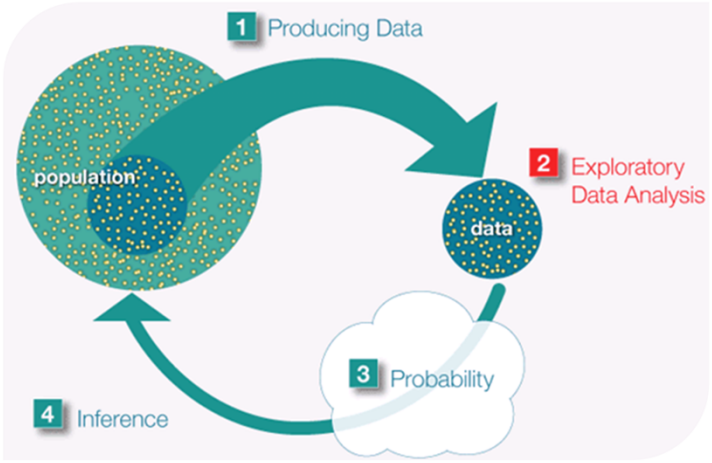
EXPLORATORY DATA ANALYSIS

Introduction to Unit 1

UF UNIVERSITY of FLORIDA

Now we begin our discussion of exploratory data analysis.

The Big Picture



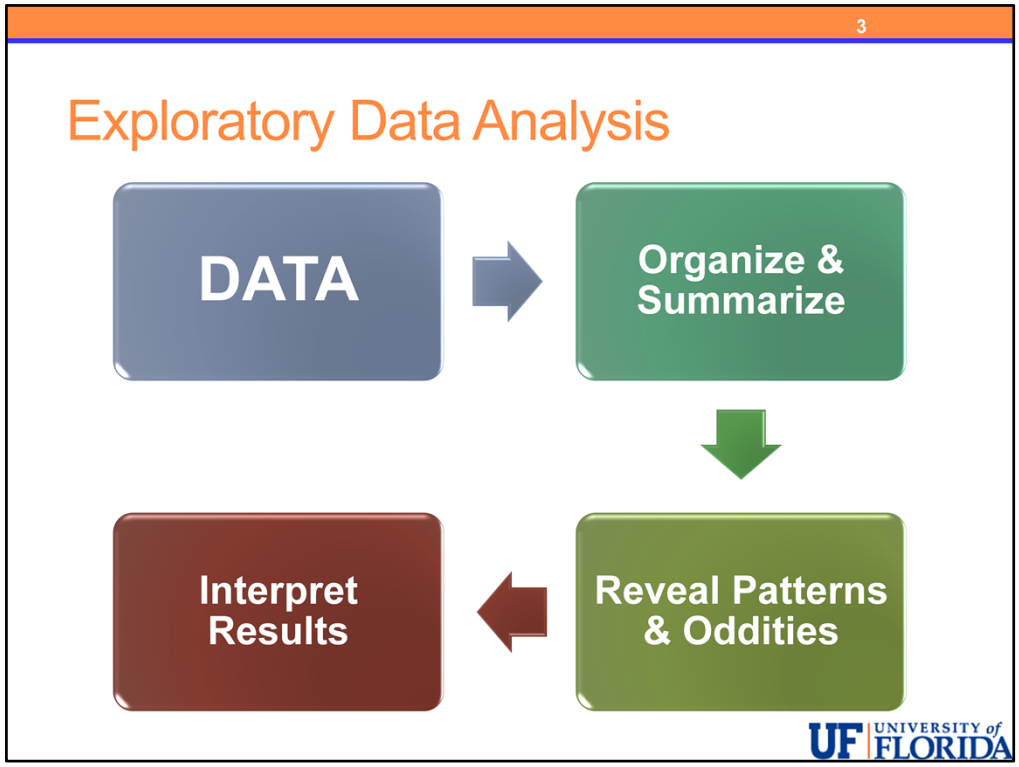
UF UNIVERSITY of FLORIDA

Remember to keep in mind where we are in the big picture.

For now, we will assume that the data we are given is a representative sample from a population of interest.

We will come back and discuss issues with Producing Data later.

