# 

**Analyzing and Understanding Distribution of Data**

Name \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_\_\_\_ Period: \_\_\_\_\_\_

**For each activity below:**

✔️ means to complete this task

✏️ means to write an answer here

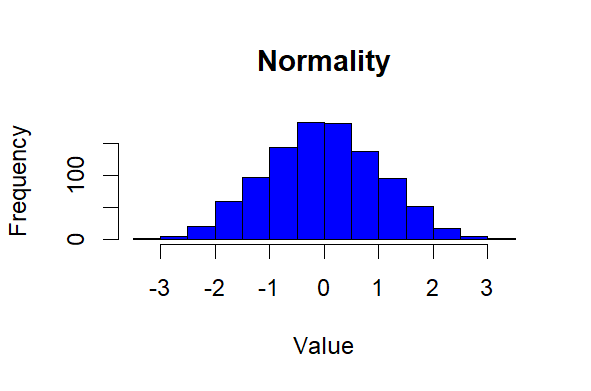
Last lesson, we explored the fascinating world of built-in datasets. Today, we are taking our data analysis skills to the next level. In this lesson, we are going to delve into the diabetes dataset that you need to upload and work with.

# **What does the Diabetes Dataset in this activity represent?**

This dataset comes from a 1997 study of 1046 subjects in central Virginia to understand obesity, diabetes, and other cardiovascular risk factors among African Americans. Dr. John Hong found that Type II Diabetes Mellitus (adult onset diabetes) is strongly linked to obesity. The waist/hip ratio might help predict the risk of diabetes and diabetes disease. Type II Diabetes Mellitus is also associated with high blood pressure, and they could both be part of a condition called "Syndrome X." The dataset includes data from 403 subjects in the study who were specifically screened for diabetes. A glycosylated **hemoglobin level above 7.0 is usually considered a positive diagnosis for diabetes**.

# **What is Normality?**

**Normality** refers to a common pattern we often find in data, where most values gather around the middle and fewer values are far from the center. Imagine a group of test scores – most students might have scores close to the mean, and fewer students would have very high or very low scores. Recognizing normality helps us make sense of data and understand how things are typically spread out.

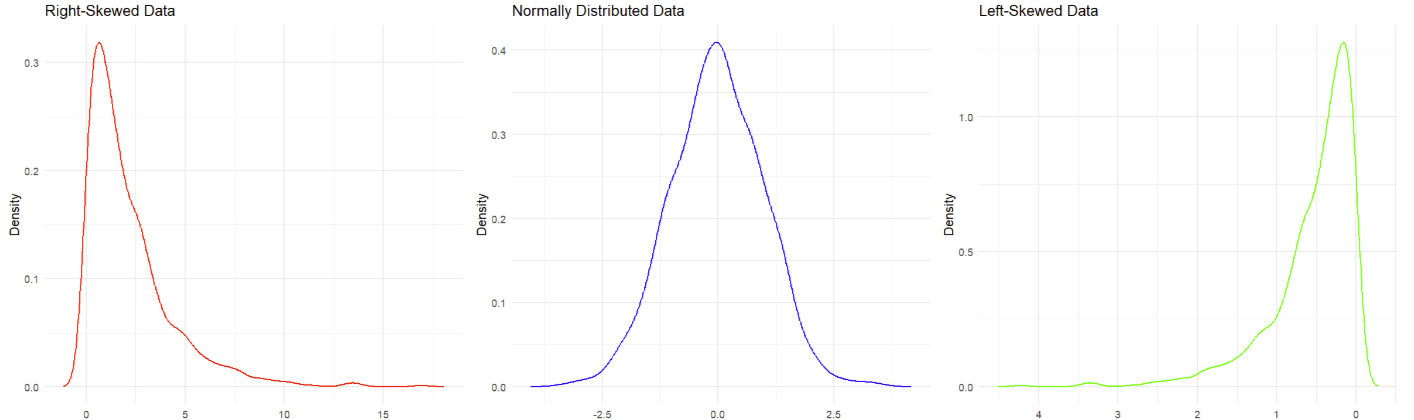


# **What is a Normal Distribution?**

A **normal distribution** is a pattern of distribution of data that is often seen in nature. It looks like a smooth, bell-shaped curve, with most data points in the middle and fewer as you go to the edges. This pattern is often found in natural things, like people's heights and test scores. Knowing about the normal distribution helps scientists make predictions and understand how things behave in real life.

**Skewness** is a pattern of distribution of data that is not symmetrical. If the data is skewed, it means the values are not evenly distributed around the average.

