Set 2. Exploring Two-Column Data

Skill 2.1: Identify the data needed to answer a question
Skill 2.2: Interpret a crosstab chart
Skill 2.3: Interpret a scatter plot
Skill 2.4: Select that best two-column visualization (crosstab chart or scatter plot) based on the data

Skill 2.1: Identify the data needed to answer a question

Skill 2.1 Concepts

Skill 2.1 Exercise 1

Previously, we learned that we could visualize one-column data in at least two ways – using a bar/column chart or histogram. How we visualized that data depended on whether the data was quantitative or qualitative.

To answer the questions in this exercise required at least two pieces of information. For example, the time of day and how happy people are. To answer such questions requires that we visualize data as two-columns. Before we can begin visualizing this data, we must first understand the type of data.

Skill 2.2: Interpret a crosstab chart

Skill 2.2 Concepts

A crosstab chart is useful for visualizing categorical data. To understand how a crosstab chart works, lets revisit our dog data from before,

A	B	C	D	E	F
id	Name	Breed Group	Bred For	Minimum Life Span	Maximum Life S
1	Affenpinscher	Toy	Small rodent	10	12
2	Afghan Hound	Hound	Coursing and	10	13
3	Airedale Terrier	Terrier	Badger, otter	10	13
4	Akbash Dog	Working	Sheep guardi	10	12
5	Akita	Working	Hunting bears	10	14
6	Alapaha Blue Bl	Mixed	Guarding	12	13
7	Alaskan Husky	Mixed	Sled pulling	10	13
8	Alaskan Malamu	Working	Hauling heavy	12	15
9	American Eskim	Non-Sporting	Circus perforr	12	15
10	American Foxho	Hound	Fox hunting, s	8	15
11	American Pit Bu	Terrier	Fighting	10	15
12	American Water	Sporting	Bird flushing a	10	12
13	Anatolian Sheph	Working	Livestock her	11	13
14	Australian Kelpie	Herding	Farm dog, Ca	10	13

Let's suppose we wanted to ask, "What Breed Group lives the longest?" or "How many herding breeds live at least 12 years?".

Below is a crosstab chart that enables us to answer these questions,

© Pluska

A	B	C	D	E	F	G	H	I	J	K	L	M
COUNTA of Bre	Maximum Life Span											
Breed Group	8	10	11	12	13	14	15	16	17	18	20	Grand Total
Herding				4	2	3	2	1				12
Hound	1	2		4	3	4	3	1				18
Mixed					2							2
Non-Sporting					2	1	5	1				9
Sporting			1	5	1	6	5	1				19
Terrier					1	2	5	1			1	10
Toy				1		4	3	1	1	2		12
Working	1	6	1	8	2	3	2					23
Grand Total	**2**	**8**	**2**	**22**	**13**	**23**	**25**	**6**	**1**	**2**	**1**	**105**

According to the chart, there are 12 herding breeds that live at least 12 years

Toy breeds appear to live longer than other breeds

According to the chart, one breed of terrier lives up to 20 years.

We can see from above that a crosstab chart records the number (or frequency) of values that have a specific characteristic. The above chart indicates the frequency that a certain breed lives a specific number of years.

Crosstab charts are useful for,

- Finding the most/least common value with a specific characteristic
- Finding patterns across two columns
- Exploring two columns when one or both column stores qualitative data

