DATA.STAT.770

Exercise set 5

2.

Numerical effect size: $\frac{50-6}{6} \approx 7.33$

Assume that the height of pillar 6 equals 1 and width equals $\frac{4}{3}$

Volume = $\pi(\frac{2}{3})^2 \approx 1.4$

Therefore height of pillar 50 is ≈ 3.2 and width ≈ 4

Volume = $\pi 2^2 * 3.2 \approx 40.2$

Visual effect size (as volume): $\frac{40.2-1.4}{1.4} \approx 27.7$

Lie factor: $\frac{27.7}{7.33} \approx 3.78$

Therefore the graph is overstating the data.

I think that the lie factor is relevant as an idea. However, I doubt its importance for (human) analyst since nowadays computer software can calculate exact measures for the graphics that reflect to numerical values.

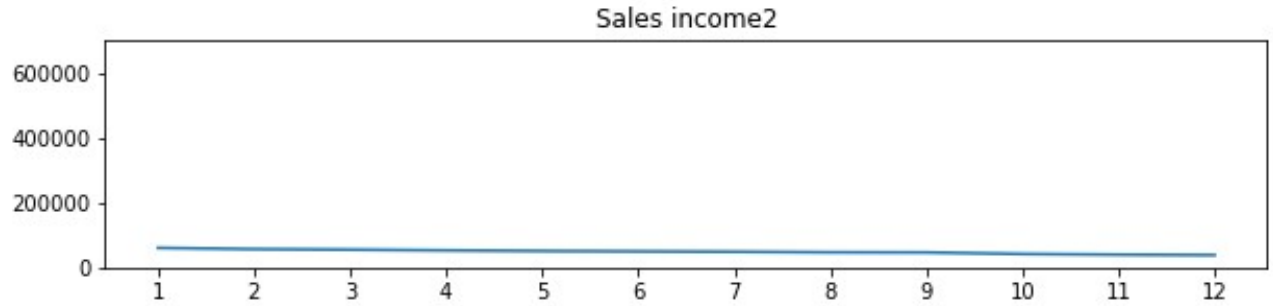
3.

a) I created a dataset which have sales/costs sample for each month of the year.

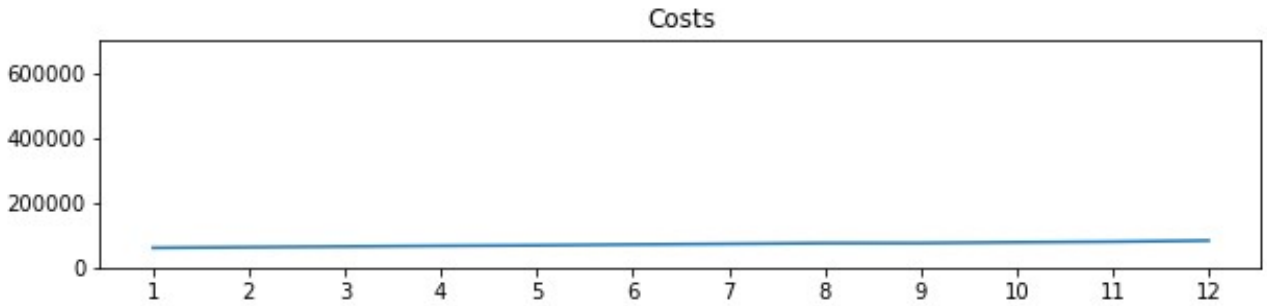
b)