# 



# 

# 

An example of what the distribution of the data can look like. The left is **right-skewed**. Middle is **normally distributed**. The right is **left-skewed**.

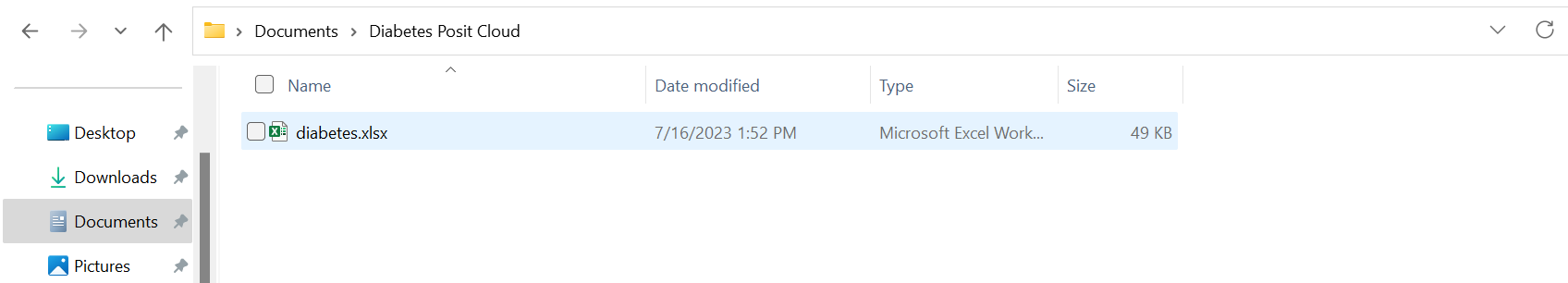
# 

# **What are you trying to figure out from the Diabetes dataset?**

From the diabetes dataset, we are trying to understand the distribution of a specific variable (ex. blood glucose levels) and see if it follows a normal distribution or is skewed. By analyzing the data's normality, we can make statistical inferences and understand how likely it is for certain values to occur. This helps us gain insights into the underlying patterns and characteristics of the data, enabling better decision-making and predictions in medical research and healthcare.

# **Activity #1: Loading the Diabetes Dataset into Posit Cloud**

1. ✔️ Let's start by creating a new folder on your computer titled “Diabetes Posit Cloud” under Documents on your computer. Save the dataset provided by your teacher to the Diabetes folder.

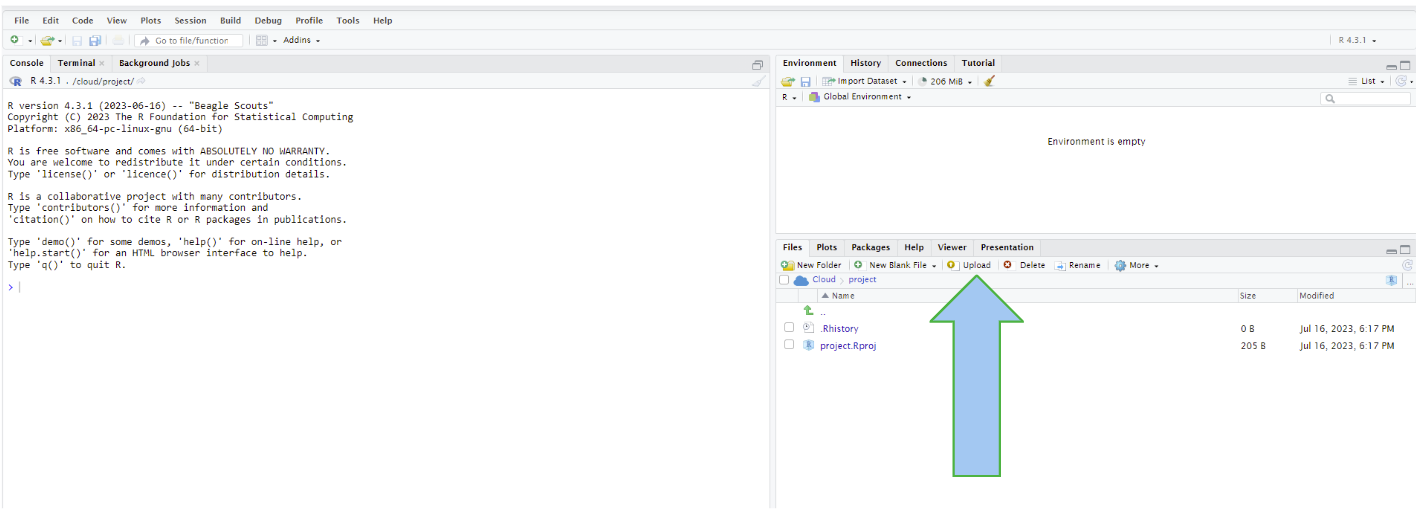


* If you are unsure if you saved the dataset in the correct location, ask your teacher for help.

