

Spark by {Examples}(https://sparkbyexamples.com/)

- PySpark Tutorial**
 - [PySpark Tutorial For Beginners \(https://sparkbyexamples.com/pyspark-tutorial/\)](https://sparkbyexamples.com/pyspark-tutorial/)
 - [PySpark – Features \(https://sparkbyexamples.com/pyspark-tutorial/#features\)](https://sparkbyexamples.com/pyspark-tutorial/#features)
 - [PySpark – Advantages \(https://sparkbyexamples.com/pyspark-tutorial/#advantages\)](https://sparkbyexamples.com/pyspark-tutorial/#advantages)
 - [PySpark – Modules & Packages \(https://sparkbyexamples.com/pyspark-tutorial/#modules-packages\)](https://sparkbyexamples.com/pyspark-tutorial/#modules-packages)
 - [PySpark – Cluster Managers \(https://sparkbyexamples.com/pyspark-tutorial/#cluster-manager\)](https://sparkbyexamples.com/pyspark-tutorial/#cluster-manager)
 - [PySpark – Install on Windows \(https://sparkbyexamples.com/pyspark-tutorial/#pyspark-installation\)](https://sparkbyexamples.com/pyspark-tutorial/#pyspark-installation)
 - [PySpark – Web/Application UI \(https://sparkbyexamples.com/spark/spark-web-ui-understanding/\)](https://sparkbyexamples.com/spark/spark-web-ui-understanding/)
 - [PySpark – SparkSession \(https://sparkbyexamples.com/pyspark/pyspark-what-is-sparksession/\)](https://sparkbyexamples.com/pyspark/pyspark-what-is-sparksession/)
 - [PySpark – RDD \(https://sparkbyexamples.com/pyspark-rdd\)](https://sparkbyexamples.com/pyspark-rdd)
 - [PySpark – Parallelize \(https://sparkbyexamples.com/pyspark/pyspark-parallelize-create-rdd/\)](https://sparkbyexamples.com/pyspark/pyspark-parallelize-create-rdd/)
 - [PySpark – repartition\(\) vs coalesce\(\) \(https://sparkbyexamples.com/pyspark/pyspark-repartition-vs-coalesce/\)](https://sparkbyexamples.com/pyspark/pyspark-repartition-vs-coalesce/)
 - [PySpark – Broadcast Variables \(https://sparkbyexamples.com/pyspark/pyspark-broadcast-variables/\)](https://sparkbyexamples.com/pyspark/pyspark-broadcast-variables/)

PySpark (https://sparkbyexamples.com/pyspark-tutorial/)

Hive (https://sparkbyexamples.com/apache-hive-tutorial/)

HBase (https://sparkbyexamples.com/apache-hbase-tutorial/)

Kafka (https://sparkbyexamples.com/apache-kafka-tutorials-with-examples/)

FAQ's (https://sparkbyexamples.com/spark-questions/)

More (https://sparkbyexamples.com/)

Start your career at HCL

Calling Class XII Students For Full Time IT Jobs A

New Markets mean new tax challenges

Automate cross-border compliance with Avalara

See how



PySpark Read CSV file into DataFrame

NNK (https://sparkbyexamples.com/author/admin/) -
 PySpark (https://sparkbyexamples.com/category/pyspark/)

PySpark provides csv("path") on DataFrameReader to read a CSV file into PySpark DataFrame and dataframeObj.write.csv("path") to save or write to the CSV file. In this tutorial, you will learn how to read a single file, multiple files, all files from a local directory into DataFrame, applying some transformations, and finally writing DataFrame back to CSV file using PySpark example.