Cumulative sales income

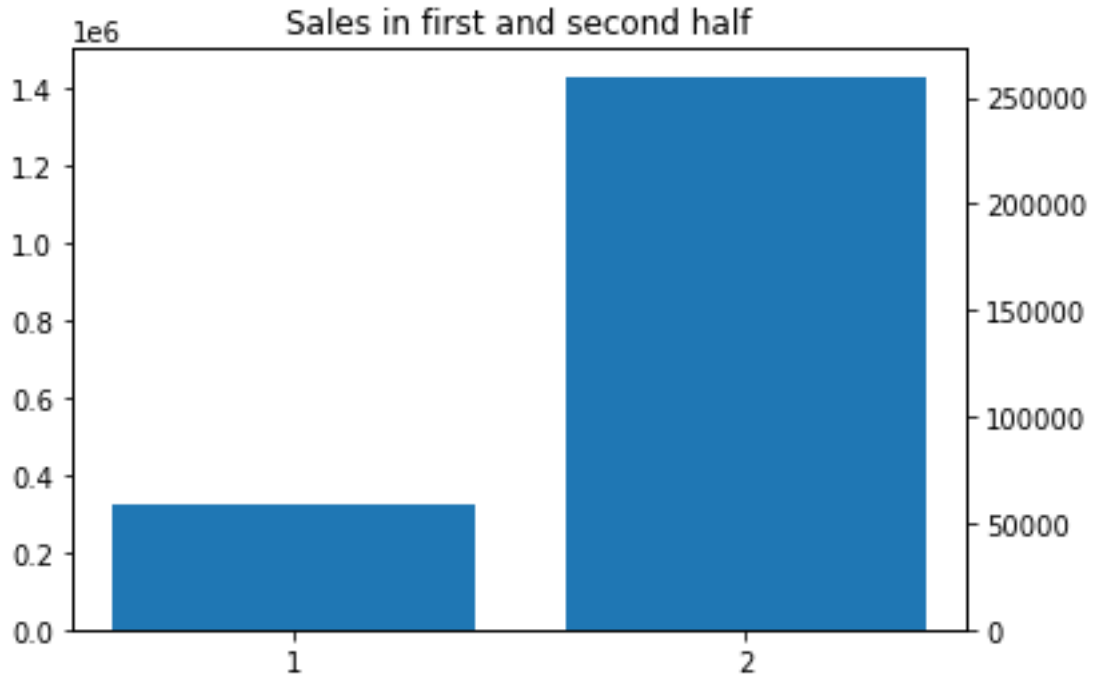
Plot of sales with upscaled y-axis and wider frames



