

---

# Big Data Analytics

## Formative Assessment 4: Data Wrangling with Spark's Structured APIs

**BILLONES**, Cristel Kaye

**ROSALES**, Frances

**SAMSON**, Justine Aizel

---

### Dataset:

NYC Yellow Taxi Trips – TLC Trip Record Data (Monthly CSV)[Links to an external site.](#)

### Objective:

Apply Spark Structured APIs to perform data loading, cleaning, transformation, and basic analysis on a real-world dataset.

---

## Importing Libraries

```
In [1]: # Install Java and Spark
!apt-get install openjdk-11-jdk-headless -qq > /dev/null
!wget -q https://archive.apache.org/dist/spark/spark-3.4.1/spark-3.4.1-bin-hadoop3.
!tar -xvzf spark-3.4.1-bin-hadoop3.tgz
!pip install -q findspark

# Set environment variables
import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-11-openjdk-amd64"
os.environ["SPARK_HOME"] = "/content/spark-3.4.1-bin-hadoop3"

# Initialize findspark
import findspark
findspark.init()
```

spark-3.4.1-bin-hadoop3/  
spark-3.4.1-bin-hadoop3/R/  
spark-3.4.1-bin-hadoop3/R/lib/  
spark-3.4.1-bin-hadoop3/R/lib/sparkr.zip  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/html/  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/html/R.css  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/html/00Index.html  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/INDEX  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/help/  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/help/aliases.rds  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/help/AnIndex  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/help/SparkR.rdx  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/help/SparkR.rdb  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/help/paths.rds  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/worker/  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/worker/worker.R  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/worker/daemon.R  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/tests/  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/tests/testthat/  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/tests/testthat/test\_basic.R  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/profile/  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/profile/shell.R  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/profile/general.R  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/doc/  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/doc/index.html  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/doc/sparkr-vignettes.html  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/doc/sparkr-vignettes.Rmd  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/doc/sparkr-vignettes.R  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/R/  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/R/SparkR.rdx  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/R/SparkR.rdb  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/R/SparkR  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/Meta/  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/Meta/nsInfo.rds  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/Meta/vignette.rds  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/Meta/Rd.rds  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/Meta/links.rds  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/Meta/hsearch.rds  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/Meta/features.rds  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/Meta/package.rds  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/DESCRIPTION  
spark-3.4.1-bin-hadoop3/R/lib/SparkR/NAMESPACE  
spark-3.4.1-bin-hadoop3/sbin/  
spark-3.4.1-bin-hadoop3/sbin/workers.sh  
spark-3.4.1-bin-hadoop3/sbin/stop-workers.sh  
spark-3.4.1-bin-hadoop3/sbin/stop-worker.sh  
spark-3.4.1-bin-hadoop3/sbin/stop-thriftserver.sh  
spark-3.4.1-bin-hadoop3/sbin/stop-slaves.sh  
spark-3.4.1-bin-hadoop3/sbin/stop-slave.sh  
spark-3.4.1-bin-hadoop3/sbin/stop-mesos-shuffle-service.sh  
spark-3.4.1-bin-hadoop3/sbin/stop-mesos-dispatcher.sh  
spark-3.4.1-bin-hadoop3/sbin/stop-master.sh  
spark-3.4.1-bin-hadoop3/sbin/stop-history-server.sh  
spark-3.4.1-bin-hadoop3/sbin/stop-connect-server.sh  
spark-3.4.1-bin-hadoop3/sbin/stop-all.sh

spark-3.4.1-bin-hadoop3/sbin/start-workers.sh  
spark-3.4.1-bin-hadoop3/sbin/start-worker.sh  
spark-3.4.1-bin-hadoop3/sbin/start-thriftserver.sh  
spark-3.4.1-bin-hadoop3/sbin/start-slaves.sh  
spark-3.4.1-bin-hadoop3/sbin/start-slave.sh  
spark-3.4.1-bin-hadoop3/sbin/start-mesos-shuffle-service.sh  
spark-3.4.1-bin-hadoop3/sbin/start-mesos-dispatcher.sh  
spark-3.4.1-bin-hadoop3/sbin/start-master.sh  
spark-3.4.1-bin-hadoop3/sbin/start-history-server.sh  
spark-3.4.1-bin-hadoop3/sbin/start-connect-server.sh  
spark-3.4.1-bin-hadoop3/sbin/start-all.sh  
spark-3.4.1-bin-hadoop3/sbin/spark-daemons.sh  
spark-3.4.1-bin-hadoop3/sbin/spark-daemon.sh  
spark-3.4.1-bin-hadoop3/sbin/spark-config.sh  
spark-3.4.1-bin-hadoop3/sbin/slaves.sh  
spark-3.4.1-bin-hadoop3/sbin/decommission-worker.sh  
spark-3.4.1-bin-hadoop3/sbin/decommission-slave.sh  
spark-3.4.1-bin-hadoop3/python/  
spark-3.4.1-bin-hadoop3/python/dist/  
spark-3.4.1-bin-hadoop3/python/pyspark.egg-info/  
spark-3.4.1-bin-hadoop3/python/pyspark.egg-info/SOURCES.txt  
spark-3.4.1-bin-hadoop3/python/pyspark.egg-info/top\_level.txt  
spark-3.4.1-bin-hadoop3/python/pyspark.egg-info/requirements.txt  
spark-3.4.1-bin-hadoop3/python/pyspark.egg-info/dependency\_links.txt  
spark-3.4.1-bin-hadoop3/python/pyspark.egg-info/PKG-INFO  
spark-3.4.1-bin-hadoop3/python/test\_support/  
spark-3.4.1-bin-hadoop3/python/test\_support/userlibrary.py  
spark-3.4.1-bin-hadoop3/python/test\_support/userlib-0.1.zip  
spark-3.4.1-bin-hadoop3/python/test\_support/test\_pytorch\_training\_file.py  
spark-3.4.1-bin-hadoop3/python/test\_support/sql/  
spark-3.4.1-bin-hadoop3/python/test\_support/sql/text-test.txt  
spark-3.4.1-bin-hadoop3/python/test\_support/sql/streaming/  
spark-3.4.1-bin-hadoop3/python/test\_support/sql/streaming/text-test.txt  
spark-3.4.1-bin-hadoop3/python/test\_support/sql/people\_array\_utf16le.json  
spark-3.4.1-bin-hadoop3/python/test\_support/sql/people\_array.json  
spark-3.4.1-bin-hadoop3/python/test\_support/sql/people1.json  
spark-3.4.1-bin-hadoop3/python/test\_support/sql/people.json  
spark-3.4.1-bin-hadoop3/python/test\_support/sql/orc\_partitioned/  
spark-3.4.1-bin-hadoop3/python/test\_support/sql/orc\_partitioned/b=1/  
spark-3.4.1-bin-hadoop3/python/test\_support/sql/orc\_partitioned/b=1/c=1/  
spark-3.4.1-bin-hadoop3/python/test\_support/sql/orc\_partitioned/b=1/c=1/part-r-00000  
-829af031-b970-49d6-ad39-30460a0be2c8.orc  
spark-3.4.1-bin-hadoop3/python/test\_support/sql/orc\_partitioned/b=1/c=1/.part-r-0000  
0-829af031-b970-49d6-ad39-30460a0be2c8.orc.crc  
spark-3.4.1-bin-hadoop3/python/test\_support/sql/orc\_partitioned/b=0/  
spark-3.4.1-bin-hadoop3/python/test\_support/sql/orc\_partitioned/b=0/c=0/  
spark-3.4.1-bin-hadoop3/python/test\_support/sql/orc\_partitioned/b=0/c=0/part-r-00000  
-829af031-b970-49d6-ad39-30460a0be2c8.orc  
spark-3.4.1-bin-hadoop3/python/test\_support/sql/orc\_partitioned/b=0/c=0/.part-r-0000  
0-829af031-b970-49d6-ad39-30460a0be2c8.orc.crc  
spark-3.4.1-bin-hadoop3/python/test\_support/sql/orc\_partitioned/\_SUCCESS  
spark-3.4.1-bin-hadoop3/python/test\_support/sql/ages\_newlines.csv  
spark-3.4.1-bin-hadoop3/python/test\_support/hello/  
spark-3.4.1-bin-hadoop3/python/test\_support/hello/sub\_hello/  
spark-3.4.1-bin-hadoop3/python/test\_support/hello/sub\_hello/sub\_hello.txt  
spark-3.4.1-bin-hadoop3/python/test\_support/hello/hello.txt

spark-3.4.1-bin-hadoop3/python/test\_support/SimpleHTTPServer.py  
spark-3.4.1-bin-hadoop3/python/test\_coverage/  
spark-3.4.1-bin-hadoop3/python/test\_coverage/sitecustomize.py  
spark-3.4.1-bin-hadoop3/python/test\_coverage/coverage\_daemon.py  
spark-3.4.1-bin-hadoop3/python/test\_coverage/conf/  
spark-3.4.1-bin-hadoop3/python/test\_coverage/conf/spark-defaults.conf  
spark-3.4.1-bin-hadoop3/python/setup.cfg  
spark-3.4.1-bin-hadoop3/python/run-tests.py  
spark-3.4.1-bin-hadoop3/python/run-tests-with-coverage  
spark-3.4.1-bin-hadoop3/python/docs/  
spark-3.4.1-bin-hadoop3/python/docs/source/  
spark-3.4.1-bin-hadoop3/python/docs/source/user\_guide/  
spark-3.4.1-bin-hadoop3/python/docs/source/user\_guide/python\_packaging.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/user\_guide/pandas\_on\_spark/  
spark-3.4.1-bin-hadoop3/python/docs/source/user\_guide/pandas\_on\_spark/types.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/user\_guide/pandas\_on\_spark/transform\_apply.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/user\_guide/pandas\_on\_spark/pandas\_pyspark.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/user\_guide/pandas\_on\_spark/options.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/user\_guide/pandas\_on\_spark/index.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/user\_guide/pandas\_on\_spark/from\_to\_dbms.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/user\_guide/pandas\_on\_spark/faq.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/user\_guide/pandas\_on\_spark/typehints.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/user\_guide/pandas\_on\_spark/best\_practices.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/user\_guide/index.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/user\_guide/arrow\_pandas.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/user\_guide/sql/  
spark-3.4.1-bin-hadoop3/python/docs/source/user\_guide/sql/index.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/user\_guide/sql/arrow\_pandas.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.streaming.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.ss/  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.ss/query\_management.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.ss/io.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.ss/index.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.ss/core\_classes.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.sql/  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.sql/window.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.sql/udf.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.sql/row.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.sql/protobuf.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.sql/observation.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.sql/io.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.sql/grouping.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.sql/configuration.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.sql/column.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.sql/catalog.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.sql/avro.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.sql/spark\_session.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.sql/index.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.sql/functions.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.sql/dataframe.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.sql/data\_types.rst

spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.sql/core\_classes.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.resource.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.pandas/  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.pandas/window.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.pandas/series.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.pandas/resampling.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.pandas/ml.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.pandas/io.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.pandas/indexing.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.pandas/index.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.pandas/groupby.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.pandas/general\_function  
s.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.pandas/extensions.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.pandas/frame.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.mllib.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.ml.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/pyspark.errors.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/reference/index.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/migration\_guide/  
spark-3.4.1-bin-hadoop3/python/docs/source/migration\_guide/koalas\_to\_pyspark.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/migration\_guide/index.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/migration\_guide/pyspark\_upgrade.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/index.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/getting\_started/  
spark-3.4.1-bin-hadoop3/python/docs/source/getting\_started/quickstart\_ps.ipynb  
spark-3.4.1-bin-hadoop3/python/docs/source/getting\_started/quickstart\_df.ipynb  
spark-3.4.1-bin-hadoop3/python/docs/source/getting\_started/quickstart\_connect.ipynb  
spark-3.4.1-bin-hadoop3/python/docs/source/getting\_started/index.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/getting\_started/install.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/development/  
spark-3.4.1-bin-hadoop3/python/docs/source/development/testing.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/development/setting\_ide.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/development/debugging.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/development/index.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/development/contributing.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/conf.py  
spark-3.4.1-bin-hadoop3/python/docs/source/\_templates/  
spark-3.4.1-bin-hadoop3/python/docs/source/\_templates/autosummary/  
spark-3.4.1-bin-hadoop3/python/docs/source/\_templates/autosummary/class\_with\_docs.rs  
t  
spark-3.4.1-bin-hadoop3/python/docs/source/\_templates/autosummary/class.rst  
spark-3.4.1-bin-hadoop3/python/docs/source/\_static/  
spark-3.4.1-bin-hadoop3/python/docs/source/\_static/css/  
spark-3.4.1-bin-hadoop3/python/docs/source/\_static/css/pyspark.css  
spark-3.4.1-bin-hadoop3/python/docs/source/\_static/copybutton.js  
spark-3.4.1-bin-hadoop3/python/docs/make2.bat  
spark-3.4.1-bin-hadoop3/python/docs/make.bat  
spark-3.4.1-bin-hadoop3/python/docs/Makefile  
spark-3.4.1-bin-hadoop3/python/README.md  
spark-3.4.1-bin-hadoop3/python/MANIFEST.in  
spark-3.4.1-bin-hadoop3/python/.gitignore  
spark-3.4.1-bin-hadoop3/python/.coveragerc  
spark-3.4.1-bin-hadoop3/python/mypy.ini  
spark-3.4.1-bin-hadoop3/python/setup.py

spark-3.4.1-bin-hadoop3/python/run-tests  
spark-3.4.1-bin-hadoop3/python/pyspark/  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/torch/  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/torch/distributor.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/torch/torch\_run\_process\_wrapper.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/torch/tests/  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/torch/tests/test\_log\_communication.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/torch/tests/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/torch/tests/test\_distributor.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/torch/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/torch/log\_communication.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/tests/  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/tests/typing/  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/tests/typing/test\_regression.yml  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/tests/typing/test\_readable.yml  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/tests/typing/test\_param.yml  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/tests/typing/test\_feature.yml  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/tests/typing/test\_evaluation.yml  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/tests/typing/test\_clustering.yml  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/tests/typing/test\_classification.yml  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/tests/test\_wrapper.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/tests/test\_util.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/tests/test\_training\_summary.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/tests/test\_stat.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/tests/test\_pipeline.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/tests/test\_persistence.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/tests/test\_param.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/tests/test\_model\_cache.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/tests/test\_linalg.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/tests/test\_image.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/tests/test\_feature.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/tests/test\_evaluation.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/tests/test\_base.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/tests/test\_algorithms.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/tests/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/tests/test\_tuning.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/tests/test\_functions.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/regression.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/recommendation.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/util.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/stat.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/functions.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/fpm.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/feature.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/pipeline.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/param/  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/param/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/param/shared.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/param/\_shared\_params\_code\_gen.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/model\_cache.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/linalg/  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/linalg/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/image.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/evaluation.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/common.py

