

Copyright Jana Schaich Borg/Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)

MySQL Exercise 11: Queries that Test Relationships Between Test Completion and Dog Characteristics

This lesson we are going to integrate all the SQL syntax we've learned so far to start addressing questions in our Dognition Analysis Plan. I summarized the reasons having an analysis plan is so important in the "Start with an Analysis Plan" video accompanying this week's materials. Analysis plans ensure that you will address questions that are relevant to your business objectives as quickly and efficiently as possible. The quickest way to narrow in the factors in your analysis plan that are likely to create new insights is to combine simple SQL calculations with visualization programs, like Tableau, to identify which factors under consideration have the strongest effects on the business metric you are tasked with improving. You can then design more nuanced statistical models in other software, such as R, based on the factors you have confirmed are likely to be important for understanding and changing your business metric.



I describe a method for designing analysis plans in the Data Visualization and Communication with Tableau course earlier in this Specialization. I call that method Structured Pyramid Analysis Plans, or "sPAPs". I have provided a skeleton of an sPAP for the Dognition data set with the materials for this course that I will use as a road map for the queries we will design and practice in the next two lessons. To orient you, the SMART goal of the analysis project is at the top of the pyramid. This is a specific, measurable, attainable, relevant, and time-bound version of the general project objective, which is to make a recommendation to Dognition about what they could do to increase the number of tests customers complete. The variables you will use to assess the goal should be filled out right under where the SMART goal is written. Then under those variables, you will see ever-widening layers of categories and sub-categories of issues that will be important to analyze in order to achieve your SMART goal.

In this lesson, we will write queries to address the issues in the left-most branch of the sPAP. These issues all relate to "Features of Dogs" that could potentially influence the number of tests the dogs will ultimately complete. We will spend a lot of time discussing and practicing how to translate

analysis questions described in words into queries written in SQL syntax.

To begin, load the sql library and database, and make the Dognition database your default database:

```
In [1]: %load_ext sql
%sql mysql://studentuser:studentpw@mysqlserver/dognitiondb
%sql USE dognitiondb
```

0 rows affected.

Out[1]: []



In order to make it easier to practice SQL queries with meaningful examples before we learned how to join tables, I added extra columns to the "dogs" table that were not in the original Dognition database. These extra columns included the "total_tests_completed" field and multiple inter-test-interval ("iti") summary fields. Please do NOT try to use these extra fields in the query exercises below. Since you now know how to join tables, we will practice writing queries as if you only had the data provided in the original Dognition database.

1. Assess whether Dognition personality dimensions are related to the number of tests completed

The first variable in the Dognition sPAP we want to investigate is Dognition personality dimensions. Recall from the "Meet Your Dognition Data" video and the written description of the Dognition Data Set included with the Week 2 materials that Dognition personality dimensions represent distinct combinations of characteristics assessed by the Dognition tests. It is certainly plausible that certain personalities of dogs might be more or less likely to complete tests. For example, "einstein" dogs might be particularly likely to complete a lot of tests.

To test the relationship between Dognition personality dimensions and test completion totals, we need a query that will output a summary of the number of tests completed by dogs that have each of the Dognition personality dimensions. The features you will need to include in your query are foreshadowed by key words in this sentence. First, the fact that you need a summary of the number of tests completed suggests you will need an aggregation function. Next, the fact that you want a different summary for each personality dimension suggests that you will need a GROUP BY clause. Third, the fact that you need a "summary of the number of tests completed" rather than just a "summary of the tests completed" suggests that you might have to have multiple stages of aggregations, which in turn might mean that you will need to use a subquery.

Let's build the query step by step.

Question 1: To get a feeling for what kind of values exist in the Dognition personality dimension column, write a query that will output all of the distinct values in the dimension column. Use your relational schema or the course materials to determine what table the dimension column is in. Your output should have 11 rows.

```
In [2]: %%sql  
SELECT DISTINCT dimension  
FROM dogs
```

11 rows affected.

Out[2]:

dimension
charmer
protodog
None
einstein
stargazer
maverick
socialite
ace
expert
renaissance-dog

The results of the query above illustrate there are NULL values (indicated by the output value "none") in the dimension column. Keep that in mind in case it is relevant to future queries.

We want a summary of the total number of tests completed by dogs with each personality dimension. In order to calculate those summaries, we first need to calculate the total number of tests completed by each dog. We can achieve this using a subquery. The subquery will require data from both the dogs and the complete_tests table, so the subquery will need to include a join. We are only interested in dogs who have completed tests, so an inner join is appropriate in this case.

Question 2: Use the equijoin syntax (described in MySQL Exercise 8) to write a query that will output the Dognition personality dimension and total number of tests completed by each unique DogID. This query will be used as an inner subquery in the next question. LIMIT your output to 100 rows for troubleshooting purposes.

```
In [3]: %%sql
SELECT d.dog_guid, d.dimension, COUNT(c.created_at) AS num_tests
FROM dogs d, complete_tests c
WHERE d.dog_guid = c.dog_guid
GROUP BY d.dog_guid
LIMIT 20
```

20 rows affected.

