

## Module 2: Survey Analysis

### Table of Contents

<b>Lesson 2-1: Survey Analysis .....</b>	<b>2</b>
Introduction to Customer Satisfaction .....	2
Psychological Constructs .....	13
Customer Satisfaction Analysis .....	19
Statistical Analysis Background Part 1 .....	27
Statistical Analysis Background Part 2 .....	33
Evaluating Airline Survey Data .....	42

## Lesson 2-1: Survey Analysis

### Introduction to Customer Satisfaction

# Definition #1

Customer satisfaction is defined as a **measurement** that determines **how happy** customers are with a **company's products, services, and capabilities.**

American Society for Quality - <https://asq.org/quality-resources/customer-satisfaction>.

A portrait of Professor Unnati Narang, a man with short dark hair, wearing a dark blue blazer over a light blue button-down shirt. He is standing with his arms crossed against a plain light gray background. A small red 'I' logo is visible in the top right corner of the slide.

Hello everyone. As we start this journey in looking into customer satisfaction, we should start with an introduction. For an introduction, we shall look at some definitions. Here's our first definition. Customer satisfaction is defined as a measurement that determines how happy customers are with a company's products, services, and capabilities. Let's stop right now. Let's look at three things here. First, customer satisfaction is a measurement. You have to figure out how to measure this. What are we trying to measure? Well, at least according to this definition, we're trying to measure how happy customers are. With regards to what? With regards to a company; their products, services, and capabilities. Let's continue on.

## Definition #1

Customer satisfaction information, including surveys and ratings, **can help a company** determine how to best improve or change its products and services.

American Society for Quality - <https://asq.org/quality-resources/customer-satisfaction>.



Customer satisfaction information including surveys and ratings can help a company determine how to best improve or change its products and services. If we are not attempting to help our company with regards to results of customer satisfaction measurement, then what are we doing? That everything we do with regards to customer satisfaction measurement needs to be in line with helping a company. Now, let's look at another definition that is a great definition.

## Definition #2

Customer satisfaction is defined as an **overall evaluation** based on the total purchase and **consumption experience** with the good or service **over time**.

Fornell, C. et al. (1996), Johnson, D.M., Anderson, W. E., Cha, J. & Bryant, E B. The American Customer Satisfaction Index: Nature, purpose, and findings, *Journal of Marketing*, 60(4), 7-18



Fornell et al. '96, they suggested, "Customer satisfaction is defined as an overall evaluation." Maybe it's not just how happy someone is, but it includes other potential psychological constructs, overall evaluation. Based on what? Based on the total purchase and consumption experience. Let's stop there too. It's an experience. It's not just a one-time thing that a consumer consumes over time and that's why this definition also says with the good or service over time. That customer satisfaction can change and it's an experience that we have to consider.

## Some Major Issues to Consider

Psychological constructs of satisfaction

Proper measurement of these constructs

Varying targets of satisfaction



With regards to looking at customer satisfaction as a whole, as we go throughout this lesson and this course, we will look at some major issues. These are major issues that we really have to consider. The psychological constructs of customer satisfaction, proper measurement of these constructs, varying targets of satisfaction,

## Some Major Issues to Consider

Differences in the impact of individuals' expressions of satisfaction

Satisfaction can change over time





differences in the impact of individuals' expressions of satisfaction as well as satisfaction changing over time. Now, let's look briefly at each one of these but we will cover these in more detail in future lessons.

## Some Major Issues to Consider

### Psychological constructs of satisfaction

Which psychological constructs signify satisfaction?

- Attitude (liking or disliking)
- Emotions (e.g., joy, anger, happiness)
- Attitude Strength (resistance to change of attitudes)
- ???

First, let's look at the psychological construct nature of customer satisfaction. What psychological constructs signify satisfaction? Now, you might think to yourself, "That's simple. It's just satisfaction." Well, it's not simple because in psych research, we know that there are at least three things that are separate variables or separate constructs when it comes to customer satisfaction. The first is what is called attitude or what we know it as liking or disliking, or some of us may know it as sentiment in computer science. It's this aspect of how much do I like something, or how much do I not like something? That's different than emotions as our second construct. There's joy, there's anger, there's happiness, so on and so forth. That's also different from our third construct called attitude strength. Now, we might not be as familiar to this construct, but it's not just how much do I like something or dislike something, how resistant am I to changing my attitude with regards to that liking or disliking of something? Then of course, there's many more constructs that we will not necessarily consider in this class, but I just want to identify that it's not that simple when it comes to what is customer satisfaction psychologically.

## Some Major Issues to Consider

Proper measurement of these constructs

Measuring psychological states of being are difficult because they are intangible



Another major issue to consider is how do we properly measure some of these psychological constructs? Because measuring psychological states of being are difficult because they are intangible. We can't just reach into someone's brain and pull out how happy they are.

## Some Major Issues to Consider

Proper measurement of these constructs

Measurement methodologies all have their strengths and weaknesses

- Surveys
- Text/Image Analysis
- Network Analysis



There's various measurement methodologies. All have their strengths and their weaknesses. We'll talk about in this class surveys. We'll talk about texts analysis, not

necessarily image analysis, but I wanted to list it here because it's definitely used to assess customer satisfaction. We'll also talk about network analysis. There's many other methods, but at least for the purposes of this class, surveys, texts analysis, network analysis, we'll go over them and they all have their strengths and their weaknesses. Another major issue to consider is varying targets of satisfaction. Customer satisfaction could be targeted towards not just a company, but it could be targeted towards a product.

## Some Major Issues to Consider

### Varying targets of satisfaction

Customer satisfaction could be targeted towards:

- Products (e.g., iPhone v. Apple Watch)
- Services (e.g., AT&T cell service v. home internet)
- The Company/Brand (e.g., revenue v. social responsibility)
- ???

