University of Toronto- Time series club
Lecture 1
Data visualization I

Lim, Kyuson

05/18/2022

Today's learning goals

- Apply data transformation and visualization tools to explore the data.

Explore the data

- Ask questions about the data.
- Looking for answers by visualizing, transforming, and modeling your data.
- Refine your questions and/or generate new questions.
- Some type of questions to ask
  - What type of variation occurs within the variables?
  - What type of covariation occurs between the variables?

Grammar of graphics

- Any plot as a combination of a data set, a geom, a set of mappings, a stat, a position adjustment, a coordinate system, and a faceting scheme.
  - Extend the plot by adding one or more additional layer.
  - Start with a dataset and then transform it into the information that you want to display.
- `ggplot2` package in R for graphical data analysis.

Simple plots

- Questions on one variable distribution or frequency distribution or questions on association between two or more variables.
- Lot of information on plots make hard to read/follow.
- Need coordinate system to make plots.

Coordinate system

- Create a coordinate system where we can add layers.
  - We can be explicit about function and corresponding package.

```
ggplot2::ggplot()
```

Coordinate system (cont.)

- If we loaded the package already, we can use the function `ggplot()`.

```
library(ggplot2)
ggplot()
```

Data

- Let's use a dataset.
  - `mpg` dataset in `ggplot2` package.
  - `mpg` dataset contains 234 observations collected by the US Environmental Protection Agency on 38 models of cars.
- Load some packages for data transformation.

```
library(magrittr)
library(dplyr)
```

- Load the data into the working environment.

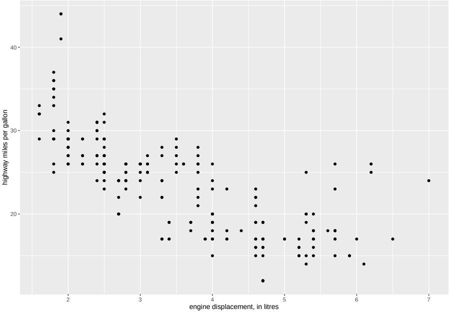
```
data(mpg)
```

- Open help page in RStudio. Read the description about the data.
- Or type in console `?mpg` to open help page.

Scatter plots

- Question: do cars with big engines use more fuel than cars with small engines?
  - Visualization method - scatter plots.

