**Learning Objectives**

* Develop multiple research questions from your own industry or interest area that can be solved with hypothesis tests, regression, classification, and clustering
* Give examples of clustering problems
* Differentiate between regression and classification problems
* Differentiate between supervised learning and unsupervised learning
* Describe, at a high-level, how machine learning can be used to solve a given scenario
* Differentiate between machine learning and other ways to solve data science questions
* Describe the challenges associated with answering research questions using only hypothesis testing
* Explain why it is important to frame questions in a way that enable them to be answered with data

# **Module and Lesson Intro**

Hi everyone, welcome to another module of Data Science Fundamentals for data analyst. [Hi everyone, welcome to another module of Data Science Fundamentals for data analyst.] We hope that you're enjoying the course so far and we want to take a few moments to review what you did in the previous module. [We hope that you're enjoying the course so far and we want to take a few moments to review what you did in the previous module.] In the previous module, you learned about descriptive and inferential statistics, probability distributions and how probability can be used to design and conduct hypothesis tests to learn from data. [In the previous module, you learned about descriptive and inferential statistics, probability distributions and how probability can be used to design and conduct hypothesis tests to learn from data.] Building these hands-on technical skills has you on the right path and developing your data science skills. [Building these hands-on technical skills has you on the right path and developing your data science skills.] While there more hands on skills to be learned, we need to spend some time considering one of the most important learning objectives in this entire course. [While there more hands on skills to be learned, we need to spend some time considering one of the most important learning objectives in this entire course.] That objective is framing real-world problems for data science and machine learning. [That objective is framing real-world problems for data science and machine learning.] That's what you'll be able to do by the end of this module. [That's what you'll be able to do by the end of this module.] Along the way, we'll describe the importance of developing answerable questions. [Along the way, we'll describe the importance of developing answerable questions.] Describe the challenges associated with hypothesis testing and also introduce some machine learning concepts. [Describe the challenges associated with hypothesis testing and also introduce some machine learning concepts.] This module will be taught in a series of videos, knowledge checks and a peer reviewed activity. [This module will be taught in a series of videos, knowledge checks and a peer reviewed activity.] When you're ready to learn how to frame real-world scenarios as question to begin to data science and machine learning process, click onto the next video. [When you're ready to learn how to frame real-world scenarios as question to begin to data science and machine learning process, click onto the next video.] Good luck. [Good luck.]

