Good morning, everyone. Good morning. Hello. Hello. I know you can hear me. Let's settle down. Did you know? Fun fact in Germany, lectures actually start 15 minutes after the poster time. So if you ever go to a lecture in Germany and it says something like start at 10 a.m., you will actually start at 1015. We call that the academic quarter. I don't know why this has never been established in the UK. I think it totally should. Because I would reduce the times that I'm late. Terribly sorry for being a bit late. In my defence, you were all standing in front of the door and I'm not entirely sure why. There's a light switch right at the wall, like when you come in on the right side, so don't feel like you have to wait for me. I think you're all grown ups. You can all go into lecture theatres without me, without smashing the place up. Or at least I hope so. Let me go ahead and close the door. Maybe that will stop latecomers from coming in. Okay. How are we all doing? First week done. Does it feel good or not so good? Not so good. Oh, gosh. What happened? Very busy or. Yeah. How many courses do you have in the first semester? Four courses. Four. Okay, good. So four lectures, four tutorials and each. So just lectures or tutorials or. Ah, interesting. So you have a great combination between them. Good. So. In this week's lecture, we'll focus mostly on data pre-processing in this lecture. So it's still a bit a bit boring. We'll start the fun stuff I think, next week or the week after. So I think that's probably the part you're all looking forward to. That being said, today, we'll also have a look at the most exciting part, which is your group coursework. So I'll have a bit of time at the end of the lecture for you to start looking for your group members and maybe have a bit of a discussion about how you want to proceed with that, because I think it makes sense to start very early on. And that would also be an opportunity for you to ask any questions about the assignment that you might have. I have uploaded the brief to the learn page. I have not yet uploaded a data set. The very simple reason for that is I try to create it and my code crashed on me. So now I have to try and figure out why it did not want to create the data set that I told it to. And I will figure that out today and I will upload the data set. But I hope you all had access to the brief, which will already tell you kind of what this is all about and what the data actually looks like is not that important for today. That being said, we will talk about why it's important to start looking at your data early on without actually thinking about the modelling process. So if you think back to our kind of little diagram of the ideal research process, we are still in the early stages. So we have discussed a little bit mostly in the principles lecture, actually how we think about research questions and how we derive research questions from a business problem that you might encounter. So we were talking a little bit about Alice, and Alice might make a little cameo appearance later in this lecture as well. And we were talking about how Alice should approach a business problem that has been posed to her by a company regarding a problem. And then she's collecting data on that. And then in the lecture we basically said and then she starts analysing the data, which is a very kind of simplified way of looking at it, because the first step is doing any type of analysis and whether that is a more traditional statistical analysis as we'll be covering in principles or whether it's more focussed on machine learning approaches as we'll focus more on this lecture series, the first step that Alice will take is always to have an initial look at a data set and then try to figure out whether there are any problems with it. And today we'll talk about problems with data. Now, before we start, a little bit of housekeeping. I've been sent an email regarding the reading list, apparently as an inconsistency between the reading list and that is posted in the course structure and the reading list that is shared via the library resources tap on learn. So if you click on the library resources, it kind of gives you a pop up where it gives you a bit of a of a structure and links to the actual work. So if in doubt, please follow the course description, which is kind of this PDF document that I uploaded on the course information unlearn. There's a table on that where it tells you exactly the topics that will be covering each week and also the chapters that I would like you to have a look at. So the table in that document is always the most up to date if I make any changes to it which are not reflected in the library resources, they will be reflected in that document. So if they inconsistencies, follow the course description. And in the meantime, I will try to also be able to update the library version. However, I can't directly update that myself, so I have to go through. I think that's the library office and I will tell them, Hey, can you please update that? So that's a bit of a lengthy process for me compared to just updating the PDF, which is why I would like you to have a look at the PDF if in any doubt. Yes, we talked a little bit about what we covered last week, both in this lecture and in the principal's lecture. So in last week's lecture, we talked about what is a predictive model. We were talking a little bit about how information can be used to make decisions and how if there's a very complex decision that you have to make, the information that you collect can't really be analysed by you on your own, which is where predictive modelling comes in. So predictive modelling basically helps you to make decisions by trying to predict what might happen in the future. We talked a little bit about the predictive modelling process by both the ideal and the more realistic version of it. We also talked about the difference between supervised and unsupervised learning. You will remember that supervised learning basically describes problems in which you have a labelled data set from which you build your model, and then you can apply that model on new and unseen data versus unsupervised learning. Where there is no two step process, you have just one step of modelling your data, which consists usually of some kind of segmentation problem. In many cases we were talking about regression and classification. Regression is where you try to predict a numeric value. For example, you might predict the price of a stock next month, and classification is where you try to classify new data points. For example, you might classify customers as being either no good customers for your company, profitable customers, or not so profitable customers. And that can also be a multi-class problem. So you can have multiple classes that you're trying to predict. We were talking a little bit about types of variables. We're talking about categorical and ordinal numeric, all of these types. And we were also briefly covering some challenges and predictive modelling and I think I went on a bit of a rant about the use of P values. And also I think I seem to remember talking about people relying too much on theory and not enough on the data. So some brands of me were in that part of the lecture. So today we'll talk a little bit about data cleaning and preparations. We'll talk about pre-processing standardisation, we'll talk about encoding variables, we'll talk about outliers. What are they? How do we detect them? What do we do about them? And we'll also talk about missing values. So let's start with an introduction to data preprocessing. So the first question that you might ask yourself is why do we have to pre-process data in the first place? And I believe me, if you've ever worked with real data, you're not asking yourself that question because it's kind of obvious. Data arrives never in the right format that you want it to. It just doesn't. I've seen data sets that are absolutely terrible, so I've seen data where there are coding issues. So for example, in different languages we have different letter systems. So sometimes if you try to translate one data set into another language, then the computer basically tries to use unique coding for those letters and it breaks because Excel is not happy about that. And then your whole data is just a mess. And then if you try to merge that with data that is actually whole, then you're creating even more of a more chaotic situation. So data does not arrive in the right format for analysis and you should always kind of start analysis with the expectation that you will have to do quite extensive preprocessing. What does that mean for you? Give yourself time to actually do that. So if you think about, for example, the course assignment, which will be kind of a little analysis in a nutshell, do not think that you can immediately just apply all the models that you want to because it's not going to work, and then you're going to wonder why it doesn't work. And then you have to go back into the data and then start the preprocessing. So give yourself time at the beginning of the analysis to figure out what do we actually have to do first before we can start. The more exciting part, which is the model building. What do we mean when we say pre-processing or cleaning? Basically, it means that you trying to create a dataset that is suitable for the model that you've chosen. So we might say we need to clean the data. We need to tidy it up a little bit. That includes all types of pre-processing steps. So handling missing values, handling or detecting outliers changing and