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  ylab("highway miles per gallon") +
  xlab("engine displacement, in litres")
```

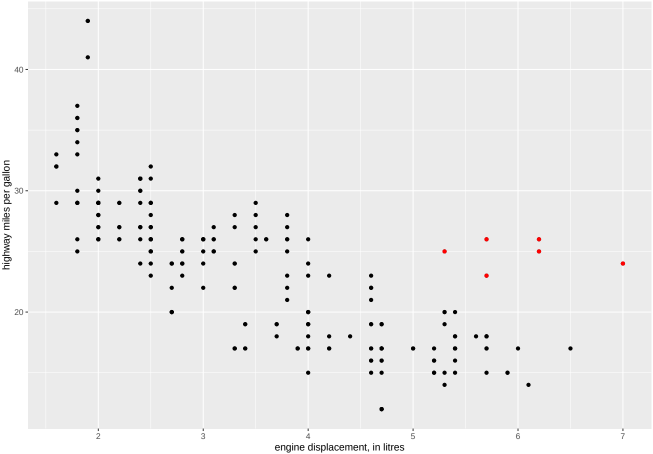


- Interpret the scatter plot.
  - What kind of association, any outliers, what is the range of variables?

Cars with big engines use more fuel.

Scatter plot with transformation

- Question: One group of points (highlighted in red) seems to fall outside of the linear trend. How can we explain these cars? * Color the points (cars) corresponding to `displ > 5 & hwy > 21`.

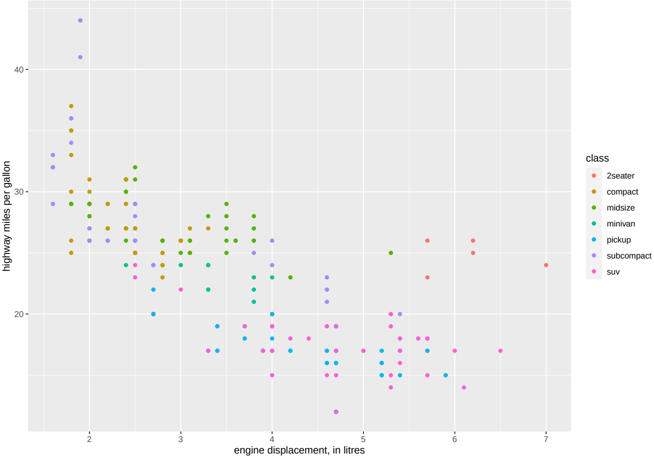


- One group of points (highlighted in red) seems to fall outside of the linear trend. How can we explain these cars?
  - Interpret the plot?
    - hybrids?
    - No answer?

- Question: how association between `displ` and `hwy` within each class of vehicle.
    - Add aesthetic - visual property of the objects in your plot (size, the shape, or the color).

Aesthetic

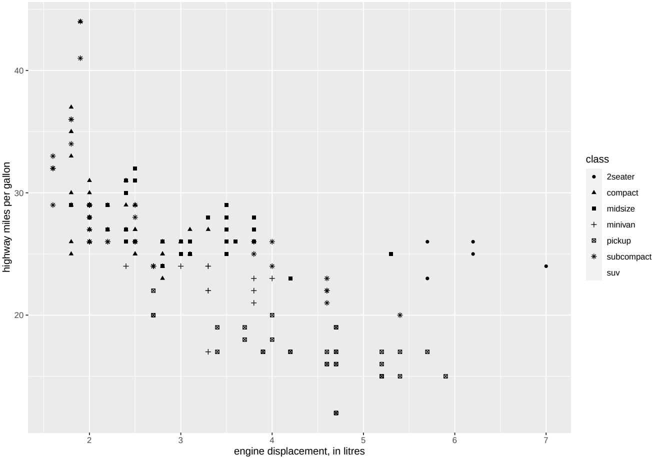
- Color points in the scatter plot by `class` variable.



- Interpret the plot?

- Unusual points are two-seater cars.
  - Hybrids? (unlikely because they have large engines).
  - Sport cars? (large engines, but small bodies).
- Not only color attribute, we can use shape as well
