Business Data Analytics Lecture 4: EDA & Visualization for Decision-Making

Assoc. Prof. Nguyen Binh Minh, Ph.D.

Learning Objectives (CLO)

- Master univariate, bivariate, and multivariate EDA.
- Recognize patterns, anomalies, and distributions.
- Choose the right visualization for data types and questions.
- Interpret correlations carefully; avoid causal pitfalls.
- Build reproducible, decision-oriented EDA workflows.

Agenda

- EDA fundamentals; Univariate; Bivariate; Multivariate;
- Correlation vs. causation; Visualization best practices;
- Mini-case & hands-on; Quick quiz; Wrap-up.

Running Case: Telco Churn

- 1,200 customers; plans, usage, charges, churn label.
- Goal: understand drivers and patterns before modeling.
- Synthetic data used for demonstration.

What is EDA?

- Systematic exploration of data to form/test hypotheses.
- Summarize distributions, relationships, anomalies.
- Bridge business questions and modeling choices.

EDA Mindset

- Start with questions and context; avoid aimless fishing.
- Iterate: profile → explore → hypothesize → validate.
- Triangulate with multiple views to reduce bias.

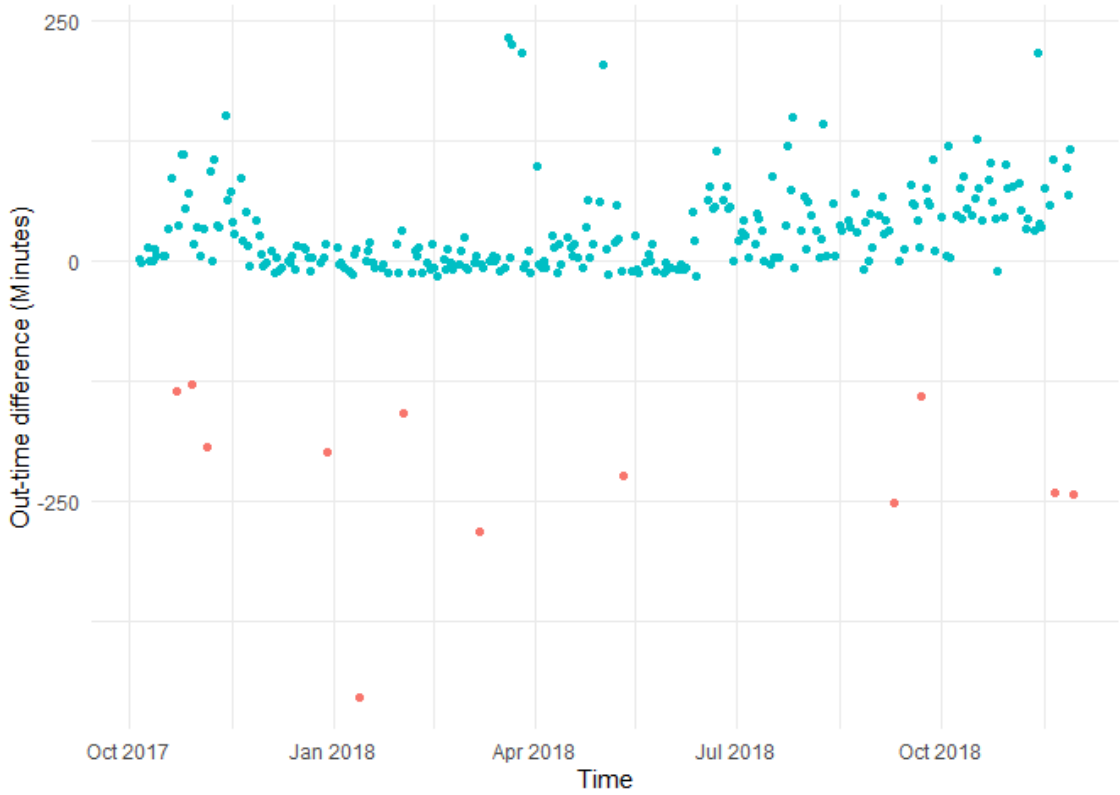
Data Types & Scales

Numeric (continuous/discrete); Categorical (nominal/ordinal).

Date-time and periodic encodings; derived features.

Univariate — Goals

- Understand shape, center, spread; spot outliers/missingness
- Inform transformations and binning choices.



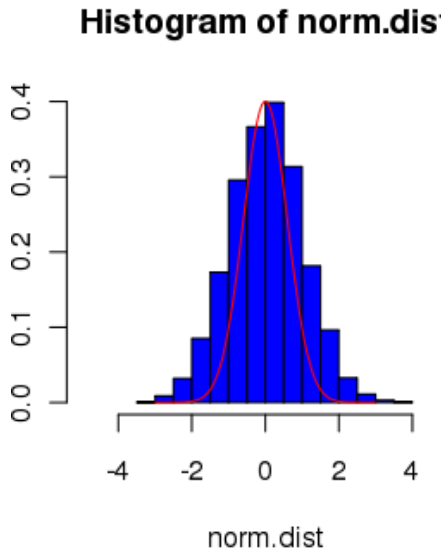
Summary Statistics

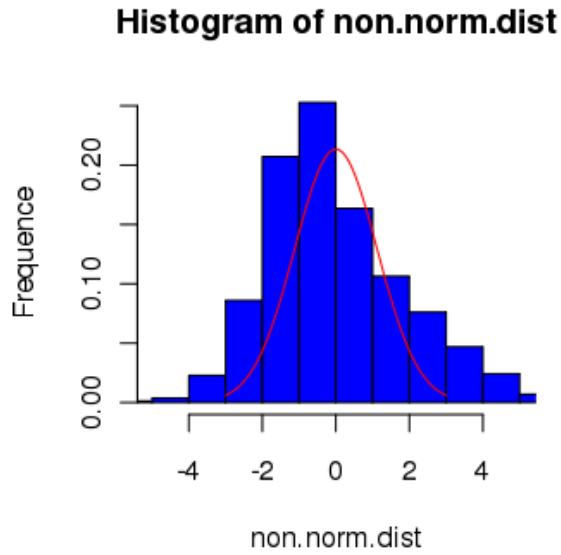
Mean, median, mode; variance, std; quantiles, IQR.

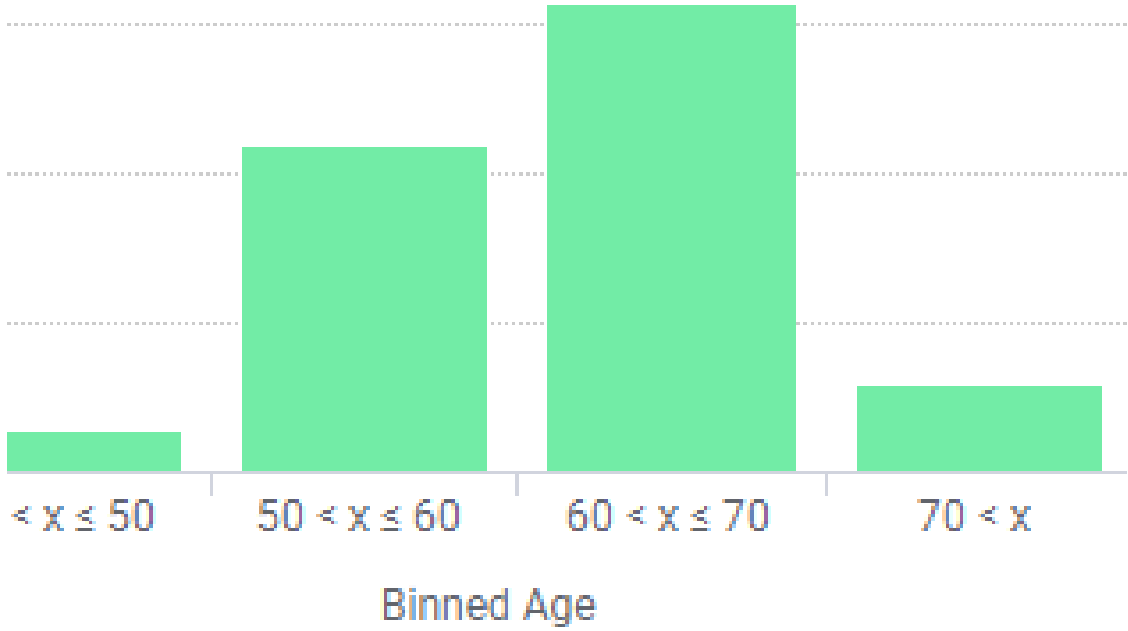
Skewness, kurtosis; robust summaries for heavy tails.