detecting unclear variable names, that's a big one, as well as inconsistent encoding of variables. That's a very big issue if you're trying to work with multiple data sets at the same time. Last time I taught you a little bit about one of my recent research projects where we look at census data and we're also trying to work on census data and at the same time different longitudinal data sets stemming from surveys. So one of the issues that I want to detect in that is I want to create something we call a synthetic population that is a data set which is statistically representative for a real population, but is artificially created using census data. So we take a census data set. Basically we take two census data sets because the way census data is published in many countries, at least in Canada where I've been working with the data, is you have two files. One of them is a household sample file. So not every household has to actually fill in that little survey that the government sends them. It's just a sample. And the government is trying to create a sample that is representative for the population. So you have a couple thousand records of households for a given city. For example, we were looking at Montreal and then you have something that is aggregated data for the whole region as well. That kind of aggregated data describes the population of the city in summarised terms. So for example, I know how many people of a specific age group live in Montreal or live in a specific census tract within Montreal. So now the question, the first question we were trying to tackle is how do we combine these two data sets to create a population of households? The aim was to create a data set which would represent all households in Montreal. Based on these two data sets, we knew what the whole population should look like in aggregated terms, and we had a sample of what these individual household records might look like. So we then extrapolated basically from these household records in such a way that the finalised data set creates the same aggregated data that we know is true. So we know what the population should look like and aggregated terms. We use household samples and sample from those and then we extrapolate. So we make a bigger data set from those and then double check whether the result is actually close to what we know should be the true population. And then we can use that population data set for all types of modelling. I mentioned that one of my interests is in financial well-being, so I'm. Twisted in how populations are, how people feel about their financial situation. So, for example, the debt levels or how anxious they are about the future, how much income they have, all sorts of different factors. So we have a representative sample about people and their financial situation also from Montreal. And now we're trying to merge the sample with our population. So we look for records which are similar in terms of some demographics, and then we try to merge in our financial well-being data into our population. And this is basically one of the research projects I've been working on recently, and we call it the synthetic ecosystem because we are layering synthetic data sets and extrapolated data sets on top of each other and create a whole population. And then we use, for example, agent based modelling to try to model human behaviour within this kind of environment of multiple connected data sets. The reason why I'm telling you that, and I promise there's a reason apart from researchers like to talk about themselves and that reason is one of the biggest challenges we had is that the data sets all came in different formats. So one of the data sources was the government. We took census data, another was the Financial Consumer Agency of Canada who gave us the financial well-being data. And then we also looked at health data coming, for example, from the Canadian Longitudinal Study on Ageing, which is a separate kind of research organisation research project going on, who were collecting different data. And all of these data had different formats, so they came with different variable names. They encoded their variables differently. For example, if you want to merge between age categories, they have to match. But if one of them collects data, for example, for people 18 to 25 and then 25 to 30, and the other has much broader categories, for example, 18 to 35, then that's not a direct match. You first have to kind of then change the encoding of each of these variables to match between the data sets. So one of the first steps we did was to create some kind of map between the variables where we're trying to explain how each of these variables from one maps onto the other. And that is one of these issues that I very simply put on the slide with inconsistent coding. But you can see how this very kind of small, seemingly small issue can become a really, really big problem in a research project. And if you underestimate it, then that takes away a lot of time and a lot of money from your research project because time is always money. You have to pay your employees who are actually working on that research project. And if it suddenly takes twice as long and that might be very expensive. Okay, let's go back to the slides. So important for this pre-processing is that it helps interpretation. Obviously, for example, if you change unclear variable names, then it helps you actually interpret what the data means. And it's also necessary for model performance in most cases. Sadly, there is no basic step by step that you can just follow. Students often ask me, Can I just have a list of all the things that I should do or things like, Well, which models should I test or which models should I run first? There's not really a step by step process because it highly depends on your research question what you're trying to find out, as well as what the data and the model needs in that specific instance. That being said, there are some common steps that are typically checked for and I put them here for you kind of to give you a bit of a guidance. If you're, for example, starting out for the first time working with real data and that might also be a good kind of guidance for you, for example, for your group project. And that gives you some some hand-holding for the beginning, let's put it that way. So initial check for the data format, you would not believe how much data that arrives is actually faulty and not I mean by accident. I received data where there was just a whole variable missing and I thought, Oh, were they not allowed to give me that variable? And I wasn't sure because surely they should. So I emailed them and it was just an upload problem on their side. They told me, Oh yeah, we kind of forgot to upload that variable to the server, we'll fix it. And then I suddenly had it. So in many cases, check whether the variables that you should have, other variables that you actually do have and also check whether there are any errors in that, any blatant issues like people have wrongly coded the postal code, for example, something really small, and that also involves checking the documentation. So in my case, that was already mentioned. I will mention them again 200 pages of census documentation. I'm still not over that. So check the documentation, check the sample size. All of these really small issues will help you later on. Don't just run into the data. Try to understand what the data is trying to tell you first, then initial check for obvious large scale problems that was dimensioned completely missing postal code. Also check for amounts of systematic missing data. We'll talk about missing values later, which are like small issues where the singular value is missing. We'll talk about why that happens. But look for systematic issues that you might have in the data. Then descriptive statistics explore the data and what it looks like. Please always start with that. Don't just throw everything into neural network and hope it works out. Please kind of try to stick to simple steps. First, explore the data through descriptive statistics. Do dummy encoding variable transformations. ET cetera. Next, handle your outliers, handle your missing values, and then, if necessary, you might want to repeat the stripped statistics. I very rarely see that mentioned, but I think that's such an important step because if you handle a lot of missing values, a lot of outliers, so you do a lot of kind of removal of data and information, surely, surely you want to double check what impact that has on your descriptive statistics and you want to see how much of an impact that removal of variables actually had on your dataset. Then check for model assumptions, choose your model and then you can actually start the fun modelling process. So there's a lot of sub steps in this kind of pre-processing umbrella. And today we'll focus on steps 3 to 7 of those. So we'll talk about descriptive statistics. First, we'll talk about dummy encoding, variable transformations, outliers and missing values, and then obviously descriptive statistics, because that is also step three already. Okay, let's start with some summary statistics. Now, this might seem a little repetitive for you because it's very similar to what we talked about in Principles of data analytics last time. The reason why I put that here is it's really important. So I'm trying to tell you multiple times, but also for kind of completeness sake so that so that you're aware that this is part of this course as well. It's not two separate courses. I remember that in the first lecture one of your