Data Science Course Bangalore Certified

Datamites - Data Science Courses in I

6-month/400 learning hours, 120-hour training, capstone & client projects



PREPARE FOR CODING INTERVIEWS

Placement Preparation I



www.prepbytes.com

New Markets mean new tax challenges

Automate cross-border compliance with Avalara

See how



[PySpark – Accumulator
\(https://sparkbyexamples.com/pyspark/pyspark-accumulator-with-example/\)](https://sparkbyexamples.com/pyspark/pyspark-accumulator-with-example/)

PySpark DataFrame

[PySpark – Create a DataFrame
\(https://sparkbyexamples.com/pyspark/different-ways-to-create-dataframe-in-pyspark/\)](https://sparkbyexamples.com/pyspark/different-ways-to-create-dataframe-in-pyspark/)

[PySpark – Create an empty DataFrame
\(https://sparkbyexamples.com/pyspark/pyspark-create-an-empty-dataframe/\)](https://sparkbyexamples.com/pyspark/pyspark-create-an-empty-dataframe/)

[PySpark – Convert RDD to DataFrame
\(https://sparkbyexamples.com/pyspark/convert-pyspark-rdd-to-dataframe/\)](https://sparkbyexamples.com/pyspark/convert-pyspark-rdd-to-dataframe/)

[PySpark – Convert DataFrame to Pandas
\(https://sparkbyexamples.com/pyspark/convert-pyspark-dataframe-to-pandas/\)](https://sparkbyexamples.com/pyspark/convert-pyspark-dataframe-to-pandas/)

[PySpark – show\(\)
\(https://sparkbyexamples.com/pyspark/pyspark-show-display-dataframe-contents-in-table/\)](https://sparkbyexamples.com/pyspark/pyspark-show-display-dataframe-contents-in-table/)

[PySpark – StructType & StructField
\(https://sparkbyexamples.com/pyspark/pyspark-structtype-and-structfield/\)](https://sparkbyexamples.com/pyspark/pyspark-structtype-and-structfield/)

[PySpark – Row Class
\(https://sparkbyexamples.com/pyspark/pyspark-row-using-rdd-dataframe/\)](https://sparkbyexamples.com/pyspark/pyspark-row-using-rdd-dataframe/)

[PySpark – Column Class
\(https://sparkbyexamples.com/pyspark/pyspark-column-functions/\)](https://sparkbyexamples.com/pyspark/pyspark-column-functions/)

[PySpark – select\(\)
\(https://sparkbyexamples.com/pyspark/select-columns-from-pyspark-dataframe/\)](https://sparkbyexamples.com/pyspark/select-columns-from-pyspark-dataframe/)

[PySpark – collect\(\)
\(https://sparkbyexamples.com/pyspark/pyspark-collect/\)](https://sparkbyexamples.com/pyspark/pyspark-collect/)

[PySpark – withColumn\(\)
\(https://sparkbyexamples.com/pyspark/pyspark-withcolumn/\)](https://sparkbyexamples.com/pyspark/pyspark-withcolumn/)

Note: PySpark out of the box supports reading files in CSV, JSON, and many more file formats into PySpark DataFrame.

Table of contents:

- [PySpark Read CSV file into DataFrame](#)
 - [Read multiple CSV files](#)
 - [Read all CSV files in a directory](#)
- [Options while reading CSV file](#)
 - [delimiter](#)
 - [InferSchema](#)
 - [header](#)
 - [quotes](#)
 - [nullValues](#)
 - [dateFormat](#)
- [Read CSV files with a user-specified schema](#)
- [Applying DataFrame transformations](#)
- [Write DataFrame to CSV file](#)
 - [Using options](#)
 - [Saving Mode](#)

1. PySpark Read CSV File into DataFrame

Using `csv("path")` or `format("csv").load("path")` of `DataFrameReader`, you can read a CSV file into a PySpark DataFrame. These methods take a file path to read from as an argument. When you use `format("csv")` method, you can also specify the Data sources by their fully qualified name, but for built-in sources, you can simply use their short names (`csv`, `json`, `parquet`, `jdbc`, `text` e.t.c).

Data Science Course Bangalore Certified

Datamites - Data Science Courses in I

6-month/400 learning hours, 120-hour training, capstone & client projects



WEBSITE



[PySpark – withColumnRenamed\(\).](#)
(<https://sparkbyexamples.com/pyspark/pyspark-rename-dataframe-column/>).

[PySpark – where\(\) & filter\(\).](#)
(<https://sparkbyexamples.com/pyspark/pyspark-where-filter/>).

[PySpark – drop\(\) & dropDuplicates\(\).](#)
(<https://sparkbyexamples.com/pyspark/pyspark-distinct-to-drop-duplicates/>).

