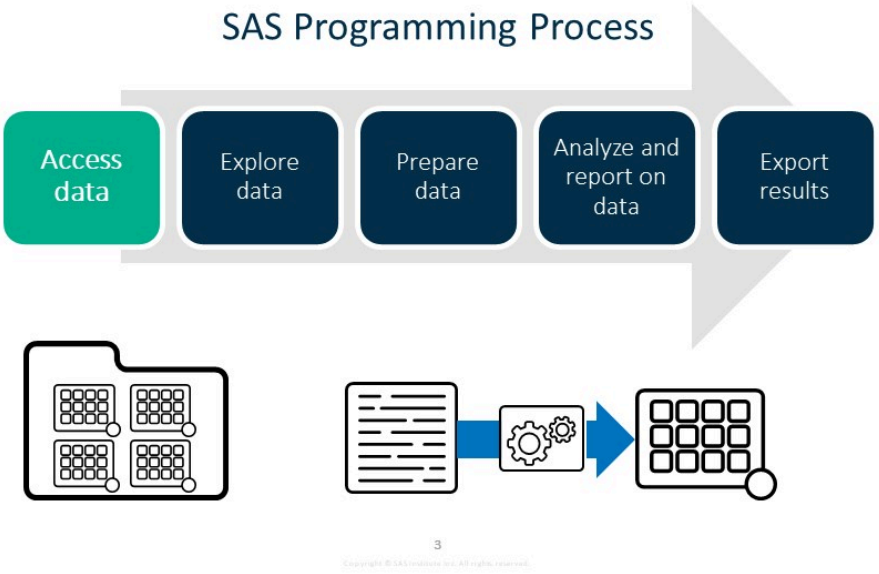
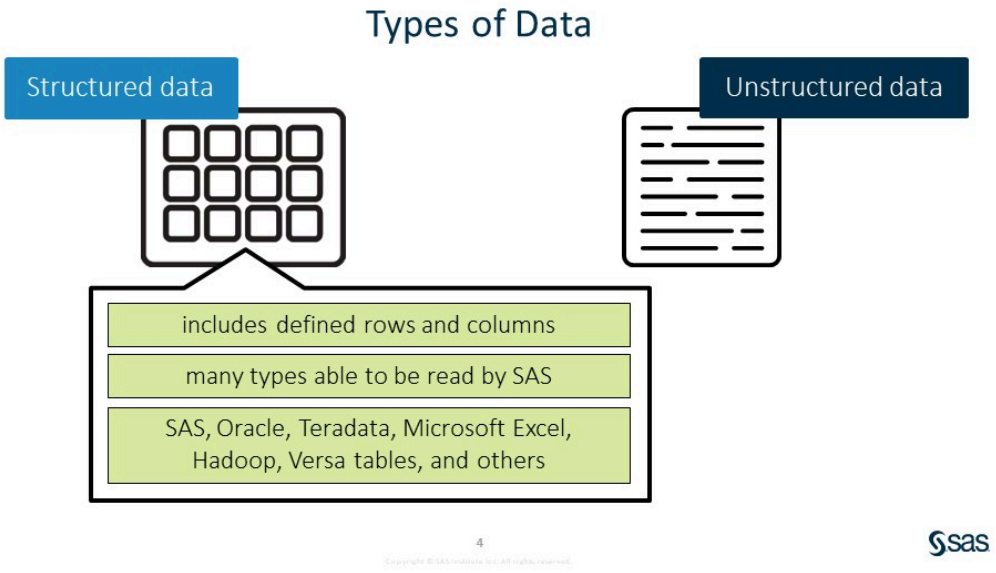
B2.1 - Understanding SAS Data



