

订阅DeepL Pro以编辑此演示文稿。  
访问[www.DeepL.com/pro](https://www.deepl.com/pro?cta=edit-document)，了解更多信息。

大家早上好 Good morning. 你好 - 你好 - Hello. Hello. 我知道你们能听见我 让我们安静下来。 你知道吗？ 在德国，有趣的事实是，讲座实际上是在海报时间后15分钟开始的。 所以，如果你在德国听讲座，上面写着 "上午 10 点开始"，实际上你会在 10 点 15 分开始。 我们称之为 "学季"。 我不知道为什么英国从来没有这样的规定。 我认为完全应该这样做。 因为我会减少迟到的次数。 非常抱歉有点迟到了 说起来，你们都站在门前，我也不知道为什么。 墙边有个电灯开关，就像你们进门时站在右边一样，所以不用觉得你们必须等我。 你们都是成年人了 你们可以自己去阶梯教室，不用等我，也不用砸场子。 至少我希望如此 我先去把门关上 也许这样能阻止迟到的人进来 好了，大家都好吗？ 第一周结束了 感觉好还是不好？ 不太好 哦，天哪 怎么了？ 很忙 第一学期有几门课？ 四门 四门 好的 四堂课 四堂辅导课 还有每堂课 所以，只是讲座或辅导或。 有意思 所以，你有一个伟大的组合 他们之间。 很好 那么 在本周的讲座中，我们将主要关注数据预处理。 所以还是有点无聊。 我想，下周或下下周，我们会开始讲有趣的内容。 所以，我想这可能是大家最期待的部分。 话虽如此，今天我们还将看看最激动人心的部分，那就是你们的小组作业。 因此，在讲座结束时，我会留出一点时间，让你们开始寻找小组成员，也许还可以讨论一下你们想如何进行，因为我认为尽早开始是有意义的。 这也是一个机会，你们可以就任务提出任何问题。 我已将简介上传到学习页面。 我还没有上传数据集。 原因很简单，我试图创建数据集，但代码崩溃了。 所以，现在我必须试着找出为什么它不想创建我让它创建的数据集。 我今天就会弄明白，然后上传数据集。 但我希望你们都能看到简介，它已经告诉了你们这一切是怎么回事，数据实际上是什么样子对今天来说并不重要。 尽管如此，我们还是要讨论一下，为什么在不考虑建模过程的情况下，尽早开始查看数据是很重要的。 如果你回想一下我们的理想研究过程小示意图，我们仍处于早期阶段。 在原理讲座中，我们已经讨论了如何思考研究问题，以及如何从可能遇到的业务问题中推导出研究问题。 因此，我们谈到了爱丽丝，爱丽丝可能会在本讲座的稍后部分客串一下。 我们在讨论爱丽丝应该如何处理一个公司向她提出的商业问题。 然后她收集相关数据。 在讲座中，我们基本上是这样说的，然后她开始分析数据，这是一种非常简化的方式，因为第一步是进行任何类型的分析，不管是我们将在原理中介绍的更传统的统计分析，还是我们将在本系列讲座中重点介绍的更侧重于机器学习的方法，爱丽丝将采取的第一步始终是初步查看数据集，然后尝试找出其中是否存在任何问题。 今天我们就来谈谈数据的问题。 现在，在我们开始之前，先整理一下内务。 我收到了一封关于阅读清单的邮件，很明显，课程结构中发布的阅读清单与通过学习上的图书馆资源共享的阅读清单不一致。 因此，如果你点击图书馆资源，它就会弹出一个窗口，为你提供一些结构和实际作品的链接。 因此，如果有疑问，请按照课程描述进行学习，这是我上传到课程信息unlearn上的PDF文档。 上面有一个表格，告诉你每周将涉及的主题，以及我希望你阅读的章节。 因此，该文件中的表格始终是最新的，如果我做了任何修改，而图书馆资源中没有反映出来，这些修改也会反映在该文件中。 因此，如果有不一致的地方，请按照课程说明进行修改。 同时，我也会尽量更新图书馆版本。 不过，我自己不能直接更新，所以必须通过图书馆。 我想那是图书馆办公室，我会告诉他们，嘿，你们能更新一下吗？ 因此，与更新 PDF 相比，这对我来说是一个有点漫长的过程，这就是为什么如果有任何疑问，我希望你能看看 PDF。 是的，我们在本讲座和校长讲座中都谈到了上周的内容。 在上周的讲座中，我们谈到了什么是预测模型。 我们谈到了如何利用信息做出决策，以及如果你要做出一个非常复杂的决策，你收集到的信息并不能真正被你自己分析出来，这就是预测建模的用武之地。 因此，预测建模基本上是通过预测未来可能发生的事情来帮助你做出决策。 我们通过理想版和现实版的预测建模过程谈了一些。 我们还谈到了有监督学习和无监督学习之间的区别。 你应该还记得，监督学习基本上描述的是这样的问题：你有一个标注过的数据集，并据此建立模型，然后你可以将该模型应用于新的、未见过的数据，而非监督学习则不同。 在无监督学习中，没有两步过程，只有一步是对数据建模，通常包括某种分割问题。 在很多情况下，我们谈论的是回归和分类。 回归是指尝试预测一个数值。 例如，您可以预测下个月的股票价格，而分类则是对新数据点进行分类。 例如，您可以将客户分类为对公司没有好处的客户、有利可图的客户或利润不高的客户。 这也可能是一个多类问题。 因此，您可以尝试预测多个类别。 我们刚才谈到了变量的类型。 我们谈到了分类变量和序数变量，所有这些类型。 我们还简要地讨论了预测建模所面临的一些挑战，我想我还对P值的使用进行了一番咆哮。 我还记得，我似乎谈到了人们过于依赖理论，而对数据依赖不够。 因此，我的一些烙印就留在了讲座的那一部分。 所以，今天我们来谈谈数据清理和准备工作。 我们将讨论预处理标准化，我们将讨论变量编码，我们将讨论异常值。 它们是什么？ 我们如何发现它们？ 我们该如何处理？ 我们还将讨论缺失值。 我们先来介绍一下数据预处理。 你可能会问自己的第一个问题是，我们为什么要对数据进行预处理？ 相信我，如果你曾经处理过真实数据，你就不会问自己这个问题，因为这很明显。 数据永远不会以你想要的正确格式到达。 就是没有。 我见过绝对糟糕的数据集，也见过存在编码问题的数据。 例如，不同的语言有不同的字母系统。 因此，有时如果你试图将一个数据集翻译成另一种语言，那么计算机基本上会试图对这些字母使用独特的编码，结果就会出现问题，因为 Excel 对此并不满意。 然后你的整个数据就是一团糟。 然后，如果你试图将这些数据与实际完整的数据合并，那么你就会造成更加混乱的局面。 因此，数据并不是以正确的格式到达进行分析的，你应该在开始分析时就预料到你将不得不进行相当广泛的预处理。 这对你意味着什么？ 给自己留出时间去做。 所以，如果你考虑到，比如说，课程作业将是一个简而言之的小分析，不要认为你可以立即应用所有你想应用的模型，因为它不会起作用，然后你会想为什么它不起作用。 然后，你必须回到数据中，开始预处理。 因此，在分析开始时，要给自己留出时间，弄清楚在开始分析之前，我们究竟要先做什么。 更令人兴奋的部分是建立模型。 我们说预处理或清理是什么意思？ 基本上，这意味着你要创建一个适合你所选模型的数据集。 因此，我们可能会说，我们需要清理数据。 我们需要稍微整理一下。 这包括所有类型的预处理步骤。 因此，处理缺失值、处理或检测异常值的变化、检测变量名是否不清晰，这是一个大问题，还有变量编码不一致的问题。 