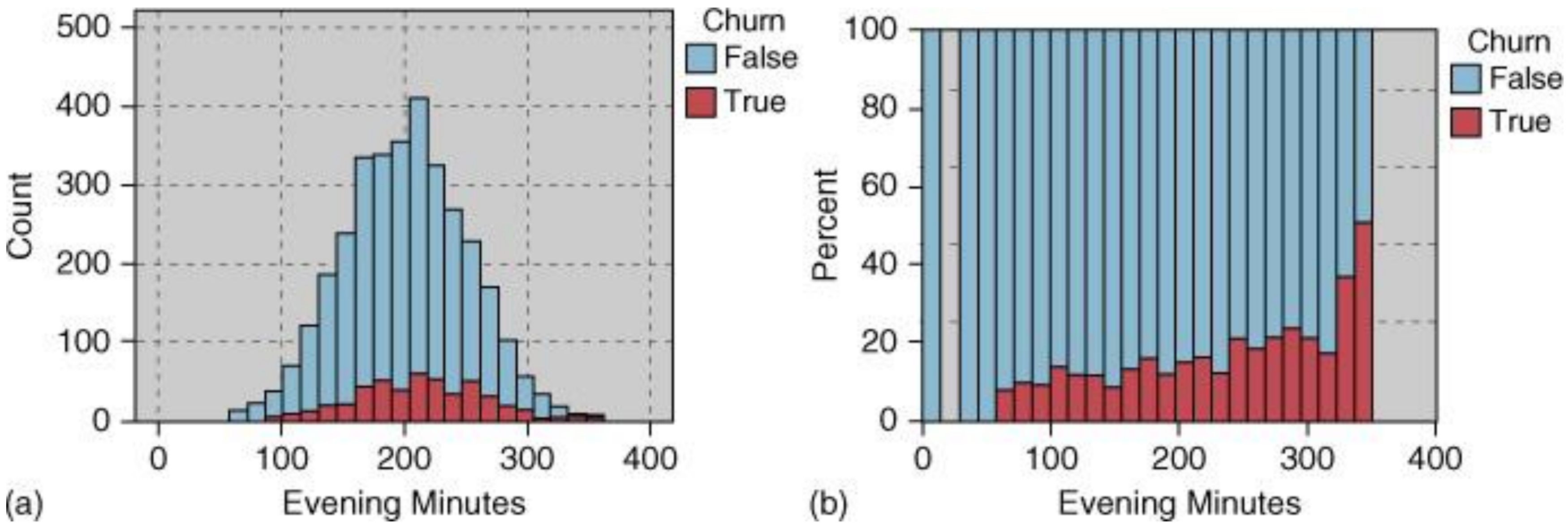

- shows a tendency for customers with higher *Day Minutes* to churn

# EXPLORING NUMERIC VARIABLES

- This EDA on the *Day Minutes* has indicated that:
  1. We should carefully track the number of day minutes used by each customer. As the number of day minutes passes 200, we should consider special incentives;
  2. We should investigate why heavy day-users are tempted to leave;
  3. We should expect that our eventual data mining model will include *day minutes* as a predictor of churn

