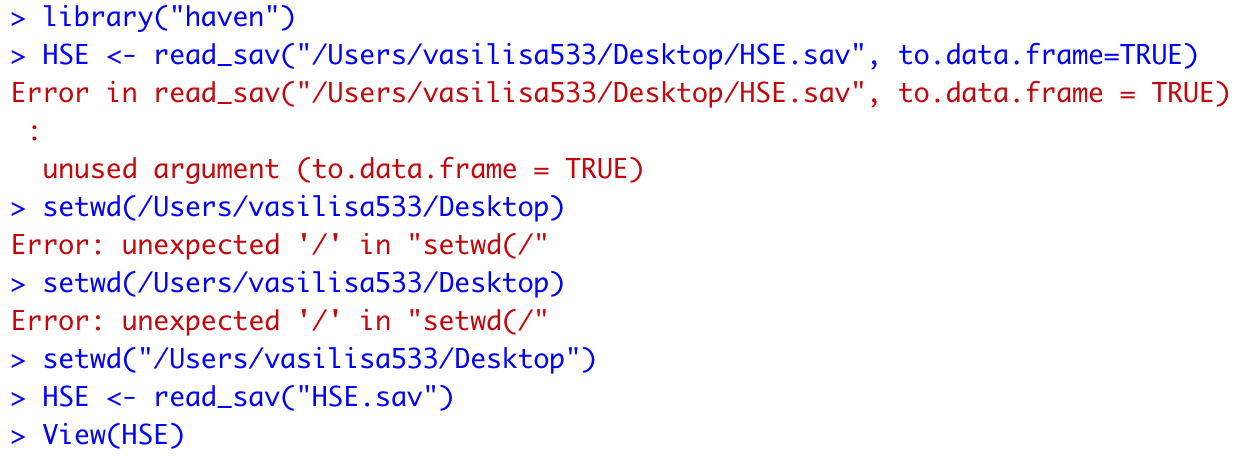
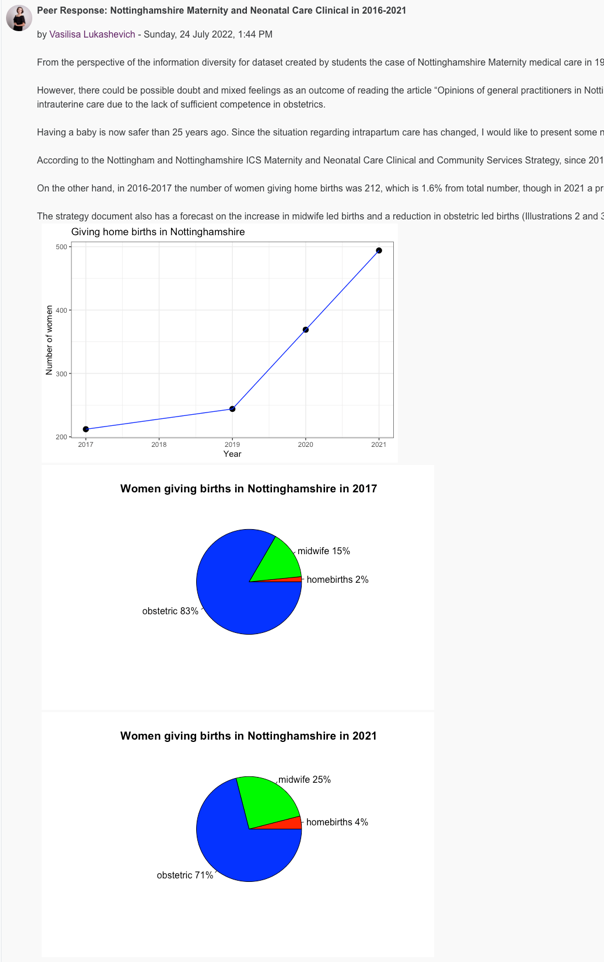
**Individual Reflection on Numeric Analysis Module: A Taste of Pepper**

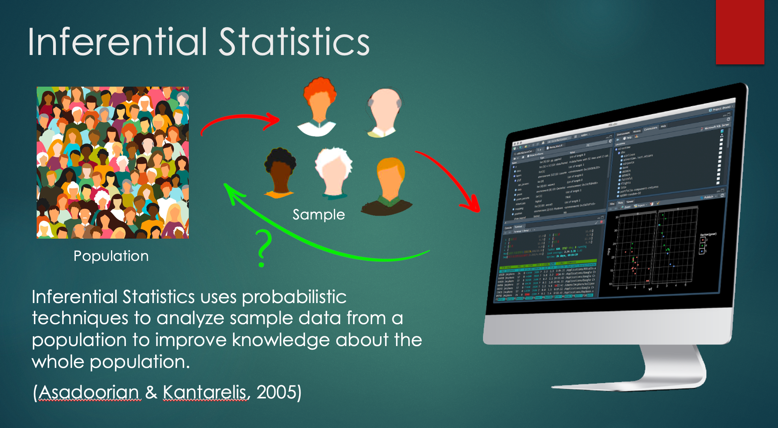
A few days ago I visited a supermarket to buy some vegetables. I noticed the same price for red bell peppers and orange bell peppers. I put them in the same bag ignoring the fact that the same price had been written on the different price tags. “It’s just a color”, I thought heading to a self-checkout counter. But store employees looked not too happy about that – they split the vegetables in two plastic bags according to the colour to weigh them separately. Before the Numerical Analysis module I would rather be disoriented and confused about this situation judging on environmentally unfriendly behavior, however now it makes sense. With all my respect I imagine the data table where red and orange peppers are different variables with the same numeric value. Maybe the vegetables were produced by different farms, hence from data scientists’ or marketing specialists’ point of view it is not a good idea to weigh two types of products together with the one tag. The advent of data mining has enhanced the customer behavior prediction accuracy (Chen et al., 2005). I want to add – sometimes it is true in case a customer understands basic data concepts too.