[PySpark – orderBy\(\) and sort\(\).](#)
(<https://sparkbyexamples.com/pyspark/pyspark-orderby-and-sort-explained/>).

[PySpark – groupBy\(\).](#)
(<https://sparkbyexamples.com/pyspark/pyspark-groupby-explained-with-example/>).

[PySpark – join\(\).](#)
(<https://sparkbyexamples.com/pyspark/pyspark-join-explained-with-examples/>).

[PySpark – union\(\) & unionAll\(\).](#)
(<https://sparkbyexamples.com/pyspark/pyspark-union-and-unionall/>).

[PySpark – unionByName\(\).](#)
(<https://sparkbyexamples.com/spark/spark-merge-two-dataframes-with-different-columns/>).

[PySpark – UDF \(User Defined Function\).](#)
(<https://sparkbyexamples.com/pyspark/pyspark-udf-user-defined-function/>).

[PySpark – map\(\).](#)
(<https://sparkbyexamples.com/pyspark/pyspark-map-transformation/>).

[PySpark – flatMap\(\).](#)
(<https://sparkbyexamples.com/pyspark/pyspark-flatmap-transformation/>).

[pyspark – foreach\(\).](#)
(<https://sparkbyexamples.com/pyspark/pyspark-loop-iterate-through-rows-in-dataframe/#use-foreach-loop-through-dataframe>).

[PySpark – sample\(\) vs sampleBy\(\).](#)

Refer dataset [zipcodes.csv](#) at [GitHub](#) (<https://github.com/spark-examples/pyspark-examples/blob/master/resources/zipcodes.csv>)

```
val spark = SparkSession.builder()
    .appName("SparkByExample")
    .getOrCreate()
df = spark.read.csv("/tmp/resources/zipcodes.csv")
df.printSchema()
```

Using fully qualified data source name, you can alternatively do the following.

```
df = spark.read.format("csv")
    .load("/tmp/resources/zipcodes.csv")
//      or
df = spark.read.format("org.apache.spark.sql.csv")
    .load("/tmp/resources/zipcodes.csv")
df.printSchema()
```

This example reads the data into DataFrame columns "_c0" for the first column and "_c1" for the second and so on. and by default data type for all these columns is treated as String.

```
root
|-- _c0: string (nullable = true)
|-- _c1: string (nullable = true)
|-- _c2: string (nullable = true)
```

1.1 Using Header Record For Column Names

If you have a header with column names on your input file, you need to explicitly specify True for header option using [option\("header", True\)](#) not mentioning this, the API treats header as a data record.

[\(https://sparkbyexamples.com/pyspark/pyspark-sampling-example/\)](https://sparkbyexamples.com/pyspark/pyspark-sampling-example/)

[PySpark – fillna\(\) & fill\(\)
\(https://sparkbyexamples.com/pyspark/pyspark-fillna-fill-replace-null-values/\)](https://sparkbyexamples.com/pyspark/pyspark-fillna-fill-replace-null-values/)

[PySpark – pivot\(\)_\(Row to Column\)
\(https://sparkbyexamples.com/pyspark/pyspark-pivot-and-unpivot-dataframe/\)](https://sparkbyexamples.com/pyspark/pyspark-pivot-and-unpivot-dataframe/)

[PySpark – partitionBy\(\).
\(https://sparkbyexamples.com/pyspark/pyspark-partitionby-example/\)](https://sparkbyexamples.com/pyspark/pyspark-partitionby-example/)

[PySpark – ArrayType Column \(Array\).
\(https://sparkbyexamples.com/pyspark/pyspark-arraytype-column-with-examples/\)](https://sparkbyexamples.com/pyspark/pyspark-arraytype-column-with-examples/)

[PySpark – MapType \(Map/Dict\).
\(https://sparkbyexamples.com/pyspark/pyspark-maptype-dict-examples/\)](https://sparkbyexamples.com/pyspark/pyspark-maptype-dict-examples/)

PySpark SQL Functions

[PySpark – Aggregate Functions
\(https://sparkbyexamples.com/pyspark/pyspark-aggregate-functions/\)](https://sparkbyexamples.com/pyspark/pyspark-aggregate-functions/)

