Unit - 5

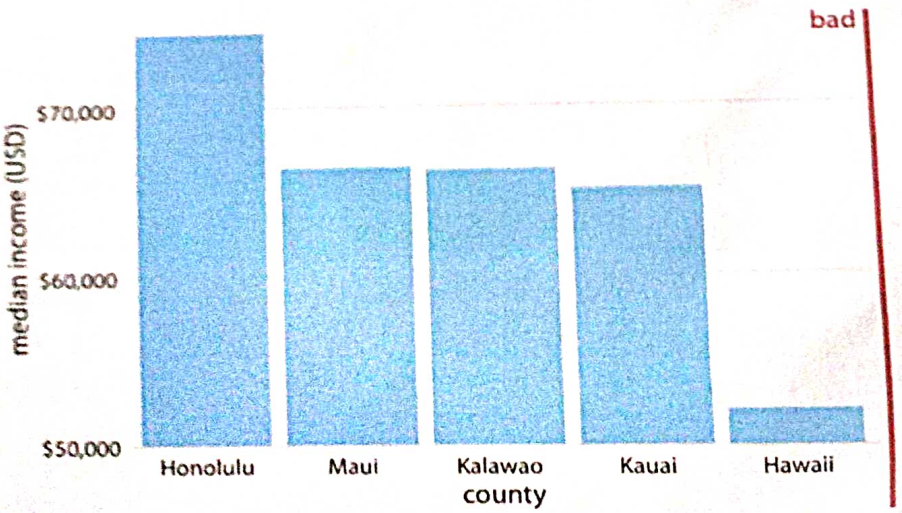
The principle of proportional ink

The principle of proportional ink is defined as the sizes of shaded area in a visualization need to be proportional to the data values they represent.

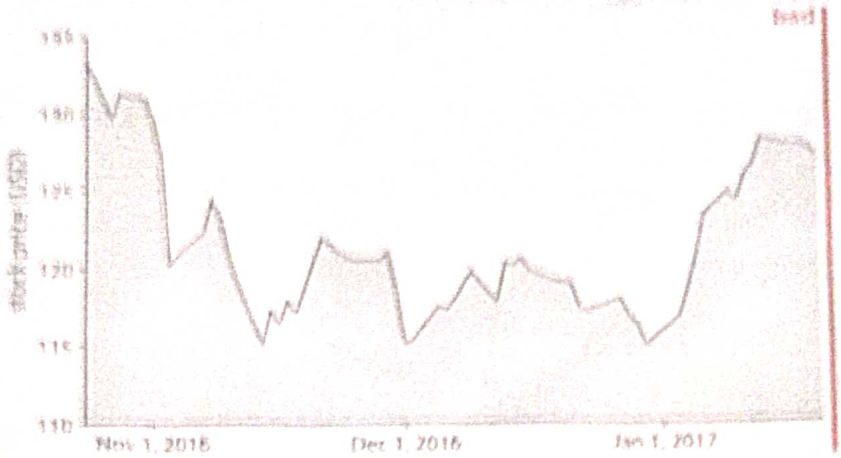
Visualizations along linear axes :

Let us consider the most common scenario visualization of amounts along a linear scale.

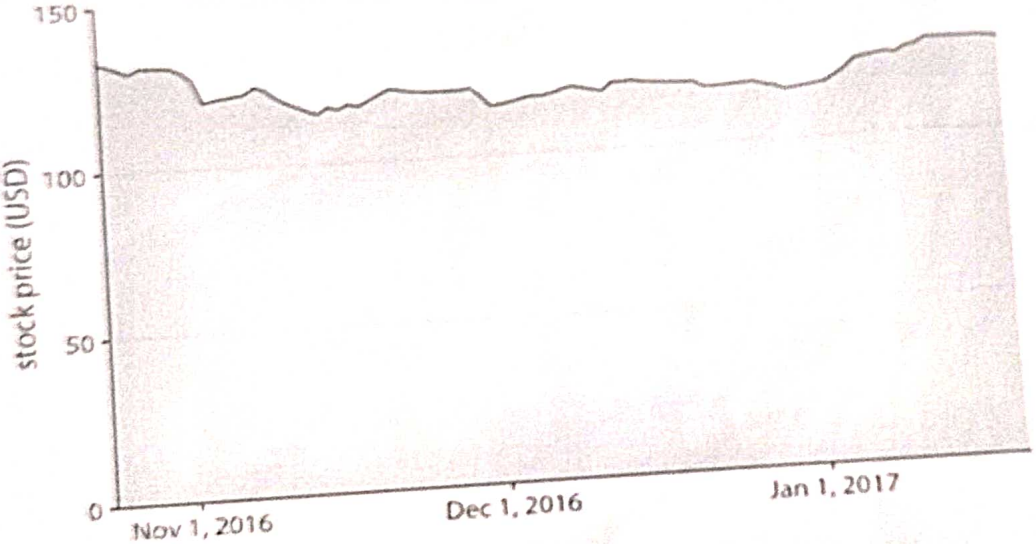
Median income in the five countries of the state of Hawaii. The figure is misleading because the y-axis scale starts at $50,000 instead of $0. As a result the bar heights are not proportional to the values shown and the income differential between the country of Hawaii and the other four countries appears much bigger than it actually is.



