Cheat Sheet for Data Analysis #5Scatter Plots with Categorical Coloring

Basic Setup and Data Loading

Import essential libraries

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Meaning: Imports necessary libraries for data analysis and visualization

Load a CSV file into a DataFrame

```
df = pd.read_csv("Diamond.csv")
```

Meaning: Reads data from a CSV file and stores it in a variable called `df`

Basic Scatter Plot (No Color)

Create a simple scatter plot

```
plt.figure(figsize=(8, 6))
plt.scatter(df['carat'], df['price'], alpha=0.7)
plt.title('Scatter Plot: Carat vs Price (No Categorical Coloring)')
plt.xlabel('Carat')
plt.ylabel('Price (Singapore$)')
plt.show()
```

Meaning: Shows the relationship between two numerical variables; serves as a baseline before adding categorical coloring

---

Scatter Plots with Categorical Coloring

Color by certification (Nominal variable)

```
plt.figure(figsize=(8, 6))
sns.scatterplot(
    data=df,
    x='carat',
    y='price',
    hue='certification',  # Color points by certification body
    alpha=0.8
)
plt.title('Scatter Plot: Carat vs Price Colored by Certification')
```

```
plt.xlabel('Carat')
plt.ylabel('Price (Singapore$)')
plt.legend(title='Certification')
plt.show()
```

Meaning: Reveals patterns across different certification bodies; GIA diamonds dominate the market while IGI diamonds are mostly in lower price ranges

Color by colour (Ordinal variable)

```
plt.figure(figsize=(10, 6))
sns.scatterplot(
    data=df,
    x='carat',
    y='price',
    hue='colour',
    palette='tab10',  # Use distinct colors for multiple categories
    alpha=0.8
)
plt.title('Scatter Plot: Carat vs Price Colored by Diamond Colour')
plt.xlabel('Carat')
plt.ylabel('Price (Singapore$)')
plt.legend(title='Colour (D=best, I=worst)')
plt.show()
```

Meaning: Shows that higher quality colors (D/E/F) tend to be priced higher than lower quality colors (G/H/I) for the same carat size

Color by clarity (Ordered categorical variable)

```
# Define order from best to worst clarity
clarity_order = ['IF', 'VVS1', 'VVS2', 'VS1', 'VS2']
df['clarity'] = pd.Categorical(df['clarity'], categories=clarity_order, ordered=True)

plt.figure(figsize=(10, 6))
sns.scatterplot(
    data=df,
    x='carat',
    y='price',
    hue='clarity',
    palette='viridis',  # Sequential palette for ordinal data
    alpha=0.8
)
plt.title('Scatter Plot: Carat vs Price Colored by Clarity')
plt.xlabel('Carat')
plt.ylabel('Price (Singapore$)')
plt.legend(title='Clarity (IF=best, VS2=worst)')
```

```
plt.show()
```

Meaning: Demonstrates that higher clarity diamonds command higher prices at the same carat weight, especially for larger diamonds (>0.7 carat)

Advanced Visualization Techniques

Faceted scatter plots (Small multiples)

```
# Create separate plots for each certification type
g = sns.FacetGrid(df, col='certification', hue='colour', palette='tab10', col_wrap=3)
g.map(plt.scatter, 'carat', 'price', alpha=0.7)
g.add_legend()
g.set_axis_labels('Carat', 'Price')
g.set_titles('Certification: {col_name}')
plt.show()
```

Meaning: Splits data by one categorical variable (certification) while coloring by another (colour); reveals subgroup patterns within each certification group

Custom color palettes

```
# Use custom colors for specific categories
custom_colors = ['#FF6B6B', '#4ECDC4', '#45B7D1', '#96CEB4', '#FFEAA7', '#DDA0DD']
plt.figure(figsize=(10, 6))
sns.scatterplot(
    data=df,
    x='carat',
    y='price',
    hue='colour',
    palette=custom_colors,
    alpha=0.8
)
plt.title('Scatter Plot with Custom Color Palette')
plt.show()
```

Meaning: Allows for brand-specific or publication-specific color schemes

---

Practical Tips for Effective Categorical Coloring

1. **Include clear legends and labels:**
   - Always include a legend with a descriptive title
   - Add explanatory text in the legend when category order matters (e.g., "D=best, I=worst")
2. **Interpretation guidelines:**

- Look for separation between groups: distinct clusters suggest strong categorical effects
- Check for interaction effects: does the relationship between carat and price differ across categories?
- Be cautious about overinterpreting patterns in sparse regions

3. **Best practices:**
   - Limit the number of categories shown simultaneously (ideally < 7)
   - Ensure colors are distinguishable for colorblind viewers
   - Use consistent color schemes across related visualizations
   - Always verify that observed patterns are statistically significant

Remember: Adding color based on a categorical variable transforms a simple 2D scatter plot into a powerful multivariate visualization that can reveal hidden subgroup patterns and interactions that would be invisible in aggregate plots.