Histograms & Binning

- Bin width selection (FD rule, sqrt rule).
- Log/Box-Cox/Yeo-Johnson for skewed distributions.

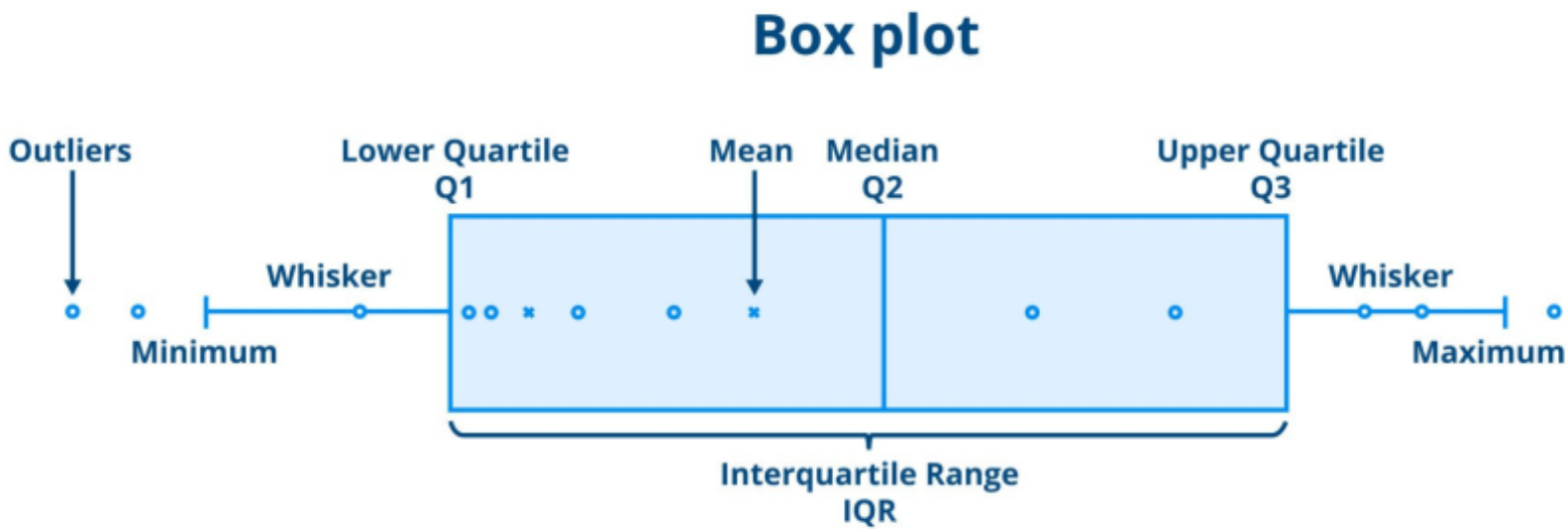






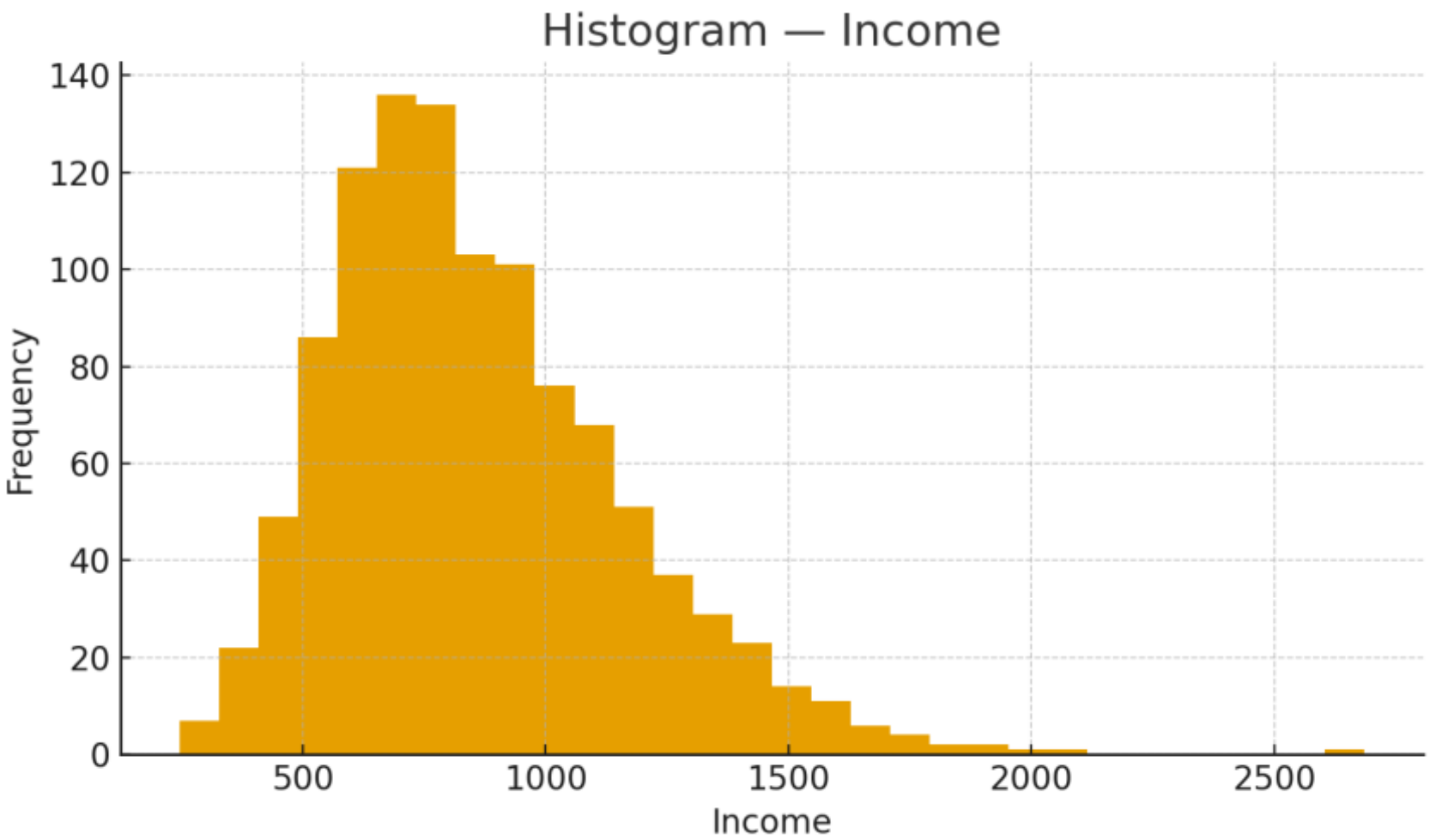
Boxplots & Distribution Views

- Boxplots show spread/outliers; violin (concept) shows density.
- Avoid misleading axis scales and overplotting.



