

# Intro to Data Science - Lab 1

DATA 1501 — Dr. Mihail  
Department of Computer Science  
Valdosta State University

September 29, 2021

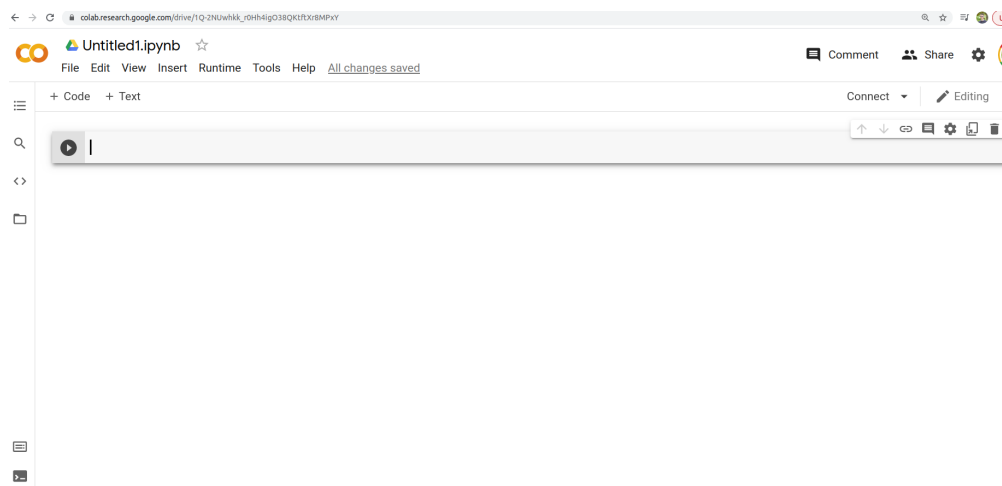
## 1 Introduction

In this lab, you will become familiar with Google’s Colab tool. This tool is freely available, using your Google account at: <https://colab.research.google.com>. It is an Interactive Python (IPython) environment hosted by Google, using their (high-end) computational resources.

## 2 Tasks

### 2.1 Task 1

Create a Google account or use your existing one to log into Colab. Create a new notebook. You should see something like the image below:



There are two types of “cells” in Colab: code cells and text cells. Code cells are used to write Python code, while Text cells are used to explain/document code.

## 2.2 Task 2

Write the following code into the first code cell in Colab:

```
# import useful libraries
import matplotlib.pyplot as plt
import numpy as np
```

Now execute that cell by clicking the circle with the triangle pointing East.

## 2.3 Task 3

Write the following code into a second code cell in Colab:

```
# create some play data
raw_data = [0.1, 0.9]
plt.figure()
plt.bar(["Group_1", "Group_2"], raw_data)
plt.ylabel("Percentage")
plt.title("Toy_example_of_bar_chart")
```

Execute that code. At this time, please let your instructor know so s/he can give you credit for the in-class portion of this lab.

# 3 Homework

In this part, you will create more Code cells and download the a dataset and its codebook.

## 3.1 Task 1

Create a cell and execute the following code:

```
# download the codebook and show it
!wget https://cs.valdosta.edu/~rpmihail/DATA1500/lab1/codebook.txt
!cat codebook.txt
```

## 3.2 Task 2

Create a cell and execute the following code:

```
# download the dataset
!wget https://cs.valdosta.edu/~rpmihail/DATA1500/lab1/data.csv
```

## 3.3 Task 3

Create a cell and execute the following code:

```
import pandas as pd
df = pd.read_csv('data.csv', error_bad_lines=False, delimiter="\t")

df.head()
```

### 3.4 Task 4

Create a cell and execute the following code:

```
# explore question 4
import numpy as np

raw_q4 = np.array(df.iloc[:, 3])
raw_q4 = raw_q4[~np.isnan(raw_q4)]
data = np.histogram(raw_q4, bins=5)
plt.figure()
plt.bar([1, 2, 3, 4, 5], data[0])
plt.xlabel('Likert scale')
plt.ylabel('Absolute count')
plt.title('Question 4 distribution')
```

## 4 Submission

**Part 1** Please create one Word document with screenshots of the results for Tasks 1 through 4. **CELL PHONE PICS ARE NOT ACCEPTABLE.** Please Google how to take a proper screenshot on the desktop/laptop computer you're working on.

**Part 2** Answer the following question: What was question 4 in the dataset you downloaded?

**Part 3** Write a paragraph interpreting the Figure obtained in the last step.

**Part 4** Write a paragraph describing what you learned.

**Due Date:** Before Midnight on Sunday, August 22nd.

Submit the Word document via Blazeview/Assignments/Lab 1.

Original dataset retrieved from:

<https://www.kaggle.com/lucasgreenwell/machivallianism-test-responses>