如果你想同时处理多个数据集，这是一个非常大的问题。 上次我向大家介绍了我最近的一个研究项目，在这个项目中，我们研究人口普查数据，同时还试图研究人口普查数据和来自调查的不同纵向数据集。 因此，我想检测的一个问题是，我想创建一个我们称之为合成人口的数据集，这个数据集在统计上可以代表真实的人口，但却是利用人口普查数据人为创建的。 因此，我们需要一个人口普查数据集。 基本上，我们需要两个人口普查数据集，因为在许多国家，至少在我使用数据的加拿大，人口普查数据的发布方式是有两个文件的。 其中一个是家庭样本文件。 因此，并非每个家庭都必须填写政府寄给他们的小调查表。 这只是一个样本。 政府正试图创建一个能够代表人口的样本。 因此，某个城市有几千个家庭记录。 例如，我们关注的是蒙特利尔，然后你也会得到一些整个地区的汇总数据。 这种汇总数据概括地描述了城市的人口。 例如，我知道有多少特定年龄段的人居住在蒙特利尔，或者居住在蒙特利尔的某个人口普查区。 因此，现在的问题是，我们要解决的第一个问题是，如何将这两个数据集结合起来，创建家庭人口？ 我们的目标是创建一个能够代表蒙特利尔所有家庭的数据集。 根据这两个数据集，我们知道了整个人口的总体情况，我们也有了这些单个家庭记录的样本。 因此，我们从这些住户记录中进行推断，最终得出的数据集与我们已知的汇总数据相同。 因此，我们知道人口应该是什么样子，也知道了汇总条件。 我们使用家庭样本并从中抽样，然后进行推断。 因此，我们从这些数据中提取出一个更大的数据集，然后仔细检查结果是否与我们所知道的真实人口相近。 然后，我们就可以使用该人口数据集进行各种建模。 我提到过，我的兴趣之一是财务福利，所以我对人口是如何扭曲的很感兴趣。 我对人口的财务状况以及人们对自己财务状况的感受很感兴趣。 例如，债务水平、对未来的焦虑程度、收入多少等各种不同因素。 因此，我们在蒙特利尔也有一个关于人们及其财务状况的代表性样本。 现在，我们正试图将样本与我们的人口合并。 因此，我们会寻找在某些人口统计数据方面相似的记录，然后尝试将我们的财务状况数据与人口数据合并。 这基本上就是我最近正在进行的研究项目之一，我们称之为合成生态系统，因为我们正在将合成数据集和推断数据集层层叠加，创建一个完整的人口。 然后，我们使用基于代理的建模等方法，尝试在这种由多个连接数据集组成的环境中模拟人类行为。 我之所以告诉你们这些，除了研究人员喜欢谈论自己之外，我保证还有一个原因，那就是我们面临的最大挑战之一就是数据集的格式各不相同。 因此，其中一个数据来源是政府。 我们使用了人口普查数据，另一个数据来源是加拿大金融消费者机构（Financial Consumer Agency of Canada），他们为我们提供了财务状况数据。 然后，我们还研究了健康数据，例如，来自加拿大老龄化纵向研究（Canadian Longitudinal Study on Ageing）的数据。 所有这些数据都有不同的格式，因此它们都有不同的变量名。 他们对变量的编码方式也不同。 例如，如果要合并年龄类别，它们必须匹配。 但如果其中一个收集的数据是 18 岁至 25 岁，然后是 25 岁至 30 岁，而另一个收集的数据则更宽泛，比如 18 岁至 35 岁，那么这就不能直接匹配。 首先，你必须改变每个变量的编码，使数据集之间相匹配。 因此，我们所做的第一步就是在变量之间建立某种映射关系，试图解释其中一个变量是如何映射到另一个变量的。 这就是我在幻灯片上简单描述的编码不一致的问题之一。 但你可以看到，这个看似很小的问题，在研究项目中是如何成为一个非常非常大的问题的。 如果你低估了这个问题，就会从研究项目中耗费大量时间和金钱，因为时间就是金钱。 因为时间就是金钱，你必须给那些真正为研究项目工作的员工发工资。 如果突然要花两倍的时间，那代价可能会非常昂贵。 好了，让我们回到幻灯片上来。 预处理的重要性在于它有助于解释。 很明显，例如，如果你修改了不清晰的变量名，那么它就能帮助你解释数据的含义。 在大多数情况下，这对模型性能也是必要的。 遗憾的是，没有一个基本的步骤可以让你照着做。 学生们经常问我，能不能给我一份清单，列出我应该做的所有事情，或者诸如，我应该测试哪些模型，或者我应该先运行哪些模型？ 这其实没有一个循序渐进的过程，因为这在很大程度上取决于你的研究问题，你想找出什么，以及在特定情况下数据和模型需要什么。 尽管如此，还是有一些常见的步骤需要检查，我把它们放在这里，算是给你一点指导。 例如，如果你是第一次接触真实数据，这对你的小组项目也是一种很好的指导。 这样说吧，这给了你一些起步阶段的指导。 因此，首先要检查数据格式，你不会相信有多少数据实际上是错误的，而不是偶然的。 我收到的数据中少了一个完整的变量，我就想，哦，难道他们不允许给我这个变量吗？ 我不确定，因为他们肯定应该给我。 于是我给他们发邮件，结果是他们那边的上传出了问题。 他们告诉我，哦，是的，我们有点忘记把那个变量上传到服务器了，我们会解决的。 然后我就突然有了。 所以，在很多情况下，要检查你应该有的变量和其他变量是否真的有，还要检查其中是否有任何错误，是否有任何明目张胆的问题，比如人们错误地编码了邮政编码，一些非常小的问题，这也涉及到检查文档。 因此，在我的案例中，已经提到了这一点。 我将再次提及他们 200 页的人口普查文件。 我还没看完。 所以，检查文档，检查样本量。 所有这些小问题都会对你以后的工作有所帮助。 不要只看数据。 首先要试着理解数据想要告诉你什么，然后初步检查是否存在明显的大规模问题，即完全缺失邮政编码的问题。 还要检查系统性缺失数据的数量。 我们稍后会讨论缺失值，这就像是奇异值缺失的小问题。 我们会讨论为什么会出现这种情况。 但要注意数据中可能存在的系统性问题。 然后，描述性统计会探索数据及其外观。 请始终从描述性统计开始。 不要把所有东西都扔进神经网络，希望它能起作用。 请尽量采用简单的步骤。 首先，通过描述性统计探索数据。 进行虚拟编码变量转换。 等等。 接下来，处理异常值，处理缺失值，然后，如果有必要，你可能需要重复进行剥离统计。 我很少看到有人提到这一点，但我认为这是非常重要的一步，因为如果你处理了大量的缺失值、异常值，所以你做了大量的数据和信息删除，当然，你肯定想仔细检查这对你的描述性统计有什么影响，你想看看变量删除对你的数据集究竟有多大影响。 然后检查模型假设，选择模型，然后就可以开始有趣的建模过程了。 因此，这种预处理总括起来有很多子步骤。 今天，我们将重点讨论其中的第 3 步至第 7 步。 我们将讨论描述性统计。 首先，我们将讨论虚拟编码、变量转换、异常值和缺失值，然后显然是描述性统计，因为这也已经是第三步了。 好的，让我们从摘要统计开始。 现在，这对你来说可能有点重复，因为这与我们上次在《数据分析原理》中讲到的内容非常相似。 我之所以把它放在这里，是因为它真的很重要。 所以，我想多次告诉你，但也是为了完整起见，让你知道这也是这门课程的一部分。 这不是两门独立的课程。 