Data Visualization Lab

Prof. Ramesh Ragala
SCSE, VIT Chennai

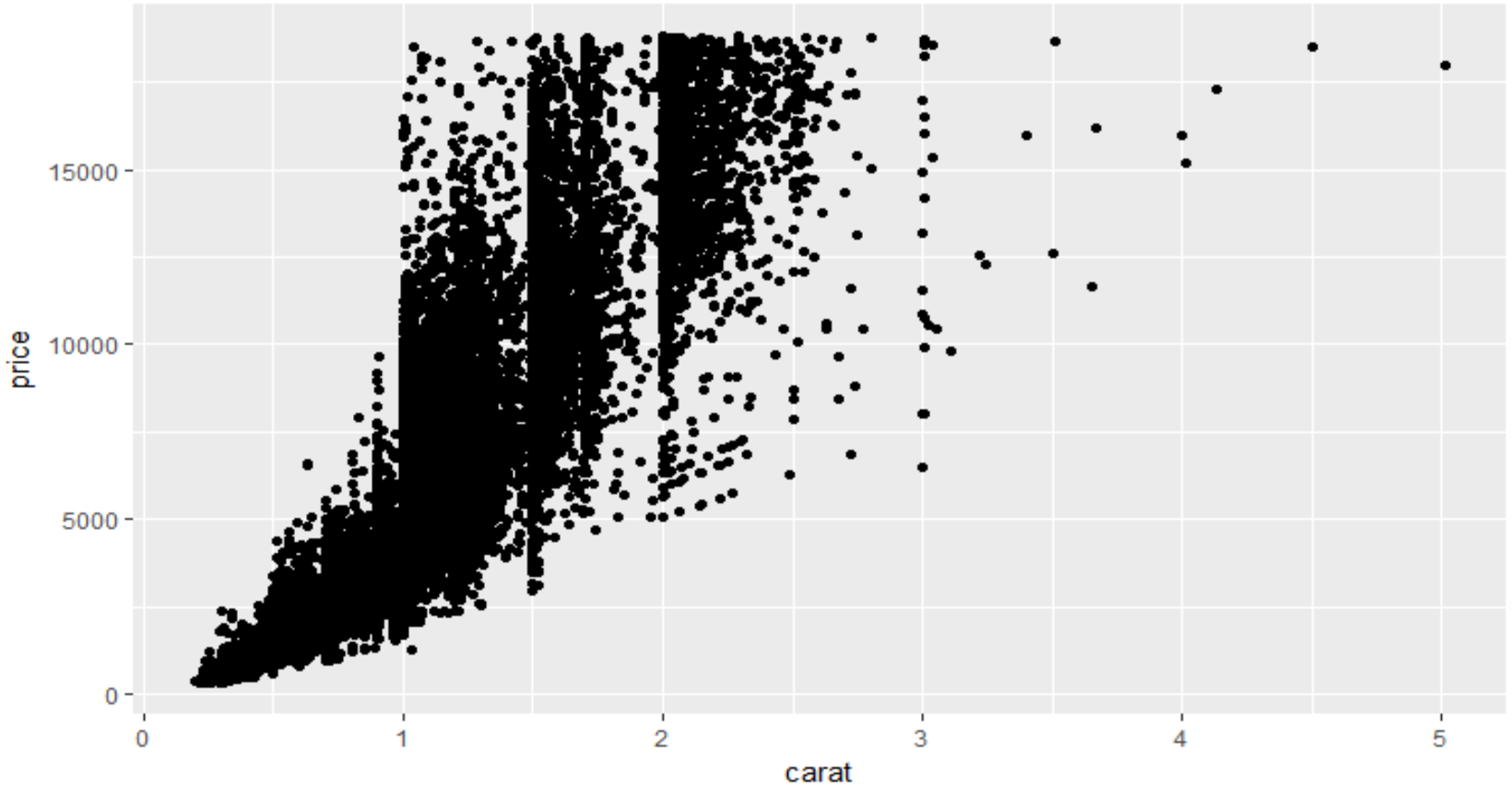
Dataset Description

- Built – in Dataset in R
  - Diamonds (dataset)
  - The diamonds dataset consists of prices and quality information about 54,000 diamonds.
  - It is included in the ggplot2 package.
  - The data contains the four C's of diamond quality, carat, cut, colour and clarity and five physical measurements depth, table, x, y and z (dimensions of diamonds)

Introduction to qplot

- qplot(carat,price,data=diamonds)
  - It produces a scatter plot showing the relationship between the price and carat(weight) of a diamond.
  - The plot shows a **strong correlation with notable outliers** and some interesting **vertical striation**.
  - The relation looks a exponential

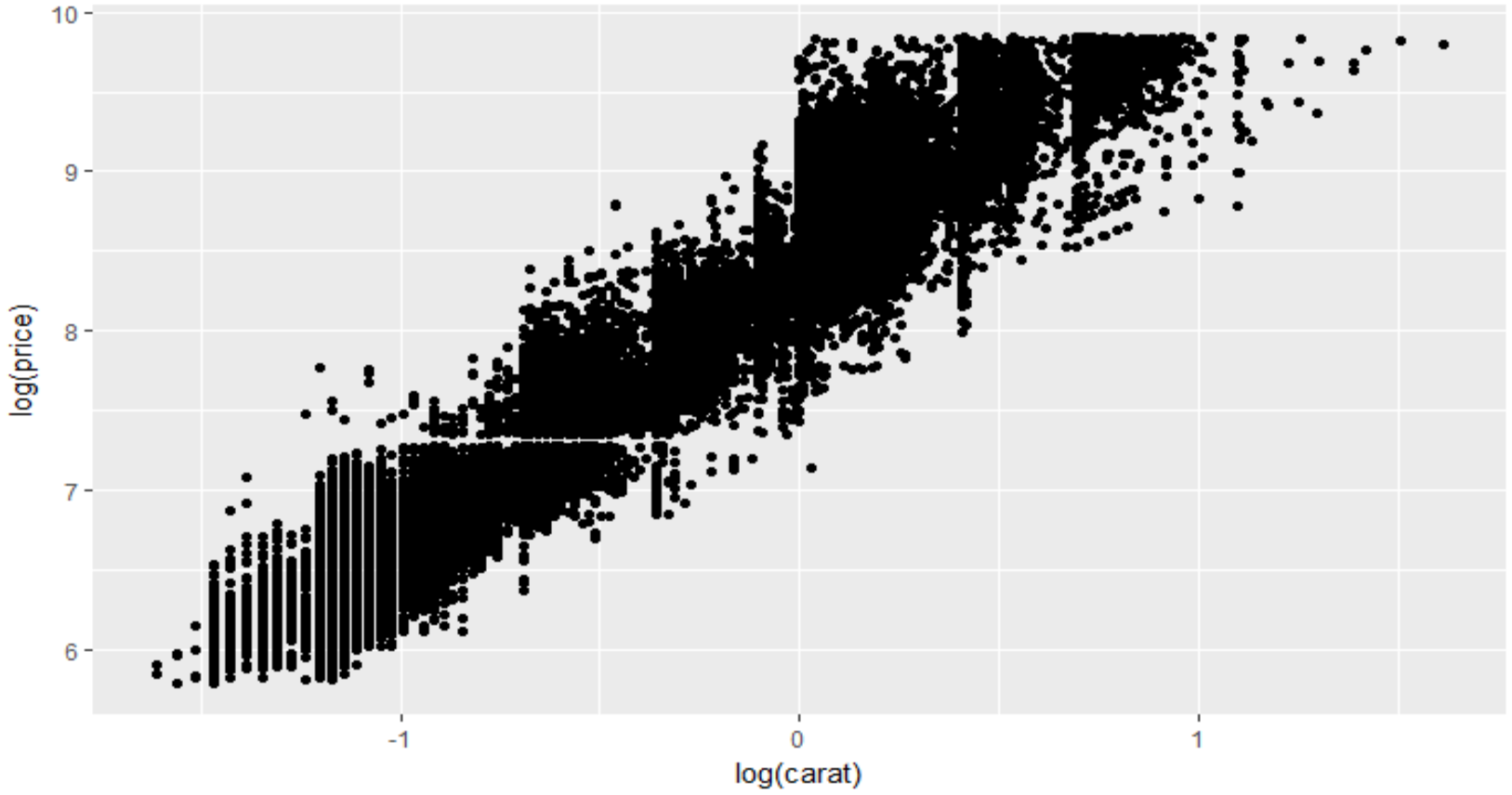
Introduction to qplot



Introduction to qplot

- qplot(log(carat), log(price), data = diamonds)
  - qplot() accepts functions of variables as arguments
  - The relationship now looks **linear**

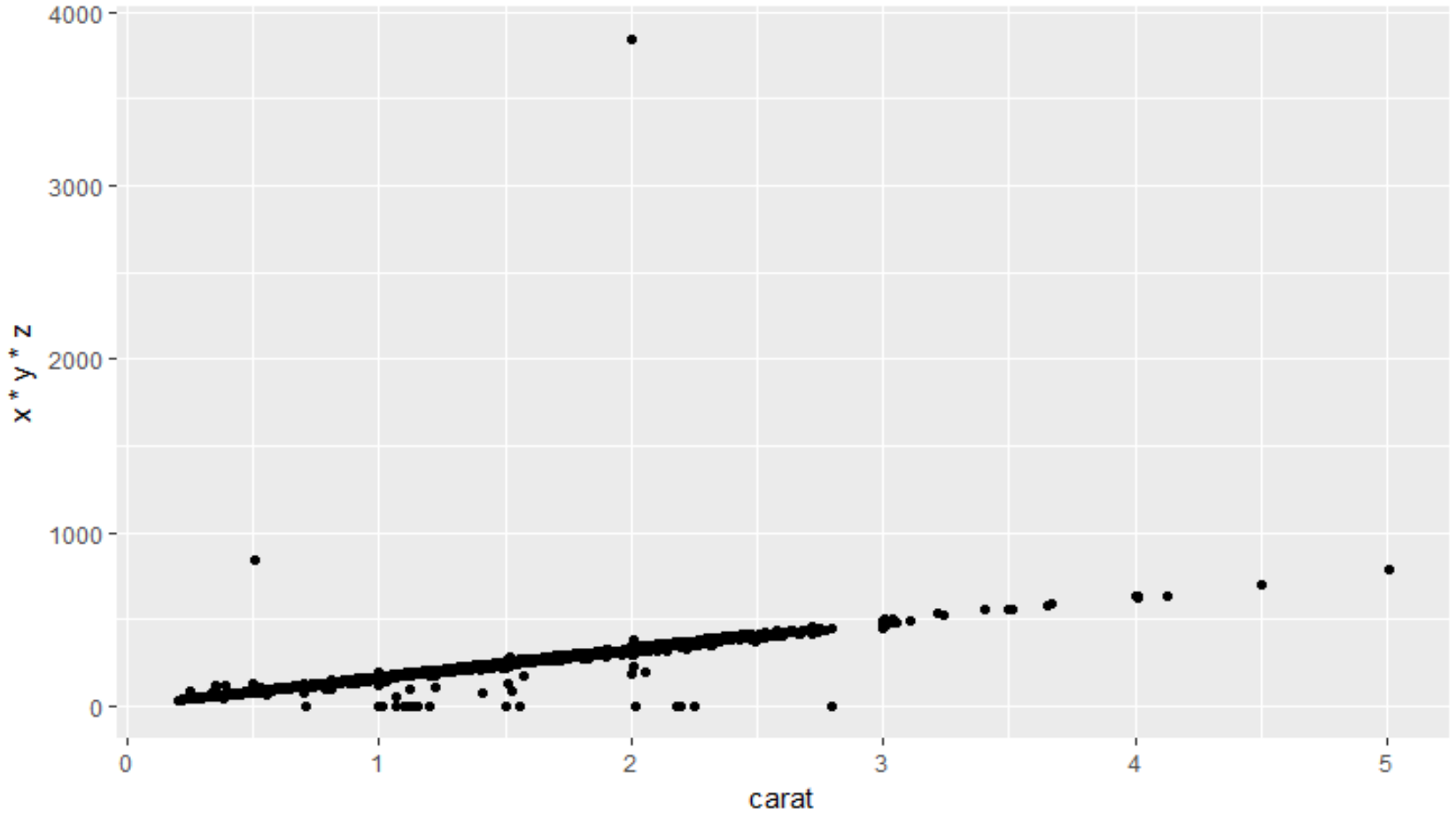
Introduction to qplot



Introduction to qplot

- qplot(carat, x * y * z, data = diamonds)
  - Relationship between the volume of the diamond (approximate by x * y * z) and its weight.
  - Density of the diamonds to be constant
  - Linear relationship between volume and weight
  - Most of the diamonds fall along the line, but there are some large outliers