spark-3.4.1-bin-hadoop3/python/pyspark/ml/clustering.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/classification.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/base.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/\_typing.pyi  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/wrapper.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/tuning.py  
spark-3.4.1-bin-hadoop3/python/pyspark/ml/tree.py  
spark-3.4.1-bin-hadoop3/python/pyspark/join.py  
spark-3.4.1-bin-hadoop3/python/pyspark/instrumentation\_utils.py  
spark-3.4.1-bin-hadoop3/python/pyspark/find\_spark\_home.py  
spark-3.4.1-bin-hadoop3/python/pyspark/files.py  
spark-3.4.1-bin-hadoop3/python/pyspark/errors/  
spark-3.4.1-bin-hadoop3/python/pyspark/errors/tests/  
spark-3.4.1-bin-hadoop3/python/pyspark/errors/tests/test\_errors.py  
spark-3.4.1-bin-hadoop3/python/pyspark/errors/tests/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/errors/exceptions/  
spark-3.4.1-bin-hadoop3/python/pyspark/errors/exceptions/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/errors/exceptions/connect.py  
spark-3.4.1-bin-hadoop3/python/pyspark/errors/exceptions/captured.py  
spark-3.4.1-bin-hadoop3/python/pyspark/errors/exceptions/base.py  
spark-3.4.1-bin-hadoop3/python/pyspark/errors/utils.py  
spark-3.4.1-bin-hadoop3/python/pyspark/errors/error\_classes.py  
spark-3.4.1-bin-hadoop3/python/pyspark/errors/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/daemon.py  
spark-3.4.1-bin-hadoop3/python/pyspark/cloudpickle/  
spark-3.4.1-bin-hadoop3/python/pyspark/cloudpickle/compat.py  
spark-3.4.1-bin-hadoop3/python/pyspark/cloudpickle/cloudpickle\_fast.py  
spark-3.4.1-bin-hadoop3/python/pyspark/cloudpickle/cloudpickle.py  
spark-3.4.1-bin-hadoop3/python/pyspark/cloudpickle/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/\_typing.pyi  
spark-3.4.1-bin-hadoop3/python/pyspark/\_globals.py  
spark-3.4.1-bin-hadoop3/python/pyspark/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/python/  
spark-3.4.1-bin-hadoop3/python/pyspark/python/pyspark/  
spark-3.4.1-bin-hadoop3/python/pyspark/python/pyspark/shell.py  
spark-3.4.1-bin-hadoop3/python/pyspark/\_\_pycache\_\_/  
spark-3.4.1-bin-hadoop3/python/pyspark/\_\_pycache\_\_/install.cpython-38.pyc  
spark-3.4.1-bin-hadoop3/python/pyspark/traceback\_utils.py  
spark-3.4.1-bin-hadoop3/python/pyspark/tests/  
spark-3.4.1-bin-hadoop3/python/pyspark/tests/typing/  
spark-3.4.1-bin-hadoop3/python/pyspark/tests/typing/test\_resultiterable.yml  
spark-3.4.1-bin-hadoop3/python/pyspark/tests/typing/test\_rdd.yml  
spark-3.4.1-bin-hadoop3/python/pyspark/tests/typing/test\_core.yml  
spark-3.4.1-bin-hadoop3/python/pyspark/tests/typing/test\_context.yml  
spark-3.4.1-bin-hadoop3/python/pyspark/tests/test\_worker.py  
spark-3.4.1-bin-hadoop3/python/pyspark/tests/test\_util.py  
spark-3.4.1-bin-hadoop3/python/pyspark/tests/test\_taskcontext.py  
spark-3.4.1-bin-hadoop3/python/pyspark/tests/test\_statcounter.py  
spark-3.4.1-bin-hadoop3/python/pyspark/tests/test\_stage\_sched.py  
spark-3.4.1-bin-hadoop3/python/pyspark/tests/test\_shuffle.py  
spark-3.4.1-bin-hadoop3/python/pyspark/tests/test\_serializers.py  
spark-3.4.1-bin-hadoop3/python/pyspark/tests/test\_readwrite.py  
spark-3.4.1-bin-hadoop3/python/pyspark/tests/test\_rddsampler.py  
spark-3.4.1-bin-hadoop3/python/pyspark/tests/test\_rddbarrier.py  
spark-3.4.1-bin-hadoop3/python/pyspark/tests/test\_rdd.py

spark-3.4.1-bin-hadoop3/python/pyspark/tests/test\_pin\_thread.py  
spark-3.4.1-bin-hadoop3/python/pyspark/tests/test\_memory\_profiler.py  
spark-3.4.1-bin-hadoop3/python/pyspark/tests/test\_join.py  
spark-3.4.1-bin-hadoop3/python/pyspark/tests/test\_daemon.py  
spark-3.4.1-bin-hadoop3/python/pyspark/tests/test\_conf.py  
spark-3.4.1-bin-hadoop3/python/pyspark/tests/test\_broadcast.py  
spark-3.4.1-bin-hadoop3/python/pyspark/tests/test\_appsubmit.py  
spark-3.4.1-bin-hadoop3/python/pyspark/tests/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/tests/test\_profiler.py  
spark-3.4.1-bin-hadoop3/python/pyspark/tests/test\_install\_spark.py  
spark-3.4.1-bin-hadoop3/python/pyspark/tests/test\_context.py  
spark-3.4.1-bin-hadoop3/python/pyspark/testing/  
spark-3.4.1-bin-hadoop3/python/pyspark/testing/utils.py  
spark-3.4.1-bin-hadoop3/python/pyspark/testing/streamingutils.py  
spark-3.4.1-bin-hadoop3/python/pyspark/testing/sqlutils.py  
spark-3.4.1-bin-hadoop3/python/pyspark/testing/pandasutils.py  
spark-3.4.1-bin-hadoop3/python/pyspark/testing/mlutils.py  
spark-3.4.1-bin-hadoop3/python/pyspark/testing/mllibutils.py  
spark-3.4.1-bin-hadoop3/python/pyspark/testing/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/testing/connectutils.py  
spark-3.4.1-bin-hadoop3/python/pyspark/streaming/  
spark-3.4.1-bin-hadoop3/python/pyspark/streaming/util.py  
spark-3.4.1-bin-hadoop3/python/pyspark/streaming/tests/  
spark-3.4.1-bin-hadoop3/python/pyspark/streaming/tests/test\_listener.py  
spark-3.4.1-bin-hadoop3/python/pyspark/streaming/tests/test\_kinesis.py  
spark-3.4.1-bin-hadoop3/python/pyspark/streaming/tests/test\_dstream.py  
spark-3.4.1-bin-hadoop3/python/pyspark/streaming/tests/test\_context.py  
spark-3.4.1-bin-hadoop3/python/pyspark/streaming/tests/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/streaming/listener.py  
spark-3.4.1-bin-hadoop3/python/pyspark/streaming/kinesis.py  
spark-3.4.1-bin-hadoop3/python/pyspark/streaming/dstream.py  
spark-3.4.1-bin-hadoop3/python/pyspark/streaming/context.py  
spark-3.4.1-bin-hadoop3/python/pyspark/streaming/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/storagelevel.py  
spark-3.4.1-bin-hadoop3/python/pyspark/status.py  
spark-3.4.1-bin-hadoop3/python/pyspark/statcounter.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/plot/  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/plot/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/plot/test\_series\_plot\_plotly.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/plot/test\_series\_plot\_matplotlib.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/plot/test\_series\_plot.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/plot/test\_frame\_plot\_plotly.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/plot/test\_frame\_plot\_matplotlib.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/plot/test\_frame\_plot.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/indexes/  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/indexes/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/indexes/test\_timedelta.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/indexes/test\_datetime.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/indexes/test\_category.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/indexes/test\_base.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/data\_type\_ops/  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/data\_type\_ops/testing\_utils.py



spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/data\_type\_ops/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/data\_type\_ops/test\_udt\_ops.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/data\_type\_ops/test\_timedelta\_ops.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/data\_type\_ops/test\_string\_ops.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/data\_type\_ops/test\_num\_ops.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/data\_type\_ops/test\_null\_ops.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/data\_type\_ops/test\_datetime\_ops.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/data\_type\_ops/test\_date\_ops.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/data\_type\_ops/test\_complex\_ops.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/data\_type\_ops/test\_categorical\_ops.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/data\_type\_ops/test\_boolean\_ops.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/data\_type\_ops/test\_binary\_ops.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/data\_type\_ops/test\_base.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_window.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_utils.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_typedef.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_stats.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_sql.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_spark\_functions.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_series\_string.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_series\_datetime.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_series\_conversion.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_series.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_scalars.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_rolling.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_reshape.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_resample.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_repr.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_ops\_on\_diff\_frames\_slow.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_ops\_on\_diff\_frames\_groupby\_rolling.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_ops\_on\_diff\_frames\_groupby\_expanding.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_ops\_on\_diff\_frames\_groupby.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_ops\_on\_diff\_frames.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_numpy\_compat.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_namespace.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_internal.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_indexops\_spark.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_indexing.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_groupby\_slow.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_groupby.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_generic\_functions.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_frame\_spark.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_extension.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_expanding.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_ewm.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_default\_index.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_dataframe\_spark\_io.py

spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_dataframe\_slow.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_dataframe\_conversion.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_dataframe.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_csv.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_config.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/tests/test\_categorical.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/spark/  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/spark/utils.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/spark/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/spark/functions.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/spark/accessors.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/resample.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/plot/  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/plot/plotly.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/plot/core.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/plot/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/plot/matplotlib.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/mlflow.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/missing/  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/missing/window.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/missing/series.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/missing/scalars.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/missing/resample.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/missing/indexes.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/missing/groupby.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/missing/general\_functions.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/missing/frame.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/missing/common.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/missing/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/indexes/  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/indexes/timedelta.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/indexes/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/indexes/numeric.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/indexes/multi.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/indexes/datetime.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/indexes/category.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/indexes/base.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/extensions.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/exceptions.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/datetime.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/data\_type\_ops/  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/data\_type\_ops/udt\_ops.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/data\_type\_ops/complex\_ops.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/data\_type\_ops/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/data\_type\_ops/timedelta\_ops.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/data\_type\_ops/string\_ops.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/data\_type\_ops/num\_ops.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/data\_type\_ops/null\_ops.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/data\_type\_ops/datetime\_ops.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/data\_type\_ops/date\_ops.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/data\_type\_ops/categorical\_ops.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/data\_type\_ops/boolean\_ops.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/data\_type\_ops/binary\_ops.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/data\_type\_ops/base.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/correlation.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/categorical.py

spark-3.4.1-bin-hadoop3/python/pyspark/pandas/accessors.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/window.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/usage\_logging/  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/usage\_logging/usage\_logger.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/usage\_logging/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/typedef/  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/typedef/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/typedef/typehints.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/utils.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/supported\_api\_gen.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/strings.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/sql\_processor.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/sql\_formatter.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/series.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/numpy\_compat.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/namespace.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/internal.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/indexing.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/groupby.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/generic.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/frame.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/config.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/base.py  
spark-3.4.1-bin-hadoop3/python/pyspark/pandas/\_typing.py  
spark-3.4.1-bin-hadoop3/python/pyspark/mllib/  
spark-3.4.1-bin-hadoop3/python/pyspark/mllib/util.py  
spark-3.4.1-bin-hadoop3/python/pyspark/mllib/tree.py  
spark-3.4.1-bin-hadoop3/python/pyspark/mllib/tests/  
spark-3.4.1-bin-hadoop3/python/pyspark/mllib/tests/test\_util.py  
spark-3.4.1-bin-hadoop3/python/pyspark/mllib/tests/test\_streaming\_algorithms.py  
spark-3.4.1-bin-hadoop3/python/pyspark/mllib/tests/test\_stat.py  
spark-3.4.1-bin-hadoop3/python/pyspark/mllib/tests/test\_linalg.py  
spark-3.4.1-bin-hadoop3/python/pyspark/mllib/tests/test\_feature.py  
spark-3.4.1-bin-hadoop3/python/pyspark/mllib/tests/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/mllib/tests/test\_algorithms.py  
spark-3.4.1-bin-hadoop3/python/pyspark/mllib/stat/  
spark-3.4.1-bin-hadoop3/python/pyspark/mllib/stat/test.py  
spark-3.4.1-bin-hadoop3/python/pyspark/mllib/stat/distribution.py  
spark-3.4.1-bin-hadoop3/python/pyspark/mllib/stat/\_statistics.py  
spark-3.4.1-bin-hadoop3/python/pyspark/mllib/stat/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/mllib/stat/KernelDensity.py  
spark-3.4.1-bin-hadoop3/python/pyspark/mllib/regression.py  
spark-3.4.1-bin-hadoop3/python/pyspark/mllib/recommendation.py  
spark-3.4.1-bin-hadoop3/python/pyspark/mllib/random.py  
spark-3.4.1-bin-hadoop3/python/pyspark/mllib/linalg/  
spark-3.4.1-bin-hadoop3/python/pyspark/mllib/linalg/distributed.py  
spark-3.4.1-bin-hadoop3/python/pyspark/mllib/linalg/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/mllib/fpm.py  
spark-3.4.1-bin-hadoop3/python/pyspark/mllib/feature.py  
spark-3.4.1-bin-hadoop3/python/pyspark/mllib/evaluation.py  
spark-3.4.1-bin-hadoop3/python/pyspark/mllib/common.py  
spark-3.4.1-bin-hadoop3/python/pyspark/mllib/classification.py  
spark-3.4.1-bin-hadoop3/python/pyspark/mllib/\_typing.pyi  
spark-3.4.1-bin-hadoop3/python/pyspark/mllib/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/mllib/clustering.py

