CSSE 386 – Data Mining with Programming
Rose-Hulman Institute of Technology

Worksheet 02

Name (Print):_____ Section:_____

1. For each definition provide the term. Also indicate whether it is a column, row, or table.

Term	Definition
	a characteristic that can be measured or observed
	a collection of information
	data points or observations

2. Define briefly 3 Vs of Big Data
   - Volume:
   - Variety:
   - Velocity:

3. Place lifecycle steps in order: explore, interpret, model, clean, gather
   1. → 2. → 3. → 4. → 5.

4. Complete the following

```
  # API from kaggle.....
import _______
# List the files in the downloaded directory
files = os.listdir(path)
print("Files in dataset:", _________________________)
csv_file = [file for file in files if file.endswith("______")][0]
file_path = os.path.join(path, csv_file)
df = pd._________________(file_path)
print(df._______________())

# Load a local file
from ____________________ import files
uploaded = files._________()
df2 = pd._____________________
print(_______.head())
```