[PySpark – Window Functions
\(https://sparkbyexamples.com/pyspark/pyspark-window-functions/\)](https://sparkbyexamples.com/pyspark/pyspark-window-functions/)

[PySpark – Date and Timestamp Functions
\(https://sparkbyexamples.com/pyspark/pyspark-sql-date-and-timestamp-functions/\)](https://sparkbyexamples.com/pyspark/pyspark-sql-date-and-timestamp-functions/)

[PySpark – JSON Functions
\(https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/\)](https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/)

PySpark Datasources

[PySpark – Read & Write CSV File
\(https://sparkbyexamples.com/pyspark/pyspark-read-csv-file-into-dataframe/\)](https://sparkbyexamples.com/pyspark/pyspark-read-csv-file-into-dataframe/)

[PySpark – Read & Write Parquet File
\(https://sparkbyexamples.com/pyspark/pyspark-read-write-parquet-file/\)](https://sparkbyexamples.com/pyspark/pyspark-read-write-parquet-file/)

```
df2 = spark.read.option("header", true) \
    .csv("/tmp/resources/zipcodes.csv")
```

As mentioned earlier, PySpark reads all columns as a string (StringType) by default. I will explain in later sections on how to read the schema (inferred schema) from the header record and derive the column type based on the data.

Our combined feed and effluent heat exchanger will improve your process



1.2 Read Multiple CSV Files

Using the `read.csv()` method you can also read multiple csv files, just pass all file names by separating comma as a path, for example :

```
df = spark.read.csv("path1,path2,path3")
```

1.3 Read all CSV Files in a Directory

We can read all CSV files from a directory into DataFrame just by passing directory as a path to the `csv()` method.

```
df = spark.read.csv("Folder path")
```

2. Options While Reading CSV File

[pyspark/pyspark-read-and-write-parquet-file/](#)

[PySpark – Read & Write JSON file](#)
(<https://sparkbyexamples.com/pyspark/pyspark-read-json-file-into-dataframe/>)

PySpark Built-In Functions

[PySpark – when\(\)](#)
(<https://sparkbyexamples.com/pyspark/pyspark-when-otherwise/>)

[PySpark – expr\(\)](#)
(<https://sparkbyexamples.com/pyspark/pyspark-sql-expr-expression-function/>)

[PySpark – lit\(\)](#)
(<https://sparkbyexamples.com/pyspark/pyspark-lit-add-literal-constant/>)

[PySpark – split\(\)](#)
(<https://sparkbyexamples.com/pyspark/pyspark-convert-string-to-array-column/>)

[PySpark – concat_ws\(\)](#)
(<https://sparkbyexamples.com/pyspark/pyspark-convert-array-column-to-string-column/>)

[Pyspark – substring\(\)](#)
(<https://sparkbyexamples.com/pyspark/pyspark-substring-from-a-column/>)

[PySpark – translate\(\)](#)
(<https://sparkbyexamples.com/pyspark/pyspark-replace-column-values/#translate-replace-character-by-character>)

[PySpark – regexp_replace\(\)](#)
(https://sparkbyexamples.com/pyspark/pyspark-replace-column-values/#regexp_replace-replace-string-columns)

[PySpark – overlay\(\)](#)
(<https://sparkbyexamples.com/pyspark/pyspark-replace-column-values/#overlay-function>)

[PySpark – to_timestamp\(\)](#)
(<https://sparkbyexamples.com/spark/pyspark-to-timestamp->

PySpark CSV dataset provides multiple options to work with CSV files. Below are some of the most important options explained with examples.

You can either use chaining `option(self, key, value)` to use multiple options or use alternate `options(self, **options)` method.

2.1 delimiter

`delimiter` option is used to specify the column delimiter of the CSV file. By default, it is **comma (,)** character, but can be set to any character like **pipe(|)**, **tab (\t)**, **space** using this option.

```
df3 = spark.read.options(delimiter='|')  
      .csv("C:/apps/sparkbyexamples/CSV/employees.csv")
```

2.2 inferSchema

The default value set to this option is `False` when setting to `true` it automatically infers column types based on the data. Note that, it requires reading the data one more time to infer the schema.

```
df4 = spark.read.options(inferSchema='true')  
      .csv("src/main/resources/zipcodes.csv")
```