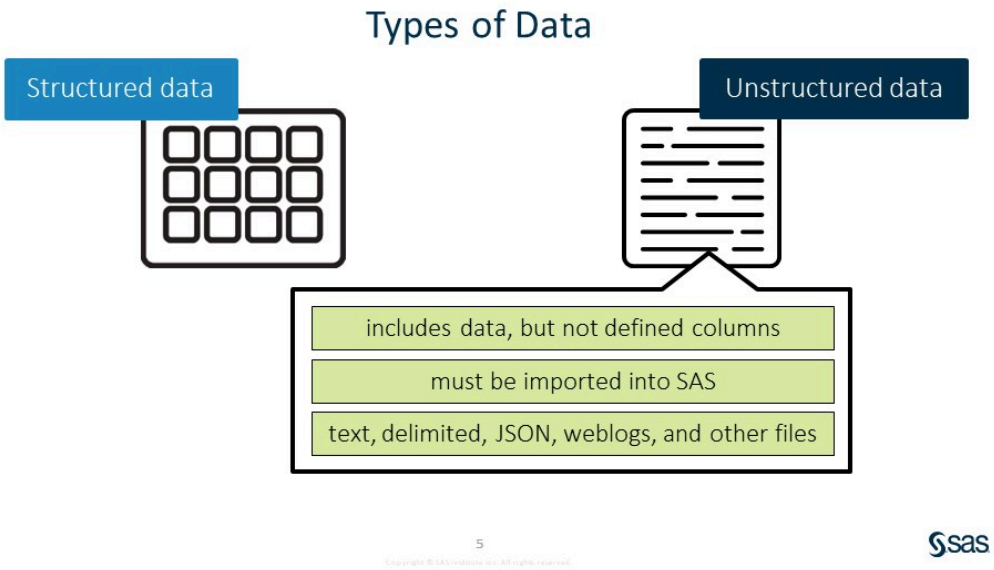
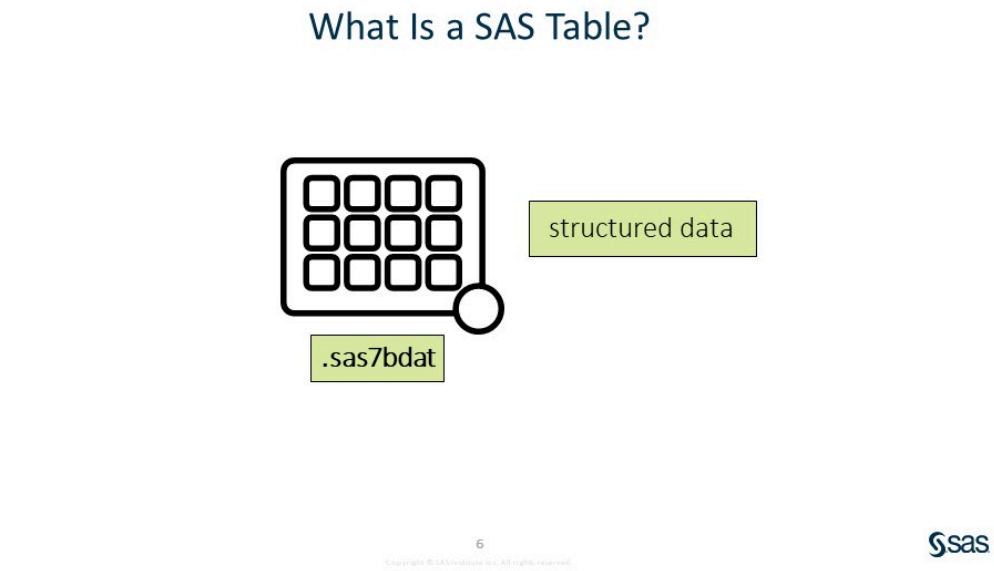
Accessing data is the first step in the SAS programming process. There are many types of data files and SAS makes it easy to access these different types of data and use them reporting and analysis.