colleagues asked me, how are the two courses actually connected? So what's the kind of connection between the two? And there are some steps that you'll see repeated through both courses. I will make that brief because I know that this can be a bit repetitive, but I do have to cover it for completeness sake. So please bear with me as we fight through that together. Very briefly, first, descriptive statistics is basically an overarching term, so you will see descriptive statistics, summary statistics, all of these words used interchangeably. Strictly speaking, descriptive statistics is its own branch of statistics itself, and creating summary statistics is one step or one part of descriptive statistics. The area of statistics it also creates. It also includes using visualisation techniques to describe a whole data set or the distribution of individual variables. So some of these statistics can or are usually calculated per variable, and they often describe some different properties of them. So for example, we look at the average, their variance or their spread, and we can also visualise multiple variables or individual variables to plots to describe the whole data set. An example for that is that you've seen last time of frequency Histograms. So as I mentioned, this is a very important early step in any data analysis project. And importantly, they can also be used to communicate initial findings. So a lot of the models that you'll be looking at later are difficult to actually visualise and describe to a lay audience. Neural networks are, for example, famously difficult to explain to a lay audience because they are a very famous black box model. We don't actually know what they do within the model. We just know what goes in and what comes out. And then we kind of try to interpret the accuracy or the errors that were created by the model. But descriptive statistics are one of your key tools to be able to communicate to, for example, a manager. So if they ask you to create some kind of research project and analyse, for example, coming back to Alice, analyse the attitude of people towards some kind of soda product, then if you just tell them, Hey, I created this really complex model, here are three numbers. This is the accuracy and this is kind of the F one statistic, and then I'm going to tell you, Yeah, okay. But what does that mean? What did the data actually look like? What do people think that they're not going to be happy if you just tell them the data set or the model says this and that. The model has the singular value or this, this, this kind of accuracy. What they want is to is a story. So what they really want is that you're able to communicate a coherent story. So a whole storyline, what does the data tell us on its own and what kind of model tell us which the data on its own can't. That's why this kind of descriptive statistics are really important. You remember back when we were talking about what is predictive modelling, we said it's a tool to make decisions, but the first step in that is we think about the information that we use to create or to make this decision. And descriptive statistics are your way of describing information, very simply put. So a couple of very basic summary statistics that you will recognise from principles are, for example, the mean to describe the central tendency of your data as well as the variance or the standard deviation to describe the spread around that around that mean. And interestingly, sometimes you you will see that whether you report the mean or the median or the mode depends on the data and it also depends on what you're trying to communicate. So in many cases you will, for example, have someone ask you, Hey, what's the average value of this variable of this question? But you have to interpret that as what's the median value of that question? Because, you know, I asked a statistician in that room that the mean is not a suitable measure to report. So if someone asks you for the average, that might in contacts also be that they're interested in the median. You just have to interpret that and you have to know what kind of measure of central tendency is the important one. So we know that the arithmetic mean which is the one I put here because it's the most common one that you'll encounter is simply calculated by summing up all of your values and then dividing them by the number of values. And we also know that the variance is calculated as the deviation of each individual value. From that mean, you sum up those differences, the squared, the squared differences of that, and you divide it by the number of measures that you have minus one. And if you take the square root of your variance, you get your sample standard deviation. You also see that we divide sample here. Reason for that is you very rarely have data on the whole population. So typically you will say that you calculate the sample mean you it's technically not correct. If you just state that you calculate the mean because it's inaccurate, you calculate the sample mean or you calculate the population mean. That being said, most cases will just state the mean or the median or the variance. And you have to infer from the text what kind of value it's referring to in the calculation. It doesn't really matter that much. The only difference is that you would sum up all elements across the population and you divide by the number of the population. Instead of having the number of the sample points, which in this case is n. Yes, I mentioned the median. You'll remember that the median is the second quartile. So it's kind of the midpoint of your data. It's split your data in half. 50% of the data will be to the left of it. 50% of the data will be to the right of it. And then simply the first and third quartile are the dividing points of that. So you can think about them as the median of the remaining data if you already took the median, if that makes sense. So you can kind of cut your data into smaller and smaller pieces. One thing that I find particularly puzzling and still keep forgetting is the difference between a quantile and a quartile or the percentile as well. So you have these three terms and especially quantile and quartile are often mixed up. So the difference is the quantile refers to the percentages. So that's kind of how you can count. You can remember it quantiles refer to the percentages quartile refer to the quarters. So they are always the 25th. So the first, the second, the third, 25th quantile, the 50th quantile and the 75 quantile. And also I can see that there's a little arrow here because that should say the 25th percent quantile is X tailed. Okay. The range we've discussed that as well is just the maximum value minus the minimum value of your whole data set. We also discussed how outlier sensitive that is. If you have an extremely large and extremely small value, then your range will be affected by that quite significantly. We also talked about the interquartile range and in that context we talked about box whisker plots. You'll remember that these boxes basically describe the spread of your data. Talking about spread of the data. We have a couple of visualisations here for you to have a look at. So we have a few different histograms put up here. One of them is kind of your your idealised beautiful histogram. So one of the top left, that would probably be what we describe as as close to as a normal distribution as you will get in real life data. So if you have something like that, that's, that's normal. That's normally distributed data. It doesn't get much better than that in real life. Sadly, we also have white skewed and left skewed distributions. You'll remember from last lecture, that's the one that I mixed up because I keep forgetting which one is right skewed and which one is left skewed. To me, it makes kind of sense to think about left skewed has more data on the left side, but no, it actually describes where the tail of the distribution is so right. Skewed data has a tail to the right side. Left skewed data has a tail to the left side. Now, before you all fall asleep, who would like to tell me what a box plot would look like for kind of each of these three data sets? Who's brave enough to try out our drawing function? Maybe, and would like to draw me a box plot for each of these three. Very rough one. Doesn't have to be pretty, but one of them has to do it. Who's brave enough. No, no, no. One of them? Yeah. Do you want to confirm? Come to me. Don't be scared. I think these are just to get off. So will you. You come here. You take this pen. And you hope that it works. It does. So for this one is like the. Mhm. I just mean it seems like. That's. Think about what the x axis actually describes. We have so many. Like. Like. All like 90 degrees. So, yeah, very good. Maybe we may have like this. Yes, exactly. At this point is the 25%. Exactly. So that's the first quartile, the 25th percent quantile. This 1st May be like still the minimum. Maximum. And since it's skew so, it has more values on the right hand side. So maybe the 75%. Look at what the boss I told him. Hmm. Hmm. All right. So as more values on the this part on the left side. Exactly. So more smaller values like this. Yeah. So think about what this bar here describes. That's your. Is it the mean or the median? I think it's the media. It is the median. Exactly. So the median in this case, you can see is, for example, described as 1.0. So it would be quite far to the left in that plot. So I personally may take the pen. I personally would probably draw the median somewhere here. All right. Because it's more close to the more value on the left. So it's more closed. So if you have a box plot and you have your data spread, then the median kind of describes where where am I doing that the wrong way around. I keep forgetting whether it's left, left to the right skewed. 