Alternatively you can also write this by chaining `option()` method.

```
df4 = spark.read.option("inferSchema", "true")  
      .option("delimiter", "|")  
      .csv("src/main/resources/zipcodes.csv")
```

2.3 header

This option is used to read the first line of the CSV file as column names. By default the value of this option

[convert-string-to-timestamp-type/](#)

[PySpark – to_date\(\) \(https://sparkbyexamples.com/pyspark/pyspark-to_date-convert-timestamp-to-date/\)](#)

[PySpark – date_format\(\) \(https://sparkbyexamples.com/pyspark/pyspark-date_format-convert-date-to-string-format/\)](#)

[PySpark – datediff\(\) \(https://sparkbyexamples.com/pyspark/pyspark-difference-between-two-dates-days-months-years/#datediff\)](#)

[PySpark – months_between\(\) \(https://sparkbyexamples.com/pyspark/pyspark-difference-between-two-dates-days-months-years/#months_between\(\)\)](#)

[PySpark – explode\(\) \(https://sparkbyexamples.com/pyspark/pyspark-explode-nested-array-into-rows/\)](#)

[PySpark – array_contains\(\) \(https://sparkbyexamples.com/pyspark/pyspark-arraytype-column-with-examples/#array_contains\)](#)

[PySpark – array\(\) \(https://sparkbyexamples.com/pyspark/pyspark-arraytype-column-with-examples/#array\)](#)

[PySpark – collect_list\(\) \(https://sparkbyexamples.com/pyspark/pyspark-aggregate-functions/#collect-list\)](#)

[PySpark – collect_set\(\) \(https://sparkbyexamples.com/pyspark/pyspark-aggregate-functions/#collect-set\)](#)

[PySpark – create_map\(\) \(https://sparkbyexamples.com/pyspark/pyspark-convert-dataframe-columns-to-maptype-dict/\)](#)

[PySpark – map_keys\(\) \(https://sparkbyexamples.com/pyspark/pyspark-maptype-dict-examples/#map_keys\)](#)

[PySpark – map_values\(\) \(https://sparkbyexamples.com/pyspark/pyspark-maptype-dict-examples/#map_values\)](#)

is False , and all column types are assumed to be a string.

```
df3 = spark.read.options(header=
.csv("/tmp/resources/zipcodes
```

2.4 quotes

When you have a column with a delimiter that used to split the columns, use quotes option to specify the quote character, by default it is " and delimiters inside quotes are ignored. but using this option you can set any character.

2.5 nullValues

Using nullValues option you can specify the string in a CSV to consider as null. For example, if you want to consider a date column with a value "1900-01-01" set null on DataFrame.

2.6 dateFormat

dateFormat option to used to set the format of the input [DateType](#) ([https://sparkbyexamples.com/pyspark/pyspark-sql-date-and-timestamp-functions/#pyspark-sql-date-functions](#)) and [TimestampType](#) ([https://sparkbyexamples.com/pyspark/pyspark-sql-date-and-timestamp-functions/#pyspark-sql-timestamp-functions](#)) columns. Supports all java.text.SimpleDateFormat formats.

Note: Besides the above options, PySpark CSV API also supports many other options, [please refer to this article for details](#) ([https://docs.databricks.com/data/data-sources/read-csv.html](#)).

[PySpark – struct\(\).](https://sparkbyexamples.com/pyspark/pyspark-structtype-and-structfield/#update-struct-function)
(<https://sparkbyexamples.com/pyspark/pyspark-structtype-and-structfield/#update-struct-function>)

[PySpark – countDistinct\(\).](https://sparkbyexamples.com/pyspark/pyspark-count-distinct-from-dataframe/)
(<https://sparkbyexamples.com/pyspark/pyspark-count-distinct-from-dataframe/>)

[PySpark – sum\(\).avg\(\).](https://sparkbyexamples.com/pyspark/pyspark-dataframe-groupby-and-sort-by-descending-order/)
(<https://sparkbyexamples.com/pyspark/pyspark-dataframe-groupby-and-sort-by-descending-order/>)

[PySpark – row_number\(\).](https://sparkbyexamples.com/pyspark/pyspark-window-functions/#row_number)
(https://sparkbyexamples.com/pyspark/pyspark-window-functions/#row_number)