Introduction to qplot

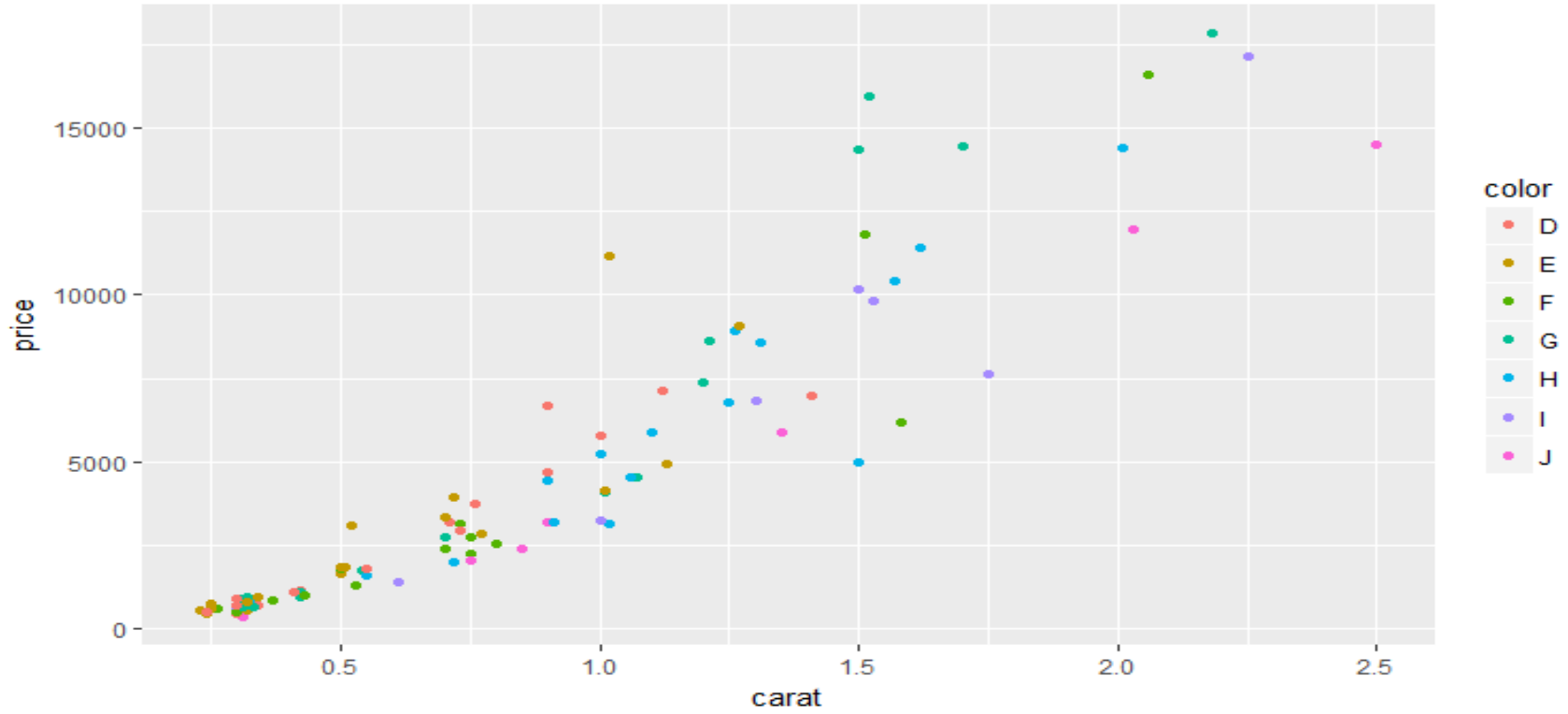


Introduction to qplot

- qplot() **converts categorical** data of dataset into something that plot knows how to use **automatically**.
- it will automatically provide a **legend** that **maps the displayed attributes** to the data values.
- Colour, size and shape are all examples of aesthetic attributes, visual properties that affect the way observations are displayed.

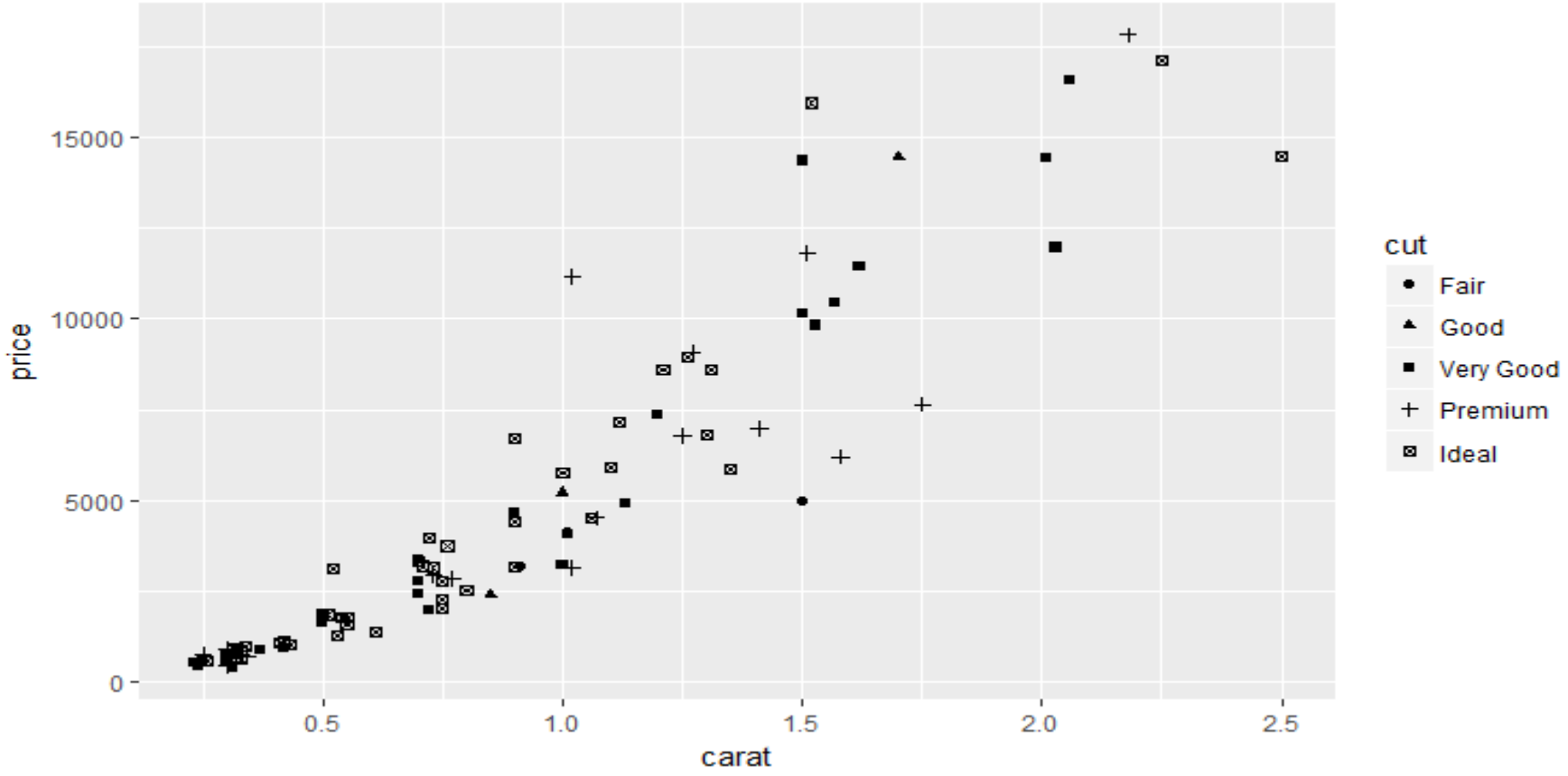
Introduction to qplot

- qplot(carat, price, data = smalldataset, colour = color)



Introduction to qplot

- qplot(carat, price, data = smalldataset, shape = cut)

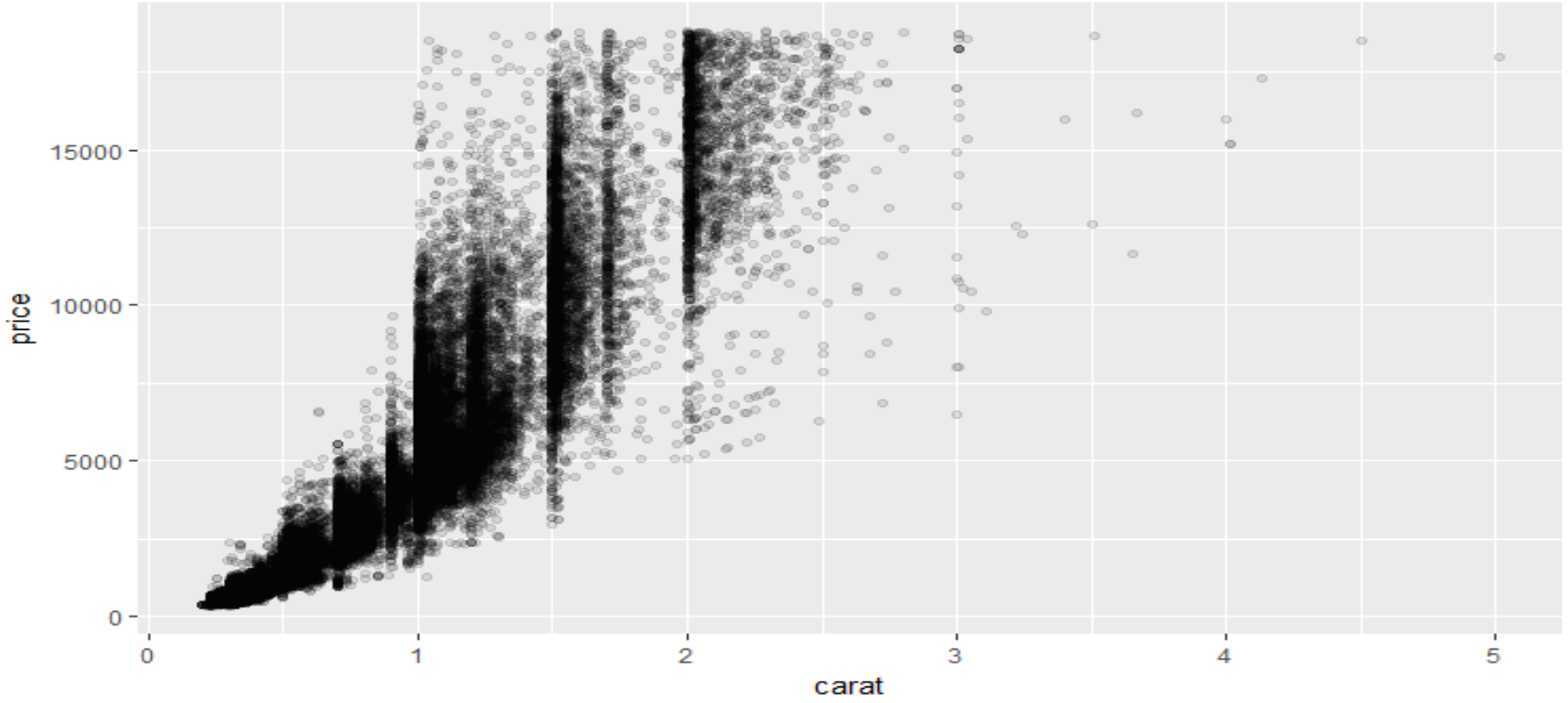


Introduction to qplot

- Scale () which maps data values to a valid values for that aesthetic.
- Scale() controls the appearance of the points and associated legend.
- manually set the aesthetics using I().
  - Colour = I("red") or Size() = I(2)
- alpha aesthetic can be used to make a semi-transparent colour.
  - 0→ completely transparent and
  - 1 → completely opaque

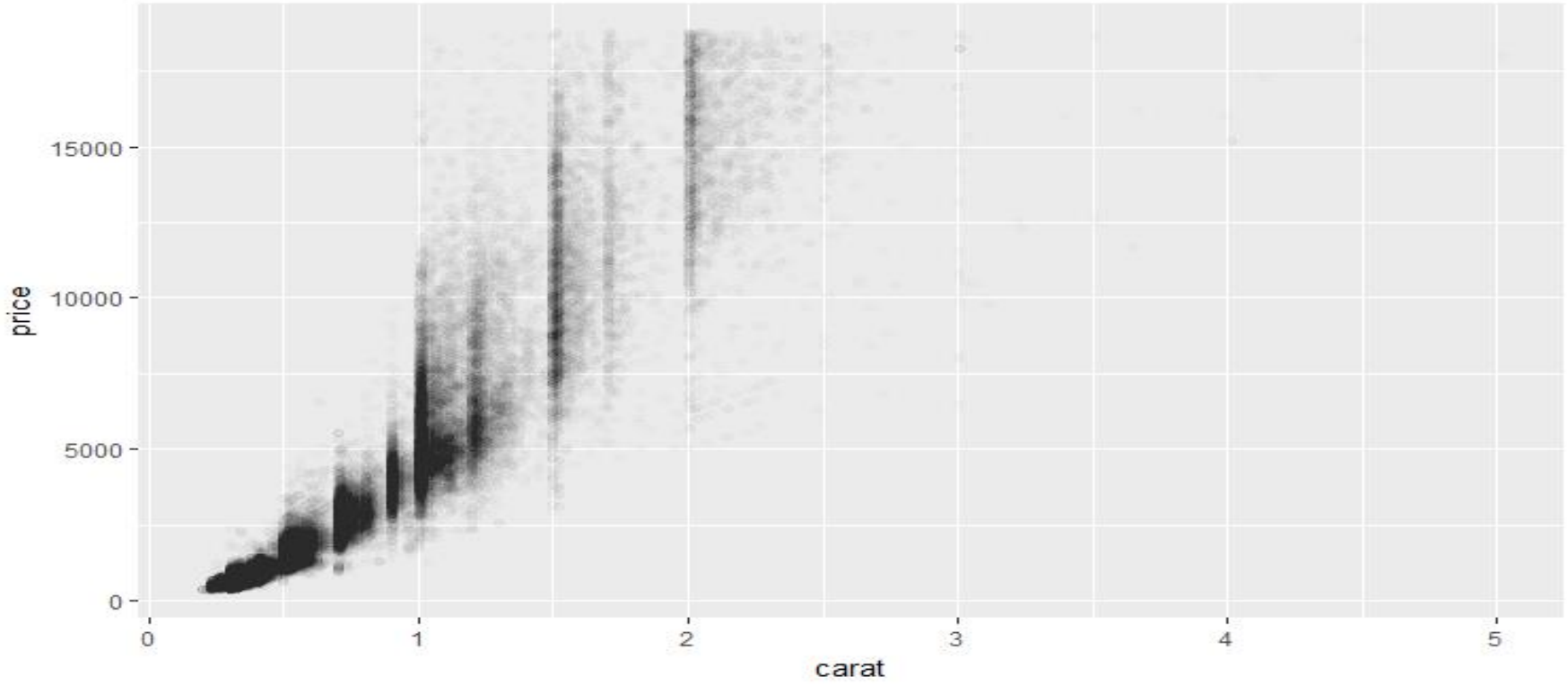
Introduction to qplot

- qplot(carat, price, data = diamonds, alpha = I(1/10))



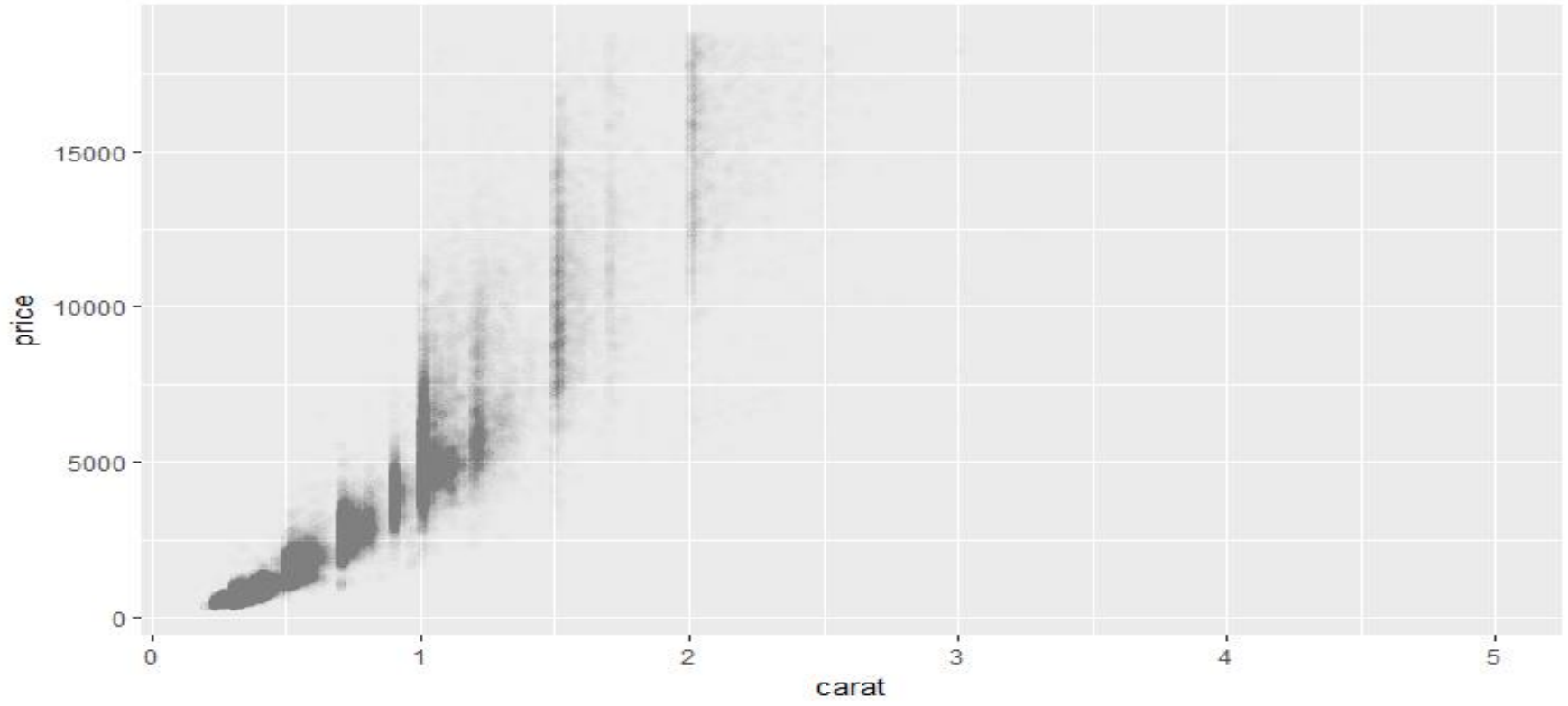
Introduction to qplot

- qplot(carat, price, data = diamonds, alpha = I(1/100))



Introduction to qplot

- qplot(carat, price, data = diamonds, alpha = I(1/200))



Introduction to qplot

- By varying geom(), qplot can produce different kinds of plots.
- Geom
  - geometric object
  - Describes the type of object that is used to display the data.
  - Some geoms have an associated statistical transformation → histograms

Introduction to qplot

- geom = "point"
  - draws points to produce a scatterplot.
  - This is the default when you supply both x and y arguments to qplot().
- geom = "smooth"
  - fits a smoother to the data and displays the smooth and its standard error.
- geom = "boxplot"
  - It produces a box-and-whisker plot to summaries the distribution of a set of points

Introduction to qplot

- geom = "path”
  - Draws a lines between the data points.
  - used to explore relationships between time and another variable.
  - Paths can go in any direction
- geom = “line”
  - Draws a lines between the data points.
  - used to explore relationships between time and another variable.
  - Lines that travel from left to right

Introduction to qplot

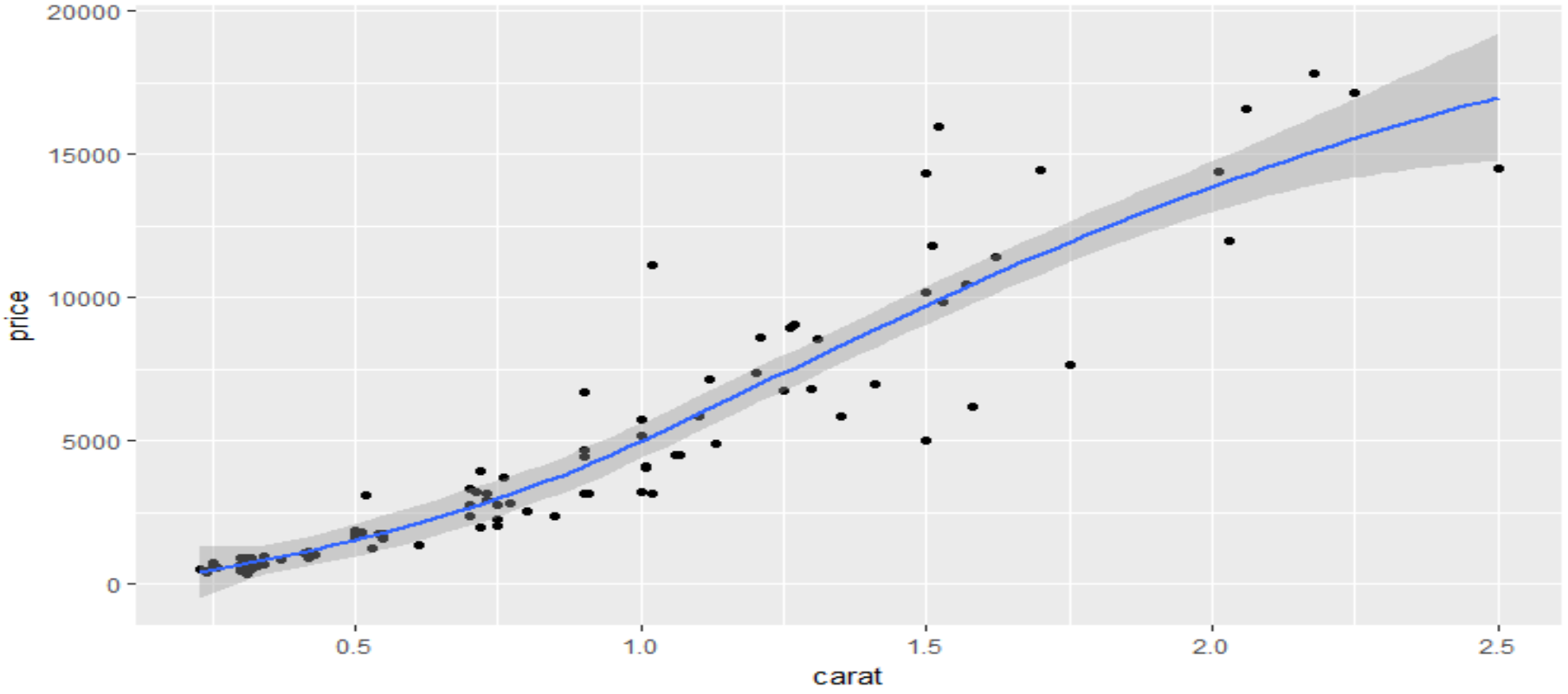
- geom = "histogram” → for one dimensional distribution
  - Draws a histogram
  - Default
- geom = “freqpoly”
  - Draws a frequency polygon
- geom = “density”
  - Draws a density plots
- geom = “bar” → discrete variables
  - Makes a bar charts