50% of the data should be left of the median. So 50%. So the larger portion should be on the left, right? Yes. Okay. So that way round. Very good. Thank you so much. Maybe I will get the wrong case for people. Like to just figure out more. No, that's perfect. So something important to kind of remember is this idea that the box can be seen as a representation of your histograms, as in the white of the box actually is the way that the histogram will also look like if you draw it. Something else that we could discuss in this context is outliers. So you remember that the whiskers of your box plot actually describe how far away from the first and the third quartile. The data is further spread. Either you can draw your whiskers all the way to the min and Max, as your colleague has demonstrated, or you can think about whether you want to stop drawing your whiskers at some point, specifically the 1.5 times interquartile range. And in that case you would indicate your outliers as little dots kind of two to the sides of your whiskers. So if we talk about outliers here, for example, you could decide to draw the box just to the length of parts of that, something like this. And then you could indicate outliers which are located here as dots in that part in that plot as well. Okay. Yes. So let's think about that a bit further In the first question, what summary statistic do you think would work best for getting the grasp on central tendency of a skewed distribution median? Who would like to tell me why in one sentence one of you. Otherwise, it's very difficult to me. To all of you. Yes. The mean is affected by outliers and it's also affected by skew, which kind of if if you think about it, a large, very strongly skewed data set is often indicated for outliers. So we have this kind of very long tail distribution and you might think some of these will be categorised as outliers. So yes, mean is sensitive to skewed distributions and outliers. So let's use the median instead. Now we have two serious down here. And now you can do very quick calculations if you have some paper and pen or some Excel sheet open and you can tell it to tell me what's the median and what's the mean of each of these series. And I will do the same. I. Oh, yeah. It's. Oh. Okay. Who is quicker than me? Who already has the mean of both of these series? Yes. 4.7, 3.4.7 and 3.4. Everyone agrees. Yes, perfect. And who can tell me the median of both of these series 3.5 and 2.5? Wonderful. So you all remember how to actually find a median? Because I think that's probably the most more tricky one that we have to remember. How did we find the median? Saw the data. And then you look for the midpoint of that data. Exactly. Oh, gosh. Why did I put that question up there? Now I have to calculate it. Okay. Based on these two series, look for the standard deviation and the interquartile range. And yes, you do have to calculate this because I have two too, So I will do so electronically. This. Huh? I learned something new. I learned how to use the quantile function. Excel. Look at me. You might love about that, but it's actually been a while since I've used this function. Okay. Who had a look at the standard deviation of both of these? Would you like me to show you how to do this? I'll share my screen. Yes. Regarding when we calculate the standard deviation. To use the populations below the. Samples formula because in the denominator. What do you think would be the difference? I mean, if it's a minus one, then it becomes an estimate. So it's for the sample and it's for the population. I was wondering if the given series is a population of the sample. It's a really good question. So I would define them as populations because to me it does not define whether there's anything going beyond kind of these numbers that we have. So for me, I think it would be a fair assumption to make to use the population. Yeah. But you could equally argue that it's a sample. So it's kind of a matter of discussion. So you could you could use either. So I'm happy with either. Do you have one for me? Yeah, I actually. Was thinking about it. Thinking about that. That's good. Does anyone have an answer? Yes. It's 2.19. If we use the population for the second one, it's 3.03. Yes. Very good. Anyone agree? Disagree. Yes. Very good. Do we also have an interquartile range? I think it was a bit trickier because it takes a little bit longer. Anyone, anyone or anyone has any quartiles that would already give us. I will have a first step in the right direction, maybe on this side of the room because you've been so quiet. Avoiding eye contact. Yeah. The. Six. I was a yes. Is it 2 or 6? I. I'm not. I'm not disagree. I'm double checking. And you do? I fear. Okay. And for the for the third. Well. IQ. Ah, yeah. My question is why? My calculation is 6.4. 6.6.6. 3.60. Yeah. 6.0. Interesting. So we'll have I will I will double check with Excel because Excel is arguing that it should be 5.5, 5.5, 5.5. Interesting. So because Excel gave me quartile 7.5 and two, so the interquartile range between the two would be 5.5 for me. Yeah. You have ten numbers within two to be two integers for each other. So counting. So third number. The eighth number. The third number. The number. It's only integers. There's no decimal. So making the decimal, unless I'm doing it wrong now, I think that might be different. I think that might actually be the difference in how Excel calculates it as the average. Yeah. Yeah. So Excel. Actually, I think that's probably the difference between why why we get six or why we get 7.5 is where we get 5.5 or 6 is exactly that. It's looking at the average because it's an even number that we're looking for. The quartiles four That's really interesting. It also makes me want to look up how Excel calculates quartiles, so I'll do that as my homework and actually try to figure out the difference in Excel versus the formula that I presented. To get into quartile. You still have to find out the two. Yeah, it's not direct. Yeah. As far as I know, there's no direct formula. Yeah. Okay. Now, briefly before the break, let's have a look at a couple of statistical distributions. First of all, I often get asked the question, why should I care about statistical distributions in the first place? A couple of simple examples where they can be helpful for you is they can help you with outlier detection. We'll see that later. In many cases, they are really important. If you make any assumptions about your models, for example, in regression and they can also be helpful for making assumptions about where new data points might fall. So that's an example for how we would use the empirical rule or Chevy Chase inequality in you were talking about this risk assumption that we were making about where should new data points realistically fall within that distribution? Now, briefly, let's have a look at a couple of the more common statistical distributions. We will go into more depth in the lecture tomorrow, which is why I'm kind of breezing through them a little, because I think if I tell you too much detail about both of them tries, then it will get very boring. So uniform distribution, this is a kind of simple it's not exact. You can see there are some deviations, but what a uniform distribution for discrete data, which in this case it is would look like is you have equal probabilities for each of your values between two set values that you have. So, for example, you would you would decide that this is your value, this is your value. A Wow, my handwriting and this is your value. B And within that interval you would have equal probability for each of these values in a continuous, in a continuous data space that's a bit more obvious where you have A and B and you have equal probability between these two values in that interval. So uniform distribution, equal distribution of probabilities, normal distribution. We already talked about that. This is kind of this nice bell shape that we have. We saw this histogram earlier, what that might look like in real in real life. And we also talked about the standard deviation and the mean. In this case, we can see this is a population mean and a population standard deviation and for a standard normal distribution also called Gaussian distribution, we would say that the mean is zero and the standard deviation around that mean is exactly one. As I said, we will talk a little bit more about that next week. Binomial distributions. That's this idea of drawing and then looking at the outcome at each individual draw. And a common example is kind of just this colour coded ball drawing that you sometimes do where balls are either black or white, and then you count how many black and how many white boards you actually draw. And the Bernoulli distribution is a special case of that where where you draw exactly once a common example for that is flipping a coin. Poisson distributions are where we look at events occurring within a time frame. And a common example for that is you look into a machine and you count how many times does the machine break within a specific time frame? So you look at event occurrence and you try to model the the fixed rate. So that's the rate at how many times does it break? So for example, five times in a one hour interval, then your lambda would be five for a sufficiently large number of