  - Shape points by `class` variable.
  - `class` - type of car.

- Exercise: Look at the plot. What is wrong?

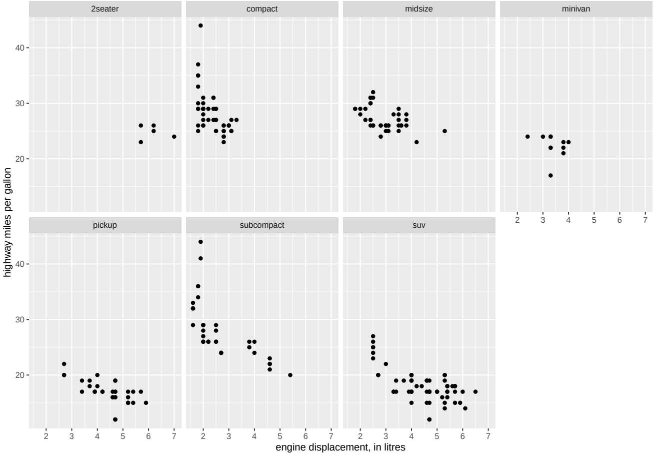


- Exercise - solution
  - We didn't have enough different shapes to refer to different classes.
  - So we may need to specify it.

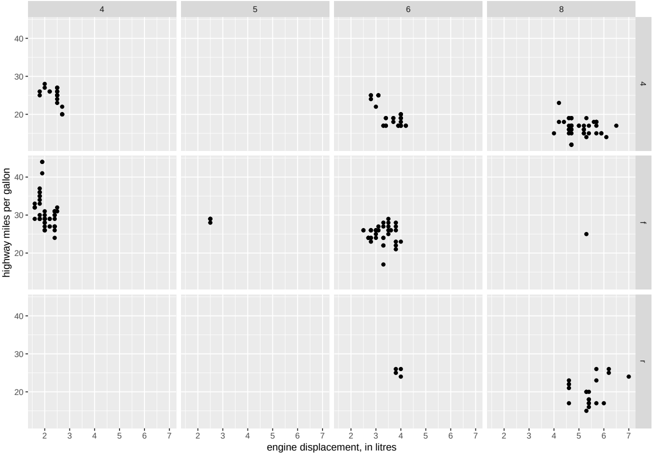
  By default, additional groups will go unplotted when you use the shape aesthetic.

Facets

- Question: how association between `displ` and `hwy` within each class of vehicle.
  - Facet the scatter plot by `class` variable.

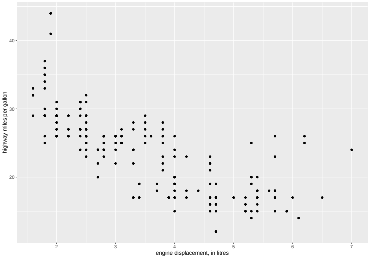


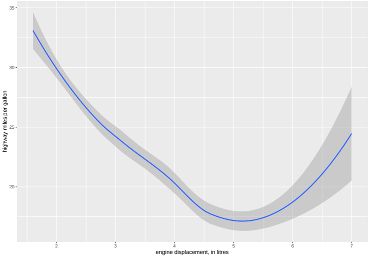
- Try to facet the plot on the combination of two variables (`drv` and `cyl`).
  - `drv` - the type of drive train, where f = front-wheel drive, r = rear wheel drive, 4 = 4wd.
  - cyl - number of cylinders.



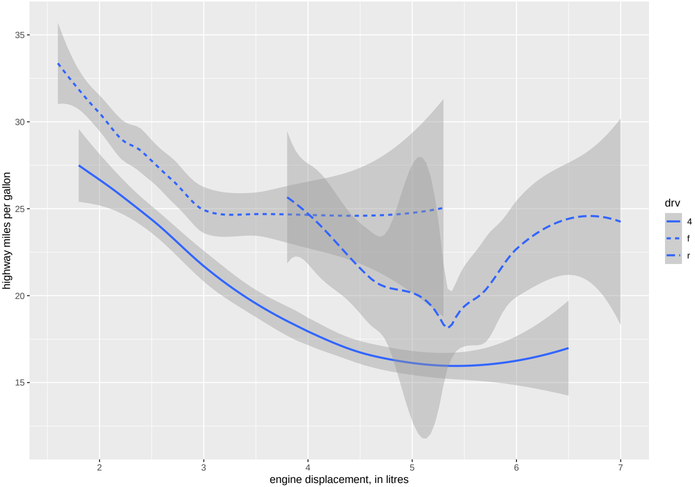
Geometric objects

- Use different geoms.
  - Draw a smooth line.



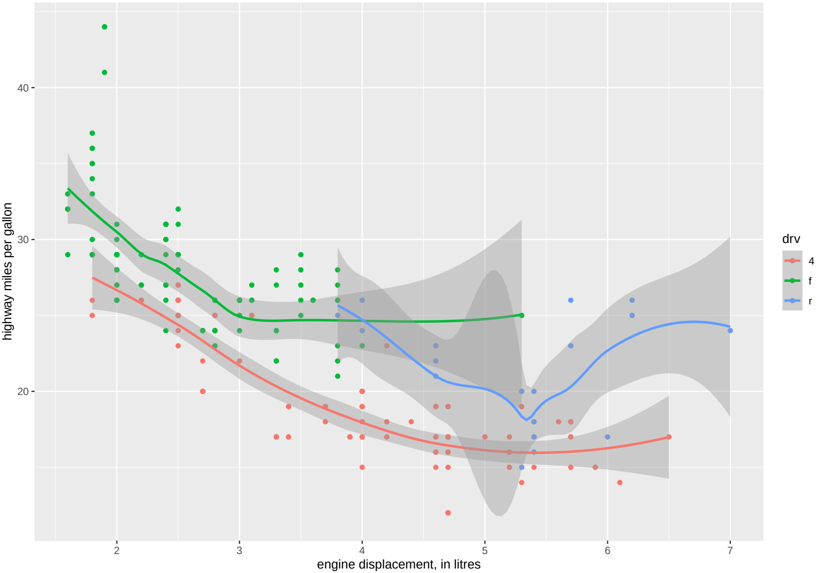


- Not every aesthetic works with every geom.
  - linetype by `drv` - the type of drive train,



This separates the cars into three lines based on their `drv` value.

▶ Two geoms in one plot.



To understand variation of categorical variable?

- Consider another dataset from ggplot package.

