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 vsualization
  - 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.

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

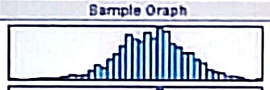
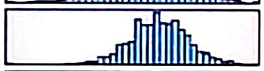
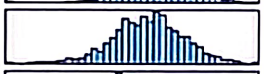
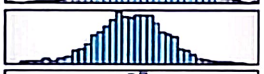
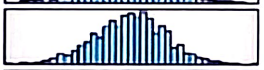
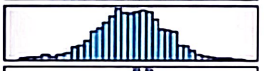
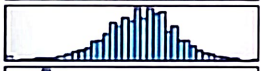
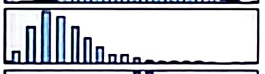
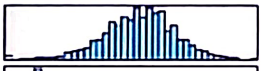
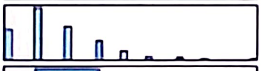
UNIVARIATE VS. MULTIVARIATE ANALYSIS

- 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

placeholder

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	169.200*	–	3333
Night Calls		Range	33	175	100.100	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

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

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

Unsupported placeholder

CHURN EXAMPLE- GETTING TO KNOW THE DATASET

Variable	Type	Description
State	Categorical	51 US states + DC
Account length	Integer	Duration of account in da
Area code	Categorical	Area classification
Phone number	Identifier	Surrogate for Customer I
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
Churn (Target)	Flag (True/False)	Customer left or stayed

Unsupported placeholder

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.

Unsupported placeholder

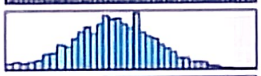
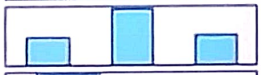
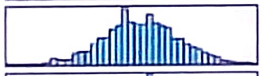
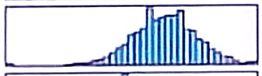
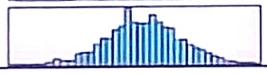
CHURN EXAMPLE- GETTING TO KNOW THE DATASET

- **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

- **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

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	0.007	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
Charge		Range	0.000	59.640	30.562	9.259	-0.029	30.530	25.180*	–	3333

Unsupported

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.

ype of Variable	**Common Techniques**	**Visualization**
gorical	Frequency counts, proportions, mode	Bar chart, pie chart
cal	Mean, median, standard deviation, skewness	Histogram, box plot, density plot

Unsaved changes Save

NIVARIATE VS. MULTIVARIATE ANALYSIS

Univa ate analysis explores **a single variable** in isolation to understand its

tion, central tendency, spread, and shape.

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.

e of Variable	Common Techniques	Visualization
orical	Frequency counts, proportions, mode	Bar chart, pie chart
al	Mean, median, standard deviation, skewness	Histogram, box plot, density plot

Unsupported placeholder

Unsaved changes Save

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

	Intl Plan = No	Intl Plan = Yes
Churn = False	2850	137
Churn = True	371	113

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

- **Interpretation:**
  - Churn rate for "Yes" = 113 / (137+113) = 42.5%
  - Churn rate for "No" = 371 / (2850+371) = 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.

Unsupported placeholder

EXPLORING CATEGORICAL VARIABLES

- **Understanding the Target Distribution**

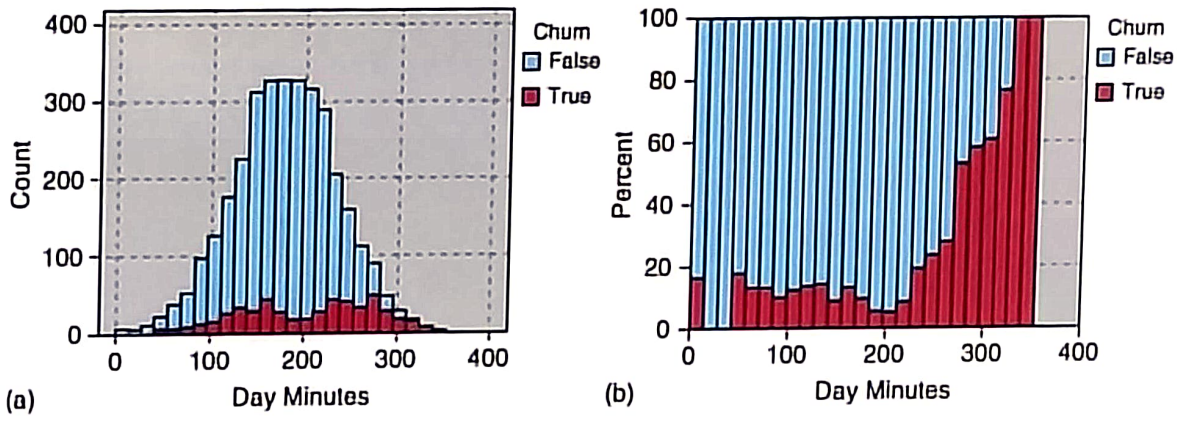
Value	Proportion	%	Count
False		85.51	2850
True		14.49	483

