Optional: Exploratory Data Analysis
Video Transcripts

Video 1: Introduction to Exploratory Data Analysis (4:19)

In this module, we're going to be talking about exploratory data analysis. Understanding your data before trying to build a machine learning model is absolutely crucial and will help you avoid making careless mistakes and erroneous conclusions. The term exploratory data analysis was coined by John Tukey in 1977. He is the super-famous statistician best known for the development of the box plot, so in this module we'll be covering both box plots and exploratory data analysis because box plots are part of exploratory data analysis and getting a feel for our data. Exploratory data analysis can also be leveraged to understand which hypotheses we might want to test in the future based on the different patterns that we're seeing. So the methods that we use for exploratory data analysis you may have heard of previously but we thought it was super important to pull everything together so you can see step by step exactly the process that you're going to go through each time you're working with a new dataset.

In exploratory data analysis, we're going to look at summary statistics and frequencies and really get a sense for what our data looks like in terms of tables and numbers. And we're also going to look at different visualizations depending on whether our data is discreet or continuous. We'll be looking at histograms, box plots and scatter plots of our data to get an understanding of the distributions. Dimensionality reduction can often be considered an exploratory data analysis technique; however, we're not going to covering that in this module. I typically use dimensionality reduction when I'm working with a large dataset. I personally do not use dimensionality reduction too frequently unless I'm doing something like a cluster analysis using unsupervised learning. We're also going to take a look at correlations, which are incredibly important understanding how our variables are correlated before we go to modeling.

Putting multiple variables that are highly correlated in a machine learning model can sometimes lead to problems; however, there are methods to deal with this. Here we're just going to be focusing on looking on how the variables are correlated with each other, generating a plot and taking a look at which variables are the most correlated with what our response variable would be. During the exploratory data analysis process, you'll often find areas where you can do more data cleaning. This is because exploratory data analysis and data cleaning are really iterative processes. I'll often open up a dataset that is new to me, look at it manually, understand that there's areas that I can clean, I will clean that data and when I go to do the exploratory data analysis, I'll often find more areas where I'm able to remove outliers, I'll identify new duplicates that I wasn't aware of, I'll find these inconsistencies and missing data and I will then go and clean that up. Then I will return back to exploratory data analysis and I will iterate between these two things until I get to a point where I feel like my data is clean and I thoroughly understand it. So, in summary, in this module, we're going to cover EDA and that's going to include looking at summary statistics and frequencies, a number of different plots and looking at correlation between variables.

****

Video 2: Descriptives, Frequencies, and Averages (10:24)

Alright. Here we're going to be looking at descriptives, frequencies, and averages. And this is often our first step in exploratory data analysis. So, the first thing that we're going to do is pull in or import our libraries. Canvas and NumPy. I'm also reading in the data set, employee attrition. And I am storing that in the variable DF, for data frame. And we'll be using this data set for our exploratory data analysis module. Okay. So, the first thing we're going to want to do is look at some descriptive statistics. So, this line right here is just formatting my floats so that they're only 2 decimal places. And it's going to do that across all of the columns. And so, if I didn't do that, then I might have decimal places to 6 places, 7 places, and it gets more difficult to read. So, this is really just for readability. I'm setting an option to format a float. And I'm telling it that it's going to be 2 decimal places. This df.describe is what gives us this table. And so, these are summary statistics for all of the variables in the data set. So, it's super handy. And when we look at this, so we see we have 1470. And so, it's going to be the same number across. If it was not, that means that we have nulls. So, that is really useful in helping us to identify where our nulls are.

We see that it gives us a mean, a standard deviation, a minimum, our quartiles, and the maximum. Alright? So, for age, we can see that the average age is 36 years, 36.92 years, so almost 37 years. We see that the minimum age is 18. The maximum age in our data set is 60. And we see that the 50th percentile, so the median, is actually quite close to the mean. Meaning that our distribution isn't too skewed. And so, I'm able to start inferring information. This also helps me understand the structure of my variables. So, if I was to look at the data set, I'd notice that job level is just a number 1 through 5, so it's probably not something that I'd want to use in a model as an actual number. I'd want to change this to a factor variable. And you know, it's an ordinal factor variable. And so, doing the descriptive statistics lets me very quickly get an understanding of what type of variables I have. Whether or not they are skewed. And in a relevant business context, this information would mean more to me. And understanding the min and max and quartiles for variables will help you get a quick understanding of if there's outliers that you need to remove.