```
?diamonds
```

carat	cut	color	clarity	depth	table	price	x
0.23	Ideal	E	SI2	61.5	55	326	3.95
0.21	Premium	E	SI1	59.8	61	326	3.89
0.23	Good	E	VS1	56.9	65	327	4.05
0.29	Premium	I	VS2	62.4	58	334	4.20
0.31	Good	J	SI2	63.3	58	335	4.34
0.24	Very Good	J	VVS2	62.8	57	336	3.94

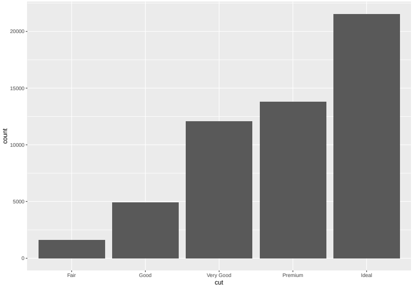
`diamonds` data set is in `ggplot2` package. This dataset contains prices and attributes of 53940 diamonds.

- What is the distribution of `cut` variable?

To understand variation of categorical variable - barplot

- Barplot - height of the bars displays how many observations occurred with each value.
- cut is a categorical variable (in R, factor or character).

```
ggplot(data = diamonds) +
 geom_bar(mapping = aes(x = cut))
```



Statistical transformations

- How barplot is created?
  - Make frequency chart, then plot frequency bar plot.
  - Make relative frequency chart, then plot relative frequency bar plot.

```
(fre_table <- diamonds %>%
   group_by(cut) %>%
   summarise(n = n()))
```

```
## # A tibble: 5 x 2
##   cut           n
##   <ord>     <int>
## 1 Fair       1610
## 2 Good       4906
## 3 Very Good 12082
## 4 Premium   13791
## 5 Ideal     21551
```

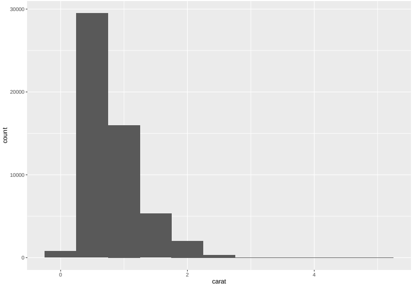
To understand variation of continuous variable

- What is the distribution of `carat` variable?

To understand variation of continuous variable - histogram

- Histogram - divides the x-axis into equally spaced bins and then uses the height of a bar to display the number of observations that fall in each bin.