[PySpark – rank\(\).](https://sparkbyexamples.com/pyspark/pyspark-window-functions/#rank)
(<https://sparkbyexamples.com/pyspark/pyspark-window-functions/#rank>)

[PySpark – dense_rank\(\).](https://sparkbyexamples.com/pyspark/pyspark-window-functions/#dense_rank)
(https://sparkbyexamples.com/pyspark/pyspark-window-functions/#dense_rank)

[PySpark – percent_rank\(\).](https://sparkbyexamples.com/pyspark/pyspark-window-functions/#percent_rank)
(https://sparkbyexamples.com/pyspark/pyspark-window-functions/#percent_rank)

[PySpark – typedLit\(\).](https://sparkbyexamples.com/pyspark/pyspark-lit-add-literal-constant/#typedlit)
(<https://sparkbyexamples.com/pyspark/pyspark-lit-add-literal-constant/#typedlit>)

[PySpark – from_json\(\).](https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#from_json)
(https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#from_json)

[PySpark – to_json\(\).](https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#to_json)
(https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#to_json)

[PySpark – json_tuple\(\).](https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#json_tuple)
(https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#json_tuple)

[PySpark – get_json_object\(\).](https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#get_json_object)
(https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#get_json_object)

[PySpark – schema_of_json\(\).](https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#schema_of_json)
(https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#schema_of_json)

3. Reading CSV files with a user-specified custom schema

If you know the schema of the file ahead and do not want to use the `inferSchema` option for column names and types, use [user-defined custom column names and type](https://sparkbyexamples.com/pyspark/pyspark-structtype-and-structfield/) (<https://sparkbyexamples.com/pyspark/pyspark-structtype-and-structfield/>) using `schema` option.

```
schema = StructType() \
    .add("RecordNumber", IntegerType()) \
    .add("Zipcode", IntegerType()) \
    .add("ZipCodeType", StringType()) \
    .add("City", StringType()) \
    .add("State", StringType()) \
    .add("LocationType", StringType()) \
    .add("Lat", DoubleType(), True) \
    .add("Long", DoubleType(), True) \
    .add("Xaxis", IntegerType()) \
    .add("Yaxis", DoubleType()) \
    .add("Zaxis", DoubleType()) \
    .add("WorldRegion", StringType()) \
    .add("Country", StringType()) \
    .add("LocationText", StringType()) \
    .add("Location", StringType()) \
    .add("Decommisioned", BooleanType()) \
    .add("TaxReturnsFiled", StringType()) \
    .add("EstimatedPopulation", IntegerType()) \
    .add("TotalWages", IntegerType()) \
    .add("Notes", StringType())

df_with_schema = spark.read.format("csv") \
    .option("header", True) \
    .schema(schema) \
    .load("/tmp/resources/zipcodes.csv")
```

4. Applying DataFrame transformations

Once you have created DataFrame from the CSV file, you can apply all transformation and actions DataFrame support. Please refer to the link for more details.

5. Write PySpark

DataFrame to CSV file

Use the `write()` method of the PySpark `DataFrameWriter` object to write PySpark `DataFrame` to a CSV file.

```
df.write.option("header", True)
    .csv("/tmp/spark_output/zipcode")
```

5.1 Options

While writing a CSV file you can use several options. for example, `header` to output the `DataFrame` column names as header record and `delimiter` to specify the delimiter on the CSV output file.

```
df2.write.options(header='True'
    .csv("/tmp/spark_output/zipcode")
```

Other options

available `quote`, `escape`, `nullValue`, `dateFormat`, `quoteMode` .

5.2 Saving modes

PySpark `DataFrameWriter` also has a method `mode()` to specify saving mode.

`overwrite` – mode is used to overwrite the existing file.

`append` – To add the data to the existing file.

`ignore` – Ignores write operation when the file already exists.