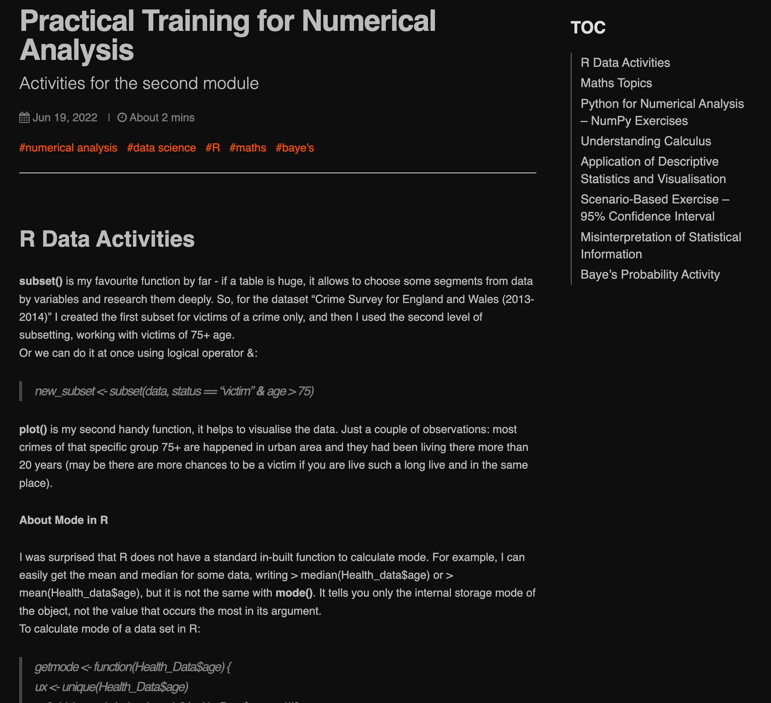
The most challenging part was the loading data to RStudio, I put the illustration of my first attempts below. I have gotten more confident about it.



The second challenge was ggplot coding process in RStudio. I did it for Assignment and for Collaborative Discussion as well (in the picture on the left), however since Microsoft Excel fits me better, I also copy-pasted data there to improve the esthetic visuality and save my nerves. To understand ggploting I watched *R Programming 101* on YouTube (Martin, 2021). Thanks to it I became more familiar with the piping operator and the way to cut the long code for *aes* to just a ‘*ggplot(BOD, aes(Time, demand)) +’ …* kind of view.

YouTube helps me a lot. Another video (Coughlin, 2020) gave me the feeling of subsetting in R. It was the first discovery I wrote on my e-portfolio page. Subset function allows you to choose some segments from data by variables and research them deeply in a new dataset. So, for the dataset “Crime Survey for England and Wales (2013-2014)” I created the first subset for victims of a crime only, and then I used the second level of subsetting, working with victims of 75+ age. Or we can do it at once using the logical operator *&*: *new\_subset <- subset(data, status == “victim” & age > 75).*

Performing analysis with null hypothesis made me crazy initially. I did not understand why we should reject the idea instead of proving it. Jeff Galak explains it in a very intuitive manner: it is one of the basic ideas in science, that we can never prove something to be true. Instead of that we provide evidence that something is not true. The example with a ball is excellent: we cannot prove that a ball **always** drops down on Earth, because **always** means in every single condition, including for instance any moment 100 years ****ago. Alternatively, we can try to prove that if we drop a ball, it will not fall to the ground (Galak, 2020). Here is the null hypothesis – if we fail with that, then the alternative hypothesis wins. To understand the concepts of p-value, confidence intervals and box plots I have watched many videos on Dr Nic's Maths and Stats channel, she explains boring topics with skill and ease. Dr Nic inspired me to illustrate Inferential Statistics for my Assignment (above).

I’m absolutely in love with Bayes’ Theorem. It looks elegant and helps calculate the probability for any case: for example, a chance to be caught in the rain according to the forecast (it is not always the same as your app predicts) or to be really pregnant with two strips on the test (50%, not 99%!). Moreover, according to my current field, it is important that Machine Learning algorithms use Bayes' Theorem. Bayesian Artificial Intelligence produces a thinking agent similar to the human process of thinking – able to absorb and learn adaptively (Korb & Nicholson, 2010).