If a variable has incorrect or invalid data in it. Like, I have seen variables before that are supposed to be a rate. And a rate is supposed to be not a daily rate in terms of dollars, but I mean a rate 0 to 1. And there were numbers greater than 1, which meant that the logic behind what was calculating the variable was incorrect. Right? So, we want to be able to look at all of these statistics for all of our data to start to identify. Not just get a feel for how the data's distributed but also get an idea if there's any more cleaning that we need to do. Okay? And so, obviously if we want to just look at the statistics for a single variable, we can do that. However, getting the summary statistics for the whole data frame at once is really useful. It’s nice to write one line. And get a ton of information back. Okay? And so, frequencies are also something that we want to be able to look at. So, I can get the count of a variable, which we already know that it's 1470 from our summary table up above. But I can get the value counts, as well. And so, this is for age, this is something that we typically want to put in a histogram.

A visual may help us more. However, I sort of do both. Because this is still going to be very useful to us, this value counts function, for our discrete variables that have only a limited number of categories. Okay? So, here I'm going to look at our categorical features. And so, this isn't necessary, but I've written some extra code. Really, what we're doing is we're looking at the value counts for the columns that have less than 30. So, up here, with age, this is really hard to read in terms of frequency output. It would be better as a histogram. But when we have a smaller number of unique entries, then it makes sense to look at them in a table. And so, here I'm just printing some formatting so that we get the column name and some information. But you can use this for your data set in the future. And now you will be able to see, okay, what was the name of the column? What were the different options? And then the count of those options. Okay? So, I can see that only a few people here are nontravel. Most of the people travel rarely. Very few, less people travel frequently. But frequent travel is more likely than nontravel. I can see that we have more men in our data than women. I can look at the marital status and over time. And so, this is a nice way to be able to look at our data and get those frequency tables really quickly without having to manually type in each variable. Which is super nice, because 10 years ago when I was doing data science, I was typing out a lot of things manually. Right? So, here is a nice function, as well. So, here I'm just writing a loop. So, for column in the data frame, for each column in the data frame, we're going to print the column name. And then we're going to print as a string number of unique values. And then we're going to use this an unique function to tell us how many unique values there are. So, 43 makes sense for age, because we knew the youngest was 18. The oldest was 60. That's 43 years.

We have our attrition variable, which it's really nice to see. Okay, we only have yes or no. There's no dirt in our data that's something other than a yes or no. Right? And we can see that real quickly by understanding that there's only 2 unique values, different values in our data set. Right? There's 71 different values for hourly rate. However, people typically have their own value for monthly income, because this is almost the total number of our data set. And so, this is relevant information for us, as well, to help us understand our data set. Okay? And then, you know, like I said. If we want the individual mean, standard deviation, or median of our data set, we can do that on individual columns. And you will do that sometimes. But really, the idea of writing a loop so that we can get a lot of information quickly is really helpful to us. So, here we wrote 2 loops, but we also had this describe function that gave us the descriptive statistics for our whole data set.

****

Video 3: Correlation (8:25)

In this video, we're going to talk about correlation. So, first we'll talk about correlation magnitude and sign. And we'll talk about the difference between correlation and causation. And then we're going to look at a plot of correlation using Python. So, in terms of magnitude and sign, correlation is always a number between negative 1 and positive 1. So, the sign can be either negative or positive. A correlation of 0 implies that there is no correlation or no linear association between 2 variables. So, a value of negative 1 means that there is a perfect negative association between 2 variables. So, as your age increases, the amount of life expectancy you have left decreases. Right? So, that is a negative relationship, negative association. A correlation of value of 1 is a perfect positive linear association. So, as 1 variable increases, the other variable is increasing. And there is a relationship there that is perfect. And so, the number can be anywhere between negative 1 and positive 1. So, a value of 0.05 is really small. And we might even consider saying that those variables aren't really correlated. Because it's so close to 0. But as the correlation gets larger, we start to comparatively say that there is a larger association between these 2 variables. As one variable increases, the other variable is increasing as well. And we could say the same for as the values become, as the magnitude becomes even greater on the negative side. Okay.