In this unit on Exploratory Data Analysis, we won't yet be able to make any inferences about the population of interest with any degree of confidence, however, we will be able to describe the sample and discuss the results we find in context.



Once we have produced our data, we begin our analysis using exploratory data analysis (also commonly called descriptive statistics).

This process consists of organizing and summarizing the data so that we can begin to understand the patterns and relationships in the data along with any striking deviations from those patterns.

We also include interpreting the results, but only in reference to the current sample. Remember that the results we see here must be understood through the “cloud” of probability before we can make any inferences about the population.

Table 73d. **Special Education—Students Ages 6-21 Served under IDEA, Part B, by Select Diagnostic Categories: Fall 2007**

Data for VT are not available. All other missing values were suppressed to protect confidentiality. U.S. values do not reflect missing values.

State	All Disabilities	Category (see Tables 63a-63c for more)				State	All Disabilities	Category (see Tables 63a-63c for more)			
		Autism	Deaf-Blindness	Traumatic Brain Injury	Developmental Delay			Autism	Deaf-Blindness	Traumatic Brain Injury	Developmental Delay
U.S.	5,905,854	256,809	1,256	23,475	88,193	MO	122,663	4,966	34	457	531
AL	77,661	2,509	5	237	2,465	MT	16,187	389	---	63	0
AK	15,581	507	8	65	1,045	NE	40,508	1,244	10	223	1,079
AZ	117,039	4,884	6	289	0	NV	42,617	2,023	9	191	0
AR	54,170	1,837	8	168	0	NH	29,751	1,155	61	60	1,460
CA	602,902	36,328	150	1,688	0	NJ	230,519	8,864	39	1,114	0
CO	72,275	1,985	27	393	0	NM	40,047	763	9	191	2,943
CT	61,327	3,800	---	117	---	NY	390,675	15,817	6	1,226	0
DE	17,171	644	49	38	28	NC	171,754	7,463	34	488	6,603
DC	10,296	245	---	---	292	ND	12,056	398	9	36	555
FL	358,273	10,582	61	588	0	OH	246,605	9,895	42	1,116	0
GA	170,970	7,839	23	466	4,898	OK	87,706	1,906	20	281	6,690
HI	17,964	918	5	64	1,111	OR	69,692	6,071	21	258	0
ID	24,013	1,209	7	137	1,591	PA	265,720	11,802	58	788	79
IL	284,711	10,682	21	785	919	RI	26,066	1,184	---	65	0
IN	159,546	8,341	19	540	0	SC	93,259	2,337	6	184	292
IA	63,332	1,057	---	182	0	SD	15,288	523	---	56	0
KS	56,104	1,700	15	208	3,928	TN	108,661	3,477	---	------	4,027
KY	88,596	2,421	10	244	8,843	TX	435,221	19,479	96	1,247	0
LA	78,002	2,240	6	234	5,619	UT	55,043	2,356	37	288	1,605
ME	30,536	1,600	---	78	--------	VT	----------	--------	---	---	--------
MD	92,833	5,693	25	298	1,101	VA	151,651	6,813	10	375	6,859
MA	150,827	6,860	194	5,826	10,202	WA	110,169	5,415	27	351	7,505
MI	212,479	10,803	6	614	1,442	WV	42,006	904	22	128	0
MN	105,046	9,876	50	434	2,101	WI	111,629	5,575	---	378	---
MS	57,295	1,081	11	155	2,365	WY	11,412	349	0	63	15

Source: Office of Special Education Programs, Data Accountability Center, Table 1-3; <https://www.ideadata.org/arc_toc9.asp>; (accessed 26 June 2009).