spark-3.4.1-bin-hadoop3/python/pyspark/sql/  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/window.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/typing/  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/typing/test\_udf.yml  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/typing/test\_session.yml  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/typing/test\_readwriter.yml  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/typing/test\_functions.yml  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/typing/test\_dataframe.yml  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/typing/test\_column.yml  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/test\_utils.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/test\_udf\_profiler.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/test\_session.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/test\_serde.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/test\_readwriter.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/test\_pandas\_sqlmetrics.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/test\_group.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/test\_errors.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/test\_datasources.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/test\_context.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/test\_conf.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/streaming/  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/streaming/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/streaming/test\_streaming\_listener.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/streaming/test\_streaming.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/pandas/  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/pandas/test\_pandas\_udf\_typehints\_with\_future\_annotations.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/pandas/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/pandas/test\_pandas\_udf\_window.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/pandas/test\_pandas\_udf\_typehints.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/pandas/test\_pandas\_udf\_scalar.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/pandas/test\_pandas\_udf\_grouped\_agg.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/pandas/test\_pandas\_udf.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/pandas/test\_pandas\_map.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/pandas/test\_pandas\_grouped\_map\_with\_state.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/pandas/test\_pandas\_grouped\_map.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/pandas/test\_pandas\_cogrouped\_map.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/connect/  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/connect/test\_parity\_udf.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/connect/test\_parity\_serde.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/connect/test\_parity\_readwriter.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/connect/test\_parity\_pandas\_map.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/connect/test\_parity\_group.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/connect/test\_parity\_functions.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/connect/test\_parity\_errors.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/connect/test\_parity\_datasources.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/connect/test\_parity\_conf.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/connect/test\_parity\_column.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/connect/test\_parity\_catalog.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/connect/test\_parity\_arrow\_map.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/connect/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/connect/test\_parity\_types.py

spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/connect/test\_parity\_pandas\_udf.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/connect/test\_parity\_pandas\_grouped\_map.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/connect/test\_parity\_pandas\_cogrouped\_map.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/connect/test\_parity\_dataframe.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/connect/test\_parity\_arrow.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/connect/test\_connect\_plan.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/connect/test\_connect\_function.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/connect/test\_connect\_column.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/connect/test\_connect\_basic.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/connect/test\_client.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/test\_udf.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/test\_types.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/test\_functions.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/test\_dataframe.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/test\_column.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/test\_catalog.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/test\_arrow\_python\_udf.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/test\_arrow\_map.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/tests/test\_arrow.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/streaming/  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/streaming/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/streaming/state.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/streaming/readwriter.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/streaming/query.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/streaming/listener.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/sql\_formatter.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/protobuf/  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/protobuf/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/protobuf/functions.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/pandas/  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/pandas/utils.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/pandas/functions.pyi  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/pandas/\_typing/  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/pandas/\_typing/protocols/  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/pandas/\_typing/protocols/series.pyi  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/pandas/\_typing/protocols/frame.pyi  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/pandas/\_typing/protocols/\_\_init\_\_.pyi  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/pandas/\_typing/\_\_init\_\_.pyi  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/pandas/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/pandas/types.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/pandas/typehints.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/pandas/serializers.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/pandas/map\_ops.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/pandas/group\_ops.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/pandas/functions.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/pandas/conversion.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/observation.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/group.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/utils.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/proto/  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/proto/types\_pb2.pyi  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/proto/example\_plugins\_pb2.pyi

spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/proto/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/proto/types\_pb2.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/proto/relations\_pb2.pyi  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/proto/relations\_pb2.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/proto/expressions\_pb2.pyi  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/proto/expressions\_pb2.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/proto/example\_plugins\_pb2.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/proto/common\_pb2.pyi  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/proto/common\_pb2.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/proto/commands\_pb2.pyi  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/proto/commands\_pb2.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/proto/catalog\_pb2.pyi  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/proto/catalog\_pb2.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/proto/base\_pb2\_grpc.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/proto/base\_pb2.pyi  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/proto/base\_pb2.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/conf.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/window.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/udf.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/types.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/session.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/readwriter.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/plan.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/group.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/functions.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/expressions.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/dataframe.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/conversion.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/column.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/client.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/catalog.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/connect/\_typing.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/conf.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/avro/  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/avro/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/avro/functions.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/utils.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/udf.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/types.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/session.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/readwriter.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/functions.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/dataframe.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/context.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/column.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/catalog.py  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/\_typing.pyi  
spark-3.4.1-bin-hadoop3/python/pyspark/sql/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/shuffle.py  
spark-3.4.1-bin-hadoop3/python/pyspark/resultiterable.py  
spark-3.4.1-bin-hadoop3/python/pyspark/resource/  
spark-3.4.1-bin-hadoop3/python/pyspark/resource/tests/  
spark-3.4.1-bin-hadoop3/python/pyspark/resource/tests/test\_resources.py  
spark-3.4.1-bin-hadoop3/python/pyspark/resource/tests/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/resource/requests.py

spark-3.4.1-bin-hadoop3/python/pyspark/resource/profile.py  
spark-3.4.1-bin-hadoop3/python/pyspark/resource/information.py  
spark-3.4.1-bin-hadoop3/python/pyspark/resource/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/python/pyspark/rddsampler.py  
spark-3.4.1-bin-hadoop3/python/pyspark/py.typed  
spark-3.4.1-bin-hadoop3/python/pyspark/rdd.py  
spark-3.4.1-bin-hadoop3/python/pyspark/profiler.py  
spark-3.4.1-bin-hadoop3/python/pyspark/java\_gateway.py  
spark-3.4.1-bin-hadoop3/python/pyspark/install.py  
spark-3.4.1-bin-hadoop3/python/pyspark/context.py  
spark-3.4.1-bin-hadoop3/python/pyspark/conf.py  
spark-3.4.1-bin-hadoop3/python/pyspark/broadcast.py  
spark-3.4.1-bin-hadoop3/python/pyspark/accumulators.py  
spark-3.4.1-bin-hadoop3/python/pyspark/worker.py  
spark-3.4.1-bin-hadoop3/python/pyspark/version.py  
spark-3.4.1-bin-hadoop3/python/pyspark/util.py  
spark-3.4.1-bin-hadoop3/python/pyspark/taskcontext.py  
spark-3.4.1-bin-hadoop3/python/pyspark/shell.py  
spark-3.4.1-bin-hadoop3/python/pyspark/serializers.py  
spark-3.4.1-bin-hadoop3/python/lib/  
spark-3.4.1-bin-hadoop3/python/lib/pyspark.zip  
spark-3.4.1-bin-hadoop3/python/lib/py4j-0.10.9.7-src.zip  
spark-3.4.1-bin-hadoop3/python/lib/PY4J\_LICENSE.txt  
spark-3.4.1-bin-hadoop3/bin/  
spark-3.4.1-bin-hadoop3/bin/sparkR2.cmd  
spark-3.4.1-bin-hadoop3/bin/sparkR.cmd  
spark-3.4.1-bin-hadoop3/bin/sparkR  
spark-3.4.1-bin-hadoop3/bin/spark-submit2.cmd  
spark-3.4.1-bin-hadoop3/bin/spark-submit.cmd  
spark-3.4.1-bin-hadoop3/bin/spark-submit  
spark-3.4.1-bin-hadoop3/bin/spark-sql2.cmd  
spark-3.4.1-bin-hadoop3/bin/spark-sql.cmd  
spark-3.4.1-bin-hadoop3/bin/spark-sql  
spark-3.4.1-bin-hadoop3/bin/spark-shell2.cmd  
spark-3.4.1-bin-hadoop3/bin/spark-shell.cmd  
spark-3.4.1-bin-hadoop3/bin/spark-shell  
spark-3.4.1-bin-hadoop3/bin/spark-connect-shell  
spark-3.4.1-bin-hadoop3/bin/spark-class2.cmd  
spark-3.4.1-bin-hadoop3/bin/spark-class.cmd  
spark-3.4.1-bin-hadoop3/bin/spark-class  
spark-3.4.1-bin-hadoop3/bin/run-example.cmd  
spark-3.4.1-bin-hadoop3/bin/run-example  
spark-3.4.1-bin-hadoop3/bin/pyspark2.cmd  
spark-3.4.1-bin-hadoop3/bin/pyspark.cmd  
spark-3.4.1-bin-hadoop3/bin/pyspark  
spark-3.4.1-bin-hadoop3/bin/load-spark-env.sh  
spark-3.4.1-bin-hadoop3/bin/load-spark-env.cmd  
spark-3.4.1-bin-hadoop3/bin/find-spark-home.cmd  
spark-3.4.1-bin-hadoop3/bin/find-spark-home  
spark-3.4.1-bin-hadoop3/bin/docker-image-tool.sh  
spark-3.4.1-bin-hadoop3/bin/beeline.cmd  
spark-3.4.1-bin-hadoop3/bin/beeline  
spark-3.4.1-bin-hadoop3/README.md  
spark-3.4.1-bin-hadoop3/conf/  
spark-3.4.1-bin-hadoop3/conf/workers.template  
spark-3.4.1-bin-hadoop3/conf/spark-env.sh.template

spark-3.4.1-bin-hadoop3/conf/spark-defaults.conf.template  
spark-3.4.1-bin-hadoop3/conf/metrics.properties.template  
spark-3.4.1-bin-hadoop3/conf/log4j2.properties.template  
spark-3.4.1-bin-hadoop3/conf/fairscheduler.xml.template  
spark-3.4.1-bin-hadoop3/data/  
spark-3.4.1-bin-hadoop3/data/streaming/  
spark-3.4.1-bin-hadoop3/data/streaming/AFINN-111.txt  
spark-3.4.1-bin-hadoop3/data/mllib/  
spark-3.4.1-bin-hadoop3/data/mllib/streaming\_kmeans\_data\_test.txt  
spark-3.4.1-bin-hadoop3/data/mllib/sample\_svm\_data.txt  
spark-3.4.1-bin-hadoop3/data/mllib/sample\_multiclass\_classification\_data.txt  
spark-3.4.1-bin-hadoop3/data/mllib/sample\_movielens\_data.txt  
spark-3.4.1-bin-hadoop3/data/mllib/sample\_linear\_regression\_data.txt  
spark-3.4.1-bin-hadoop3/data/mllib/sample\_libsvm\_data.txt  
spark-3.4.1-bin-hadoop3/data/mllib/sample\_lda\_libsvm\_data.txt  
spark-3.4.1-bin-hadoop3/data/mllib/sample\_lda\_data.txt  
spark-3.4.1-bin-hadoop3/data/mllib/sample\_kmeans\_data.txt  
spark-3.4.1-bin-hadoop3/data/mllib/sample\_isotonic\_regression\_libsvm\_data.txt  
spark-3.4.1-bin-hadoop3/data/mllib/sample\_fpgrowth.txt  
spark-3.4.1-bin-hadoop3/data/mllib/sample\_binary\_classification\_data.txt  
spark-3.4.1-bin-hadoop3/data/mllib/ridge-data/  
spark-3.4.1-bin-hadoop3/data/mllib/ridge-data/lpsa.data  
spark-3.4.1-bin-hadoop3/data/mllib/pic\_data.txt  
spark-3.4.1-bin-hadoop3/data/mllib/pagerank\_data.txt  
spark-3.4.1-bin-hadoop3/data/mllib/kmeans\_data.txt  
spark-3.4.1-bin-hadoop3/data/mllib/images/  
spark-3.4.1-bin-hadoop3/data/mllib/images/origin/  
spark-3.4.1-bin-hadoop3/data/mllib/images/origin/multi-channel/  
spark-3.4.1-bin-hadoop3/data/mllib/images/origin/multi-channel/grayscale.jpg  
spark-3.4.1-bin-hadoop3/data/mllib/images/origin/multi-channel/chr30.4.184.jpg  
spark-3.4.1-bin-hadoop3/data/mllib/images/origin/multi-channel/BGRA\_alpha\_60.png  
spark-3.4.1-bin-hadoop3/data/mllib/images/origin/multi-channel/BGRA.png  
spark-3.4.1-bin-hadoop3/data/mllib/images/origin/license.txt  
spark-3.4.1-bin-hadoop3/data/mllib/images/origin/kittens/  
spark-3.4.1-bin-hadoop3/data/mllib/images/origin/kittens/not-image.txt  
spark-3.4.1-bin-hadoop3/data/mllib/images/origin/kittens/DP802813.jpg  
spark-3.4.1-bin-hadoop3/data/mllib/images/origin/kittens/DP153539.jpg  
spark-3.4.1-bin-hadoop3/data/mllib/images/origin/kittens/54893.jpg  
spark-3.4.1-bin-hadoop3/data/mllib/images/origin/kittens/29.5.a\_b\_EGDP022204.jpg  
spark-3.4.1-bin-hadoop3/data/mllib/images/license.txt  
spark-3.4.1-bin-hadoop3/data/mllib/gmm\_data.txt  
spark-3.4.1-bin-hadoop3/data/mllib/als/  
spark-3.4.1-bin-hadoop3/data/mllib/als/test.data  
spark-3.4.1-bin-hadoop3/data/mllib/als/sample\_movielens\_ratings.txt  
spark-3.4.1-bin-hadoop3/data/graphx/  
spark-3.4.1-bin-hadoop3/data/graphx/users.txt  
spark-3.4.1-bin-hadoop3/data/graphx/followers.txt  
spark-3.4.1-bin-hadoop3/NOTICE  
spark-3.4.1-bin-hadoop3/licenses/  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-zstd.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-zstd-jni.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-xmlenc.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-vis-timeline.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-spire.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-sorttable.js.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-slf4j.txt