However this is misleading because the y-axis starts at $110 instead of $0.



The following fig is more accurately reflects the magnitude of the FB price drop because the y-axis starts at $0.

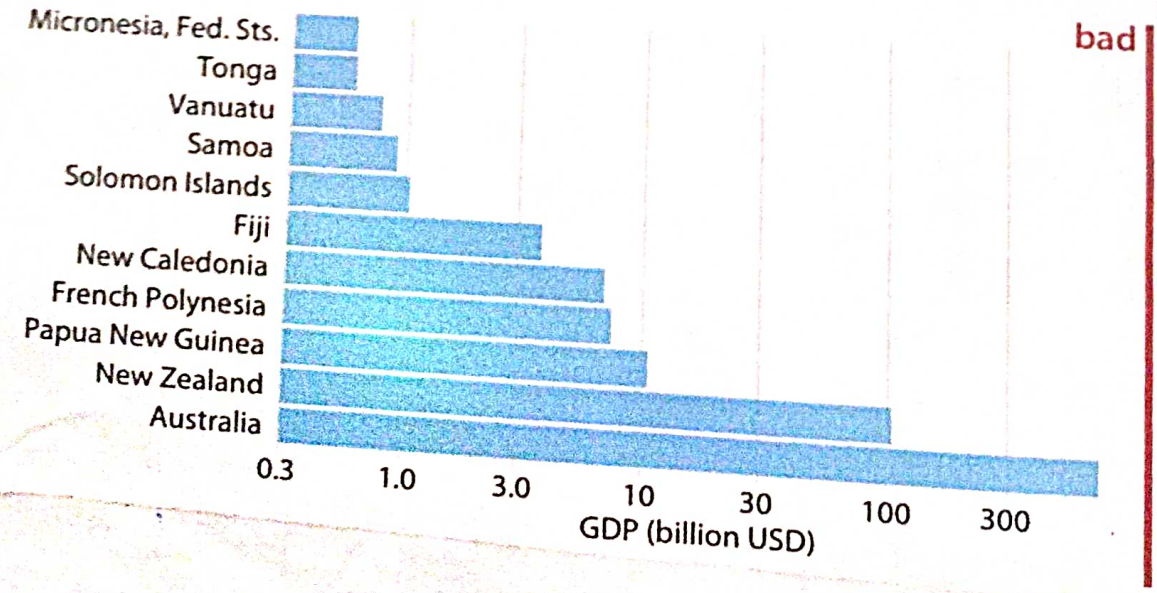


and shaded areas are not useful to represent small changes overtime or difference between conditions, since we always have to draw the whole bar or area starting from zero '0'.

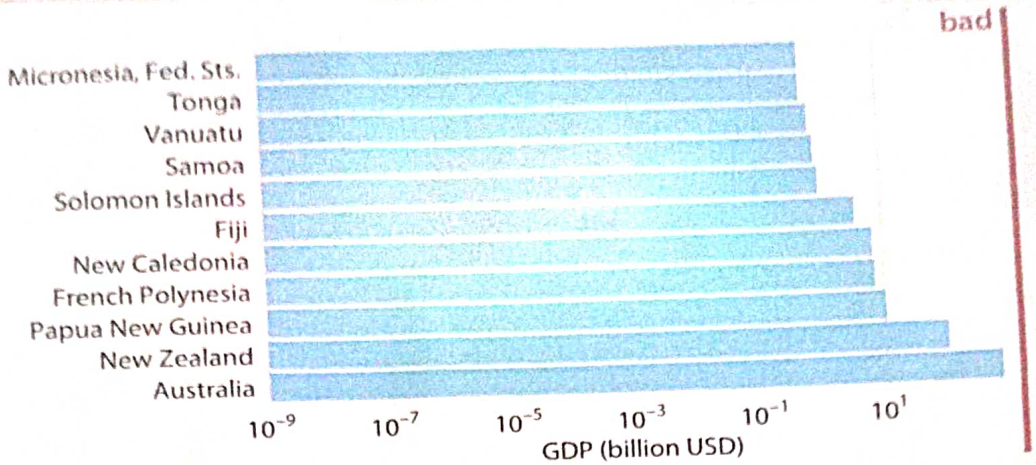
Visualizing along logarithmic axes

A log scale is the natural scale to visualize ratios, because a unit step along a log scale corresponds to multiplication or division by a constant factor. The area of each bar will be proportional to the logarithm of the data value and thus bar graphs on a log scale satisfy the principle of proportional ink in log-transformed coordinates.