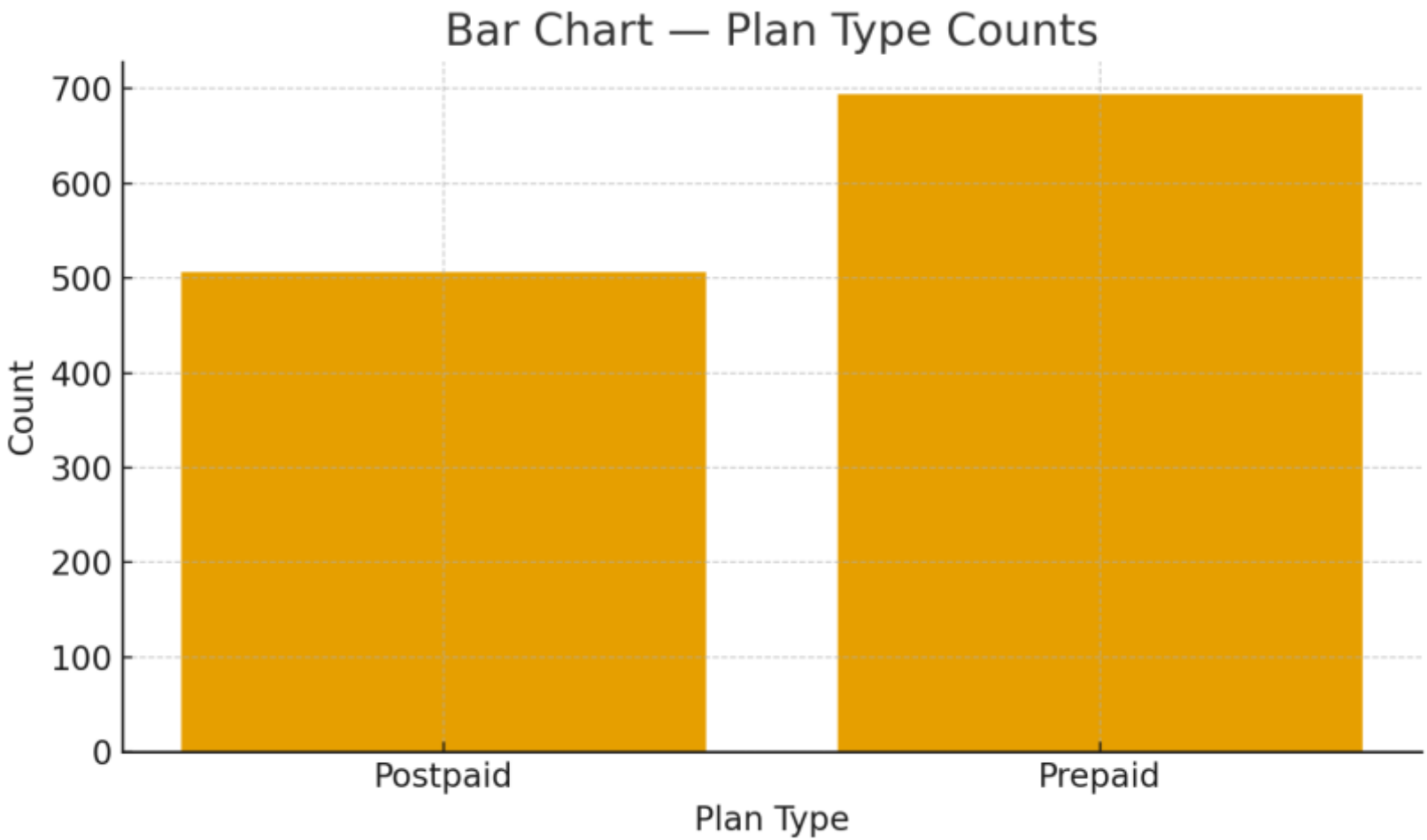
Univariate Visual — Histogram (Income)

Distribution of Income



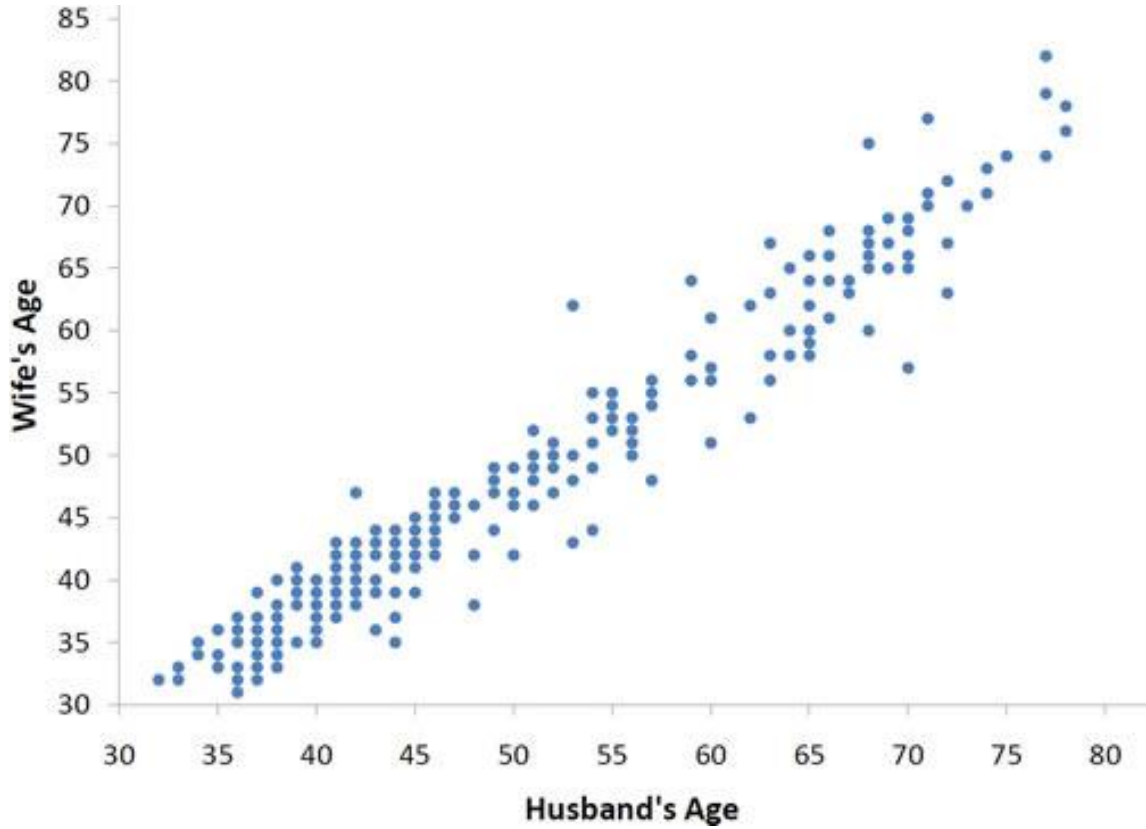
Univariate Visual — Bar (Plan Type)

Counts by Plan Type



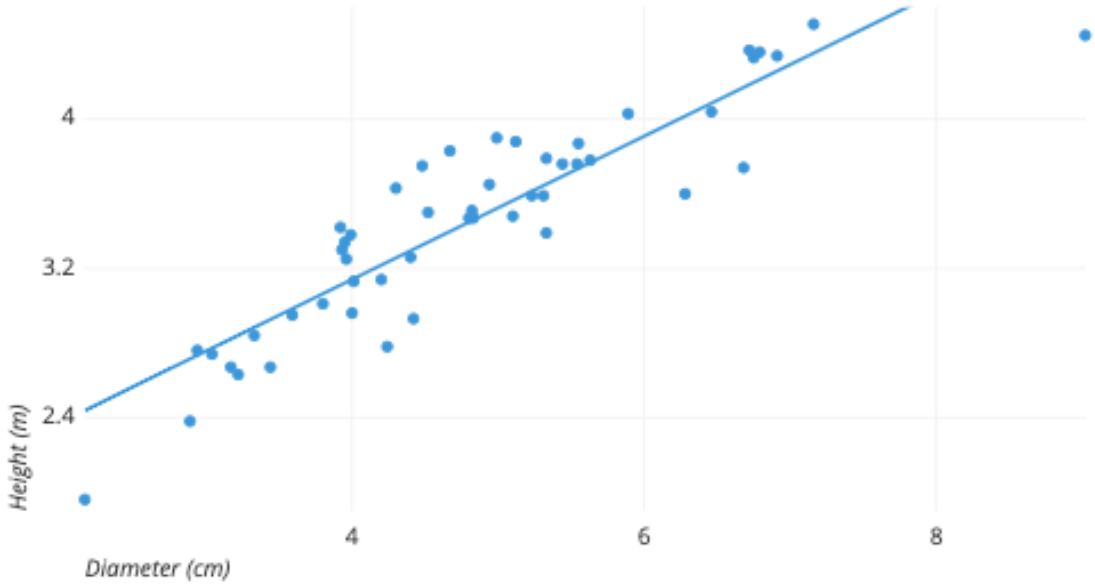
Bivariate — Goals

- Quantify pairwise relationships.
- Compare distributions across categories.
- Guide transformations and interaction terms.