And so, we want to be careful that we don't mistake correlation for causation. Right. So, being out in the sun with no sunscreen causes a sunburn. That is scientifically proven. However, when there's a really hot day, people are more likely to eat ice cream potentially. However, we know that the sun being out does not cause someone to eat ice cream. And so, this is really important in modeling, because we will see that we have highly significant variables in our model. That is correlation. That is not causation. The way that we would determine causation is to then say, okay, there's a relationship here. We are going to set up a hypothesis test. And we're going to run that test and see if there is a statistically significant difference. And because we would have randomized to teste and control. And done this in a scientific way. At the end of that test, we'd be able to determine whether or not there is a causal relationship. But here we are just talking about 2 variables being correlated. And were not making any assumptions that one variable necessarily causes the other variable. So, to actually create a plot of correlation or to look at correlations in Python, first we are going to call in Pandas, NumPy, and then our visualization libraries. Matplotlib and Seaborn. We are going to use our percent Matplotlib inline that allows us to render a plot. And we're going to read in our data set. Because although we've read in this data set before, every time we're in a new notebook, we need to call in the data again. So, I am again storing the employee attrition data set in this variable named DF. And then I am going to set up my plot area, giving it a size of 15 by 15. And honestly, I had to play with that a little bit to get it so that it would fit on the screen and allow us to see the labels in both axes. So, you may want to make it smaller or larger. But this was literally just generating a plot with the dimensions that allowed me to look at the heat map of correlation. Okay? And so, I'm able to, using the Seaborn library, create a heat map, which is really handy for correlation.

We can very quickly understand with the coloring what has a high correlation or correlation closer to 0. So, you'll notice that on the diagonal axes, that the values are all going to be 1. Because of course age is perfectly correlated with itself. So, that is not super interesting. But what is more interesting is to look at these. So, job level is highly correlated with the total number of years that you were working. Which makes perfect sense, right? We expect that as the number of years that you've been working increases, your job level is also going to increase. We also see, and this is very interesting, that the years with current manager is positively correlated with the years since your last promotion. And the years since last promotion is also highly correlated with the number of years at the company. Which also makes sense, and we need to use our critical thinking skills here, right? So, if somebody has only been with the company for a short period of time, they haven't had the opportunity yet to get promoted, right? So, part of that is going to be biased. Just the fact that if you've been with the company for less than a year, you don't really have the opportunity to get promoted. And so, as you're with a company for a longer period of time, you then have the opportunity to be promoted. And as your years at the company increases, you're also able to have a longer number of years since your last promotion, right? And so, we want to start thinking about these things.

It's not just about the data, but it's about what the data tells us. And this information is going to be relevant when we go to do modeling. Again, I have mentioned that, you know, for certain types of models, you don't want independent variables that are too highly correlated with each other. But now you're able to generate a really nice plot that very clearly identifies to you which variables are correlated with each other and what variables have little or not linear association between them.

