Scott Berry

```python
from pyspark.sql import SparkSession

if __name__ == '__main__':
    spark = (SparkSession
             .builder
             .enableHiveSupport()
             .appName("Assignment3")
             .getOrCreate())

    people_json = spark.read.json("people.json")
    people_json.printSchema()
    people_json.createOrReplaceTempView("people_json")
    people_json.show()
    distinct_names = spark.sql("SELECT DISTINCT(name) FROM people_json")
    distinct_names.show()

    people_csv = spark.read.csv("people.txt")
    people_csv.show()
```

For ease, I moved “people.json” and “people.txt” to my working directory. I used the initially recommended methods of native python csv and json libraries to import the data, but Spark would throw many WARNs about how the methods had become deprecated, recommending the new Spark-native methods to Spark 2.0+. My final code uses these Spark-native libraries. The four outputs below are (in order): 1) the schema of the DF created from “people.json”, 2) the aforementioned DF, 3) the distinct names from said DF, and 4) the DF created from “people.txt”.

```
Thu Feb 24 12:40:16 berry@DESKTOP-SCOTT MINGW64 ~/OneDrive/Documents/Indiana University/Spring 2022/Data Science On Ramp/DsOnRampSpark (master)
$ python assignment_3.py
2022-02-24 12:40:26,864 WARN util.Shell: Did not find winutils.exe: {}
java.io.FileNotFoundException: Could not locate Hadoop executable: C:\Hadoop\hadoop-3.3.1\bin\winutils.exe -see https://wiki.apache.org/hadoop/WindowsProblems
        at org.apache.hadoop.util.Shell.getQualifiedBinInner(Shell.java:619)
        at org.apache.hadoop.util.Shell.getQualifiedBin(Shell.java:592)
        at org.apache.hadoop.util.Shell.<clinit>(Shell.java:689)
        at org.apache.hadoop.util.StringUtils.<clinit>(StringUtils.java:79)
        at org.apache.hadoop.conf.Configuration.getTimeDurationHelper(Configuration.java:1886)
        at org.apache.hadoop.conf.Configuration.getTimeDuration(Configuration.java:1846)
        at org.apache.hadoop.conf.Configuration.getTimeDuration(Configuration.java:1819)
        at org.apache.hadoop.util.ShutdownHookManager.getShutdownTimeout(ShutdownHookManager.java:183)
        at org.apache.hadoop.util.ShutdownHookManager$HookEntry.<init>(ShutdownHookManager.java:207)
        at org.apache.hadoop.util.ShutdownHookManager.addShutdownHook(ShutdownHookManager.java:304)
        at org.apache.spark.util.SparkShutdownHookManager.install(ShutdownHookManager.scala:181)
        at org.apache.spark.util.ShutdownHookManager$.shutdownHooks$lzycompute(ShutdownHookManager.scala:50)
        at org.apache.spark.util.ShutdownHookManager$.shutdownHooks(ShutdownHookManager.scala:48)
        at org.apache.spark.util.ShutdownHookManager$.addShutdownHook(ShutdownHookManager.scala:153)
        at org.apache.spark.util.ShutdownHookManager$.<init>(ShutdownHookManager.scala:58)
        at org.apache.spark.util.ShutdownHookManager$.<clinit>(ShutdownHookManager.scala)
        at org.apache.spark.util.Utils$.createTempDir(Utils.scala:335)
        at org.apache.spark.deploy.SparkSubmit.prepareSubmitEnvironment(SparkSubmit.scala:344)
        at org.apache.spark.deploy.SparkSubmit.org$apache$spark$deploy$SparkSubmit$$runMain(SparkSubmit.scala:898)
        at org.apache.spark.deploy.SparkSubmit.doRunMain$1(SparkSubmit.scala:180)
        at org.apache.spark.deploy.SparkSubmit.submit(SparkSubmit.scala:203)
        at org.apache.spark.deploy.SparkSubmit.doSubmit(SparkSubmit.scala:90)
        at org.apache.spark.deploy.SparkSubmit$$anon$2.doSubmit(SparkSubmit.scala:1043)
        at org.apache.spark.deploy.SparkSubmit$.main(SparkSubmit.scala:1052)
        at org.apache.spark.deploy.SparkSubmit.main(SparkSubmit.scala)
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
2022-02-24 12:40:27,041 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
root
 |-- age: long (nullable = true)
 |-- name: string (nullable = true)

1

+----+-------+
| age|   name|
+----+-------+
|null|Michael|
|  30|   Andy|
|  19| Justin|
+----+-------+

2

+-------+
|   name|
+-------+
|Michael|
|   Andy|
| Justin|
+-------+

3

+-------+---+
|    _c0|_c1|
+-------+---+
|Michael| 29|
|   Andy| 30|
| Justin| 19|
+-------+---+

4
```