我记得在第一堂课上，你们的一位同事问我：这两门课究竟有什么联系？ 这两门课之间有什么联系？ 在这两门课程中，你会看到一些重复的步骤。 我会简短地说一下，因为我知道这可能有点重复，但为了完整起见，我确实必须说一下。 所以，请耐心听我讲完。 首先，非常简短地说，描述性统计基本上是一个总括性术语，所以你会看到描述性统计、汇总统计，所有这些词都可以互换使用。 严格来说，描述性统计是统计本身的一个分支，而创建汇总统计是描述性统计的一个步骤或一个部分。 它还创建了统计领域。 它还包括使用可视化技术来描述整个数据集或单个变量的分布。 因此，其中一些统计量可以或通常是按变量计算的，它们通常描述变量的一些不同属性。 例如，我们可以查看平均值、方差或价差，还可以将多个变量或单个变量可视化到图中，以描述整个数据集。 你上次看到的频率直方图就是一个例子。 因此，正如我提到的，这是任何数据分析项目中非常重要的第一步。 重要的是，它们还可以用来交流初步发现。 因此，你稍后会看到的很多模型都很难可视化，也很难向普通观众描述。 例如，神经网络就是一个很难向普通观众解释的模型，因为它是一个非常著名的黑盒子模型。 我们实际上并不知道它们在模型中做了什么。 我们只知道什么东西进去了，什么东西出来了。 然后我们试图解释模型的准确性或误差。 但是，描述性统计是与经理等人沟通的重要工具之一。 因此，如果他们要求你创建某种研究项目并进行分析，例如，回到爱丽丝的话题，分析人们对某种苏打水产品的态度，那么如果你只是告诉他们，嘿，我创建了这个非常复杂的模型，这里有三个数字。 这是准确率，这是 F 统计量，然后我会告诉你，是的，好的。 但这意味着什么呢？ 数据到底是什么样的？ 如果你只告诉他们数据集或模型是这样那样的，人们会怎么想？ 模型有奇异值或这样、这样、这样的准确性。 他们想要的是一个故事。 因此，他们真正想要的是你能够传达一个连贯的故事。 所以，一个完整的故事情节，数据本身能告诉我们什么，什么样的模型能告诉我们数据本身无法告诉我们的东西。 这就是为什么描述性统计非常重要。 还记得我们在讨论什么是预测建模时，我们说它是一种决策工具，但其中的第一步是我们要思考我们用来创建或做出决策的信息。 描述性统计就是描述信息的方法，非常简单。 因此，你会从原理中认识到一些非常基本的概括性统计，例如，用平均值来描述数据的中心倾向，用方差或标准差来描述平均值周围的散布情况。 有趣的是，有时你会发现，报告平均值、中位数还是模式取决于数据，也取决于你想要传达的信息。 所以在很多情况下，比如有人会问你：嘿，这个问题的这个变量的平均值是多少？ 但你必须把它理解为这个问题的中值是多少？ 因为，你知道，我问过在场的一位统计学家，平均值并不是一个合适的报告指标。 因此，如果有人问你平均值，那么在接触中，他们也可能对中位数感兴趣。 你只需对此加以解释，并知道哪种中心倾向的测量方法才是重要的。 因此，我们知道算术平均数，我把它放在这里是因为它是最常见的，计算方法很简单，就是将所有数值相加，然后除以数值的个数。 我们还知道，方差是根据每个数值的偏差计算出来的。 从平均值出发，将这些差异的平方和相加，再除以测量值的数量减一。 如果取方差的平方根，就得到了样本标准偏差。 你还可以看到，我们在这里对样本进行了划分。 这样做的原因是，你很少有关于整个人群的数据。 因此，通常情况下，你会说你计算了样本平均数，但这在技术上是不正确的。 如果你只是说你计算了平均值，因为这是不准确的，你计算的是样本平均值，或者你计算的是总体平均值。 尽管如此，大多数情况下都只说平均数、中位数或方差。 你必须从文本中推断出计算中所指的是哪种值。 这其实并不重要。 唯一的区别是，你要把整个群体的所有元素加起来，然后除以群体的数量。 是的，我提到了中位数。 你应该记得，中位数是第二四分位数。 因此，它是数据的中点。 它将数据分成两半。 50%的数据在它的左边。 50%的数据在其右边。 然后，第一四分位数和第三四分位数就是数据的分界点。 所以，如果你已经取了中位数，你可以把它们看成是剩余数据的中位数，如果这样做有意义的话。 因此，你可以把数据切成越来越小的碎片。 有一件事让我觉得特别困惑，也一直忘不了，那就是四分位数和四分位数或百分位数之间的区别。 这三个词经常被混淆，尤其是量值和四分位数。 区别在于，四分位数指的是百分比。 因此，你可以这样来计算。 你可以记住量级指的是百分比，四分位数指的是季度。 所以它们总是第 25 位。 因此，第一个、第二个、第三个、第 25 个四分位数、第 50 个四分位数和第 75 个四分位数。 我还可以看到这里有一个小箭头，因为这应该表示第 25 个百分位数是 X 尾数。 好的。 我们讨论过的范围也是整个数据集的最大值减去最小值。 我们还讨论了对离群值的敏感度。 如果你有一个极大值和极小值，那么你的范围就会受到极大的影响。 我们还讨论了四分位数间距，并在此基础上讨论了盒须图。 你应该记得，这些方框基本上描述了数据的分布。 说到数据的分布。 我们这里有几种可视化图供你参考。 这里有几个不同的直方图。 其中一个是你理想中的漂亮直方图。 左上角的一个直方图，可能就是我们所说的接近正态分布的直方图。 所以，如果你有这样的数据，那就是正常的。 这就是正态分布数据。 在现实生活中，没有比这更好的了。 遗憾的是，我们也有白斜和左斜分布。 你们应该还记得上一堂课，我就是把这两种分布搞混了，因为我总是忘记哪种是右偏分布，哪种是左偏分布。 在我看来，左偏分布的左侧数据较多，这是有道理的，但实际上左偏分布描述的是分布尾部偏右的情况。 偏斜数据的尾部在右侧。 左偏数据的尾部在左边。 现在，在你们都睡着之前，谁愿意告诉我这三个数据集的箱形图是什么样的？ 谁有勇气试试我们的绘图功能？ 也许吧，谁愿意给我画出这三个数据的方框图？ 非常粗糙的。 不一定要漂亮，但必须有一个人完成。 谁够勇敢 不，不，不 其中一个？ 其中一个？ 你想确认吗？ 来我这 别害怕 我觉得这些只是为了脱身 你也是 你过来 拿着这支笔 并希望它的工作原理。 确实有用 所以，这一个就像。 嗯。 我的意思是它看起来像。 就是这样 想想 X 轴到底描述了什么？ 我们有这么多。 比如说 像 都是90度 所以，是的，非常好。 也许我们可以有这样的。 是的，没错。 在这一点上是25％。 没错 所以这是第一个四分位数，25%的四分位数。 第1个可能还是最小值 最大值 因为它是倾斜的，所以右边的值更多。 所以可能是 75%。 看看我是怎么跟老板说的 嗯 - 嗯 - Hmm. 嗯 好吧 所以，作为更多的值 在这部分上的左侧。 没错 所以像这样更小的值。 对 Yeah. 想想这个条形图描述了什么？ 这是你的 是平均数还是中位数？ 我认为是媒体。 是中位数 没错 所以在这种情况下，你可以看到中位数被描述为1.0。 所以在这幅图中，中位数应该在左边。 所以，我个人可能会拿起笔。 我个人可能会把中位数画在这里的某个地方。 好的 因为它更接近左边的数值。 所以比较封闭。 所以，如果你有一个方框图，你的数据是分散的，那么中位数就描述了我在哪里做错了。 