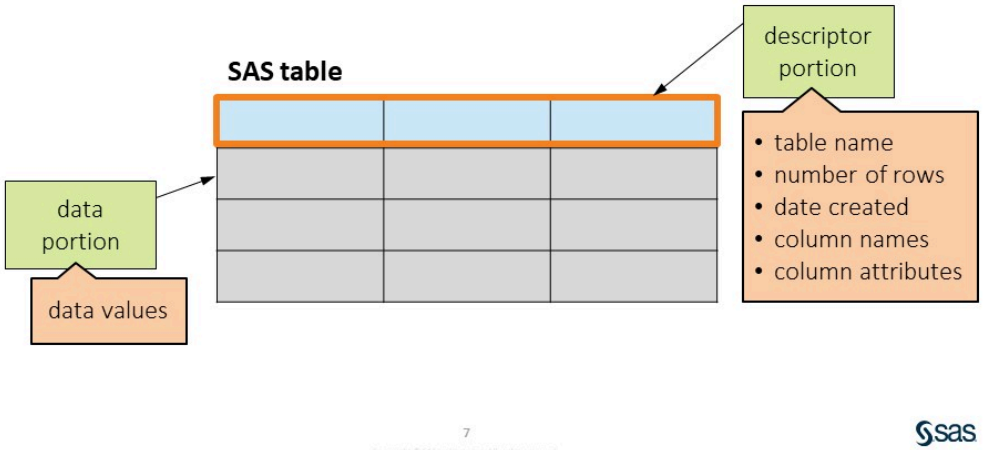


I like to think of data as falling into one of two broad categories: structured and unstructured.

A structured table has defined columns with certain attributes that help SAS or other applications know how to read and display the values. Structured data could include SAS tables, Microsoft Access tables, other DBMS tables, such as Oracle or Teradata, or even cloud-based data such as Versa or Hadoop.

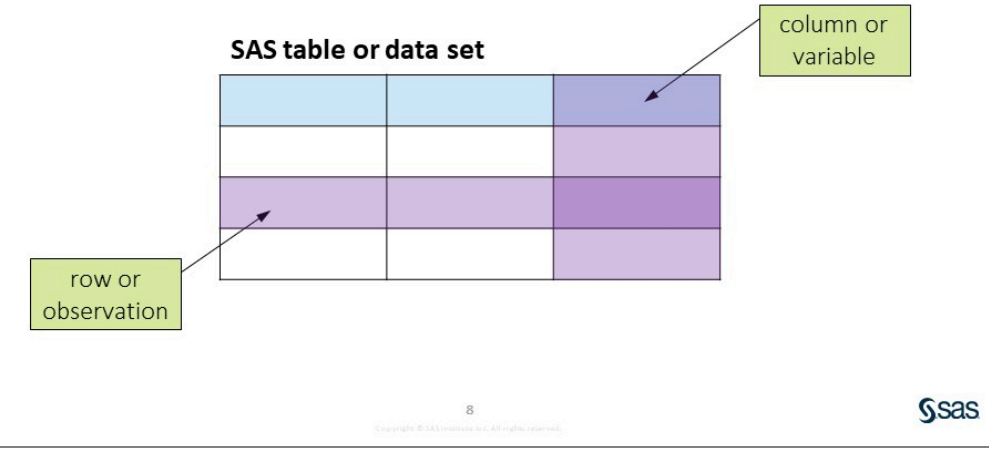
SAS has engines to enable it to understand

	and read various types of structured data.
	By contrast, unstructured data does not have defined columns. The data might be in a format that appears to be in columns, such as tab-delimited data, but to your computer it's just strings of text. Examples of unstructured data are text files or comma-delimited files, JSON files, and weblogs. There is important information in these files, but it's not in a defined structure. You must import unstructured data into SAS before you can analyze or report on it. SAS makes importing data easy, too.
	A SAS table is a structured data file that has defined columns and rows. SAS tables have the file extension .sas7bdat.