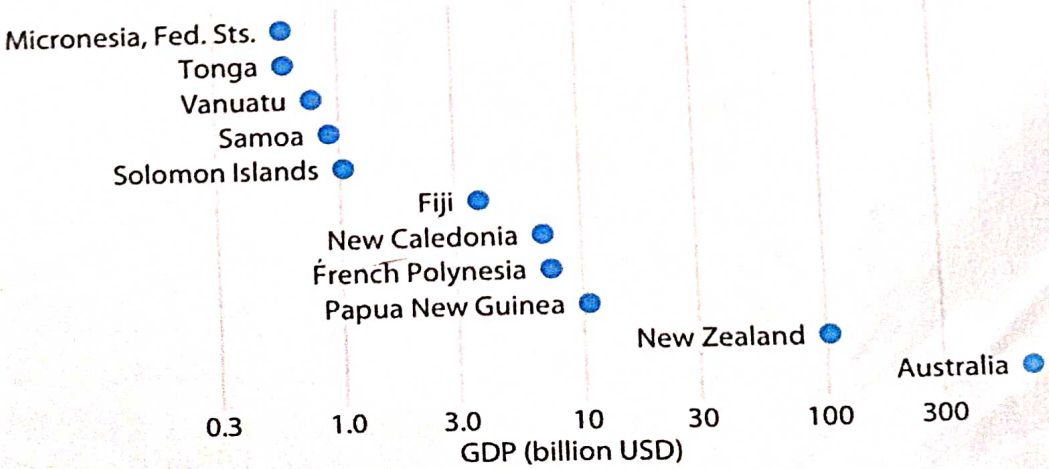
Let us consider the example of GDPs of countries in oceania. In 2007, these varied from less than a billion US dollars to over 300 billion USD. Visualizing these numbers on a linear scale could not work, because the two countries with the biggest GDPs would dominate the figure.



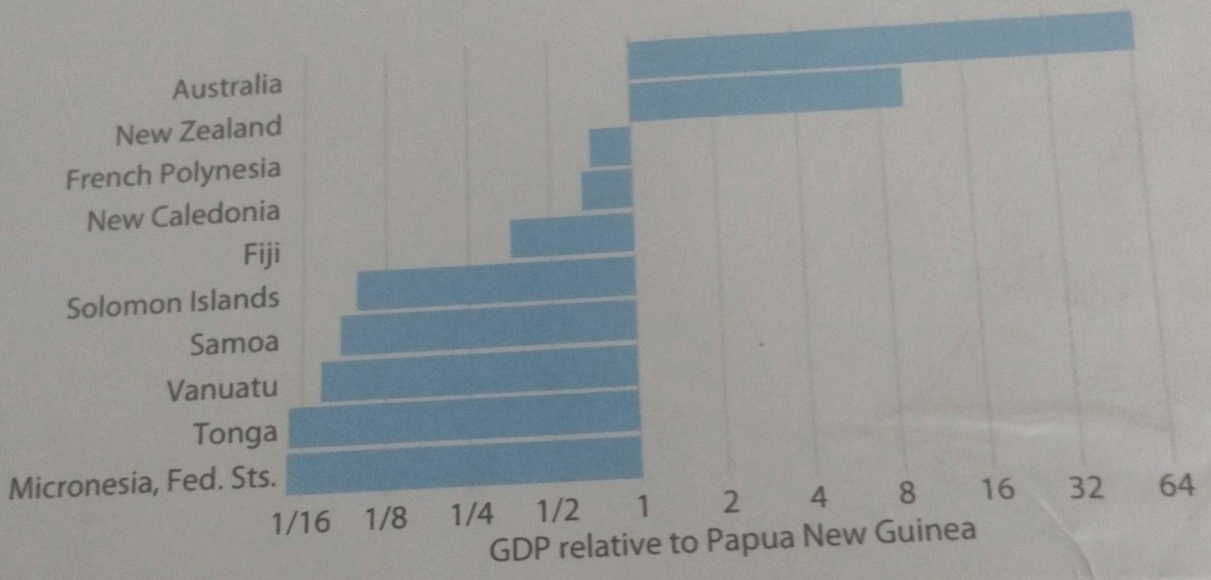
However, the visualization with bars on a log scale does not work either. The bar starts out at an arbitrary value of 0.3 billion USD. This problem always arises when we try to visualize amounts on a log scale.



By placing the country names right next to the dots rather than along the y-axis we avoid generating the visual perception of a magnitude conveyed by the distance from the country name to the dot.