****

Video 4: Visualizing and Plotting Data in Exploratory Analysis (8:42)

Here we're going to be talking about data visualization for exploratory data analysis. So, we will be going over plotting continuous and discrete variables. This is different from what we learned in best practices in data visualization. Where we're really trying to convey a story to our stakeholders. And we're going to use color in a certain way and we're going to make sure that our plots are pretty. In exploratory data analysis, we're looking to get an understanding of all of our variables. So, we're going to look at a bunch of plots. All of these plots are not going to go the stakeholders because they won't all show information that is relevant, interesting, or even useful. But we don't know that it doesn't contain relevant or useful information until we look at it. But, so here, we're going to be generating a lot of plots and not worrying too much about the formatting. We're trying to get an understanding of the story for ourselves. And then we'd worry about formatting and making the relevant plots to share pretty, later. Okay. So, I am going to be calling in pandas, and numpy, and our visualization libraries. And we are still using the employee attrition dataset and assigning that to the variable DF. Okay, before I start plotting, I'm going to look at my data, and we have mentioned before that we have these variables like environment, satisfaction, job involvement that are really a rating of 1 to 4 and aren't a true numeric variable.

So, I am going to classify these as the correct data type. They are not actually int 64. These should be classified as objects. And so, having our data in the correct datatype is going to allow us to quickly create our visualization because these would be more relevant as bar charts than as our continuous plots. So, as I said we want to be able to get lots of plots quickly. So, here what I'm doing, is I am going to write a loop. So, 4 I in columns. And I've created columns as a list, including some of the variables that are continuous. That would be appropriate for a histogram. So, I have created this loop and now I am going to get a number of histograms back. And so, I would take my continuous data. I would put it in there. And I'm able to look and say okay years in current role is bimodal. Why might that be? You know of course, some people it's going to be that they're new to the company. Some people it's going to be that it's been a long time since their last promotion. And I may want to dig into this deep. But I can understand that this is not normal looking data. There's two areas of time that appear to be popular, like this six-year time period. And what's probably two or three years here. Okay, so that's going to be relevant information to me when I go to modeling. I can see the years at company is really high in the beginning, and it sort of tapers off. And so, I'm starting to get an idea of the structure of the data. So, it's interesting that this is pretty evenly distributed. There's a lot of people that live super close. I wonder if IBM office is in the city. Because these people are very close. And so, we were quickly able to generate a number of plots. We'll also want to look at scatter plots. And doing the same you could create a loop to get these scatter plots, which you're looking for. However, we did already look at correlation as well. So, we do know already which variables are highly correlated and which are not. Right, we'll want to look at fox plots. And so, if we didn't create a distinction that we wanted this by attrition, then we'd be getting information that wasn't much more than what was in our summary statistics table. But by looking at these different continuous variables by attrition, we're able to just quickly got the idea of the differences there.

What we just talked about were all four continuous variables. And then, we're going to also want to look at all of our discrete variables. And so, again, I would set up a loop. I would give the columns that have the discrete data. So, this is the data type object and so I've added environment satisfaction here. And I could very easily add the other variables that we were covering as well. But the idea is that you know, I didn't really set the plot figure size. I didn't worry too much about what the colors were. I'm just trying to get a lot of plots quickly so that I can understand my data. Okay. And we want to be careful how we interpret this here. We know that because these are pounds and we know that there are more men than women in our dataset. But comparatively, we can see here that divorced people are much less likely to leave the company that single people. And people who scored their work-life balance as 1, there's not a lot of them, but by comparison, attrition compared to people who were not in attrition is quite high. And you can scroll down to see this last variable here, environmental satisfaction. So, we see people with and environmental satisfaction of 4 having a much lower rate of attrition, comparing these two bars to each other than if somebody only had an environmental satisfaction of 1. So, I can sort of assume from looking at this that environment satisfaction of 1 is probably poor and 4 is better. Okay, and so each time we go to do exploratory data analysis we'll want to do a series of plots. And it's really good that we created this video and put it all together, right. So, you're going to create histograms, scatter plots and box plots for your continuous variables. And you can do that in a look to make it more quick. And you're going to want to create bar charts for your discrete variables. And take a look and intuitively understand what is going on with your data.