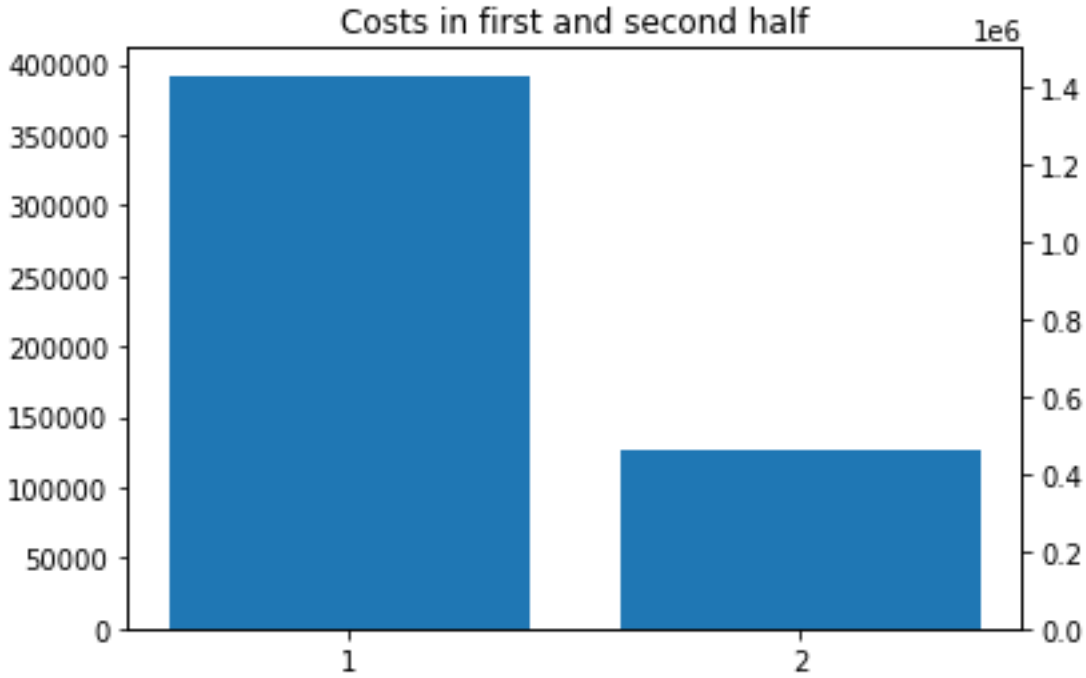
Same with costs



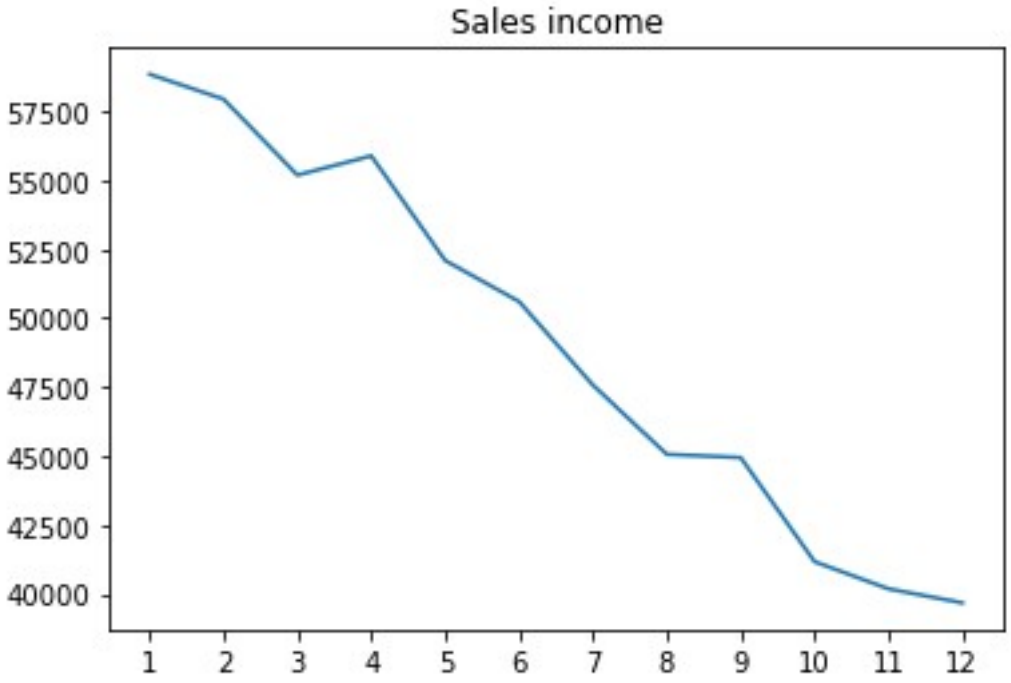
Sales in first and second part of the year

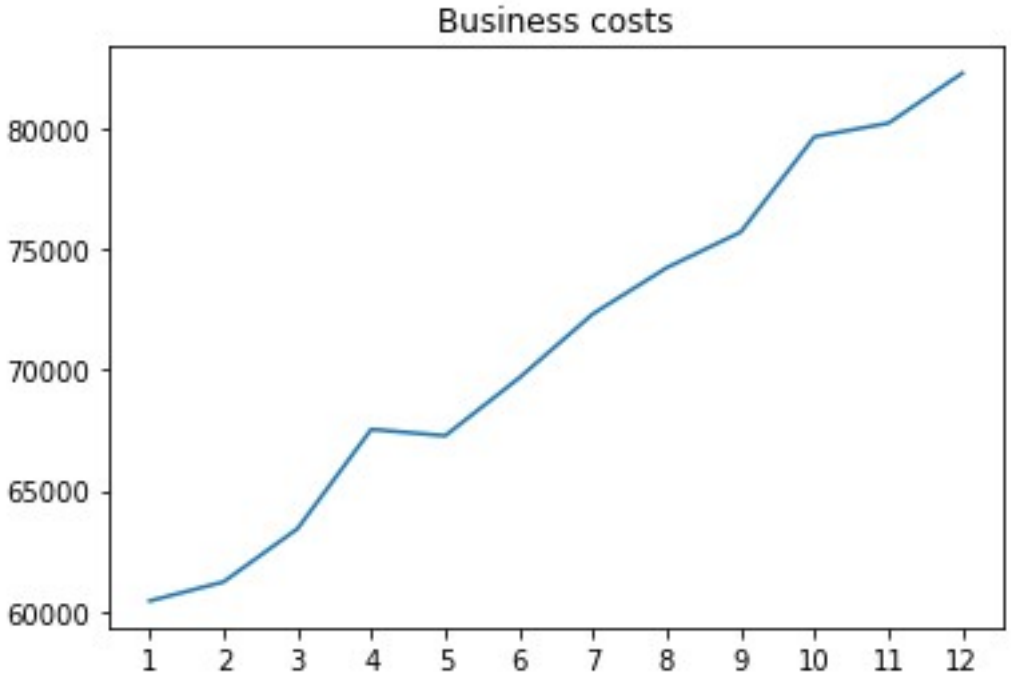


Costs in first and second part of the year



c) Original data





4.

a)

By my analysis, there are something wrong in each figure. If we start from the first one (that I think is most correct), I understood that the white circle represent the whole 24h day and the yellow region inside represent the proportion of daylight hours. For myself, the figure raises contradictory thoughts as terms of chatterjunk. Firstly I think that the figure is already very simple but it doesn't correspond to my mental image of time interval, actually, first I thought that it represents a lunar phase, and therefore the selected shape is not self-explanatory. Secondly, two dimensions are used to present a one-dimensional time interval and therefore the extra dimension can be thought as chatterjunk (maybe this isn't fair to count since histograms eg. are 2d as well).