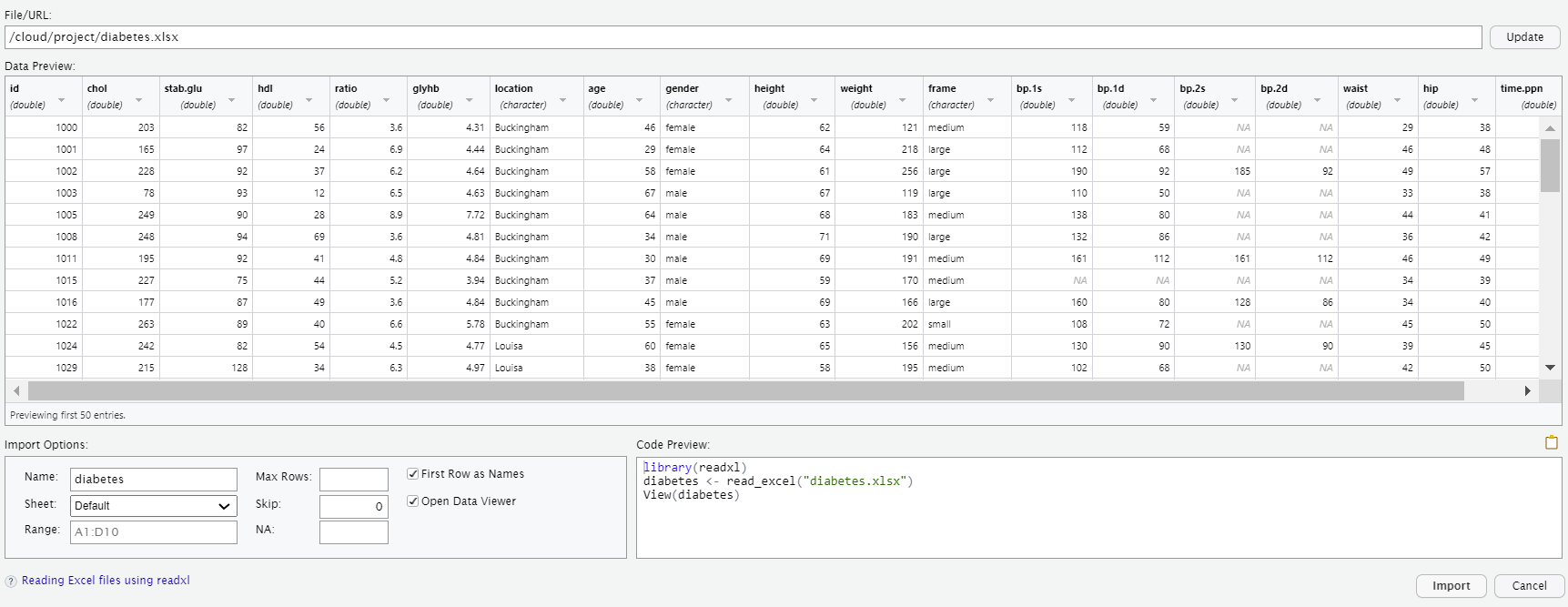
1. ✔️ Open up Posit Cloud. Create a new script by going to the top left and selecting **File** → **New File** → **R Script. Name the script** “FirstName\_LastName\_Diabetes\_Activity\_Code”. This script will contain commands so that you can repeat your analysis at any time without having to type it all in again.

* Replace “FirstName” with your actual first name and your “LastName” with your actual last name.