Crosstab charts are not useful,

- If either column has too many unique values

Let's consider another data set: *100 influential African Americans*

© Pluska

	A	B	C	D	E	F	G	H	I
1	id	Rank	Name	Profession	Sector	Age	Influence	Reach	Twitter Followers
2	1	1	Stacey Abrams	Politician	Politics	45	250.5	15.31	548861
3	2	2	Nipsey Hussle	Rapper, Entrepr	Entertainment	33	247.6	16.56	929681
4	3	3	Lizzo	Singer, Flutist	Entertainment	31	243.6	13.12	1080757
5	4	4	Steven Canals	Co-creator, Exec	Entertainment	38	234.8	10.08	22921
6	5	5	Oronike Odeleye	Activist	Community	39	231.4	9.79	3697
7	6	6	Colin Kaepernicl	Activist	Community	31	224.8	16.85	2169649
8	7	7	Ilhan Omar	US Representati	Politics	37	219.4	16.05	473422
9	8	8	Rodney Robinsc	Educator	Community	40	218.5	8.33	5815
10	9	9	Janet Mock	Writer, Producer	Entertainment	36	218.4	13.56	189795
11	10	10	Karen Attiah	Journalist	Media	33	213.3	11.68	143419
12	11	11	Alysia Montaño	Track Champion	Sports	33	207.3	11.61	17408
13	12	12	Crystal Dunn	Soccer Star	Sports	27	204.4	13.93	128430
14	13	13	Jeremy O Harris	Playwright	Arts	30	203.7	10.65	10187
15	14	14	LeBron James	NBA Player, Ent	Sports	34	203.4	19.34	43695389
16	15	15	Ibram X Kendi	Author, Professc	Community	37	203.1	12.36	66489
17	16	16	Serena Williams	Tennis Athlete, E	Sports	38	202.5	18.09	10943793
18	17	17	Tomi Adeymei	Author	Arts	26	202.3	12.27	62181
19	18	18	Beyoncé	Entertainer, Entr	Entertainment	38	200	18.72	15360340
20	19	19	Jackie Aina	Beauty Expert, A	Community	32	199.8	14.06	0
21	20	20	Janelle Monáe	Singer, Actor, Ac	Entertainment	33	199	16.5	1192873
22	21	21	Nikole Hannah-J	Journalist	Media	43	196.2	9.4	177442
23	22	22	Stephen Curry	NBA Player, Ent	Sports	31	193.4	18.56	13716022
24	23	23	Blair Imani	Activist	Community	25	191.4	11.57	72614

Suppose we wanted as the question, “What is the relationship between age and twitter followers?”. Below is a portion of a Crosstab chart that attempts to answer this question,

COUNTA of Age	*Twitter Followers*							
Age	0	106	135	523	539	565	687	717
0								
25								
26								
27	1							
28					1			
29						1		
30								
31	1							
32	1							
33	1							
34	1	1						
35								1
36	1			1				
37								
38	1							
39							1	
40	1							
41			1					
42								
43								
44	1							

The crosstab chart above is very large. In fact, the chart is 89 columns long! Furthermore, the majority of the entries are “1’s”, which doesn’t convey very meaningful information.

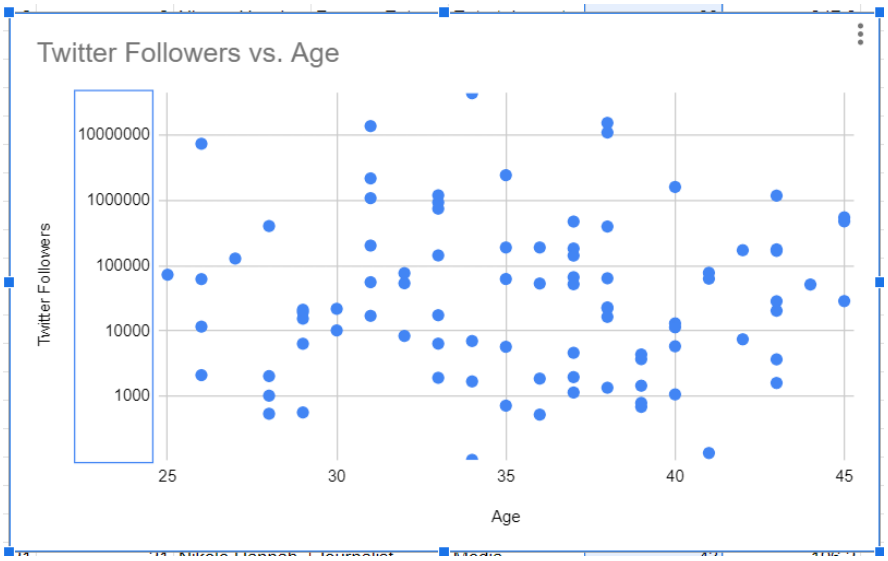
Skill 2.2 Exercise 1

© Pluska

Skill 2.3: Interpret a Scatter Plot

Skill 2.3 Concepts

In a previous example, we looked at a crosstab chart created from many unique values. This visualization did not convey very meaningful information – or was difficult to read. Let's revisit this data and visualize it in a different way,