spark-3.4.1-bin-hadoop3/licenses/LICENSE-scopt.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-sbt-launch-lib.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-respond.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-reflectasm.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-re2j.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-pyrolite.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-py4j.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-protobuf.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-pmml-model.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-paranamer.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-mustache.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-modernizr.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-minlog.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-matchMedia-polyfill.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-machinist.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-leveldbjni.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-kryo.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-jsp-api.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-json-formatter.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-jquery.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-join.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-jodd.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-jline.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-jaxb-runtime.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-javolution.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-javax-transaction-transaction-api.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-javassist.html  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-janino.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-jakarta.xml.bind-api.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-jakarta.activation-api.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-jakarta.ws.rs-api  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-jakarta.annotation-api  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-istack-commons-runtime.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-graphlib-dot.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-f2j.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-dnsjava.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-datatables.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-dagre-d3.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-d3.min.js.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-cloudpickle.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-bootstrap.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-blas.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-automaton.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-arpack.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-antlr.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-JTransforms.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-JLargeArrays.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-CC0.txt  
spark-3.4.1-bin-hadoop3/licenses/LICENSE-AnchorJS.txt  
spark-3.4.1-bin-hadoop3/LICENSE  
spark-3.4.1-bin-hadoop3/examples/  
spark-3.4.1-bin-hadoop3/examples/src/  
spark-3.4.1-bin-hadoop3/examples/src/main/  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/

spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/GroupByTes  
t.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ExceptionH  
andlingTest.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/DriverSubm  
issionTest.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/DFSReadWri  
teTest.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/BroadcastT  
est.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/Accumulato  
rMetricsTest.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/streaming/  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/streaming/  
clickstream/  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/streaming/  
clickstream/PageViewStream.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/streaming/  
clickstream/PageViewGenerator.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/streaming/  
StreamingExamples.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/streaming/  
StatefulNetworkWordCount.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/streaming/  
SqlNetworkWordCount.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/streaming/  
RecoverableNetworkWordCount.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/streaming/  
RawNetworkGrep.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/streaming/  
QueueStream.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/streaming/  
NetworkWordCount.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/streaming/  
HdfsWordCount.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/streaming/  
DirectKerberizedKafkaWordCount.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/streaming/  
DirectKafkaWordCount.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/streaming/  
CustomReceiver.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/stream  
ing/  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/stream  
ing/StructuredSessionization.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/stream  
ing/StructuredNetworkWordCountWindowed.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/stream  
ing/StructuredNetworkWordCount.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/stream  
ing/StructuredKerberizedKafkaWordCount.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/stream  
ing/StructuredKafkaWordCount.scala

spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/streaming/StructuredComplexSessionization.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/jdbc/spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/jdbc/ExampleJdbcConnectionProvider.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/hive/spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/hive/SparkHiveExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/UserDefinedUntypedAggregation.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/UserDefinedTypedAggregation.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/UserDefinedScalar.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/SparkSQLExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/SimpleTypedAggregator.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/SQLDataSourceExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/RDDRelation.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/pythonconverters/spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/pythonconverters/AvroConverters.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/SparseNaiveBayes.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/SimpleFPGrowth.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/SampledRDDs.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/SVMWithSGDExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/SVDExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/RecommendationExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/RankingMetricsExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/RandomRDDGeneration.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/RandomForestRegressionExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/RandomForestClassificationExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/PrefixSpanExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/PowerIterationClusteringExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/PMMLModelExportExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/PCANSourceVectorExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/PCAO

nRowMatrixExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/NormalizerExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/NaiveBayesExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/MultivariateSummarizer.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/MulticlassMetricsExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/MultilabelMetricsExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/MovieLensALS.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/LogisticRegressionWithLBFGSExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/LatentDirichletAllocationExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/LDAExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/LBFGSExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/KernelDensityEstimationExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/KMeansExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/IsotonicRegressionExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/HypothesisTestingKolmogorovSmirnovTestExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/HypothesisTestingExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/GradientBoostingRegressionExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/GradientBoostingClassificationExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/GradientBoostedTreesRunner.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/GaussianMixtureExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/FPGrowthExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/ElementwiseProductExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/DenseKMeans.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/DecisionTreeRunner.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/DecisionTreeRegressionExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/DecisionTreeClassificationExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/CosineSimilarity.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/CorrelationsExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/Corr

relations.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/ChiSqSelectorExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/BisectingKMeansExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/BinaryClassificationMetricsExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/BinaryClassification.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/AssociationRulesExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/AbstractParams.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/Word2VecExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/TallSkinnySVD.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/TallSkinnyPCA.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/TFIDFExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/SummaryStatisticsExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/StreamingTestExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/StreamingLogisticRegression.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/StreamingLinearRegressionExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/StreamingKMeansExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/StratifiedSamplingExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/StandardScalerExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Word2VecExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/VectorSlicerExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/VectorSizeHintExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/VectorIndexerExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/VectorAssemblerExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/VariantThresholdSelectorExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/UnivariateFeatureSelectorExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/UnaryTransformerExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/TokenizerExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/TfIdfExample.scala

spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/SummarizerExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/StringIndexerExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/StopWordsRemoverExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/StandardScalerExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/SQLTransformerExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/RobustScalerExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/RandomForestRegressorExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/RandomForestExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/RandomForestClassifierExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/RFormulaExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/QuantileDiscretizerExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/PrefixSpanExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/PowerIterationClusteringExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/PolynomialExpansionExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/PipelineExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/PCAExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/OneVsRestExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/OneHotEncoderExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/NormalizerExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/NaiveBayesExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/NGramExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/MultilayerPerceptronClassifierExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/MulticlassLogisticRegressionWithElasticNetExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/ModelSelectionViaTrainValidationSplitExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/ModelSelectionViaCrossValidationExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/MinMaxScalerExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/MinHashLSHExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/MaxAbsScalerExample.scala

spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/LogisticRegressionWithElasticNetExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/LogisticRegressionSummaryExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/LogisticRegressionExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/LinearSVCExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/LinearRegressionWithElasticNetExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/LinearRegressionExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/LDAExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/KMeansExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/IsotonicRegressionExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/InteractionExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/IndexToStringExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/ImputerExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/GradientBoostedTreeRegressorExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/GradientBoostedTreeClassifierExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/GeneralizedLinearRegressionExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/GaussianMixtureExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/GBTEExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/FeatureHasherExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/FPGrowthExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/FMRegressorExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/FMClassifierExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/EstimatorTransformerParamExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/ElementwiseProductExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/DeveloperApiExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/DecisionTreeRegressionExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/DecisionTreeExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/DecisionTreeClassificationExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/DataFrameExample.scala

spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/DCTExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/CountVectorizerExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/CorrelationExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/ChiSquareTestExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/ChiSqSelectorExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/BucketizerExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/BucketedRandomProjectionLSHExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/BisectingKMeansExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/BinarizerExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/ALSExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/AFTSurvivalRegressionExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/graphx/  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/graphx/TriangleCountingExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/graphx/SynthBenchmark.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/graphx/SSSPExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/graphx/PageRankExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/graphx/LiveJournalPageRank.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/graphx/ConnectedComponentsExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/graphx/ComprehensiveExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/graphx/Analytics.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/graphx/AggregateMessagesExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/extensions/  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/extensions/SparkSessionExtensionsTest.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/extensions/SessionExtensionsWithoutLoader.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/extensions/SessionExtensionsWithLoader.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/extensions/AgeExample.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/SparkTC.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/SparkRemoteFileTest.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/SparkPi.sc



ala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/SparkPageRank.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/SparkLR.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/SparkKMeans.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/SparkHdfsLR.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/SparkALS.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/SkewedGroupByTest.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/SimpleSkewedGroupByTest.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/MultiBroadcastTest.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/MiniReadWriteTest.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/LogQuery.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/LocalPi.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/LocalLR.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/LocalKMeans.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/LocalFileLR.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/LocalALS.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/HdfsTest.scala  
spark-3.4.1-bin-hadoop3/examples/src/main/resources/  
spark-3.4.1-bin-hadoop3/examples/src/main/resources/users.parquet  
spark-3.4.1-bin-hadoop3/examples/src/main/resources/users.orc  
spark-3.4.1-bin-hadoop3/examples/src/main/resources/users.avro  
spark-3.4.1-bin-hadoop3/examples/src/main/resources/user.avsc  
spark-3.4.1-bin-hadoop3/examples/src/main/resources/people.txt  
spark-3.4.1-bin-hadoop3/examples/src/main/resources/people.json  
spark-3.4.1-bin-hadoop3/examples/src/main/resources/people.csv  
spark-3.4.1-bin-hadoop3/examples/src/main/resources/kv1.txt  
spark-3.4.1-bin-hadoop3/examples/src/main/resources/full\_user.avsc  
spark-3.4.1-bin-hadoop3/examples/src/main/resources/employees.json  
spark-3.4.1-bin-hadoop3/examples/src/main/resources/dir1/  
spark-3.4.1-bin-hadoop3/examples/src/main/resources/dir1/file3.json  
spark-3.4.1-bin-hadoop3/examples/src/main/resources/dir1/file1.parquet  
spark-3.4.1-bin-hadoop3/examples/src/main/resources/dir1/dir2/  
spark-3.4.1-bin-hadoop3/examples/src/main/resources/dir1/dir2/file2.parquet  
spark-3.4.1-bin-hadoop3/examples/src/main/resources/META-INF/  
spark-3.4.1-bin-hadoop3/examples/src/main/resources/META-INF/services/  
spark-3.4.1-bin-hadoop3/examples/src/main/resources/META-INF/services/org.apache.spark.sql.jdbc.JdbcConnectionProvider  
spark-3.4.1-bin-hadoop3/examples/src/main/resources/META-INF/services/org.apache.spark.sql.SessionExtensionsProvider  
spark-3.4.1-bin-hadoop3/examples/src/main/r/

spark-3.4.1-bin-hadoop3/examples/src/main/r/streaming/  
spark-3.4.1-bin-hadoop3/examples/src/main/r/streaming/structured\_network\_wordcount.R  
spark-3.4.1-bin-hadoop3/examples/src/main/r/ml/  
spark-3.4.1-bin-hadoop3/examples/src/main/r/ml/svmLinear.R  
spark-3.4.1-bin-hadoop3/examples/src/main/r/ml/survreg.R  
spark-3.4.1-bin-hadoop3/examples/src/main/r/ml/randomForest.R  
spark-3.4.1-bin-hadoop3/examples/src/main/r/ml/prefixSpan.R  
spark-3.4.1-bin-hadoop3/examples/src/main/r/ml/powerIterationClustering.R  
spark-3.4.1-bin-hadoop3/examples/src/main/r/ml/naiveBayes.R  
spark-3.4.1-bin-hadoop3/examples/src/main/r/ml/mlp.R  
spark-3.4.1-bin-hadoop3/examples/src/main/r/ml/ml.R  
spark-3.4.1-bin-hadoop3/examples/src/main/r/ml/logit.R  
spark-3.4.1-bin-hadoop3/examples/src/main/r/ml/lm\_with\_elastic\_net.R  
spark-3.4.1-bin-hadoop3/examples/src/main/r/ml/lda.R  
spark-3.4.1-bin-hadoop3/examples/src/main/r/ml/kstest.R  
spark-3.4.1-bin-hadoop3/examples/src/main/r/ml/kmeans.R  
spark-3.4.1-bin-hadoop3/examples/src/main/r/ml/isoreg.R  
spark-3.4.1-bin-hadoop3/examples/src/main/r/ml/glm.R  
spark-3.4.1-bin-hadoop3/examples/src/main/r/ml/gbt.R  
spark-3.4.1-bin-hadoop3/examples/src/main/r/ml/gaussianMixture.R  
spark-3.4.1-bin-hadoop3/examples/src/main/r/ml/fpm.R  
spark-3.4.1-bin-hadoop3/examples/src/main/r/ml/fmRegressor.R  
spark-3.4.1-bin-hadoop3/examples/src/main/r/ml/fmClassifier.R  
spark-3.4.1-bin-hadoop3/examples/src/main/r/ml/decisionTree.R  
spark-3.4.1-bin-hadoop3/examples/src/main/r/ml/bisectingKmeans.R  
spark-3.4.1-bin-hadoop3/examples/src/main/r/ml/als.R  
spark-3.4.1-bin-hadoop3/examples/src/main/r/dataframe.R  
spark-3.4.1-bin-hadoop3/examples/src/main/r/data-manipulation.R  
spark-3.4.1-bin-hadoop3/examples/src/main/r/RSparkSQLExample.R  
spark-3.4.1-bin-hadoop3/examples/src/main/python/  
spark-3.4.1-bin-hadoop3/examples/src/main/python/wordcount.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/transitive\_closure.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/streaming/  
spark-3.4.1-bin-hadoop3/examples/src/main/python/streaming/stateful\_network\_wordcount.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/streaming/sql\_network\_wordcount.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/streaming/recoverable\_network\_wordcount.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/streaming/queue\_stream.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/streaming/network\_wordjoinsentiments.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/streaming/network\_wordcount.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/streaming/hdfs\_wordcount.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/streaming/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/status\_api\_demo.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/sql/  
spark-3.4.1-bin-hadoop3/examples/src/main/python/sql/streaming/  
spark-3.4.1-bin-hadoop3/examples/src/main/python/sql/streaming/structured\_sessionization.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/sql/streaming/structured\_network\_wordcount\_windowed.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/sql/streaming/structured\_network\_wordcount\_session\_window.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/sql/streaming/structured\_network\_wordcount.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/sql/streaming/structured\_kafka\_word

count.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/sql/streaming/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/sql/hive.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/sql/datasource.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/sql/basic.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/sql/arrow.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/sql/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/sort.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/pi.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/parquet\_inputformat.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/pagerank.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/word2vec\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/word2vec.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/tf\_idf\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/svm\_with\_sgd\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/svd\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/summary\_statistics\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/streaming\_linear\_regression\_e  
xample.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/streaming\_k\_means\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/stratified\_sampling\_example.p  
y  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/standard\_scaler\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/sampled\_rdds.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/regression\_metrics\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/recommendation\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/ranking\_metrics\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/random\_rdd\_generation.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/random\_forest\_regression\_exam  
ple.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/random\_forest\_classification\_  
example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/power\_iteration\_clustering\_ex  
ample.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/pca\_rowmatrix\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/normalizer\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/naive\_bayes\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/multi\_label\_metrics\_example.p  
y  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/multi\_class\_metrics\_example.p  
y  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/logistic\_regression\_with\_lbfg  
s\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/logistic\_regression.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/linear\_regression\_with\_sgd\_ex  
ample.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/latent\_dirichlet\_allocation\_e  
xample.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/kmeans.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/kernel\_density\_estimation\_exa  
mple.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/k\_means\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/isotonic\_regression\_example.p  
y  
spark-3.4.1-bin-hadoop3/examples/src/main/python/mllib/hypothesis\_testing\_kolmogorov