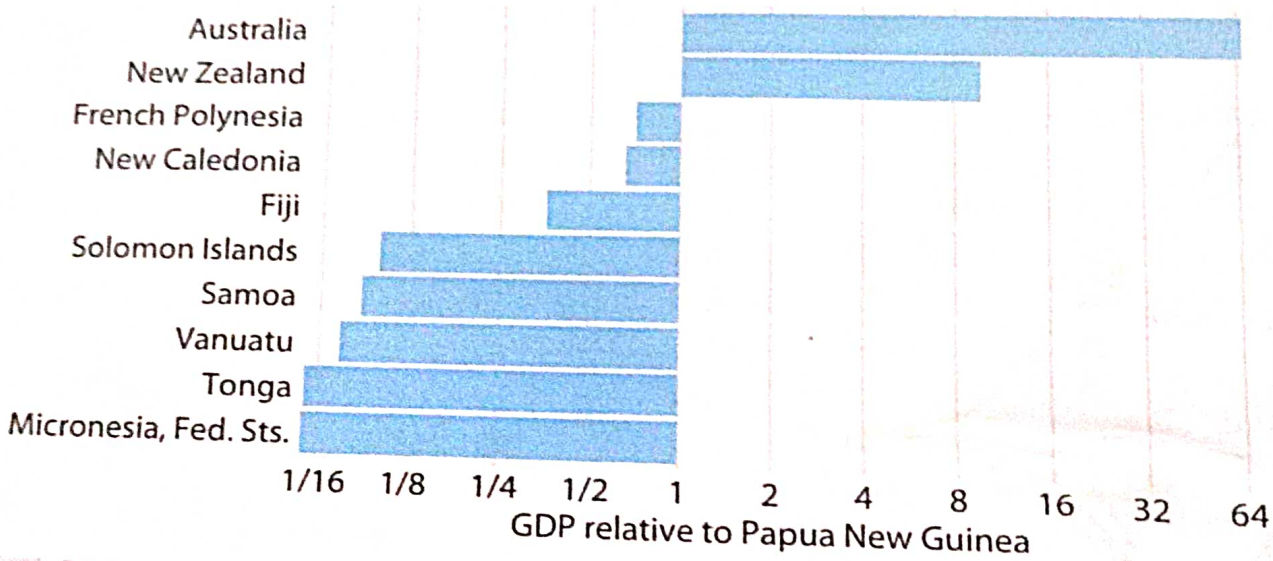


If we want to visualize ratios rather than amounts, however, bars on a log scale are a perfectly good option. In the below figure highlights that the natural midpoint of a log scale is 1, with bars representing numbers above 1 going in one direction and bars representing numbers below one going in the other direction.



When bars are given on a log scale they represent ratio and need to be drawn starting from 1 not 0

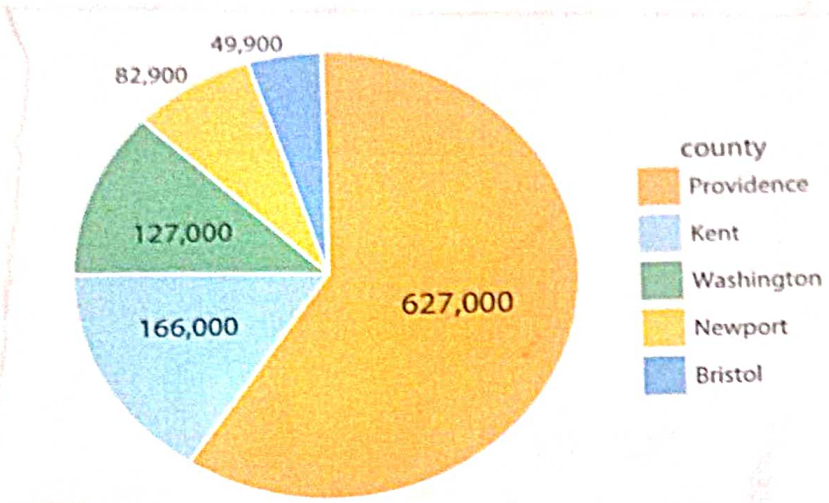
If we want to visualize ratios rather than amounts, however, bars on a log scale are a perfectly good option. In the below figure highlights that the natural midpoint of a log scale is 1, with bar representing numbers above 1 going in one direction and bars representing numbers below one going in the other direction.



When bars are given direction on a log scale they represent ratio and need to be drawn starting from 1 not 0

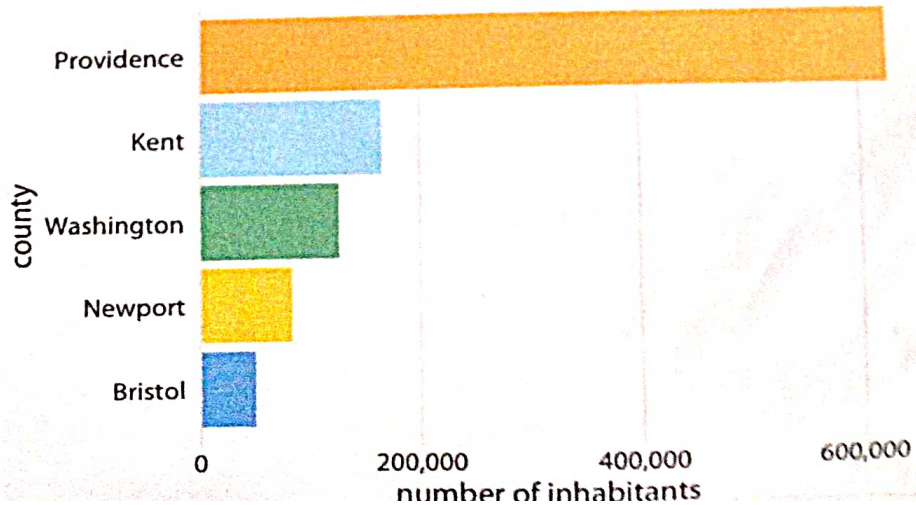
Direct area visualizations

The most common one is the piechart Even though technically the data values values are mapped onto angles, which are represented by location along a circular axis. Instead, the dominant visual property we notice is the size of the areas of each pie wedge.