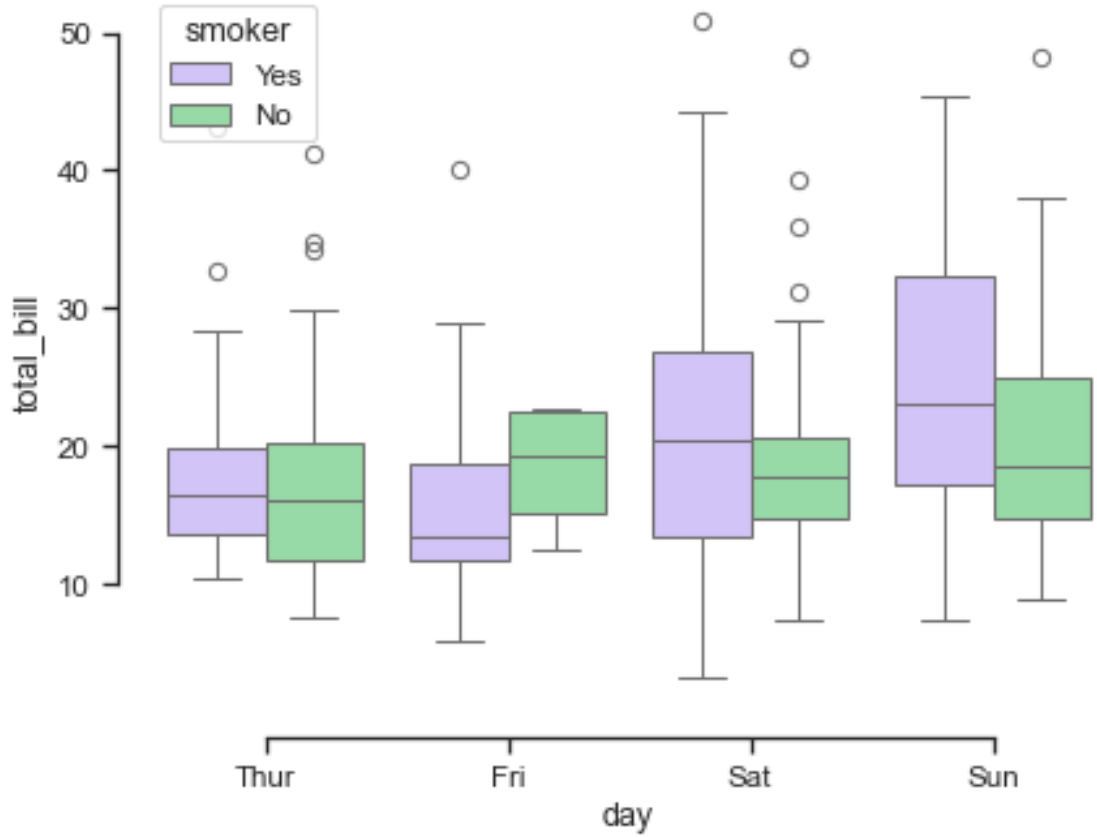
Numeric–Numeric

- Scatter plots; Pearson vs. Spearman correlations.
- Look for nonlinearity and heteroscedasticity.



Numeric–Categorical

- Grouped boxplots; strip/dot plots for raw observations.
- Aggregate summaries with confidence intervals.