我老是忘记数据是否偏左、偏左到偏右。 50%的数据应该在中位数的左边。 所以是 50%。 所以更大的部分应该在左边，对吗？ 对 好的，就这样 很好 非常感谢 Very good. 非常感谢 也许我会给别人弄错案例。 喜欢只是想出更多。 不，这很好。 要记住的重要一点是，方框可以看作是直方图的代表，因为方框的白色实际上就是直方图的样子，如果你画出来的话。 在这种情况下，我们还可以讨论离群值。 你应该还记得，方框图的边线实际上描述了离第一和第三四分位数有多远。 数据更加分散。 你可以像你的同事演示的那样，一直画到最小值和最大值，也可以考虑是否要在某个点上停止画须，特别是 1.5 倍四分位数范围。 在这种情况下，你可以将离群值标注为胡须两侧的小圆点。 因此，如果我们在这里讨论离群值，你可以决定在离群值的部分长度上画一个方框，就像这样。 然后，你也可以在图中的这一部分用点来表示异常值。 好的。 在第一个问题中，你认为哪种汇总统计量最适合用来掌握偏态分布中位数的中心倾向？ 谁愿意用一句话告诉我原因？ 否则，我很难回答。 给你们所有人。 是的。 平均值会受到异常值的影响，同时也会受到倾斜度的影响，如果你仔细想想，一个大型的、非常强烈倾斜的数据集通常会显示出异常值。 因此，我们有这种非常长的尾部分布，你可能会认为其中一些会被归类为异常值。 所以是的，均值对偏斜分布和异常值很敏感。 所以，我们还是用中位数来代替吧。 现在我们有两个严重的问题。 现在，如果你有纸和笔，或者打开 Excel 表，就可以进行快速计算，告诉我这些序列的中位数和平均数分别是多少。 我也会这样做。 好的 是的 Oh. Okay. 谁比我更快？ 谁已经有了这两个系列的平均数？ 是的，4. 7，3. 4. 大家都同意 是的，完美 谁能告诉我这两个数列的中位数分别是 3. 5 和 2. 5？ 好极了 你们都记得如何求中位数吗？ 因为我觉得这可能是我们要记住的最棘手的一个问题。 我们是如何找到中位数的？ 看到数据。 然后找数据的中点。 没错 哦，天哪 我为什么要把这个问题放在上面？ 现在我得计算了 好的，根据这两个序列，查找标准偏差和四分位距。 是的，你必须计算这个，因为我也有两个，所以我会用电子方式计算。 这个 咦？ 我学到了新东西。 我学会了如何使用量化函数。 Excel 看着我 你可能会喜欢，但实际上我已经有一段时间没用过这个函数了。 好吧。 谁看看这两个数据的标准偏差？ 要不要我演示一下？ 我会分享我的屏幕 好的。 在计算标准差时 要使用下面的种群。 样本公式，因为在分母中。 你认为会有什么区别？ 我的意思是，如果它是负一，那么它就变成了一个估计值。 所以它是针对样本的，也是针对人群的。 我想知道给定的序列是否是样本的总体。 这真是个好问题。 所以，我会把它们定义为群体，因为对我来说，它并没有定义我们所拥有的这些数字之外是否还有其他东西。 所以对我来说，我认为使用总体是一个合理的假设。 是的。 但你也可以说这只是个样本。 所以这是一个需要讨论的问题。 所以你可以用其中任何一种。 所以我对两者都很满意 你有一个给我吗？ 是啊，我其实。 正在考虑 正在考虑 很好，有人有答案吗？ That's good. 有人有答案吗？ 有 Yes. 是2. 19 如果我们用人口数量来算第二个，是3. 03 03 - 对 很好 有人同意吗? Very good. 有人同意吗？ 不同意 同意 - 很好 - Yes. 很好 Very good. 我们还有四分位距吗？ 我认为这是一个有点棘手，因为它需要一点点时间。 任何人，任何人或任何人有任何四分位数，已经给我们。 我会有一个正确的方向的第一步，也许在房间的这一边，因为你一直很安静。 避免眼神接触 是啊。 对 六。 我是 "是"。 是2还是6？ 我不同意 我不是不同意 I'm double checking. 你呢？ 我害怕 好吧 而对于第三个。 好吧。 IQ. 啊，是的。 我的问题是，为什么呢？ 我的计算是6. 6.6.6. 3.60. 对，6. 0 有意思 所以我得再和Excel核对一遍 因为Excel认为应该是5. 5 5. 5 5.5, 5.5. 因为Excel给我的四分位数是7.5和2，所以两者之间的四分位数范围是5.5。 5. 你有十个数字在两个之内，彼此是两个整数。 所以要数 第三个数 第八个数 第三个数 这个数 只有整数 没有小数。 因此，除非我现在做错了，否则小数点的计算可能会有所不同。 我想这可能就是 Excel 计算平均值的不同之处。 是啊，是啊。 所以Excel 所以 Excel 实际上，我认为这可能就是为什么我们得到 6 或为什么我们得到 7.5 的区别所在，我们得到 5.5 或 6 正是如此。 我们要找的是平均值，因为它是偶数。 四分位数 这真的很有趣。 这也让我想查一查 Excel 是如何计算四分位数的，所以我会把这作为我的家庭作业，并尝试找出 Excel 与我给出的公式的区别。 要进入四分位数。 你还是得找出两者的区别。 是啊，这不是直接的。 是啊 据我所知 没有直接的公式 好吧 - Yeah. 好吧，在休息之前 现在，在休息之前，让我们来简单了解一下统计分布。 首先，我经常被问到这样一个问题：我为什么要关心统计分布？ 举几个简单的例子，它们可以帮助你检测离群值。 我们稍后会看到。 在很多情况下，它们真的很重要。 如果你对模型做出任何假设，例如在回归中，它们也可以帮助你对新数据点的位置做出假设。 这就是我们如何使用经验法则或切维-切斯不等式的一个例子。 现在，让我们简要地看看几个更常见的统计分布。 我们将在明天的讲座中进行更深入的讲解，这也是为什么我对它们略显轻描淡写的原因，因为我觉得如果我对这两种分布讲得太详细，会让人觉得很无聊。 均匀分布，这是一种简单的分布，并不精确。 你可以看到有一些偏差，但离散数据的均匀分布，也就是本例中的离散数据，看起来就像你在两个设定值之间的每个值都有相等的概率。 因此，举例来说，你会决定这是你的值，这是你的值。 A 哇，我的笔迹，这是你的值。 B 在这个区间内，在一个连续的数据空间中，每一个值的概率都是相等的，这一点比较明显，你有 A 和 B，在这个区间内，这两个值之间的概率是相等的。 因此，均匀分布、等概率分布、正态分布。 我们已经说过了。 这就是我们的钟形分布。 我们之前看到过这个直方图，它在现实生活中可能是什么样子。 我们还谈到了标准偏差和平均值。 在这种情况下，我们可以看到这是一个群体均值和一个群体标准差，对于标准正态分布也称为高斯分布，我们会说均值为零，而均值周围的标准差正好为一。 正如我所说的，下周我们将进一步讨论这个问题。 二项分布。 二项分布的概念是先画图，然后观察每次画图的结果。 一个常见的例子就是你有时会做的颜色编码抽球，球要么是黑色的，要么是白色的，然后你数一数你到底抽到了几块黑板和几块白板。 伯努利分布是其中的一种特例，即你只抽一次，常见的例子就是掷硬币。 泊松分布是指在一定时间内发生的事件。 一个常见的例子是，你观察一台机器，数一数在特定时间内机器坏了多少次？ 