Assignments Overview

Goal

For each type of visualization introduced in the slides, you will:

- Implement the plot on the Titanic training data.
- Save the figure with a clear filename.
- Write a short interpretation (2–4 bullet points) answering:
  - What pattern does the plot show?
  - How is this pattern useful for modeling or feature engineering?
- Include all plots and interpretations in a short report (PDF or notebook).

Assignment 1: Univariate Visualizations

Histograms, Boxplots, Countplots, KDE

Using the Titanic training data:

1. **Histogram (Age)**
   - Create a histogram of `Age` (e.g., using `sns.histplot`).
   - Try at least two different bin sizes (e.g., 10 and 30).
   - Compare how the shape changes and which bin size you prefer and why.
2. **Histogram or KDE (Fare)**
   - Plot the distribution of `Fare` with a histogram, optionally with KDE.
   - Comment on skewness and outliers; suggest a possible transformation (e.g., log).
3. **Boxplot (Fare)**
   - Produce a boxplot of `Fare`.
   - Identify approximate median and presence of extreme outliers.
4. **Countplot (Sex, Pclass)**
   - Create countplots for `Sex` and `Pclass`.
   - Discuss class imbalance and how it might affect survival analysis.

Assignment 1 (cont.): Univariate Visualizations

Refinements

5. **KDE Plot (Age)**
   - Plot a KDE of `Age`.
   - Compare it to the histogram from Task 1: what does KDE reveal that the histogram does not?
6. **Summary Paragraph**
   - Write a short paragraph (5–7 sentences) summarizing what you learned about the marginal distributions of Age, Fare, Sex, and Pclass.
   - Explicitly state which variable(s) you expect to be most predictive of survival and why.

Assignment 2: Bivariate Visualizations

Target-Oriented Analysis

Using `Survived` as the target:

1. **Barplot: Survival Rate by Sex**
   - Plot survival rate by `Sex` using a barplot (`sns.barplot`).
   - Quantify the difference in survival rate between males and females (approximate percentages).
2. **Barplot or Pointplot: Survival Rate by Pclass**
   - Plot survival rate vs. `Pclass` (`sns.pointplot` or `sns.barplot`).
   - Comment on how survival probability changes with passenger class.
3. **Violinplot or Boxplot: Age by Survival**
   - Plot `Age` (y-axis) versus `Survived` (x-axis) using a violinplot or boxplot.
   - Discuss differences in the distributions of age for survivors vs. non-survivors (e.g., medians, spread, presence of children).

Assignment 2 (cont.): Bivariate Visualizations

Continuous Interactions

4. **Scatterplot: Age vs. Fare (colored by Survived)**
   - Create a scatterplot of `Age` vs. `Fare`, with color (`hue`) indicating `Survived`.
   - Optionally, adjust point transparency (`alpha`) for dense regions.
   - Identify any visible cluster(s) associated with higher survival rates.
5. **Crosstab and Heatmap: Pclass vs. Survived**
   - Create a crosstab of `Pclass` and `Survived`.
   - Visualize it with a heatmap.
   - Interpret which combinations of class and survival are most/least frequent.
6. **Short Reflection**
   - Write 3–5 bullet points summarizing the strongest bivariate relationships with survival (e.g., Sex, Pclass, Fare, Age).

Assignment 3: Multivariate Visualizations

Heatmaps, Facets, Pairplots, Catplots

1. **Correlation Heatmap**
   - Compute a correlation matrix for numeric features (e.g., Survived, Pclass, Age, SibSp, Parch, Fare).
   - Plot it using a heatmap with annotations.
   - Identify the top two positive and top two negative correlations involving `Survived`.
2. **FacetGrid: Age by Sex and Survival**
   - Use a facet grid (e.g., `sns.FacetGrid`) to plot Age histograms conditioned on `Sex` (rows) and `Survived` (columns).
   - Describe how the Age distribution differs across the four panels (male/female vs. survived/died).
3. **Pairplot**
   - Create a pairplot for a subset of variables (e.g., Survived, Pclass, Age, Fare).
   - Use `hue='Survived'` to color points.
   - Identify any pair of variables where survivors and non-survivors appear well separated.

Assignment 3 (cont.): Multivariate Visualizations

Conditioned Categorical Effects

4. **Catplot: Survival by Pclass and Sex**
   - Create a categorical plot (e.g., `sns.catplot` with `kind='bar'`) showing survival rate by `Pclass` and `Sex` (as hue).
   - Discuss which subgroup (e.g., 1st class females) has the highest survival rate and which has the lowest.
5. **Design Your Own Multivariate Plot**
   - Propose and implement one additional multivariate visualization not directly shown in the slides (e.g., a facet scatterplot, stacked bar chart, or jointplot).
   - Explain the design choices (axes, hues, facets) and what new insight it provides beyond previous plots.

Assignment 4: Evaluation Visualizations

Model Performance and Comparison

Assume you have trained at least two models (e.g., Logistic Regression and Random Forest).

1. **Model Score Barplot**
   - Create a small DataFrame with model names and their cross-validation scores.
   - Plot a barplot comparing model scores.
   - Comment on which model you would choose and why.
2. **Confusion Matrix**
   - For one chosen model, compute and plot a confusion matrix on the training or validation set.
   - Interpret each cell (TP, TN, FP, FN) in the context of Titanic survival.
   - Discuss the trade-off between correctly predicting survivors vs. non-survivors.

Assignment 4 (cont.): Evaluation Visualizations

Advanced Evaluation (Optional Bonus)

3. **ROC Curve (Bonus)**
   - Plot the ROC curve for one model using predicted probabilities.
   - Compute and report the AUC.
   - Briefly discuss how the ROC curve complements accuracy and confusion matrix.
4. **Feature Importance Plot (Bonus)**
   - For a tree-based model (e.g., Random Forest), plot feature importances as a horizontal bar chart.
   - Compare the visual importance ranking with the relationships you observed in your earlier visualizations.