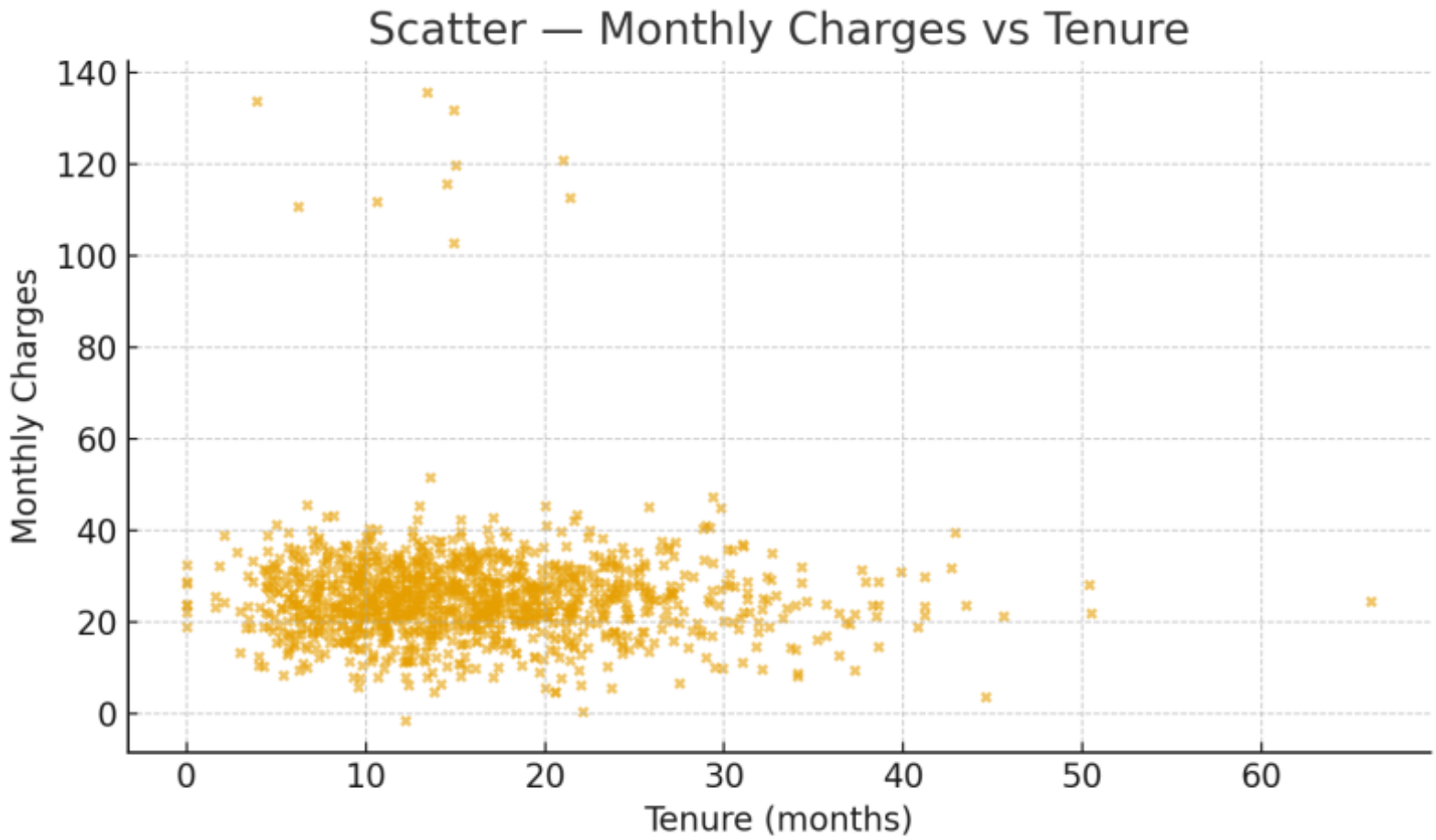


Categorical–Categorical

	Dog	Cat	Total
Male	42	10	52
Female	9	39	48
Total	51	49	100

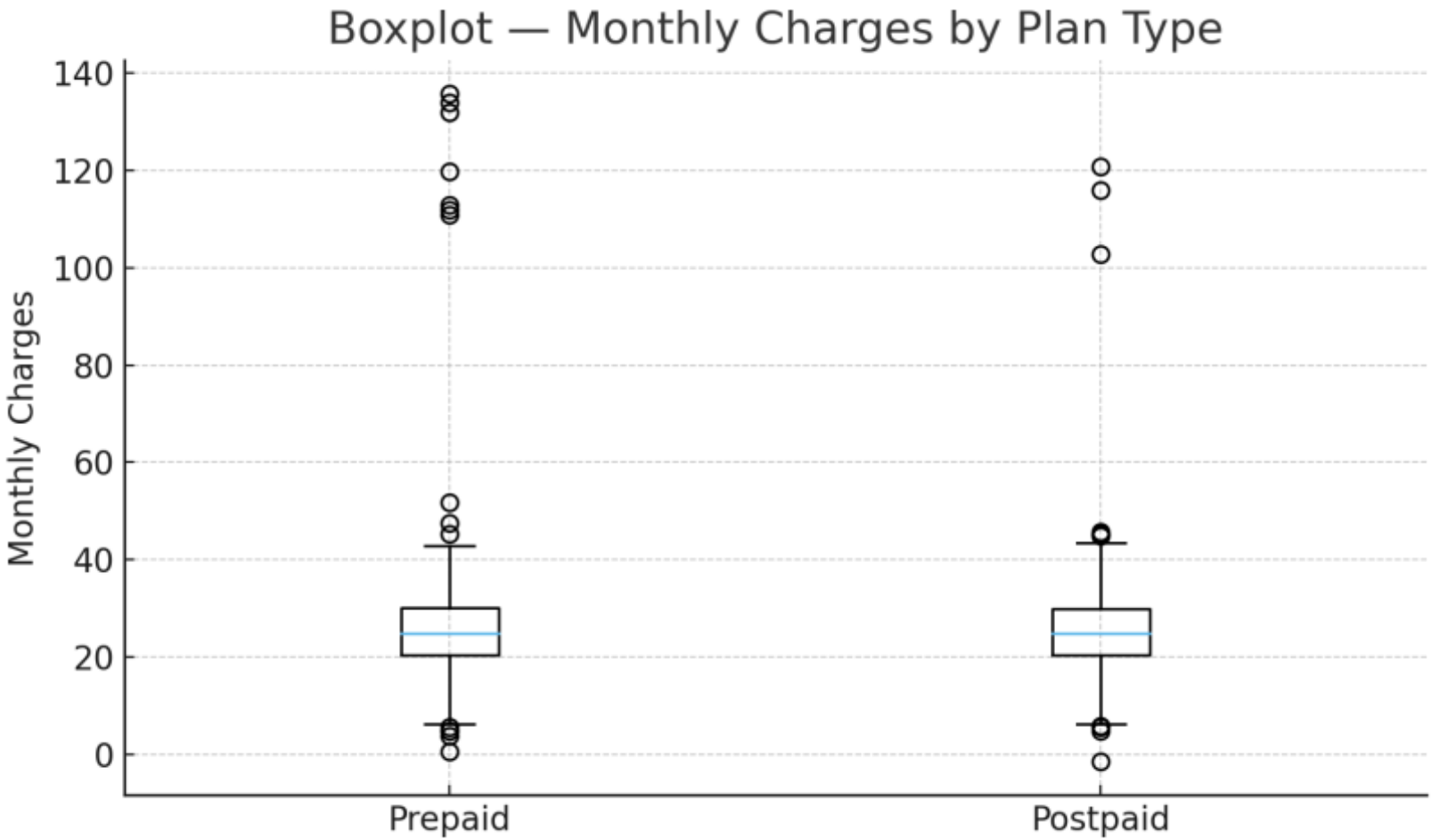
- Contingency tables; stacked/grouped bars.
- Chi-square test overview; expected vs observed.

Bivariate Visual — Scatter (Charges vs Tenure)



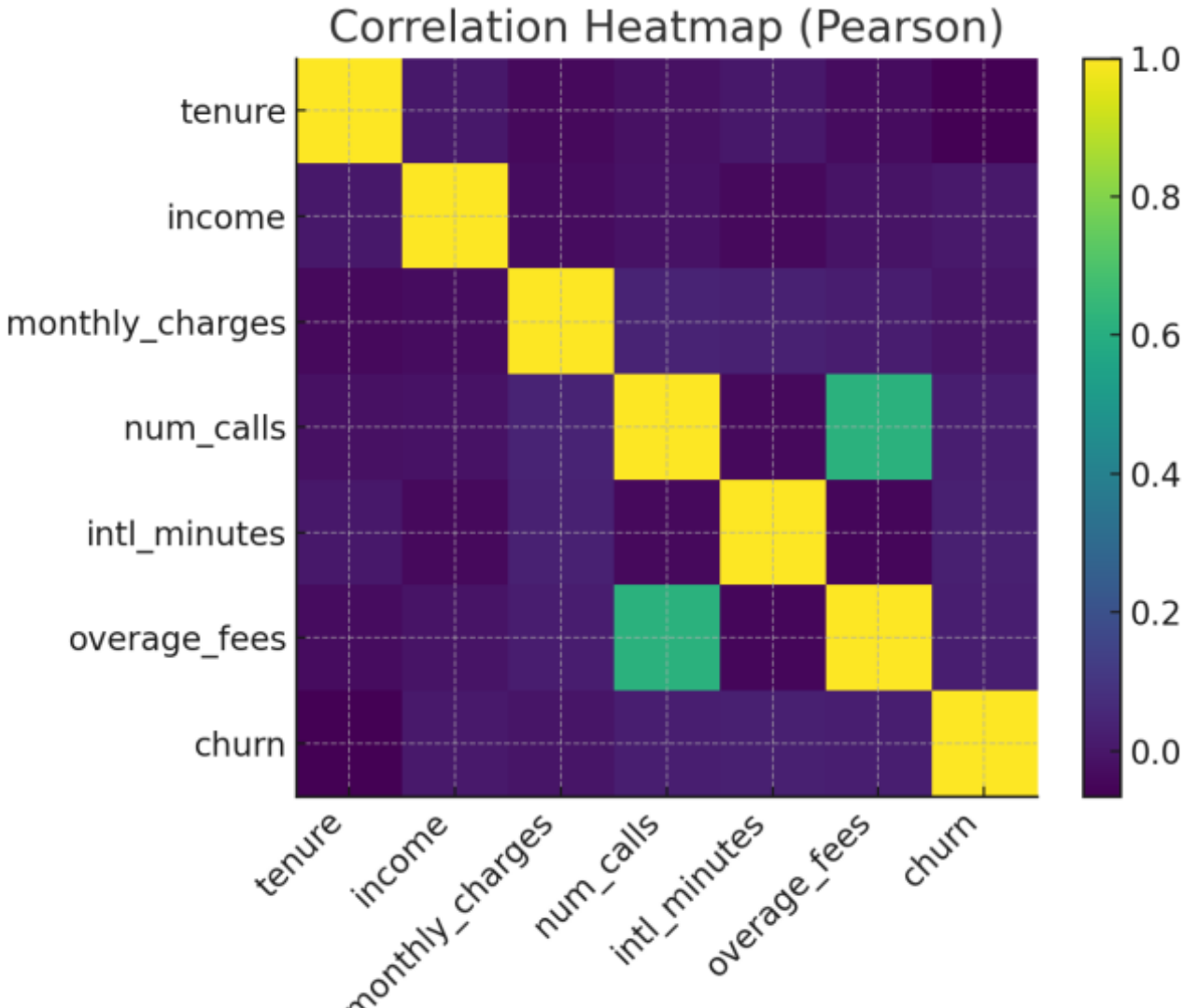
Nonlinear spread and heteroscedasticity

Bivariate Visual — Boxplot (Charges by Plan)