# EXPLORING NUMERIC VARIABLES

- *Evening Minutes Vs. Churn*



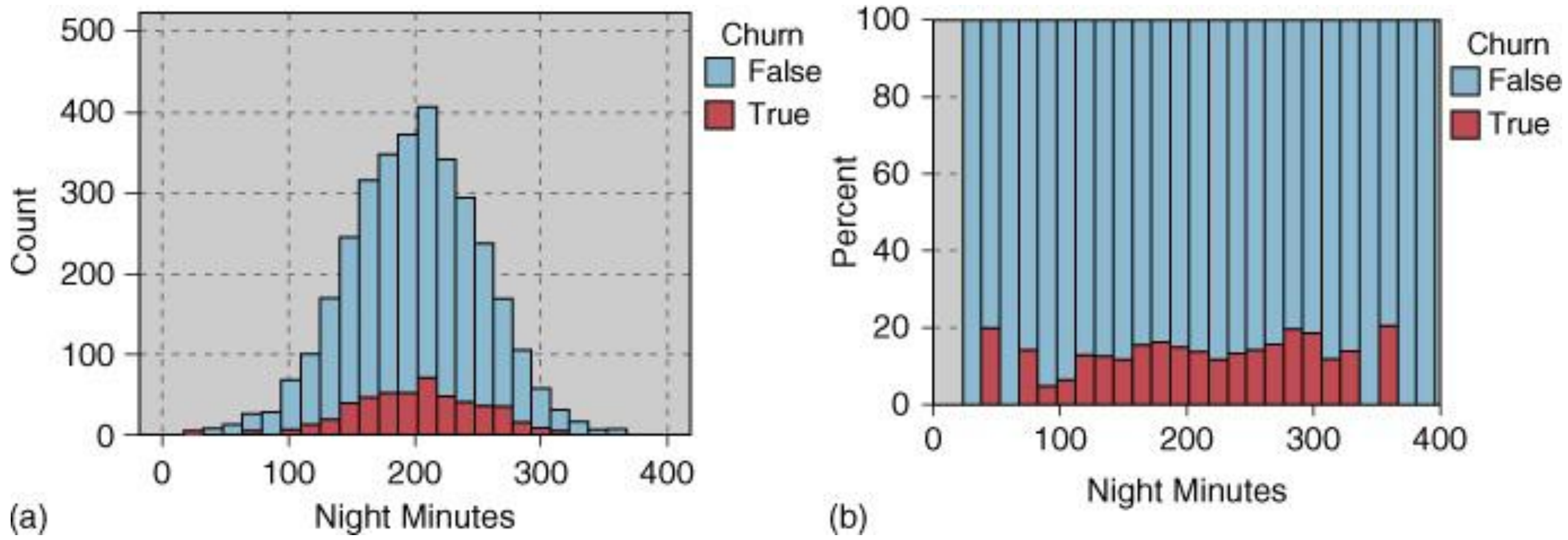

- Shows a slight tendency for customers with higher *evening minutes* to churn.

# EXPLORING NUMERIC VARIABLES

- *Evening Minutes Vs. Churn*
  - Based solely on the graphical evidence, however, we cannot conclude beyond a reasonable doubt that such an effect exists.
  - Therefore, we shall hold off on formulating policy recommendations on evening cell-phone use until our data mining models offer firmer evidence that the presumed effect is in fact present.

# EXPLORING NUMERIC VARIABLES

- *Night minutes Vs. Churn*




- This indicates that there is no obvious association between churn and *night minutes*, as the pattern is relatively flat.

# EXPLORING NUMERIC VARIABLES

- During the **Exploratory Data Analysis (EDA)** stage, a lack of clear or visible association between a **predictor variable** and the **target variable** does **not automatically justify removing** that predictor from the model.
- For example, if there is no obvious relationship between **customer churn** and **night call minutes**, it does not mean that **night minutes** is useless as a predictor.
- Even when a variable does not show an evident association at the overall level, it might still contain **valuable predictive information** for **specific subsets** of the data.

# EXPLORING NUMERIC VARIABLES

- Some predictors may also participate in **complex, higher-dimensional interactions** with other variables that are not visible in simple pairwise analysis.
- Therefore, it is generally advisable to **retain such variables** for the **data mining or modeling stage**, allowing the model to evaluate their predictive contribution.
- In summary, **variables should only be excluded** before modeling if there is a **strong and justified reason**, such as redundancy, irrelevance, or data quality issues