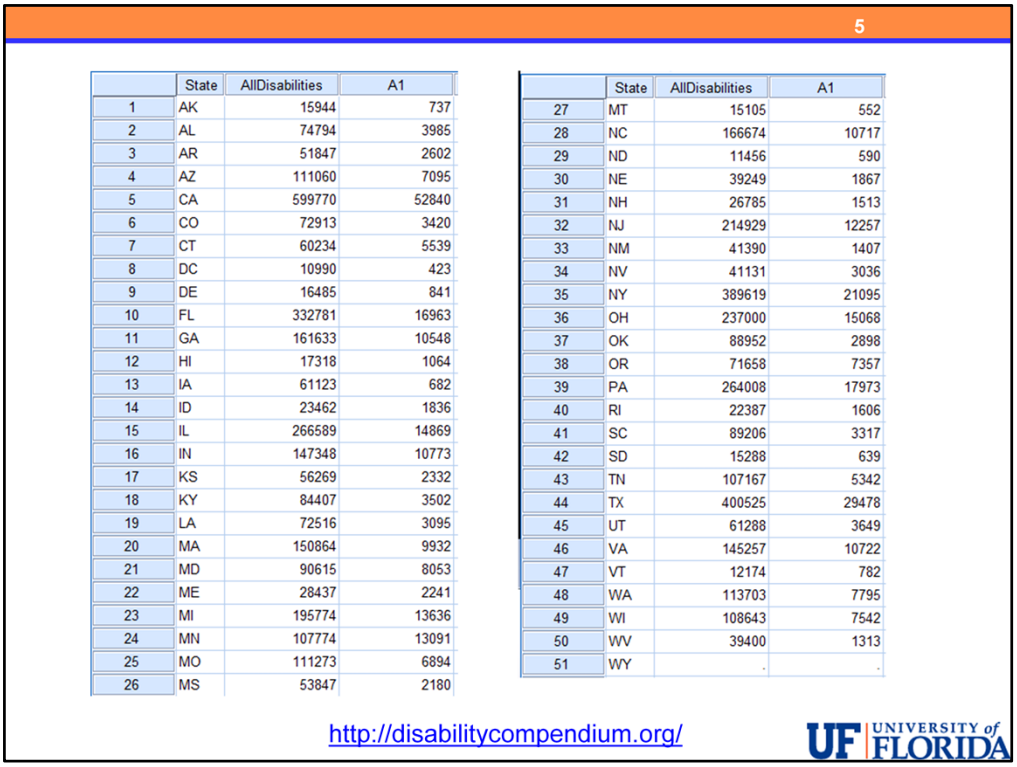
Annual Disability Statistics Compendium: 2009

113

UF UNIVERSITY of FLORIDA

Here is a sample dataset on Special Education which gives the total number of students enrolled in special education and the number of students in four categories (Autism, Deaf-Blindness, Traumatic Brain Injury, Developmental Delay).

Although we could understand something of the situation just using these numbers. It is best to summarize and organize them to help us see any patterns which may exist.



5

	State	AllDisabilities	A1
1	AK	15944	737
2	AL	74794	3985
3	AR	51847	2602
4	AZ	111060	7095
5	CA	599770	52840
6	CO	72913	3420
7	CT	60234	5539
8	DC	10990	423
9	DE	16485	841
10	FL	332781	16963
11	GA	161633	10548
12	HI	17318	1064
13	IA	61123	682
14	ID	23462	1836
15	IL	266589	14869
16	IN	147348	10773
17	KS	56269	2332
18	KY	84407	3502
19	LA	72516	3095
20	MA	150864	9932
21	MD	90615	8053
22	ME	28437	2241
23	MI	195774	13636
24	MN	107774	13091
25	MO	111273	6894
26	MS	53847	2180

	State	AllDisabilities	A1
27	MT	15105	552
28	NC	166674	10717
29	ND	11456	590
30	NE	39249	1867
31	NH	26785	1513
32	NJ	214929	12257
33	NM	41390	1407
34	NV	41131	3036
35	NY	389619	21095
36	OH	237000	15068
37	OK	88952	2898
38	OR	71658	7357
39	PA	264008	17973
40	RI	22387	1606
41	SC	89206	3317
42	SD	15288	639
43	TN	107167	5342
44	TX	400525	29478
45	UT	61288	3649
46	VA	145257	10722
47	VT	12174	782
48	WA	113703	7795
49	WI	108643	7542
50	WV	39400	1313
51	WY	.	.