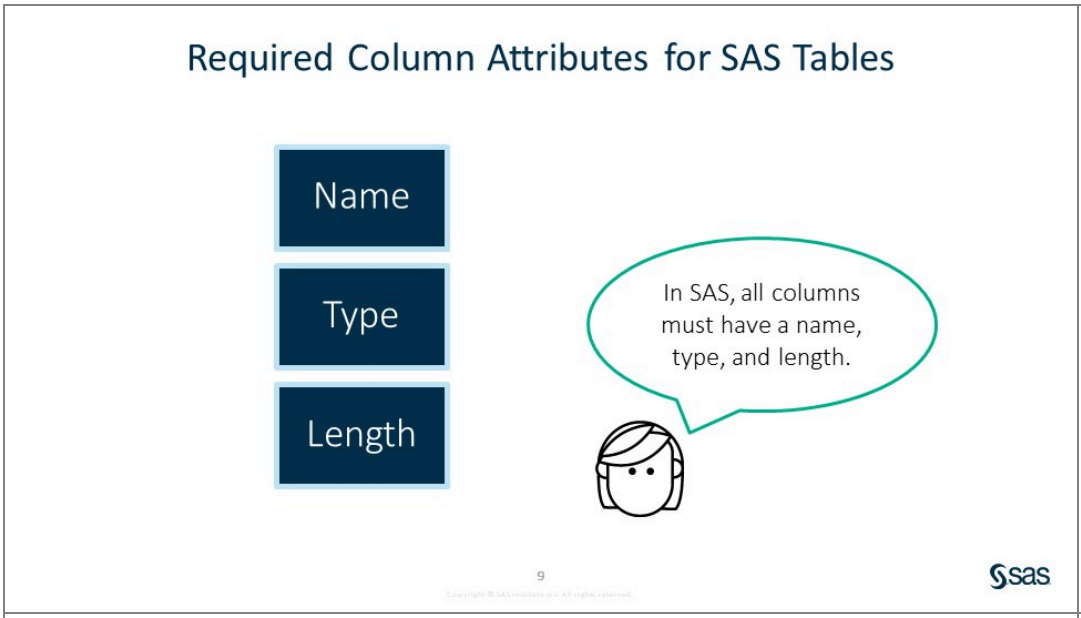


There are two parts to a SAS table: **a descriptor portion** and a **data portion**. The descriptor portion contains information about the attributes of the table, or metadata. The metadata includes general properties such as the table name, the number of rows, and the date and time that the table was created. The descriptor portion also includes the column definitions.

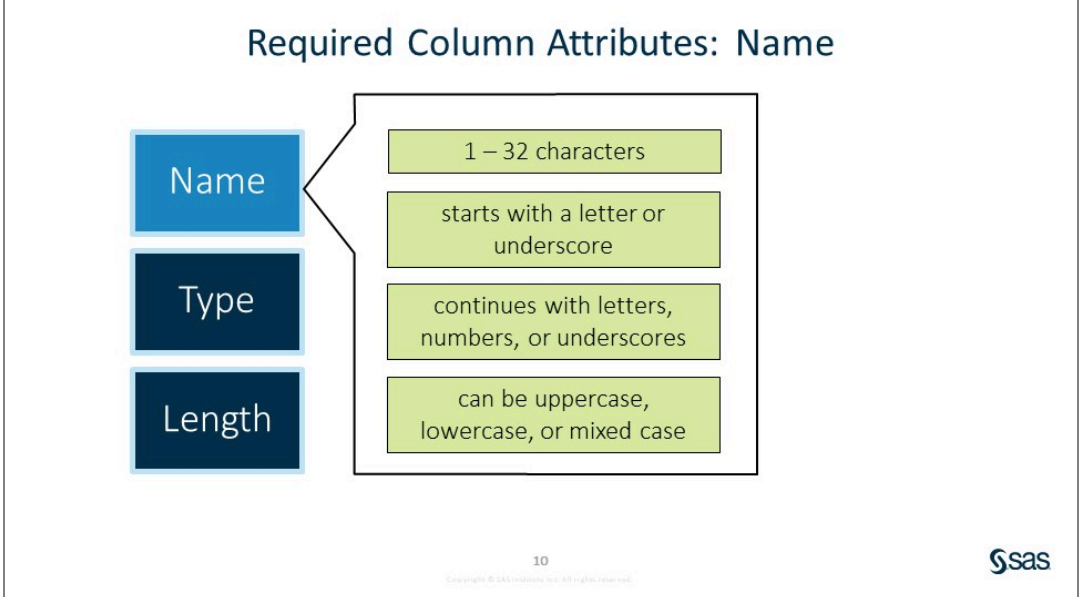
The data portion of a SAS table contains the data values, stored in columns.