`error` – This is a default option when the file already exists, it returns an error.


```
df2.write.mode('overwrite').csv  
//you can also use this  
df2.write.format("csv").mode('o
```

6. PySpark Read CSV

Complete Example

```

import pyspark
from pyspark.sql import SparkSession
from pyspark.sql.types import StructType, StructField
from pyspark.sql.types import ArrayType, StringType, IntegerType
from pyspark.sql.functions import col

spark = SparkSession.builder.appName("Zipcode").getOrCreate()

df = spark.read.csv("/tmp/resources/zipcodes.csv")

df.printSchema()

df2 = spark.read.option("header", True) \
    .csv("/tmp/resources/zipcodes.csv")
df2.printSchema()

df3 = spark.read.options(headers="true") \
    .csv("/tmp/resources/zipcodes.csv")
df3.printSchema()

schema = StructType() \
    .add("RecordNumber", IntegerType()) \
    .add("Zipcode", IntegerType()) \
    .add("ZipCodeType", StringType()) \
    .add("City", StringType()) \
    .add("State", StringType()) \
    .add("LocationType", StringType()) \
    .add("Lat", DoubleType(), True) \
    .add("Long", DoubleType(), True) \
    .add("Xaxis", IntegerType()) \
    .add("Yaxis", DoubleType()) \
    .add("Zaxis", DoubleType()) \
    .add("WorldRegion", StringType()) \
    .add("Country", StringType()) \
    .add("LocationText", StringType()) \
    .add("Location", StringType()) \
    .add("Decommisioned", BooleanType()) \
    .add("TaxReturnsFiled", StringType()) \
    .add("EstimatedPopulation", IntegerType()) \
    .add("TotalWages", IntegerType()) \
    .add("Notes", StringType())

df_with_schema = spark.read.format("csv") \
    .option("header", True) \
    .schema(schema) \
    .load("/tmp/resources/zipcodes.csv")
df_with_schema.printSchema()

df2.write.option("header", True) \
    .csv("/tmp/spark_output/zipcodes.csv")

```

7. Conclusion:

In this tutorial, you have learned how to read a CSV file, multiple CSV files and all files from a local folder into PySpark DataFrame, using multiple options to change the default behavior and write CSV files back to DataFrame using different save options.

Happy Learning !!

References:

- [Databricks read CSV](https://docs.databricks.com/data/data-sources/read-csv.html)
(<https://docs.databricks.com/data/data-sources/read-csv.html>)
- [PySpark CSV library](https://github.com/databricks/spark-csv)
(<https://github.com/databricks/spark-csv>)

Share this:



(<https://sparkbyexamples.com/pyspark/pyspark-read-csv-file-into-dataframe/?share=facebook&nb=1>)



(<https://sparkbyexamples.com/pyspark/pyspark-read-csv-file-into-dataframe/?share=reddit&nb=1>)



(<https://sparkbyexamples.com/pyspark/pyspark-read-csv-file-into-dataframe/?share=pinterest&nb=1>)

1



(<https://sparkbyexamples.com/pyspark/pyspark-read-csv-file-into-dataframe/?share=tumblr&nb=1>)



(<https://sparkbyexamples.com/pyspark/pyspark-read-csv-file-into-dataframe/?share=pocket&nb=1>)



(<https://sparkbyexamples.com/pyspark/pyspark-read-csv-file-into-dataframe/?share=linkedin&nb=1>)



(<https://sparkbyexamples.com/pyspark/pyspark-read-csv-file-into-dataframe/?share=twitter&nb=1>)

TAGS: [CSV](https://sparkbyexamples.com/tag/csv/),
([HTTPS://SPARKBYEXAMPLES.COM/TAG/CSV/](https://sparkbyexamples.com/tag/csv/)),
[HEADER](https://sparkbyexamples.com/tag/header/)
([HTTPS://SPARKBYEXAMPLES.COM/TAG/HEADER/](https://sparkbyexamples.com/tag/header/)),
[PYSARK WRITE CSV](https://sparkbyexamples.com/tag/pyspark-write-csv/)
([HTTPS://SPARKBYEXAMPLES.COM/TAG/PYSARK-WRITE-CSV/](https://sparkbyexamples.com/tag/pyspark-write-csv/)), [SCHEMA](https://sparkbyexamples.com/tag/schema/)
([HTTPS://SPARKBYEXAMPLES.COM/TAG/SCHEMA/](https://sparkbyexamples.com/tag/schema/))
).