Introduction to qplot

- Adding smoother to a plot:
  - If a scatter plot have many data points, then it is difficult to see the trend shown by the data points
  - Add smoothed line to plot

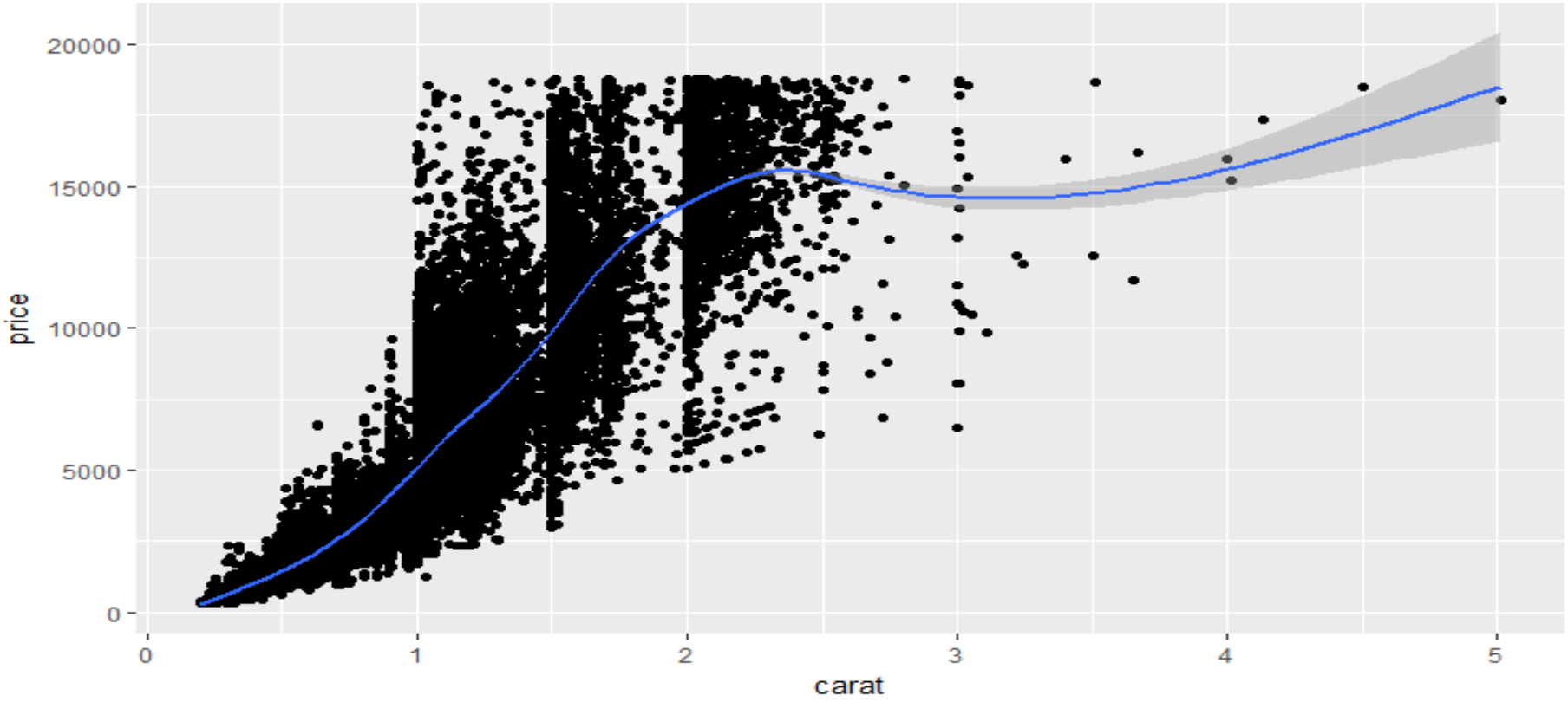
Introduction to qplot

- qplot(carat, price, data = smalldataset, geom = c("point", "smooth"))



Introduction to qplot

- qplot(carat, price, data = diamonds, geom = c("point", "smooth"))

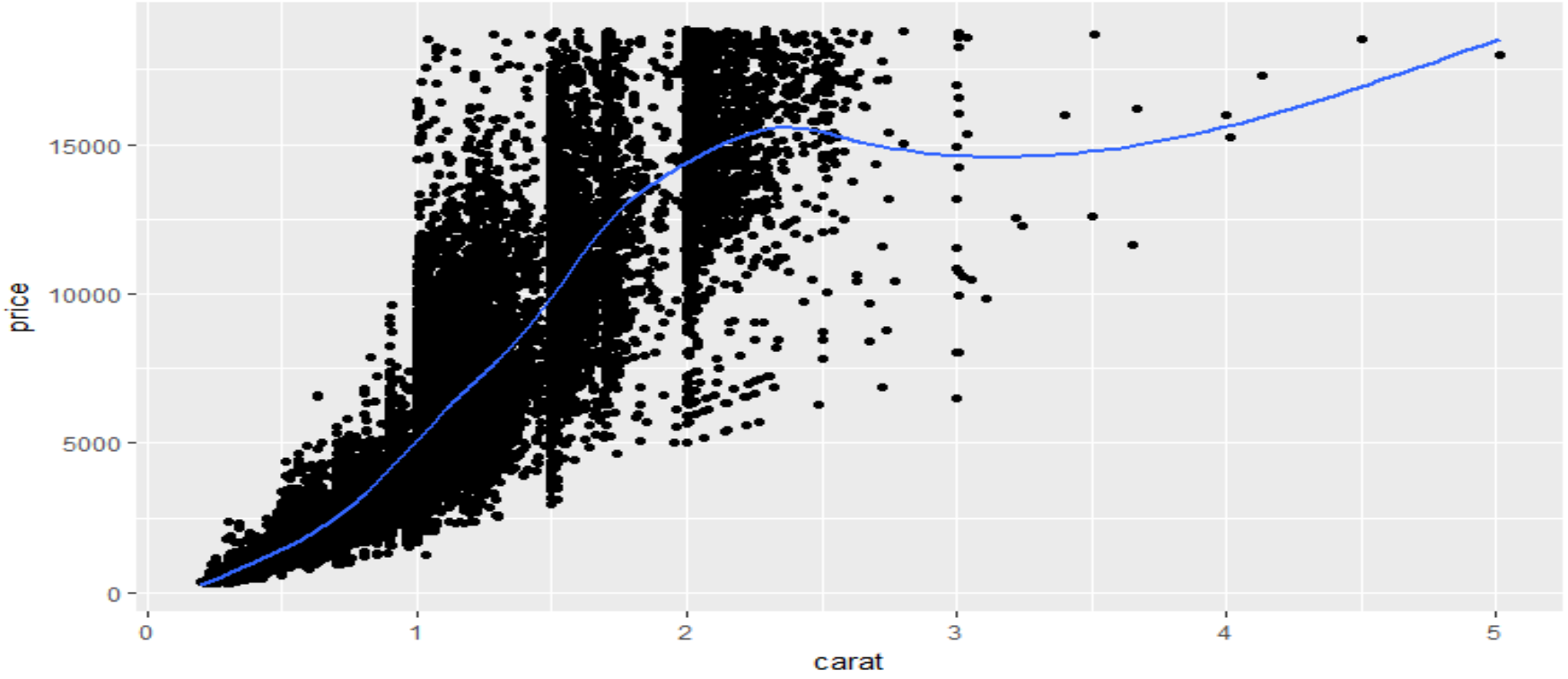


Introduction to qplot

- An **exponential relationship** between price and carat was correct.
- There are few diamonds bigger than three carats.
- uncertainty in the form of the relationship increases as illustrated by the point-wise confidence interval shown in grey.
- Use se=FALSE to turn off confidence interval

Introduction to qplot

- qplot(carat, price, data = diamonds, geom = c("point", "smooth"), se=FALSE)



Introduction to qplot

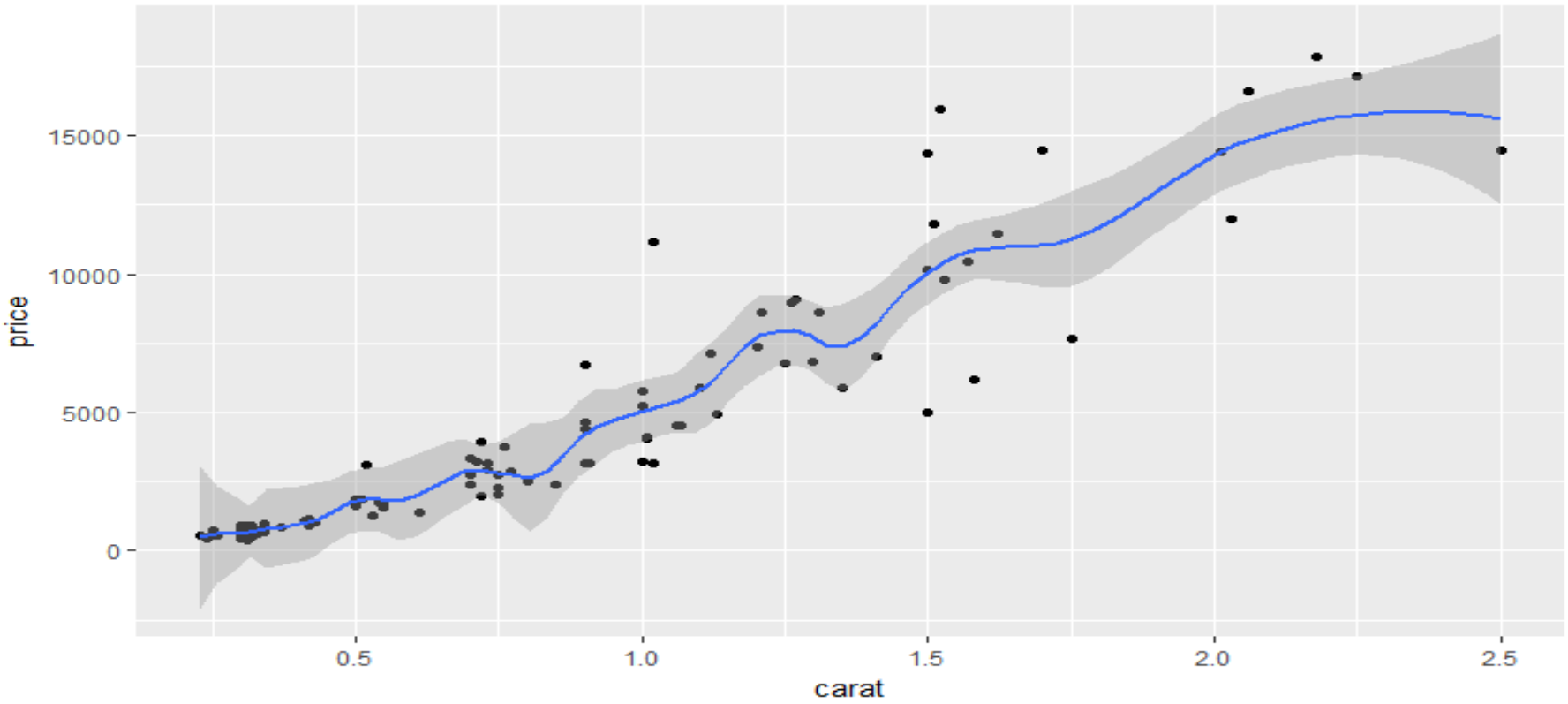
- An **exponential relationship** between price and carat was correct.
- There are few diamonds bigger than three carats.
- uncertainty in the form of the relationship increases as illustrated by the point-wise confidence interval shown in grey.
- Use se=FALSE to turn off confidence interval

