# Inspecting a Dataset

**Activity Overview**

Welcome to this lab on accessing data and importing data! As a data analyst, you'll need to use data to answer questions and solve problems. When you analyze data and draw conclusions, you are generating insights. It is not the goal to do tons of analyses and create reports that contain lots of information. The goal is to effectively use data and analytics to your business' advantage. To do this you will need to learn how to "ask the right questions" to provide the insights that are most important to the goals of your business.

Data sets the stage for the discovery of insights that can then influence decisions and drive change. There are two basic types of data sources used by data analysts: internal and external. Depending on the project, you might need to use both. Whichever source you choose, you must know how to access the data you need.

Using spreadsheets is one of the first things aspiring data analysts must learn. Spreadsheets are important because they combine data storage, presentation, and tools for analysis. Once you've identified the data for your project, you'll need to know how to import and analyze it on a spreadsheet.

Let's look at a real-world situation. You are a Junior Data Analyst for a small ice cream company. Management is interested in a sales analysis that will help them understand the company's ice cream sales. The sales analysis is mining your data to evaluate the performance of your sales against your company's goals. It provides insights about the top-performing and underperforming products and services, the problems in selling and market opportunities, sales forecasting, and sales activities that generate revenue.

The company has been collecting some information about its sales, but not a lot. There are three data sets available from an **internal data source**. You are given this data and are asked to provide insight into the sales. The data sets you will use contain only 2019 information and are located here [SalesByTemp](https://drive.google.com/u/0/uc?id=1_syLRjBs85d0IfaUxb_KeSa7aPPgw8Bi&export=download), [SalesByDay](https://drive.google.com/u/0/uc?id=1bmlCMz7sCi5eL9U5hMFoSl5Ya7hrHZtS&export=download), [SalesByFlavor](https://drive.google.com/u/0/uc?id=1RHzmPkLSPFujuq5WcZWElaLYRNMl0odA&export=download). Click the links to **download them to your computer**.

What you'll do

For this lab, we will use Google Sheets. To do this, you will need to set up a Google Workspace account. There are other options in the data analyst's toolbox as well, such as Microsoft Excel and Libra Office 3. No matter what tool you use, the steps we'll follow will be similar. You will perform the following spreadsheet actions:

* Import data from a given file into a workbook
* Scroll to inspect spreadsheet data

**You have 60 minutes to complete this lab.**

Sign in to Google Workspace

**Note:** Make sure you followed the instructions on Coursera to open this Qwiklab in an incognito window. If you did not return to the Coursera instructions and follow them now. Remember that if you logged in to Coursera with your Google account credentials, you must open a new tab, go to Gmail, and log out of your Google account in the incognito window.

1. Click **Start Lab** at the top of the screen. You may need to wait a few minutes for the lab resources.

2. Right click **Open Workspace Admin Console** in the upper-left panel. Then, select to open it in an incognito window. Do not proceed with your **Google credentials**. Instead, return to the previous page and copy the **username**.

3. Now, in the sign-in tab, paste the **username**. Click Next. Then, repeat this process for the **password**.

4. Read and **Accept** the terms of service.

5. Start the free Google Workspace Trial.

6. You may need to verify the domain to protect your identity. Then, the **Admin console** will open. Click **Next**

**Note:** Wait a few minutes before using any Google Apps. The system usually needs 2-3 minutes to process your access, but could take as long as 5-10 minutes.

You can navigate to any Google workspace application from this page by clicking the **Google apps grid icon** in the **top right** . If you do not see the icon, widen the browser window.

Do not click **End Lab** unless you are finished with the lab or want to restart it. This clears your work and removes the project.

**Importing the data**

First, Click on the **Google apps grid icon** in the top right and select the **Sheets** app. Check again to make sure that you are not logged in to your personal Google account. You may be invited to take a tour of Google Sheets. You can choose to take the tour or close the pop-up window.

Start a new blank spreadsheet by clicking on **Blank**.

Then, you need to import your data into a spreadsheet. In Google Sheets, go to the top menu bar and click **File → Import**.

You will be prompted to upload a file. Click **Upload** and then click **Select a file from your device** and then navigate to your downloaded file. If you did not rename the file, it should be *SalesByTemp.xlsx*.

Next, in the pop-up window, make the following selection:

Click Import data. Now the *SalesByTemp* dataset has been imported!

Follow the same steps to import the data file named *SalesByDay.xlsx* but this time, on the pop-up window, you will make the following selection:

Click **Import data**. Now the *SalesByDay* dataset has been imported!