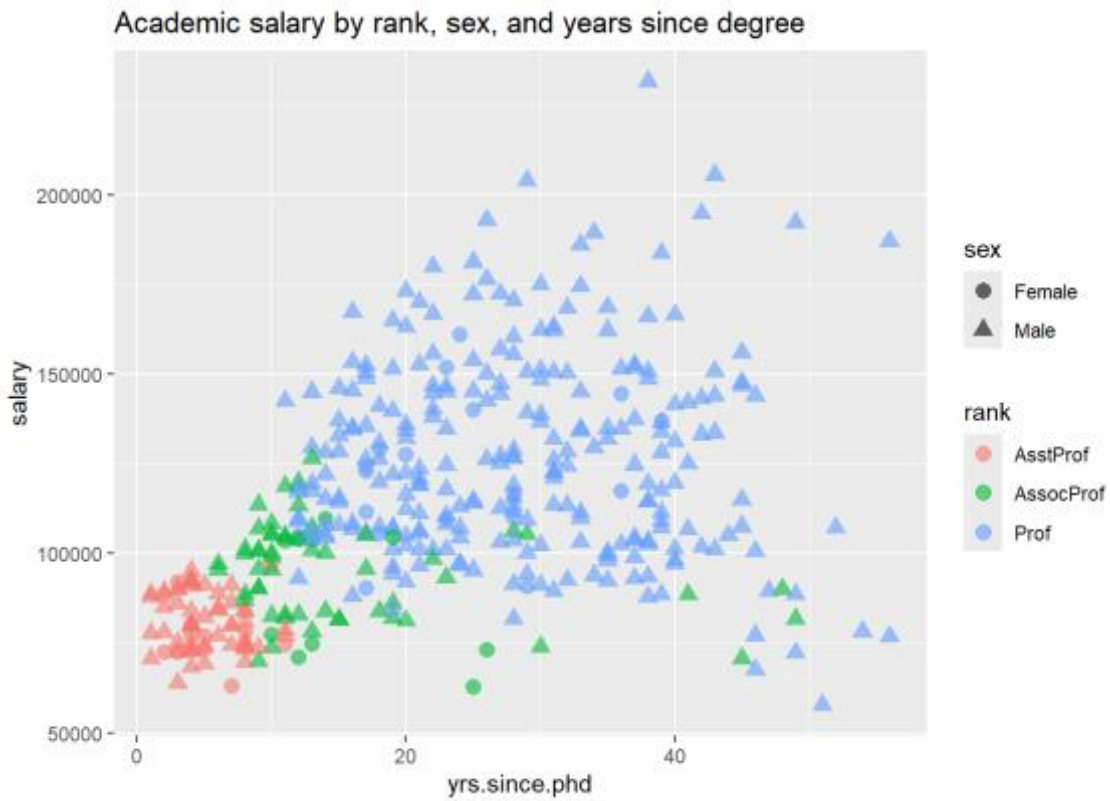
Group comparison by plan type

Correlation Heatmap — Numeric Features



Pearson correlations

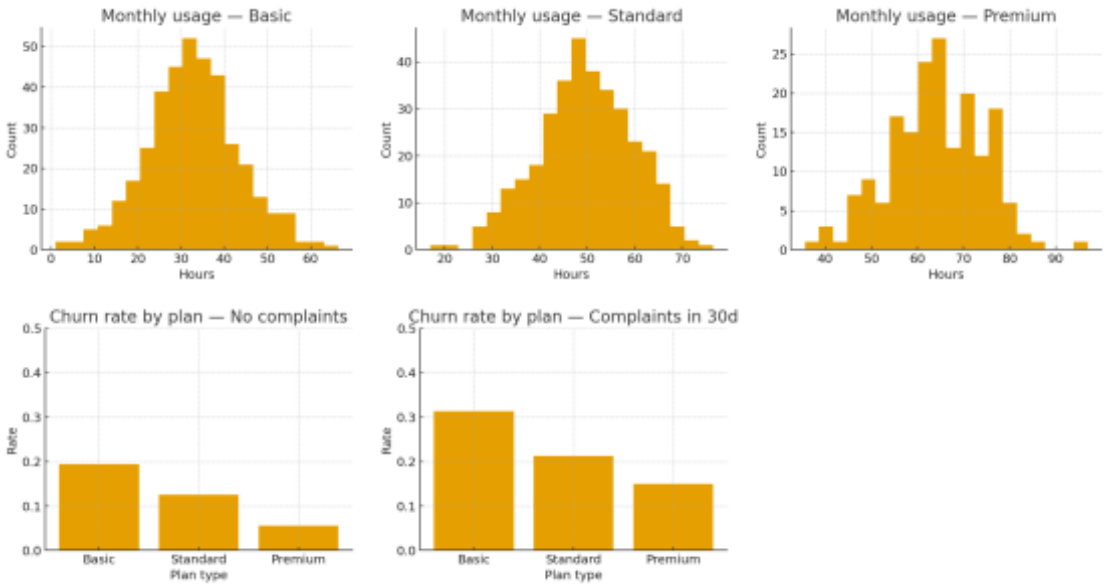
Multivariate — Goals



- Reveal joint structures and interactions beyond pairs.
- Identify segments and conditional patterns.

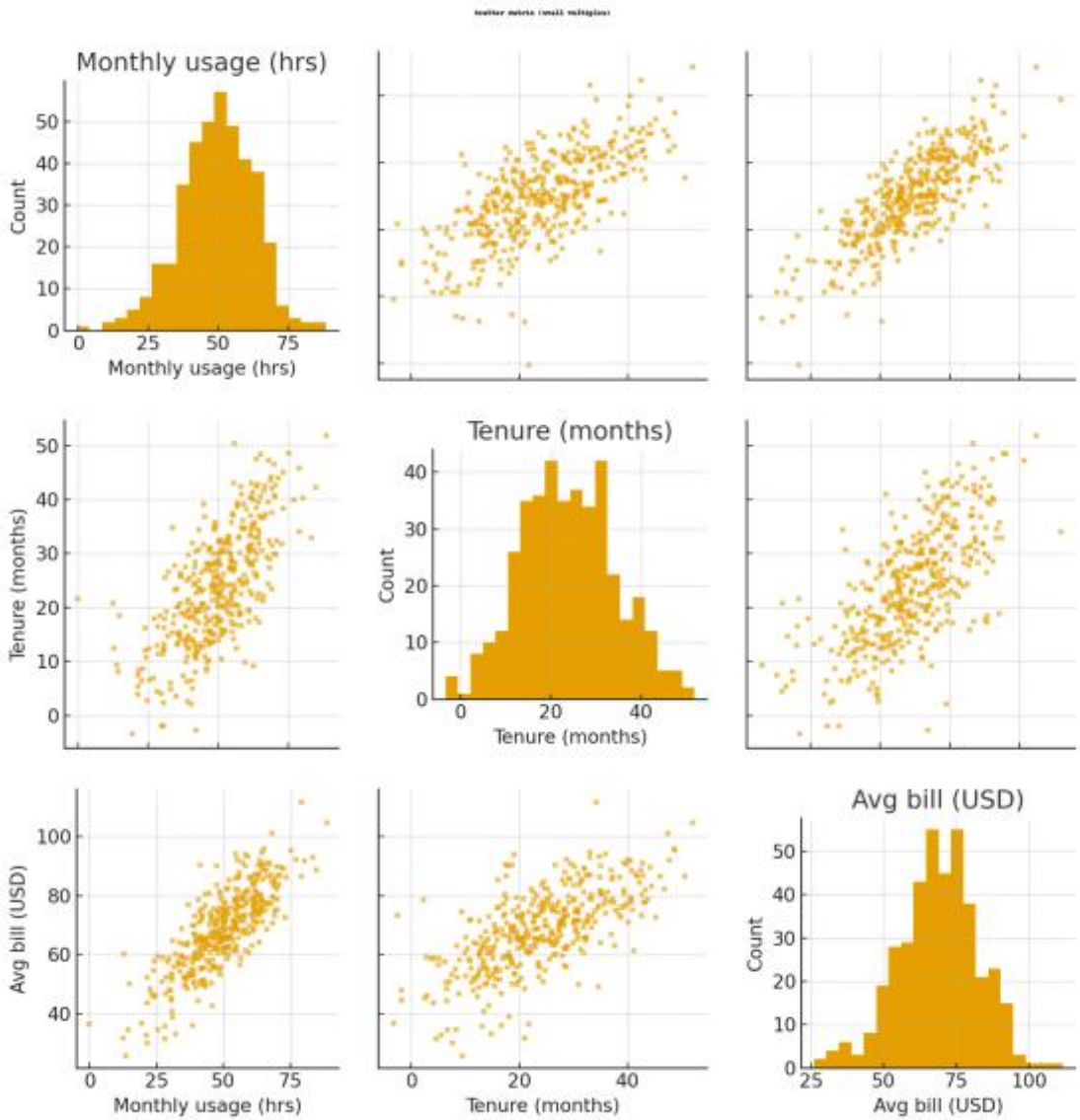
Segmented Views & Faceting

- Split by plan_type or complaints_30d.
- Use small multiples to compare segments.



