ChatGPT

Expert-Level Data Visualization Q&A Guide

Author: Senior Data Scientist | 10+ Years of Experience | 150+ Interviews Conducted

---

Section 1: Advanced Interview Questions and Answers

Q1. What is the best visualization technique for detecting outliers in numerical data?

Answer: - **Box Plot** is most commonly used to identify outliers using IQR (Interquartile Range). - **Violin Plot** is preferred when you want both distribution and outliers. - **Z-score and scatter plots** can be used for multidimensional outlier detection. - **Parameters:** - IQR multiplier (commonly 1.5) - Color coding for outliers

Q2. How would you visualize and compare the distribution of a numerical variable across multiple categories?

Answer: - Use **Box Plots** for clean comparison and spotting medians/outliers. - Use **Violin Plots** to see KDE (Kernel Density Estimation). - Use **FacetGrid (Seaborn)** for comparison across subplots. - **Scenario:** Analyzing salaries across departments.

Q3. What kind of plots are suitable for mixed data types in bivariate/multivariate analysis?

Answer: - **Categorical vs Numerical:** Box Plot, Violin Plot, Swarm Plot - **Numerical vs Numerical:** Scatter Plot, Hexbin Plot (for dense data), Bubble Plot - **Categorical vs Categorical:** Grouped Bar Charts, Mosaic Plots, Heatmaps - **Mixed Case:** Use **Pair Plot** with `hue` on a categorical target.

Q4. How do you visualize relationships when the data is heavily overlapping in scatter plots?

Answer: - Use **Hexbin Plots** or **2D KDE plots** to show density. - Use **Transparency (alpha)** to reduce clutter. - Consider **jittering** for categorical variables. - **Alternative: Strip Plot** with jitter or **Swarm Plot**.

Q5. How would you choose between a Pair Plot and a Correlation Heatmap?

Answer: - **Pair Plot:** Good for smaller datasets, gives actual scatter comparisons, but not scalable. - **Correlation Heatmap:** Efficient for summarizing large numerical datasets quickly. - Use **Correlation Coefficients (Pearson, Spearman)**.

Section 2: Scenario-Based Expert Q&A

Q6. You are analyzing customer purchase behavior with 10 numerical features. How do you summarize the relationships?

Answer: - Start with a **Correlation Matrix** and **Heatmap**. - Use **Pair Plots** to explore non-linear relationships. - If dimensionality is too high, use **PCA** for visualization. - **Plotly Scatter Matrix** for interactivity in dashboards.

Q7. Your dataset includes date/time. What plots would you use to find trends or seasonality?

Answer: - **Line Plots** for time series. - **Seasonal Decompose Plot (statsmodels)** to break into trend/seasonality/residual. - **Heatmaps** with weekdays vs hour for hourly trend. - **Box Plots** grouped by month to compare seasonal distributions.

Q8. How would you visualize a classification model's feature impact?

Answer: - **SHAP Summary Plot** or **Partial Dependence Plots (PDP)**. - **Bar Plot of feature importances** from model output. - Use **Clustered Box Plots** for each class.

Q9. Which plots help in detecting multicollinearity?

Answer: - **Correlation Heatmap** to identify high-correlation pairs. - **Variance Inflation Factor (VIF)** values visualized in bar plots. - **Condition Index Plot** for multi-collinearity issues.

Q10. What visualization would you use for anomaly detection in time series?

Answer: - **Rolling Mean and Standard Deviation plots**. - **Control Charts** (used in industrial analytics). - **Time-series decomposition plots** with confidence intervals. - **Isolation Forest/LOF score overlaid** on time-series for visual insight.

Section 3: Specialized Plots and Contexts

Plot Type	Use Case	Parameters/Controls
Radar Chart	Compare multiple KPIs per category	Axis scaling, normalization
Treemap	Visualizing proportions in hierarchical data	Color/size encoding, hierarchy depth
Sunburst Plot	Interactive tree visualizations	Drill-down enabled
Word Cloud	Textual data visualization	Stopwords, frequency cutoff
Dendrogram	Clustering visualization	Linkage method, distance metric
Parallel Coordinates	Multivariate data profiling and class separation	Z-score scaling, axis ordering

Plot Type	Use Case	Parameters/Controls
Facet Grid	Categorical slicing for large multivariate datasets	Rows/cols mapping, plot types
Ternary Plot	Visualizing compositional data (3-part constraints)	Sum-to-one normalization
JointGrid	Combine scatter plot with marginal histograms and KDE	Set marginal and joint axes

Section 4: Visualization Parameters Often Asked in Interviews

Parameter/Setting	Relevance
Color Palette	Colorblind-friendly, aesthetic grouping, branding alignment
Bins (Histogram)	Affects skewness perception, ideal via Freedman–Diaconis rule
Alpha (Transparency)	Useful for overlap, data density clarity
Figure Size	Needed for presentation scaling, subplots readability
Hue / Style / Size	Used for encoding additional variables (Seaborn)
Log Scale	Handle skewed data, especially in income, population, etc.
Annotations	Value labeling, decision explanation in plots
Axis Formatting	Currency, percentages, log formatting

Section 5: Commonly Asked Plot Comparison Questions

Q11. When should you use a Violin Plot instead of a Box Plot?

Answer: - Use **Violin Plot** when you want to understand the **distribution shape** (KDE) in addition to summary statistics. - **Box Plot** is more straightforward and less detailed, better for simpler insights. - Violin plots are useful when comparing multiple groups and you want to see modality or skewness.

Q12. FacetGrid vs JointGrid – When to Use Each?

Answer: - **FacetGrid** is best for slicing the dataset across multiple **subsets of a categorical variable**. - Example: Plotting the same distribution across different genders. - **JointGrid** is ideal for **bivariate analysis** with **marginal histograms or KDE**. - Example: Understanding relationship between height and weight.

Q13. Heatmap vs Cluster Map – Key Differences?

Answer: - **Heatmap** shows value intensities (correlations, frequencies) in matrix form. - **Clustermap** adds **hierarchical clustering** on rows/columns to group similar patterns. - Clustermap is useful in gene expression, customer segmentation.

Q14. Swarm Plot vs Strip Plot – Which one to use?

Answer: - **Swarm Plot** adjusts points to avoid overlap – better for smaller datasets. - **Strip Plot** can be jittered but might overlap – suitable for larger samples. - Swarm is visually cleaner but more computationally expensive.

Q15. When should you use a Pair Plot, and what are its limitations?

Answer: - **Pair Plot** gives a matrix of bivariate relationships and histograms. - Great for **initial EDA** but **inefficient for large feature sets** (too many subplots). - Use `hue` for classification insights.

Would you like this content enriched with example plots and Python code using Seaborn, Matplotlib, or Plotly?