# EXPLORING MULTIVARIATE RELATIONSHIPS

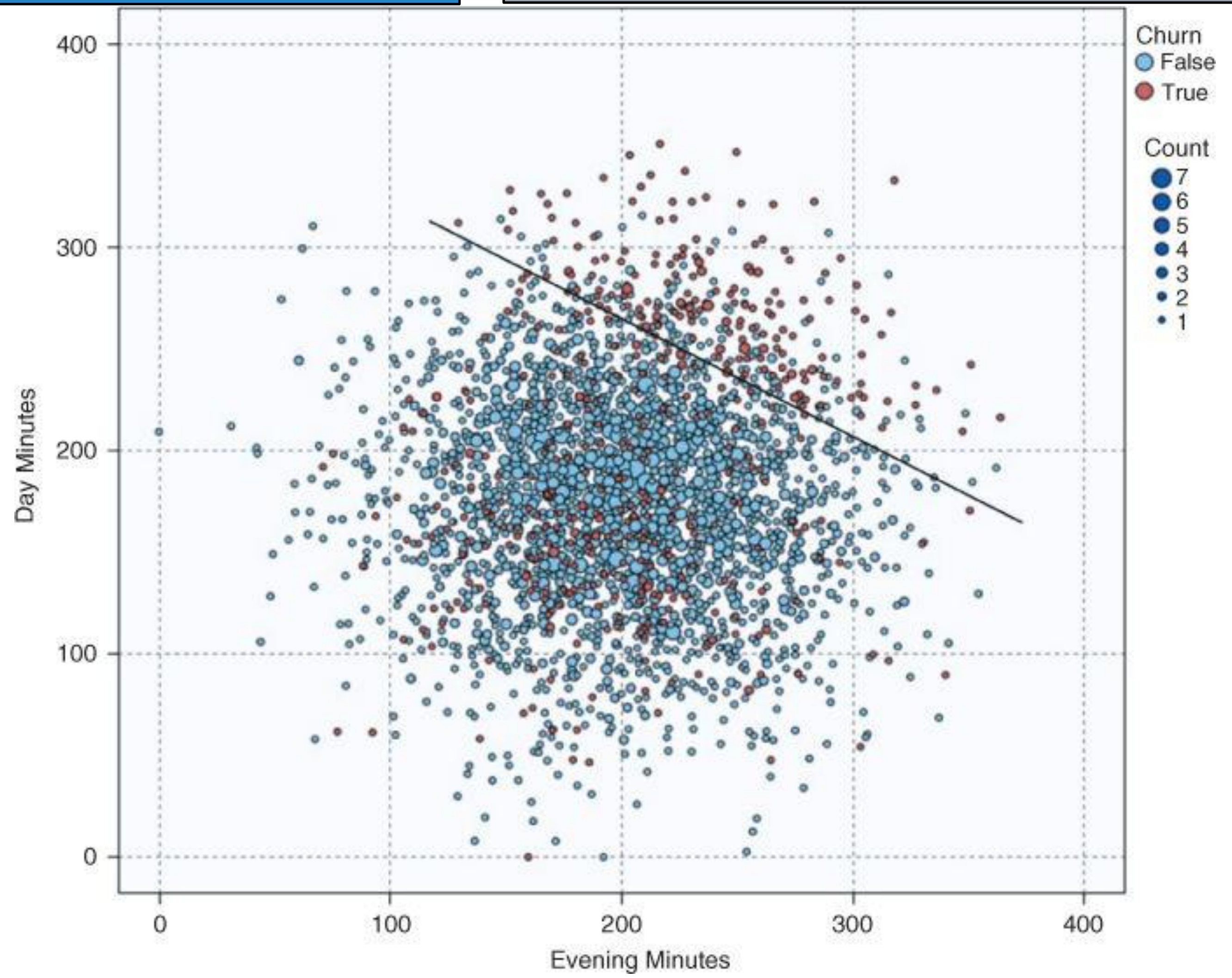
- We next turn to an examination of the possible **multivariate associations** of numeric variables with *churn*, using **scatter plots**.
- Multivariate analysis investigates **two or more variables simultaneously** to detect **patterns, relationships, correlations, and interactions** between them.
- **Purpose**
  - Find **dependencies** and **interaction effects** between variables.
  - Identify **predictors** for a target variable.
- Multivariate graphics can uncover new interaction effects which our univariate exploration missed.

# EXPLORING MULTIVARIATE RELATIONSHIPS

- ***Day minutes* and *Evenings minutes* Vs. *Churn***
  - The univariate evidence for a high churn rate for high evening minutes was not conclusive (in previous univariate analysis)
  - Hence, it is nice to have a multivariate graph that supports the association, at least for customers with high day minutes.
- The EDA confirms
  - Customers with both high *day minutes* and high *evening minutes* are at greater risk of churning.