observations and lambda that is larger than ten. So a high kind of occurrence of events, the Poisson distribution and starts looking like a normal distribution. We also have exponential distributions that describes the waiting times between independent events and you can see the relationship between exponential distributions and Poisson distributions. And we'll talk a little bit more about the exact relationship between the two tomorrow in the lecture as well. So you have your beautiful probability theory lecture to look forward to at 9 a.m. tomorrow. Okay. I think now is actually a good time to take a little break. So we'll take a ten minute break for some water. And then we come back to talk about transformation of variables, outliers, etcetera. Now that we've actually covered the basics, see you in ten. Okay, let's come back and sit down. Challenges. Okay. Two questions that came up that I found very interesting. One is. What do we actually have? So especially considering the reading. What concepts are important to take away from this? Or what do we actually basically have to study and understand and be able to reiterate for the exam, for example? I think we will talk more about the exam later in the course. The important thing that I want you to remember is that it is the exam will be essay based, so you don't have to remember formulas and then do calculations or anything like that, because the purpose of this course is not to teach you how to calculate probability values for distributions or something like that. That's not the purpose. The purpose of this course is to give you critical thinking skills in being able to follow a modelling process and evaluate the results and discuss these results in a practical context. So what I really want you to take away from this course is. More on the abstract side, so more being able to understand what type of concepts exist and how you can actually use them in real life. That means you don't have to sit in the exam. And I'm not giving you like a data sheet and a calculator and you have to calculate numbers or something like that. I don't I don't find that useful. If you find formulas useful to study concepts, then you're very welcome to use them in the exam to explain something. So a lot of students actually find it very helpful to, for example, if they talk about regression models to give me the formula of a regression and then explain each component and then explain in which concepts or considerations of regression is useful, why and what is the relationship to the formula for that? Others find it more more easy to remember concepts in an abstract way, so more conceptually thinking. And for me, both are. Both are good as long as you're on the show. A good theoretical understanding of the model and the practical implications of that theory. So I hope that that answers your question. The other question I had was about the statistical distributions and the reason why I breezed through them. So we will talk in a lot of detail about them tomorrow. And I also got a question about what is kind of implication of those specifically also in relation to real world applications or exams or for example, the question how do we actually know which distribution I should use for a step in my analysis? So there are a couple of steps in that. One is in many cases it's more a visual question. So in many cases you actually plot your data plot, for example, your histogram of your data, and then you think, what is a possible distribution That could explain that data, which helps me in my modelling process. And then there are statistical tests that you can run where you actually check whether the data does indeed conform to the distribution. We will talk a bit about tests for variable distributions next week in the Principles of Data Analytics series. When we talk, for example, about a hypothesis, testing for sample mean differences, but also for conformity with distributions. So that will be statistical testing, which we'll do next week. So the reason why we cover statistical distributions in this specific lecture at this time is really coming in for the data preprocessing, which we'll talk about now because one of the ways of detecting outliers is checking whether they conform to the distribution of the rest of your of your observations. So if all observations follow a normal distribution, for example, then you can check which of these values lie outside of the expected values for the distribution and therefore outliers. So that's why we covered statistical distributions here as a tool to detect values which are outside of what you would expect them to. Before we talk about missing outliers, though, I would like to talk about variable transformations. And to me, this is one of the most important pre-processing steps, especially in the social sciences. You get data in a lot of odd kind of and for a mixed data types, for example, you might get categorical data, ordinal data and numeric data all in the same data set. And if you want to do any more serious modelling on those, you first. We first have to transform those variables, especially the categorical and numeric, categorical and ordinal ones into data that is actually possible to model with. We touched upon that a little bit when we were talking about quantitative data versus qualitative data. So now we will look at how actually we bridge the gap between these two. So why do we transform data models, require it, require data to be within certain bounds, for example, and variables are measured in different scales, which is difficult for model interpretation and estimation. So first steps for standardisation. We did cover that briefly last time, either here or in the principles lecture where we talked about how we can make our data basically dimensionless. So standardisation is one of the most important preprocessing steps for almost all modelling purposes. So in almost all. All models that you have a look at. The first steps that people will do is standardise your numeric data. And what does that mean? It basically means that for each value you create their z-score so you deduct their mean. In this case we use the population mean and population standard deviation. In this formula. It would work the same way with sample though you deduct the mean and you divide it by the standard deviation. So that means if you do that to every variable in your data set, all of these values then have a mean of zero and a standard deviation of one while still retaining the shape of your variable distribution. That means that the data variables actually get become comparable. If you, for example, have one variable like income which is measured in thousands of pounds, and then you have another much smaller variable like for example, a small numeric variable, something age perfect. So then you have age, which is a much smaller variable, probably measured on a scale from 0 to 100 ish. You can see the scale differences. If you don't actually use a model, then the larger values would overpower the model and the smallest would become insignificant. But if you want to make sure that can be handled on the same scale, so with equal importance, especially you would use standardisation to then make these values comparable. And a very similar approach to standardisation is minmax scaling, which we also call feature scaling. Some people call it normalisation. Normalisation can also be used for standardisation. There's a lot of terms which are used for very similar concepts very differently. The difference here is that instead of using the mean value and the standard deviation, we actually use the minimum and maximum value of the distribution. The that also scales variables to be between 0 and 1 and still retains shape. However, because we're using the minimum at the maximum value of that, it's actually quite sensitive to outliers. And you will see that these two in comparison tomorrow in your computer lab. So we'll actually implement both of these. And then you can have a look at the comparison between these and what that looks like. Now, I mentioned encoding for categorical variables. Usually if you have a categorical variable in your data set, you would use something we call dummy encoding and we also call that one hot encoding. The reason why we call it one hot encoding is because for each observation, for example, here we have five people who report their favourite colour. Then for each of these we would create a dummy variable and we would give them one value. So one hot. That's why it's called one hot encoding. For example, if someone reports their favourite number favourite colour is green, we would give value one to the green dummy. You will also see that there's only three dummies, even though we have four possible colour values. And the reason for that is that we want to avoid that the all dummies are perfectly correlated, so we always create one dummy less than we have values In this case that we create three dummies for four possible values. And you can still see that we can code that someone could select Blue as the option by giving them all zeros. So this is still a perfect representation of someone voting blue, namely as someone who does not vote red, green or yellow. That being said, this is only the case. Obviously if these are the four perfect options. So if someone then says, but my favourite colour is for example, purple, then this would not accurately represent that. In that case we would have to have another dummy variable for purple, or we would have to create a missing value for them because in that case they