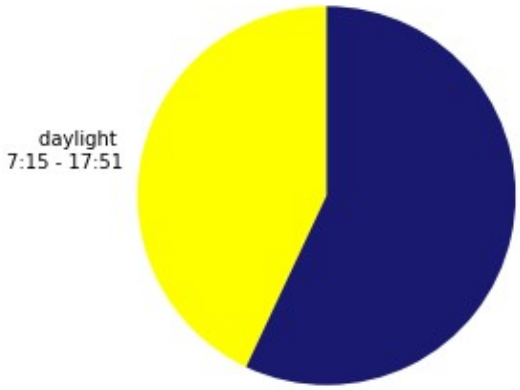
If we calculate the numerical size of effect, I think that it is most intuitive to calculate the difference 17.51 – 7.15 = 10.36 and compare it to the duration of the day, 24. Doing like this we get $\frac{10.36}{24} = 0.43$ and therefore the area of yellow part should be 0.43 times the area of the whole circle. Without going into calculating areas of circle subregions, I would think that the visual effect size is roughly equal to the numerical one.

The point where second figure went wrong is the way that it jointly represents two disjoint properties. If the yellow sphere represents alternative power sources, why the Benzene is visualized as subset of them? If this kind of interpretation is plausible, then there isn't anything that visualizes the effect of percentage growth that was stated in the text, yellow region simply represents the amount of 3767 new registrations of 'green' cars, and white region the (unknown) portion of benzene cars.

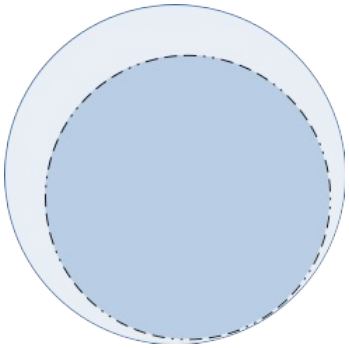
For the third figure, same thing apply that were stated in previous paragraph, differing only by the joint-disjoint part, since I think that it is fair to assume that daily consumer goods sales belong to the subset of total department store sales. Again, there isn't any effect that we can calculate since the numbers to compare these two sales was not presented in text.

b)

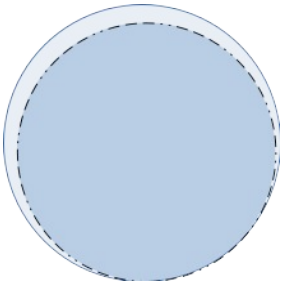
First figure. The yellow area is exactly 0.43 times the blue area.



Second figure. Darker sphere represents old portions and light sphere effect size (actual measures are incorrect).

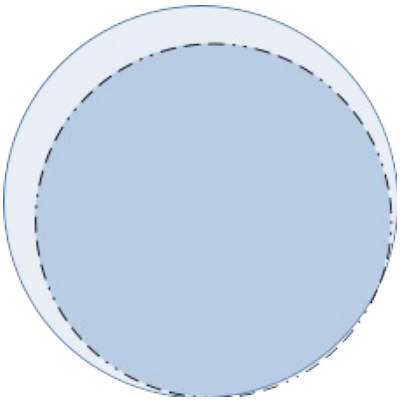


40% increase in alternatively powered cars

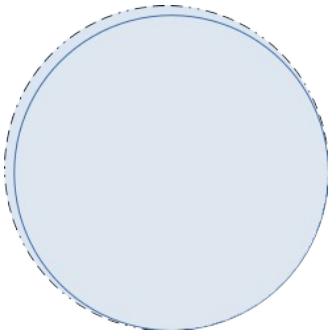


6% increase in benzene powered cars

Third figure (measures are incorrect).



8.2% increase in department store sales



0.9% decrease in daily consumer goods sales

c)

The scale used for open job positions is ten times smaller than the scales used for job seekers. Therefore the figure is misleading because at first glance it might seem that there is more positions than possible applicants, even though in reality the conclusion is opposite. It is reasonable that unemployment doesn't decrease since the growth of it is far more superior compared to growth of open positions. As an improvement, I suggest that both curves use same y-axis. Alternatively, I would suggest a following histogram: