In [ ]:

```
from pyspark.sql import SparkSession

from pyspark.sql.functions import *
from pyspark.sql.types import *

spark = SparkSession \
    .builder \
    .appName("how to read csv file") \
    .enableHiveSupport() \
    .getOrCreate()
```

```
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
2023-11-26T12:54:11,660 WARN [Thread-4] org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop librar
y for your platform... using builtin-java classes where applicable
```

In [ ]:

```
refined_monthly_pays = spark.sql("select * from obligatorio.refined_monthly_pays")
```

```
2023-11-26T12:54:18,094 INFO [Thread-4] org.apache.hadoop.hive.conf.HiveConf - Found configuration file file:/home/ort/spark/conf/hive-site.xml
2023-11-26T12:54:18,500 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.metastore.wm.default.pool.size does not exist
2023-11-26T12:54:18,500 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.llap.task.scheduler.preempt.independent does not exist
2023-11-26T12:54:18,500 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.llap.output.format.arrow does not exist
2023-11-26T12:54:18,500 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.tez.llap.min.reducer.per.executor does not exist
2023-11-26T12:54:18,500 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.arrow.root.allocator.limit does not exist
2023-11-26T12:54:18,500 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.vectorized.use.checked.expressions does not exist
2023-11-26T12:54:18,500 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.tez.dynamic.semijoin.reduction.for.mapjoin does not exist
2023-11-26T12:54:18,500 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.vectorized.complex.types.enabled does not exist
2023-11-26T12:54:18,500 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.server2.wm.worker.threads does not exist
2023-11-26T12:54:18,500 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.repl.partitions.dump.parallelism does not exist
2023-11-26T12:54:18,500 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.metastore.uri.selection does not exist
2023-11-26T12:54:18,500 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.strict.checks.no.partition.filter does not exist
2023-11-26T12:54:18,500 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.tez.dynamic.semijoin.reduction.for.dpp.factor does not exist
2023-11-26T12:54:18,500 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.stats.filter.in.min.ratio does not exist
2023-11-26T12:54:18,501 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.metastore.client.cache.initial.capacity does not exist
2023-11-26T12:54:18,501 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.stats.ndv.estimate.percent does not exist
2023-11-26T12:54:18,501 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.server2.webui.cors.allowed.methods does not exist
2023-11-26T12:54:18,501 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.optimize.joinreducededuplication does not exist
2023-11-26T12:54:18,501 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.metastore.client.cache.enabled does not exist
2023-11-26T12:54:18,501 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.stats.fetch.bitvector does not exist
```

```
2023-11-26T12:54:18,501 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.disable.unsafe.e
xternal.table.operations does not exist
2023-11-26T12:54:18,501 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.materializedvie
w.rewriting.incremental does not exist
2023-11-26T12:54:18,501 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.server2.material
izedviews.registry.impl does not exist
2023-11-26T12:54:18,501 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.metastore.event.
db.notification.api.auth does not exist
2023-11-26T12:54:18,501 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.exec.orc.delta.s
treaming.optimizations.enabled does not exist
2023-11-26T12:54:18,501 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.stats.ndv.algo d
oes not exist
2023-11-26T12:54:18,501 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.spark.job.max.ta
sks does not exist
2023-11-26T12:54:18,501 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.msck.repair.batc
h.max.retries does not exist
2023-11-26T12:54:18,501 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.prewarm.spark.ti
meout does not exist
2023-11-26T12:54:18,501 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.optimize.update.
table.properties.from.serde.list does not exist
2023-11-26T12:54:18,502 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.llap.plugin.clie
nt.num.threads does not exist
2023-11-26T12:54:18,502 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.test.bucketcode
c.version does not exist
2023-11-26T12:54:18,502 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.query.reexecutio
n.enabled does not exist
2023-11-26T12:54:18,502 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.materializedvie
w.rewriting.time.window does not exist
2023-11-26T12:54:18,502 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.query.reexecutio
n.stats.cache.batch.size does not exist
2023-11-26T12:54:18,502 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.server2.webui.co
rs.allowed.headers does not exist
2023-11-26T12:54:18,502 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.join.inner.resid
ual does not exist
2023-11-26T12:54:18,502 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.server2.active.p
assive.ha.enable does not exist
2023-11-26T12:54:18,502 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.llap.io.trace.al
ways.dump does not exist
2023-11-26T12:54:18,502 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.query.reexecutio
n.stats.persist.scope does not exist
2023-11-26T12:54:18,502 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.mm.allow.origina
ls does not exist
```

```
2023-11-26T12:54:18,502 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.compactor.compact.insert.only does not exist
2023-11-26T12:54:18,502 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.txn.xlock.iow does not exist
2023-11-26T12:54:18,502 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.spark.rsc.conf.list does not exist
2023-11-26T12:54:18,502 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.llap.cache.defaultfs.only.native.fileid does not exist
2023-11-26T12:54:18,502 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.spark.optimize.shuffle.serde does not exist
2023-11-26T12:54:18,502 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.testing.remove.logs does not exist
2023-11-26T12:54:18,502 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.distcp.privileged.doAs does not exist
2023-11-26T12:54:18,502 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.strict.checks.orderby.no.limit does not exist
2023-11-26T12:54:18,502 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.metastore.client.cache.expiry.time does not exist
2023-11-26T12:54:18,503 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.llap.io.allocator.defrag.headroom does not exist
2023-11-26T12:54:18,503 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.notification.event.consumers does not exist
2023-11-26T12:54:18,503 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.vectorized.input.format.supports.enabled does not exist
2023-11-26T12:54:18,503 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.metastore.client.cache.max.capacity does not exist
2023-11-26T12:54:18,503 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.repl.dumpdir.clean.freq does not exist
2023-11-26T12:54:18,503 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.spark.use.ts.stats.for.mapjoin does not exist
2023-11-26T12:54:18,503 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.repl.dump.include.acid.tables does not exist
2023-11-26T12:54:18,503 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.server2.webui.use.pam does not exist
2023-11-26T12:54:18,503 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.query.reexecution.max.count does not exist
2023-11-26T12:54:18,503 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.llap.io.share.object.pools does not exist
2023-11-26T12:54:18,503 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.optimize.update.table.properties.from.serde does not exist
2023-11-26T12:54:18,503 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.service.metrics.codahale.reporter.classes does not exist
```

```
2023-11-26T12:54:18,503 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.tez.session.even
ts.print.summary does not exist
2023-11-26T12:54:18,503 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.llap.io.vrb.queu
e.limit.base does not exist
2023-11-26T12:54:18,503 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.mm.avoid.s3.glob
status does not exist
2023-11-26T12:54:18,503 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.repl.replica.fun
ctions.root.dir does not exist
2023-11-26T12:54:18,504 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.query.results.ca
che.max.entry.lifetime does not exist
2023-11-26T12:54:18,504 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.server2.limit.co
nnections.per.user does not exist
2023-11-26T12:54:18,504 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.server2.thrift.h
ttp.compression.enabled does not exist
2023-11-26T12:54:18,504 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.vectorized.execu
tion.ptf.enabled does not exist
2023-11-26T12:54:18,504 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.optimize.shared.
work.extended does not exist
2023-11-26T12:54:18,504 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.vectorized.row.i
dentifier.enabled does not exist
2023-11-26T12:54:18,504 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.query.reexecutio
n.always.collect.operator.stats does not exist
2023-11-26T12:54:18,504 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.repl.dumpdir.ttl
does not exist
2023-11-26T12:54:18,504 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.local.time.zone
does not exist
2023-11-26T12:54:18,504 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.server2.tez.wm.a
m.registry.timeout does not exist
2023-11-26T12:54:18,504 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.server2.active.p
assive.ha.registry.namespace does not exist
2023-11-26T12:54:18,504 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.create.as.inser
t.only does not exist
2023-11-26T12:54:18,504 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.llap.mapjoin.mem
ory.oversubscribe.factor does not exist