# EXPLORING MULTIVARIATE RELATIONSHIPS

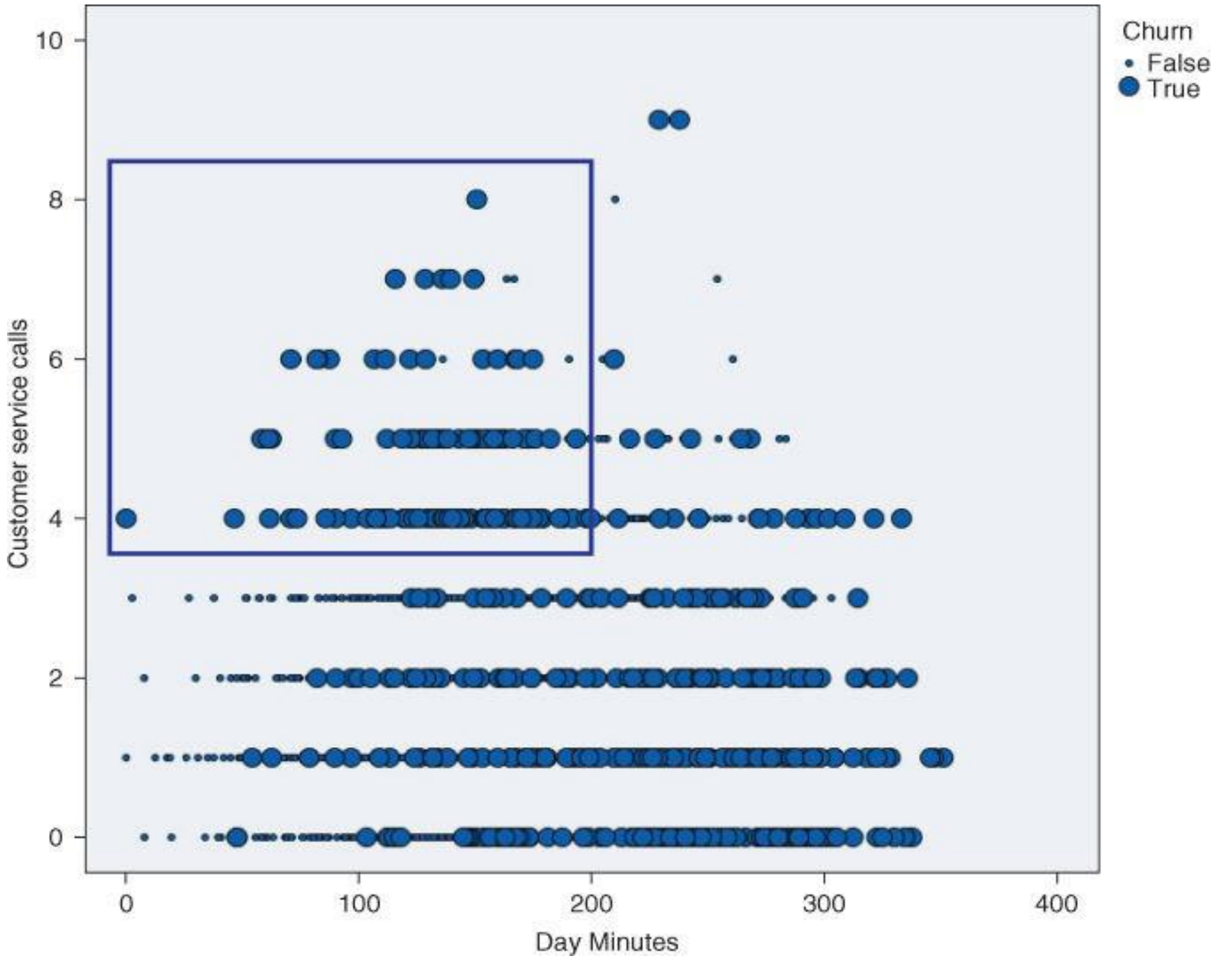
- Note the straight line partitioning off the upper right section of the graph.
- Records above this diagonal line, representing customers with both high *day minutes* and high *evening minutes*, appear to have a higher proportion of churners than records below the line.




scatter plot of *day minutes* versus *evenings minutes*, with churners indicated by the darker circles.

# EXPLORING MULTIVARIATE RELATIONSHIPS

- ***Customer service and Day minutes VS. Churn***
- records inside the rectangle partition in the scatter plot, which indicates a high-churn area in the upper left section of the graph.
- These records represent customers who have a combination of a high number of customer service calls and a low number of day minutes used.




scatter plot of *customer service calls* versus *day minutes*.

# EXPLORING MULTIVARIATE RELATIONSHIPS

- In general, customers with higher numbers of *customer service calls* tend to churn at a higher rate (as we learned earlier in the univariate analysis)
- However, this analysis shows that, of these customers with high numbers of *customer service calls*, those who also have high *day minutes* are somewhat “protected” from this high churn rate.
- The customers in the upper right of the scatter plot exhibit a lower churn rate than those in the upper left.
- This group of customers **could not** have been identified had we restricted ourselves to univariate exploration

# BINNING BASED ON PREDICTIVE VALUE

- Binning the *Customer Service Calls*
  - Earlier we saw that - customers with less than four calls to *customer service* had a lower churn rate than customers who had four or more calls to *customer service*.
  - We may therefore decide to bin the *customer service calls* variable into two classes, *low* (fewer than four) and *high* (four or more)

