Exercise: Multiple Distributions

This exercise uses:

- `{ggplot2}`, `{ggridges}`, and `{ggExtra}` functions
- Your knowledge from past in-class exercises, videos, homework, etc. and corresponding modules from the course site.

Overview

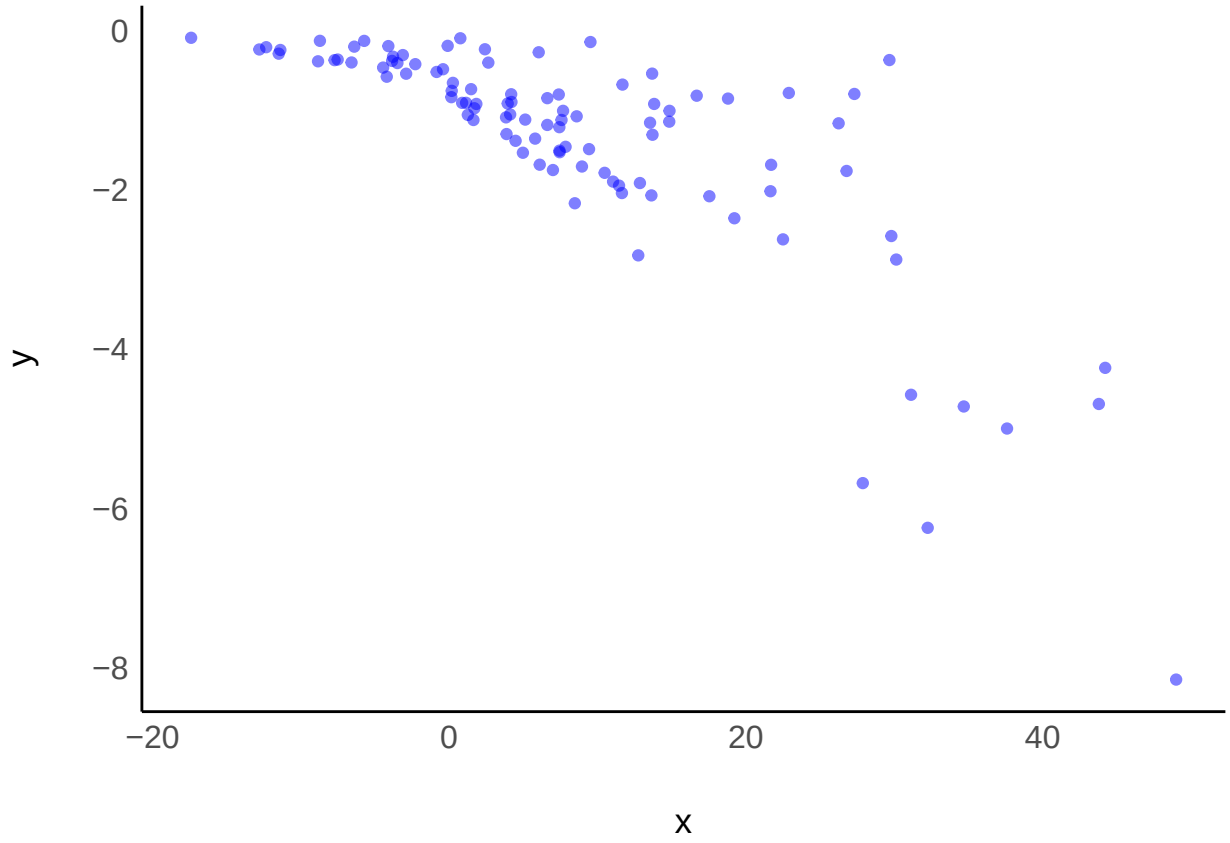
This exercises provides some practice creating visualizations that allow you to compare distributions of data across multiple categorical groups. Whereas `geom_histogram()` and `geom_density()` are useful for visualizing a single distribution, they are compromised in their utility for making comparisons. Similarly, whereas `geom_bar()` or `geom_col()` *may* be useful for communicating summary statistics like the mean or median, they fail to communicate the underlying data distribution which informs those measures of central tendency. Adding a `geom_point()` layer to those bar plots will provide some crude visual details pertaining to the range. For the *trained eye*, some basic density information may be extracted but this will likely not be comprehensive and representative of the true distribution. Adding distributions to plots can help with interpreting data and appropriateness of statistical models.

Data Set

If appropriate, use your team project data. If you dot not have numeric variables that are conducive to visualizing data distributions like histograms or density plots, use `ggplot2::diamonds` data. In this exercise, you will think about distributions of data as well as create visualizations containing multiple distributions in order to compare them easily.

Problem 1: Distribution Analysis

Consider the following point plot.



Along the axis below, draw a density distribution (or histogram) that might resemble the underlying data for the y axis.

Are you confident in your attempt to represent the distribution?

Problem 2: Ridgeline Distributions

Using **{ggridges}**, visualize the density distributions for a numeric variable for your project across levels of a grouping/factor variable.

Problem 3: Adding Marginal Distributions to Scatterplots

In general, scatterplots do not include distributions of the data. You may attempt to extract their distributional shape mentally (problem #1) but this is extremely difficult to do and will likely lead to errors.

The {ggExtra} library allows you to add distributions in the margins of a plot.

- create a scatterplot with two numeric variables
- try to visualize the shape of a density distribution for the x and y variables
- assign that plot object to a name
- pass that object to the `p` (stands for plot) parameter of `ggExtra::ggMarginal(p = ?)`;
- by default the distribution `type = "density"`; you can, however, pass one of these as arguments to `type`: "density", "histogram", "box plot", "violin", "densigram"
- determine how accurately you were able to visualize the distributions