http://disabilitycompendium.org/

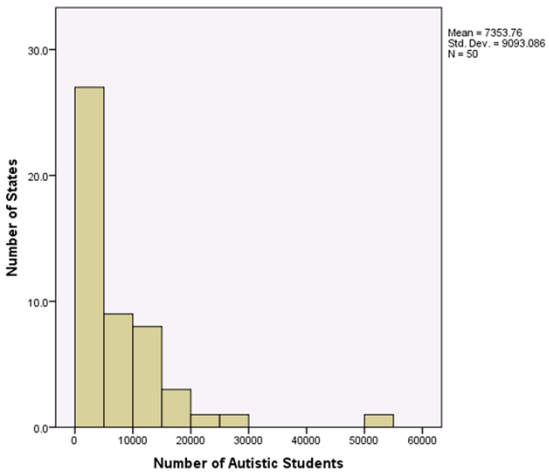
UF UNIVERSITY of FLORIDA

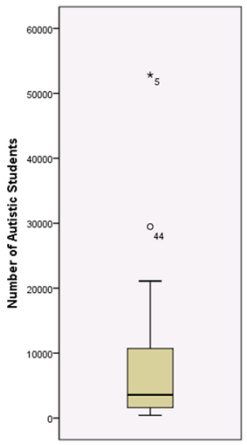
Here we have similar data from a more recent survey. The data can be found at http://disabilitycompendium.org/.

Next we will preview some of the descriptive statistics on quantitative variables that we will learn in this unit.

Don’t be concerned yet with any details!

Number of Autistic Students

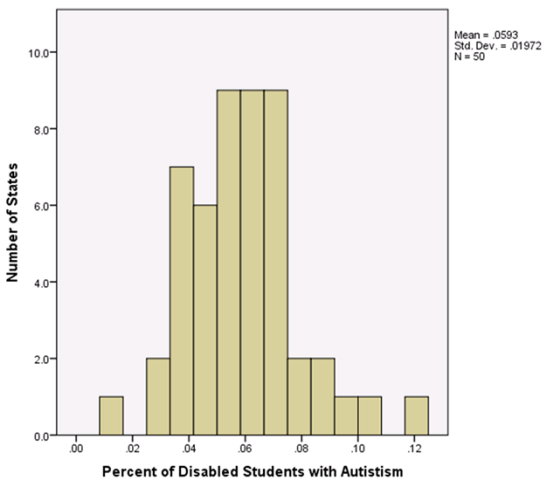


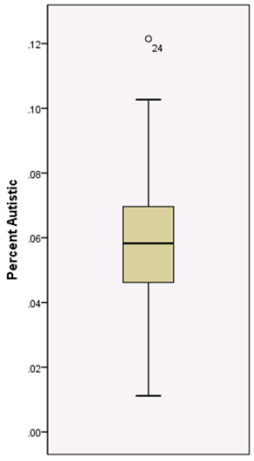


UF UNIVERSITY of FLORIDA

When we look at the number of autistic students in each state, we seem to be seeing a pattern related to the size of the state as the unusually large values for this variable represent the populous states of California (# 5 in the dataset) and Texas (# 44).

Relative to Total Number Disabled





UF UNIVERSITY of FLORIDA

If we look at the percentage of Autistic students relative to the total number of students enrolled in special education then we see a different pattern to the distribution.

Here, there is one somewhat unusually large value, Minnesota (# 24 in the dataset) for which 12% of its special education students are Autistic.

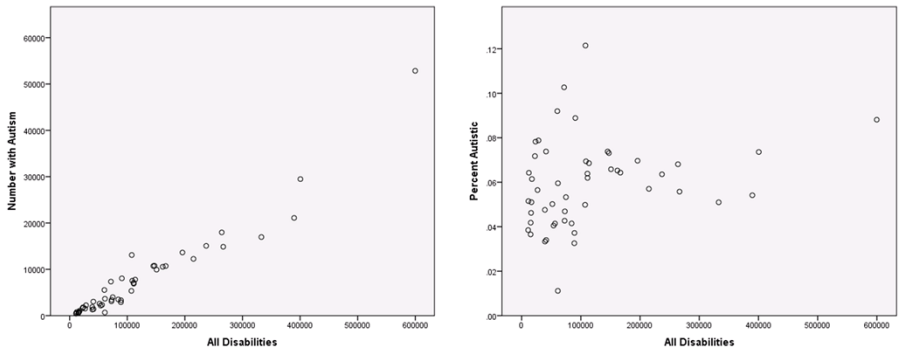
Descriptives

			Statistic	Std. Error
Percent Autistic	Mean		.0593	.00279
	95% Confidence Interval for Mean	Lower Bound	.0537	
		Upper Bound	.0649	
	5% Trimmed Mean		.0585	
	Median		.0583	
	Variance		.000	
	Std. Deviation		.01972	
	Minimum		.01	
	Maximum		.12	
	Range		.11	
	Interquartile Range		.02	
	Skewness		.572	.337
	Kurtosis		1.324	.662