# BINNING BASED ON PREDICTIVE VALUE

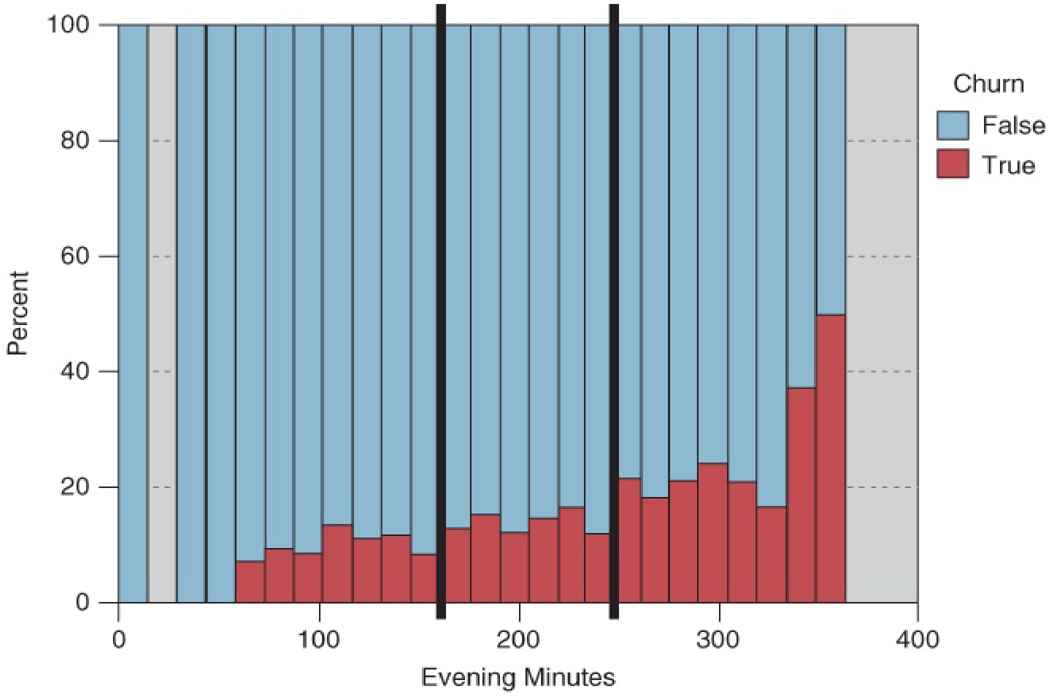
- The churn rate for customers with a low number of *customer service call* is 11.3%,
- While, the churn rate for customers with a high number of *customer service call* is 51.7%, more than four times higher.

| | Cust. Serv. Call= No | Cust. Serv. Call= Yes |
|---|---|---|
| Churn = False | 2721 (88.7%) | 129 (48.3%) |
| Churn = True | 345 (11.3%) | 138 (51.7%) |

- This binning of *customer service calls* created a **flag variable** with two values, high and low.

# BINNING BASED ON PREDICTIVE VALUE

- **Binning the *Evening Minutes***
  - Recall that – relationship between *evening minutes* and *churn* was **inconclusive.**




*Deciding Bin boundaries that will maximize the difference in churn proportions?*

- The first boundary = 160, as the group of rectangles to the right of this boundary seem to have a higher proportion of churners than the group of rectangles to the left.
- The second boundary = 240 for the same reason.

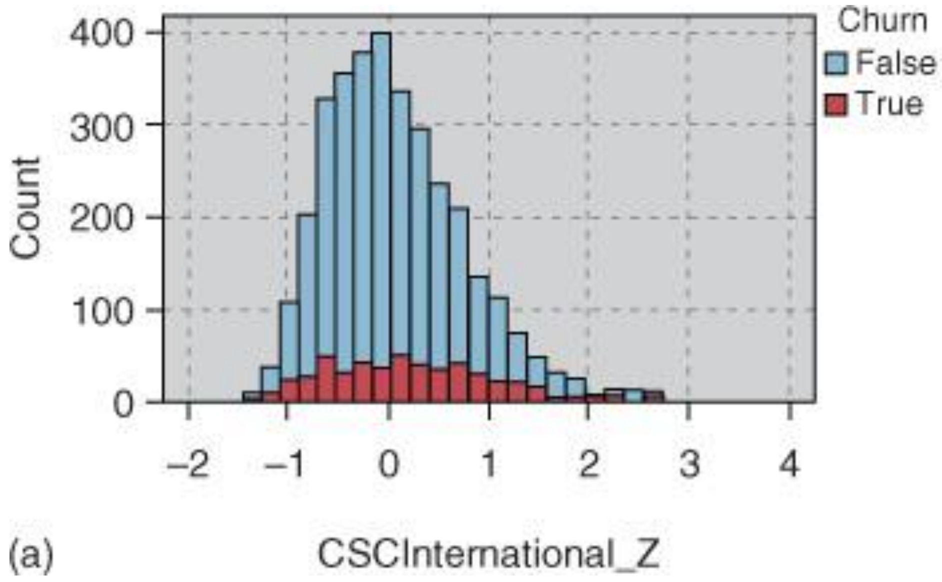
- Binning *Evening Minutes* creates an ordinal categorical variable with three values, low, medium, and high.

