**Part 1 Code**

print('Printing reserves with specified schema')

reserves.printSchema()

#question\_1

print('Question 1 Output')

q1\_query = spark.sql('SELECT sid, sname, age FROM sailors WHERE rating > 6')

q1\_query.show()

#question\_2

print('Question 2 Output')

q2 = reserves.filter(reserves.bid != 101)\

.groupBy(reserves.sid)\

.agg(f.count(reserves.bid))

q2.show()

#question\_3

print('Question 3 Output')

q3 = sailors.join(reserves, ['sid'], 'inner')\

.groupBy(sailors.sid, sailors.sname)\

.agg(f.countDistinct(reserves.bid))

q3.show()

#question\_4

print('Question 4 Output')

q4 = tracks.join(artist\_term, ['artistID'], 'left')\

.groupBy(artist\_term.term)\

.agg(f.min(tracks.year), f.avg(tracks.duration).alias('avg\_duration'), f.count(tracks.artistID))\

.orderBy(f.col('avg\_duration').desc())

q4.show(10)

#question\_5

print('Question 5 Output: top 10')

q5\_top = artist\_term.join(tracks, ['artistID'], 'left')\

.groupBy(artist\_term.term)\

.agg(f.countDistinct(tracks.trackID).alias('num\_track'))\

.orderBy(f.col('num\_track').desc())

q5\_top.show(10)

print('Question 5 Output: bottom 10')

q5\_bottom = artist\_term.join(tracks, ['artistID'], 'left')\

.groupBy(artist\_term.term)\

.agg(f.countDistinct(tracks.trackID).alias('num\_track'))\

.orderBy('num\_track')

q5\_bottom.show(10)

# Only enter this block if we're in main

if \_\_name\_\_ == "\_\_main\_\_":

# Create the spark session object

spark = SparkSession.builder.appName('part1').getOrCreate()

# Get user netID from the command line

netID = getpass.getuser()

# Call our main routine

main(spark, 'yl7576')

**Part 1 Output**

LogContents:

Lab 3 Example dataframe loading and SQL query

Printing boats inferred schema

root

|-- \_c0: string (nullable = true)

|-- \_c1: string (nullable = true)

|-- \_c2: string (nullable = true)

Printing sailors inferred schema

root

|-- age: double (nullable = true)

|-- rating: long (nullable = true)

|-- sid: long (nullable = true)

|-- sname: string (nullable = true)

Reading boats.txt and specifying schema

Printing boats with specified schema

root

|-- bid: integer (nullable = true)

|-- bname: string (nullable = true)

|-- color: string (nullable = true)

Example 1: Executing SELECT count(\*) FROM boats with SparkSQL

+--------+

|count(1)|

+--------+

| 4|

+--------+

Printing artist\_term with specified schema

root

|-- artistID: string (nullable = true)

|-- term: string (nullable = true)

Printing tracks with specified schema

root

|-- trackID: string (nullable = true)

|-- title: string (nullable = true)

|-- release: string (nullable = true)

|-- year: integer (nullable = true)

|-- duration: float (nullable = true)

|-- artistID: string (nullable = true)

Printing reserves with specified schema

root

|-- bid: long (nullable = true)

|-- date: string (nullable = true)

|-- sid: long (nullable = true)

Question 1 Output

+---+-------+----+

|sid| sname| age|

+---+-------+----+

| 22|dusting|45.0|

| 31| lubber|55.5|

| 32| andy|25.5|

| 58| rusty|35.0|

| 64|horatio|16.0|

| 71| zorba|35.0|

| 74|horatio|25.5|

+---+-------+----+

Question 2 Output

+---+----------+

|sid|count(bid)|

+---+----------+

| 22| 3|

| 31| 3|

| 74| 1|

| 64| 1|

+---+----------+

Question 3 Output

+---+-------+-------------------+

|sid| sname|count(DISTINCT bid)|

+---+-------+-------------------+

| 64|horatio| 2|

| 22|dusting| 4|

| 31| lubber| 3|

| 74|horatio| 1|

+---+-------+-------------------+

Question 4 Output

+--------------------+---------+------------------+---------------+

| term|min(year)| avg\_duration|count(artistID)|

+--------------------+---------+------------------+---------------+

| bhajan| 0|1147.2648391723633| 4|

| svart metal| 2010| 1003.232177734375| 1|

| experimental noise| 0|1001.5255482991537| 3|

| mixed jazz| 1998| 950.281982421875| 1|

| heavy psychedelic| 1996| 917.0281372070312| 1|

| middle eastern pop| 0| 845.2958374023438| 2|

| pakistani| 0| 810.4124545142764| 21|

| nippon| 0| 805.9815979003906| 2|

|electronic music ...| 0| 768.0648345947266| 6|

| early electronic| 0| 768.0648345947266| 6|

+--------------------+---------+------------------+---------------+

only showing top 10 rows

Question 5 Output: top 10

+----------------+---------+

| term|num\_track|

+----------------+---------+

| rock| 21796|

| electronic| 17740|

| pop| 17129|

|alternative rock| 11402|

| hip hop| 10926|

| jazz| 10714|

| united states| 10345|

| pop rock| 9236|

| alternative| 9209|

| indie| 8569|

+----------------+---------+

only showing top 10 rows

Question 5 Output: bottom 10

+--------------------+---------+

| term|num\_track|

+--------------------+---------+

| sense| 0|

| trance melodic| 0|

| funky jazz| 0|

| milled pavement| 0|

| stonersludgecore| 0|

| fonal records| 0|

| ambient metal| 0|

| finnish death metal| 0|

|progressive melod...| 0|

| techcore| 0|

+--------------------+---------+

only showing top 10 rows