are not actually reporting blue as the absence of votes, but they reporting purple here are missing value in that case. Another way of coding categorical variables, specifically ordinal variables is through integer encoding. So you will also often see this where we have quantitative, a qualitative value reported. So for example, here we have the highest degree a person has achieved and then we assign an integer to each of these values which still represents the ordering of these variables. So for example, if someone has a master's degree that would give them a higher value than someone who has a high school degree, high school degree, which in this case high school is assigned a letter of one or the integer one and a master's degree is assigned the number three. Very, very important. These are still ordinal variables. They might look numeric and your model might think they are numeric and then try to treat them as numeric variables. But it can't you can't calculate the mean, for example, of an ordinal variable because it's not meaningful. And the reason for that is in ordinal variables. We know the ordering of things. We know that, for example, a master's degree is more than a high school degree, but we do not know how much more. For example, an undergraduate degree in May is not twice as much as a high school degree, even though a two is twice as much as of one. So be very careful if you do have ordinal variables and use integer encoding because models do and will think they are numbers. They are not intelligent enough to not think so. So be very careful about that. In many cases, dummy encoding is a safer choice because you can also encode dummy encode ordinal variables that loses the order of things though. So there is always kind of a bit of a trade off between the two. Now let's talk about outliers. And this is where it comes back to. Why should we care about distributions? Extreme values are typically found by using probability distributions. So if we have excess, if we have a sufficiently large data set, typically we use a standard normal distribution. We assume that if we collect a lot of data, kind of all of the data on that subject, in most cases it roughly conforms to a normal distribution. Well, that's always true in real life is a completely different question, but it gives us a bit of a guideline to work with. So in this case, we have our normal distribution here, and this is even a standard normal distribution. And we can then look at how far away each observation is from the mean. And in many cases we kind of make a cut-off So we say something like we have a Three Sigma rule, which is a common rule of thumb, but it could be any two Sigma for sigma, it could be any of those. We usually use three and we say any observation which is further away from the mean than Three Sigma is considered an outlier. That comes back to your empirical rule, which we were talking about, where we think that most data is actually within 99, 99% of the data is actually within these three standard deviations away from the mean. If we think 99% of data is within that, then we can safely cut off the values that are outside of that by making use of that empirical rule which relies on the statistical distribution. So this is a common way of kind of cutting off the values. That being said, if you do cut off outliers, it's always important to first double check. How many outliers are you actually cutting off because you might have a mistake in assumption about the distribution of your data. So not every data set actually conforms to this, to that distribution. Oh yeah, with small data sets. In many cases we use student's t distribution instead of the normal distribution. We will talk a bit more about that when we talk about t tests in next week's lecture as well. When we cover more in depth the idea of using distributions to check whether data is actually conforming to our assumptions about that data. But same principle. We would still use a cut-off value of Three Sigma, for example. Multivariate outliers. What would we do with them? Exact same thing. We would just use multivariate Gaussians. So instead of using two dimensional Gaussian, you would have three or 4 or 5 dimensional Gaussian, You would still have a cut-off value, so you would still estimate your parameters of your Gaussian the same way you would do with much smaller dimensions. This is the same approach that we have. So you can still see we are still deducting the mean from each variable. We are still summing it up. We are still dividing it by the numbers of values that we have. The only difference now is that we are actually looking at a bold X will be looking at many more dimensions than just two or just one variable, but it's the same principle. Yes, this is one of these formulas that I don't expect you to study and kind of be able to to copy paste on your exam paper. Unless you find that topic really interesting and you find it very intuitive to remember that formula, in which case go for it. I would be happy about that. But this gives you the actual calculation for the likelihood that a pop up sample X is within that population, that distribution with Mu and your. What's the English word for that? Yes. Thank you. Yes. So same calculation, but for calculated likelihood that this is actually in that. So you can see we take basically this term here. I put it in here. Do you have to remember that? Not really. The reason why it's here is mostly for completeness sake, but I want you to take away from this is large enough data set Gaussian distribution, check the margins of it, large enough multivariate data, set Gaussian distribution, check the margins of it. That's the basic concept that I would like you to take away. So there's a second approach which is a bit more intuitive, and it's called the Mahalanobis Distance, which is a very, very interesting idea of thinking about distances of observations from the mean vector. The idea is very similar, so you can still see that we look at each of these observations or a matrix of observations. We look at their deviation from the mean and we calculate that as a distance. I personally find that quite an intuitive approach, the reason of which I will explain a little bit later when we talk about clustering, because a lot in clustering is all about distance and dissimilarity. So I find the concept of thinking of how far away, what is the distance of an observation to a vector or a collection of observations to a vector, quite intuitive. And you also can see that this distance then follows a square distribution and an outlier would be flagged very similarly to the idea of is it a way from the Gaussian distribution in this case, Is it a way from our square distribution? So similar concept, similar idea behind that. And this is again, where your distributions come in, where you think about if the data follows a distribution, what does that mean for data that does not follow this distribution? And there's a third factor, which I will very briefly cover because we'll cover that in more depth tomorrow in the tutorial because it is basically a really convenient scikit learn implementation where we look at the local outlier factor. So if we have a set of neighbours of point A and we look at the number of how many of these points are actually in the neighbourhood of our object. A So this comes back to this idea of distance, dissimilarity and neighbourhoods. The interesting thing about that, we will talk more about that when we talk about K nearest neighbour approaches as well. But the interesting thing about that is it adapts to the idea that how far a point can be away without being an outlier depends on the spread of the points in that area. So if we look at that visually, you can see that we have these dense points that are kind of clustered together in some of these areas. And if you then think about what does it mean if a point is far away from an area which is so densely clustered, that basically means they are likely an outlier. So the idea is that if there are areas which are very dense and you would you would assume that all points behave similarly because you have these density clusters, then if someone does not behave like these clusters, then they must be an outlier. That's kind of the idea behind that. And the beauty of it is that it's very adaptable. So what is dense is different for each data set. So in this case, this is a really, really dense point here. So we have this area which is extremely densely clustered on this side as well, this extremely densely clustered area. But normal density could look differently in a different data set. So this is adaptable to actually thinking about how far is normal spread. And it's also kind of similar to this idea if looking at Gaussian distributions again, because it brings back this idea of how far how far can the Gaussian spread to to the data. And if we use a standard normal distribution, we say this is exactly how far it should spread and no further, because the standard deviation is always exactly one. So this is what it should look like. Whereas this approach allows you for more kind of flexibility. So sometimes maybe the spread is further, sometimes it's a bit narrower. It depends on your data. I will go into a long rant when we talk about clustering about that because it's very closely related to a clustering approach called Dbscan, which stands for density based something clustering around points with noise. Something like that. It has a very complicated name. I call it Dbscan. The idea is that you look at neighbourhoods of points