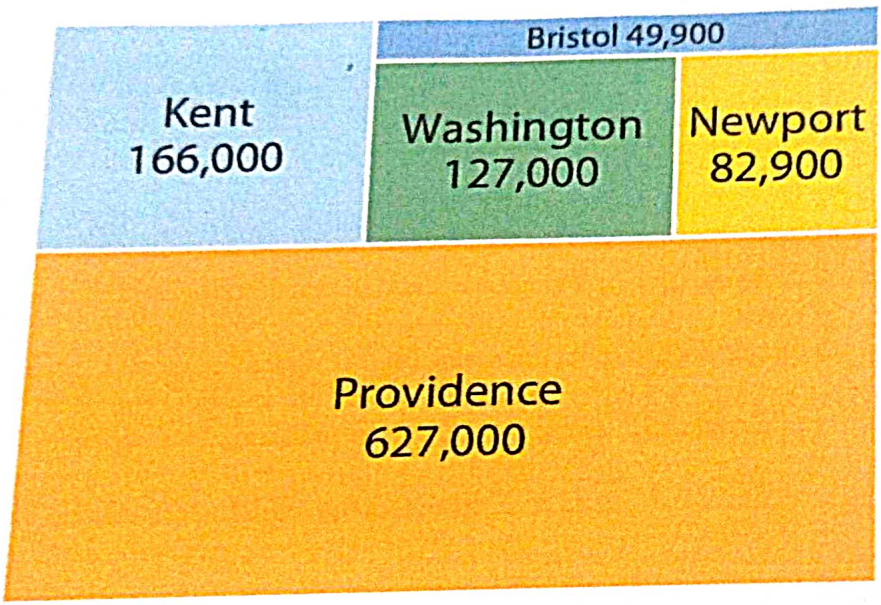


B

Because the area of each pie wedge is proportional to its angle which is proportional to the data value the wedge represent, pie charts satisfy the principle of proportional ink. We perceive the area in a piechart differently from the same area in a boxplot.



The problem that human perception is better at judging distance than at judging areas also arises in treemaps, which can be thought of as a square version of pie charts.



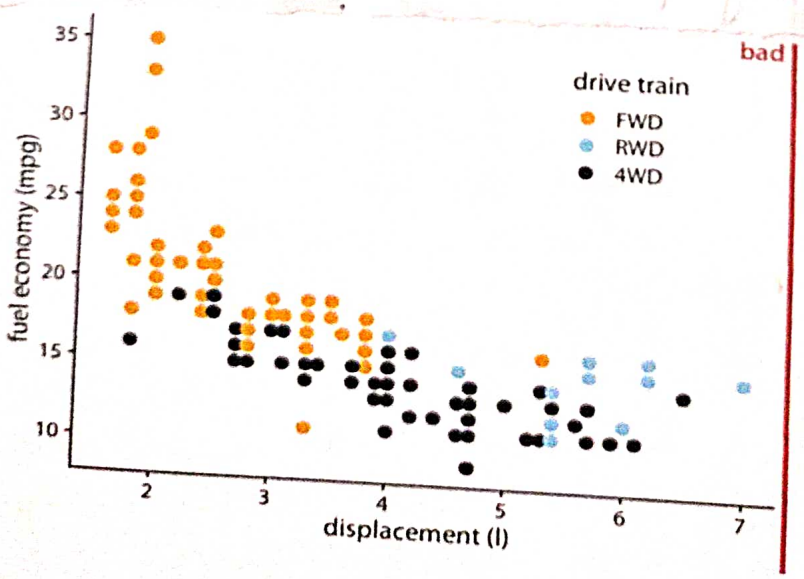
Handling overlapping points

When the multiple observations have exactly the same numeric values the technical term used to describe this situation is "overplotting" i.e. plotting many points on top of each other. The strategies used when we encounter this situation is

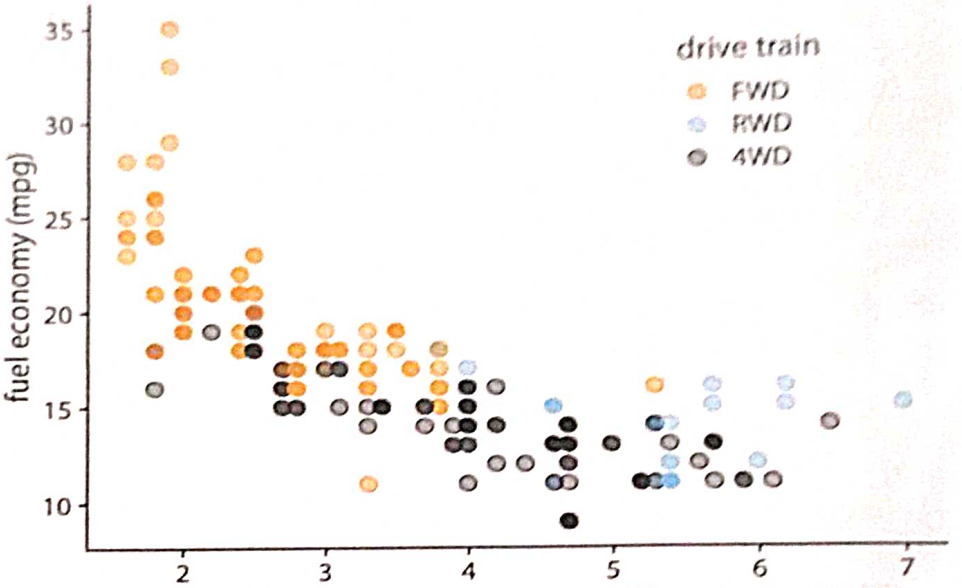
1. partial transparency and jittering
2. 2D histograms
3. contour lines.