UF UNIVERSITY of FLORIDA

On average, around 6% of students enrolled in special education are Autistic but this percentage ranged from 1% to 12% among states during this reporting cycle.

Relationships



Now, let's visualize the relationship between two quantitative variables.

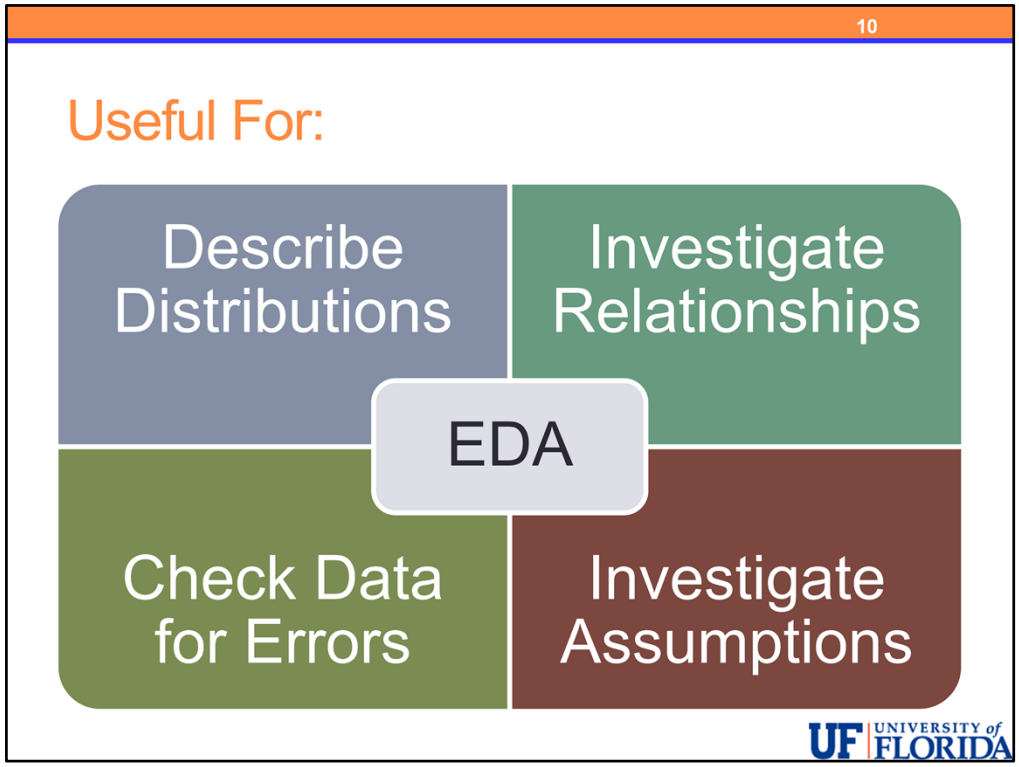
On the left we see that, not surprisingly, as the number of students in special education increases, the number of students with autism tends to increase as well.

In this case, most of the variation in the number of autistic students can be explained by the total number of special education students.

However, on the right, when we look at the total number of students vs. the percent autistic, we cannot see a clear trend.