SAS has its own terminology, and you'll see these terms used in results and the log. A SAS table is also called a SAS data set, a column is also called a variable, and a row is also called an observation.



So what does it mean for a column to be *defined*? In SAS, a column must have three attributes: a name, a type, and a length. Let's look at each of these in more depth.



Column Names

Let's start with column names.

- SAS column names can be 1 to 32 characters long.
- The name must start with a letter or underscore and can continue with any combination of numbers,letters, or underscores.
- SAS column names can be uppercase, lowercase, or mixed case.
- Column names are stored in the case that you use when you create the column, and that's the way the column name appears in reports.

	• After a column has been created, you can type it in any case in your code without affecting the way that it is stored. These same naming conventions should be followed for SAS table names. Depending on the environment you use to submit your SAS code, SAS might allow for spaces and special symbols other than underscores in column and table names. If you use data sources other than SAS that have flexible column-name rules, SAS can make allowances for that. However, for simplicity and consistency, we recommend following these SAS naming conventions.

Note: The following activity is for you to answer questions and for you to get your response. It also serves as a checking point for if you are understanding the material. The following activity is not graded.

Answer the following question:

2.01 Question

Which column names are valid? (Select all that apply) 1 point

- [ ] month6
- [ ] 6month
- [ ] month#6
- [ ] month 6 (contains a space between month and 6)
- [ ] month_6
- [ ] Month6

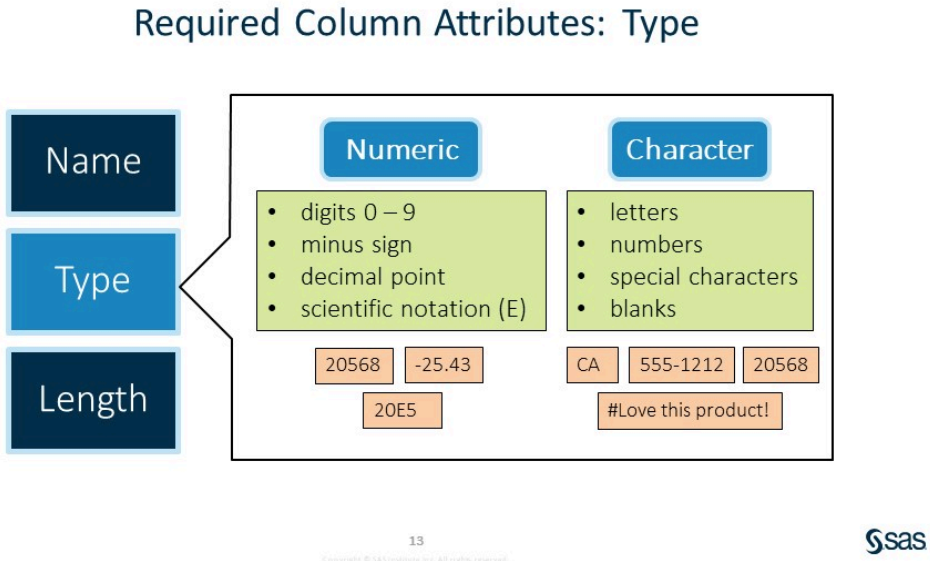
Submit Page 1 of 1 Clear form

Column Type

The next required column attribute is the type. There are two types of columns: **character** and **numeric**.