Introduction to qplot

- There are many different smoothers you can choose between by using the **method():**
- method = “loess”
  - Default for small data points
  - Uses smooth local regression
  - Algorithm details are available in ?loess
  - The wiggliness of the line is controlled by the **span** parameter, which ranges from 0 (exceedingly wiggly) to 1 (not so wiggly)

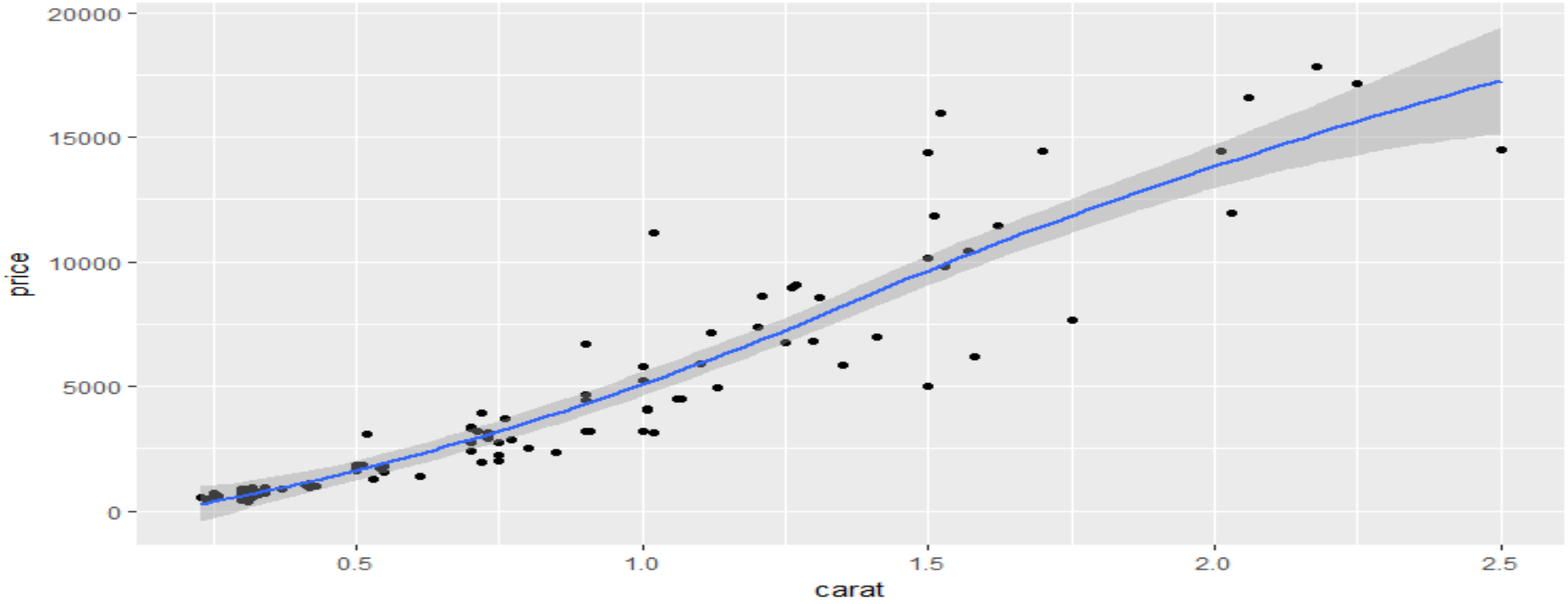
Introduction to qplot

- qplot(carat, price, data = smalldataset, geom = c("point", "smooth"), span = 0.2)



Introduction to qplot

- qplot(carat, price, data = smalldataset, geom = c("point", "smooth"), span = 1)



Introduction to qplot

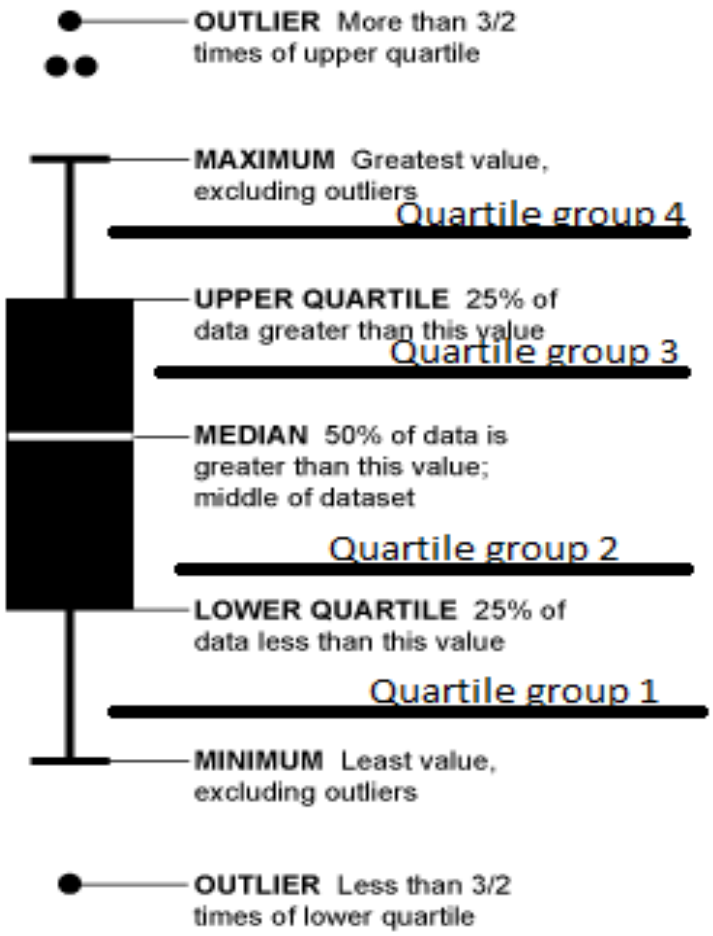
- method = “gam”
  - Uses the mgcv library
  - Works for large data point, n>1000

Introduction to qplot

- When a set of data includes a categorical variable and one or more continuous variables, you will probably be interested to know how the values of the continuous variables vary with the levels of the categorical variable.
- Box-plots and jittered points are in this scenario
- Box-plots:
  - It summaries the bulk of the distribution with only five numbers.

Introduction to qplot

- Type of data used for hist, same data can be used
- summary(fivenum(diamons $carat))
- Median
- Inter-quartile range
- Upper quartile range
- Lower quartile range
- whiskers

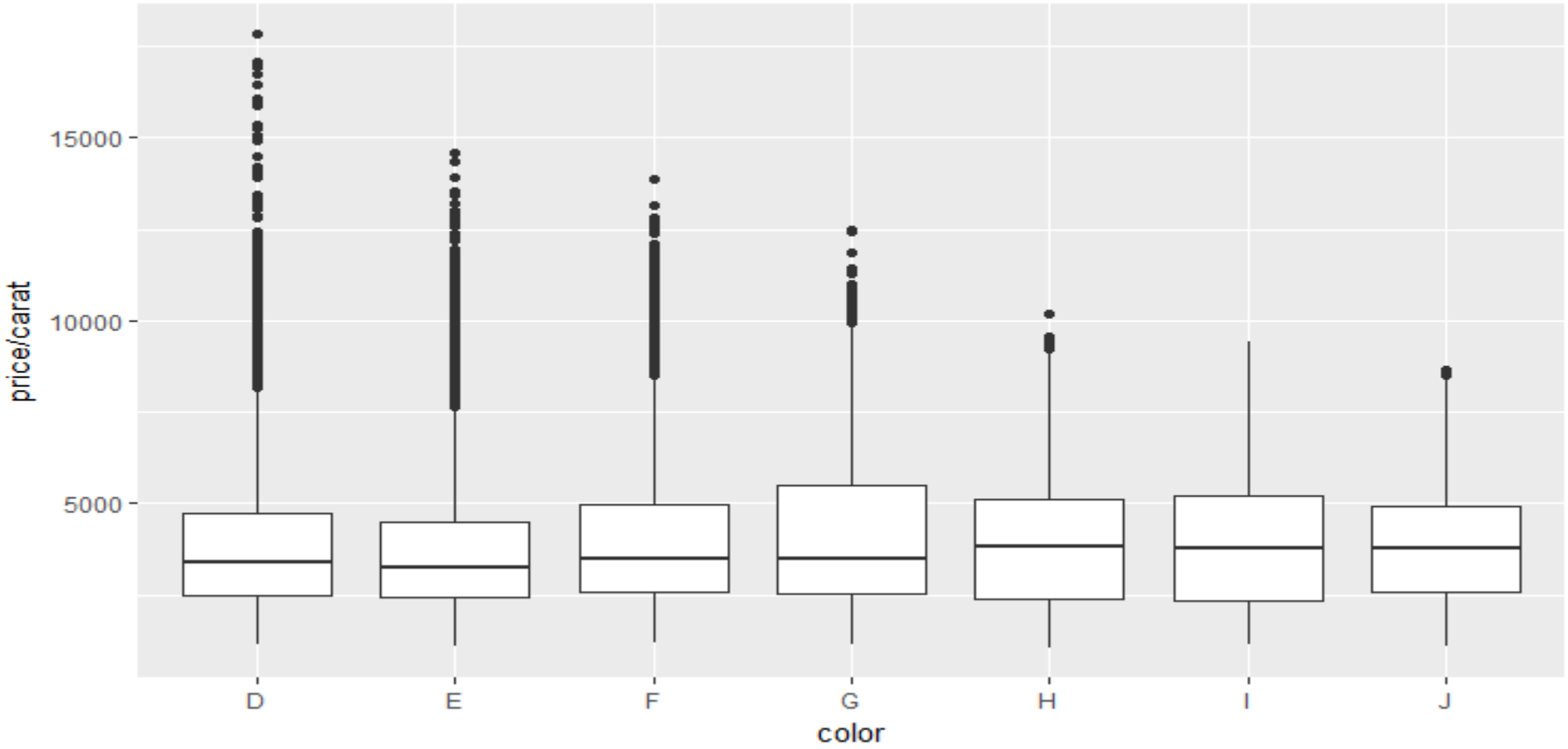


Introduction to qplot

- The box plot is comparatively short
  - Data points are closer to median
- The box plot is comparatively tall
  - Data points are not closer to median
- One box plot is much higher or lower than another
  - Differences between groups
- The 4 sections of the box plot are uneven in size
  - Long whisker means that data points are varied amongst most positive quartile group
  - Low whisker means similar for the least positive quartile group

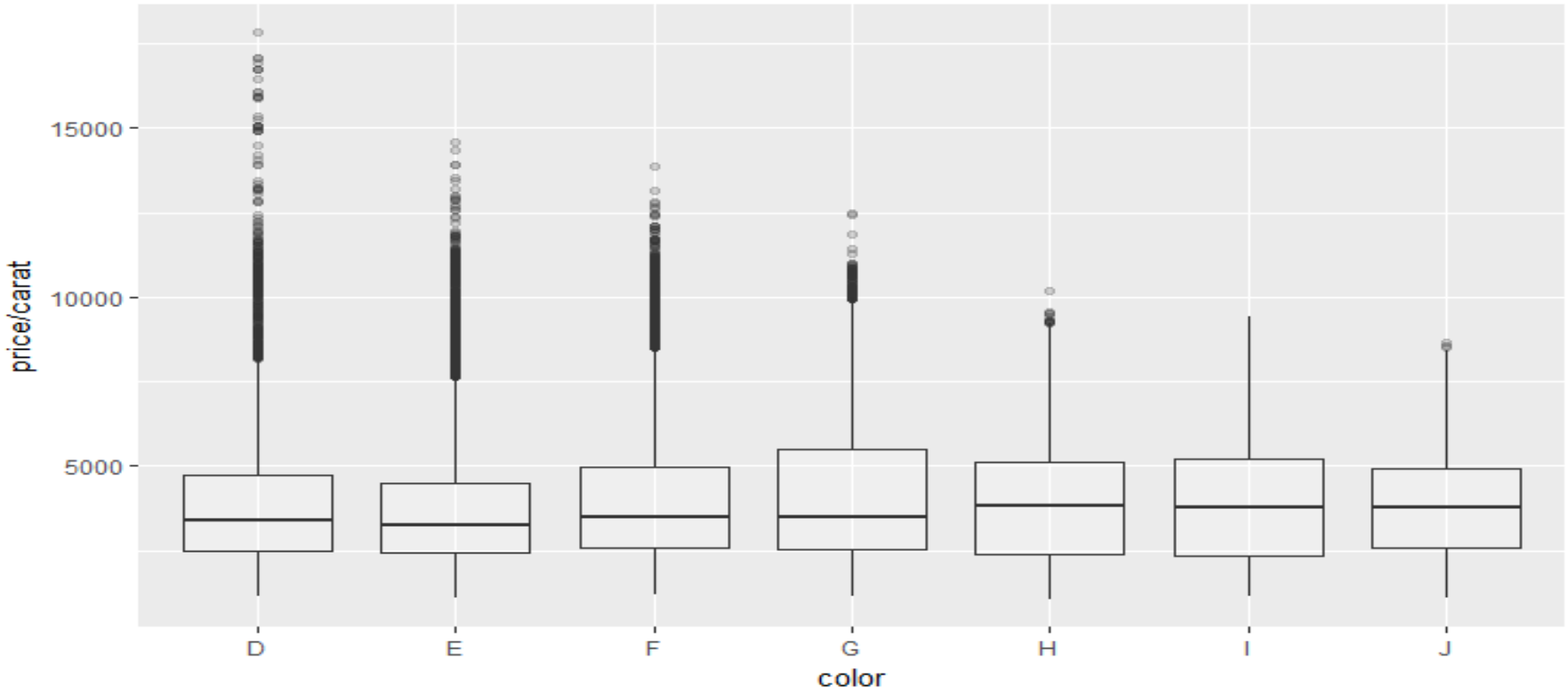
Introduction to qplot

- qplot(color, price/carat, data=diamonds, geom = "boxplot")



Introduction to qplot

- qplot(color, price / carat, data = diamonds, geom = "boxplot", alpha = I(1 / 5))

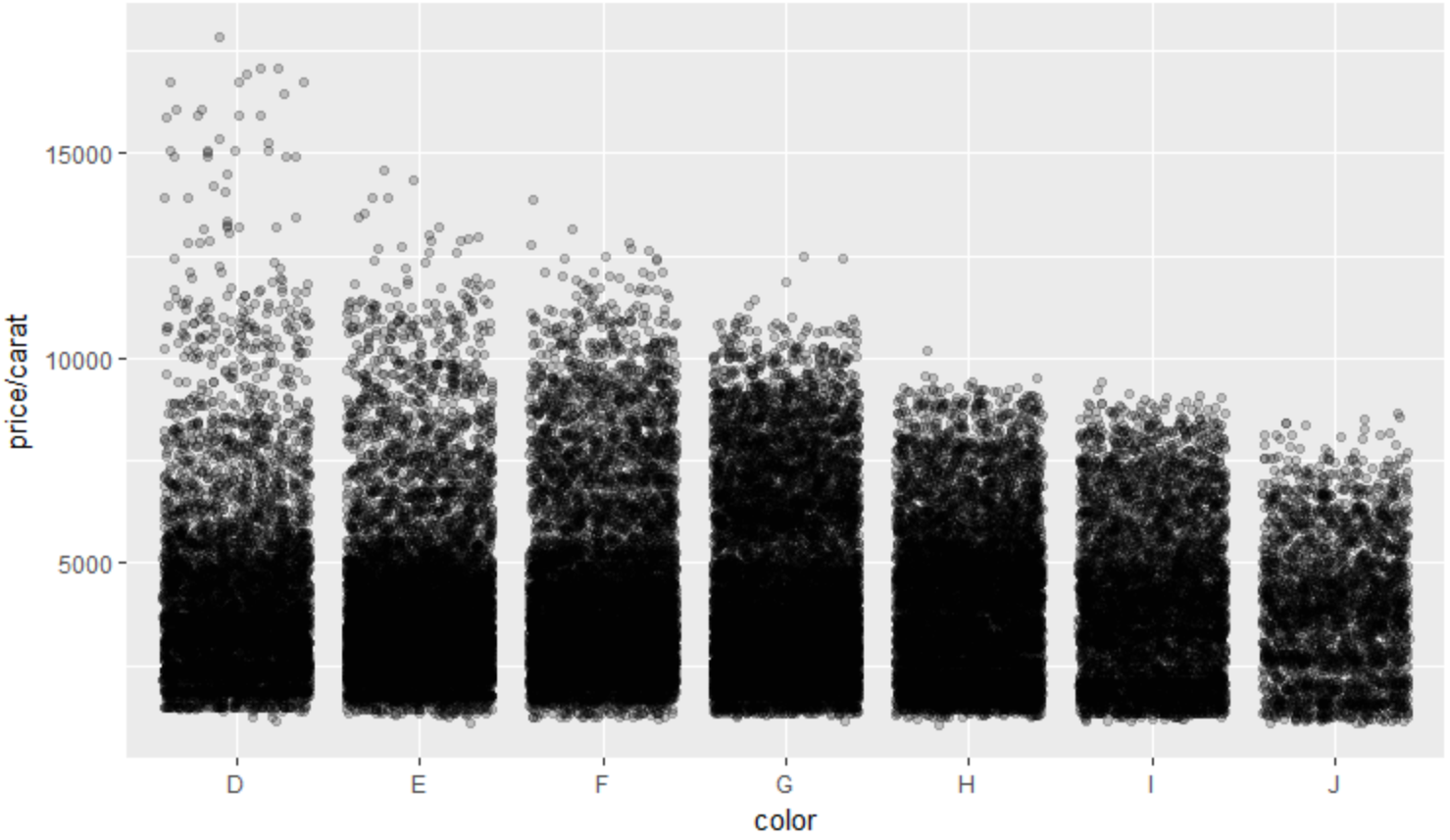


Introduction to qplot

- The dependency of the spread of price per carat on diamond colour.
- there is very little change in the median and adjacent quartiles. → generally close to average
- Box-plot has control over outline colour, the internal fill colour and the size of the lines.

Introduction to qplot

- J



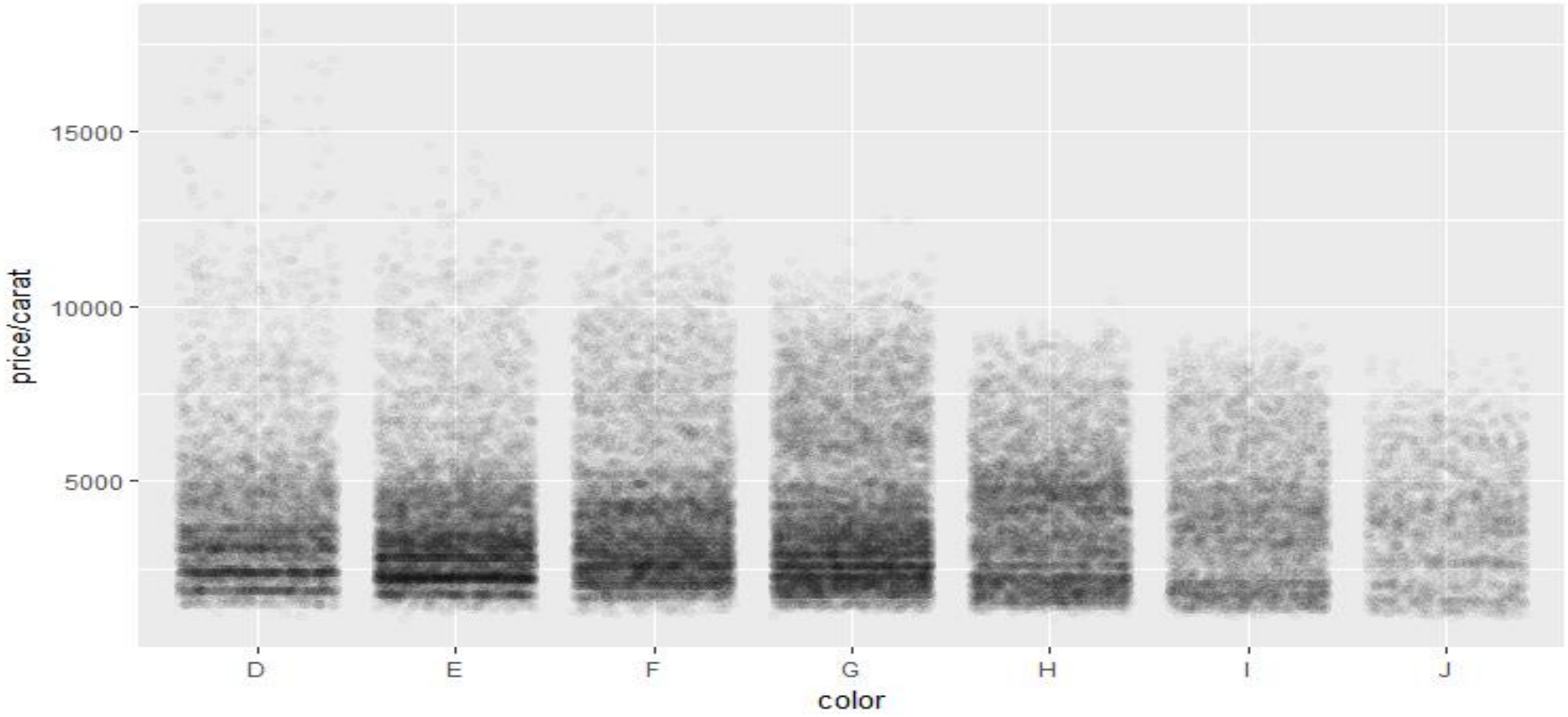
but

can

size

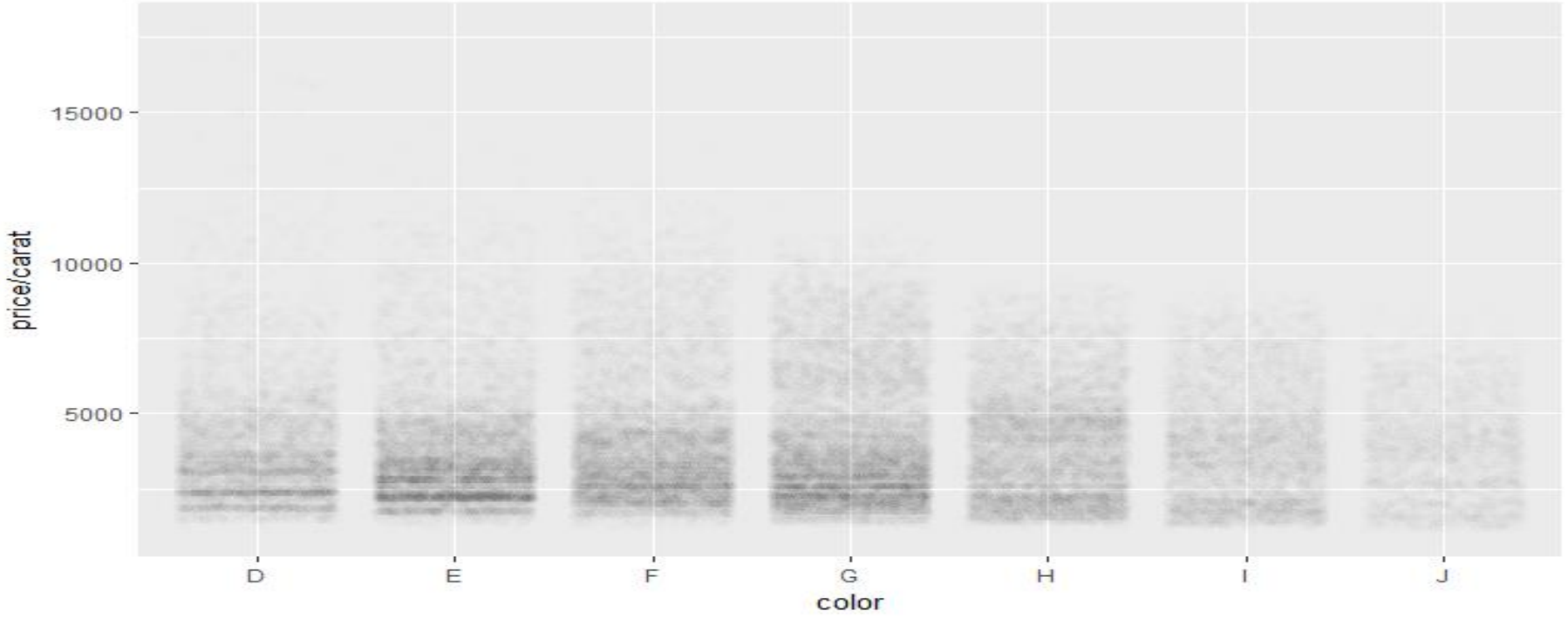
Introduction to qplot

- qplot(color, price / carat, data = diamonds, geom = "jitter", alpha = I(1 / 50))



Introduction to qplot

- qplot(color, price / carat, data = diamonds, geom = "jitter", alpha = I(1 / 200)) → sees bulk of data.