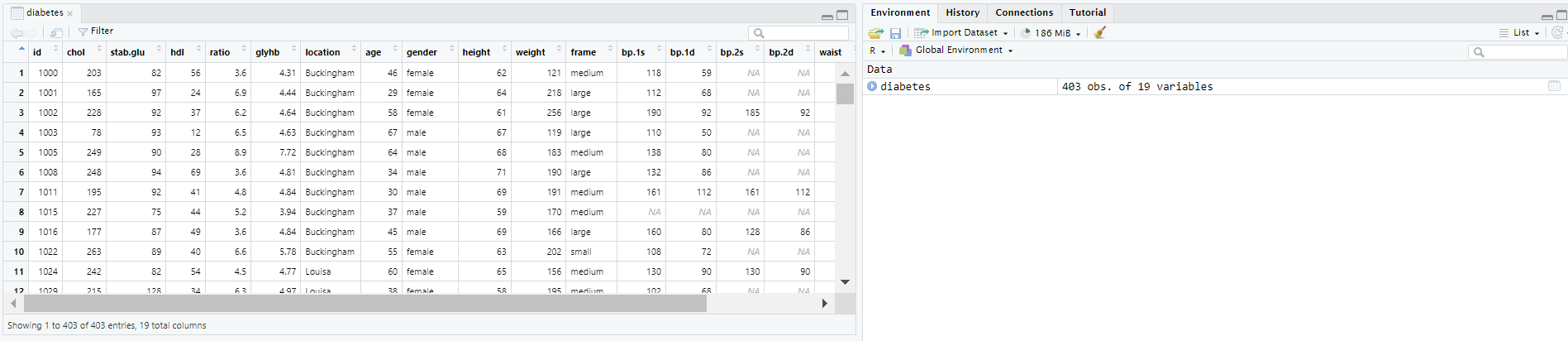
1. ✔️ On Posit Cloud, go to the upload button at the lower right (blue arrow below), select **Upload** → **Choose File,** and select the Diabetes dataset you saved earlier. Click ok once you have selected the Diabetes dataset.



1. ✔️ At the bottom right, you will see a file that says “diabetes.xlsx”. Click on it and then click on the pop-up that says “Import Dataset”. You’ll be brought to a pop-up asking if you want to download a few packages. Say yes to this as these packages will help to load the dataset**.** This will take a few seconds to complete. You’ll be brought up to this screen:



1. ✔️ The term “diabetes” will become a dataframe that represents that dataset as shown by **diabetes <- read\_excel("diabetes.xlsx")** line in the code preview at the bottom right. Click the “import” button on the bottom right. This will cause the preview code to run, which will import the dataset into a dataframe, ‘diabetes’ on Posit Cloud
2. ✔️ You will then see the dataset appear in a new tab under the Source panel and the “diabetes”dataframe will be added under the Environment panel. To look at the dataset, type the code **View(diabetes)** into your script, put the cursor to the right of the line of code, and click “Ctrl + Enter” at the same time on your keyboard. You should get a page pop up that looks like this. This is what the diabetes dataset looks like:



Each column in the dataframe is a vector that contains different information about each subject in the diabetes dataset, as defined below.

* + id - subject ID
  + chol - Cholesterol level
  + stab.glu - Stabilized Glucose
  + hdl - High Density Lipoprotein
  + ratio - Cholestrol/hdl
  + glyhb - Glycosolated Hemoglobin
  + bp.1s - First Systolic Blood Pressure
  + bp.1d - First Diastolic Blood Pressure
  + bp.2s - Second Systolic Blood Pressure
  + bp.2d - Second Diastolic Blood Pressure
  + time.ppn - Postprandial Time when Labs were Drawn

"First" refers to the initial measurement of the recorded blood pressure, taken at the beginning of a medical examination. The "Second" Blood Pressure refers to a subsequent measurement of the systolic or diastolic blood pressure, taken at a later time during the same medical visit or examination.

1. ✔️ Take a look at the dataset. You can also look at the first few lines of the dataset by typing and running the **head(diabetes)** command.
2. ✏️ Write down 3 observations from the dataset.
3. ✏️ How many patients are in the dataset? (Remember how to count the number of observations from the previous lesson. If you are stuck, ask the teacher for help).

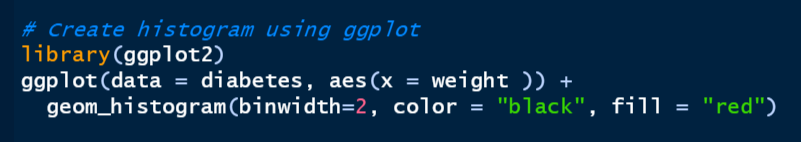
# **Activity #2: Visualizing data in the Diabetes Dataset using Histograms**

In this activity, you will begin exploratory data analysis by plotting histogram(s) of variables in the diabetes dataset to see if they appear to be normally distributed.

To create data visualizations like histograms, we can use a special package called **ggplot2.** ggplot2 is a powerful data visualization package in Posit Cloud that allows for the flexible and intuitive creation of visually appealing plots. With its code-driven approach, extensive customization options, and integration with data manipulation tools, ggplot2 facilitates reproducible and informative data visualizations.