2023-11-26T12:54:18,504 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.arrow.batch.size
does not exist
2023-11-26T12:54:18,504 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.notification.seq
uence.lock.retry.sleep.interval does not exist
2023-11-26T12:54:18,504 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.repl.approx.max.
load.tasks does not exist
2023-11-26T12:54:18,504 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.query.results.ca
che.enabled does not exist
```

```
2023-11-26T12:54:18,504 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.legacy.schema.for.all.serdes does not exist
2023-11-26T12:54:18,504 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.tez.dag.status.check.interval does not exist
2023-11-26T12:54:18,504 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.druid.bitmap.type does not exist
2023-11-26T12:54:18,505 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.spark.dynamic.partition.pruning.map.join.only does not exist
2023-11-26T12:54:18,505 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.llap.memory.oversubscription.max.executors.per.query does not exist
2023-11-26T12:54:18,505 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.llap.io.trace.size does not exist
2023-11-26T12:54:18,505 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.llap.plugin.rpc.num.handlers does not exist
2023-11-26T12:54:18,505 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.server2.wm.allow.any.pool.via.jdbc does not exist
2023-11-26T12:54:18,505 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.vectorized.groupby.complex.types.enabled does not exist
2023-11-26T12:54:18,505 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.avro.timestamp.skip.conversion does not exist
2023-11-26T12:54:18,505 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.query.results.cache.nontransactional.tables.enabled does not exist
2023-11-26T12:54:18,505 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.stats.correlated.multi.key.joins does not exist
2023-11-26T12:54:18,505 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.metastore.db.type does not exist
2023-11-26T12:54:18,505 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.streaming.auto.flush.check.interval.size does not exist
2023-11-26T12:54:18,505 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.zookeeper.connection.timeout does not exist
2023-11-26T12:54:18,505 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.query.reexecution.strategies does not exist
2023-11-26T12:54:18,505 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.server2.limit.connections.per.user.ipaddress does not exist
2023-11-26T12:54:18,505 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.llap.mapjoin.memory.monitor.check.interval does not exist
2023-11-26T12:54:18,505 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.optimize.shared.work does not exist
2023-11-26T12:54:18,505 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.stats.estimate does not exist
2023-11-26T12:54:18,505 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.llap.io.allocator.discard.method does not exist
```

```
2023-11-26T12:54:18,505 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.tez.cartesian-pr
oduct.enabled does not exist
2023-11-26T12:54:18,506 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.notification.seq
uence.lock.max.retries does not exist
2023-11-26T12:54:18,506 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.heap.memory.moni
tor.usage.threshold does not exist
2023-11-26T12:54:18,506 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.privilege.synchr
onizer.interval does not exist
2023-11-26T12:54:18,506 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.vectorized.adapt
or.suppress.evaluate.exceptions does not exist
2023-11-26T12:54:18,506 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.materializedvie
w.rebuild.incremental does not exist
2023-11-26T12:54:18,506 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.query.results.ca
che.max.entry.size does not exist
2023-11-26T12:54:18,506 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.spark.stage.max.
tasks does not exist
2023-11-26T12:54:18,506 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.testing.short.lo
gs does not exist
2023-11-26T12:54:18,506 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.streaming.auto.f
lush.enabled does not exist
2023-11-26T12:54:18,506 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.spark.explain.us
er does not exist
2023-11-26T12:54:18,506 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.describe.partiti
onedtable.ignore.stats does not exist
2023-11-26T12:54:18,506 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.server2.operatio
n.log.cleanup.delay does not exist
2023-11-26T12:54:18,506 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.repl.dump.metada
ta.only does not exist
2023-11-26T12:54:18,506 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.optimize.countdi
stinct does not exist
2023-11-26T12:54:18,506 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.auto.convert.joi
n.shuffle.max.size does not exist
2023-11-26T12:54:18,506 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.llap.plugin.acl
does not exist
2023-11-26T12:54:18,506 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.metastore.schem
a.info.class does not exist
2023-11-26T12:54:18,507 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.server2.tez.queu
e.access.check does not exist
2023-11-26T12:54:18,507 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.llap.external.sp
lits.temp.table.storage.format does not exist
2023-11-26T12:54:18,507 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.llap.io.row.wrap
per.enabled does not exist
```

```
2023-11-26T12:54:18,507 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.constraint.notnu
ll.enforce does not exist
2023-11-26T12:54:18,507 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.cli.print.escap
e.crlf does not exist
2023-11-26T12:54:18,507 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.trigger.validati
on.interval does not exist
2023-11-26T12:54:18,507 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.server2.webui.co
rs.allowed.origins does not exist
2023-11-26T12:54:18,507 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.server2.limit.co
nnections.per.ipaddress does not exist
2023-11-26T12:54:18,507 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.llap.external.sp
lits.order.by.force.single.split does not exist
2023-11-26T12:54:18,507 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.metastore.clien
t.cache.stats.enabled does not exist
2023-11-26T12:54:18,507 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.notification.eve
nt.poll.interval does not exist
2023-11-26T12:54:18,507 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.transactional.co
ncatenate.noblock does not exist
2023-11-26T12:54:18,507 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.materializedvie
w.rewriting.strategy does not exist
2023-11-26T12:54:18,507 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.vectorized.if.ex
pr.mode does not exist
2023-11-26T12:54:18,507 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.exim.test.mode d
oes not exist
2023-11-26T12:54:18,507 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.query.results.ca
che.directory does not exist
2023-11-26T12:54:18,507 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.query.results.ca
che.wait.for.pending.results does not exist
2023-11-26T12:54:18,508 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.remove.orderby.i
n.subquery does not exist
2023-11-26T12:54:18,508 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.tez.bmj.use.subc
ache does not exist
2023-11-26T12:54:18,508 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.llap.io.vrb.queu
e.limit.min does not exist
2023-11-26T12:54:18,508 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.server2.wm.pool.
metrics does not exist
2023-11-26T12:54:18,508 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.repl.add.raw.res
erved.namespace does not exist
2023-11-26T12:54:18,508 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.resource.use.hdf
s.location does not exist
2023-11-26T12:54:18,508 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.stats.num.nulls.
estimate.percent does not exist
```

```
2023-11-26T12:54:18,508 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.llap.io.acid does not exist
2023-11-26T12:54:18,508 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.llap.zk.sm.session.timeout does not exist
2023-11-26T12:54:18,508 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.vectorized.ptf.max.memory.buffering.batch.count does not exist
2023-11-26T12:54:18,508 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.llap.task.scheduler.am.registry does not exist
2023-11-26T12:54:18,508 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.druid.overlord.address.default does not exist
2023-11-26T12:54:18,508 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.optimize.remove.sq_count_check does not exist
2023-11-26T12:54:18,508 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.server2.webui.enable.cors does not exist
2023-11-26T12:54:18,508 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.vectorized.row.serde.inputformat.excludes does not exist
2023-11-26T12:54:18,508 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.query.reexecution.stats.cache.size does not exist
2023-11-26T12:54:18,508 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.combine.equivalent.work.optimization does not exist
2023-11-26T12:54:18,508 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.lock.query.string.max.length does not exist
2023-11-26T12:54:18,508 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.llap.io.track.cache.usage does not exist
2023-11-26T12:54:18,509 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.use.orc.codec.pool does not exist
2023-11-26T12:54:18,509 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.query.results.cache.max.size does not exist
2023-11-26T12:54:18,509 WARN [Thread-4] org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.repl.bootstrap.dump.open.txn.timeout does not exist
2023-11-26T12:54:18,695 INFO [Thread-4] hive.metastore - Trying to connect to metastore with URI thrift://localhost:9083
2023-11-26T12:54:18,720 INFO [Thread-4] hive.metastore - Opened a connection to metastore, current connections: 1
2023-11-26T12:54:18,761 INFO [Thread-4] hive.metastore - Connected to metastore.
```

In [ ]:

```
refined_monthly_pays.show()
```

```
[Stage 0:>                                                          (0 + 1) / 1]
```

```
+-------+----+----------------+--------------+-------------+---------------------+-----------------------------+
|user_id|plan|mensual_cost_usd|local_currency|      country|has_monthly_purchases|purchase_value_local_currency|
+-------+----+----------------+--------------+-------------+---------------------+-----------------------------+
|    783|   B|             113|       Afghani|  Afghanistan|                 true|                       1209.0|
|    829|   C|             140|       Afghani|  Afghanistan|                false|                          0.0|
|    835|   D|             100|       Afghani|  Afghanistan|                 true|                      2785.16|
|    412|   C|              36|          Euro|Aland Islands|                 true|                      2320.89|
|     67|   B|             149|           Lek|      Albania|                 true|                      1535.58|
|    104|   D|              48|           Lek|      Albania|                 true|                      2433.05|
|    879|   A|              78|           Lek|      Albania|                 true|                      1465.89|
|     47|   B|              99|          Peso|    Argentina|                false|                          0.0|
|     52|   C|             111|          Peso|    Argentina|                 true|                       259.11|
|    111|   D|             121|          Peso|    Argentina|                false|                          0.0|
|    143|   C|             104|          Peso|    Argentina|                 true|                       699.79|
|    164|   D|              66|          Peso|    Argentina|                 true|                      1272.38|
|    262|   B|              92|          Peso|    Argentina|                 true|                       2472.3|
|    306|   A|              76|          Peso|    Argentina|                false|                          0.0|
|    323|   C|              41|          Peso|    Argentina|                false|                          0.0|
|    339|   D|             143|          Peso|    Argentina|                 true|                       157.22|
|    463|   C|              42|          Peso|    Argentina|                false|                          0.0|
|    512|   D|             122|          Peso|    Argentina|                 true|                      2220.36|
|    518|   D|              90|          Peso|    Argentina|                 true|                       2297.9|
|    561|   B|             119|          Peso|    Argentina|                 true|                      2002.05|
+-------+----+----------------+--------------+-------------+---------------------+-----------------------------+
only showing top 20 rows
```

In [ ]:

```
refined_monthly_pays.printSchema()
```

```
root
 |-- user_id: string (nullable = true)
 |-- plan: string (nullable = true)
 |-- mensual_cost_usd: integer (nullable = true)
 |-- local_currency: string (nullable = true)
 |-- country: string (nullable = true)
 |-- has_monthly_purchases: boolean (nullable = true)
 |-- purchase_value_local_currency: float (nullable = true)
```

In [ ]:

```
refined_fecha = spark.sql("select * from obligatorio.refined_fecha")
```

In [ ]:

```
refined_fecha.show()
```

```
+---+----------+----------+
| id|fecha_alta|fecha_baja|
+---+----------+----------+
|  1|2021-06-02|      null|
|  2|2021-06-02|      null|
|  3|2021-06-03|      null|
|  4|2021-06-03|2022-03-13|
|  5|2021-06-04|2023-11-05|
|  6|2021-06-04|2022-07-03|
|  7|2021-06-05|      null|
|  8|2021-06-08|      null|
|  9|2021-06-09|      null|
| 10|2021-06-11|      null|
| 11|2021-06-12|      null|
| 12|2021-06-12|      null|
| 13|2021-06-13|      null|
| 14|2021-06-13|      null|
| 15|2021-06-15|      null|
| 16|2021-06-15|      null|
| 17|2021-06-16|      null|
| 18|2021-06-16|      null|
| 19|2021-06-18|      null|
| 20|2021-06-19|      null|
+---+----------+----------+
only showing top 20 rows
```

In [ ]:

```
refined_plans = spark.sql("select * from obligatorio.refined_plans")
```

In [ ]:

```
refined_plans.show()
```

```
+----+--------+----------+
|plan|  nombre|created_at|
+----+--------+----------+
|   A|   Basic|2021-01-04|
|   B|Original|2021-01-08|
|   C|  Family|2021-01-02|
|   D| Premium|2021-01-02|
+----+--------+----------+
```

In [ ]:

```
refined_users = spark.sql("select * from obligatorio.refined_users")
```

In [ ]:

```
refined_users.show()
```

```
2023-11-26T12:54:28,681 WARN [Executor task launch worker for task 0.0 in stage 3.0 (TID 3)] org.apache.hadoop.hive.s
erde2.lazy.LazyStruct - Extra bytes detected at the end of the row! Ignoring similar problems.
+-------+----------+-----------+--------------------+-----------+---+---------------------------+--------------------
-----+
|user_id|first_name|  last_name|               email|     gender|age|principal_sport_of_interest|principal_sport_frec
uency|
+-------+----------+-----------+--------------------+-----------+---+---------------------------+--------------------
-----+
|      1|      Alic|    Kaemena|akaemena0@4shared...|       Male| 18|                 Senderismo|                 Han
dball|
|      2|    Dorrie|   Fountian|dfountian1@craigs...|     Female| 51|                    Running|               Sende
rismo|
|      3|   Spenser|  Fernandes|sfernandes2@devhu...|       Male| 31|                   Natación|                   T
ennis|
|      4|    Andris|    Moakson|amoakson3@rediff.com|       Male| 22|                     Tennis|                   V
olley|
|      5|     Elsey|   Rathbone|erathbone4@bloglo...|     Female| 23|                   Natación|                   T
ennis|
|      6|  Vivianna|     Naerup|vnaerup5@sourcefo...|     Female| 46|                    Rafting|                   F
útbol|
|      7|    Jobina|     Toller|jtoller6@cbslocal...|     Female| 56|                   Ciclismo|                   T
ennis|
|      8|    Debbie|    Ambrodi|dambrodi7@seattle...|     Female| 51|                   Trecking|                  Ra
fting|
|      9|   Kathryn|  Lissenden|klissenden8@quant...|     Female| 53|                     Tennis|                  Ru
nning|
|     10|   Jeannie|Prettejohns|jprettejohns9@the...| Non-binary| 24|                   Trecking|                  Ru
nning|
|     11|      Nike|     Clelle|nclellea@csmonito...|Genderfluid| 33|                   Natación|                   F
útbol|
|     12|  Shoshana|    Tenaunt|stenauntb@mozilla...|     Female| 49|                    Running|               Sende
rismo|
|     13|        Em|    Albrook|ealbrookc@netscap...|     Female| 29|                    Running|                 Cic
lismo|
|     14|     Judah|     Fearon|  jfearond@phpbb.com|       Male| 36|                   Handball|                   F
útbol|
|     15|      Jock|Mothersdale|jmothersdalee@wik...|       Male| 58|                   Handball|                   V
olley|
|     16| Kristofor|   Holstein|kholsteinf@nature...|       Male| 51|                   Natación|                   F
útbol|
|     17|     Edwin|    Battams|ebattamsg@pagineg...|       Male| 38|                     Fútbol|                 Han
dball|
|     18|   Stefano|     Truman|strumanh@xinhuane...|       Male| 24|                    Natación|                    V
olley|
|     19|  Mohammed|    Kincaid|mkincaidi@columbi...|       Male| 69|                      Volley|                   Ru
nning|
|     20|      Tova|    Chazier|  tchazierj@nasa.gov|     Female| 65|                    Natación|                    F
útbol|
+-------+----------+-----------+--------------------+-----------+---+----------------------------+--------------------
-----+
only showing top 20 rows
```

¿Cuáles son las actividades más requeridas?

In [ ]:

```
spark.sql("""
    SELECT principal_sport_of_interest,
    COUNT (DISTINCT user_id) as users_by_sports
    FROM obligatorio.refined_users
    GROUP BY principal_sport_of_interest
    ORDER BY users_by_sports DESC
""").show(truncate=False)
```

```
2023-11-26T12:54:29,906 WARN [Executor task launch worker for task 0.0 in stage 4.0 (TID 4)] org.apache.hadoop.hive.s
erde2.lazy.LazyStruct - Extra bytes detected at the end of the row! Ignoring similar problems.
+---------------------------+---------------+
|principal_sport_of_interest|users_by_sports|
+---------------------------+---------------+
|Running                    |112            |
|Tennis                     |105            |
|Ciclismo                   |105            |
|Natación                   |104            |
|Senderismo                 |99             |
|Fútbol                     |99             |
|Rafting                    |98             |
|Handball                   |95             |
|Trecking                   |93             |
|Volley                     |90             |
+---------------------------+---------------+
```

¿Cuál es la evolución mensual en cuanto a cantidad de nuevos usuarios y cual fue la distribución del ingreso por plan en el último mes?

Se asume que la información de la tabla monthly_pays corresponde al último mes de analisis.

Evolución a nivel de nuevos usuarios.

In [ ]:

```
spark.sql("""
    SELECT TRUNC(fecha_alta,'MM') as mes_alta,
    COUNT(DISTINCT id) as users
    FROM obligatorio.refined_fecha
    GROUP BY 1
    ORDER BY mes_alta ASC
""").show(truncate=False)
```

```
+----------+-----+
|mes_alta  |users|
+----------+-----+
|2021-06-01|43   |
|2021-07-01|73   |
|2021-08-01|66   |
|2021-09-01|52   |
|2021-10-01|80   |
|2021-11-01|71   |
|2021-12-01|49   |
|2022-01-01|44   |
|2022-02-01|56   |
|2022-03-01|39   |
|2022-04-01|48   |
|2022-05-01|34   |
|2022-06-01|52   |
|2022-07-01|38   |
|2022-08-01|24   |
|2022-09-01|28   |
|2022-10-01|27   |
|2022-11-01|21   |
|2022-12-01|30   |
|2023-01-01|28   |
+----------+-----+
only showing top 20 rows
```

Evolución a nivel de ingresos por plan.

In [ ]:

```
spark.sql("""
    SELECT p.nombre as plan,
    SUM(m.mensual_cost_usd) as revenue
    FROM obligatorio.refined_monthly_pays as m
    LEFT JOIN obligatorio.refined_plans as p on p.plan = m.plan
    GROUP BY p.nombre
""").show(truncate=False)
```

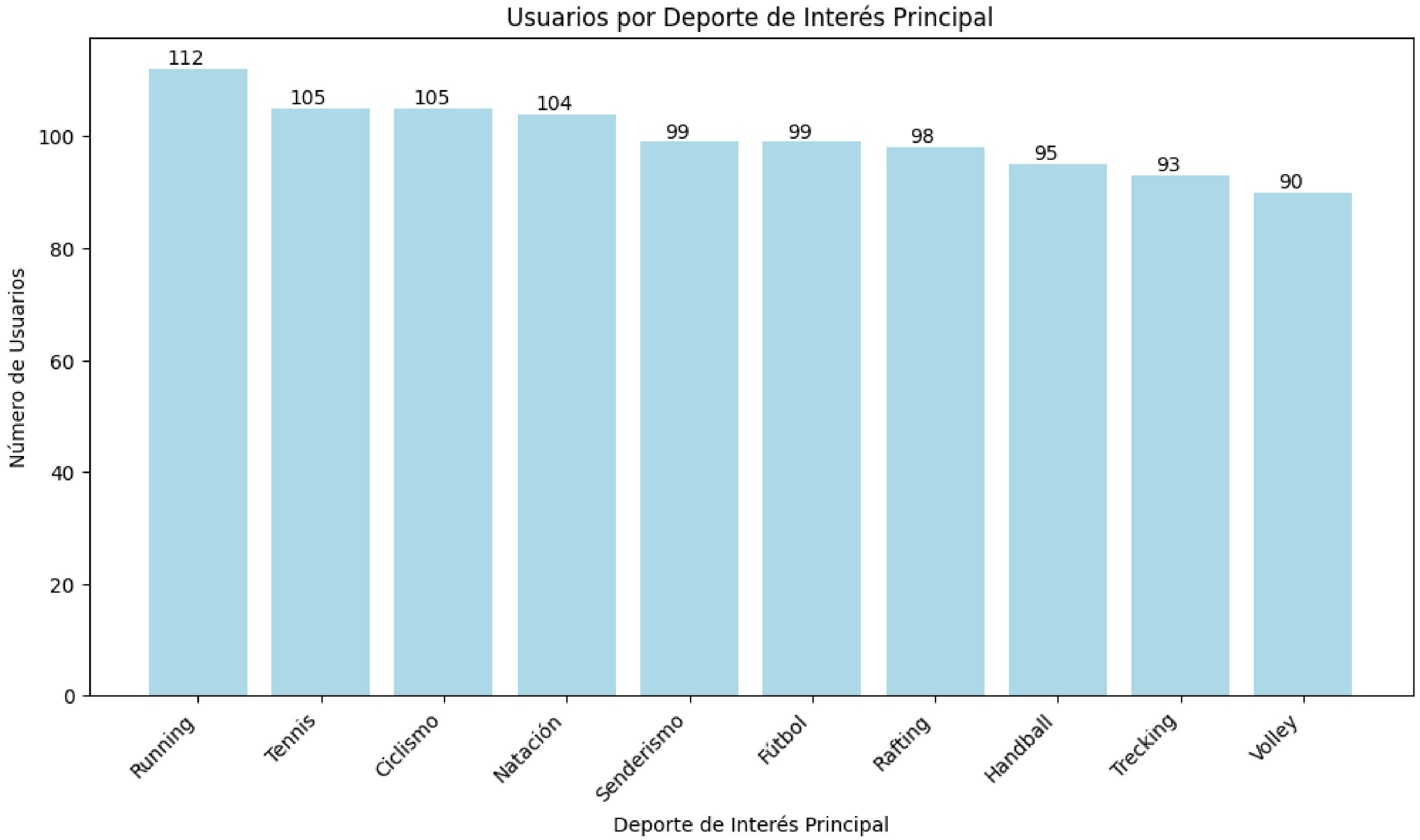
```
+--------+-------+
|plan    |revenue|
+--------+-------+
|Premium |22816  |
|Original|23990  |
|Family  |22600  |
|Basic   |22776  |
+--------+-------+
```

Visualizaciones

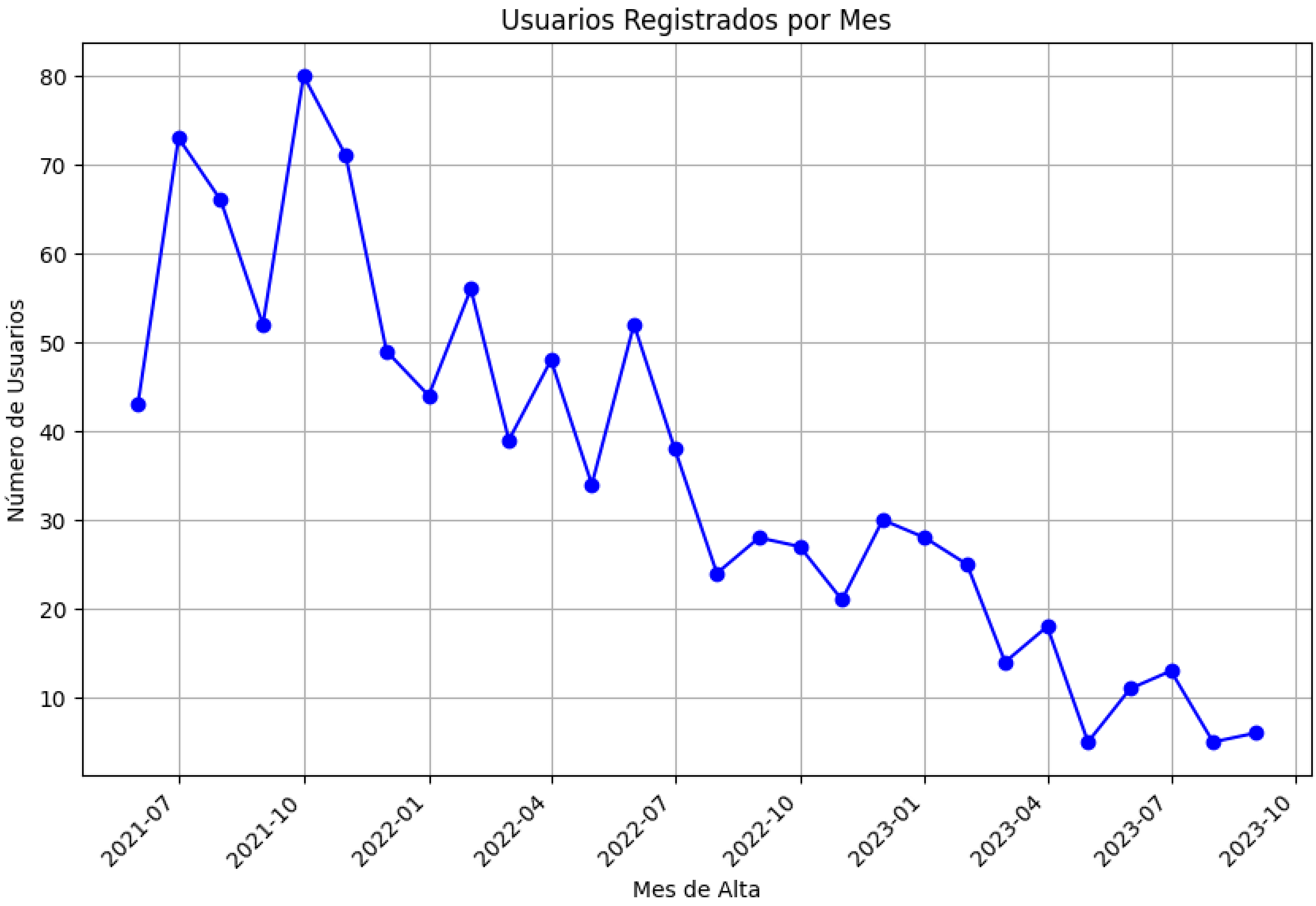
In [ ]:

```
import matplotlib.pyplot as plt
import pandas as pd

result_df = spark.sql("""
    SELECT principal_sport_of_interest,
    COUNT(DISTINCT user_id) as users_by_sports
    FROM obligatorio.refined_users
    GROUP BY principal_sport_of_interest
    ORDER BY users_by_sports DESC
""").toPandas()

plt.figure(figsize=(12, 6))
bars = plt.bar(result_df['principal_sport_of_interest'], result_df['users_by_sports'], color='lightblue')
plt.xlabel('Deporte de Interés Principal')
plt.ylabel('Número de Usuarios')
plt.title('Usuarios por Deporte de Interés Principal')
plt.xticks(rotation=45, ha='right')  # Rotar etiquetas del eje x para mayor legibilidad


for bar, label in zip(bars, result_df['users_by_sports']):
    plt.text(bar.get_x() + bar.get_width() / 2 - 0.1, bar.get_height() + 0.1, label, ha='center', va='bottom')
plt.show()
```