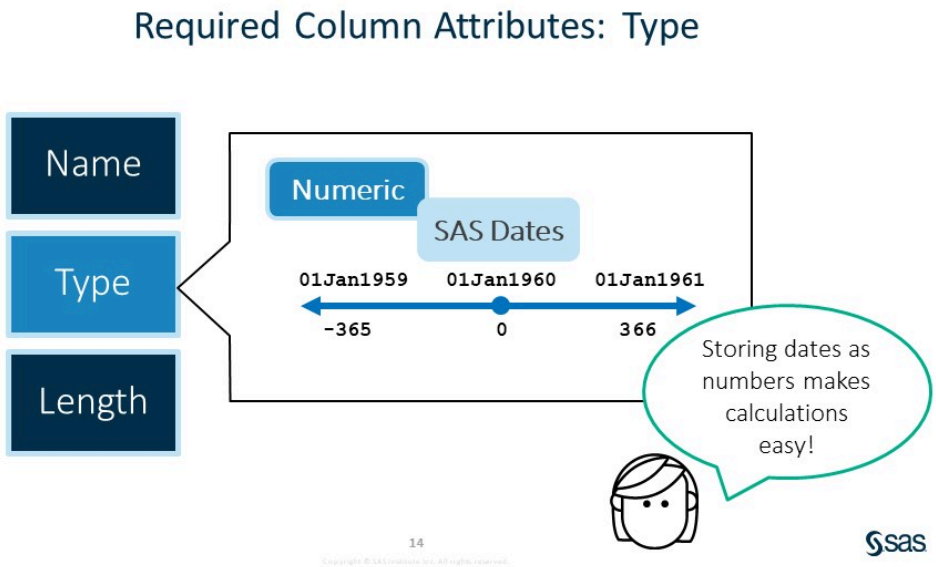
Numeric columns can store only numeric values, which can include the digits 0 through 9, a minus sign, a single decimal point, and E for scientific notation.

Character columns can store letters, numbers, special characters, and blanks.



Date Values

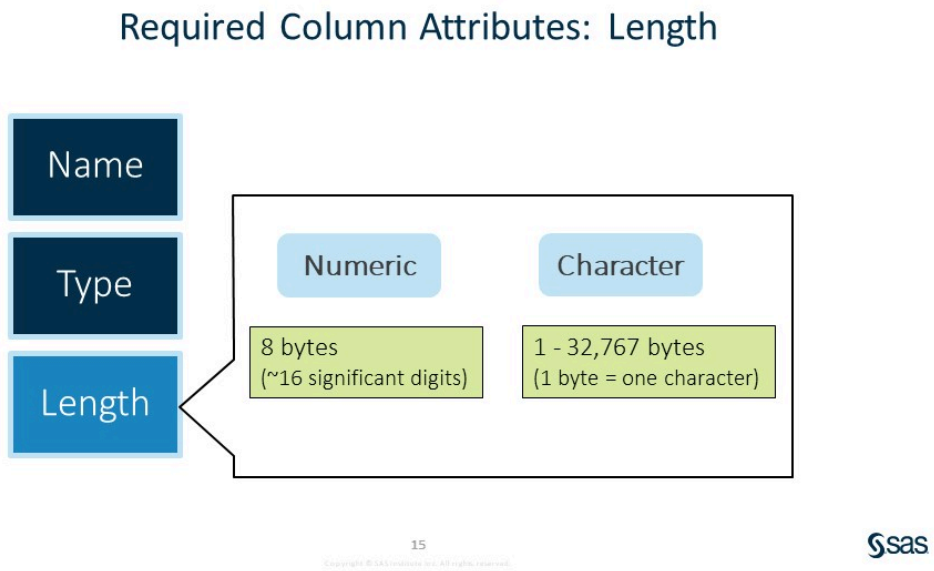
Let's talk about a particular kind of numeric value: SAS dates. SAS stores date values as the number of days between January 1, 1960, and a specific date. Dates before January 1, 1960, are stored as negative values. SAS keeps track of leap years, so you can see that 1960 was a leap year and has 366 days. This way of storing date values makes



calculations and sorting possible. Don't worry – there are numerous ways to display these values so they look like dates you can understand!	

