|  |
| --- |
| **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**](https://hpluska.github.io/APCompSciPrinciples/ticketOutTheDoor/set0/Set0TicketOutTheDoorAPCompSciPrinciples.pdf) **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 screenshot of a computer

Description automatically generated

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,

A screenshot of a computer

Description automatically generated



Toy breeds appear to live longer than other breeds

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

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*

A screenshot of a computer

Description automatically generated

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,

A screenshot of a computer

Description automatically generated

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**](https://hpluska.github.io/APCompSciPrinciples/ticketOutTheDoor/set0/Set0TicketOutTheDoorAPCompSciPrinciples.pdf) **1**

|  |
| --- |
| **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,

A screen shot of a graph

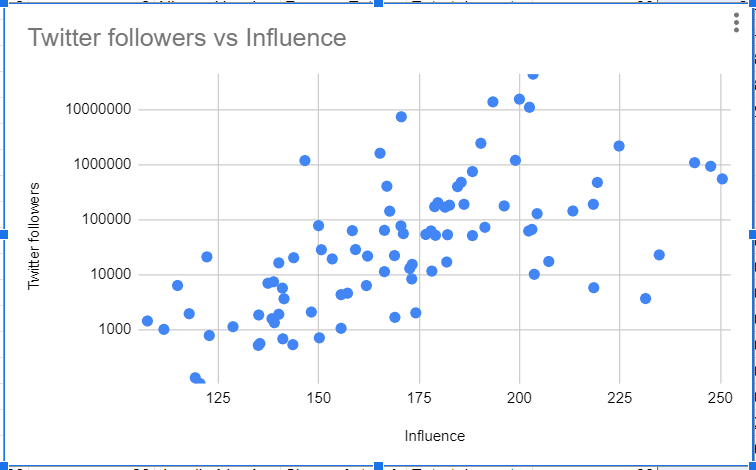
Description automatically generated

There is *no correlation* between Age and Twitter followers

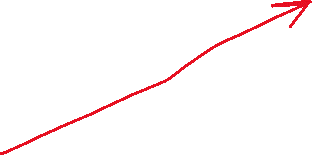
The visualization above is called a scatter plot. Each dot on the graph represents a different person.

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*.



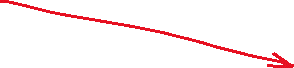
According to the graph above, the number of Twitter followers generally increases as the Influence rating increases. Such a relationship is referred to as a positive correlation.

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

A graph with blue dots

Description automatically generated

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:

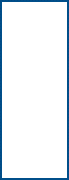
* If either column has lots of repeated values

[**Skill 2.3 Exercise**](https://hpluska.github.io/APCompSciPrinciples/ticketOutTheDoor/set0/Set0TicketOutTheDoorAPCompSciPrinciples.pdf) **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**](https://hpluska.github.io/APCompSciPrinciples/ticketOutTheDoor/set0/Set0TicketOutTheDoorAPCompSciPrinciples.pdf) **1**