- 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 NUMERIC VARIABLES

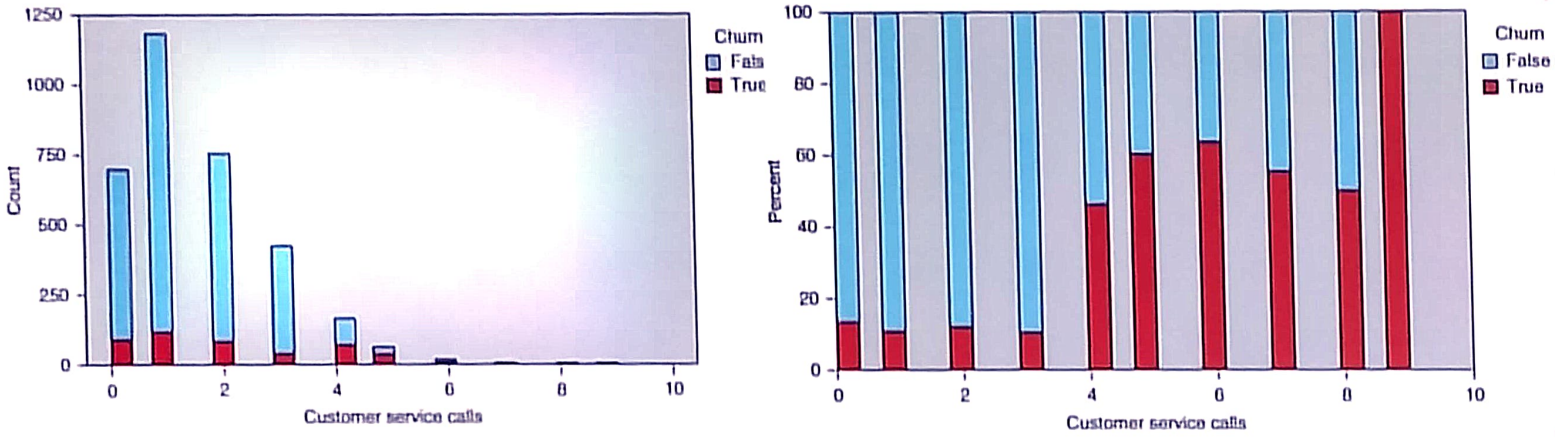
- *Day Minutes* Vs. *Churn*



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

EXPLORING NUMERIC VARIABLES

- *Customer Service Calls* With *Churn* Overlay

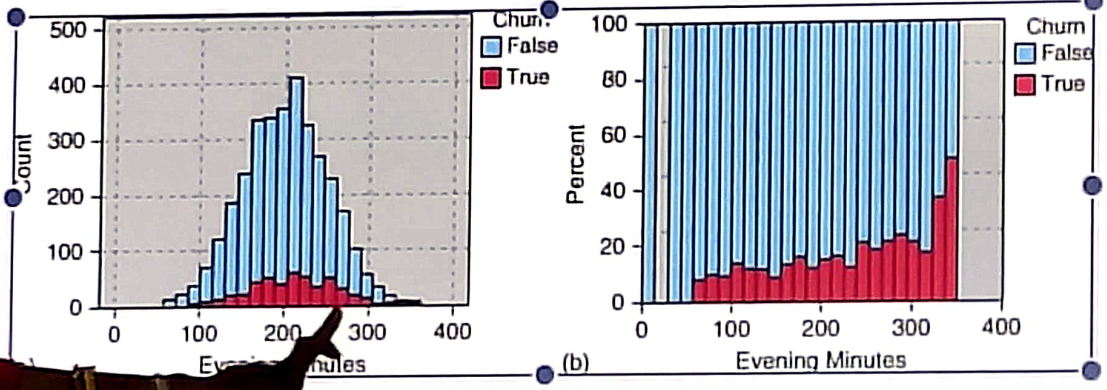


- Customers who have called customer service three times or less have a markedly lower churn rate (red part of the rectangle) than customers who have called customer service four or more times.