* Tip: When using ggplot2, the code to type to use the package is **ggplot** instead of “ggplot2”.

1. ✔️ To create your first data visualization, we need to use ggplot2. To make sure it is implemented into the Posit Cloud, run **install.packages("ggplot2")** to install the package and then run **library(ggplot2)**. The library function is used to load packages that the user wants to use in their code to perform certain functions
2. ✔️ Let’s create a histogram to give us a visualization of how the data is distributed. To do this, type the code in this format

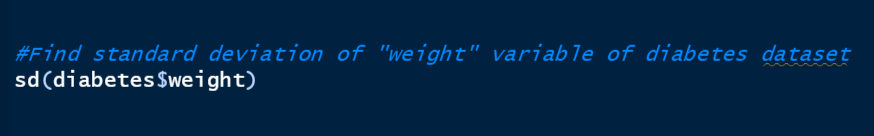


1. ✏️ Paste the plot you got here:
2. ✏️ Briefly explain what the aes, geom\_histogram, color, and fill commands do. Use the **?** command to help you.
3. ✏️ Change the **binwidth** to a different number of your choice and run the code. Paste your code and plot here. What do you notice happened to the histogram?
4. ✏️ Create a histogram with a variable other than weight. Paste your code and plot here.

Congratulations, you coded your first visualization!

# **Activity #3: Calculate the Standard Deviation of the Diabetes Dataset**

1. ✔️ You have created a histogram. But you want to know how spread out the data is. Calculating standard deviation can help us find out how spread out the data is. To do this, type the code in the format below.

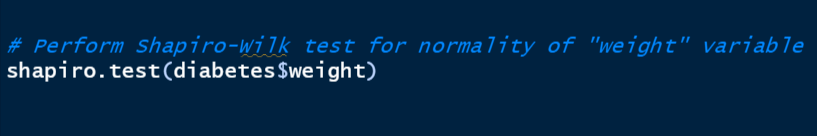


1. ✏️ What is the standard deviation of the weight variable of the diabetes dataset? Is the standard deviation bigger or smaller than you expected?
2. ✔️ Look back at the histogram you generated in Activity #2. Notice that if the standard deviation is large, the histogram is more spread out. If the standard deviation is small, the histogram is less spread out.
3. ✔️ Try the code above but use a different variable instead of weight.
4. ✏️ What variable(s) did you find the standard deviation for? What were their standard deviations? Is the standard deviation bigger or smaller than you expected? Did the standard deviations match how spread out the data was in your histograms?

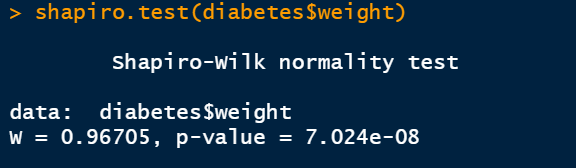
# **Activity #4: Testing for Normality in the Diabetes Dataset**

Testing for normality after creating a histogram is a common practice in statistics to assess the distribution of data. While a histogram provides a visual representation of the data's distribution, it might not be sufficient to conclusively determine if the data follows a normal distribution. By conducting a formal normality test, such as the Shapiro-Wilk test, you can obtain a quantitative measure of how closely the data follows a normal distribution.

1. ✔️ To perform a Shapiro-Wiki test, write the code formatted like this:



1. ✔️ After performing the code, you should be given an output in the console panel (lower left panel):



* Test Statistic (W): The test statistic W is a value that measures the departure of the sample data from a normal distribution. In this case, the calculated W value is 0.96705.
* p-value: The p-value is a measure of evidence against the null hypothesis that the data is normally distributed. A small p-value (typically less than 0.05) suggests that the data significantly deviates from a normal distribution. In this case, the p-value is 7.024e-08 (scientific notation for a very small value close to zero).
* Based on the test results, the data in the "weight" variable of the diabetes dataset does not appear to follow a normal distribution. The Shapiro-Wilk test result suggests that the data may have significant departures from normality, potentially having skewed or non-normal characteristics.

1. ✔️ Now perform more normality tests but use a different variable instead of weight.
2. ✏️ Perform the normality test on a variable you used to create a histogram in Activity #2. Paste the test results below.
3. ✏️ What were the p-values of the normality tests you performed?
4. ✏️ Does the test result fit with how the data looked in the histogram? Why or why not?
5. ✏️ Based on the results, did the variable you tested follow a normal distribution?