Partial transparency and Jittering

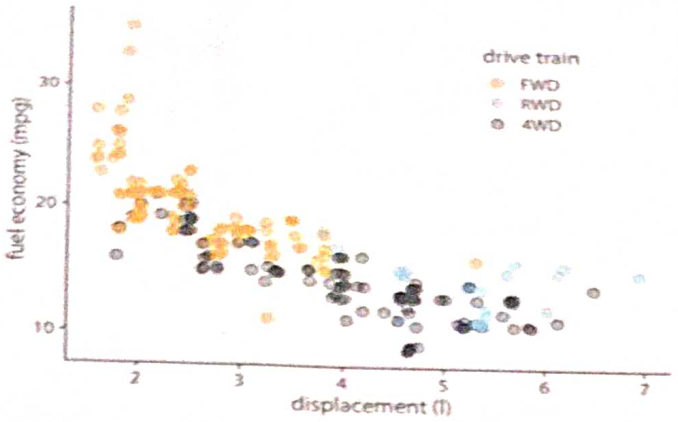
our dataset is city fuel economy versus engine displacement for popular cars released between 1999 and 2008. Each point represent one car. The point color encodes the drive train ie front-wheel drive (FWD), rear-wheel drive (RWD) or four-wheel drive (4WD). The figure is labeled "bad" because many points are plotted on top of others and obscure them.



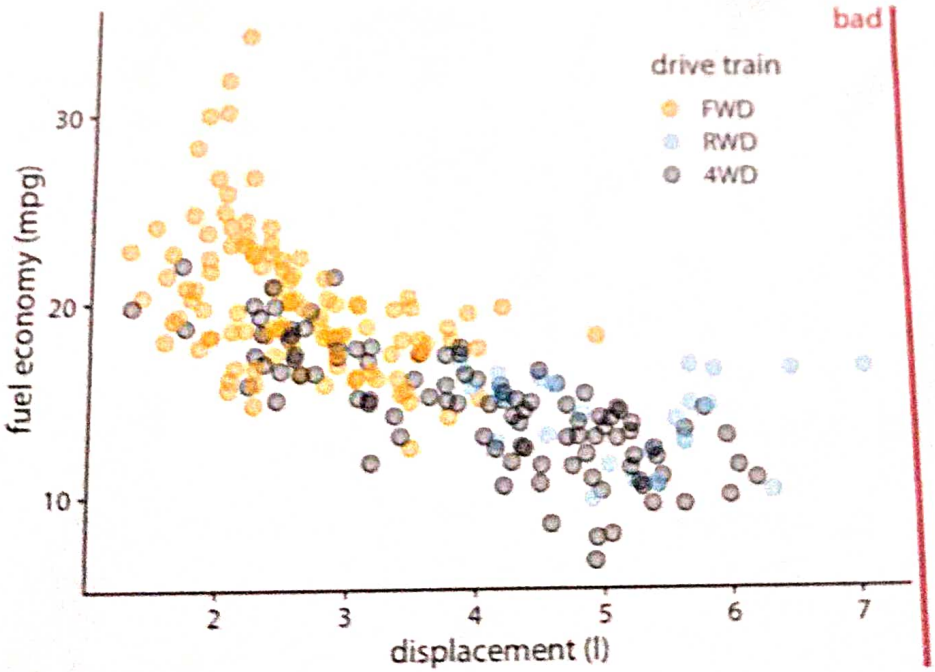
one way to overcome this problem is to use partial transparency. If we make individual points partially transparent then overplotted points appear as darker points and thus the shade of the point reflects the density of points in that location of the graph.



Making points partially transparent is not always sufficient to solve the issue of overplotting; it is difficult to estimate how many points were plotted on top of each other in each location. A simple trick that helps in this situation is to apply a small amount of jitter to the points, i.e. to displace each point randomly by a small amount in either the x or the y direction or both.