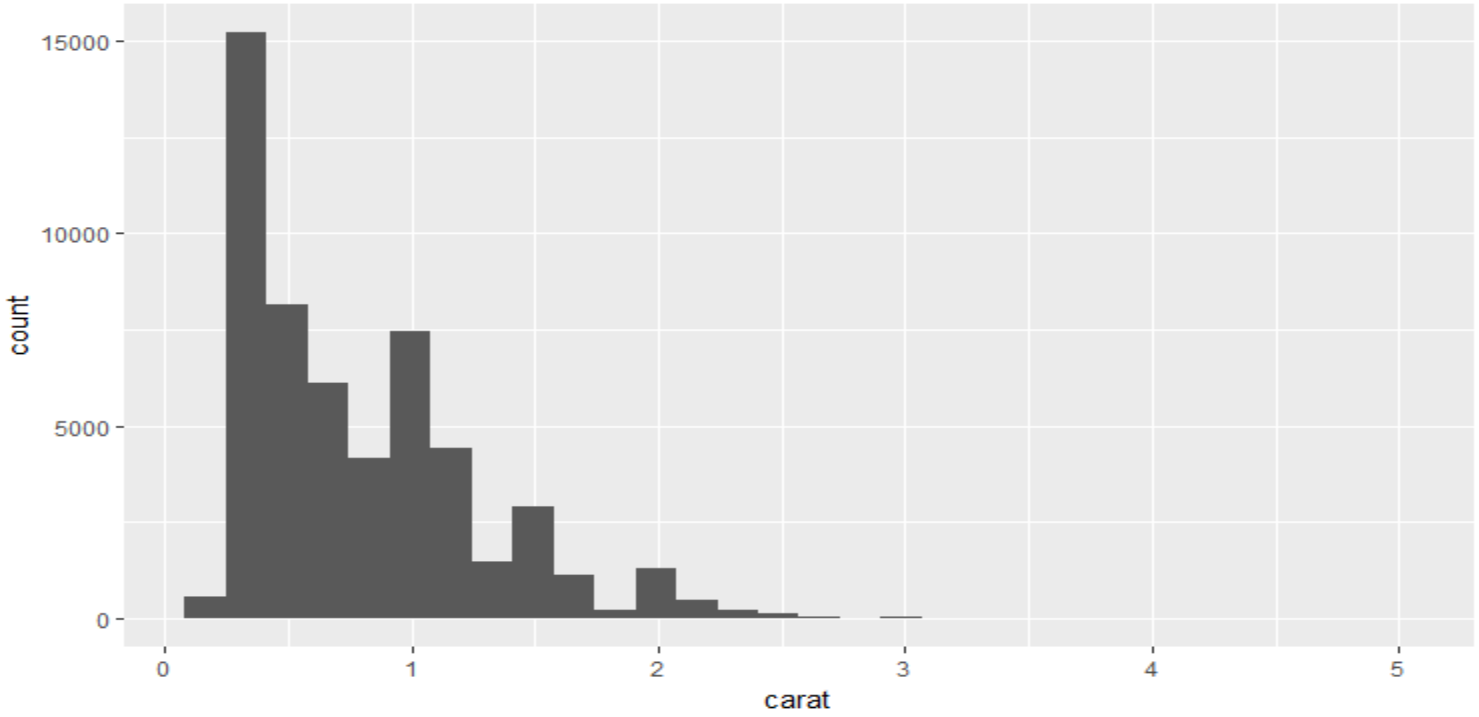


Introduction to qplot

- Histogram and density plots:
  - Distribution of single variable
  - Difficult to compare many groups
- Histograms:
  - The binwidth argument controls the amount of smoothing by setting the bin size
  - Use aesthetic variable to compare the distribution of different **sub-groups**.

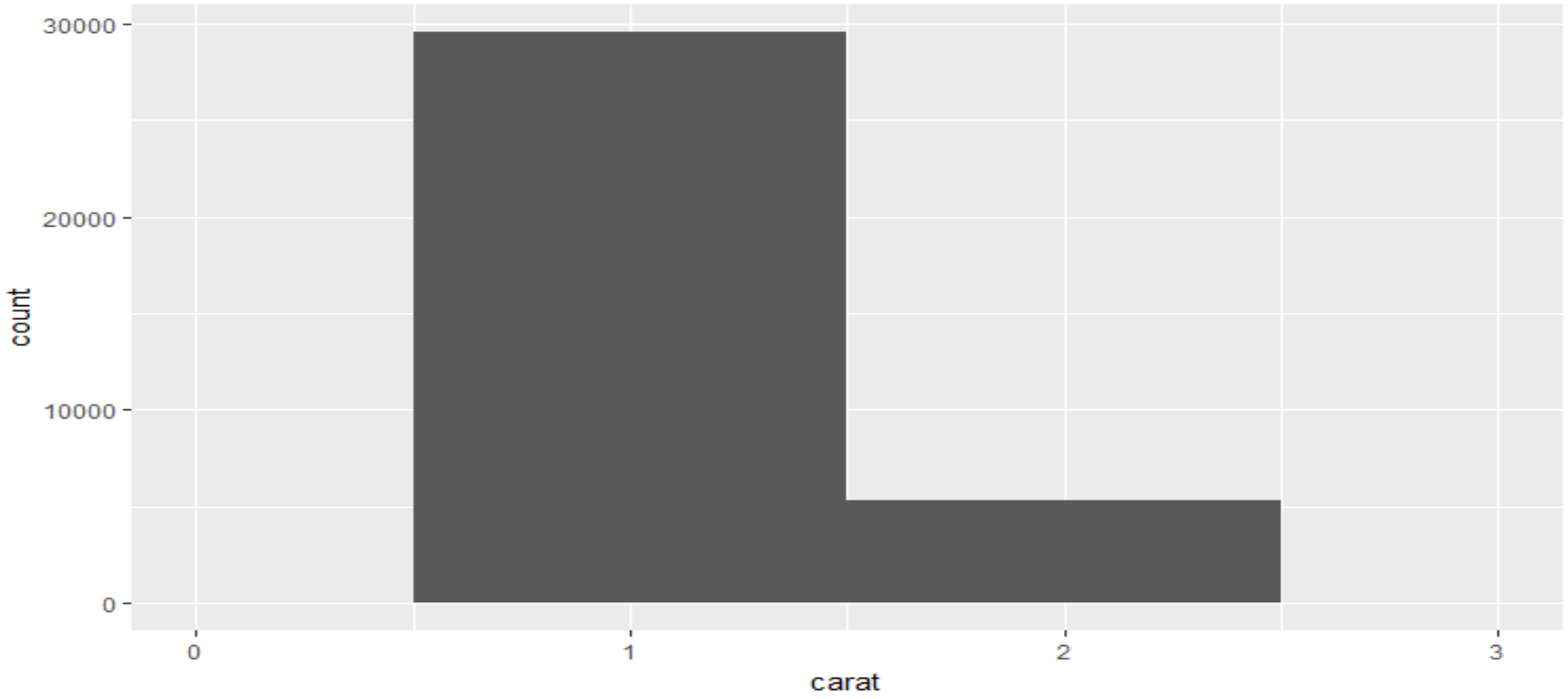
Introduction to qplot

- qplot(carat, data = diamonds, geom = "histogram") → distribution of carat variable



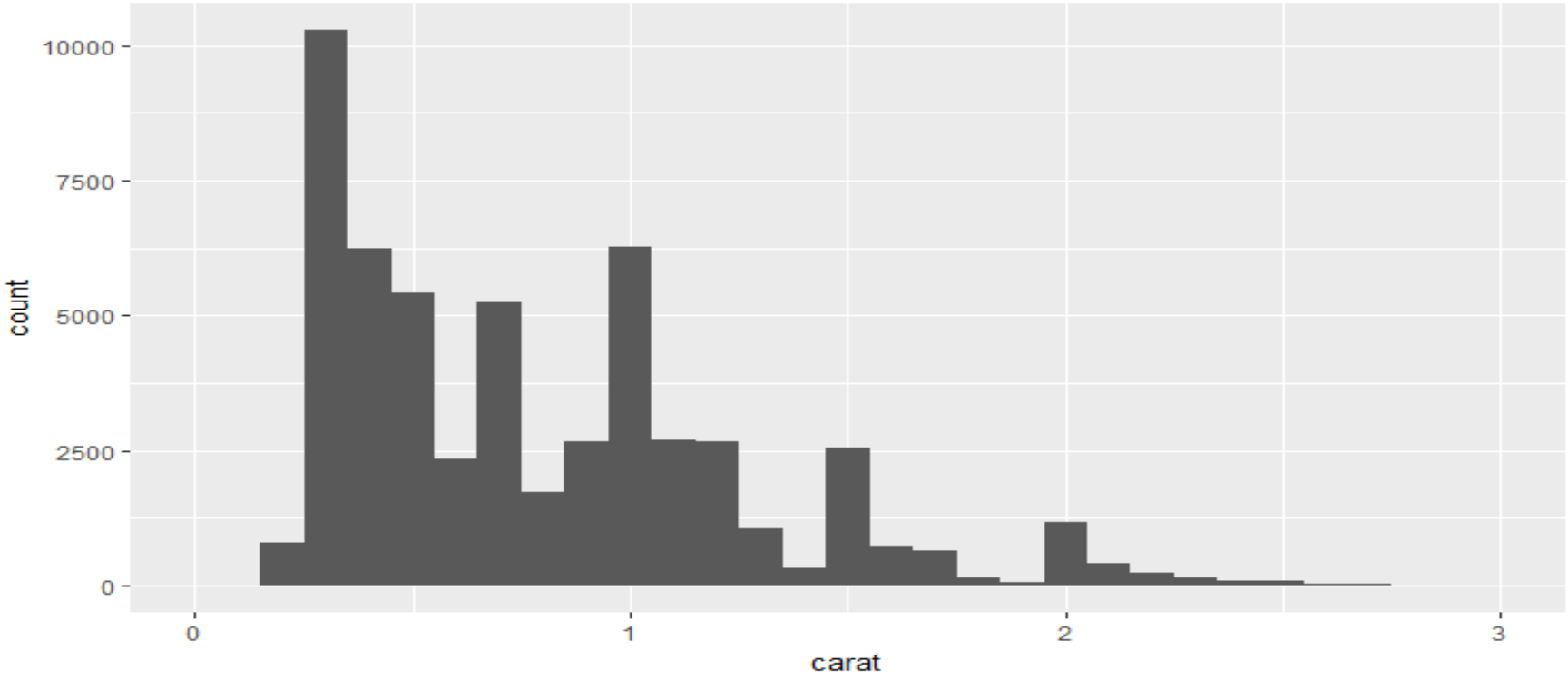
Introduction to qplot

- qplot(carat, data = diamonds, geom = "histogram", binwidth = 1, xlim = c(0,3))