# **Why Good Questions Matter**

Hello and welcome back, in this video we're going to discuss why it's important to frame real-world scenarios so that they can be easily answered with data. [Hello and welcome back, in this video we're going to discuss why it's important to frame real-world scenarios so that they can be easily answered with data.] Remember the scientific method that we've been using to design data science processes. [Remember the scientific method that we've been using to design data science processes.] We talked about how the entire data science process starts with a question, and for that reason it's crucial for that question to be good. [We talked about how the entire data science process starts with a question, and for that reason it's crucial for that question to be good.] But what actually makes a question good? [But what actually makes a question good?] In the first module, we said that good questions provide answers that are objectively measurable. [In the first module, we said that good questions provide answers that are objectively measurable.] In data science, we want to apply our own knowledge in the areas of applied statistics, computer science, and within the domain. [In data science, we want to apply our own knowledge in the areas of applied statistics, computer science, and within the domain.] But we want our answers to come from the data we have available. [But we want our answers to come from the data we have available.] This is where tools like hypothesis tests come in handy. [This is where tools like hypothesis tests come in handy.] When we ask questions like, are out of stock items a problem? [When we ask questions like, are out of stock items a problem?] Or does medicine work? [Or does medicine work?] We're asking questions that are really large in scope and they're tough to answer objectively. [We're asking questions that are really large in scope and they're tough to answer objectively.] What is a problem when it comes to out of stock items? [What is a problem when it comes to out of stock items?] Is it a problem if a single item is out of stock? [Is it a problem if a single item is out of stock?] What about if 100 items are out of stock? [What about if 100 items are out of stock?] What about if items are more likely to be out of stock in certain communities, is that a problem? [What about if items are more likely to be out of stock in certain communities, is that a problem?] We can ask similar questions about medicine. [We can ask similar questions about medicine.] What does it mean for medicine to work? [What does it mean for medicine to work?] Is medicine supposed to eliminate all health issues? [Is medicine supposed to eliminate all health issues?] Is it supposed to extend life expectancy? [Is it supposed to extend life expectancy?] And is medicine the entire medical field or is it just something that's administered to people who need treatment for some illness? [And is medicine the entire medical field or is it just something that's administered to people who need treatment for some illness?] The major issue with these questions is their lack of specificity. [The major issue with these questions is their lack of specificity.] When our questions are too broad in scope, it's difficult to arrive at answers that are objective, easily understood and actionable. [When our questions are too broad in scope, it's difficult to arrive at answers that are objective, easily understood and actionable.] Even if we deemed that medicine does not work, what would we actually recommend be done about that? [Even if we deemed that medicine does not work, what would we actually recommend be done about that?] We wouldn't know where to start, and one of the key goals of data science is to further our understanding in application within a specific domain. [We wouldn't know where to start, and one of the key goals of data science is to further our understanding in application within a specific domain.] Remember, that's right there in our data science definition, so what do we do? [Remember, that's right there in our data science definition, so what do we do?] We ask more pointed questions, when we ask questions like do we have more out of stock items in rural communities than in suburban communities? [We ask more pointed questions, when we ask questions like do we have more out of stock items in rural communities than in suburban communities?] Or does this new medical treatment reduced the average recovery time for a patient? [Or does this new medical treatment reduced the average recovery time for a patient?] We're asking questions that can be answered objectively using hypothesis tests. [We're asking questions that can be answered objectively using hypothesis tests.] We can design and carry out experiments, collect data, and perform analysis on that data to objectively answer those questions. [We can design and carry out experiments, collect data, and perform analysis on that data to objectively answer those questions.] This is because they're specific in their descriptions and their potential answers don't allow any room for interpretation. [This is because they're specific in their descriptions and their potential answers don't allow any room for interpretation.] Another key point to make about these questions is that they're actionable. [Another key point to make about these questions is that they're actionable.] If we find that there are more out of stock items in rural communities than suburban communities, we can design more data science projects to determine why that might be. [If we find that there are more out of stock items in rural communities than suburban communities, we can design more data science projects to determine why that might be.] And then, we can take action to make product distribution more equitable. [And then, we can take action to make product distribution more equitable.] This might look like investing more in delivery trucks and delivery drivers in rural or urban communities. [This might look like investing more in delivery trucks and delivery drivers in rural or urban communities.] Or it might even mean investing in a distribution center more central to the affected communities. [Or it might even mean investing in a distribution center more central to the affected communities.] Similarly, if we find that the new medical treatment does reduce the average recovery time for a patient, we could begin educating medical professionals on that treatment and distributing the treatment to medical centers around the world. [Similarly, if we find that the new medical treatment does reduce the average recovery time for a patient, we could begin educating medical professionals on that treatment and distributing the treatment to medical centers around the world.] This is why good questions matter, they enable the use of data, the use of analysis techniques and objective in actionable results. [This is why good questions matter, they enable the use of data, the use of analysis techniques and objective in actionable results.] Many data science projects fail because they don't start with the right question. [Many data science projects fail because they don't start with the right question.] By following these questions guidelines, you can ensure your data science projects don't share that same thing. [By following these questions guidelines, you can ensure your data science projects don't share that same thing.] In the next video, we'll talk about how even the most powerful hypothesis can answer all data science questions. [In the next video, we'll talk about how even the most powerful hypothesis can answer all data science questions.]