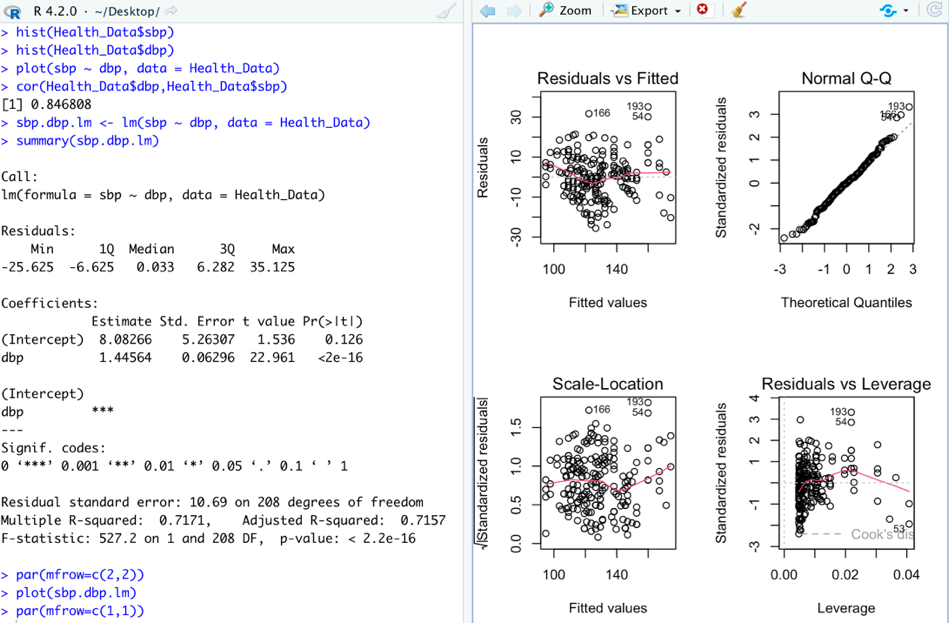
Let’s go back to R. I was surprised that RStudio does not have a standard in-built function to calculate mode. I point it out in my e-portfolio. For example, I can easily get the mean and median for some data, writing *‘median(Health\_data$age)’* or ‘*mean(Health\_data$age)’,* but it is not the same with *mode()*. It tells you only the internal storage mode of the object, not the value that occurs the most in its argument. To calculate mode for a data set in R, I use *unique, which.max* and *match* functions:

*>ux <- unique(Health\_Data$age)*

*>ux[which.max(tabulate(match(Health\_Data$age, ux)))].*

Additionally, I faced a problem with finding the mean. R showed “NA” only, then I solved it like this: *mean(HSE$bmival, trim = 0, na.rm = TRUE).*

To understand Linear Regression in R better I was reading an *Easy Step-by-Step Guide* by Scribbr. Working out the Health Dataset, I noticed a linear relationship between systolic and diastolic blood pressure (r = 0,85). The idea was not new for me, but it is quite exciting to find the statistically significant results with numbers and figures of plots and lines. According to other research, correlation between SBP and DBP is r = 0.74 ± 0.14, r > 0.5 in 95% of patients (Gavish et al., 2008). So, probably I was on the right track. However, I still do not understand a lot of things. For example, how much statistical calculations AI actually uses creating a picture based on a text query. I think from that aspect there is a huge room for growth.



**References:**

Chen, M.C., Chiu, A.L. & Chang, H.H. (2005) Mining changes in customer behavior in retail marketing. *Expert Systems with Applications*, 28(4), pp.773-781.

Martin, G. (2021) Ggplot for plots and graphs. An introduction to data visualization using R programming Available at: https://youtu.be/HPJn1CMvtml [Accessed 20 August 2022]

Coughlin, D. (2020) Subsetting Cases From a Data Frame in R | Using the subset Function. Available at: https://youtu.be/iM1e0wxUMrs [Accessed 25 August 2022]

Galak, J. (2020) Statistical Significance and p-Values Explained Intuitively. Available at: https://youtu.be/DAkJhY2zQ3c [Accessed 1 September 2022]

Hypothesis testing: step-by-step, p-value, t-test for difference of two means (2011)

Available at: https://youtu.be/0zZYBALbZgg [Accessed 31 August 2022]

Korb, K.B. & Nicholson, A.E. (2010) *Bayesian artificial intelligence*. CRC press.

MSc AI E-Portfolio: Vasilisa Lukashevich's experience. Available at: https://vasilisalook.github.io/practice/2022/06/19/practical-module2.html. [Accessed 2 September 2022]

Bevans, R. (2020) Linear Regression in R | An Easy Step-by-Step Guide. Available at: https://www.scribbr.com/statistics/linear-regression-in-r/ [Accessed 30 August 2022]

Gavish, B., Ben-Dov, I.Z. & Bursztyn, M. (2008) Linear relationship between systolic and diastolic blood pressure monitored over 24 h: assessment and correlates. *Journal of hypertension*, 26(2), pp.199-209.