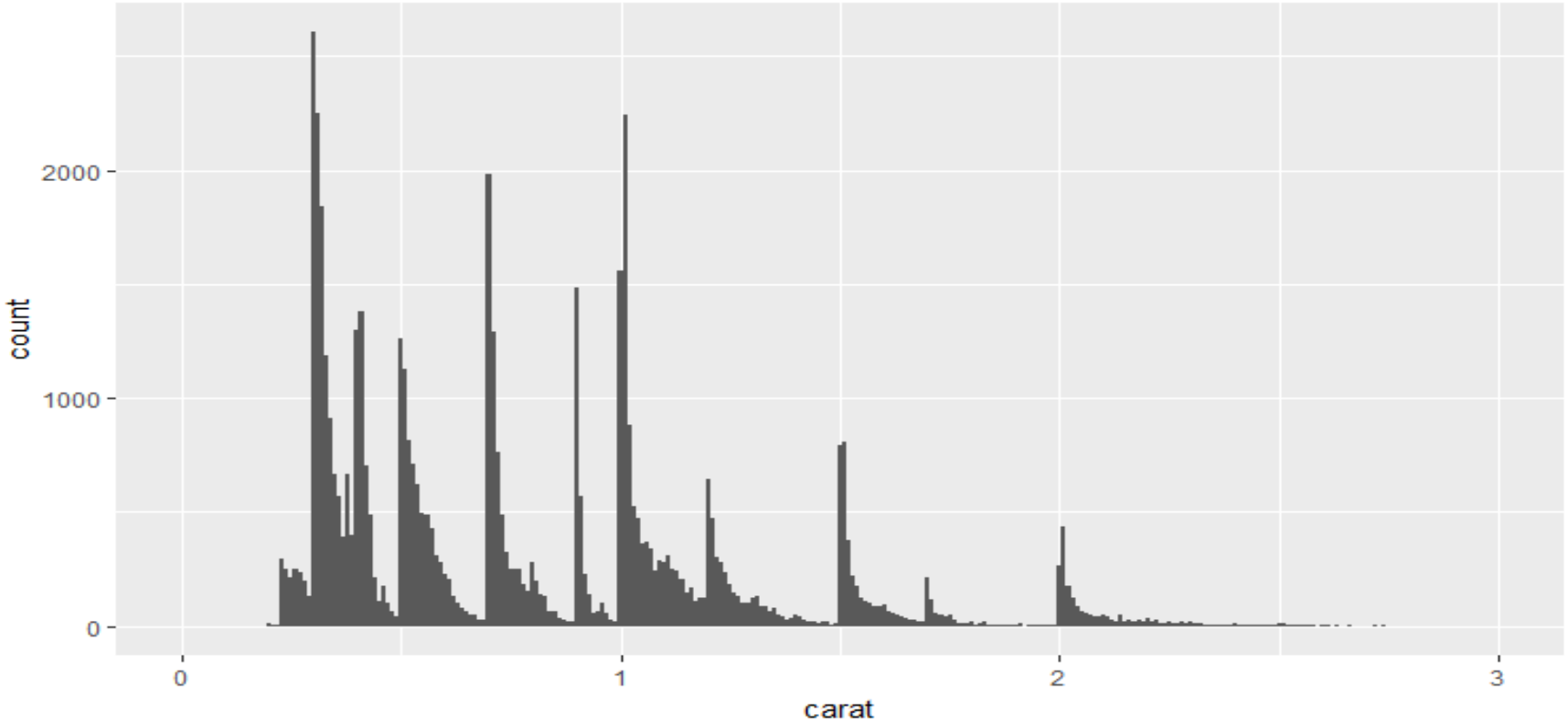
Introduction to qplot

- qplot(carat, data = diamonds, geom = "histogram", binwidth = 0.1, xlim = c(0,3))



Introduction to qplot

- qplot(carat, data = diamonds, geom = "histogram", binwidth = 0.01, xlim = c(0,3))

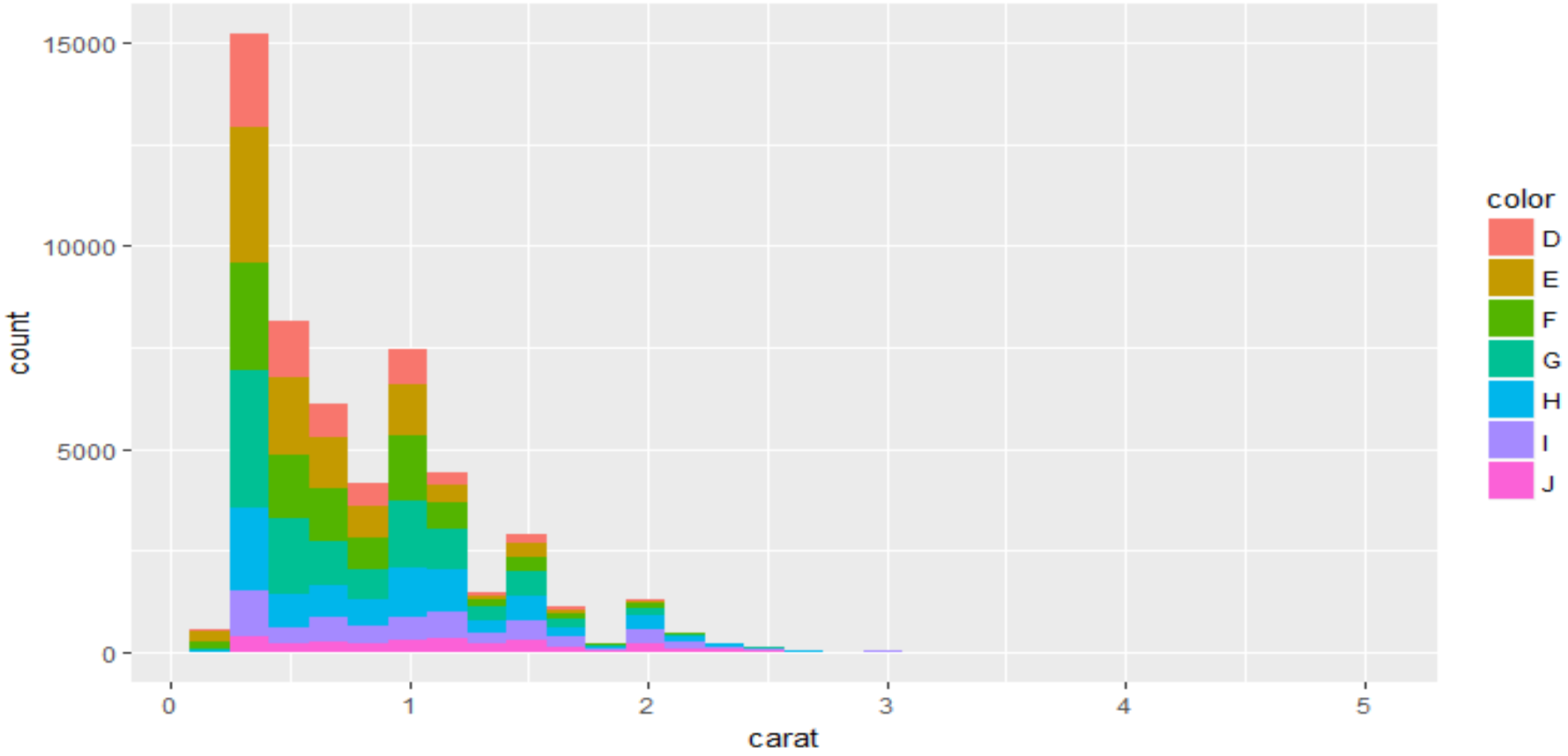


Introduction to qplot

- Histograms:
  - Varying the bin width on a histogram of carat reveals interesting patterns.
  - Only diamonds between 0 and 3 carats shown

Introduction to qplot

- qplot(carat, data = diamonds, geom = "histogram", fill = color)

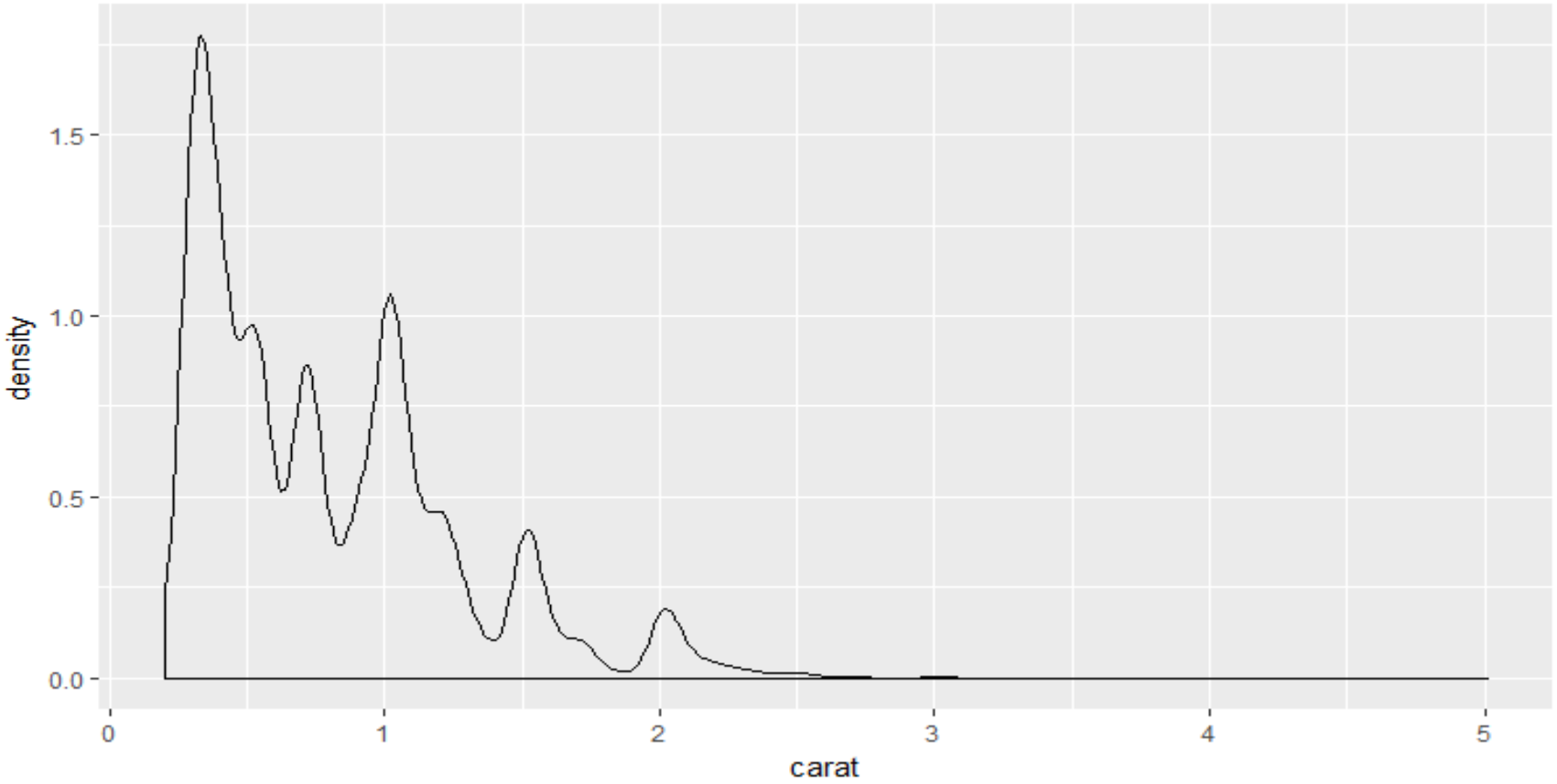


Introduction to qplot

- Density Plots:
  - Distribution of single variable. → like histogram
  - it seems easy to read and compare the various curves.
  - But, it is more difficult to understand exactly what a density plot is showing.
  - Density plot makes some assumptions that may not be true for our data
    - It is unbounded
    - Continuous
    - smooth

Introduction to qplot

- qplot(carat, data = diamonds, geom = "density")



Introduction to qplot

- qplot(carat, data = diamonds, geom = "density", colour = color)