Stat 231 – SQL and R Example Solution

Programming/Statistical Language "Bingo"

At your tables, see how many squares you and your peers can fill in. Fill in individually and then compare around the table.

Instructions:

- Checkmark a square if you've heard of the language/program.
- Shade it in ¼ of the way if you've seen code from the language.
- Shade it in ½ of the way if you've learned some of the language yourself (have or could code in it).
- Shade it in nearly completely (make sure we can see the name of the language/program) if you have plenty of experience with the language / could teach someone else about it (you all should be at this level with R).

Prof. W.

Matlab / Octave (free version)	✓ Stata	Minitab	C++ or C
SAS	R	Python	HTML and/or CSS (not really considered programming languages)
Fortran	Java	SQL	SPSS
✓ S / S-Plus	Excel (not recommended but you can write scripts and do analysis)	Maple or Mathematica (more math-centric)	Spark / HTCondor / Bright Cluster (or any other various programs related to cluster computing)

Note: You are NOT expected to be able to fill all this in. It's to give you a sense of what's out there!

SQL versus R

Answer the following questions with your peers at your tables.

1. Instead of the R command glimpse(), you could use this command to see what is inside a table in SQL.

DESCRIBE and EXPLAIN (very similar)

2. The JOIN command in SQL does which type(s) of joins?

Left-join Right-join Full-join inner-join Anti-join

may be specific to MySQL

3. The SQL command LIMIT acts like (or can be used like) what R command(s)?

filter() head() separate() slice() tail()

→ with some extra info yes

4. When writing a query in SQL, instead of specifying the dataset like in R with "data =" or by piping a dataset into a command, you must specify the dataset using this SQL command.

FROM

5. In R, you can use the mean() function, but you better use this in SQL.

AVG

6. The following SQL query performs actions similar to what R commands? See how many you can list.

```
SELECT
  dest, SUM(1) AS numFlights,
  MIN(arr_delay) AS min_arr_delay
FROM flights
```

Ideally would have GROUP BY dest @ end

select, rename, mutate / summarize depending on how you think about it

7. SQL uses `AND` and `OR` but in R we are used to seeing these instead.

& and | ampersand and bar

8. In R, suppose you want to arrange(). Then in SQL, you would use this.

ORDER BY

9. Examine the following SQL query and output (which uses the familiar airlines database). Explain what the query does in a few sentences.

```
SELECT
  carrier, SUM(1) AS numFlights,
  MAX(dep_delay) AS max_dep_delay
FROM flights
WHERE year = 2016
  AND origin = 'BDL'
GROUP BY carrier
HAVING numFlights > 365
ORDER BY max_dep_delay DESC
LIMIT 0, 5
```

Table 1: 5 records

carrier	numFlights	max_dep_delay
DL	4337	969
WN	6414	809
UA	1286	685
AA	3686	429
OO	488	408

Finds top 5 carriers in terms of max departure delay from Hartford in 2016 out of carriers with an avg of 1 flight out a day or more. Reports carrier, # of flights, and max dep. delay.

10. Prepare a SQL query based on the following instructions:

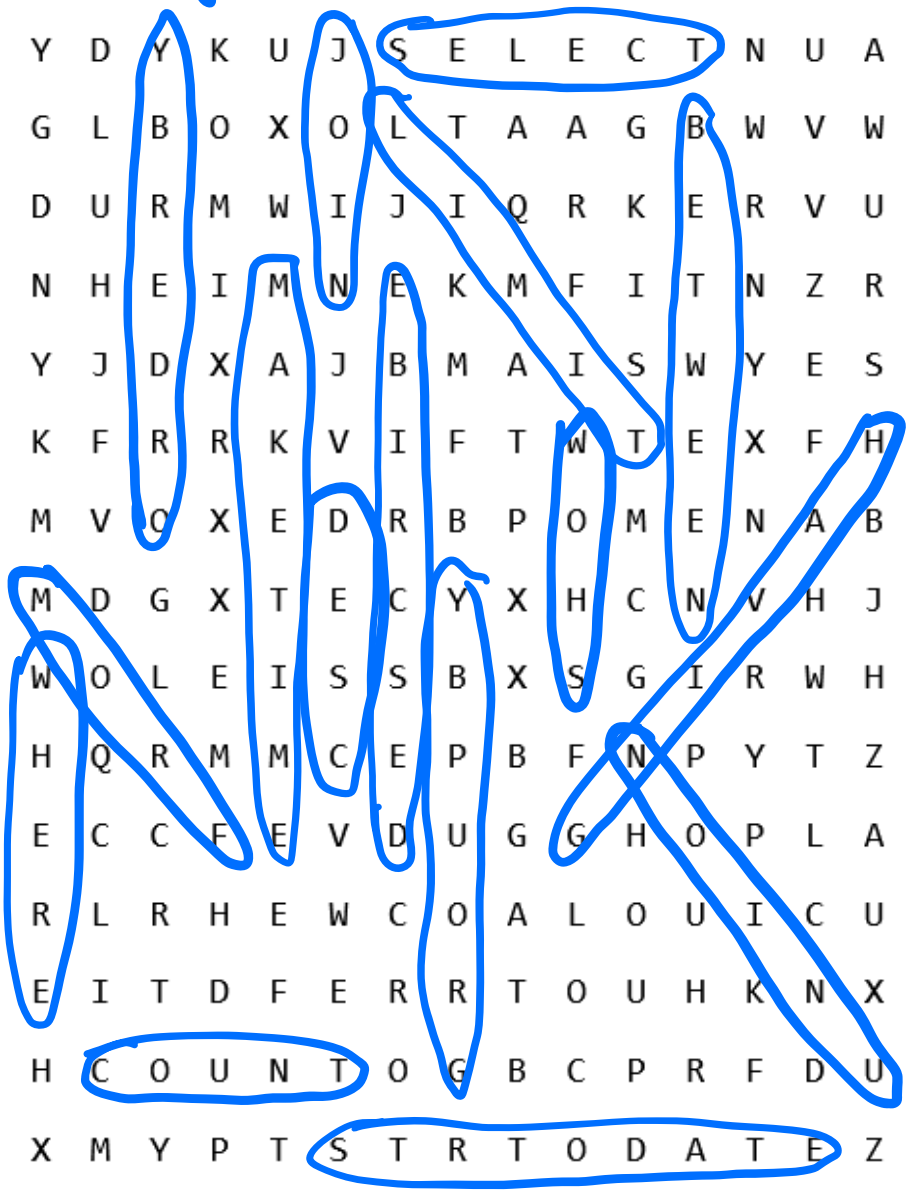
- We want to examine flights into BDL in 2017 from airline carriers with more than 2 flights on average in a day.
- We want to find the carriers, if any, with an average arrival delay (arr_delay) of more than 30 minutes.

Note: This will likely return 0 carriers, but what would the query look like?

you could constrain on an avg value instead like this

```
SELECT
  carrier, SUM(1) AS numFlights,
  AVG(arr_delay) AS avg_arr_delay
FROM flights
WHERE year=2017 AND dest='BDL'
GROUP BY carrier
HAVING numFlights > 365*2 AND
  avg_arr_delay > 30
```

SQL Command and Keyword (Word) Search (Optional!) – Do you know what each command or keyword is used for? They were all demo-ed in the text!



BETWEEN
COUNT
DESC
DESCRIBE
FROM
GROUPBY
HAVING
JOIN
LIMIT
MAKETIME
ORDERBY
SELECT
SHOW
STRTODATE
UNION
WHERE