Socio-Informatics 348
Practical 3

Submission Instructions

- Submit your completed practical as `studentnumber.qmd` on SocSciLearn.
- Submissions are checked for completeness, not correctness.
- At least 80% of exercises must be attempted to receive 1% towards AF assessment.
- Attendance of at least one practical session per week is required to earn the 1% for that week's practical.

Deadline

Friday 29 August, 17:00 (submit on SocSciLearn)

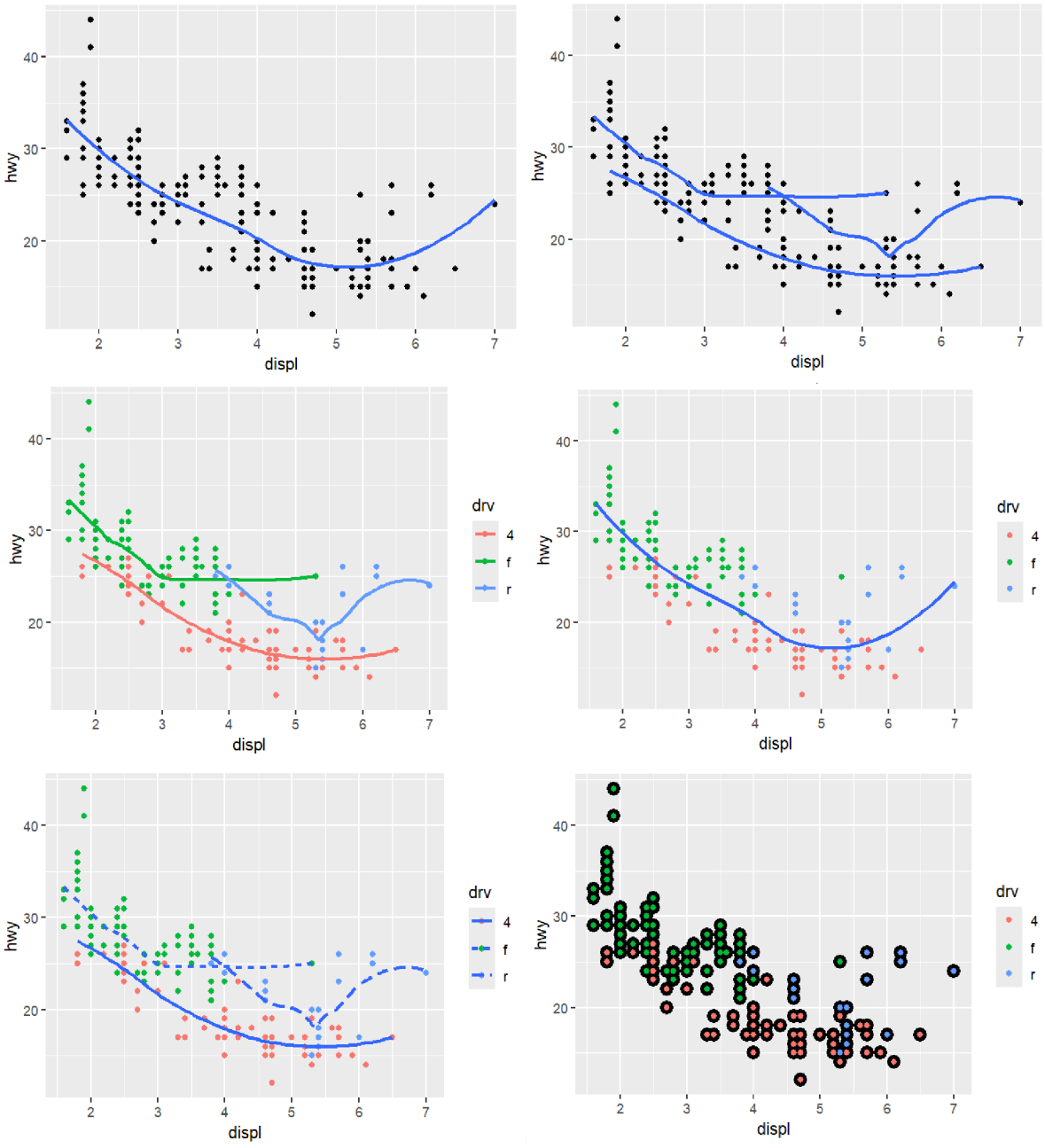
Exercises

Section 1: `mpg` dataset from `ggplot2` package

1. Create a scatterplot of `hwy` vs. `displ` where the points are pink filled in triangles.

   Note: Make sure you understand the difference between assigning constant values (such as 'pink') to the colours and shapes of geometries, and mapping variables from the dataset to aesthetics. **The `aes()` function is used for mapping variables to aesthetics**, while constant values are assigned outside of `aes()`.

2. Write the R code necessary to separately generate each of the following plots. Note that wherever a categorical variable is used in the plot, it is `drv`.


hwy
displ
drv
4
f
r

Section 2: `diamonds` datasset from `ggplot2` package

3. Create a single stacked bar chart of the number of diamonds by `cut`.

    - In your aesthetic mappings, set `x = ""` and `fill = cut`.
    - Using `x = ""` tells ggplot2 to create a single bar.
    - Add a `coord_polar()` layer to the plot to convert your single bar into a pie chart.
    - The `coord_polar()` requires a `theta` argument, which specifies the variable to be mapped to the angle of the pie chart.
    - Set `theta = "y"` to map the y aesthetic (the count of diamonds in each cut category) to the angle of the pie chart.

4. Create a visualisation of diamond prices vs. a categorical variable from the diamonds dataset using `geom_violin()`, then a faceted `geom_histogram()`, then a `geom_freqpoly()`, and then a `geom_density()`. Where necessary, map the categorical variable to the `color` mapping. Compare and contrast the four plots. What are the pros and cons of each method of visualising the distribution of a numerical variable based on the levels of a categorical variable?

5. What happens to missing values in a histogram? What happens to missing values in a bar chart? Why is there a difference in how missing values are handled in histograms and bar charts?

6. Based on EDA, what variable in the diamonds dataset appears to be most important for predicting the price of a diamond? How is that variable correlated with cut? Why does the combination of those two relationships lead to lower quality diamonds being more expensive?

    - Plot the relationships between `price` and the following variables: carat, clarity, and color.
    - Which of these variables appears to be most important for predicting the price of a diamond?
    - Plot the relationship between this variable and `cut`.
    - Explain how the combination of these two relationships leads to lower quality diamonds being more expensive.

7. Create a `geom_tile()` plot of clarity vs color, where the fill colour of each tile represents the number of observations that fall into each combination of clarity and color.

8. Create a `geom_tile()` plot of clarity vs color, where the fill colour of each tile represents the median price of diamonds for each combination of clarity and color.