\_smirnov\_test\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/lib/hypothesis\_testing\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/lib/gradient\_boosting\_regression\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/lib/gradient\_boosting\_classification\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/lib/gaussian\_mixture\_model.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/lib/gaussian\_mixture\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/lib/fpgrowth\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/lib/elementwise\_product\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/lib/decision\_tree\_regression\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/lib/decision\_tree\_classification\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/lib/correlations\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/lib/correlations.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/lib/bisecting\_k\_means\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/lib/binary\_classification\_metrics\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/lib/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/word2vec\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/vector\_slicer\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/vector\_size\_hint\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/vector\_indexer\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/vector\_assembler\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/variance\_threshold\_selector\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/univariate\_feature\_selector\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/train\_validation\_split.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/tokenizer\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/tf\_idf\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/summarizer\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/string\_indexer\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/stopwords\_remover\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/standard\_scaler\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/sql\_transformer.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/robust\_scaler\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/rformula\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/random\_forest\_regressor\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/random\_forest\_classifier\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/quantile\_discretizer\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/prefixspan\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/power\_iteration\_clustering\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/polynomial\_expansion\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/pipeline\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/pca\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/onehot\_encoder\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/one\_vs\_rest\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/normalizer\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/naive\_bayes\_example.py

spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/n\_gram\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/multilayer\_perceptron\_classification.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/multiclass\_logistic\_regression\_with\_elastic\_net.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/min\_max\_scaler\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/min\_hash\_lsh\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/max\_abs\_scaler\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/logistic\_regression\_with\_elastic\_net.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/logistic\_regression\_summary\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/linearsvc.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/linear\_regression\_with\_elastic\_net.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/lda\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/kmeans\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/isotonic\_regression\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/interaction\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/index\_to\_string\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/imputer\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/gradient\_boosted\_tree\_regressor\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/gradient\_boosted\_tree\_classifier\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/generalized\_linear\_regression\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/gaussian\_mixture\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/fpgrowth\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/fm\_regressor\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/fm\_classifier\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/feature\_hasher\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/estimator\_transformer\_param\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/elementwise\_product\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/decision\_tree\_regression\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/decision\_tree\_classification\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/dct\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/dataframe\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/cross\_validator.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/count\_vectorizer\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/correlation\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/chisq\_selector\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/chi\_square\_test\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/bucketizer\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/bucketed\_random\_projection\_lsh\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/bisecting\_k\_means\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/binarizer\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/als\_example.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/aft\_survival\_regression.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/ml/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/logistic\_regression.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/kmeans.py

spark-3.4.1-bin-hadoop3/examples/src/main/python/avro\_inputformat.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/als.py  
spark-3.4.1-bin-hadoop3/examples/src/main/python/\_\_init\_\_.py  
spark-3.4.1-bin-hadoop3/examples/src/main/java/  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/streaming/  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/streaming/J  
avaStatefulNetworkWordCount.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/streaming/J  
avaSqlNetworkWordCount.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/streaming/J  
avaRecoverableNetworkWordCount.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/streaming/J  
avaRecord.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/streaming/J  
avaQueueStream.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/streaming/J  
avaNetworkWordCount.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/streaming/J  
avaDirectKerberizedKafkaWordCount.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/streaming/J  
avaDirectKafkaWordCount.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/streaming/J  
avaCustomReceiver.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/sql/  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/sql/streami  
ng/  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/sql/streami  
ng/JavaStructuredSessionization.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/sql/streami  
ng/JavaStructuredNetworkWordCountWindowed.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/sql/streami  
ng/JavaStructuredNetworkWordCount.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/sql/streami  
ng/JavaStructuredKerberizedKafkaWordCount.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/sql/streami  
ng/JavaStructuredKafkaWordCount.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/sql/streami  
ng/JavaStructuredComplexSessionization.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/sql/hive/  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/sql/hive/Ja  
vaSparkHiveExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/sql/JavaUse  
rDefinedUntypedAggregation.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/sql/JavaUse  
rDefinedTypedAggregation.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/sql/JavaUse  
rDefinedScalar.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/sql/JavaSpa  
rkSQLExceptionExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/sql/JavaSQL  
DataSourceExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/

spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaSummaryStatisticsExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaStreamingTestExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaStratifiedSamplingExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaSimpleFPGrowth.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaSVMWithSGDExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaSVDEExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaRecommendationExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaRankingMetricsExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaRandomForestRegressionExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaRandomForestClassificationExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaPrefixSpanExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaPowerIterationClusteringExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaPCAExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaNaiveBayesExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaMulticlassClassificationMetricsExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaMultilabelClassificationMetricsExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaLogisticRegressionWithLBFGSExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaLatentDirichletAllocationExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaLBFGSExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaKernelDensityEstimationExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaKMeansExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaIsotonicRegressionExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaHypothesisTestingKolmogorovSmirnovTestExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaHypothesisTestingExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaGradientBoostingRegressionExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaGradientBoostingClassificationExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaGaussianMixtureExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaElementwiseProductExample.java

spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaDecisionTreeRegressionExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaDecisionTreeClassificationExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaCorrelationsExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaChiSqSelectorExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaBisectingKMeansExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaBinaryClassificationMetricsExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaAssociationRulesExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaALS.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaWord2VecExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaVectorSlicerExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaVectorSizeHintExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaVectorIndexerExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaVectorAssemblerExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaVarianceThresholdSelectorExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaUnivariateFeatureSelectorExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaTokenizerExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaTfidfExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaSummarizerExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaStringIndexerExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaStopWordsRemoverExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaStandardScalerExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaSQLTransformerExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaRobustScalerExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaRandomForestRegressorExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaRandomForestClassifierExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaRFormulaExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaQuantileDiscretizerExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaPref



ixSpanExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaPowerIterationClusteringExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaPolynomialExpansionExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaPipelineExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaPCAExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaOneVsRestExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaOneHotEncoderExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaNormalizerExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaNaiveBayesExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaNGramExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaMultilayerPerceptronClassifierExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaMulticlassLogisticRegressionWithElasticNetExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaModelSelectionViaTrainValidationSplitExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaModelSelectionViaCrossValidationExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaMinMaxScalerExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaMinHashLSHExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaMaxAbsScalerExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaLogisticRegressionWithElasticNetExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaLogisticRegressionSummaryExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaLinearSVCExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaLinearRegressionWithElasticNetExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaLabeledDocument.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaLDAExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaKMeansExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaIsotonicRegressionExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaInteractionExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaIndexToStringExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaImputerExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaGrad

ientBoostedTreeRegressorExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaGradientBoostedTreeClassifierExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaGeneralizedLinearRegressionExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaGaussianMixtureExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaFeatureHasherExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaFPGrowthExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaFMRegressorExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaFMClassifierExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaEstimatorTransformerParamExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaElementwiseProductExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaDocument.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaDecisionTreeRegressionExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaDecisionTreeClassificationExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaDCTExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaCountVectorizerExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaCorrelationExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaChiSquareTestExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaChiSquareSelectorExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaBucketerExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaBucketedRandomProjectionLSHExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaBisectingKMeansExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaBinomializerExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaALSEExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaAFTSurvivalRegressionExample.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/JavaWordCount.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/JavaTC.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/JavaStatusTrackerDemo.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/JavaSparkPi.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/JavaPageRank.java

spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/JavaLogQuery.java  
spark-3.4.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/JavaHdfsLR.java  
spark-3.4.1-bin-hadoop3/examples/src/main/scripts/  
spark-3.4.1-bin-hadoop3/examples/src/main/scripts/getGpusResources.sh  
spark-3.4.1-bin-hadoop3/examples/jars/  
spark-3.4.1-bin-hadoop3/examples/jars/spark-examples\_2.12-3.4.1.jar  
spark-3.4.1-bin-hadoop3/examples/jars/scopt\_2.12-3.7.1.jar  
spark-3.4.1-bin-hadoop3/kubernetes/  
spark-3.4.1-bin-hadoop3/kubernetes/tests/  
spark-3.4.1-bin-hadoop3/kubernetes/tests/worker\_memory\_check.py  
spark-3.4.1-bin-hadoop3/kubernetes/tests/python\_executable\_check.py  
spark-3.4.1-bin-hadoop3/kubernetes/tests/pyfiles.py  
spark-3.4.1-bin-hadoop3/kubernetes/tests/py\_container\_checks.py  
spark-3.4.1-bin-hadoop3/kubernetes/tests/decommissioning\_cleanup.py  
spark-3.4.1-bin-hadoop3/kubernetes/tests/decommissioning.py  
spark-3.4.1-bin-hadoop3/kubernetes/tests/autoscale.py  
spark-3.4.1-bin-hadoop3/kubernetes/dockerfiles/  
spark-3.4.1-bin-hadoop3/kubernetes/dockerfiles/spark/  
spark-3.4.1-bin-hadoop3/kubernetes/dockerfiles/spark/entrypoint.sh  
spark-3.4.1-bin-hadoop3/kubernetes/dockerfiles/spark/decom.sh  
spark-3.4.1-bin-hadoop3/kubernetes/dockerfiles/spark/bindings/  
spark-3.4.1-bin-hadoop3/kubernetes/dockerfiles/spark/bindings/python/  
spark-3.4.1-bin-hadoop3/kubernetes/dockerfiles/spark/bindings/python/Dockerfile  
spark-3.4.1-bin-hadoop3/kubernetes/dockerfiles/spark/bindings/R/  
spark-3.4.1-bin-hadoop3/kubernetes/dockerfiles/spark/bindings/R/Dockerfile  
spark-3.4.1-bin-hadoop3/kubernetes/dockerfiles/spark/Dockerfile  
spark-3.4.1-bin-hadoop3/yarn/  
spark-3.4.1-bin-hadoop3/yarn/spark-3.4.1-yarn-shuffle.jar  
spark-3.4.1-bin-hadoop3/jars/  
spark-3.4.1-bin-hadoop3/jars/rocksdbjni-7.9.2.jar  
spark-3.4.1-bin-hadoop3/jars/py4j-0.10.9.7.jar  
spark-3.4.1-bin-hadoop3/jars/protobuf-java-2.5.0.jar  
spark-3.4.1-bin-hadoop3/jars/pickle-1.3.jar  
spark-3.4.1-bin-hadoop3/jars/parquet-jackson-1.12.3.jar  
spark-3.4.1-bin-hadoop3/jars/parquet-hadoop-1.12.3.jar  
spark-3.4.1-bin-hadoop3/jars/parquet-format-structures-1.12.3.jar  
spark-3.4.1-bin-hadoop3/jars/parquet-encoding-1.12.3.jar  
spark-3.4.1-bin-hadoop3/jars/parquet-common-1.12.3.jar  
spark-3.4.1-bin-hadoop3/jars/parquet-column-1.12.3.jar  
spark-3.4.1-bin-hadoop3/jars/paranamer-2.8.jar  
spark-3.4.1-bin-hadoop3/jars/osgi-resource-locator-1.0.3.jar  
spark-3.4.1-bin-hadoop3/jars/oro-2.0.8.jar  
spark-3.4.1-bin-hadoop3/jars/orc-shims-1.8.4.jar  
spark-3.4.1-bin-hadoop3/jars/orc-mapreduce-1.8.4-shaded-protobuf.jar  
spark-3.4.1-bin-hadoop3/jars/orc-core-1.8.4-shaded-protobuf.jar  
spark-3.4.1-bin-hadoop3/jars/opencsv-2.3.jar  
spark-3.4.1-bin-hadoop3/jars/okio-1.15.0.jar  
spark-3.4.1-bin-hadoop3/jars/okhttp-3.12.12.jar  
spark-3.4.1-bin-hadoop3/jars/objenesis-3.2.jar  
spark-3.4.1-bin-hadoop3/jars/netty-transport-native-unix-common-4.1.87.Final.jar  
spark-3.4.1-bin-hadoop3/jars/netty-transport-native-kqueue-4.1.87.Final-osx-x86\_64.jar  
spark-3.4.1-bin-hadoop3/jars/netty-transport-native-kqueue-4.1.87.Final-osx-aarch\_64.jar