```
ggplot(data = diamonds) +
  geom_histogram(mapping = aes(x = carat),
                 binwidth = 0.5)
```



- Histogram - sensitive to width of the intervals (in R, `binwidth`).
- Do we see outliers?
  - Use `carat < 3`.

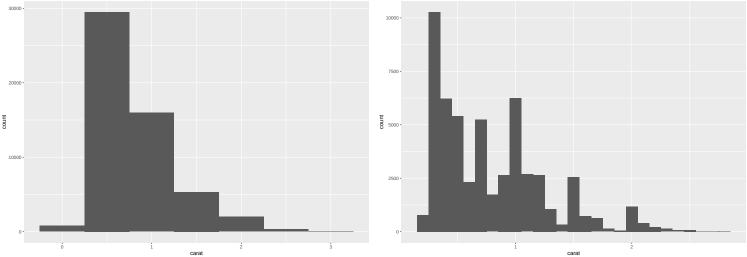


Figure 1: binwidth = 0.5 and binwidth = 0.1

- A Simply Statistics blog by Jeff Leek, Roger Peng, and Rafa Irizarry on evidence-based data analysis point out the methods in R for computing number of bins.
  - `R` uses Sturges' formula[1] to find a bin width.
  - David Scott[2] derive integrated mean squared error-based optimal histogram bin width.

[1] The Choice of a Class Interval [@sturges1926choice]

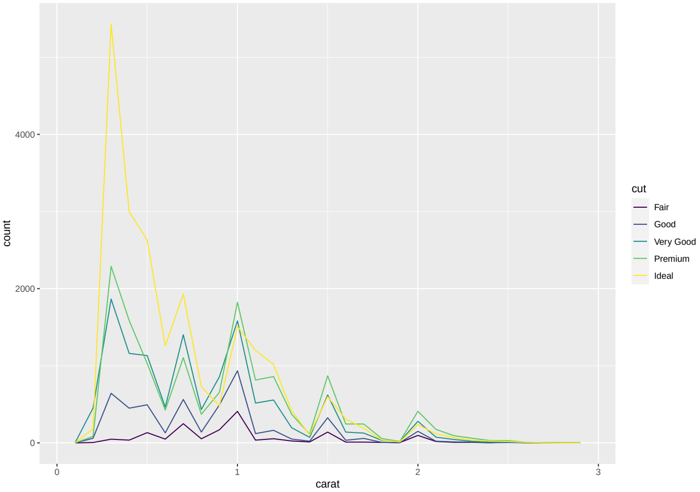
[2] On Optimal and Data-Based Histograms [@scott1979optimal]

Multiple histograms in the same plot

- What is the distribution of `carat` within each `cut` type of diamond?

Multiple histograms in the same plot - Frequency polygone

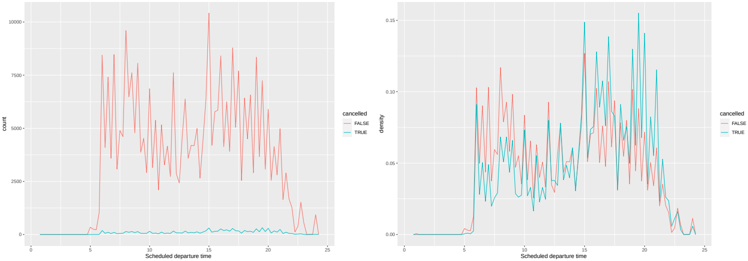
- Displaying the counts with lines.
  - `geom_freqpoly()`
- Display frequency polygon of `carat` for different cuts of diamonds.
  - Use color for different cut of diamonds


count
4000
2000
0
0
1
2
3
carat
cut
Fair
Good
Very Good
Premium
Ideal

Missing values

- If there is an unusual value
  - Drop the entire row (not recommended).
  - Replace the value with NA.
- `ggplot2` will warn you about missing values (can suppress using na.rm = TRUE).
- Sometimes `NA` has meaning
  - In `flights` data, `dept_time` is `NA` if the flight was cancelled.
  - Create a new variable from it.
  - Then, plot cancelled and not-cancelled by time.

Example - frequency polygone with missing vlaues



- Caution: one category has more count than other. Hard to compare it.
- Resolve: Use density plot (right plot) - count standardized so that the area under each frequency polygon is one.

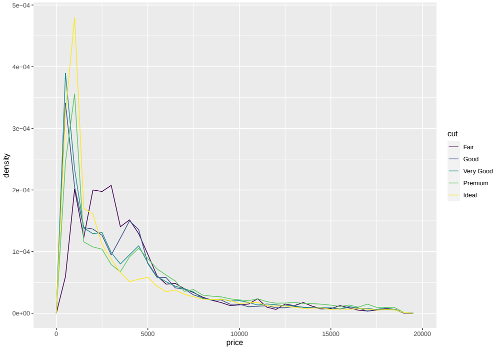
Covariation

- Some questions to ask.
  - Could this pattern be due by random chance?