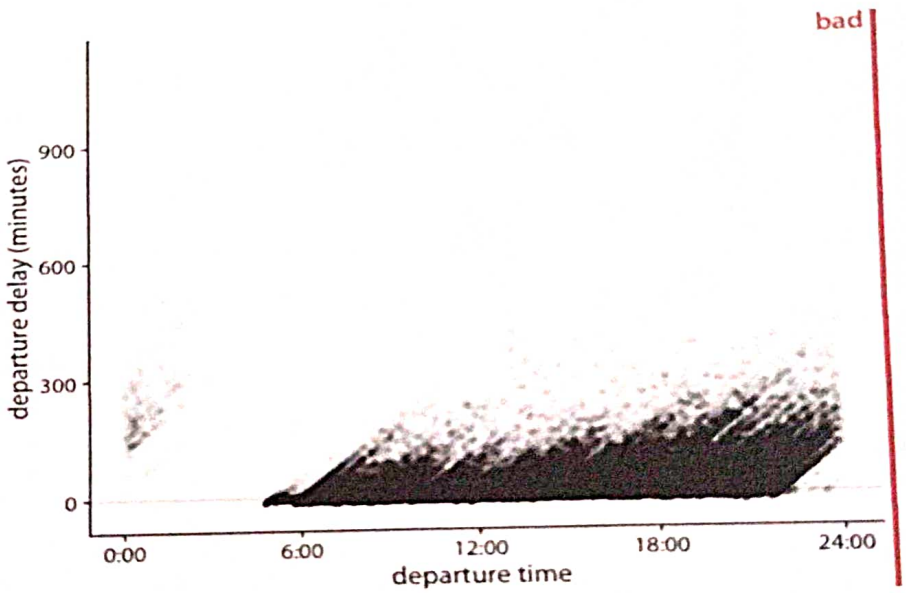
If we jitter too much, we end up placing points in locations that are not ~~representative~~ of the underlying dataset. The result is a misleading visualization of the data.



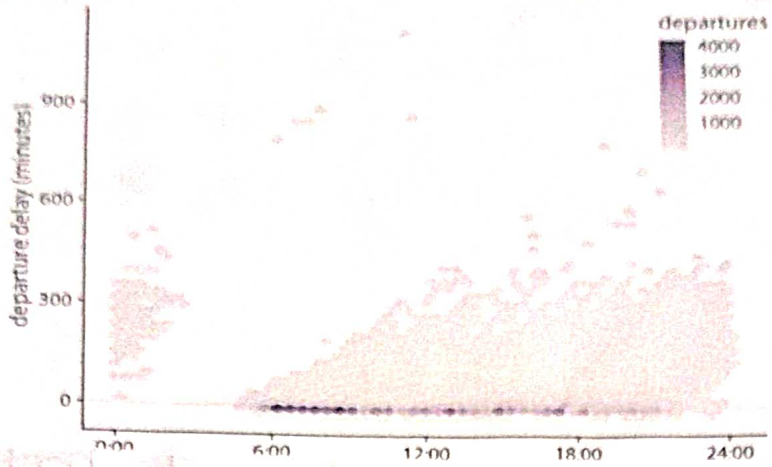
2D histograms

When the number of individual points gets very large, partial transparency with or without jittering will not be sufficient to resolve the overlapping issue.

Let us consider the dataset depar delay in minutes versus the flight departure time, for the all flights depar Newyork airport in 2013. Each dot repres one departure.

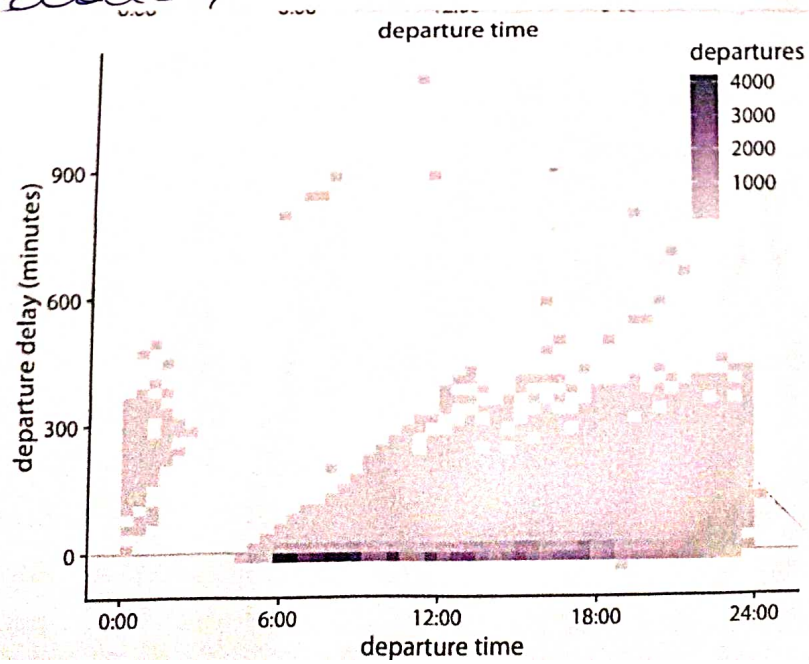


It is not visible properly. In such cases, instead of plotting individual points, we can make a 2D histogram. We subdivide the entire x-y plane into small rectangles, count how many observations fall into each one and then color the rectangles by that count.