spark-3.4.1-bin-hadoop3/jars/netty-transport-native-epoll-4.1.87.Final-linux-x86\_64.jar  
spark-3.4.1-bin-hadoop3/jars/netty-transport-native-epoll-4.1.87.Final-linux-aarch\_64.jar  
spark-3.4.1-bin-hadoop3/jars/netty-transport-classes-kqueue-4.1.87.Final.jar  
spark-3.4.1-bin-hadoop3/jars/netty-transport-classes-epoll-4.1.87.Final.jar  
spark-3.4.1-bin-hadoop3/jars/netty-transport-4.1.87.Final.jar  
spark-3.4.1-bin-hadoop3/jars/netty-resolver-4.1.87.Final.jar  
spark-3.4.1-bin-hadoop3/jars/netty-handler-proxy-4.1.87.Final.jar  
spark-3.4.1-bin-hadoop3/jars/netty-handler-4.1.87.Final.jar  
spark-3.4.1-bin-hadoop3/jars/netty-common-4.1.87.Final.jar  
spark-3.4.1-bin-hadoop3/jars/netty-codec-socks-4.1.87.Final.jar  
spark-3.4.1-bin-hadoop3/jars/netty-codec-http2-4.1.87.Final.jar  
spark-3.4.1-bin-hadoop3/jars/netty-codec-http-4.1.87.Final.jar  
spark-3.4.1-bin-hadoop3/jars/netty-codec-4.1.87.Final.jar  
spark-3.4.1-bin-hadoop3/jars/netty-buffer-4.1.87.Final.jar  
spark-3.4.1-bin-hadoop3/jars/netty-all-4.1.87.Final.jar  
spark-3.4.1-bin-hadoop3/jars/minlog-1.3.0.jar  
spark-3.4.1-bin-hadoop3/jars/metrics-jvm-4.2.15.jar  
spark-3.4.1-bin-hadoop3/jars/metrics-json-4.2.15.jar  
spark-3.4.1-bin-hadoop3/jars/metrics-jmx-4.2.15.jar  
spark-3.4.1-bin-hadoop3/jars/metrics-graphite-4.2.15.jar  
spark-3.4.1-bin-hadoop3/jars/metrics-core-4.2.15.jar  
spark-3.4.1-bin-hadoop3/jars/mesos-1.4.3-shaded-protobuf.jar  
spark-3.4.1-bin-hadoop3/jars/lz4-java-1.8.0.jar  
spark-3.4.1-bin-hadoop3/jars/logging-interceptor-3.12.12.jar  
spark-3.4.1-bin-hadoop3/jars/log4j-slf4j2-impl-2.19.0.jar  
spark-3.4.1-bin-hadoop3/jars/log4j-core-2.19.0.jar  
spark-3.4.1-bin-hadoop3/jars/log4j-api-2.19.0.jar  
spark-3.4.1-bin-hadoop3/jars/log4j-1.2-api-2.19.0.jar  
spark-3.4.1-bin-hadoop3/jars/libthrift-0.12.0.jar  
spark-3.4.1-bin-hadoop3/jars/libfb303-0.9.3.jar  
spark-3.4.1-bin-hadoop3/jars/leveldbjni-all-1.8.jar  
spark-3.4.1-bin-hadoop3/jars/lapack-3.0.3.jar  
spark-3.4.1-bin-hadoop3/jars/kubernetes-model-storageclass-6.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/kubernetes-model-scheduling-6.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/kubernetes-model-rbac-6.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/kubernetes-model-policy-6.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/kubernetes-model-node-6.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/kubernetes-model-networking-6.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/kubernetes-model-metrics-6.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/kubernetes-model-gatewayapi-6.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/kubernetes-model-flowcontrol-6.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/kubernetes-model-extensions-6.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/kubernetes-model-events-6.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/kubernetes-model-discovery-6.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/kubernetes-model-core-6.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/kubernetes-model-coordination-6.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/kubernetes-model-common-6.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/kubernetes-model-certificates-6.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/kubernetes-model-batch-6.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/kubernetes-model-autoscaling-6.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/kubernetes-model-apps-6.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/kubernetes-model-apiextensions-6.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/kubernetes-model-admissionregistration-6.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/kubernetes-httpclient-okhttp-6.4.1.jar

spark-3.4.1-bin-hadoop3/jars/kubernetes-client-api-6.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/kubernetes-client-6.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/kryo-shaded-4.0.2.jar  
spark-3.4.1-bin-hadoop3/jars/jul-to-slf4j-2.0.6.jar  
spark-3.4.1-bin-hadoop3/jars/jta-1.1.jar  
spark-3.4.1-bin-hadoop3/jars/jsr305-3.0.0.jar  
spark-3.4.1-bin-hadoop3/jars/json4s-scalap\_2.12-3.7.0-M11.jar  
spark-3.4.1-bin-hadoop3/jars/json4s-jackson\_2.12-3.7.0-M11.jar  
spark-3.4.1-bin-hadoop3/jars/json4s-core\_2.12-3.7.0-M11.jar  
spark-3.4.1-bin-hadoop3/jars/json4s-ast\_2.12-3.7.0-M11.jar  
spark-3.4.1-bin-hadoop3/jars/json-1.8.jar  
spark-3.4.1-bin-hadoop3/jars/jpam-1.1.jar  
spark-3.4.1-bin-hadoop3/jars/jodd-core-3.5.2.jar  
spark-3.4.1-bin-hadoop3/jars/joda-time-2.12.2.jar  
spark-3.4.1-bin-hadoop3/jars/jline-2.14.6.jar  
spark-3.4.1-bin-hadoop3/jars/jersey-server-2.36.jar  
spark-3.4.1-bin-hadoop3/jars/jersey-hk2-2.36.jar  
spark-3.4.1-bin-hadoop3/jars/jersey-container-servlet-core-2.36.jar  
spark-3.4.1-bin-hadoop3/jars/jersey-container-servlet-2.36.jar  
spark-3.4.1-bin-hadoop3/jars/jersey-common-2.36.jar  
spark-3.4.1-bin-hadoop3/jars/jersey-client-2.36.jar  
spark-3.4.1-bin-hadoop3/jars/jdo-api-3.0.1.jar  
spark-3.4.1-bin-hadoop3/jars/jcl-over-slf4j-2.0.6.jar  
spark-3.4.1-bin-hadoop3/jars/jaxb-runtime-2.3.2.jar  
spark-3.4.1-bin-hadoop3/jars/javolution-5.5.1.jar  
spark-3.4.1-bin-hadoop3/jars/javax.jdo-3.2.0-m3.jar  
spark-3.4.1-bin-hadoop3/jars/javassist-3.25.0-GA.jar  
spark-3.4.1-bin-hadoop3/jars/janino-3.1.9.jar  
spark-3.4.1-bin-hadoop3/jars/jakarta.xml.bind-api-2.3.2.jar  
spark-3.4.1-bin-hadoop3/jars/jakarta.ws.rs-api-2.1.6.jar  
spark-3.4.1-bin-hadoop3/jars/jakarta.validation-api-2.0.2.jar  
spark-3.4.1-bin-hadoop3/jars/jakarta.servlet-api-4.0.3.jar  
spark-3.4.1-bin-hadoop3/jars/jakarta.inject-2.6.1.jar  
spark-3.4.1-bin-hadoop3/jars/jakarta.annotation-api-1.3.5.jar  
spark-3.4.1-bin-hadoop3/jars/jackson-module-scala\_2.12-2.14.2.jar  
spark-3.4.1-bin-hadoop3/jars/jackson-mapper-asl-1.9.13.jar  
spark-3.4.1-bin-hadoop3/jars/jackson-datatype-jsr310-2.14.2.jar  
spark-3.4.1-bin-hadoop3/jars/jackson-dataformat-yaml-2.14.2.jar  
spark-3.4.1-bin-hadoop3/jars/jackson-databind-2.14.2.jar  
spark-3.4.1-bin-hadoop3/jars/jackson-core-asl-1.9.13.jar  
spark-3.4.1-bin-hadoop3/jars/jackson-core-2.14.2.jar  
spark-3.4.1-bin-hadoop3/jars/jackson-annotations-2.14.2.jar  
spark-3.4.1-bin-hadoop3/jars/ivy-2.5.1.jar  
spark-3.4.1-bin-hadoop3/jars/istack-commons-runtime-3.0.8.jar  
spark-3.4.1-bin-hadoop3/jars/httpcore-4.4.16.jar  
spark-3.4.1-bin-hadoop3/jars/httpclient-4.5.14.jar  
spark-3.4.1-bin-hadoop3/jars/hk2-utils-2.6.1.jar  
spark-3.4.1-bin-hadoop3/jars/hk2-locator-2.6.1.jar  
spark-3.4.1-bin-hadoop3/jars/hk2-api-2.6.1.jar  
spark-3.4.1-bin-hadoop3/jars/hive-storage-api-2.8.1.jar  
spark-3.4.1-bin-hadoop3/jars/hive-shims-scheduler-2.3.9.jar  
spark-3.4.1-bin-hadoop3/jars/hive-shims-common-2.3.9.jar  
spark-3.4.1-bin-hadoop3/jars/hive-shims-2.3.9.jar  
spark-3.4.1-bin-hadoop3/jars/hive-shims-0.23-2.3.9.jar  
spark-3.4.1-bin-hadoop3/jars/hive-service-rpc-3.1.3.jar  
spark-3.4.1-bin-hadoop3/jars/hive-serde-2.3.9.jar

spark-3.4.1-bin-hadoop3/jars/hive-metastore-2.3.9.jar  
spark-3.4.1-bin-hadoop3/jars/hive-llap-common-2.3.9.jar  
spark-3.4.1-bin-hadoop3/jars/hive-jdbc-2.3.9.jar  
spark-3.4.1-bin-hadoop3/jars/hive-exec-2.3.9-core.jar  
spark-3.4.1-bin-hadoop3/jars/hive-common-2.3.9.jar  
spark-3.4.1-bin-hadoop3/jars/hive-cli-2.3.9.jar  
spark-3.4.1-bin-hadoop3/jars/hive-beeline-2.3.9.jar  
spark-3.4.1-bin-hadoop3/jars/hadoop-yarn-server-web-proxy-3.3.4.jar  
spark-3.4.1-bin-hadoop3/jars/hadoop-shaded-guava-1.1.1.jar  
spark-3.4.1-bin-hadoop3/jars/hadoop-client-runtime-3.3.4.jar  
spark-3.4.1-bin-hadoop3/jars/hadoop-client-api-3.3.4.jar  
spark-3.4.1-bin-hadoop3/jars/guava-14.0.1.jar  
spark-3.4.1-bin-hadoop3/jars/gson-2.2.4.jar  
spark-3.4.1-bin-hadoop3/jars/flatbuffers-java-1.12.0.jar  
spark-3.4.1-bin-hadoop3/jars/dropwizard-metrics-hadoop-metrics2-reporter-0.1.2.jar  
spark-3.4.1-bin-hadoop3/jars/derby-10.14.2.0.jar  
spark-3.4.1-bin-hadoop3/jars/datanucleus-rdbms-4.1.19.jar  
spark-3.4.1-bin-hadoop3/jars/datanucleus-core-4.1.17.jar  
spark-3.4.1-bin-hadoop3/jars/datanucleus-api-jdo-4.2.4.jar  
spark-3.4.1-bin-hadoop3/jars/curator-recipes-2.13.0.jar  
spark-3.4.1-bin-hadoop3/jars/curator-framework-2.13.0.jar  
spark-3.4.1-bin-hadoop3/jars/curator-client-2.13.0.jar  
spark-3.4.1-bin-hadoop3/jars/compress-lzf-1.1.2.jar  
spark-3.4.1-bin-hadoop3/jars/commons-text-1.10.0.jar  
spark-3.4.1-bin-hadoop3/jars/commons-pool-1.5.4.jar  
spark-3.4.1-bin-hadoop3/jars/commons-math3-3.6.1.jar  
spark-3.4.1-bin-hadoop3/jars/commons-logging-1.1.3.jar  
spark-3.4.1-bin-hadoop3/jars/commons-lang3-3.12.0.jar  
spark-3.4.1-bin-hadoop3/jars/commons-lang-2.6.jar  
spark-3.4.1-bin-hadoop3/jars/commons-io-2.11.0.jar  
spark-3.4.1-bin-hadoop3/jars/commons-dbcp-1.4.jar  
spark-3.4.1-bin-hadoop3/jars/commons-crypto-1.1.0.jar  
spark-3.4.1-bin-hadoop3/jars/commons-compress-1.22.jar  
spark-3.4.1-bin-hadoop3/jars/commons-compiler-3.1.9.jar  
spark-3.4.1-bin-hadoop3/jars/commons-collections4-4.4.jar  
spark-3.4.1-bin-hadoop3/jars/commons-collections-3.2.2.jar  
spark-3.4.1-bin-hadoop3/jars/commons-codec-1.15.jar  
spark-3.4.1-bin-hadoop3/jars/commons-cli-1.5.0.jar  
spark-3.4.1-bin-hadoop3/jars/chill\_2.12-0.10.0.jar  
spark-3.4.1-bin-hadoop3/jars/chill-java-0.10.0.jar  
spark-3.4.1-bin-hadoop3/jars/cats-kernel\_2.12-2.1.1.jar  
spark-3.4.1-bin-hadoop3/jars/breeze\_2.12-2.1.0.jar  
spark-3.4.1-bin-hadoop3/jars/breeze-macros\_2.12-2.1.0.jar  
spark-3.4.1-bin-hadoop3/jars/bonecp-0.8.0.RELEASE.jar  
spark-3.4.1-bin-hadoop3/jars/blas-3.0.3.jar  
spark-3.4.1-bin-hadoop3/jars/avro-mapred-1.11.1.jar  
spark-3.4.1-bin-hadoop3/jars/avro-ipc-1.11.1.jar  
spark-3.4.1-bin-hadoop3/jars/avro-1.11.1.jar  
spark-3.4.1-bin-hadoop3/jars/audience-annotations-0.5.0.jar  
spark-3.4.1-bin-hadoop3/jars/arrow-vector-11.0.0.jar  
spark-3.4.1-bin-hadoop3/jars/arrow-memory-netty-11.0.0.jar  
spark-3.4.1-bin-hadoop3/jars/arrow-memory-core-11.0.0.jar  
spark-3.4.1-bin-hadoop3/jars/arrow-format-11.0.0.jar  
spark-3.4.1-bin-hadoop3/jars/arpack\_combined\_all-0.1.jar  
spark-3.4.1-bin-hadoop3/jars/arpack-3.0.3.jar  
spark-3.4.1-bin-hadoop3/jars/aopalliance-repackaged-2.6.1.jar