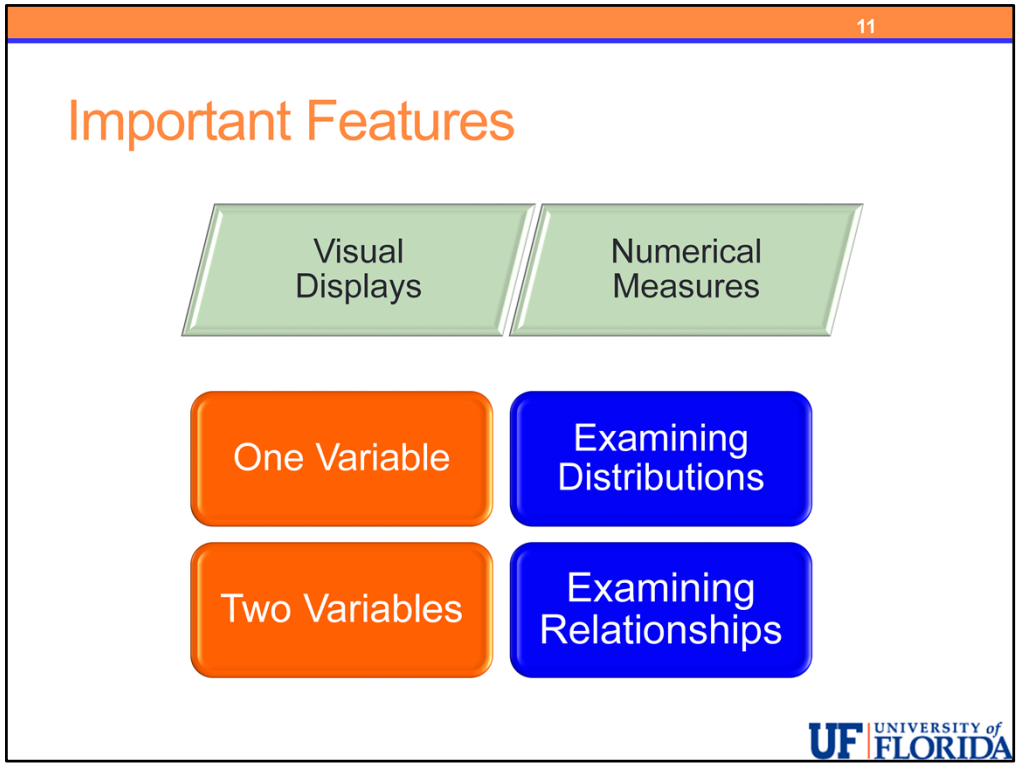
The variation in percent autistic does not seem to be explained by the total number of students.

We might be interested in determining what variables do explain the variation in percent autistic.



Exploratory Data Analysis is particularly useful for describing the distribution of one variable and investigating relationships between two variables (these topics will be the focus of this Unit).

Exploratory data analysis is very helpful for checking data for errors and investigating the validity of assumptions in more complex analyses.



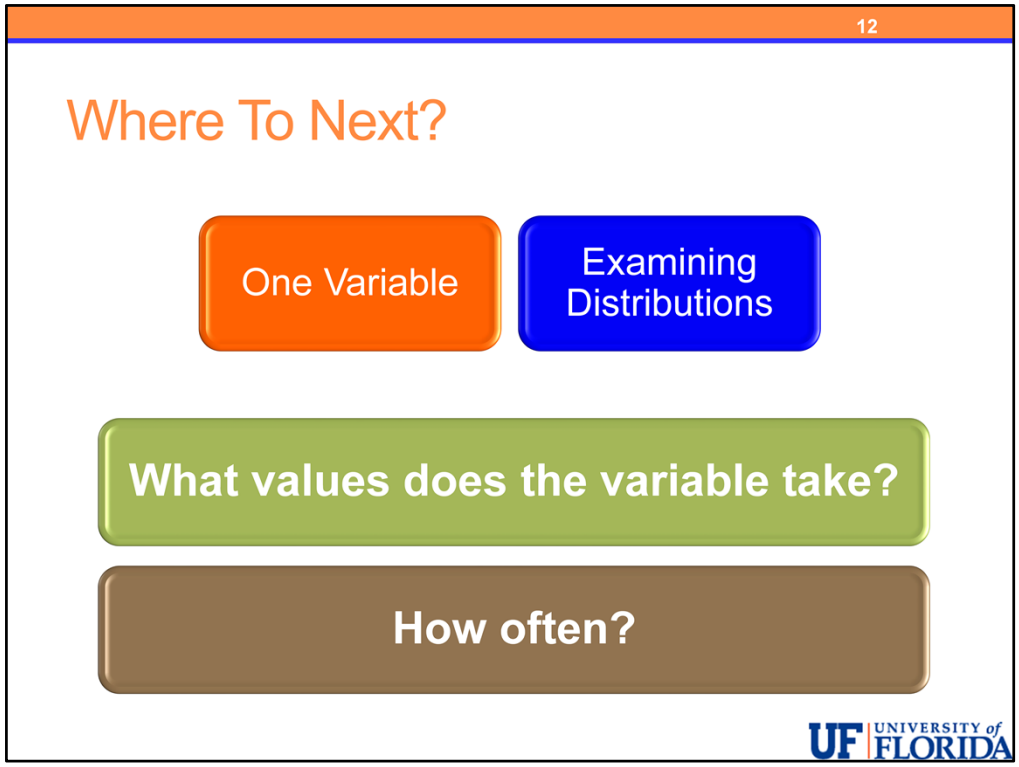
In general, a full exploratory data analysis will always consist of two elements