  - How can we describe the relationship implied by the pattern?
  - How strong is the relationship implied by the pattern?
  - What other variables might affect the relationship?
  - Does the relationship change if we look at individual subgroups of the data?

A categorical and continuous variable

- What is the diamond price distirbution by cut type?

A categorical and continuous variable - frequency polygon

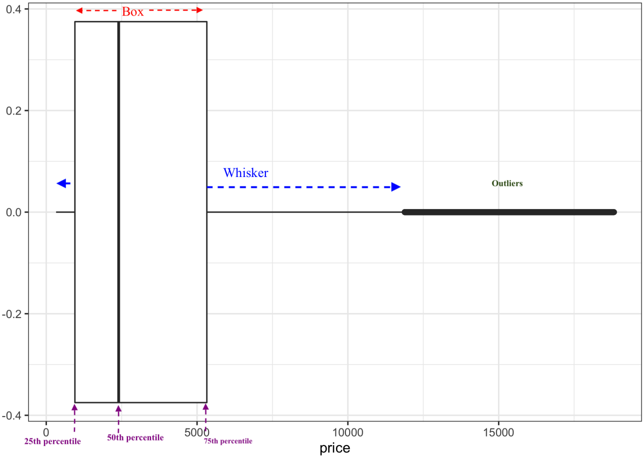


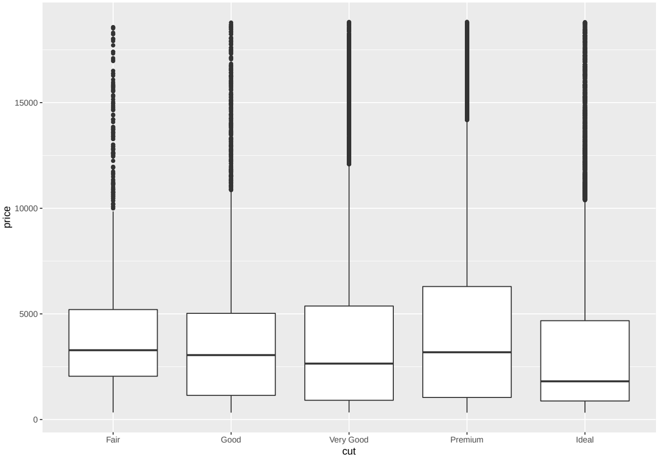
Fair diamonds (the lowest quality) have the highest average price.

- Little hard to interpret - shape depends on the binwidth.
  - Any other tools?

A categorical and continuous variable - boxplot

- A box, Whisker, 25th percentile, 50th percentile, 75th percentile, outliers

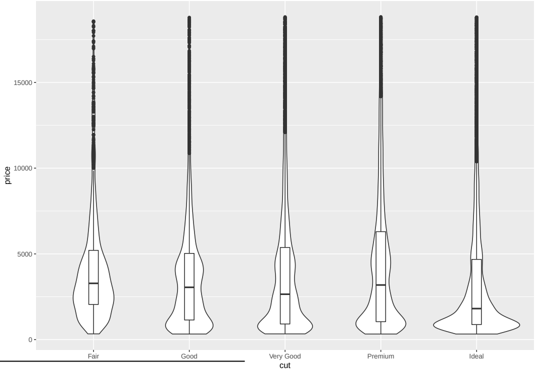




Counter intuitive finding that better quality diamonds are cheaper on average.
Is there any other variables determine the price?

A categorical and continuous variable - violine plot

- Violin plot - similar to a box plot, includes rotated kernel density plot.
  - Kernel density plot - nonparametric way of estimating the density of a random variable[3]
  - Violin plot - can add a marker for the median and a box or marker showing the interquartile range



[3] A reference to my lecture slides

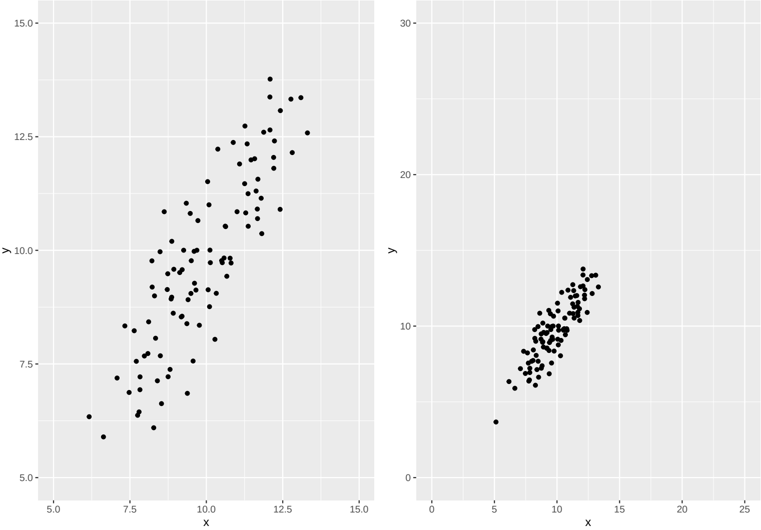
Human interpretation about patterns

- `Magical thinking`[4]: natural human inclination to over-interpret connections between random events .

[4] Theories of data analysis: From magical thinking through classical statistics [@thinking1985]

Two continuous variables - scatterplot

- Which plot shows the higher degree of association? (left/right)



Scatterplot

- Scale does change the perception of a viewer.

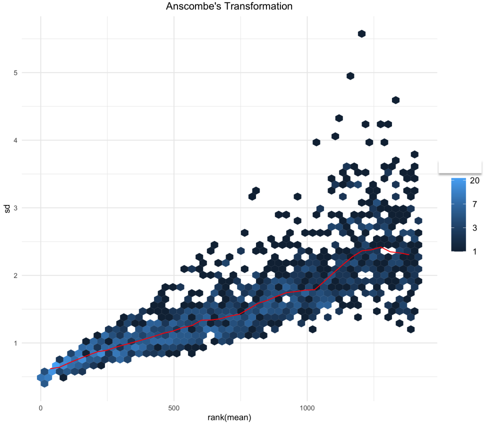
Two continuous variables - scatterplot

- Concern: overplot for large dataset .
  - For example - microbiome data, neuro data (MRI), spatial transcriptomics data.
- Resolve