# DERIVING NEW VARIABLES: NUMERICAL VARIABLES

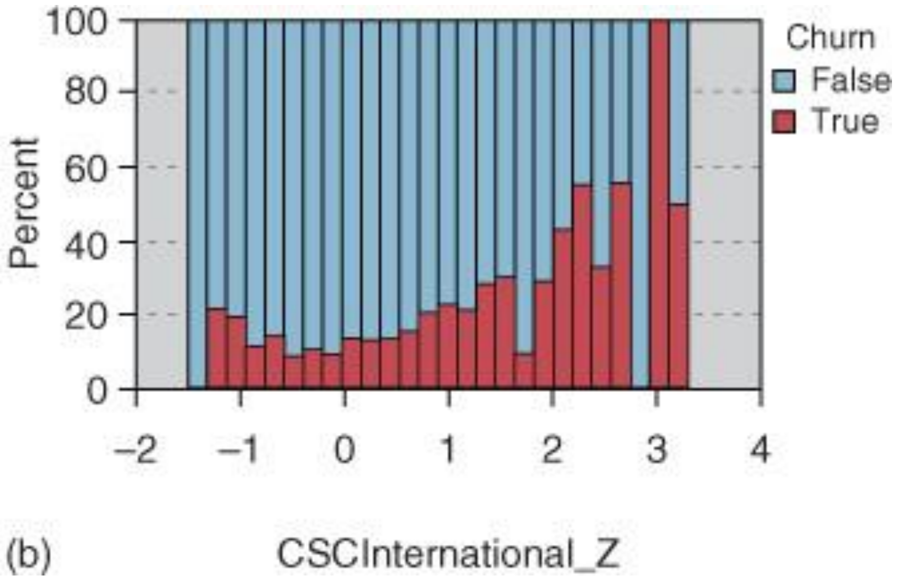
- Combining *Customer Service Calls* and *International Calls*
  - Suppose we derive a new numerical variable combining *Customer Service Calls* and *International Calls*, and whose values will be the mean of the two fields.
  - *International Calls* have a larger mean and SD than *Customer Service Calls*,
  - Unwise to take the mean of the raw field values, as *International Calls* would thereby be more heavily weighted.
  - Instead, when combining numerical variables, we first need to standardize.
  - The new derived variable therefore takes the form:

$$CSCInternational\_Z = \frac{(CSC\_Z + International\_Z)}{2}$$

# DERIVING NEW VARIABLES: NUMERICAL VARIABLES




Non-normalized histogram of *CSCInternational_Z*.




Normalized histogram of *CSCInternational_Z*.

# EDA TO INVESTIGATE CORRELATED PREDICTOR VARIABLES

- ## What is Correlation?
  - Two variables **x** and **y** are **linearly correlated** when changes in one are associated with changes in the other.
  - **Positive correlation:** $x \uparrow \rightarrow y \uparrow$
  - **Negative correlation:** $x \uparrow \rightarrow y \downarrow$
- ## Correlation Coefficient (r)
  - Measures the **strength and direction** of the linear relationship between two variables.
  - **Range:** $-1 \leq r \leq +1$
    - **r = +1:** perfect positive correlation
    - **r = –1:** perfect negative correlation
    - **r = 0:** no linear relationship
- ## Statistical Significance
  - For **large datasets (n > 1000)**, even small values of *r* (e.g., 0.05 or 0.1) can be **statistically significant**.

# EDA TO INVESTIGATE CORRELATED PREDICTOR VARIABLES

- **Impact of Using Correlated Variables**
  - **Overemphasis:** Repeated information exaggerates one data component.
  - **Model instability:** Leads to unreliable or fluctuating model parameters.
  - **Multicollinearity:** Makes it hard to determine which variable truly influences the target.
- **Example**
  - If *Day Minutes* and *Day Charge* are perfectly correlated (one is a linear function of the other), including both:
    - Inflates the importance of that predictor.
    - Can distort regression coefficients or split criteria in tree models.