– visual displays which may include graphs and tables
And
– numerical measures which will include values such as frequencies, percentages, means, standard deviations, etc.

We also need to consider whether we are interested in only one variable at a time – examining distributions

Or if we have two variables where we wish to examine the relationship between them.

In many research problems involving statistics, we often study relationships between more than two variables, however, in this course, we are primarily concerned with covering situations involving one or two variables.



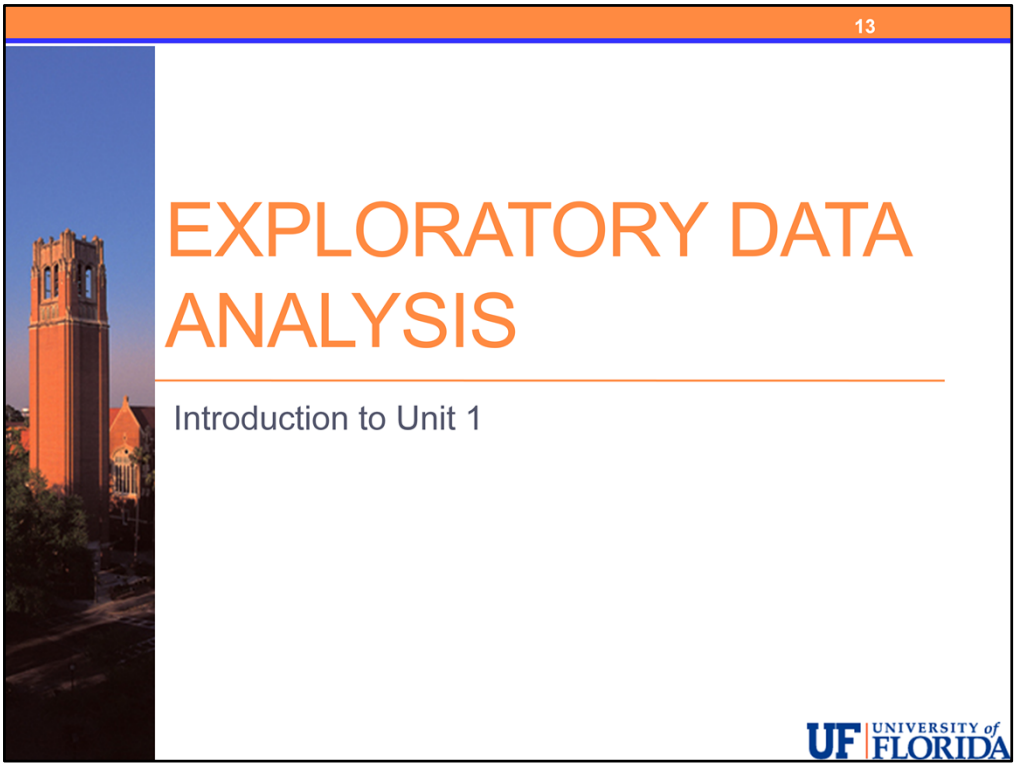
We will begin with exploratory data analysis for one variable at a time, examining distributions.

By Distribution we mean

What values the variable can take

And

How often the variable takes those values.



As you learn more about exploratory data analysis, think about how probability, chance, and randomness might be at work behind the scenes.