For example, I like iPhones, but I had an Apple watch, I didn't find much use to it, and so my customer satisfaction towards Apple watch is definitely different than my satisfaction towards an iPhone. Customer satisfaction can be targeted towards products in different ways. What about services? I have AT&T cell service, I don't necessarily have their home Internet but I've tried it, and I can say I have a much better experience with cell service and home Internet. Customer satisfaction is different depending on the service that a consumer is using. Then of course, going back to a company and the brand, you might think to yourself, "I really like brand X because their revenue is Stellar. I invest in their stock maybe, and their financials are great." But then when I look at their social responsibility, I say to myself, "Well, I don't really like that aspect of that brand." There's varying targets of satisfaction even to a company with regards to dimensions of that company. Then of course, there's question marks because it could be targeted towards other things as well. This is another major issue to consider when thinking about customer satisfaction.



## Some Major Issues to Consider

Differences in the impact of individuals' expressions of satisfaction

When people express or record their satisfaction or dissatisfaction, they do not all have equal impact on others (e.g., me v. Kim Kardashian)



Now, how about differences in the impact of individuals' expression of satisfaction? What do I mean by that? When people express or record their satisfaction or dissatisfaction towards something, we don't all have equal impact on others. The example I'll give is the difference between Kim Kardashian and myself. Now, the differences are abounding. But with regards to customer satisfaction, if Kim Kardashian really likes a brand and promotes it, that's going to have a much bigger impact on other people's view and customer satisfaction towards that brand than if I went on social media and said I really love a brand. There are differences in the impact of individuals' expressions of satisfaction.

## Some Major Issues to Consider

Satisfaction can change over time

Time = Change (e.g.,  
progression of Apple Arcade)



Another major issue is that satisfaction can change over time. I'd like to give this example of Apple Arcade. You can tell I'm an Apple fan. But Apple Arcade, when it first came out, I really didn't like it. I was like, "There's not enough games and the games are not that fun." But over time, they started pumping out more games and I said, "Wow, I really love my Apple Arcade." Customer satisfaction over time can change. That's another major issue to consider.

## Some Major Issues to Consider

“As the cornerstone of the marketing concept, customer satisfaction has been embraced by practitioners and academics alike as the highest-order goal of a company” (Peterson & Wilson, 1992).

Peterson, R. A., & Wilson, W. R. (1992). Measuring customer satisfaction: Fact and artifact. *Journal of the Academy of Marketing Science*, 20(1), 61-71.  
<https://doi.org/10.1007/bf02723476>

Now, with all of these issues, we ask ourselves, why are we even trying to measure customer satisfaction? Peterson and Wilson in '92 put it this way, "As the cornerstone of the marketing concept, customer satisfaction has been embraced by practitioners and academics alike as the highest order goal of a company."



If our company is not trying to figure out how satisfied a customer is with our company, our brand, our products, our services, then really what are we doing as a business?

That it is about the consumers, it is about the customers. Even though there's so many issues with regards to measuring customer satisfaction, we really need to focus on how can we do it and get better at doing it?

Psychological Constructs

## Psychological Question

What is customer satisfaction psychologically?  
Is it an attitude?  
Is it emotions?  
Is it strength of opinion?  
Is it something else or a combination of many constructs



What is customer satisfaction psychologically? We talked about it in a previous lesson, but is it an attitude? Is it emotions? Is it a strength of opinion? Or is it something else, or a combination of many constructs? Well, a really good definition that I found is a definition from Eagly and Chaiken in 93'. It's simple but comprehensive. They suggest that an attitude is a psychological tendency towards favor or disfavor.





**Attitude**

Psychological tendency  
towards **favor** or **disfavor**  
(Eagly & Chaiken, 1993).

How satisfied are you with your iPhone?  
(-5 strongly dissatisfied, 0 neither dissatisfied  
or satisfied, +5 strongly satisfied)

Eagly, A. H., & Chaiken, S. (1993). *The Psychology of Attitudes*. Orlando: Harcourt Brace Jovanovich Co.

Now, this concept of favor or disfavor, you could also say liking or disliking, we've probably seen this in various surveys in the past. For example, with regards to an iPhone, you might have seen a survey question if you have an iPhone where it says, how satisfied are you with your iPhone? The range may be negative five, strongly dissatisfied to zero, neither too satisfied or satisfied, or positive five, strongly satisfied, and you may be somewhere in-between there negative three, positive two, so on and so forth. But what we're trying to measure here is, what is our attitude towards an iPhone, our level of favor or disfavor, liking or disliking.

## Attitude & Emotions

An attitude is "a relatively **enduring** organization of beliefs, **feelings**, and behavioral tendencies towards socially significant objects, groups, events or symbols" (Hogg & Vaughan, 2005, p. 150).

Hogg, M., & Vaughan, G. (2005). *Social psychology* (4th ed.). London: Prentice-Hall.



Now what about emotions? Well, in this definition by Hogg and Vaughan in 2005, it suggests that an attitude is a relatively enduring. Now let's stop right there. An attitude is something that has staying power, and so it's not necessarily just momentary. Emotions are momentary. In fact, if an emotion is longer than just a moment, it's usually called a mood. This concept of endurance is important with regards to an attitude. An attitude is a relatively enduring organization of beliefs, feelings, and behavioral tendencies towards socially significant objects, groups, events, or symbols. What I would like to suggest is that emotions are somewhat of a subset, at least for our dealing with customer satisfaction. It's included within this definition of attitude.

## Attitude Strength

“Attitude persistence, resistance, impact on information processing and judgements, and **guiding behavior**” (Petty & Krosnick, 1995).

How likely are you to switch from an iPhone to a Samsung Galaxy?  
(-5 very unlikely, 0 neither unlikely or likely, and +5 very likely)

Petty, R. E., & Krosnick, J. A. (1995). *Attitude strength : antecedents and consequences*. Mahwah, NJ: Erlbaum.