因此，你要观察事件发生的情况，并尝试建立固定比率模型。 这就是机器损坏的次数？ 例如，一小时内发生五次，那么在观测次数足够多的情况下，你的 lambda 值就是五，而 lambda 值则大于十。 因此，如果事件发生率较高，泊松分布就会开始看起来像正态分布。 我们还有指数分布来描述独立事件之间的等待时间，你可以看到指数分布和泊松分布之间的关系。 明天的讲座中，我们还将进一步讨论两者之间的确切关系。 所以，明天上午 9 点，你们就可以期待美丽的概率论讲座了。 好了，我想现在是休息一下的好时机。 我们休息十分钟，喝点水。 然后我们再回来讨论变量转换、离群值等问题。 现在我们已经讲完了基础知识，10分钟后见。 好，我们回来坐下。 挑战 好的 有两个问题我觉得非常有趣。 一个是 我们到底有什么？ 特别是考虑到阅读。 哪些概念对我们来说很重要？ 或者说，比如说，我们实际上基本上必须学习和理解哪些内容，并能够在考试中重申哪些内容？ 我想我们会在课程的后半部分详细讨论考试。 我希望你们记住的重要一点是，考试将以论文为基础，所以你们不必记住公式，然后进行计算或类似的事情，因为本课程的目的不是教你们如何计算分布的概率值或类似的东西。 这不是目的。 这门课程的目的是让你掌握批判性思维技能，能够遵循建模过程，评估结果，并在实际环境中讨论这些结果。 因此，我真正希望你们从这门课程中获得的是。 更抽象，所以更多的是能够理解存在哪些类型的概念，以及如何在实际生活中使用这些概念。 这意味着你不必参加考试。 我不会给你一张数据表和一个计算器，让你计算数字或类似的东西。 我不觉得那有用。 如果你觉得公式对学习概念有用，那么我非常欢迎你在考试中使用它们来解释一些东西。 因此，很多学生实际上觉得这很有帮助，比如说，如果他们谈到回归模型，给我一个回归的公式，然后解释每一个组成部分，然后解释回归的哪些概念或考虑是有用的，为什么，以及与这个公式的关系是什么？ 另一些人则认为抽象的概念更容易记忆，因此更注重概念思维。 对我来说，两者都是。 只要你在节目中，两者都是好的。 对模型有很好的理论理解，并能将理论应用于实践。 希望这能回答你的问题。 我的另一个问题是关于统计分布的，也是我轻而易举就能解决的原因。 明天我们将详细讨论这些问题。 我还有一个问题是，具体到现实世界的应用或考试中，这些分布的含义是什么？ 这其中有几个步骤。 其一，在很多情况下，这更像是一个视觉问题。 因此，在很多情况下，你实际上是在绘制数据图，例如数据的直方图，然后你会想，什么是可以解释该数据的可能分布，这有助于我的建模过程。 然后，你可以进行统计检验，检查数据是否符合分布。 下周，我们将在《数据分析原理》系列中讨论变量分布测试。 例如，当我们谈到假设时，不仅要测试样本平均值的差异，还要测试其是否符合分布。 下周我们将进行统计测试。 因此，我们之所以要在这一特定讲座中介绍统计分布，实际上是为了进行数据预处理，我们现在就来讨论这个问题，因为检测异常值的方法之一就是检查异常值是否符合其余观测值的分布。 举例来说，如果所有观测值都遵循正态分布，那么你就可以检查这些值中有哪些超出了分布的预期值，从而成为异常值。 这就是为什么我们在这里介绍统计分布，将其作为检测超出预期值的工具。 在讨论异常值缺失之前，我想先谈谈变量变换。 对我来说，这是最重要的预处理步骤之一，尤其是在社会科学领域。 例如，你可能会在同一个数据集中得到分类数据、序数数据和数值数据。 如果你想对这些数据进行更严肃的建模，你首先要做的就是转换这些变量。 首先，我们必须将这些变量，尤其是分类变量、数值变量、分类变量和序数变量转化为可以建模的数据。 在讨论定量数据与定性数据时，我们已经略微提到了这一点。 现在，我们来看看如何在这两者之间架起一座桥梁。 那么，我们为什么要转换数据模型，要求它，要求数据在一定的范围内，例如，变量以不同的尺度测量，这对模型的解释和估算是很困难的。 因此，标准化是第一步。 我们上次在这里或在原理讲座中简单介绍过这一点，我们谈到了如何使我们的数据基本无量纲化。 因此，标准化是几乎所有建模过程中最重要的预处理步骤之一。 因此，在几乎所有 所有模型中。 人们首先要做的就是将数值数据标准化。 这意味着什么？ 基本上就是为每个数值创建 Z 值，从而扣除其平均值。 在这种情况下，我们使用群体平均值和群体标准差。 在这个公式中。 虽然扣除平均值后再除以标准差，但与样本的计算方法相同。 因此，这意味着如果对数据集中的每个变量都这样做，那么所有这些值的均值都是 0，标准差都是 1，同时还能保持变量分布的形状。 这意味着数据变量实际上变得具有可比性。 举例来说，如果你有一个变量，比如以数千英镑为单位的收入，然后你有另一个小得多的变量，比如一个小的数值变量，比如完美年龄。 年龄是一个更小的变量，可能以 0 到 100 的范围来衡量。 你可以看到刻度上的差异。 如果不使用模型，那么较大的数值就会压倒模型，而最小的数值就会变得微不足道。 但是，如果你想确保在相同的尺度上进行处理，使其具有同等的重要性，特别是你会使用标准化来使这些数值具有可比性。 与标准化非常相似的一种方法是最小值缩放，我们也称之为特征缩放。 也有人称之为归一化。 归一化也可用于标准化。 有很多术语用于非常相似的概念，但使用方法却大相径庭。 这里的区别在于，我们不使用平均值和标准差，而是使用分布的最小值和最大值。 这也可以将变量缩放至 0 和 1 之间，并保持形状。 不过，由于我们使用的是最小值和最大值，因此它对异常值相当敏感。 明天，你将在计算机实验室看到这两种方法的对比。 我们将同时实现这两种方法。 然后，你可以看看这两者之间的比较，以及看起来像什么。 现在，我提到了分类变量的编码。 通常，如果你的数据集中有一个分类变量，你会使用我们称之为假编码的东西，我们也称之为一热编码。 之所以称其为一热编码，是因为对于每个观测值，例如这里有五个人报告了他们最喜欢的颜色。 我们会为每一个人创建一个虚拟变量，并给他们一个值。 所以是一个 "热"。 这就是为什么叫一热编码。 例如，如果某人报告说他最喜欢的颜色是绿色，我们就会给绿色虚拟变量赋值 1。 你还会发现，虽然我们有四个可能的颜色值，但只有三个假值。 这样做的原因是，我们要避免所有的虚拟值都完全相关，所以我们总是比我们的值少创建一个虚拟值。 你仍然可以看到，我们可以通过把所有值都赋为 0 来说明有人会选择 "蓝色 "作为选项。 因此，这仍然完美地代表了某人投票给蓝色，即某人不投票给红色、绿色或黄色。 尽管如此，这只是一种情况。 显然，如果这四个选项都是完美的。 因此，如果有人说，但我最喜欢的颜色是紫色，那么这就不能准确地代表这一点。 在这种情况下，我们就必须为紫色设置另一个虚拟变量，或者为他们创建一个缺失值，因为在这种情况下，他们实际上并不是将蓝色作为没有选票的颜色，而是将紫色作为缺失值。 另一种对分类变量，特别是序数变量进行编码的方法是整数编码。 因此，在报告定量、定性值时，你也会经常看到这种情况。 