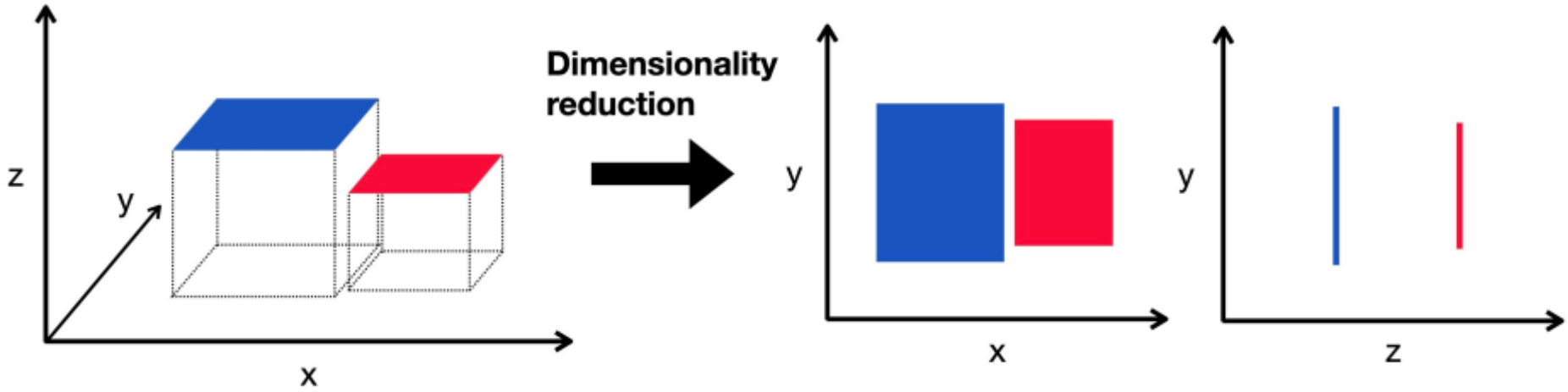
Scatter Matrix (concept)

- Pairwise scatter grid; use alpha/jitter to reduce overplotting.
- Sanity check before modeling.



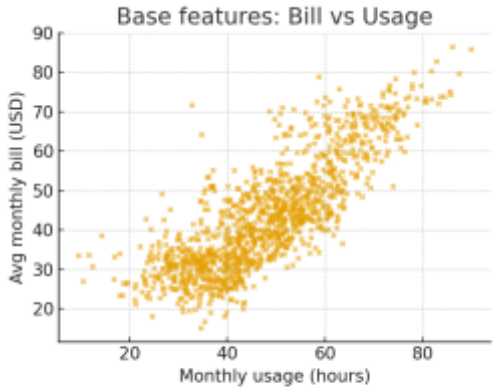
Dimensionality Reduction (EDA lens)

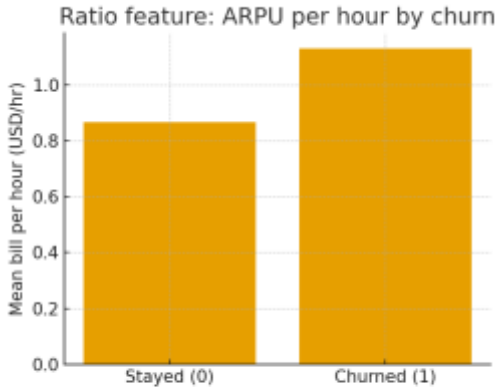
PCA for linear structure; t-SNE/UMAP for visualization only.

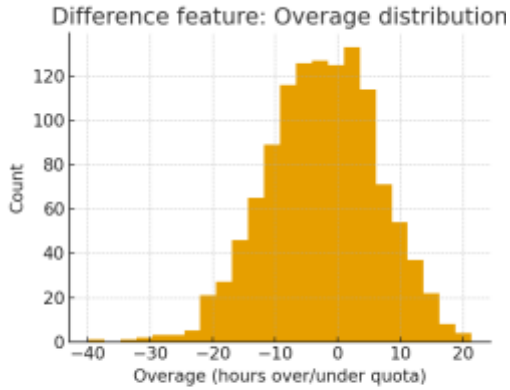


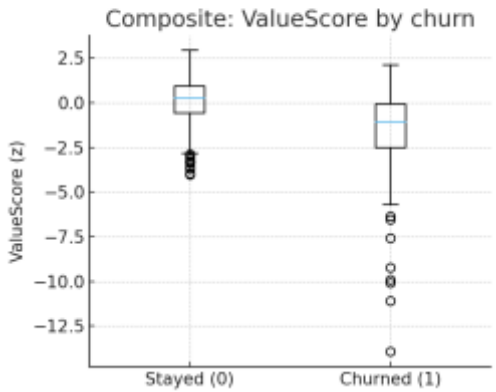
Feature Interactions (EDA cues)

- Ratios/differences; domain-informed composites.
- Record hypotheses for later tests.



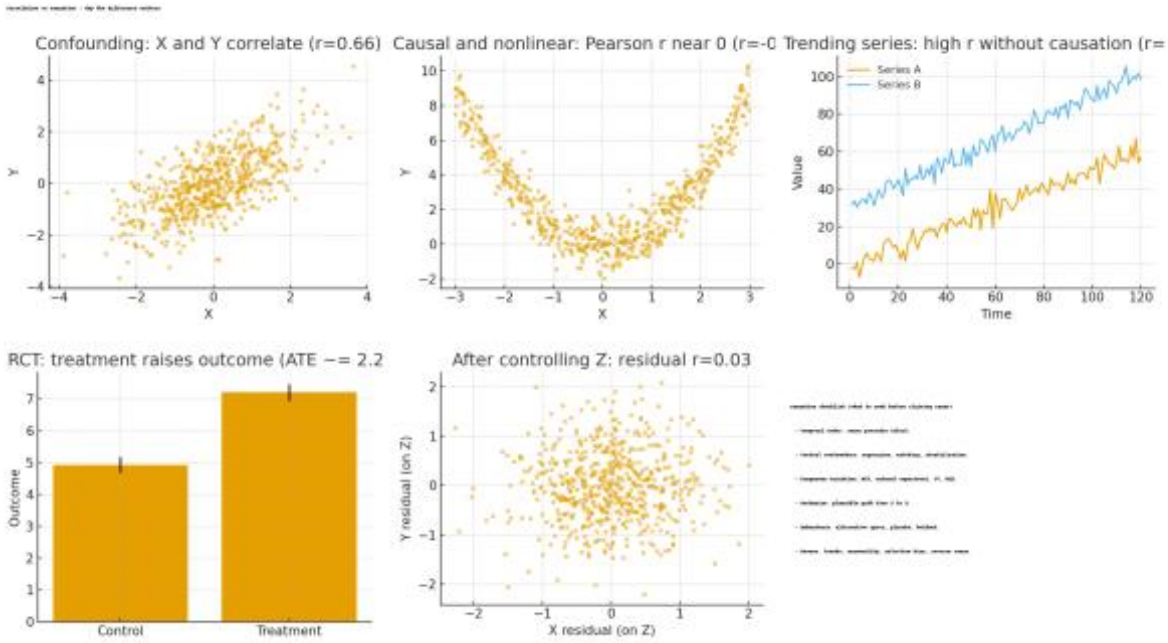






Correlation vs. Causation — Importance

- Decisions require causal reasoning.
- Spurious correlations are common; EDA suggests, tests confirm.



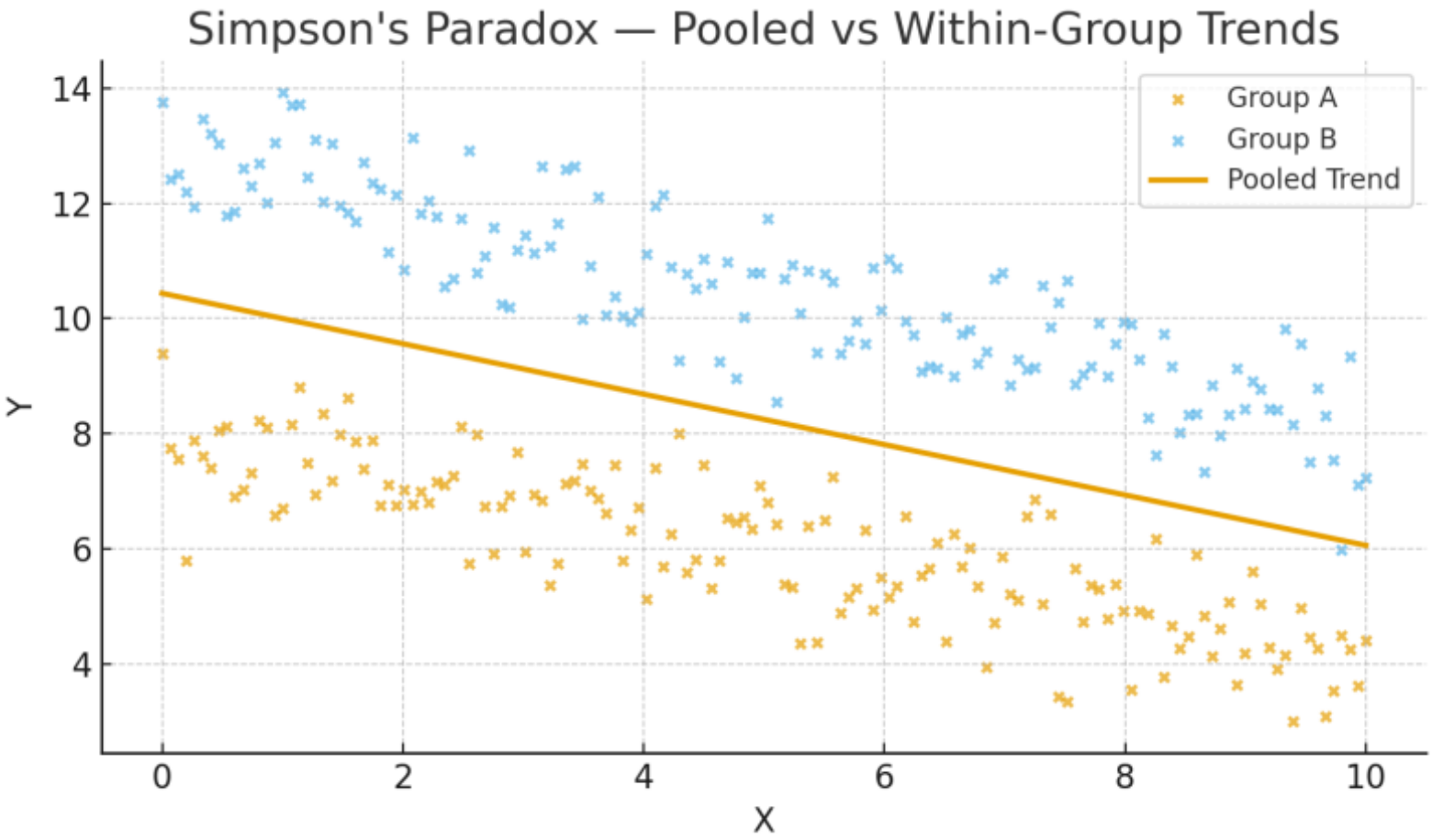
Confounding & Colliders (intuition)