Then the one thing that may be I believe we are missing is attitude strength. This definition by Petty and Krosnick in 95' suggests that attitude strength is attitude persistence, resistance, impact on information processing and judgments, and guiding behavior. This is a key phrase, "Guiding behavior." Here's an example with regards to our iPhone. Let's say that we rated an iPhone positive five, but then this next question gets at the attitude strength, not just the attitude, because this question says, how likely are you to switch from an iPhone to a Samsung Galaxy? We have negative five very unlikely, zero, neither unlikely or likely, and positive five very likely. Imagine if we had a positive five liking or attitude towards our iPhone, but then with regard to this question, we say maybe plus three with regards to how likely we are to switch from an iPhone to a Samsung Galaxy. You can imagine that putting in that attitude strength measure alongside attitude is so important, because what scholars have found is that when you add the measure of attitude strength with attitude, you can help predict behavior.

## What Should We Measure?

Customer satisfaction as a psychological construct is flexible/vague.



What should we measure when it comes to customer satisfaction? Remember customer satisfaction is a psychological construct, it's flexible. It's vague. It's intangible. We should consider what our goals are.

## What Should We Measure?

Attitude strength is likely the ultimate goal.



Is our goals to measure behavior? Is our goals to get an emotion for whatever reason? So on and so forth, and that will affect what we try to measure. But I would like to suggest that attitude strength is likely the ultimate goal. Why? Because with regards to



measuring customer satisfaction for our various organizations, we're likely looking at behavior change, because it's not our goal to produce behavior change.

## **Some Major Issues to Consider**

“As the cornerstone of the marketing concept, customer satisfaction has been embraced by practitioners and academics alike as the highest-order goal of a company” (Peterson & Wilson, 1992).

**Is not our goal to produce behavior change?**

Peterson, R. A., & Wilson, W. R. (1992). Measuring customer satisfaction: Fact and artifact. *Journal of the Academy of Marketing Science*, 20(1), 61-71.  
<https://doi.org/10.1007/BF02723476>

Going back to this definition by Peterson and Wilson in 92', as a cornerstone of the marketing concept, customer satisfaction has been embraced by practitioners and academics alike as a highest order goal of a company, because a company is trying to produce behavior change towards their products or services, or even their brand.



Customer Satisfaction Analysis

## Types of Data to Analyze

Survey data

Social media data

Online behavioral data

More...



Okay everyone, we are now going to look at customer satisfaction analysis, and of course start with introduction. Let's first think about, what are the various types of data that we can analyze when we think about customer satisfaction? Now, we've already talked about survey data, but that's a major way that companies already look at data to analyze customer satisfaction, so we will definitely look at survey data as a source of measurement of customer satisfaction. But here's another dataset that's been increasingly popular, which is social media data, that consumers are going every day to social media platforms in announcing their liking or disliking of various brands, products and services, and so that's another dataset that we can look at to measure customer satisfaction. Now, if your company also has online behavioral data, so for example, how people are interacting with your website or how people are interacting with your marketplace, how people are commenting on things that you might be selling online, this is all online behavioral data, and that's also another dataset that you can analyze to assess customer satisfaction. Now, this third point, online behavioral data, we won't talk too much in this class because that's probably proprietary data to your company, and so we won't have public datasets to be looking at. But I just want to identify that that is a major way to assess customer satisfaction with regards to datasets that you own. Then of course, both now and into the future, there will be many more different types of data to analyze with regards to customer satisfaction, but we'll focus on survey and social media data. Now, let's talk about survey data.

## Survey Data

Collect via physical or online means.

Build survey instruments with psychological and analysis concerns in mind.

Analyze using statistical or machine learned methods (for numerical data).



Survey data, you can collect either via physical or online means. Back in the day it was mostly physical, but of course now it's mostly online. Now, we need to build survey instruments with psychological and analysis concerns in mind. Remember from the first module, there is psychological construct considerations as well as measurement issue considerations with regards to constructs, and we need to keep all that in mind. But this class is not a class in how to build the survey because that could be its very much own class. In fact, many people's whole careers are based on trying to figure out how to create a proper survey. I'll just say that survey instruments need to be really carefully created with psychological and analysis concerns in mind. But then after you have created your survey and you've collected your data, usually, you analyze surveys with statistical or machine learning methods, especially for numerical data. We'll talk a little bit about statistical methods in this class.

## Survey Data

Typically measured by how satisfied someone is with a brand, product, or service.

Rating specific components of your brand, product, or service can help to identify what aspects drive customer satisfaction more heavily than others (e.g., airline dataset).



Now, with regards to customer satisfaction surveys, it's typically measured by how satisfied someone is with regards to a brand, a product, or a service. You may even rate specific components of your brand, product, or service, to identify which aspects drive customer satisfaction more heavily than others. We will look at an airline dataset in this course, and you might rate how comfortable the seat is with regards to your airline experience. Or you might ask for how people rate how good the food was. There's various components with regards to your product or your service, and these can all be measured with regards to customer satisfaction surveys.

## Survey Data

Net promotor score (NPS) is attempting to get closer to attitude strength because it deals with behavior



Now, a really popular survey, and I wanted to identify this one, because it's really powerful, a type of survey called a Net Promoter Score, or you might have heard of it as an NPS survey. An NPS survey is attempting to get closer to attitude strength because it deals with behavior. Now, we've probably seen this everywhere, we just didn't know it was called a Net Promoter Score Survey, where you're just asked one question, how likely are you to recommend this service or this product or this company to a friend? It's probably something from 0-10 scale, whereas that positive 10 is very likely to recommend, and that zero is very unlikely to recommend. The numbers may be different, but you get the gist. It's this range from bad to good, very unlikely to very likely. We've seen this before. Now, why is this a very powerful survey, this NPS survey?