which are closely related to each other, close in some sense of distance or dissimilarity and points which are then not close to that are labelled as outliers. So Dbscan is one of the few clustering methods which automatically label outliers basically like that. So very closely related concepts. How do we then deal with these outliers, which we've just discovered most commonly? You will simply remove them from your analysis. We've talked again and again about method methods and measures which are sensitive towards outliers. The simplest way is to then remove them. However, don't just cut them off and be done with it. Always check how many outliers you actually have, because in many cases there might be a systematic recording issue. For example, if your data set comes from different sources and one of them measures income in a different scale and the other, then that might actually lead to you detecting more outliers than there actually are. How do you avoid that? You standardise your data, but you also always check whether there's a systematic issue. For example, all the outliers that you remove are from this one data set that you merged in. So in that case, you would think, okay, there's something wrong with the data set, not necessarily that they are all outliers. So are they actually outliers? Should they be considered in model? We briefly talked about that last week as well when we were thinking about these kind of customers and their behaviour with the mobile data mobile banking data. You might remember that scatterplot that we had and we had these two outliers who were elderly, people who were really using mobile banking a lot, and we said, should we move them? Should we kind of consider them as a model? And it depends on a lot of expert knowledge, objective oriented data processing. So don't just throw them away. Think about whether it makes sense to include them or not. If you do remove them. To me, it's best practice to record how many of them move in your analysis that can be in percentage or it can be in the actual absolute number. And that's for transparency sake that you don't just remove the data. In many cases. You also want to report specific details about them. For example, are they all outliers in one specific dimension, one specific data set? Are they all outliers towards one direction, or are they kind of equally cut off from the lower and the upper end of your distribution? And that comes back to this argumentation and making judgement, judgemental or judgements about each step of the process. Okay. Data quality. If you think back to our tourist visitor example from the last lecture. So let's think about what could actually impact the data quality of that tourist visitor data that we were collecting. And so, for example, if we think about we're collecting data on people who are visiting some local tourist attraction in Edinburgh. And we also already have our own data set, which we were collecting last time. Then there are there can be incompatibility issues between these two data sets. I was talking about that earlier when I was talking about the census data and the bringing in of the financial wellbeing data. So in many cases, these data sets are recorded differently and then you have to figure out what is the difference between the two. There are human errors, imprecise measurements, improper conversion of data sources, language differences, translation, technical infrastructure. All of that can be quite tricky in companies. In many cases, that means data that was collected years ago. But someone who is not in the company anymore and can't explain to you what the variables actually mean. So that can be quite tricky to handle. Okay, now you have to do a bit more thinking and have a look at the two data sets here. So we have, let's assume the upper one that was collected by by us and someone else gives us additional data that we can merge into that to create one large data set of 12 records, which is huge and amazing. So try to spot a couple of differences, a couple of difficulties that you can see when you start trying to merge in these these different data. I can spontaneously see three. Who has the first one? The date? Yes. So in one case, we have the date completely in kind of date time format. And this is probably saved as a string variable. Thankfully, many software packages now offer you quite easy day transformation functions, but to double check what they actually work. Okay. Second problem. Nationality. What's the problem with nationality? In the UK. It's different when you have granularity. Yes. So in one case we actually have each country of the UK listed and the other is just the UK. So if you try to merge them, you have a lot of data or information that is not available. And if you try to merge them, there's two ways that you could actually do that. One is you would create a new kind of column and these would all be missing values for your nationality variable. And the other is that you say this is actually not meaningful. I will ignore it and make these into UK. What would be the better watch? It depends on whether you think nationality is a meaningful and important variable for your problem. So if you think it doesn't tell you much about the behaviour, then transforming that into UK is valid because you're not losing important information. But it would be kind of a decision you would have to make. Okay. Third problem. We have others that. Could indicate other data that can indicate what subgroup within the. Mhm. Mhm. Yeah. Yeah, yeah. Absolutely. So if you have additional data you could use that to then infer a likely nationality for example for them. The problem is we do say nationality. So even if we have postal code we don't necessarily know where that postal code is coming from. Is it the home address, their work address? Is it kind of the nationality? Even it just says nationality? I don't ask whether this is whether you live, what do you identify as, for example? So if we have someone who lives in Scotland but identifies as English, they would have a Scottish postal code. So be careful with that. But yes, if you have additional information, do go ahead. Who sees more problems? Yes, sex and gender are recorded differently. And we also have missing values. That's quite common if you collect demographic data for some people, it's kind of sensitive, sensitive information that they are less likely to share with you. So that's why for things like age, income, gender, all of these variables are more difficult to collect in surveys, so you're more likely to have missing values. And also we have this encoding problem. So we have this as a numeric variable for sex and on behalf latter variable for gender. And we can also see that the the mapping between the categories that we have will not be very easy in this case because, for example, here on gender, we apparently gave more options for the people to actually take than in sex, which in this case is a binary variable. Okay. There's one more issue. Yes, the last one visited or not? It can be. That's why people do not visit and. They put one for now, but for the first one. Yes, I did it. And I will put one. Yeah. That comes back to whether the questions and two in these two data sets were collected exactly the same. So whether the definitions that people assumed in both data collections were the same, whether they were formulated, the questions were formulated in the same way. So how do we actually interpret the visited variable? Do each of these people agree on the definition of visited? There's one more issue. So I think they just the age to increase. So then I think. There is a person here who is just two years old but visited. Now, the question is, did they visit with a parent, for example, Is that just kind of, for example, the grand child of this person above them, or is that a recording issue? Should it be like 26 or something like that? It's impossible to tell and it's really difficult to then to then handle. And there might be one more issue. You did. The tender. Tender? Yes, we already covered that. Where we say we have the missing value. Oh, What do you think that might mean? Yeah. So it might be just a third category that we say. For example, they would like to add other category for that. Or we might also have someone who wants to self specify their gender or just another form of coding missing values. So all of those are possibilities. Yeah. No. 9292. Exactly. That's much older data than all the other records. So if we then say, for example, do we really want to include these extremely old records in our data set or would that skew our model? Yes. So that's a lot of issue for very small amounts of data that you might all encounter in your modelling. Let's talk briefly about missing values again. They can be caused by a lack of entering the value, so someone refuses to add. There might be an issue with data that is just left empty. So people sometimes start surveys and then just stop them in the middle of it or refuse to give you an answer. There also, there's also the issue that missing values can mean very different things. So if you have, for example, a survey, there are someone asks you what was your income last month? There are three ways that can lead to a missing value. They could completely refuse to answer the question. Leave it blank. Just not take anything. They might take something like don't know or don't remember, which is not exactly the same thing as saying I don't want to answer the question. So these kind of three options, refusal or forgetting the exact value or not