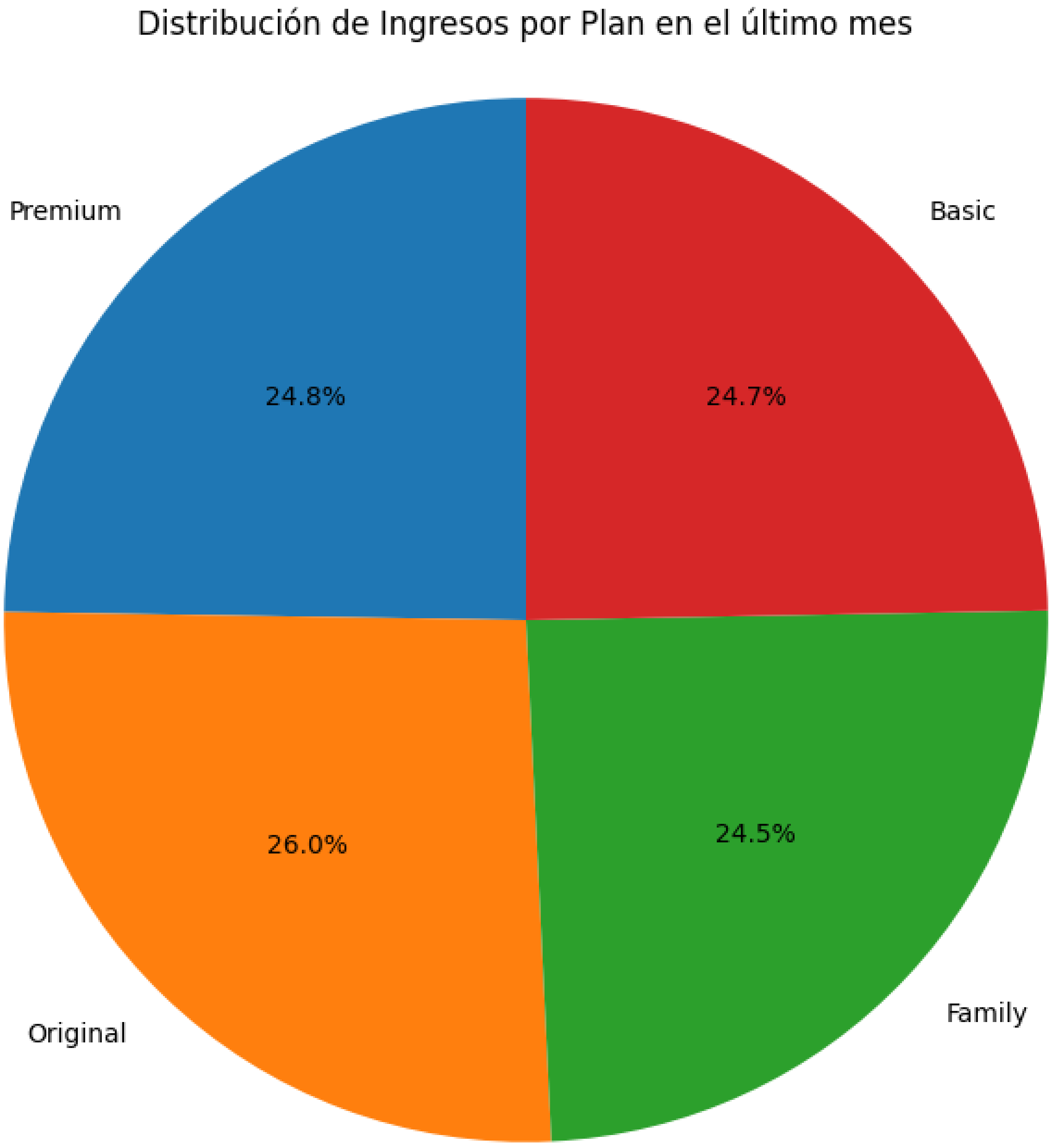
```
2023-11-26T12:54:34,956 WARN [Executor task launch worker for task 0.0 in stage 20.0 (TID 13)] org.apache.hadoop.hiv
e.serde2.lazy.LazyStruct - Extra bytes detected at the end of the row! Ignoring similar problems.
```



```
In [ ]: result_df = spark.sql("""
            SELECT TRUNC(fecha_alta,'MM') as mes_alta,
            COUNT(DISTINCT id) as users
            FROM obligatorio.refined_fecha
            GROUP BY 1
            ORDER BY mes_alta ASC
        """).toPandas()

        plt.figure(figsize=(10, 6))
        plt.plot(result_df['mes_alta'], result_df['users'], marker='o', linestyle='-', color='b')
        plt.xlabel('Mes de Alta')
        plt.ylabel('Número de Usuarios')
        plt.title('Usuarios Registrados por Mes')
        plt.xticks(rotation=45, ha='right')  # Rotar etiquetas del eje x para mayor legibilidad
        plt.grid(True)
        plt.show()
```



```
In [ ]: result_df = spark.sql("""
            SELECT p.nombre as plan,
            SUM(m.mensual_cost_usd) as revenue
            FROM obligatorio.refined_monthly_pays as m
            LEFT JOIN obligatorio.refined_plans as p on p.plan = m.plan
            GROUP BY p.nombre
        """).toPandas()

        plt.figure(figsize=(8, 8))
        plt.pie(result_df['revenue'], labels=result_df['plan'], autopct='%1.1f%%', startangle=90)
        plt.axis('equal')
        plt.title('Distribución de Ingresos por Plan en el último mes')
        plt.show()
```


Distribución de Ingresos por Plan en el último mes
Premium
24.8%
Basic
24.7%
Original
26.0%
Family
24.5%