**Survey Data**

**Net promotor score (NPS)**

**Pros**

- Simple to set up and simple to use
- Some evidence that it is a good indicator of growth
- Adaptable to different aspects of your company (e.g., brand, product, service)

Reference: <https://www.slideshare.net/aramshaw/net-promoter-score-a-10-slide-introduction>

Well, because some of the pros is that it's very simple to set up and very simple to use. Some places I've even seen where they have these as physical setups where you're leaving a bathroom in an airport, and in that experience, there's some mechanism that you push where you either say positive or negative and you could even give potential variation of how positive and how negative, and it's some form of this Net Promoter Score. There's some evidence that NPS surveys are really a good indicator of company growth and company satisfaction. It's also adaptable to different aspects of your company. You can do a Net Promoter Score for your brand, for your product, for your service, and so it's really simple, it's really nice.



# Survey Data

## Net promotor score (NPS)

### Cons

- Lacks context behind the scores
- Usually takes quite some time to get full results
- Little ability to apply statistical checks



But some of the negatives is that it lacks contexts behind the scores. Because you're only collecting one score, you have no idea if you're catching the same thing for each person. Let's say you say to somebody how likely they're to recommend Apple Music to a friend. You don't know if they're recommending it based on the catalog of songs or how quality the sound quality is with regards to each one of the songs, it just lacks context. Also, usually, Net Promoter Scores are collected over time, you see these emails that are sent to you, and so usually it takes quite some time to get full results, as well as there's little ability to apply statistical checks. You don't know if over time people are using that instrument in different ways. Remember this issue of reliability. With other types of surveys where you're asking a ton of questions, you can statistically check question to question to see, are they gaming the survey? Are they taking the survey seriously? Whereas regards to NPS, you have no idea because it's just one score. But what's really nice about the Net Promoter Score is it's getting closer to this concept of attitudes strength. Now, we also have this dataset that is social media.

## Social Media Data

Collect via social media platform APIs (Application Programming Interface)

For satisfaction analysis, analyze using psychological text models



I mentioned before that people are more and more going on to social media platforms and talking about companies, brands, products, services. We can use this data to measure customer satisfaction. Now, usually you can click social media data via social media platform APIs, application programming interfaces, and we won't talk really at all about APIs in this class because that's not the focus of this class. But if you wanted to go to a social media platform and get data, you're likely going to have to interact with what is called an API. For satisfaction analysis, once you have data, which we'll give you some data in this class and tools to access that data without APIs, you can analyze that data using psychological texts models, and we'll talk about that in this class.

## Social Media Data

For influence analysis, analyze using network analysis techniques

For content analysis, analyze using text mining, image processing, and/or video processing techniques



For influence analysis, you can analyze using network analysis techniques, which we will also talk about in this class. For content analysis, you can analyze using text-mining, image processing, and or video processing techniques. Now, we won't talk about image processing or video processing techniques in this class, but we will definitely talk about text mining techniques to get at customer satisfaction.

Statistical Analysis Background Part 1

## Various Ways to Analyze Survey Data

T-tests (comparing two averages)

Multiple linear/logistic regression (comparing how multiple predictor variables impact target variable)



Okay, everyone. Now that we've had the intro to customer satisfaction analysis, we will go into stats. This is part 1 of two parts, to give you a background on statistical analysis. There's various ways to analyze survey data. This is not a comprehensive list. These are just three different ways. But these are very common ways to analyze survey data. We'll talk about them. First, let's think about a t-test. What is a t-test? It's a statistical test comparing two averages. Let's say in the airline example you have seat comfort and customer satisfaction towards seat comfort, and you have customer satisfaction towards food quality. Let's say that you take those two average scores and you want to see if their customer satisfaction is similar to both or different. To accomplish this you would use, the statistical test called a t-test. The t-test will tell you whether they are same or different. Now another way to analyze survey data is what is called a multiple linear or logistic regression. Now that sounds really fancy. But really what we're talking about here is comparing how multiple predictor variables impact a target variable. We're trying to figure out, for example, in this same example, how seat comfort plus food quality may affect overall customer satisfaction. That they might have individual ratings of customer satisfaction towards see comfort, and towards food quality. But then we ask them another question that is just, how in aggregate is your customer satisfaction towards this airline, or this airplane. Seat comfort and food quality are predictor variables, and they impact this target variable of overall customer satisfaction. You use a statistical technique called a multiple linear or logistic regression for that.

## Various Ways to Analyze Survey Data

Machine learning (variety of linear and non-linear prediction/classification techniques)



Then third, there is a body of tools or algorithms called machine learning, which is a variety of linear and non-linear prediction and classification techniques. We won't go over that in this class, but that's probably one of the major ways to analyze survey data. I wanted to make sure that I identified that with regards to this introduction. There's various ways to analyze survey data. Now let's jump into a multiple linear and logistic regression. As a simple primer to linear regression, think back to when we were in grade school, maybe junior high, and we were looking at the equation of a line. Now if you remember the equation of a line,  $y$  equals  $m \cdot x$  plus  $b$ . Remember  $y$  is the target variable,  $m$  is the slope,  $x$  is again your  $x$ -coordinate, and then  $b$  is your  $y$ -intercept. Let's say in this example of customer satisfaction with regards to seat comfort in an airplane,  $x$  is the predictor variable. It's seat comfort.



## Linear Regression – Simple Primer

Based on the equation of a line:

$$y = mx + b$$

**x** – the predictor variable (e.g.,  
seat comfort in airplane)

**y** – the target variable (customer  
satisfaction)



We're trying to predict what? We're trying to predict the target variable, customer satisfaction.

## Linear Regression – Simple Primer

Based on the equation of a line:

$$y = mx + b$$

**b** – the y-intercept

**m** – the slope (how much **x**  
affects **y**)



B is the y-intercept still, and m is the slope. But what does that slope represent? The slope represents how much seat comfort, **x**, affects customer satisfaction, **y**.