This visualization already highlights several important features of the flight departure data. As an alternative to binning the data into rectangles we can also bin into hexagons. The advantage that the points in a hexagon are, on average closer to the hexagon center than the points in a equal-area square are to the center of the square.

∴ the hexagon represents the data slightly more accurately than the rectangle does.

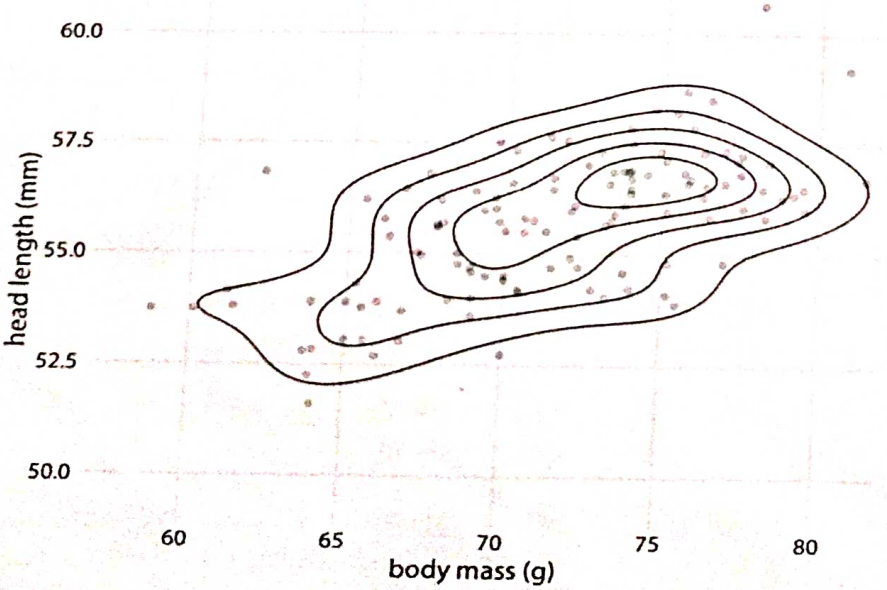


Contous Lines

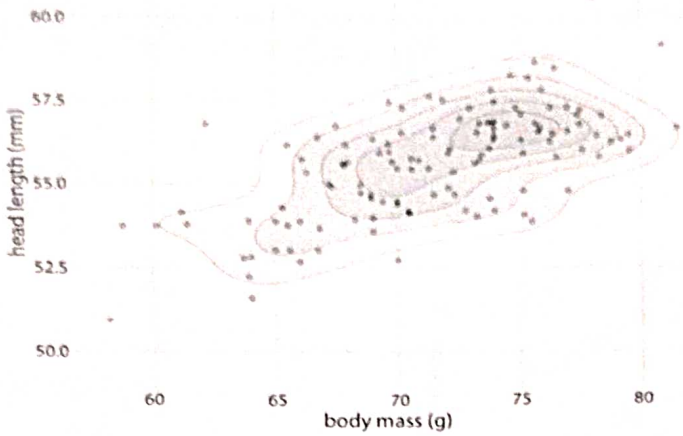
Instead of binning data points into rectangles as hexagons, we can also estimate the point density across the plot area and indicate regions of diffe -rent point densities with contour lines -This technique works well when the point density changes steady across th both the x and the y dimensions.

For this approach consider the data blue jays. To showcase the relationship between head length and body mass for 123 blue jays and there was some amount of overlap among the points.

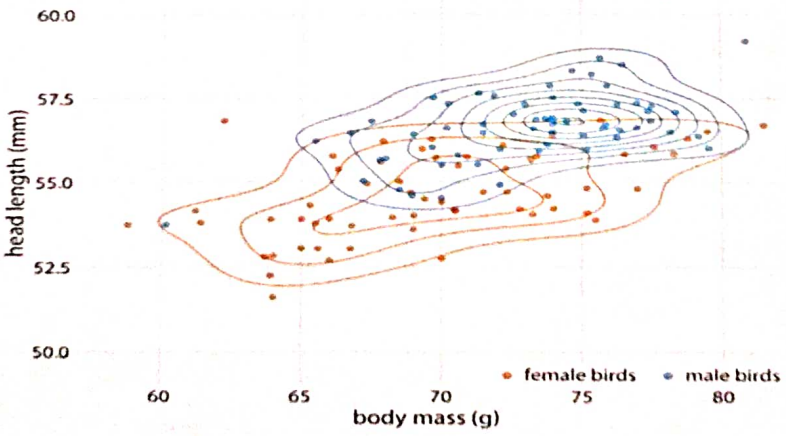
We can highlight the distribution of points more clearly by making the points smaller and partially transparent and plotting them on top of contous lines that delineate regions of similar point density



We can further enhance the perception of changes in the point density by shading the regions enclosed by the contour lines, using darker colors for regions representing higher point densities.



We can do the same with color contour lines, by drawing separately colored contour lines for male and female birds.



Assignment

1) Define principle of proportional line? Explain visualization along linear axis?

2) Explain visualization along logarithmic axes

3) Explain the method of transparency and filtering.

4) Explain 2D histograms & contour lines?