例如，我们这里有一个人获得的最高学位，然后我们给每个值分配一个整数，这仍然代表了这些变量的排序。 因此，举例来说，如果某人拥有硕士学位，那么他的学位值就会高于拥有高中学位的人，在这种情况下，高中学位会被赋予一个字母1或整数1，而硕士学位会被赋予数字3。 非常非常重要。 这些仍然是顺序变量。 它们看起来可能是数字变量，你的模型可能会认为它们是数字变量，然后试图将它们作为数字变量处理。 但这是不可能的，例如，你无法计算一个顺序变量的平均值，因为这没有意义。 原因就在于顺序变量。 我们知道事物的排序。 例如，我们知道硕士学位比高中学位高，但不知道高多少。 例如，5 月份的本科学位并不是高中学位的两倍，尽管 2 是 1 的两倍。 因此，如果你使用顺序变量和整数编码，一定要非常小心，因为模型会认为它们是数字。 它们还没有聪明到不这么认为。 所以要非常小心。 在很多情况下，虚数编码是一个更安全的选择，因为你也可以对失去顺序的序数变量进行虚数编码。 因此，这两者之间总要有所取舍。 现在我们来谈谈异常值。 这又回到了原点。 我们为什么要关心分布？ 极值通常是通过概率分布发现的。 因此，如果我们有过量的数据，如果我们有足够大的数据集，通常我们会使用标准正态分布。 我们假设，如果我们收集了大量数据，也就是该主题的所有数据，那么在大多数情况下，这些数据都大致符合正态分布。 至于在现实生活中是否总是如此，这完全是另外一个问题，但它给了我们一个工作准则。 因此，在这种情况下，我们有了正态分布，这甚至是一个标准的正态分布。 然后，我们可以看看每个观测值离平均值有多远。 在很多情况下，我们会设定一个分界线，比如说我们有一个 "三西格玛规则"，这是一个常用的经验法则，但也可以是任何两个西格玛，也可以是任何一个西格玛。 我们通常使用三个西格玛，并且我们会说，任何偏离平均值超过三个西格玛的观测结果都会被视为离群值。 这又回到了我们刚才谈到的经验法则，我们认为大多数数据实际上都在 99% 的范围内，99% 的数据实际上都在这三个标准差的范围内。 如果我们认为 99% 的数据都在这一范围内，那么我们就可以利用这一依赖于统计分布的经验法则，安全地切断超出这一范围的数值。 因此，这是一种常见的切分方法。 尽管如此，如果你真的要切除异常值，首先要仔细检查。 因为你对数据分布的假设可能有误。 因此，并不是每个数据集都符合这种分布。 是的，对于小数据集。 在很多情况下，我们使用学生 t 分布来代替正态分布。 我们将在下周的讲座中讨论 t 检验时再详细介绍。 届时，我们会更深入地介绍使用分布来检查数据是否符合我们对数据的假设。 但原理是一样的。 例如，我们仍然会使用三西格玛的临界值。 多变量异常值。 我们该如何处理？ 完全一样。 我们只是使用多元高斯。 因此，我们不再使用二维高斯，而是使用三维、四维或五维高斯，但仍会有一个截止值，因此我们仍会以与更小维度相同的方式估计高斯参数。 这和我们的方法是一样的。 所以你仍然可以看到，我们仍然在从每个变量中扣除平均值。 我们仍在求和。 我们还是要除以我们所拥有的数值。 现在唯一不同的是，我们实际上是在研究一个大胆的 X，我们将研究更多的维度，而不仅仅是两个或一个变量，但原理是一样的。 是的，我不指望你去研究这些公式，也不指望你能把它们复制粘贴到试卷上。 除非你觉得这个题目非常有趣，而且你觉得记住这个公式非常直观，在这种情况下，你就去学吧。 我会很高兴的。 但这给了你一个实际的计算结果，即一个弹出样本 X 在该人群中的可能性，即穆和你的分布。 英文怎么说来着？ 英文怎么说来着？ 谢谢 谢谢 所以计算方法是一样的，但计算的是这个样本确实在其中的可能性。 所以你可以看到，我们基本上把这个词放在这里。 我把它放在这里。 你一定要记住吗？ 不需要。 之所以把它放在这里，主要是为了完整起见，但我想让你从中明白的是，足够大的数据集高斯分布，检查它的边际，足够大的多元数据，集高斯分布，检查它的边际。 这就是我想让你记住的基本概念。 还有一种更直观的方法，叫做马哈拉诺比斯距离（Mahalanobis Distance）。 这个想法非常相似，所以你仍然可以看到，我们查看每个观测值或观测值矩阵。 我们查看它们与平均值的偏差，并将其计算为距离。 我个人认为这是一种相当直观的方法，稍后在谈到聚类时我会解释其中的原因，因为聚类的很多内容都与距离和相似性有关。 因此，我认为思考一个观测值与一个向量或观测值集合与一个向量之间的距离有多远，是一个非常直观的概念。 你还可以看到，这个距离遵循正方形分布，离群点会被标记出来，这与 "在这种情况下，它是否与高斯分布有一定距离"、"它是否与我们的正方形分布有一定距离 "的想法非常相似。 因此，概念和背后的想法是相似的。 同样，这也是分布的作用所在，即如果数据遵循某种分布，那么对于不遵循这种分布的数据来说，这意味着什么？ 还有第三个因素，我将简要介绍一下，因为明天的教程中我们会更深入地介绍，因为它基本上是一个非常方便的 scikit learn 实现，我们可以在其中查看局部离群因子。 因此，如果我们有一组 A 点的邻域，我们可以查看这些点中有多少点是在我们对象的邻域中。 因此，这又回到了距离、相似度和邻域的概念上。 有趣的是，我们在讨论 K 最近邻方法时还会进一步讨论这个问题。 但它的有趣之处在于，它适应了这样一种理念，即一个点距离多远才不会成为离群点，这取决于该区域内点的分布情况。 因此，如果我们直观地看一下，就会发现在其中一些区域内，有一些密集的点聚集在一起。 如果你再想一想，如果一个点远离一个如此密集的区域，这意味着什么呢？ 因此，我们的想法是，如果有一些区域的密度非常高，而你会假设所有点的表现都差不多，因为你有这些密度集群，那么如果有人的表现与这些集群不一样，那么他们一定是离群点。 这就是它背后的理念。 它的优点在于适应性很强。 因此，每个数据集的密度都是不同的。 因此，在这种情况下，这里就是一个非常非常密集的点。 因此，我们这边的聚类密度也非常高，这个聚类密度非常高的区域。 但在不同的数据集中，正常密度可能会有不同的表现。 因此，这可以用来思考正态分布的范围。 如果再看一遍高斯分布，这也和这个想法有点类似，因为它又带回了高斯分布能在数据中扩散多远的想法。 如果我们使用标准正态分布，我们就会说，这就是它应该扩散的范围，不能再扩散了，因为标准偏差总是正好为一。 所以它应该是这个样子。 而这种方法允许你有更多的灵活性。 因此，有时可能扩散得更远，有时可能更窄。 这取决于你的数据。 当我们谈到聚类时，我会长篇大论一番，因为它与一种叫做 Dbscan 的聚类方法关系密切，Dbscan 是基于密度的聚类，围绕有噪声的点进行聚类。 差不多就是这个意思。 它有一个非常复杂的名字。 我称之为 Dbscan。 这种方法的原理是，你可以查看相互之间关系密切的邻近点，这些邻近点在距离或相似度上比较接近，而不接近的点则会被标记为异常值。 因此，Dbscan 是为数不多的几种聚类方法之一，基本上可以像这样自动标记异常值。 因此，这两个概念密切相关。 那么我们如何处理这些我们刚刚发现的最常见的离群值呢？ 