## Linear Regression – Simple Primer

There can be multiple x's, so

**x1** = seat comfort,

**x2** = snack variety,

**x3** = flight length, etc.

So, your equation could be

$$y = m_1x_1 + m_2x_2 + m_3x_3 + b$$



Now there could be multiple xs, remember.  $x_1$  could be seat comfort,  $x_2$  could be snack variety,  $x_3$  could be flight length, etc.. Your equation could now be rewritten to  $y$  equals  $m_1x_1$ , so it's the slope with regards to how much seat comfort  $x_1$  affects  $y$ , plus  $m_2x_2$  plus  $m_3x_3$  plus  $b$ . All we're doing is we're taking that equation of a line  $y$  equals  $m_x$  plus  $b$ , and we're adding more xs, and therefore we need more ms. Now normally in regression,  $m$ , it's called the Beta coefficient.

## Linear Regression – Simple Primer

Normally in regression,  $m$  is called  $\beta$ , so each  $\beta$  (e.g.,  $\beta_1$ ,  $\beta_2$ ,  $\beta_3$ ) is called a  $\beta$  coefficient



Each Beta is called Beta coefficient 1, Beta coefficient 2, Beta coefficient 3, so on and so forth.

## Linear Regression – Simple Primer I

$$\text{Customer satisfaction} = \beta_1(\text{seat comfort}) + \beta_2(\text{snack variety}) + \beta_3(\text{flight length}) + \beta_0$$

With the rights of linear regression, we could take that equation of a line, and convert it into this with regards to an equation for customer satisfaction. This is a linear regression formula, in our case, for customer satisfaction. Where customer satisfaction, which is  $y$ , equals  $b_1$ , which is slope 1, times seat comfort plus  $b_2$  times snack variety plus  $b_3$  times flight length plus  $b_0$ , which is actually just a replacement for our  $y$ -intercept. Now if you just take this equation at face value, you might say, this looks complicated. But if you think about it back towards the traditional equation of a line,  $y$  equals  $m_x$  plus  $b$ , you can say, this makes sense. This is simply just changing or adjusting the equation of a line to be the equation for customer satisfaction. This is called a linear regression, in regards to customer satisfaction.

Statistical Analysis Background Part 2



Hello everyone, this is now part two of a stats background for customer satisfaction analysis. In our previous lesson, we broke down the equation for linear regression with regards to customer satisfaction.

## Linear v. Logistic Regression

The difference between linear and logistic regression is what type of variable is your target variable (in our case customer satisfaction) — and it is a different formula based on probabilities.

$$\text{Log}[\text{Cust sat} / (1 - \text{Cust sat})] = \beta_1(\text{seat comfort}) + \beta_2(\text{snack variety}) + \beta_3(\text{flight length}) + \beta_0$$



So, now we need to talk about logistic regression, now the difference between linear and logistic regression is. What type of variable is your target variable, in our case customer satisfaction. And so, logistic regression is a different formula based on probabilities.

## Linear v. Logistic Regression

If customer satisfaction is continuous (-5, -4, ..., 4, 5), then you should use linear regression.

If customer satisfaction is ordinal (not satisfied = 0, satisfied = 1), then you should use logistic regression.



What do I mean by this, what I mean is that if your customer satisfaction variable is continuous. So negative five, negative four, positive four, or positive five positive two, etcetera, then you should use linear regression.

## Evaluating Linear Regression

First and foremost, it is all about how well the line (model) predicts the data



$R^2$  - provides a measure of how well the model is fitting the actual data. It takes the form of a proportion of variance. Values range from 0 to 1.

But if your target variable customer satisfaction is ordinal meaning that it's just zero or one not satisfied or satisfied then you should use logistic regression. If you're evaluating linear regression, you remember trying to create this equation for a line. And so, first and foremost it's all about how well that line or that model predicts the data. Now this is a two dimensional graph and so this is not totally accurate with regard to the equation that we had in our previous lesson. But just follow along here with regards to just generally how you assess or evaluate linear regression. So we have two charts here, we have one plot on the left in which this line. Does a pretty good job of representing the position of all the dots on that graph on the right side. We have a bunch of dots and a line that approximates the average of these dots but it doesn't do as well of a job as the graph on the left. And so, we typically use a measure called R squared to understand how well a line fits this model, how well this model fits the actual data. And so, it takes the form of a proportion of variants, so values range from zero to one and so on. The left hand side, R squared might be theoretically point eight five and to remember the ranges from zero to one. So, point eight one five is pretty good and we can even see it in the graph that that line does a pretty good job of approximating almost all of those points. But on the right hand side our R squared is let's say a point three two. So, it's doing a much worse job with regards to the left graph at approximating the dots because why the docks are everywhere. And so it's going to be hard for a line to approximate or create a model that represents all of those dots. And so R squared is one way to evaluate linear regression, it's one of the major ways. Secondly, you can identify which variables contribute more to customer satisfaction than others.

## Evaluating Linear Regression

Customer satisfaction =  $\beta_1(\text{seat comfort}) + \beta_2(\text{snack variety}) + \beta_3(\text{flight length}) + \beta_0$

$\beta_1 = 40; p = .02$

$\beta_2 = 1; p = .24$

$\beta_3 = 1; p = .04$

Snack variety has no significant effect on customer satisfaction ( $p > .05$ )

Seat comfort has 40X as much impact on customer satisfaction versus flight length (40 vs. 1)

When you're evaluating linear regression, let's go back to this equation that we had from lesson one. Customer satisfaction equals beta one time seat comfort plus beta two times snack variety plus beta three times flight length plus beta zero. Now I've just theoretically put some beta values here under as well as some P values. So you see beta one equals 40, P equals point zero two, so on and so forth. And so, what we can see from these beta values and these P values first is that snack variety has no significant effect on customer satisfaction why? Because the P value is greater than point zero five, in fact the P value equals point two four.



Now just as a general note and we won't have time to go too much into this. But in statistical analysis, when something has a greater than .05p value, it means that there's likely a high chance. That any results coming from the values associated with that P value are due to chance and so they're usually thrown out.