****

Video 5: Data Preprocessing (7:21)

Here we're going to talk about data preprocessing. Specifically, we're going to look at one-hot encoding and feature engineering. So, after you have done your exploratory data analysis. You've looked at plots. You understand what data type your different variables are, you may have some thoughts on how you can transform your data to create relevant variables for modeling. So, here we are going to pull in Pandas and NumPy and our data set, as we have been doing. And let's look at our data set real quick. So, we have this variable, environment satisfaction, that we know is not really numeric. These are people who, you know, on a scale of 1 to 4, their environment satisfaction is a 2. So, this should actually be an object variable. And so, when it comes to modeling, these type of variables should be one-hot encoded. To do this, we have a nice get_dummies function. And so, I'm going to run this. And then I'm going to show you on the data set exactly what happened here. But so, instead of having a single variable where each of the entries is a number 1 through 4. Instead, we're going to create 3 variables on that single variable. And then we're going to drop the original variable. Because at that point, the information will be redundant in our data frame. Okay, so, I'm able to do that for a number of the factor variables that we had that were in our data set. And business travel was not a scale of 1 to 4. But it is an object variable. And it was, you know, are you traveling frequently, rarely, whatever. That, again, is a discrete categorical variable and would be a good use case for one-hot encoding, right? So, after I run that, I can look at the data, and I see that I no longer have a job satisfaction variable up at the top. Instead, I now have job 2, job 3, and job 4. And involvement is a similar variable. And environment satisfaction was the same, so these have all gone from being a single variable to now being 3 variables. And with business travel, there is now a rarely and a frequently. You may be wondering why job satisfaction is 3 variables and not 4. And it's because you only need 3 variables to get that information.

If somebody is set to 0 for job satisfaction 2 and 3 and 4, that means that they must be 1. So, if we were to have 4 variables here, we'd actually have redundant information. Because we're able to infer who was a 1 by looking across the 3 variables. And so, we can take a look at what this actually looks like. If we go over to the end, we see that for job, we'll look at job involvement. So, for job involvement, we now have 3 variables. You're either a 3 or a 2, 3, 3, 3, 4. Here is somebody who would have been a 1 in terms of environment satisfaction. And we can see that by looking across the other 3 environment satisfaction variables. And so, this is going to make it much easier when we get to building machine learning models to have our object variables, or our discrete variables. And so, one-hot encoding is considered feature engineering. Another type of feature engineering is when we create a variable, a new variable, that is a combination of previous variables. So, the example that is being given here is that percent salary hike, which is actually an integer in the data set. You know, it may be 11% or 13%. Well, we can assume that somebody making 300 grand a year might get a different percentage than somebody who was maybe making 60 grand a year. Maybe the executives get a higher percentage. Maybe they get a lower percentage since it would be a higher dollar amount. Regardless, one way that we may want to take our variables that we have and create a new variable based on this data would be to take the percent salary hike, which was an integer. So, I divide by a hundred to get it as an actual decimal. And then I'm going to multiply by the monthly income to see how much extra money these people actually got in terms of increases in their salary.

Alright, so, this is how I can create a new variable in the data frame that I'm working with. And so, I just give it a name. And then I literally just put a mathematical function using the syntax that we've come to know. And I am able to take a look at this. And I will now have an extra variable that is new feature dollar hike. And it's the dollar amount that each person's monthly income has increased by. Which may be a relevant way to look at the data. And again, after I do this, I might go back and look at another chart of this data. Alright? So, that is how you would go about one-hot encoding or feature engineering to create new variables for your data set.

****

Video 6: Exploratory Data Analysis: Summary (1:26)

In the exploratory data analysis module, we looked at ways to quickly generate a table of summary statistics and analyse those. We also looked at how to loop through plots so that we could quickly get a sense of our different variables, whether it was continuous or discrete data. But to be able to quickly generate all the plots that we need to get a better understanding of our data. We also looked at Pearson correlation and a heat map of those different correlations. And what correlation means. And we also looked at some data preprocessing techniques, including feature engineering and one-hot encoding our data. Again, exploratory data analysis is going to be an iterative process when you're doing it in industry. And you're cleaning messy data first. Moving to exploratory data analysis, you're often going to find more dirt in your data that you weren't aware of. You'll go back and clean it up and then come back and look at it again. I hope this module was helpful.

****

--END--