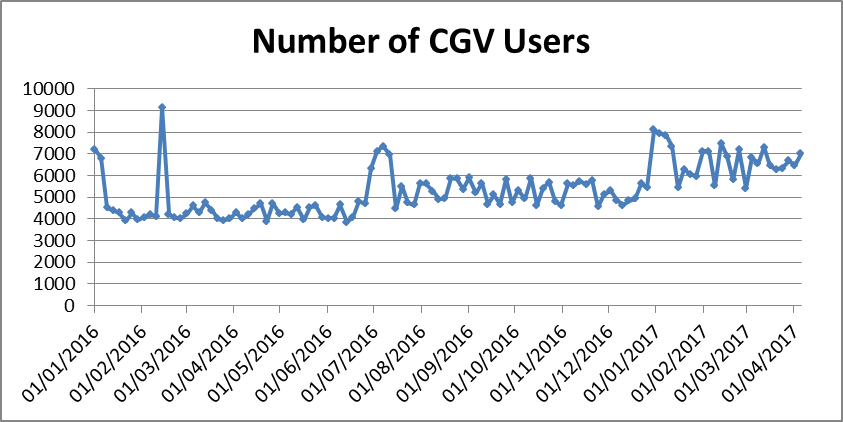
Questions:

1. In the following graph what would be trends you would want to comment/model?



*There is an outlier in mid-February – more than 100 increase in CGV in a single period. Perhaps this was a promotion*

*There looks to be an uplift every 6 months, once around Christmas / new year and once around July (summer holidays?)*

*There also looks like there are step changes after the uplift in summer 2016 and the uplift at Christmas 2016. There seems to be no trend upwards other than these step changes*

*The variance seems to increase during H2 of 2016 and again during 2017*

1. You are watching a game of football and the commentators say that on average a goal is scored every 30’. How long would you wait to see a goal being scored?

*It is difficult to think of probabilities in this way. If you watched 100 games of football back to back then you would, on average, see a goal being scored every 30 minutes.*

*But let’s assume that it’s the case that there is an equal probability of scoring a goal at any point of the match. Therefore, a goal being scored in the 30th minutes is just as likely as a goal being score in the 29th minute, and that is independent on whether a goal has been scored up until then.*

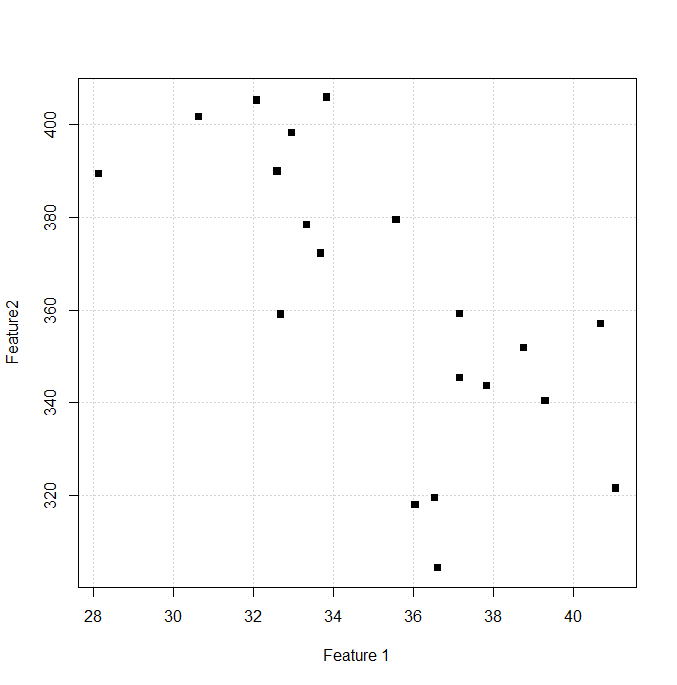
*Some games you watch you will see a goal being scored in the first minute, some in the 90th minute. But this will average out to 30 minutes after many games*

1. A co-worker from another department is interested in particular variable’s effect on customer attrition. You fit a complex statistical model to examine the effect of a particular variable on customer attrition. The variable comes back insignificant, but your co-worker says that this is impossible as it is known to have a strong effect. How do you explain this?

*Perhaps there is another variable in the model that correlates strongly with the variable in question. For example, let’s say the variable is promotional spend which has just increased. Customer attrition has just reduced*

*Let’s say we have another variable that takes the same shape as promotional spend - maybe the football season has just started. The model contains this variable too, and this is explaining the entire decrease in customer attrition*

*The fact that these two variables are highly colinear will mean that the model will not show both as being significant. This is true even if the promotional spend was actually a driver. In reality it may have been a combination of the football season and promo spend that decreased the attrition*

1. In the following graph Algorithm A suggested the existence of 1 cluster while Algorithm B the existence of 2 clusters. Which algorithm you think is correct and why?

*The correct answer depends on a couple of things:*

* *What feature 1 and feature 2 is. For example, perhaps a simple negative relationship could explain the variance – in which case one cluster would be appropriate*
* *What we are using the analysis for. Do we need to have two buckets? Will the two be treated differently?*

Please keep your answers succinct (<150 words).