Column Length

Finally, lets talk about the last required attribute: the length. The column length is the number of bytes allocated to store column values. The length is related to the column type. Numeric columns, by default, are always stored as 8 bytes, which is enough for about 16 significant digits. Character columns can be any length between 1 and 32,767 bytes, and a byte stores one character. A column such as **Country Code** that is always a two-letter code might be assigned a length of 2. A column such as **Country Name** that could have a varying number of characters must have a length at least as long as the longest country name.



2.02 Activity

Use the **storm_summary.sas7bdat** SAS table into your **data** folder.

1. Navigate to the location of your course files and open the **data** folder.
   SAS Studio: Expand **Files and Folders.**
2. Double-click the **storm_summary.sas7bdat** SAS table to view it.

How are missing character and numeric values represented in the data?

Click here for Solution.

Note: The following activity is for you to answer questions and for you to get your response. It also serves as a checking point for if you are understanding the material. The following activity is not graded.

Answer the following question:

2.03 Question

* Indicates required question

Email *

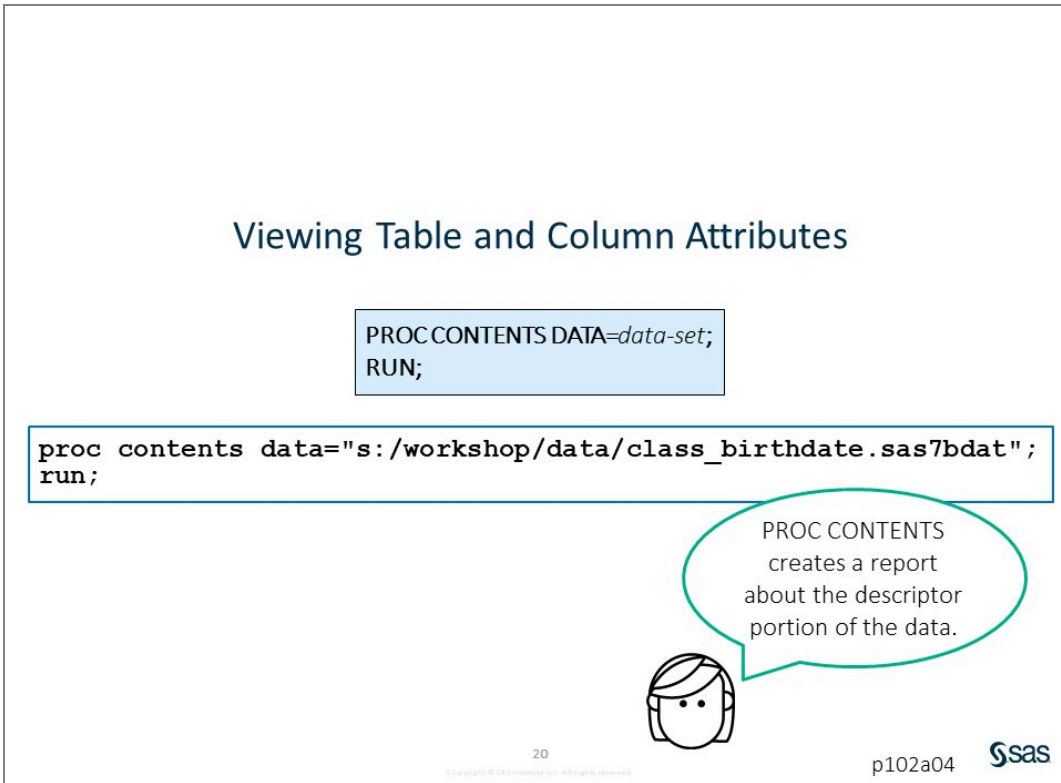
Your email address

Using the storm_summary data. Choose Table Properties (click on 3 vertical dots) then select Column Properties column attributes. Examine the length of the Basin column. Could "East Pacific" be properly stored as a data value in the Basin column? 1 point