spark-3.4.1-bin-hadoop3/jars/antlr4-runtime-4.9.3.jar  
spark-3.4.1-bin-hadoop3/jars/antlr-runtime-3.5.2.jar  
spark-3.4.1-bin-hadoop3/jars/annotations-17.0.0.jar  
spark-3.4.1-bin-hadoop3/jars/algebra\_2.12-2.0.1.jar  
spark-3.4.1-bin-hadoop3/jars/aircompressor-0.21.jar  
spark-3.4.1-bin-hadoop3/jars/activation-1.1.1.jar  
spark-3.4.1-bin-hadoop3/jars/ST4-4.0.4.jar  
spark-3.4.1-bin-hadoop3/jars/RoaringBitmap-0.9.38.jar  
spark-3.4.1-bin-hadoop3/jars/JTransforms-3.1.jar  
spark-3.4.1-bin-hadoop3/jars/JLargeArrays-1.5.jar  
spark-3.4.1-bin-hadoop3/jars/HikariCP-2.5.1.jar  
spark-3.4.1-bin-hadoop3/jars/zstd-jni-1.5.2-5.jar  
spark-3.4.1-bin-hadoop3/jars/zookeeper-jute-3.6.3.jar  
spark-3.4.1-bin-hadoop3/jars/zookeeper-3.6.3.jar  
spark-3.4.1-bin-hadoop3/jars/zjsonpatch-0.3.0.jar  
spark-3.4.1-bin-hadoop3/jars/xz-1.9.jar  
spark-3.4.1-bin-hadoop3/jars/xbean-asm9-shaded-4.22.jar  
spark-3.4.1-bin-hadoop3/jars/univocity-parsers-2.9.1.jar  
spark-3.4.1-bin-hadoop3/jars/transaction-api-1.1.jar  
spark-3.4.1-bin-hadoop3/jars/tink-1.7.0.jar  
spark-3.4.1-bin-hadoop3/jars/threeten-extra-1.7.1.jar  
spark-3.4.1-bin-hadoop3/jars/super-csv-2.2.0.jar  
spark-3.4.1-bin-hadoop3/jars/stream-2.9.6.jar  
spark-3.4.1-bin-hadoop3/jars/stax-api-1.0.1.jar  
spark-3.4.1-bin-hadoop3/jars/spire\_2.12-0.17.0.jar  
spark-3.4.1-bin-hadoop3/jars/spire-util\_2.12-0.17.0.jar  
spark-3.4.1-bin-hadoop3/jars/spire-platform\_2.12-0.17.0.jar  
spark-3.4.1-bin-hadoop3/jars/spire-macros\_2.12-0.17.0.jar  
spark-3.4.1-bin-hadoop3/jars/spark-yarn\_2.12-3.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/spark-unsafe\_2.12-3.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/spark-tags\_2.12-3.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/spark-streaming\_2.12-3.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/spark-sql\_2.12-3.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/spark-sketch\_2.12-3.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/spark-repl\_2.12-3.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/spark-network-shuffle\_2.12-3.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/spark-network-common\_2.12-3.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/spark-mllib\_2.12-3.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/spark-mllib-local\_2.12-3.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/spark-mesos\_2.12-3.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/spark-launcher\_2.12-3.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/spark-kvstore\_2.12-3.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/spark-kubernetes\_2.12-3.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/spark-hive\_2.12-3.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/spark-hive-thriftserver\_2.12-3.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/spark-graphx\_2.12-3.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/spark-core\_2.12-3.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/spark-catalyst\_2.12-3.4.1.jar  
spark-3.4.1-bin-hadoop3/jars/snappy-java-1.1.10.1.jar  
spark-3.4.1-bin-hadoop3/jars/snakeyaml-1.33.jar  
spark-3.4.1-bin-hadoop3/jars/slf4j-api-2.0.6.jar  
spark-3.4.1-bin-hadoop3/jars/shims-0.9.38.jar  
spark-3.4.1-bin-hadoop3/jars/scala-xml\_2.12-2.1.0.jar  
spark-3.4.1-bin-hadoop3/jars/scala-reflect-2.12.17.jar  
spark-3.4.1-bin-hadoop3/jars/scala-parser-combinators\_2.12-2.1.1.jar  
spark-3.4.1-bin-hadoop3/jars/scala-library-2.12.17.jar

```
spark-3.4.1-bin-hadoop3/jars/scala-compiler-2.12.17.jar
spark-3.4.1-bin-hadoop3/jars/scala-collection-compat_2.12-2.7.0.jar
spark-3.4.1-bin-hadoop3/RELEASE
```

```
In [2]: from pyspark.sql import SparkSession

spark = SparkSession.builder \
    .appName("NYC Yellow Taxi Data Wrangling") \
    .getOrCreate()
```

---

## Installing Java and Apache Spark for Data Wrangling with Spark's Structured APIs

To enable data processing using Spark's Structured APIs, a suitable environment must be configured. The following steps outline the process of setting up Java, Apache Spark, and the necessary Python libraries to begin working with the NYC Yellow Taxi Trips dataset.

1. **Install Java** Java is required to run Apache Spark because Spark operates on the Java Virtual Machine (JVM). OpenJDK 11 is installed as it is a stable and widely supported version. The installation is performed quietly to minimize unnecessary console output.
  2. **Download and Extract Apache Spark** Apache Spark version 3.4.1, bundled with Hadoop 3, is downloaded from the official Apache archive. This version supports Spark's Structured APIs and integrates with various data storage systems. Once downloaded, the Spark archive is extracted to make its contents accessible for further use.
  3. **Install the findspark Library** The `findspark` Python package is installed to simplify the integration of Spark with Python. It allows Python scripts to locate and initialize the Spark environment without manual configuration of paths and environment settings.
  4. **Set Environment Variables** Two key environment variables are configured:
    - `JAVA_HOME` is set to the directory of the Java installation, enabling Spark to find and use the installed Java environment.
    - `SPARK_HOME` is set to the directory where Spark has been extracted. This allows `findspark` and other tools to locate Spark's core components.
  5. **Initialize findspark** The `findspark` library is initialized to ensure that Spark can be properly loaded into the Python runtime. This step is crucial for enabling PySpark, the Python API for Spark, to function correctly.
  6. **Create a SparkSession** A `SparkSession` is created to establish the main entry point for using Spark's DataFrame and SQL APIs. The session is named appropriately to reflect the context of the project. Once initialized, this session allows for the execution of Spark operations such as data loading, transformation, and analysis.
-



# Hands-on Coding Task

## Importing Dataset

```
In [3]: from google.colab import drive
drive.mount('/content/drive')

parquet_path = "/content/drive/My Drive/Colab Notebooks/yellow_tripdata_2025-01.parquet"
df = spark.read.parquet(parquet_path)
df.show(10)

''' output_path = "/content/drive/My Drive/Colab Notebooks/yellow_tripdata_2025-01_
df.write.option("header", True).csv(output_path)'''
```

Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force\_remount=True).

| VendorID tpep_pickup_datetime tpep_dropoff_datetime passenger_count trip_distance RatecodeID store_and_fwd_flag PULocationID DOLocationID payment_type fare_amount extra mta_tax tip_amount tolls_amount improvement_surcharge total_amount congestion_surcharge Airport_fee cbd_congestion_fee |   |     |   |                     |                     |   |      |     |     |      |     |     |     |       |     |
|---|---|-----|---|---------------------|---------------------|---|------|-----|-----|------|-----|-----|-----|-------|-----|
| 1   | 5 | 0.0 | 1 | 2025-01-01 00:18:38 | 2025-01-01 00:26:59 | 1 | 10.0 | 3.5 | 0.0 | 3.0  | 0.0 | 0.0 | 1.0 | 18.0  | 2.5 |
| 1   | 5 | 0.0 | 1 | 2025-01-01 00:32:40 | 2025-01-01 00:35:13 | 1 | 5.1  | 3.5 | 0.0 | 2.02 | 0.0 | 0.0 | 1.0 | 12.12 | 2.5 |
| 1   | 5 | 0.0 | 1 | 2025-01-01 00:44:04 | 2025-01-01 00:46:01 | 1 | 5.1  | 3.5 | 0.0 | 2.0  | 0.0 | 0.0 | 1.0 | 12.1  | 2.5 |
| 1   | 5 | 0.0 | 2 | 2025-01-01 00:14:27 | 2025-01-01 00:20:01 | 3 | 7.2  | 1.0 | 0.0 | 0.0  | 0.0 | 0.0 | 1.0 | 9.7   | 0.0 |
| 1   | 5 | 0.0 | 2 | 2025-01-01 00:21:34 | 2025-01-01 00:25:06 | 3 | 5.8  | 1.0 | 0.0 | 0.0  | 0.0 | 0.0 | 1.0 | 8.3   | 0.0 |
| 1   | 5 | 0.0 | 2 | 2025-01-01 00:48:24 | 2025-01-01 01:08:26 | 2 | 19.1 | 1.0 | 0.0 | 0.0  | 0.0 | 0.0 | 1.0 | 24.1  | 2.5 |
| 1   | 5 | 0.0 | 1 | 2025-01-01 00:14:47 | 2025-01-01 00:16:15 | 0 | 4.4  | 3.5 | 0.0 | 2.35 | 0.0 | 0.0 | 1.0 | 11.75 | 2.5 |
| 1   | 5 | 0.0 | 1 | 2025-01-01 00:39:27 | 2025-01-01 00:51:51 | 0 | 12.1 | 3.5 | 0.0 | 2.0  | 0.0 | 0.0 | 1.0 | 19.1  | 2.5 |
| 1   | 5 | 0.0 | 1 | 2025-01-01 00:53:43 | 2025-01-01 01:13:23 | 0 | 19.1 | 3.5 | 0.0 | 3.0  | 0.0 | 0.0 | 1.0 | 27.1  | 2.5 |
| 1   | 5 | 0.0 | 2 | 2025-01-01 00:00:02 | 2025-01-01 00:09:36 | 1 | 11.4 | 1.0 | 0.0 | 0.0  | 0.0 | 0.0 | 1.0 | 16.4  | 2.5 |

```
-+-----+-----+-----+-----+-----+-----+-----+-----+-----+
-----+-----+-----+-----+
only showing top 10 rows
```

```
Out[3]: ' output_path = "/content/drive/My Drive/Colab Notebooks/yellow_tripdata_2025-01_c
sv"\ndf.write.option("header", True).csv(output_path)'
```

---

This dataset contains detailed records of taxi trips in New York City for January 2025. It includes information such as pickup and drop-off times and locations, fare amounts, passenger counts, and additional charges like tips, tolls, and congestion surcharges. The data provides insights into travel behavior within the city, including average trip lengths, common payment methods, and pricing components. Notably, many trips are short-distance and frequently occur within the central parts of the city. This dataset can be valuable for analyzing urban transportation trends, rider habits, and operational efficiency of taxi services.

Key observations from the dataset:

- Most trips are under 3 miles and last less than 15 minutes.
- Credit card is the most common payment method, and only these trips show tip amounts.
- Congestion surcharges (typically \$2.50) are frequently applied, reflecting travel in high-traffic zones.
- Several entries have a **passenger count of 0**, which may indicate missing or incorrect data.
- Fare breakdown includes base fare, taxes, surcharges, and optional tips, allowing for full revenue analysis.

---

## Importing Dataset

1. **Mount Google Drive** Google Drive is mounted to access the dataset stored in the user's Drive from within the Colab environment.
  2. **Specify File Path** The path to the Parquet file ( `yellow_tripdata_2025-01.parquet` ) is defined based on its location in Google Drive.
  3. **Read Dataset with Spark** The Parquet file is read into a Spark DataFrame using Spark's `read.parquet` method, allowing for efficient handling of large-scale structured data.
  4. **Preview Data** The first 10 rows of the DataFrame are displayed using `.show(10)` to verify successful loading.
  5. **(Optional) Export as CSV** An optional (commented-out) command shows how the DataFrame can be saved as a CSV file with headers.
-

## Cleaning Dataset and Displaying First 10 Rows

```
In [4]: # Drop rows with null values in critical columns
df_clean = df.dropna(subset=["tpep_pickup_datetime", "tpep_dropoff_datetime", "fare_amount"])

# Filter out records with trip_distance <= 0 or fare_amount <= 0
df_clean = df_clean.filter((df_clean["trip_distance"] > 0) & (df_clean["fare_amount"] > 0))

# Rename two columns for clarity
df_clean = df_clean.withColumnRenamed("tpep_pickup_datetime", "pickup_datetime") \
                  .withColumnRenamed("tpep_dropoff_datetime", "dropoff_datetime")

df = df.withColumnRenamed("pickup_datetime", "pickup_time") \
        .withColumnRenamed("dropoff_datetime", "dropoff_time")

df_clean.show(10)
```

| VendorID       | pickup_datetime       | dropoff_datetime    | passenger_count       | trip_distance | Rate               |         |
|----------------|-----------------------|---------------------|-----------------------|---------------|--------------------|---------|
| codeID         | store_and_fwd_flag    | PULocationID        | DOLocationID          | payment_type  | fare_amount        | extra_m |
| ta_tax         | tip_amount            | tolls_amount        | improvement_surcharge | total_amount  | congestion_surchar |         |
| ge Airport_fee | cdb_congestion_fee    |                     |                       |               |                    |         |
|                | 1 2025-01-01 00:18:38 | 2025-01-01 00:26:59 |                       | 1             | 1.6                |         |
| 1              | N                     | 229                 | 237                   | 1             | 10.0               | 3.5  0. |
| 5              | 3.0                   | 0.0                 | 1.0                   | 18.0          |                    | 2.5     |
| 0.0            | 0.0                   |                     |                       |               |                    |         |
|                | 1 2025-01-01 00:32:40 | 2025-01-01 00:35:13 |                       | 1             | 0.5                |         |
| 1              | N                     | 236                 | 237                   | 1             | 5.1                | 3.5  0. |
| 5              | 2.02                  | 0.0                 | 1.0                   | 12.12         |                    | 2.5     |
| 0.0            | 0.0                   |                     |                       |               |                    |         |
|                | 1 2025-01-01 00:44:04 | 2025-01-01 00:46:01 |                       | 1             | 0.6                |         |
| 1              | N                     | 141                 | 141                   | 1             | 5.1                | 3.5  0. |
| 5              | 2.0                   | 0.0                 | 1.0                   | 12.1          |                    | 2.5     |
| 0.0            | 0.0                   |                     |                       |               |                    |         |
|                | 2 2025-01-01 00:14:27 | 2025-01-01 00:20:01 |                       | 3             | 0.52               |         |
| 1              | N                     | 244                 | 244                   | 2             | 7.2                | 1.0  0. |
| 5              | 0.0                   | 0.0                 | 1.0                   | 9.7           |                    | 0.0     |
| 0.0            | 0.0                   |                     |                       |               |                    |         |
|                | 2 2025-01-01 00:21:34 | 2025-01-01 00:25:06 |                       | 3             | 0.66               |         |
| 1              | N                     | 244                 | 116                   | 2             | 5.8                | 1.0  0. |
| 5              | 0.0                   | 0.0                 | 1.0                   | 8.3           |                    | 0.0     |
| 0.0            | 0.0                   |                     |                       |               |                    |         |
|                | 2 2025-01-01 00:48:24 | 2025-01-01 01:08:26 |                       | 2             | 2.63               |         |
| 1              | N                     | 239                 | 68                    | 2             | 19.1               | 1.0  0. |
| 5              | 0.0                   | 0.0                 | 1.0                   | 24.1          |                    | 2.5     |
| 0.0            | 0.0                   |                     |                       |               |                    |         |
|                | 1 2025-01-01 00:14:47 | 2025-01-01 00:16:15 |                       | 0             | 0.4                |         |
| 1              | N                     | 170                 | 170                   | 1             | 4.4                | 3.5  0. |
| 5              | 2.35                  | 0.0                 | 1.0                   | 11.75         |                    | 2.5     |
| 0.0            | 0.0                   |                     |                       |               |                    |         |
|                | 1 2025-01-01 00:39:27 | 2025-01-01 00:51:51 |                       | 0             | 1.6                |         |
| 1              | N                     | 234                 | 148                   | 1             | 12.1               | 3.5  0. |
| 5              | 2.0                   | 0.0                 | 1.0                   | 19.1          |                    | 2.5     |
| 0.0            | 0.0                   |                     |                       |               |                    |         |
|                | 1 2025-01-01 00:53:43 | 2025-01-01 01:13:23 |                       | 0             | 2.8                |         |
| 1              | N                     | 148                 | 170                   | 1             | 19.1               | 3.5  0. |
| 5              | 3.0                   | 0.0                 | 1.0                   | 27.1          |                    | 2.5     |
| 0.0            | 0.0                   |                     |                       |               |                    |         |
|                | 2 2025-01-01 00:00:02 | 2025-01-01 00:09:36 |                       | 1             | 1.71               |         |
| 1              | N                     | 237                 | 262                   | 2             | 11.4               | 1.0  0. |
| 5              | 0.0                   | 0.0                 | 1.0                   | 16.4          |                    | 2.5     |
| 0.0            | 0.0                   |                     |                       |               |                    |         |