taking anything, all of these would be coded as missing values in your data set. So, for example, in this case, we have someone who wants to apply for a loan and is being asked, did you have any previous loans? And they may leave that blank. But if you then change that question to have a more limited scope. So for example, if you have if you don't ask them how large of a loan that you last applied to, when did you apply for your last loan? Or you simply ask them, did you ever had a loan in the past? Yes or no? They might be more likely or more willing to actually answer that question. So there are ways in survey design that you can then minimise the missing value that you actually collect. In other cases, we are just not allowed to record anything in credit scoring famously. For example, we can't use gender as a variable in predicting the likelihood of the default of that person. One of the researchers in this school, which I would really recommend checking out the papers to her name is Professor Andreeva, and I don't know whether you've met her, but she publishes a lot in credit scoring. And some of the research is really interesting because it looks into the impact of leaving out gender as a variable in credit scoring models. And her research actually found that in many cases that is detrimental to women who it was included or introduced to actually protect. So the original idea was that the gender of the person should not be included because the thought process was credit scoring agencies might rate women lower, for example, because women historically were less likely to be full time employed and have had a lower income than men. Now, the actual funny or interesting story about that is that in most research actually showed that women had a better credit history than men. So including gender as a variable might actually be beneficial for women. But we are still legally not allowed to use it as a variable because it can lead to discrimination. So that's one of the interesting stories that you will also hear more about in your guest lecture in week five. Yes. Week five. Um, yes. Missing values. So how do we deal with them? One simple way is simply remove the observation. So row wise deletion of that, and it's usually useful if there are multiple missing values for the same observation, because in that case there might be something wrong. The person did not want to actually fill out that survey, for example, or there is a recording issue or a technical issue. In some cases it makes sense to completely remove it. Be careful not to remove too many, especially if there's a systematic issue behind that. So if you notice, for example, the example I made earlier, if all the missing values come from one data source, then removing all of them might actually skew your results because then you're training your model on just a select few sources and you're excluding a potentially important data source in that. And some models do allow for you to just keep in the missing values. So we'll see which models actually are able to handle missing values better than others. Neural networks, for example, do not like missing values. Decision trees are a bit more flexible with that regard. Imputation is another way of dealing with missing values. In that case, you would just try to guess what is the most likely value for that person. And there are tuple of different ways you can use the mean or the median to simply impute that value, which has the advantage that it doesn't really skew your model because you're basically just adding one more record with the same number as many others in their mean. And that's, for example, often used in neural networks. But it only really works for numeric values. You can't just impute categorical variables. You can create a new category for them, for example, not available or not applicable. Some people also impute categorical values as the most common option across the data set, which is kind of like the mean of a categorical variable, I guess. And in other cases people might also choose to model the most likely mean using, for example, segmentation processes. So you would look at, for example, the most likely categorical value for a woman aged 20 to 30. And then you would just look at that subset of your data, look at the most likely categorical value for that category or for that group of people, and then impute that value for your categorical value. Nearest neighbour correction coming back to this nearest neighbour approach. So that's why I actually cover and quite extensively in one of the future lectures because it's close to a lot of different, a lot of different concepts that we cover here. The idea here is that you calculate the nearest neighbours to your observations. So which records in the data set are the most similar to them? And then you impute the missing values through those neighbours. Problem is the number of neighbours is quite tricky to set. We'll see that when we talk about N and there's a couple of ways to try to overcome that. But in the end it's a subjective choice and any subjective choice that you bring into your model is kind of potentially skewing your results because it brings in subjectivity. You can also build a whole regression model to actually predict the missing value. So in that case, you would, for example, use all the remaining values to predict what that one value would be for that combination of factors. Yes. So I think this is I'm really good in timing today. So I would now like to take a couple minutes at the end to talk briefly about the assessment. So you've been notified of your course work groups on Learn. I hope you all had a look at those groups because they will be your group mates for the remainder of the semester And please check your number if you haven't done so already. And I have uploaded the assessment brief on unlearn for you to have a look at. The deadline is early December. Now that looks very, very far away to you, but it's really, really not. So what I want you to do now because now I actually have you all in the same room. I will now. Divide you into your groups. And then I would like you to start thinking about a structure for the process that you want to follow. The workflow structure. You're free to decide that and discuss that with your Groupmates There is no assessment on that part, on that part of the project. You're completely free to decide that. And I think project management skills are kind of part of what I hope you get out of that course. So if you feel like it's difficult to actually find a structure on the workflow with everyone else, see it as a learning opportunity because it really is. That being said, I would recommend that you write down your workflow with your group this week or by the latest next week because that will help you to structure the work for the rest of the semester and you hold each other accountable for actually doing what you write down. So some groups decide to write a proper contract out of that, so they write condition and it works the right conditions. This is our workflow. We will follow this workflow. If we are sick, we will let everyone know in advance via email. We will talk to each other on these days of the week and then you all sign that document and then you can hold each other accountable. And you don't have to argue when it gets really stressful. And in the middle of the semester I will recommend that process. That being said, you can all figure it out on your own. You're all old enough. This is what of workflow document can look like. You can make it as pretty as you want to, but this is kind of my recommendation. So get to know your group members over the next coming weeks. It's week two now, so next week is week three. Get to know your group members. Double check whether you have access to the information, double check whether everything works. Start reading a little bit about the background of the of the topic and get familiar with the data, etcetera, etcetera. So this is a proposed group work, a poll. My gosh, my English. I've been talking all day. I have to drink some water. Apple post structure. There you go. Very good. So let me pull up my trusty document. About a groups. If I can type my password. There we go. So first. Okay. There are 11 groups in total. We have 4 to 5 people in each group. You all have a phone or a laptop or tablet or similar. So if you don't know your group number yet, look it up now. And I mean now go and learn. Open up the course. Open up your assessment folder and you will have a document which gives you the exact number that you have just. Do you all have your group number? Yeah. Yeah. Okay. Listen to me now. Can all members of Group one raise their hand, please? Look around. Look at each other. Okay. Group one. Can all members of Group two raise their hands? Okay, good. Group three, please raise your hands. Look at each other. Wonderful. I see happy faces. That's good. Group number four. Can we see group number four, please? Very good grip number five. Get to know your group members. Good Group number six. Who is in group number six? Good. Make connections. Group number seven, please. Hey. Group them by eight. Very good. Group number nine, please. Okay. Group number ten. Who is group number ten? Very good. And last but not least, group number 11, please. Get it. Okay. Sorry. I hope you remember your group members. So please get together. Organise yourselves. You have ten minutes. Find your group members. Find a corner in the room, sit together, start talking. How do you want to structure your project over the next weeks?