|  |
| --- |
| **Set 1. Exploring One-Column Data** |

|  |
| --- |
| **Skill 1.1: Discuss the importance of visualizing data**  **Skill 1.2: Review the data analysis process**  **Skill 1.3: Differentiate between quantitative and qualitative**  **Skill 1.4: Visualize one column data**  **Skill 1.5: Clean dirty data** |

|  |
| --- |
| **Skill 1.1: Discuss the importance of visualizing data** |

**Skill 1.1 Concepts**

[**Skill 1.1 Exercise**](https://hpluska.github.io/APCompSciPrinciples/ticketOutTheDoor/set0/Set0TicketOutTheDoorAPCompSciPrinciples.pdf) **1**

The previous exercise helped us understand the way data visualizations can serve to,

* Answer questions
* Look at lots of data at once
* See patterns that are "invisible" if you just look at a data set

Today we're going to learn how to make two different types of visualizations

|  |
| --- |
| **Skill 1.2: Review the data analysis process** |

**Skill 1.2 Concepts**

The data analysis process starts with identifying a problem that can be solved with data. Once you’ve identified this problem, you can collect, clean, process, and analyze data. The purpose of analyzing this data is to identify trends, patterns, and meaningful insights, with the ultimate goal of solving the original problem. This is summarized below,

A diagram of a diagram

Description automatically generated

The video below discusses how the data analysis process is used to help oceanographers study artic sea ice,

|  |
| --- |
|  |
| <https://www.youtube.com/watch?v=uzESOw7tmzw> |

[**Skill 1.2 Exercise**](https://hpluska.github.io/APCompSciPrinciples/ticketOutTheDoor/set0/Set0TicketOutTheDoorAPCompSciPrinciples.pdf) **1**

|  |
| --- |
| **Skill 1.3: Differentiate between quantitative and qualitative** |

**Skill 1.3 Concepts**

At its simplest, data can be broken down into two different categories: *quantitative data*and *qualitative data*. But what’s the difference between the two? And how can we visualize them?

**Quantitative data** refers to any information that can be quantified, counted or measured, and given a numerical value.

**Qualitative data** is descriptive in nature, expressed in terms of language rather than numerical values.

Examples of qualitative and quantitative data are shown below. And as you can see, both provide immense value for any data collection and are key to truly finding answers and patterns.

A close-up of a paper

Description automatically generated

[**Skill 1.3 Exercise**](https://hpluska.github.io/APCompSciPrinciples/ticketOutTheDoor/set0/Set0TicketOutTheDoorAPCompSciPrinciples.pdf) **1**

|  |
| --- |
| **Skill 1.4: Visualize one-column data** |

**Skill 1.4 Concepts**

Determining the type of data we want to visualize is the first step in the data visualization process. In a previous exercise you determine whether the data stored in a given column is qualitative or quantitative. If the data we want to visualize is qualitative, a bar or column chart should be used.

**Bar Charts** count how many times each value in the column appears and make a bar or column at that length.

Let’s revisit the dog data, <https://docs.google.com/spreadsheets/d/1dy2TrqRqXNcq-0k4ciLcATcPINv8u1eNPWhbJRuYYjU/edit?usp=sharing>

Below is a snippet from this data table,

Most of the values for each dog in these columns are different. A bar graph would not display very useful information.



Many of the values in this column are repeated. A bar graph can be used to count the number of times each value appears.

Notice in the table above that *Name*, *Breed Group*, and *Bred For* columns all store qualitative data. The most of the values in the *Name* and *Bred For* columns are different for every dog and displaying them as a bar graph would not communicate very useful information. The *Breed Group* column on the hand, has values that are repeated. We can use a bar graph to visualize how many times each value is repeated. This is illustrated below,

|  |  |
| --- | --- |
|  |  |

Notice what happens when we try to plot the Minimum Lifespan as a bar graph. A different line is created for every dog! Visualizing this data as bar chart is difficult to read.

A blue line graph with black text

Description automatically generated

The data above, because it is quantitative, is better visualized as a histogram.

**Histogram Charts** aresimilar to a bar chart, but first all numbers in a range or "bucket" are grouped together.  For example, the chart below has a bucket size of 2 so dogs with a minimum lifespan of 10 or 11 would all be placed in the same bucket.

A graph of a number of life span

Description automatically generated

Histograms can only be created with numeric data but can be useful when a normal bar chart may be difficult to read.

Below is a summary of the types of information we can get from bar and histogram charts.

|  |  |
| --- | --- |
| Information we can get from bar and histography charts. | |
| **Bar Charts** | **Histogram Charts** |
| * What value(s) are most common in this column? * What value(s) are least common in this column? * What is the unique list of values in this column? | * What range of value(s) are most common in this column? * What range value(s) are least common in this column? * What ranges of values do or do not appear in this column? |

[**Skill 1.4 Exercise**](https://hpluska.github.io/APCompSciPrinciples/ticketOutTheDoor/set0/Set0TicketOutTheDoorAPCompSciPrinciples.pdf) **1**

|  |
| --- |
| **Skill 1.5: Clean dirty data** |

**Skill 1.5 Concepts**

We've started to explore how to use charts to visualize one column data, but there are often times challenges with doing this.  One challenge that often occurs with survey data is “dirty” data. The data set below for example represents data collected via a student survey. Notice that there are several instances where students typed a word (or String) instead of an integer to represent a number; there are also instances where students abbreviated their favorite subject (CS instead of Computer Science for example).

A screenshot of a computer

Description automatically generated

Before we can adequately analyze the data, we must first clean it – the third step in the data analysis process.  Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

[**Skill 1.5 Exercise**](https://hpluska.github.io/APCompSciPrinciples/ticketOutTheDoor/set0/Set0TicketOutTheDoorAPCompSciPrinciples.pdf) **1**