## Evaluating Linear Regression

Customer satisfaction =  $\beta_1(\text{seat comfort}) + \beta_2(\text{snack variety}) + \beta_3(\text{flight length}) + \beta_0$

$\beta_1 = 40; p = .02$   
 $\beta_2 = 1; p = .24$   
 $\beta_3 = 1; p = .04$

Snack variety has no significant effect on customer satisfaction ( $p > .05$ )

Seat comfort has 40X as much impact on customer satisfaction versus flight length (40 vs. 1)

And so, beta two equals one but the p value equals point two four, we just consider that as not significant. So, the only players in the game right now are beta one and beta



three which are seat comfort and flight length. So now, beta one and beta three they have  $P$  equals point zero two and  $P$  equals point zero four, so we know they're significant. So now we can compare the beta coefficients, beta one is 40 and beta three is one. So, what we can say is seat comfort which is beta one has 40 times as much impact on customer satisfaction versus flight length. And so this is another way to evaluate the results of a linear regression.

## Evaluating Logistic Regression

$P$ -values still hold as important per  $\beta$  coefficient, but the coefficient signifies an odds ratio.



Now, what about evaluating a logistic regression,  $P$  values still hold is important for beta coefficient. But the coefficient signifies something else in a logistic regression, it's what's called an odds ratio. Remember with regard to logistic regression, you only have two states for the customer satisfaction. You have zero or one not satisfied or satisfied, so what you're really trying to do with the logistic regression is. Predict how likely a person is to flip from not satisfied to satisfied or flip from satisfied to not satisfied.



## Evaluating Logistic Regression

$$\text{Log}[\text{Cust sat} / (1 - \text{Cust sat})] = \beta_1(\text{seat comfort}) + \beta_2(\text{snack variety}) + \beta_3(\text{flight length}) + \beta_0$$

$$\beta_1 = 4.0; p = .02$$

$$\beta_2 = 0.1; p = .24$$

$$\beta_3 = 0.1; p = .04$$

Snack variety has no significant effect on customer satisfaction ( $p > .05$ )

Each rating step (e.g., 1 to 2) of seat comfort increases the log odds of a customer being satisfied by 4.0

Each rating step (e.g., 100 to 101) of flight length increases the log odds of a customer being satisfied by 0.1

And I did talk about this in the second slide but this is an actual modified version of lesson once customer satisfaction equation. That is modified for legit logistic regression and you see that here you see the log times customer satisfaction over etcetera etcetera. That's a bit modified of an equation because this is a logistic regression equation. But we see similar beta one, beta two, beta three P values for each one of those and so, let's look at those as well. We also see in this case, that snack variety has no significant effect on customer satisfaction because P equals point two four. So the only players in the game are seat comfort and flight length, what we see is both of them are significant. And so beta one is four, not 40 is in the last example, beta three is 0.1. so what is this saying?

## Evaluating Logistic Regression

Each rating step (e.g., 100 to 101) of flight length increases the log odds of a customer being satisfied by 0.1.

**Log odds** – log (odds ratio)

**Odds ratio** – odds of success  
odds of failure (80% chance of rain/20% chance of rain)



It's saying that each rating step, for example, one to two of seat comfort plus one to plus two. Increases the log odds of a customer being satisfied by four point one times. It's also saying that each rating step for example, 100 to 101 miles of flight length or kilometers, depending on how you're measuring. Increases the log odds of a customer being satisfied by zero point one. Now again, this probably all sounds a bit confusing if you don't have a background in statistics. But what we're basically trying to say is this last bullet point here, the key is that seat comfort.


## Evaluating Logistic Regression

The key is that seat comfort has more impact than flight length on the probability of flipping customer satisfaction from unsatisfied to satisfied.



Has more impact than flight length on the probability of flipping a person's customer satisfaction from unsatisfied to satisfied. And so, these logistic regression formulas, although they look complex, they're really trying to get at the same thing as linear regression. We're just trying to understand, what are the things or what are the factors that contribute to people's customer satisfaction. So, this is a statistical way to analyze via linear regression and a logistic regression getting at customer satisfaction.

Evaluating Airline Survey Data



**Dataset Background**

Dataset pulled from Kaggle competition

US Airline – identity unknown

Creating a survey has many nuanced considerations, but we are focusing on the post-survey statistical evaluation.

Dataset can be accessed from: <https://www.kaggle.com/teejmahal20/airline-survey-data>

Okay everyone, now that we've talked about the stats behind all of this. Let's dive deep and evaluate airline survey data that we have from the Internet. So here's some dataset background. This dataset was pulled from a Kaggle competition. The identity of the US Airline is unknown. And I don't actually know how this survey was created and distributed, but remember creating a survey has many nuanced considerations. We are focusing on the post survey statistical evaluation with regard to this data.

## Dataset Fields

### **Categorical/nominal variables (No measure of distance between values and order does not matter)**

**Gender:** Gender of the passengers (Female, Male)

**Customer Type:** The customer type (Loyal customer, disloyal customer)

**Type of Travel:** Purpose of the flight of the passengers (Personal Travel, Business Travel)

**Class:** Travel class in the plane of the passengers (Business, Eco, Eco Plus)

Now, I've included these slides here as somewhat of a key so that you can understand your data. So that, please use this slide deck as your way of understanding what each one of the variables in the dataset mean. But I'll talk about the general categories of the data that we have in this survey dataset. First, we've got a bunch of categorical or nominal variables in this airline dataset. What does that mean? Categorical and nominal variables, there's no measure of distance between the values and order does not matter. For example, gender of the passengers. There's no distance between a female and a male and order doesn't matter, can be male, female, female, male. Customer type, loyal customer, disloyal customer. They are categorical and they are nominal variables. And so, there're four types of those variables gender, customer type, type of travel, and class.



## Dataset Fields



### **Interval variables (Measure of distance between values are equal and meaningful)**

**Age:** The actual age of the passengers

**Flight distance:** The flight distance of this journey

**Departure Delay in Minutes:** Minutes delayed when departure

**Arrival Delay in Minutes:** Minutes delayed when arrival

Another type of variable set that is in this dataset is interval variables. Where the measure of distance between the values are equal and meaningful. So they're not just meaningful, they're equal. So for example age, the difference between 23 and 24, and age 45 and 46. It might feel different but it's actually equal. Flight distance, the difference between 101 kilometers, and 102 kilometers, 103 kilometers. The distance between those values are equal and meaningful. Departure delay and arrival delay, also they fall under these interval variable categories.