- Yes
- No

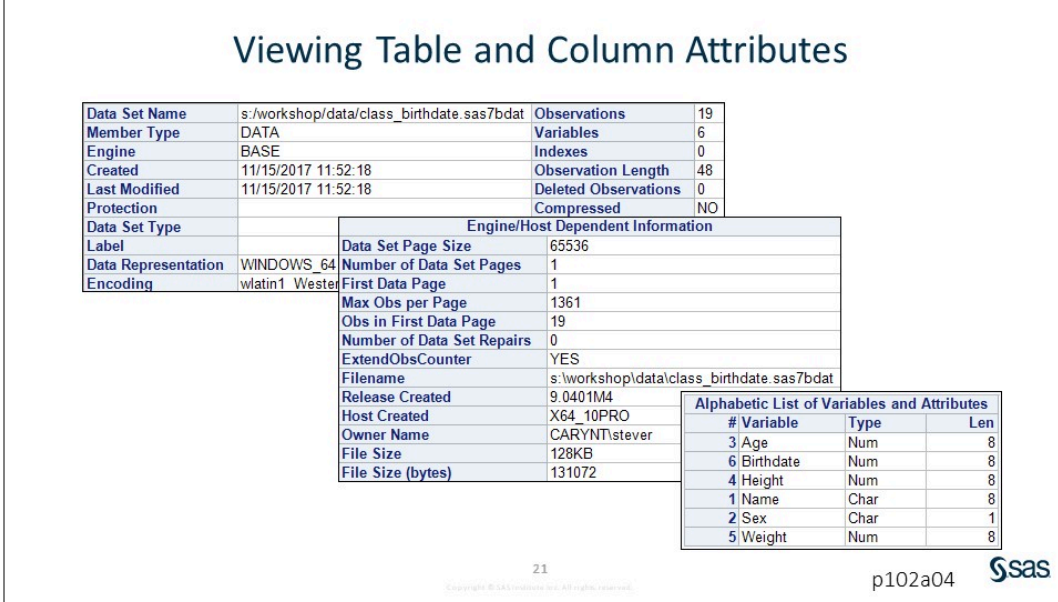
Send me a copy of my response.

Submit Clear form



PROC Contents

You've used your software to view the **storm_summary** table and its properties, but another way to view table attributes is to write a little program using a PROC CONTENTS step. The syntax is simple: just PROC CONTENTS and DATA=, followed by the location and table name that you want to examine. The RUN statement ends the step.



PROC Contents

The output of PROC CONTENTS is a listing of the information in the descriptor portion of the table. You can also think of this as the metadata or properties of the table. The first two sections of the report give general information about the table, including where the table is stored, when it was created and modified, and the number of rows and columns. The variable list

is where I'll focus my attention. The column names are listed in this table along with their type and length.

In the **class_birthdate** table, we have the numeric columns **Age**, **Birthdate, Height**, and **Weight** that all have a length of 8. We also have two character columns. **Name** is 8 bytes, meaning we can store names with up to eight characters. **Sex** has a length of 1, which is appropriate because it contains only one-letter codes.

2.04 Activity

Use **p102a04.sas** in your **activities** folder and perform the following task:

1. Write a PROC CONTENTS step to generate a report of the **storm_summary.sas7bdat** table properties. Highlight the step and run only the selected code. (Note: Use the full location of the data file)
2. How many observations are in the table?
3. How is the table sorted?

```
PROC CONTENTS DATA=data-set;
RUN;
```

How are missing character and numeric values represented in the data?

Click here for Solution.