Unsupported placeholder

EXPLORING NUMERIC VARIABLES

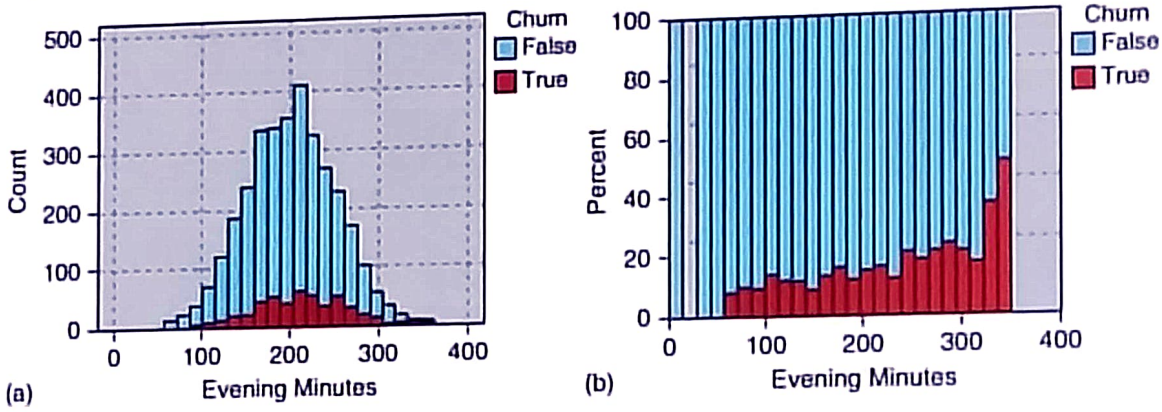
- *Evening Minutes* Vs. *Churn*



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

EXPLORING NUMERIC VARIABLES

- *Evening Minutes* Vs. *Churn*

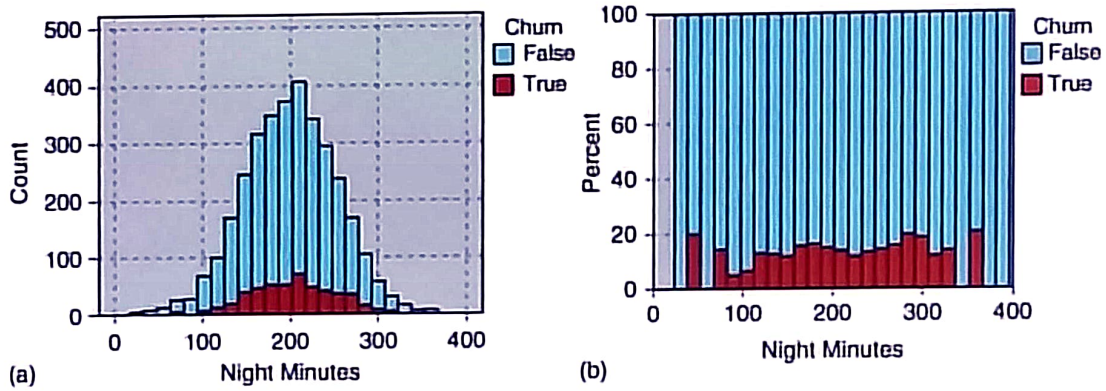


- Shows a slight tendency for customers with higher *evening minu*
to churn.

Unsupported placeholder

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.