# EDA TO INVESTIGATE CORRELATED PREDICTOR VARIABLES

- ## Strategy for Handling Correlated Predictors (During EDA)
  - **Avoid feeding correlated variables to any data mining and statistical models**

1. **Identify Perfectly Correlated Variables**
   - When |r| = 1 (e.g., *Day Minutes* and *Day Charge*).
   - **Action:** Remove one of the two; both carry the same information.
2. **Identify Groups of Correlated Variables**
   - Variables that move together but not perfectly.
   - **Action:** Keep for now, but handle later using **dimension reduction** (e.g., **PCA**).

- **Note:**
  - This strategy applies to **correlations among predictor variables**, not between **predictors and the target variable** (which are essential for modeling).

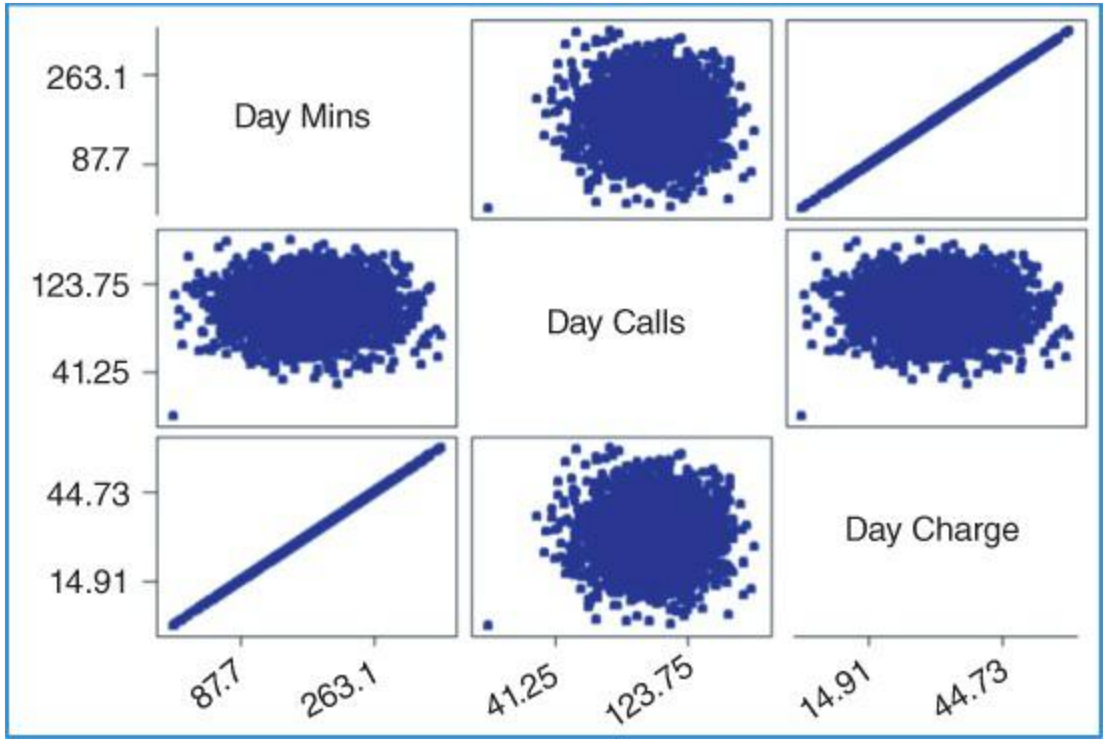
# EDA TO INVESTIGATE CORRELATED PREDICTOR VARIABLES

- **Key Takeaways**
  - Correlation helps detect **redundancy** among variables.
  - Avoid feeding highly correlated predictors into models — it causes **multicollinearity**.
  - Perfectly correlated variables → remove duplicates.
  - Moderately correlated groups → reduce dimensionality later (e.g., PCA).
  - Proper correlation handling ensures **model stability, interpretability, and efficiency**.

# EDA TO INVESTIGATE CORRELATED PREDICTOR VARIABLES

- ## Understanding Variable Relationships in the Churn Dataset
  - In the *Churn Dataset*, each of the four time periods — **Day, Evening, Night, and International** — includes three variables:
    - **Minutes** (continuous usage time)
    - **Calls** (number of calls)
    - **Charge** (total billed amount)
  - Intuitively, one might expect these to be **mutually correlated**, since higher call duration should lead to higher charges or call counts.
  - To investigate this assumption, analysts used
    - **Matrix Plot** — a grid of scatter plots for numerical variable pairs.
    - **Correlation coefficients (r)** and their **p-values**.

# EDA TO INVESTIGATE CORRELATED PREDICTOR VARIABLES




Matrix plot of *day minutes*, *day calls*, and *day charge*.