NNK

[\(Https://Sparkbyexamples.Com/Author/Admin/\)](https://Sparkbyexamples.Com/Author/Admin/)

[\(https://sparkbyexamples.com/author/admin/\)](https://sparkbyexamples.com/author/admin/)

SparkByExamples.com is a Big Data and Spark examples community page, all examples are simple and easy to understand and well tested in our development environment [Read more ..](#)

[\(https://sparkbyexamples.com/about-sparkbyexamples/\)](https://sparkbyexamples.com/about-sparkbyexamples/)

➤ THIS POST HAS 8 COMMENTS



Anonymous

18 APR 2021 [REPLY](#)

where is zipcode CSV file ?



NNK

18 APR 2021 [REPLY](#)

You can find it at [zipcodes.csv @ GitHub \(https://github.com/spark-examples/pyspark-examples/blob/master/resources/zipcodes.csv\)](https://github.com/spark-examples/pyspark-examples/blob/master/resources/zipcodes.csv).



Anonymous

11 FEB 2021 [REPLY](#)

Really very helpful pyspark example..Thanks for the details!!



Anonymous

5 JAN 2021 [REPLY](#)

Very much helpful!! Thank you for the article!!

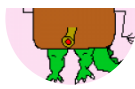


NNK

5 JAN 2021 [REPLY](#)

Thank you!!





Anonymous

2 NOV 2020 [REPLY](#)

Thank you so much for
this article 😊



NNK 2 NOV 2020 [REPLY](#)

Glad you like it.



Anonymous

19 SEP 2020 [REPLY](#)

Thanks for the example.
could you please explain
how to define/initialise
the “spark” in the above
example (e.g.
spark.read.csv)?

Leave a Reply



Apache Hadoop
(<https://sparkbyexamples.com/category/hadoop/>)

Apache Spark
(<https://sparkbyexamples.com/category/spark/>)

Apache Spark Streaming
(<https://sparkbyexamples.com/category/spark/apache-spark-streaming/>)

Apache Kafka
(<https://sparkbyexamples.com/category/kafka/>)

Apache HBase
(<https://sparkbyexamples.com/category/hbase/>)

Apache Cassandra
(<https://sparkbyexamples.com/category/cassandra/>)

Snowflake Database
(<https://sparkbyexamples.com/category/snowflake/>)

H2O Sparkling Water
(<https://sparkbyexamples.com/category/h2o-sparkling-water/>)

PySpark
(<https://sparkbyexamples.com/category/pyspark/>)

Spark regexp_replace() – Replace String Value

(https://sparkbyexamples.com/spark/spark-regexp_replace-replace-string-value/)

How to Run a PySpark Script from Python?

(<https://sparkbyexamples.com/pyspark/r/un-pyspark-script-from-python-subprocess/>)

Spark SQL like() Using Wildcard Example

(<https://sparkbyexamples.com/spark/spark-sql-like-using-wildcard-example/>)

Spark isin() & IS NOT IN Operator Example

(<https://sparkbyexamples.com/spark/spark-isin-is-not-in-operator-example/>)

Spark – Get Size/Length of Array & Map Column

(<https://sparkbyexamples.com/spark/spark-get-size-length-of-array-map-column/>)

Spark Using Length/Size Of a DataFrame Column

(<https://sparkbyexamples.com/spark/spark-using-length-size-of-a-dataframe-column/>)

Spark rlike() Working with Regex
Matching Examples
(<https://sparkbyexamples.com/spark/spark-rlike-regex-matching-examples/>)

Spark Check String Column Has Numeric Values

(<https://sparkbyexamples.com/spark/spark-check-string-column-has-numeric-values/>)

Spark Check Column Data Type is Integer or String
(<https://sparkbyexamples.com/spark/spark-check-column-data-type-is-integer-or-string/>)

SparkByExamples.com is a Big Data and Spark examples community page, all examples are simple and easy to understand, and well tested in our development environment Read more .. (<https://sparkbyexamples.com/about-sparkbyexamples/>)

(https: (https:

//www. //www.

(https: facebo linkedi (https:

//twitter ok.co n.com/ //github

r.com/ m/spar in/n- b.com/

sparkb kbyex nk- spark-

yexam ample b860a exam

ples) s/) 8193/) les/)