# **Challenges with Solving Real-World Problems with Hypothesis Testing**

Hello again, in our last video we discussed why it's important to frame real-world scenarios so that they can be easily answered with data. [Hello again, in our last video we discussed why it's important to frame real-world scenarios so that they can be easily answered with data.] In this video, we're going to talk about some of the challenges associated with answering these types of research questions using hypothesis tests. [In this video, we're going to talk about some of the challenges associated with answering these types of research questions using hypothesis tests.] >> Think back to the questions that we asked in our previous video. [>> Think back to the questions that we asked in our previous video.] Do we have more out of stock items in rural or urban communities than we do in suburban communities? [Do we have more out of stock items in rural or urban communities than we do in suburban communities?] And does this new medical treatment reduce the average recovery time for a patient? [And does this new medical treatment reduce the average recovery time for a patient?] What do these questions have in common? [What do these questions have in common?] Well, we know they're both good because they're focused, objectively answerable, and actionable, but they have a few other similarities too. [Well, we know they're both good because they're focused, objectively answerable, and actionable, but they have a few other similarities too.] Both of these questions are yes or no questions. [Both of these questions are yes or no questions.] The answer to each of these questions is either going to be yes, or it's going to be no. [The answer to each of these questions is either going to be yes, or it's going to be no.] While we can provide various explanations, the answers can vary a whole lot. [While we can provide various explanations, the answers can vary a whole lot.] And this is partially what makes them objectively answerable using data and tools like hypothesis tests. [And this is partially what makes them objectively answerable using data and tools like hypothesis tests.] These questions are also focused on understanding simple truths about observed data. [These questions are also focused on understanding simple truths about observed data.] They are asking about simple aggregations like counts and averages for various samples of data. [They are asking about simple aggregations like counts and averages for various samples of data.] And in the case of hypothesis testing, they're using those aggregations and statistical inference to make claims about a population. [And in the case of hypothesis testing, they're using those aggregations and statistical inference to make claims about a population.] These similarities are what make these questions suitable for hypothesis testing. [These similarities are what make these questions suitable for hypothesis testing.] But this is also pretty limiting. [But this is also pretty limiting.] What if we have a different data science question? [What if we have a different data science question?] What if we want to know something more, or something else? [What if we want to know something more, or something else?] For example, what if we want to estimate when a specific item is going to be out of stock in a specific store? [For example, what if we want to estimate when a specific item is going to be out of stock in a specific store?] Or what if we want to understand what causes items to be out of stock in the first place? [Or what if we want to understand what causes items to be out of stock in the first place?] In the medical example, we might want to estimate the exact probability that a new medical treatment will affect the recovery time for a specific patient. [In the medical example, we might want to estimate the exact probability that a new medical treatment will affect the recovery time for a specific patient.] These are complex questions, but they're still good questions. [These are complex questions, but they're still good questions.] They're still specific, they're still objectively answerable, and they're still actionable. [They're still specific, they're still objectively answerable, and they're still actionable.] They're also incredibly similar to countless data science projects that you'll come across in the real world. [They're also incredibly similar to countless data science projects that you'll come across in the real world.] These questions are focused less on the what of the data, and more of the when, the how, how much, and why. [These questions are focused less on the what of the data, and more of the when, the how, how much, and why.] Questions that data scientists and key stakeholders ask all the time to learn more about their specific domains and industries. [Questions that data scientists and key stakeholders ask all the time to learn more about their specific domains and industries.] We can and should set hypotheses for these types of projects. [We can and should set hypotheses for these types of projects.] We can hypothesize that items go out of stock at regular intervals. [We can hypothesize that items go out of stock at regular intervals.] We can hypothesize that items go out of stock on specific days of the week or months of the year. [We can hypothesize that items go out of stock on specific days of the week or months of the year.] But hypotheses can't help us with these more complex data science projects. [But hypotheses can't help us with these more complex data science projects.] How would we use a hypothesis test to estimate when a certain item

is going to be out of stock? [How would we use a hypothesis test to estimate when a certain item is going to be out of stock?] There's no realistic mechanism for that, especially when the scenario doesn't appropriately match with one of the probability distributions we learned about. [There's no realistic mechanism for that, especially when the scenario doesn't appropriately match with one of the probability distributions we learned about.] The same can be said for understanding why items go out of stock, or estimating the exact probability that a new medical treatment will affect the recovery time of a specific patient. [The same can be said for understanding why items go out of stock, or estimating the exact probability that a new medical treatment will affect the recovery time of a specific patient.] For these questions, hypothesis tests just aren't enough. [For these questions, hypothesis tests just aren't enough.] We need more technical tools in our data science tool belt. [We need more technical tools in our data science tool belt.] Throughout the rest of this lesson, we'll start to look into the basic concepts of machine learning. [Throughout the rest of this lesson, we'll start to look into the basic concepts of machine learning.] And learn how to frame these more complex data science questions. [And learn how to frame these more complex data science questions.]