**Correlations: Day Mins, Day Calls, Day Charge**

| | Day Mins | Day Calls |
|---|---|---|
| Day Calls | 0.007 | |
| | 0.697 | |
| Day Charge | 1.000 | 0.007 |
| | 0.000 | 0.697 |

Cell Contents: Pearson correlation
P-Value

Correlations and *p*-values

# EDA TO INVESTIGATE CORRELATED PREDICTOR VARIABLES

- **Day Minutes vs. Day Calls:**
  - *$r = 0.07$, $p = 0.697$* → No meaningful linear relationship.
  - Interpretation: Number of calls does not directly depend on total call minutes.
- **Day Calls vs. Day Charge:**
  - *$r = 0.07$, $p = 0.697$* → Also weak, non-significant.
  - Unexpected, since more calls should theoretically increase total charge.
- **Day Minutes vs. Day Charge:**
  - *Perfect linear relationship found* ($r = 1.0$).
  - Indicates that charge is **a direct linear function of minutes**.

# EDA TO INVESTIGATE CORRELATED PREDICTOR VARIABLES

- **Action Taken in EDA**
  - Because *Day Charge* and *Day Minutes* convey identical information:
    - ➢ **One must be eliminated (Arbitrarily)** to prevent redundancy.
  - The analysts **retained “Day Minutes”** and **dropped “Day Charge.”**
- **Applied Consistently**
  - Similar findings appeared for:
    - Evening Charge vs. Evening Minutes
    - Night Charge vs. Night Minutes
    - International Charge vs. International Minutes
- → Therefore, **four “Charge” variables** were removed (Dimensionality reduced)

# EDA TO INVESTIGATE CORRELATED PREDICTOR VARIABLES

- **Detecting Other Correlated Predictorsfor (for later handling with PCA)**
  - After removing perfectly correlated variables, analysts proceed to identify **weaker correlations** for possible **dimension reduction** during modeling.
  - The correlation of each numerical predictor with every other numerical predictor should be checked, if feasible.
  - Correlations with small *p*-values should be identified **(Weak but statistically significant correlation)**
  - A subset of this procedure is shown next:

# EDA TO INVESTIGATE CORRELATED PREDICTOR VARIABLES

- Note that the correlation coefficient 0.038 between *account length* and *day calls* has a **small *p*-value** of 0.026, telling us that *account length* and *day calls* are **positively correlated**.
- The data analyst should note this, and prepare to apply the PCA during the modeling phase.

**Correlations: Account Leng, VMail Messag, Day Mins, Day Calls, CustServ Cal**

| | Account Length | VMail Message | Day Mins | Day Calls |
|---|---|---|---|---|
| VMail Message | -0.005 | | | |
| | 0.789 | | | |
| Day Mins | 0.006 | 0.001 | | |
| | 0.720 | 0.964 | | |
| Day Calls | 0.038 | -0.010 | 0.007 | |
| | 0.026 | 0.582 | 0.697 | |
| CustServ Calls | -0.004 | -0.013 | -0.013 | -0.019 |
| | 0.827 | 0.444 | 0.439 | 0.274 |

Cell Contents: Pearson correlation
P-Value

*Account length* is positively correlated with *day calls*

# SUMMARY OF EDA

- Following are some of the insights we have gained into the *churn* data set through the useof EDA.
  - The four *charge* fields are linear functions of the *minute* fields, and should be omitted.
  - The *area code* field and/or the *state* field are anomalous, and should be omitted until further clarification is obtained.
- Insights with respect to churn are as follows:
  - Customers with the International Plan tend to churn more frequently.

# SUMMARY OF EDA

- Customers with the Voice Mail Plan tend to churn less frequently
- Customers with four or more *Customer Service Calls* churn more than four times as often as the other customers.
- Customers with both high *Day Minutes* and high *Evening Minutes* tend to churn at a higher rate than the other customers.
- Customers with both high *Day Minutes* and high *Evening Minutes* churn at a rate about six times greater than the other customers.
- Customers with low *Day Minutes* and high *Customer Service Calls* churn at a higher rate than the other customers

# SUMMARY OF EDA

- Customers with lower numbers of *International Calls* churn at a higher rate than do customers with more international calls.
- For the remaining predictors, EDA uncovers no obvious association of *churn*. However, these variables are still retained for input to data mining models and techniques.
- Note the power of EDA.
  - Even without complex algorithms (like decision trees or neural networks), EDA can reveal deep insights.
  - Careful exploration helps identify key factors linked to customer churn.
  - These insights can be turned into actionable strategies to reduce churn and improve retention.