  - Add alpha aesthetic to the plot, but transparency can be challenging for very large datasets.

Two continuous variables (large dataset)

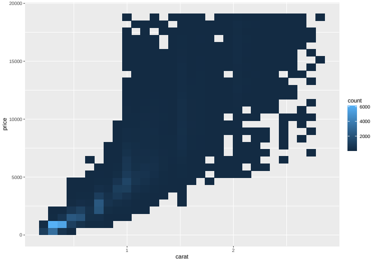
Two continuous variables (large dataset) - hex plots

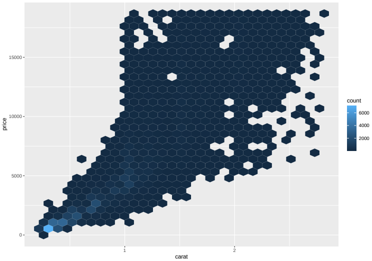
- Association between mean and standard deviation of biomarkers in high-throughput data (thousands or millions of variables).
  - Use `geom_hex()`.



Two continuous variables - geom_bin2d() and geom_hex()

- Divide the coordinate plane into 2d bins.
- Use a fill color to display how many points fall into each bin.

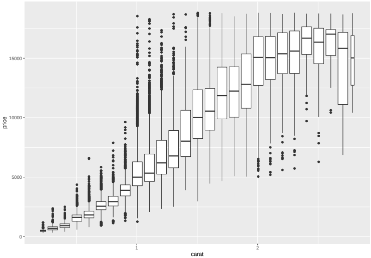


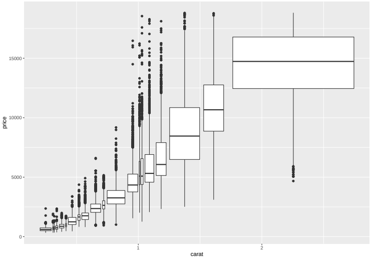


Two continuous variables - bin one continuous variable

- Now use the techniques for one continuous and categorical.
- For example, bin `carat` (by width 0.1), then boxplot of price for each bin.
  - Difficult to tell that each boxplot summarizes a different number of points.
  - Set the width of the boxplot proportional to the number of points with `varwidth = TRUE`.

Two continuous variables - bin one continuous variable



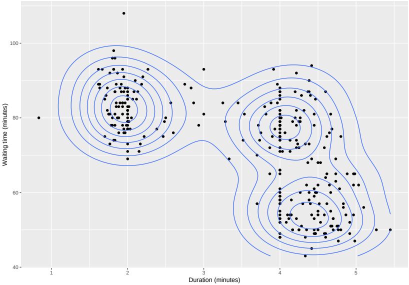


To understand variation of two continuous variables (any clusters?) - contour plot

- The geyser data set contains a total of 299 observations of eruption duration (in minutes) and waiting time (in minutes, for this eruption) for the Old Faithful geyser[5].
- Available as geyser in MASS package.

[5] A look at some data on the Old Faithful geyser [@azzalini1990look]

Contour plot - identify patterns



- Observation - three clusters.
- Suppose variation increases uncertainty, covariation reduces it - so we can use one to predict the other.
- Causal - if the covariation is due to a causal relationship (a special case), then we can use the value of one variable to control the value of the second.
  - Example, design of experiment.

Next

- See some R codes

Next lecture

- More on data visualization (categorical data, interactive plots, Shiny, networks, word cloud)

References