只需将它们从分析中删除即可。 我们已经反复讨论过对异常值敏感的方法和测量手段。 最简单的方法就是将它们删除。 但是，不要一刀切就了事。 一定要检查到底有多少离群值，因为在很多情况下，可能存在系统性记录问题。 例如，如果您的数据集来自不同的来源，而其中一个来源和另一个来源测量收入的尺度不同，那么这实际上可能会导致您检测到比实际情况更多的异常值。 如何避免这种情况？ 你可以将数据标准化，但也要经常检查是否存在系统性问题。 例如，你移除的所有异常值都来自你合并进来的一个数据集。 因此，在这种情况下，你会想，好吧，数据集有问题，不一定它们都是离群值。 那么，它们真的是异常值吗？ 模型中是否应该考虑它们？ 上周，我们在考虑这类客户及其使用移动数据的行为时，也简要地谈到了移动银行数据。 你可能还记得我们的散点图，我们有两个离群值，他们都是老年人，是真正经常使用手机银行的人，我们说，我们应该移动他们吗？ 我们是否应该把他们作为一个模型？ 这取决于大量的专家知识和客观的数据处理。 所以，不要把它们扔掉。 想想是否有必要纳入它们。 如果有必要，就将其删除。 在我看来，最好的做法是在分析中记录有多少人移动了，可以是百分比，也可以是实际的绝对数字。 这样做的目的是为了提高透明度，而不是直接删除数据。 在很多情况下 你还想报告它们的具体细节。 例如，它们是否都是某个特定维度、某个特定数据集中的异常值？ 它们是否都是朝一个方向的异常值，还是从分布的上端和下端都同样分离出来的异常值？ 这又回到了论证和判断的问题上，对过程中的每一步进行判断。 好的。 数据质量。 如果你回想一下上次讲座中游客的例子。 那么，让我们思考一下，我们收集的游客数据质量究竟会受到哪些因素的影响。 举例来说，我们正在收集爱丁堡当地旅游景点游客的数据。 而且我们也已经有了自己的数据集，这也是我们上次收集的数据。 那么，这两个数据集之间就可能存在不兼容问题。 我早些时候在谈到人口普查数据和引入经济福利数据时就提到过这个问题。 因此，在很多情况下，这些数据集的记录方式是不同的，然后你就必须找出两者之间的区别。 这其中有人为错误、测量不精确、数据源转换不当、语言差异、翻译、技术基础设施等问题。 所有这些在企业中都可能相当棘手。 在很多情况下，这意味着数据是多年前收集的。 但有些人已经不在公司了，无法向你解释这些变量的实际含义。 因此，处理起来会相当棘手。 好了，现在你得多动动脑子，看看这里的两组数据。 假设上面的数据是由我们收集的，其他人给了我们额外的数据，我们可以将这些数据合并到一起，创建一个包含 12 条记录的大型数据集，这是一个巨大而惊人的数据集。 所以，当你开始尝试合并这些不同的数据时，试着找出其中的一些不同点和困难。 我能自发地发现三个。 谁是第一个？ 日期？ 日期？ 在一种情况下，我们的日期完全是日期时间格式。 而这很可能被保存为一个字符串变量。 值得庆幸的是，现在很多软件包都提供了非常简单的日期转换功能，但要仔细检查它们的实际效果。 好的，第二个问题。 国籍。 国籍有什么问题？ 在英国。 在英国，如果你有细分，情况就不一样了。 是的 因此，在一种情况下，我们实际上列出了英国的每个国家，而另一种情况下，我们只列出了英国。 因此，如果尝试将它们合并，就会有很多数据或信息无法获得。 如果尝试合并，实际上有两种方法。 一种是创建一个新的列，这些都是国籍变量的缺失值。 另一种方法是，你说这实际上没有意义。 我会忽略它，把这些数据变成英国数据。 什么是更好的表呢？ 这取决于你是否认为国籍对你的问题来说是一个有意义的重要变量。 因此，如果您认为国籍并不能说明什么问题，那么将其转换为英国国籍也是有效的，因为您不会丢失重要信息。 但这是你必须做出的决定。 好吧，第三个问题。 第三个问题 我们还有其他问题。 可以指出其他数据，可以指出哪些亚群。 Mhm. Mhm. Mhm. - Yeah. Yeah, yeah. Absolutely. 所以，如果你有额外的数据，你可以用这些数据推断出他们可能的国籍。 问题是，我们说的是国籍。 因此，即使我们有邮政编码，我们也不一定知道邮政编码来自哪里。 是家庭地址还是工作地址？ 是国籍吗？ 甚至只写国籍？ 我不问这是否是你的居住地，比如你的身份是什么？ 因此，如果有人住在苏格兰，但他的身份是英国人，他们就会有一个苏格兰邮政编码。 所以要注意这一点。 但是，如果您有其他信息，请继续填写。 谁会遇到更多问题？ 是的，性和性别的记录方式不同。 我们还有缺失值。 如果你收集一些人的人口统计学数据，这很常见，因为这是一种敏感信息，他们不太可能与你分享。 这就是为什么年龄、收入、性别等变量在调查中更难收集，因此更容易出现缺失值。 此外，我们还有编码问题。 因此，我们将性别作为一个数字变量，并用后者代表性别变量。 我们还可以看到，在这种情况下，我们所拥有的类别之间的映射并不是很容易，因为，例如，在性别方面，我们为人们提供了比性别更多的选项，在这种情况下，性别是一个二进制变量。 好吧，还有一个问题。 还有一个问题。 是的，最后一个是否访问过？ 可以。 这就是为什么人们不访问和。 他们把一个现在，但对于第一个。 是的，我做到了。 我会把一个。 好的 这又回到了这两个数据集中的问题和二是否完全相同的问题。 因此，人们在这两个数据集中所假设的定义是否相同，问题的提法是否相同。 那么，我们究竟该如何解释访问变量呢？ 这些人是否都同意 "访问 "的定义？ 还有一个问题。 所以我认为他们只是把年龄提高了。 所以我认为 这里有一个人只有两岁，但却被访问过。 现在的问题是，他们是和父母一起探视的，比如说，他们是和上面这个人的孙子一起探视的，还是说这是一个记录的问题？ 应该是 26 岁之类的吗？ 这些都说不清楚，也很难处理。 可能还有一个问题。 你说了 投标。 投标？ 是的，我们已经说过了。 我们说我们有缺失的价值 你觉得那是什么意思？ 你觉得那是什么意思？ 所以可能只是我们说的第三类。 例如，他们希望增加其他类别。 或者也可能有人想自己指定性别，或者只是另一种形式的编码缺失值。 所以所有这些都是可能的。 都有可能。 9292号 没错 这比其他所有记录的数据都要早得多。 那么，如果我们说，比如说，我们真的想把这些极老的记录纳入我们的数据集，还是说，这会让我们的模型出现偏差？ 会的。 这就是你们在建模过程中可能会遇到的少量数据的很多问题。 我们再来简单谈谈缺失值。 缺失的原因可能是没有输入数值，因此有人拒绝添加。 也可能是数据为空。 因此，人们有时会开始调查，然后中途停止，或拒绝给你答案。 还有一个问题是，缺失值可能意味着非常不同的事情。 例如，如果有人问你上个月的收入是多少？ 有三种情况会导致缺失值。 他们可能完全拒绝回答这个问题。 留白。 什么都不回答。 他们可能会回答 "不知道 "或 "不记得"，这与 "我不想回答这个问题 "并不完全相同。 因此，这三个选项，即拒绝回答、忘记确切数值或什么都不回答，都将在数据集中被编码为缺失值。 因此，举例来说，在这种情况下，有人想申请贷款，并被问到 "您以前是否有贷款记录？ 