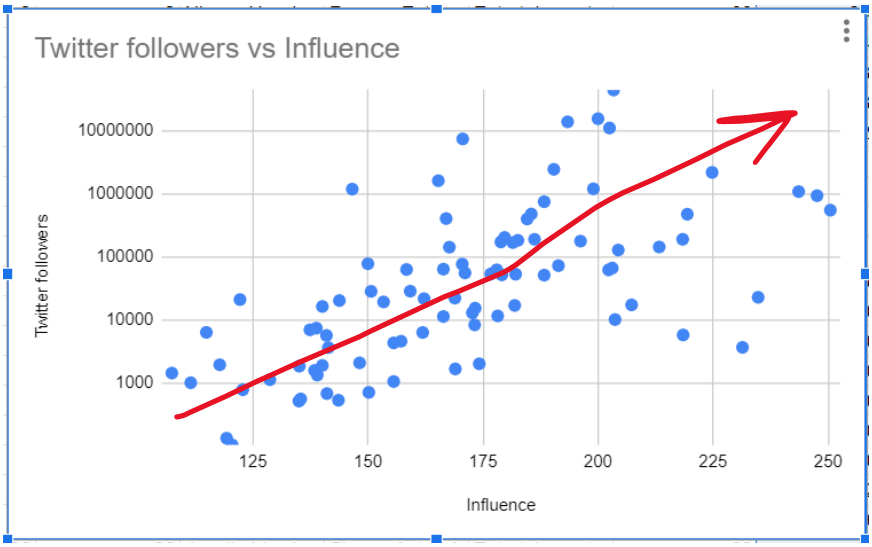


There is *no correlation* between Age and Twitter followers

The visualization above is called a scatter plot. Each dot on the graph represents a different person.

One benefit of visualizing this data as a scatter plot, versus a crosstab is that we can easily see all the data at once. And it can be easily deduced that there is no relationship between age and Twitter followers.

Let's consider another graph from the same data set which compares the influence rating to Twitter followers.

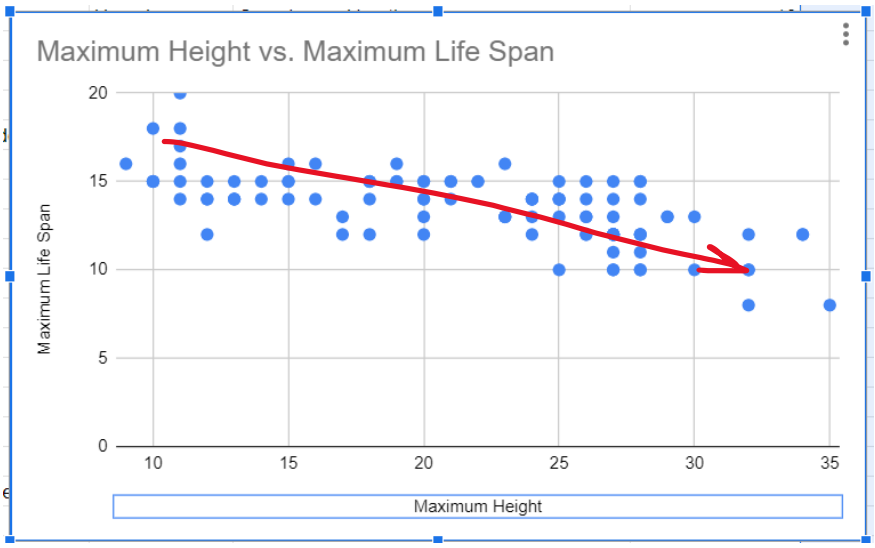


As the influence rating increases, the number of Twitter followers increases. This is a *positive correlation*.

© Pluska

According to the graph above, the number of Twitter followers generally increases as the Influence rating increases. Such a relationship is referred to as a positive correlation.

Finally, check out the scatter plot below which compares maximum dog height to maximum dog life span,



As the maximum height of dogs increases, the maximum life span of dogs decreases. This is a *negative correlation*.

Just like crosstab charts, scatter plots show relationships between two columns of data.

Scatter plots are useful for,

- Seeing patterns and trends between two values
- Numeric data with lots of different values

Scatter plots are not useful:

- If either column has lots of repeated values

Skill 2.3 Exercise 1

Skill 2.4: Select that best two-column visualization (crosstab chart or scatter plot) based on the data

Skill 2.4 Concepts

We have explored two ways to visualize two-column data. Below is a summary,

Crosstab Chart	Scatter plot
• Displays the number (or frequency) of values that have a specific characteristic • Useful for descriptive data (sophomore, junior, senior) • Useful for nominal scales (a 1-5 star rating scale for example)	• Displays the relationship between two sets of values • Useful for large data sets • Useful for numeric data sets that have lots of different values

Skill 2.4 Exercise 1

© Pluska