- Confounder affects both X and Y; collider distorts when conditioned.
- Use stratification and DAG-thinking.

Simpson’s Paradox (concept)

- Pooled trend differs from within-group trends.
- Always check stratified views.

Simpson’s Paradox — Visual Demo



Pooled vs within-group trends can disagree

Observational vs Experimental

- Randomized experiments identify causal effects.
- Quasi-experiments: DID, IV, RDD (high-level).

Practical A/B Testing

- Define metrics (primary & guardrails), sample size, duration.
- Randomization and bucketing; avoid peeking.

From EDA to Decisions

- Move from patterns to testable hypotheses.
- Quantify uncertainty; document assumptions.

Visualization Principles — Marks & Channels

Position/length are most precise; avoid misleading area/volume.

Clear titles, labels, and context.

Choosing the Right Chart

- Distribution: histogram/boxplot; Relationship: scatter/line;
- Composition: stacked bars; avoid 3D pies.

Scales & Axes

Zero baseline for bars; log scales for heavy tails;

Avoid dual y-axes unless necessary.

Binning & Smoothing

Tradeoff between noise and detail; keep bins consistent.

Show raw points with jitter/alpha when possible.

Small Multiples & Faceting

- Compare patterns across groups/time with repeated layouts.

Avoiding Chartjunk

- Remove non-data ink; limit encodings per chart;
- Use annotations sparingly to guide the story.

Dashboards for Decision-Making

- Tie visuals to questions and actions; highlight targets/thresholds.

EDA Workflow & Reproducibility

- Notebook scripts; versioned data; saved figures/tables.
- Share EDA artifacts for review.

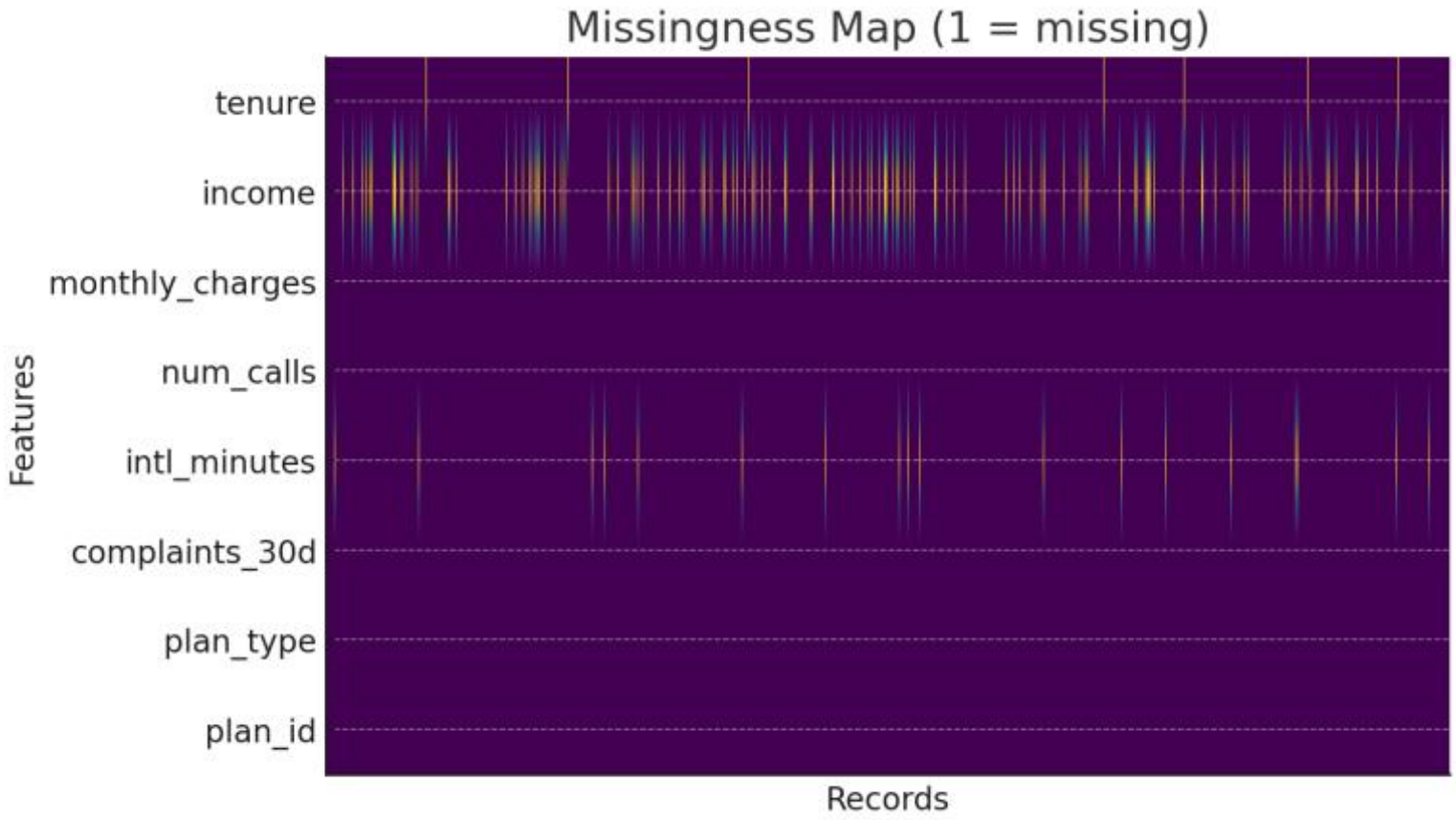
EDA Checklist

- Types/ranges/missingness/duplicates;
- Univariate, bivariate, multivariate patterns;
- Stratify by key segments; log insights.

Common Anti-Patterns

- Cherry-picking; overplotting;
- Causal claims from pure correlation.

Missingness Map (1 = missing)



Diagnose mechanisms before imputation

Hands-on Tasks

- Recreate provided visuals; add one new plot per section.
- Write a one-page executive EDA summary.

Quick Quiz (10)

- Name two ways histograms can mislead.
- Explain Simpson’s paradox in your own words.
- Suggest a visualization for numeric–categorical comparison.
- Why can correlation heatmaps be misinterpreted?
- When should you use a log scale?

Key Takeaways

- EDA sharpens questions; visuals should inform decisions.
- Stratify and consider confounding; don’t infer causality blindly.

Recommended References

- Anscombe (1973); Tufte; Cleveland;
- scikit-learn & matplotlib documentation.