他们可能会留下空白。 但是，如果你把这个问题改成一个更有限的范围。 例如，如果你不问他们你上一次申请的贷款额度是多少，而是问他们你上一次申请贷款是什么时候？ 或者你干脆问他们，你过去有贷款吗？ 有还是没有？ 他们可能更愿意回答这个问题。 因此，在调查设计中，有一些方法可以最大限度地减少实际收集到的缺失值。 在其他情况下，我们只是不允许在信用评分中记录任何东西。 例如，我们不能将性别作为预测个人违约可能性的变量。 安德烈娃教授是本校的研究人员之一，我非常推荐大家去看看她的论文，我不知道你们是否见过她，但她发表了很多信用评分方面的论文。 她的一些研究非常有趣，因为她研究了在信用评分模型中剔除性别变量的影响。 而她的研究实际上发现，在很多情况下，这对女性是不利的，因为纳入或引入性别变量实际上是为了保护女性。 因此，最初的想法是，不应将个人的性别包括在内，因为当时的想法是，信用评分机构可能会降低女性的评分，例如，因为从历史上看，女性不太可能全职工作，收入也低于男性。 有趣的是，大多数研究表明，女性的信用记录要好于男性。 因此，将性别作为一个变量可能实际上对女性有利。 但在法律上，我们仍然不能将其作为变量，因为这会导致歧视。 这就是其中一个有趣的故事，你也将在第五周的嘉宾讲座中听到更多关于这个故事的内容。 是的，第五周。 第五周 嗯，是的。 缺失值 那我们该如何处理呢？ 一种简单的方法是直接删除观测值。 如果同一观察结果存在多个缺失值，那么删除观察结果通常会很有用，因为在这种情况下，可能会出现一些问题。 例如，受访者不想填写调查问卷，或者存在记录问题或技术问题。 在某些情况下，完全删除是有意义的。 注意不要删除太多，尤其是如果这背后存在系统性问题。 因此，如果你注意到，比如我之前举的例子，如果所有的缺失值都来自一个数据源，那么删除所有的缺失值实际上可能会使你的结果出现偏差，因为这样你就只能在少数几个数据源上训练你的模型，而你却排除了其中一个潜在的重要数据源。 有些模型允许你只保留缺失值。 因此，我们将看到哪些模型能够比其他模型更好地处理缺失值。 例如，神经网络就不喜欢缺失值。 在这方面，决策树要灵活一些。 估算是处理缺失值的另一种方法。 在这种情况下，您只需尝试猜测该人最有可能的值。 有很多不同的方法可以使用平均值或中位数来简单地估算该值，这样做的好处是不会对你的模型造成真正的偏斜，因为你基本上只是增加了一条记录，而这条记录的平均值与其他记录的平均值相同。 例如，神经网络中就经常使用这种方法。 但这只对数值有效。 你不能只对分类变量进行估算。 你可以为它们创建一个新的类别，例如，不可用或不适用。 有些人还会将分类值作为整个数据集中最常见的选项进行推算，我猜这有点像分类变量的平均值。 在其他情况下，人们可能还会选择使用细分过程等方法来模拟最可能的平均值。 例如，你可以查看 20-30 岁女性最可能的分类值。 然后，您只需查看数据的子集，查看该类别或该人群最可能的分类值，然后将该值推算为您的分类值。 最近邻校正又回到了最近邻方法。 这就是为什么我在今后的一次讲座中会广泛介绍这种方法，因为它与我们这里涉及的很多不同的概念都很接近。 这里的想法是，你要计算观测值的近邻。 因此，数据集中哪些记录与它们最相似？ 然后通过这些相邻记录来计算缺失值。 问题是邻域数的设置相当棘手。 我们在讨论 N 时会看到，有几种方法可以克服这个问题。 但归根结底，这是一个主观的选择，而任何主观的选择如果带入到模型中，都有可能使结果出现偏差，因为它会带来主观性。 你也可以建立一个完整的回归模型来预测缺失值。 因此，在这种情况下，例如，你可以使用所有剩余值来预测该因素组合中的一个值。 是的。 我觉得我今天的时机掌握得很好。 最后，我想用几分钟时间简单谈谈评估。 你们已经在 Learn 上收到了课程工作小组的通知。 我希望你们都看了这些小组，因为他们将是你们在本学期剩余时间里的小组伙伴。 我已经将评估简报上传到了unlearn上，供大家查看。 截止日期是 12 月初。 对你们来说，这看起来非常非常遥远，但其实并不遥远。 所以，现在我要你们做的是，现在你们都在同一个房间里。 我现在 把你们分成小组 然后，我希望你们开始思考你们想要遵循的流程结构。 工作流程结构。 你们可以自由决定，并与你们的组员讨论。 你完全可以自由决定。 我认为，项目管理技能是我希望你们从课程中获得的一部分。 所以，如果你觉得在工作流程上很难和其他人一起找到一个结构，那就把它看作是一个学习的机会，因为这确实是一个学习的机会。 话虽如此，我还是建议你们在本周或最迟下周与小组成员一起写下工作流程，因为这将有助于你们在本学期余下的时间里安排工作，而且你们还能相互问责，真正做到写下的内容。 因此，有些小组决定在此基础上写一份适当的合同，这样他们就能写出条件，并在正确的条件下工作。 这就是我们的工作流程。 我们将遵循这个工作流程。 如果我们生病了，我们会提前通过邮件通知大家。 我们会在一周中的这几天相互沟通，然后你们都在这份文件上签字，这样你们就可以相互问责。 压力大的时候，你们就不用争吵了。 在学期中间，我会推荐这个过程。 话虽如此，你们还是可以自己解决的。 你们都长大了。 这就是工作流程文件的样子。 你可以把它做得越漂亮越好，但这是我的建议。 在接下来的几周里，认识一下你们的小组成员吧。 现在是第二周，下周是第三周。 了解你的小组成员。 仔细检查你是否能获取信息，仔细检查一切是否正常。 开始阅读一些关于主题背景的资料，熟悉数据，等等。 因此，这是一项拟议的小组工作，一项民意调查。 天哪，我的英语。 我已经说了一整天了。 我得喝点水。 苹果职位结构 好了 很好 Very good. 让我调出我最信任的文件 So let me pull up my trusty document. 关于一个组 请输入密码 好了 所以，首先。 好的 一共有11个小组 每组4到5人 你们都有手机、笔记本电脑、平板电脑或类似设备。 如果你们还不知道自己的组号，现在就去查。 我的意思是现在就去学习。 打开课程。 打开你的评估文件夹，你会看到一个文档，里面有你的准确编号。 你们都有自己的组号吗？ 都有 都有 好的 听我说 Okay. 听我说 第一组的所有成员请举手 Can all members of Group one raise their hand, please? 环顾四周。 看看对方。 好的 第一组 Okay. 第一组 第二组的所有成员能举手吗？ 好的 第三组 请举手 看着对方 好极了 我看到了快乐的脸庞 很好 第四组 我们能看看第四组吗？ 很好，第五组 认识一下你们的组员 很好 第六组 第六组有谁？ 很好 建立联系 请到第七组 嘿 按八分组 很好 第九组，请。 好的 十号组 Okay. 十号组 谁是十号组？ 很好 最后，但并非最不重要的， 11号组，请。 好的 好的 抱歉 希望你们还记得自己的组员 请大家集合 组织好自己 你们有十分钟 找到你们的组员 在房间里找个角落，坐在一起，开始讨论。 在接下来的几周里，你们想如何安排你们的项目？