Finally, follow the same steps to import the *SalesByFlavor.xlsx* data file. Your Google Sheets workbook should now have three sheets: *temperatures, sales*, and *flavors*.

**Asking the right questions**

The job of a data analyst is to provide insights that are *most important* to the goals of the business. The process of weeding out what you don't want to measure changes with each project. Every problem is different and every business has different goals.

For this ice cream company, you probably want to know:

* What is the most popular flavor?
* What is the cost of selling as a percent of the revenue generated?
* Does temperature affect sales?
* What percentage of sales come from new customers?
* Are sales on the weekends/holidays greater than other days?

Let's make sure the data provides us with the information we need to answer these questions.

**To find out what the most popular flavor is,** we first need to define what we mean by "popular." Is the most popular flavor the one that generated the most money in sales in 2019? Or is it the flavor that has the largest number of units sold in 2019? Sometimes our measurement choices are limited by what data we have.

Click on the flavors tab to view the *flavors* spreadsheet data. You should see 3 columns and 209 rows of data. A quick way to check this is to select the row labeled 1. The status bar at the bottom should say **Count: 3**. Then select the column labeled A. The status bar should say **Count: 209**. If another statistic is shown, click the down arrow next to the number and select the **Count: 209** option from the menu.

You should also see that the column headers are: *week*, *units sold,* and *flavor*. Though this dataset did not come with a data description, you can deduce that these columns provide information about the number of units sold for each ice cream flavor, by week, in 2019.

From this dataset, you can find out what the most popular flavor is by using the *units sold column to calculate the total number of units sold during the year for each flavor.* Since the dataset does not provide the annual sales amount by flavor you won’t be able to use that as a measure. This would be useful insight, nonetheless. Because it is part of the junior analyst's job to make recommendations to improve data collection strategies, you can suggest to the stakeholders that it be collected in the future.

To find out whether temperature affects sales, you will need data that connects temperature to company sales. Click on the *temperatures* tab to view the *temperatures* spreadsheet data. You should see 2 columns and 366 rows of data. The column headers are *temperature* and *sales*. This data is most likely total company sales by temperature. But, it could also be total company sales and temperature highs recorded daily in 2019. So which is it? It is impossible to be sure without a data description or help from the person who collected the data.

You still have enough information to find out whether temperature affects sales. You just need a little more information. With more details, if you find that temperature does affect sales, you can say something like: when daily temperature highs are above 75o, average ice cream sales increase.

Next, click on the *sales* tab to view the last spreadsheet. You should see 2 columns and 366 rows of data. The column headers are *date* and *sales*. This data is most likely 2019 total daily sales. The sales are recorded for both weekends and holidays. This should make it easier to find out whether sales on the weekends/holidays are greater than on other days.

Though you were able to uncover some valuable insights regarding company ice cream sales, you were not able to answer all of your important questions. The data just didn't have all the information. Without cost information and sales related to new customers, you are not able to answer questions like:

* **What is the cost of selling as a percent of the revenue generated?**
* **What percentage of sales come from new customers?**

Not having all the data does sometimes happen when working on analytics projects. In many of these cases, you can turn to external data sources to find more information. Since all unanswered questions are specific to your ice cream company's sales, external data won't be much help. You may not have all the answers but you can make some wonderful recommendations for improving your company's future data collection efforts.

**Choosing your own dataset**

Now it's time to use **external data** and find your own dataset! You can leave ice cream behind if you’d like and instead think about a topic that interests you. Perhaps you are fascinated with business and economics, education, employment, health, or crime. Those are all great topics! There are so many options to choose from.

Now, pretend you are your stakeholders. Come up with some questions your stakeholders would want you to answer about the topic you chose. Then, find a dataset that might give you the information you need to solve problems related to your topic. Here are some datasets to help you in your search:

* <https://www.freecodecamp.org/news/https-medium-freecodecamp-org-best-free-open-data-sources-anyone-can-use-a65b514b0f2d>
* <https://cloud.google.com/bigquery/public-data>

When choosing your dataset, make sure it is not too large. Datasets larger than 20MB might be slow to import, open, and analyze in Google Sheets. Additionally, ensure your dataset is:

* Open
* Related to your topic
* Recent and updated

Next, download and import your dataset into Google Sheets.

Finally, return to that list of questions you need to answer and inspect the dataset you chose. Does your chosen dataset give you the information needed to answer your questions? Does it have relevant columns containing relevant data? If it doesn't, find another dataset or brainstorm ways you can fix this.