## Dataset Fields

**Interval variables on a 1-5 scale (low to high), with 0 as not applicable**

**Inflight Wi-Fi service:** Satisfaction  
level of the inflight Wi-Fi service

**Departure/Arrival time convenient:**  
Satisfaction level of Departure/  
Arrival time convenient

**Ease of Online booking:**  
Satisfaction level of online  
booking



And then we've got a bunch of interval variables that are on a very specific scale, on a 1 to 5 scale of satisfaction, where 1 is low satisfaction and 5 is high satisfaction with 0 as not being applicable. And we have a bunch of these, so for example, satisfaction towards inflight WiFi service.

## Dataset Fields

Interval variables on a 1-5 scale (low to high), with 0 as not applicable

**Gate location:** Satisfaction level of gate location

**Food and drink:** Satisfaction level of food and drink

**Online boarding:** Satisfaction level of online boarding

**Seat comfort:** Satisfaction level of seat comfort



Or satisfaction level of food and drink, satisfaction of the example we've been using seat comfort, so on and so forth.

## Dataset Fields

Interval variables on a 1-5 scale (low to high), with 0 as not applicable

**Check-in service:** Satisfaction level of check-in service

**Inflight service:** Satisfaction level of inflight service

**Cleanliness:** Satisfaction level of cleanliness



## Dataset Fields

**Interval variables on a 1-5 scale (low to high), with 0 as not applicable**

**Inflight entertainment:** Satisfaction

level of inflight entertainment

**On-board service:** Satisfaction

level of on-board service

**Leg room service:** Satisfaction

level of leg room service

**Baggage handling:** Satisfaction

level of baggage handling



And so you can use this as a key for understanding all the customer satisfaction variables that are in this dataset.

## Dataset Fields

**Categorical/nominal target variable**

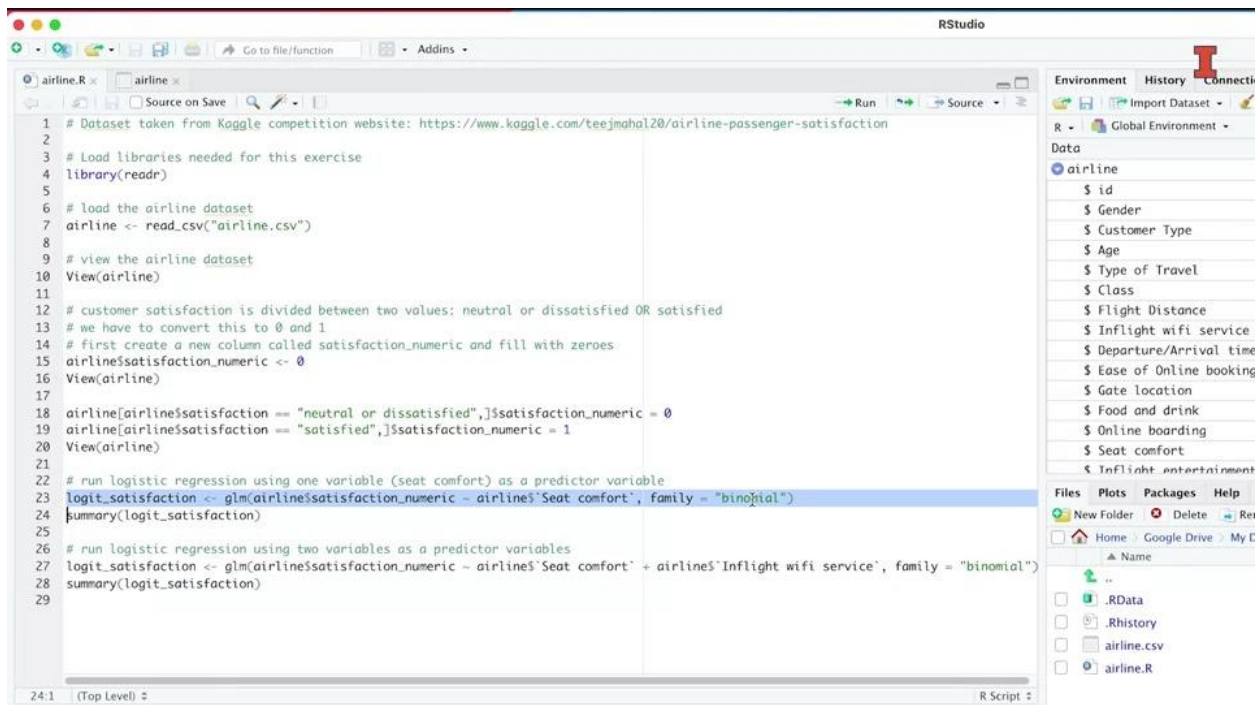
**Satisfaction:** Airline satisfaction level (satisfied, neutral or dissatisfied)