only showing top 10 rows

---

This cleaned dataset presents comprehensive trip-level records of NYC yellow taxi rides from January 2025. Each entry includes the vendor ID, timestamps for pickup and drop-off, trip distance, number of passengers, and a detailed breakdown of fare components including base fare, taxes, tips, tolls, and surcharges. Also captured are pickup and drop-off location IDs, payment type, and congestion or airport-related fees. The dataset is especially useful for analyzing fare structures, passenger distribution, ride frequency, and the financial impact of surcharges and fees in urban transport.

Key highlights from the cleaned dataset:

- All rides have valid timestamps, trip distances, and location IDs for both pickup and drop-off points.
- The **majority of trips are short**, with distances often below 3 miles and durations under 15 minutes.
- **Tip amounts appear only for payment type 1 (credit card)**, confirming tipping data is recorded only for card transactions.
- **Passenger count includes some zero values**, which may represent data entry errors or non-passenger trips (e.g., vehicle repositioning).
- The **“congestion\_surcharge”** field is frequently populated with \$2.50, indicating rides through congested zones.
- Extra charges such as **improvement surcharge (\$1.00)** and **MTA tax (\$0.50)** are consistent across trips.
- No entries in this sample include **airport fees** or **CBD congestion fees**, suggesting rides were likely not airport-bound or outside central business district surcharge hours.

---

## Cleaning Dataset and Displaying First 10 Rows

1. **Drop Null Values** Rows with missing values in key columns—pickup time, dropoff time, and fare amount—are removed to ensure data integrity.
  2. **Filter Invalid Records** Records with non-positive trip distances or fare amounts are filtered out, as these are likely errors or irrelevant entries.
  3. **Rename Columns** Pickup and dropoff datetime columns are renamed to more readable names for clarity during analysis.
  4. **Display Cleaned Data** The first 10 rows of the cleaned dataset are displayed to verify that the cleaning steps were applied successfully.
-

## Transform and analyze the data

```
In [5]: df_clean.groupBy("passenger_count") \
        .avg("fare_amount") \
        .withColumnRenamed("avg(fare_amount)", "avg_fare_amount") \
        .orderBy("passenger_count") \
        .show()
```

```
+-----+-----+
|passenger_count| avg_fare_amount|
+-----+-----+
|          null|18.143937865569782|
|             0|15.504962978633374|
|             1| 17.77854678492162|
|             2|20.166688025345344|
|             3|19.855320701866237|
|             4| 22.18569644975821|
|             5| 16.85873063610059|
|             6| 17.34644684956796|
|             7|              79.0|
|             8|              69.73|
|             9| 91.33333333333333|
+-----+-----+
```

| passenger_count | avg_fare_amount | Insights   |
|-----------------|-----------------|--|
| null            | 18.14           | Possibly missing or unrecorded passenger counts, still has an average fare value.      |
| 0               | 15.50           | Could represent unoccupied taxis (maybe test rides or errors). Still charges recorded. |
| 1               | 17.78           | Most likely standard single-passenger trips.   |
| 2–4             | ~20–22.18       | Slightly higher fares — maybe due to longer/multi-person shared trips.                 |
| 5–6             | ~16.85–17.35    | Lower again — possibly short rides or shared trips with limited distance.              |
| 7–9             | 69.73–91.33     | Outliers — very high average fares, could be long-distance or erroneous entries.       |

## Transforming and Analyzing the Data

- 1. Group by Passenger Count** The dataset is grouped based on the number of passengers per trip to analyze trends related to fare distribution.
- 2. Calculate Average Fare** For each passenger count group, the average fare amount is calculated to understand how fare values vary with the number of passengers.

3. **Rename Column for Clarity** The resulting column showing average fare values is renamed to `avg_fare_amount` for readability.
  4. **Sort and Display Results** The grouped data is sorted by passenger count and displayed to observe patterns and draw insights.
- 

## Insights

- **Null Passenger Count (Avg: \$18.14)** Entries with no recorded passenger count may indicate missing data or input errors, but still show fare values.
  - **Passenger Count = 0 (Avg: \$15.50)** Could reflect system tests or errors, showing trips recorded without passengers.
  - **Passenger Count = 1 (Avg: \$17.78)** Likely the most common and standard type of ride — solo passengers.
  - **Passenger Count 2–4 (Avg: ~20.00–22.18)** Higher fares possibly due to longer trips or rides shared among multiple people.
  - **Passenger Count 5–6 (Avg: ~16.85–17.35)** Slightly lower fares; these may be short-distance group rides or anomalies.
  - **Passenger Count 7–9 (Avg: 69.73–91.33)** These are likely outliers or erroneous data entries, given the unusually high average fares.
- 
- 

```
In [6]: df_clean.groupBy("PULocationID") \
        .count() \
        .orderBy("count", ascending=False) \
        .limit(5) \
        .show()
```

```
+-----+-----+
|PULocationID| count|
+-----+-----+
|          161|161414|
|          237|158098|
|          236|149839|
|          132|134836|
|          230|118268|
+-----+-----+
```

| Rank | PULocationID | Number of Pickups |
|------|--------------|-------------------|
| 1    | 161          | 161,414           |
| 2    | 237          | 158,098           |



| Rank | PULocationID | Number of Pickups |
|------|--------------|-------------------|
| 3    | 236          | 149,839           |
| 4    | 132          | 134,836           |
| 5    | 230          | 118,268           |

---

## Analyzing Pickup Locations by Frequency

1. **Group by Pickup Location ID** The dataset is grouped according to the pickup location identifier ( `PULocationID` ) to analyze which locations have the highest number of taxi pickups.
  2. **Count Trips per Location** For each pickup location, the total number of trips originating there is counted.
  3. **Sort and Limit Results** The results are sorted in descending order by the number of pickups, and only the top five locations are displayed.
- 

Insights:

- These five PULocationIDs are the busiest pickup locations in the dataset.
  - Pickup counts in these locations are significantly higher than average,
- 
- 

```
In [7]: from pyspark.sql.functions import col, to_timestamp, to_date, avg, sum as _sum, uni

# Transform datetime columns and compute duration in minutes
df_transformed = df_clean.withColumn("pickup_datetime", to_timestamp("pickup_dateti
    .withColumn("dropoff_datetime", to_timestamp("dropoff_date
    .withColumn("trip_duration_minutes",
        (unix_timestamp("dropoff_datetime") - unix_tim
    .withColumn("trip_date", to_date("pickup_datetime"))

# Group by trip_date and calculate total trip distance and average trip duration
df_transformed.groupBy("trip_date") \
    .agg(
        _sum("trip_distance").alias("total_trip_distance"),
        avg("trip_duration_minutes").alias("avg_trip_duration_minutes")
    ) \
    .orderBy("trip_date") \
    .show()
```

| trip_date  | total_trip_distance | avg_trip_duration_minutes |
|------------|---------------------|---------------------------|
| 2024-12-31 | 76.81               | 16.805555555555557        |
| 2025-01-01 | 391678.28999999666  | 15.749463393966307        |
| 2025-01-02 | 302003.8099999971   | 16.919175767191987        |
| 2025-01-03 | 538893.4300000002   | 16.09171787126701         |
| 2025-01-04 | 307149.05999999866  | 15.134161549177637        |
| 2025-01-05 | 295058.6999999997   | 14.840657839428191        |
| 2025-01-06 | 274068.42000000016  | 15.240369651916128        |
| 2025-01-07 | 336074.5500000024   | 14.923981477614614        |
| 2025-01-08 | 318620.56000000116  | 14.771554037873857        |
| 2025-01-09 | 608176.5600000052   | 15.464664529227802        |
| 2025-01-10 | 544009.3200000017   | 15.169033019917526        |
| 2025-01-11 | 835741.7400000031   | 13.846707316019868        |
| 2025-01-12 | 802457.2899999972   | 13.956667871278771        |
| 2025-01-13 | 545132.5200000003   | 15.001899916797006        |
| 2025-01-14 | 788489.4100000055   | 15.063421423722941        |
| 2025-01-15 | 777530.0000000021   | 15.387044141267769        |
| 2025-01-16 | 1062429.6400000053  | 15.69611736473662         |
| 2025-01-17 | 848059.6699999978   | 16.066730181787868        |
| 2025-01-18 | 424121.83000000275  | 14.38218315375618         |
| 2025-01-19 | 637236.9100000036   | 13.023908918695026        |

only showing top 20 rows

## Transforming Datetime and Analyzing Trip Duration and Distance

- Convert Columns to Timestamp and Date** The pickup and dropoff datetime columns are converted into proper timestamp formats. Additionally, a new column representing the trip date (without time) is created for daily aggregation.
- Calculate Trip Duration** A new column is added to calculate the trip duration in minutes by subtracting the pickup time from the dropoff time using Unix timestamps.
- Aggregate by Trip Date** The data is grouped by the trip date to summarize daily taxi activity. For each day:
  - The total trip distance is summed.
  - The average trip duration is calculated.
- Sort and Display Results** The aggregated daily data is ordered chronologically and displayed to reveal trends over time.

## Key Observations

- Peak Distance Day:** January 16, 2025, recorded the highest total trip distance, indicating a day of particularly heavy taxi usage.

- **Longest Average Trips:** January 2, 2025, had the longest average trip duration at nearly 17 minutes.
  - **Shortest Average Trips:** January 19, 2025, showed the shortest average trip duration around 13 minutes, despite a high total distance, suggesting many short trips were made.
  - **Low Activity Day:** December 31, 2024, had unusually low total trip distance, possibly due to incomplete data for that day.
- 
- 

## Export Results

```
In [ ]: output_path = "/content/drive/My Drive/Colab Notebooks/yellow_tripdata_2025-01_clean_df_clean.write.mode("overwrite").parquet(output_path)
```

```
In [ ]: csv_output_path = "/content/drive/My Drive/Colab Notebooks/yellow_tripdata_2025-01_clean_df_clean.write.mode("overwrite").option("header", True).csv(csv_output_path)
```

---

## Exporting Results

The cleaned dataset is saved to Google Drive in two formats:

1. **Parquet Format** The DataFrame `df_clean` is written to the specified path in Parquet format. The mode is set to overwrite any existing files to ensure the latest version is saved efficiently in a columnar storage format.
2. **CSV Format** The same cleaned DataFrame is also saved as CSV files to a different path. The export includes headers for clarity, and the overwrite mode ensures previous files at the location are replaced.

Both formats allow for easy sharing and further analysis outside of Spark.

---

---

---

---

## Short Answer Reflection

---

a. What benefits does Spark offer when processing large-scale datasets like the NYC Taxi data compared to traditional pandas-based processing?

- Spark processes large datasets by distributing tasks across multiple machines, allowing it to handle much more data than pandas, which works on a single computer. This makes Spark faster and more scalable for big data.
- 

### **b. Why is it important to perform data cleaning (e.g., removing invalid trips) before aggregation and analysis?**

- Data cleaning is important because invalid or wrong data, like trips with zero distance or missing times, can give incorrect results. Cleaning ensures that the analysis is accurate and meaningful.
- 

### **c. How do Spark's lazy evaluation and distributed computation contribute to performance efficiency?**

- Spark's lazy evaluation means it waits to run tasks until absolutely needed, which helps combine steps and save time. Distributed computation splits the work across many machines, speeding up processing by doing many things at once.
- 
- 

```
In [ ]: import warnings
from pandas.errors import PerformanceWarning # Use this if PerformanceWarning need
from statsmodels.tools.sm_exceptions import ValueWarning # Import ValueWarning fro

# Suppress specific warnings
warnings.filterwarnings("ignore", category=UserWarning) # General warnings
warnings.filterwarnings("ignore", category=FutureWarning) # FutureWarning from stat
warnings.filterwarnings("ignore", category=ValueWarning) # ValueWarning for unsupp
warnings.filterwarnings("ignore", category=PerformanceWarning) # PerformanceWarnin

# Mount Google Drive
from google.colab import drive
drive.mount('/content/drive')

!jupyter nbconvert --to html "/content/drive/My Drive/Colab Notebooks/FA4_Billoness_
```