And then lastly, we have this and I pulled this out even though this is a categorical nominal variable. I pulled this out as a separate slide because this is our most important variable. This is our target variable, what we're overall trying to measure, which is airline satisfaction level. And you'll notice that there's only two categories or two values



possible for this overall satisfaction, which is satisfied. And then they've taken neutral or dissatisfied and put it together. So it's either satisfied, or neutral, or dissatisfied, or we could say one or zero. And so you can already imagine that how we're going to look at this dataset is via the lens of logistic regression because our target variable only has two options satisfied, or neutral, or dissatisfied. Now let's jump into our studio and get into the data.



```

1 # Dataset taken from Kaggle competition website: https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction
2
3 # Load libraries needed for this exercise
4 library(readr)
5
6 # load the airline dataset
7 airline <- read_csv("airline.csv")
8
9 # view the airline dataset
10 View(airline)
11
12 # customer satisfaction is divided between two values: neutral or dissatisfied OR satisfied
13 # we have to convert this to 0 and 1
14 # first create a new column called satisfaction_numeric and fill with zeroes
15 airline$satisfaction_numeric <- 0
16 View(airline)
17
18 airline[airline$satisfaction == "neutral or dissatisfied",]$satisfaction_numeric = 0
19 airline[airline$satisfaction == "satisfied",]$satisfaction_numeric = 1
20 View(airline)
21
22 # run logistic regression using one variable (seat comfort) as a predictor variable
23 logit_satisfaction <- glm(airline$satisfaction_numeric ~ airline$Seat comfort", family = "binomial")
24 summary(logit_satisfaction)
25
26 # run logistic regression using two variables as a predictor variables
27 logit_satisfaction <- glm(airline$satisfaction_numeric ~ airline$Seat comfort" + airline$Inflight wifi service", family = "binomial")
28 summary(logit_satisfaction)
29

```

So we'll just go directly to my working directory that includes the airline.CSV dataset and the airline.R R script that I created for this module. And so I'll load this R script and we see it right here. And just as a note, this dataset was taken from this Kaggle website competition, so you can find the dataset source right there. Let's load this dataset in by pulling in the library that allows us to pull in CSVs. And let's run this chunk of code. And we see that this dataset has been turned into a data frame here, and we can actually view that data frame. So you see right here. Remember the key point with this is that, satisfaction is stored as a two value variable neutral, or dissatisfied, or satisfied. Now remember with linear regression you need a continuous target variable. And in this case, we have a target variable that only has two values. And so that leads us to use logistic regression to be able to predict customer satisfaction in the case of this airline dataset. Well, the problem with this target variable is that to perform logistic regression we need to convert these string values to 0 or 1. So neutral or dissatisfied needs to be converted to 0 and satisfied needs to be converted to 1. So this snippet of code is doing exactly that. First what we're doing is, we're creating a whole new column. I'd rather instead of rewriting over that satisfaction column to just create a new column called satisfaction numeric and populate it with zeros first. So let's do that. And so then now, if



we look at the new airline updated dataset with a new column, we still have the satisfaction column here. But we have a new satisfaction underscore numeric column with just zeros. Now let's replace all the zeros or keep the zeros as zeros if the satisfaction column shows a neutral or dissatisfied in that same row. But if in that same row, if there is a satisfied value in the satisfaction column, then let's change satisfaction numeric value to be 1. And so let's run this. And we see that this worked. So now we have satisfied as 1, there's another satisfied as 1, but neutral or dissatisfied state as 0. So now we just have to run the logistic regression. Let's try our first logistic regression using just one variable as a predictor which is seat comfort. And so what we need to run is this function called GLM, which stands for general linear model. And the way that we do logistic regression is by indicating that the family of GLM functions that we're pulling from is the binomial family. Binomial means two values which in this case we have 0 and 1, which indicates logistic regression. So the binomial family of general linear models means we're doing a logistic regression. And then the way that this function works is you're saying that, I want to predict the target variable which is satisfaction underscore numeric found in the airline data frame. And I want to predict it using seat comfort which is found in the airline data frame. And you use this little indicator here to talk to say that, we're using these right side predictor variables to try to predict this left side target variable. And that we're storing all of the output of this function into this variable called logic\_satisfaction. And then we're going to ask for a summary of all the output from that logic\_satisfaction. So let's run this. Okay so now we've got the output here in the bottom and so let's go through just pieces of this because we've got a lot of output here that we haven't even really covered in this class. But what we see here is that we have this predictor variable seat comfort. And the first thing that we can see is that the p value is traumatically less than 0.05. In fact it's a dramatically less either the - 16.

```

29
24:28 (Top Level)
R Script
/Volumes/GoogleDrive/My Drive/Teaching/MBA564/Module 2/Dataset/

> View(airline)
> airline[airline$satisfaction == "neutral or dissatisfied",]$satisfaction_numeric = 0
> airline[airline$satisfaction == "satisfied",]$satisfaction_numeric = 1
> View(airline)
> logit_satisfaction <- glm(airline$satisfaction_numeric ~ airline$Seat comfort', family = "binomial")
> summary(logit_satisfaction)

Call:
glm(formula = airline$satisfaction_numeric ~ airline$Seat comfort',
    family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4495   -0.9406   -0.5550    1.1702    1.9732

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.395997   0.019340  -123.9  <2e-16 ***
airline$Seat comfort'  0.603242   0.005049   119.5  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 177814  on 129879  degrees of freedom
Residual deviance: 160973  on 129878  degrees of freedom
AIC: 160977

Number of Fisher Scoring iterations: 4
> |

```

So we're talking about 0.00000 and just continue to go on and so dramatically less than 0.05. And so we know that this is a statistically significant result and that we can see the coefficient remember the beta coefficient. We can see that it is 0.6 which is again the odds ratio. And so we can see that seat comfort is a significant predictor of customer satisfaction. Well let's try using two variables as predictive variables. Let's use seat comfort and let's use in flight wifi service. But the way that we use to variables is that we still keep it on the right hand side of this indicator in GLM. But what we do is we put one variable and then we use the plus sign to add a second variable. And then we could use another plus sign to add a 3rd and 1/4 another plus sign and so on and so forth. So that's how you add a multi variant logistic regression variables. And so let's run this again to see if seat comfort and in flight wifi service are predictors of customer satisfaction. So you run this and we see that both of these variables have p values that are highly significant. So they're both predictors, we can see here that seat comfort is a slightly better predictor of customer satisfaction than in flight wifi service, although both are predictors. There are significant predictors, but it seems that seat comfort has more of an impact than in flight wifi service on customer satisfaction. And so this is just a really quick and brief primer on how to use are, and specifically this our script to analyze this airline survey data using logistic regression.