Out[3]:

dog_guid	dimension	num_tests
fd27b272-7144-11e5-ba71-058fbc01cf0b	charmer	21
fd27b5ba-7144-11e5-ba71-058fbc01cf0b	protodog	20
fd27b6b4-7144-11e5-ba71-058fbc01cf0b	None	2
fd27b79a-7144-11e5-ba71-058fbc01cf0b	None	11
fd27b86c-7144-11e5-ba71-058fbc01cf0b	einstein	31
fd27b948-7144-11e5-ba71-058fbc01cf0b	stargazer	20
fd27ba1a-7144-11e5-ba71-058fbc01cf0b	maverick	27
fd27bbe-7144-11e5-ba71-058fbc01cf0b	protodog	20
fd27c1c2-7144-11e5-ba71-058fbc01cf0b	einstein	20
fd27c5be-7144-11e5-ba71-058fbc01cf0b	socialite	20
fd27c74e-7144-11e5-ba71-058fbc01cf0b	None	14
fd27c7d0-7144-11e5-ba71-058fbc01cf0b	socialite	20
fd27c852-7144-11e5-ba71-058fbc01cf0b	stargazer	20
fd27c8d4-7144-11e5-ba71-058fbc01cf0b	ace	20
fd27c956-7144-11e5-ba71-058fbc01cf0b	None	11
fd27cb72-7144-11e5-ba71-058fbc01cf0b	protodog	20
fd27cd98-7144-11e5-ba71-058fbc01cf0b	expert	20
fd27ce1a-7144-11e5-ba71-058fbc01cf0b	None	7
fd27cea6-7144-11e5-ba71-058fbc01cf0b	None	2
fd27cf28-7144-11e5-ba71-058fbc01cf0b	charmer	20

Question 3: Re-write the query in Question 2 using traditional join syntax (described in MySQL Exercise 8).

In [4]:

```
%%sql
SELECT d.dog_guid, d.dimension, COUNT(c.created_at) AS num_tests
FROM dogs d
JOIN complete_tests c
ON d.dog_guid = c.dog_guid
GROUP BY d.dog_guid
LIMIT 20
```

20 rows affected.

Out[4]:

dog_guid	dimension	num_tests
fd27b272-7144-11e5-ba71-058fbc01cf0b	charmer	21
fd27b5ba-7144-11e5-ba71-058fbc01cf0b	protodog	20
fd27b6b4-7144-11e5-ba71-058fbc01cf0b	None	2
fd27b79a-7144-11e5-ba71-058fbc01cf0b	None	11
fd27b86c-7144-11e5-ba71-058fbc01cf0b	einstein	31
fd27b948-7144-11e5-ba71-058fbc01cf0b	stargazer	20
fd27ba1a-7144-11e5-ba71-058fbc01cf0b	maverick	27
fd27bbbe-7144-11e5-ba71-058fbc01cf0b	protodog	20
fd27c1c2-7144-11e5-ba71-058fbc01cf0b	einstein	20
fd27c5be-7144-11e5-ba71-058fbc01cf0b	socialite	20
fd27c74e-7144-11e5-ba71-058fbc01cf0b	None	14
fd27c7d0-7144-11e5-ba71-058fbc01cf0b	socialite	20
fd27c852-7144-11e5-ba71-058fbc01cf0b	stargazer	20
fd27c8d4-7144-11e5-ba71-058fbc01cf0b	ace	20
fd27c956-7144-11e5-ba71-058fbc01cf0b	None	11
fd27cb72-7144-11e5-ba71-058fbc01cf0b	protodog	20
fd27cd98-7144-11e5-ba71-058fbc01cf0b	expert	20
fd27ce1a-7144-11e5-ba71-058fbc01cf0b	None	7
fd27cea6-7144-11e5-ba71-058fbc01cf0b	None	2
fd27cf28-7144-11e5-ba71-058fbc01cf0b	charmer	20

Now we need to summarize the total number of tests completed by each unique DogID within each Dognition personality dimension. To do this we will need to choose an appropriate aggregation function for the count column of the query we just wrote.

Question 4: To start, write a query that will output the average number of tests completed by unique dogs in each Dognition personality dimension. Choose either the query in Question 2 or 3 to serve as an inner query in your main query. If you have trouble, make sure you use the appropriate aliases in your GROUP BY and SELECT statements.

In [26]:

```
%%sql
SELECT dimension, AVG(num_tests) AS avgnum_tests
FROM (SELECT d.dog_guid, d.dimension, COUNT(c.created_at) AS num_tests
      FROM dogs d
      JOIN complete_tests c
        ON d.dog_guid = c.dog_guid
     GROUP BY d.dog_guid) AS dog_numtests
GROUP BY dimension
```

11 rows affected.

Out[26]:

dimension	avgnum_tests
None	6.9416
	9.5352
ace	23.3878
charmer	23.2594
einstein	23.2171
expert	23.3926
maverick	22.8199
protodog	22.9336
renaissance-dog	23.0157
socialite	23.1194
stargazer	22.7368

You should retrieve an output of 11 rows with one of the dimensions labeled "None" and another labeled "" (nothing is between the quotation marks).

Question 5: How many unique DogIDs are summarized in the Dognition dimensions labeled "None" or ""? (You should retrieve values of 13,705 and 71)

In [17]:

```
%%sql
SELECT dimension, COUNT(dog_guid) AS num_dogs
FROM (SELECT DISTINCT d.dog_guid, d.dimension
      FROM dogs d
      JOIN complete_tests c
        ON d.dog_guid = c.dog_guid) AS dog_completetests
WHERE dimension IS NULL OR dimension = ''
GROUP BY dog_completetests.dimension
```

2 rows affected.

Out[17]:

dimension	num_dogs
None	13705
	71

It makes sense there would be many dogs with NULL values in the dimension column, because we learned from Dognition that personality dimensions can only be assigned after the initial "Dognition Assessment" is completed, which is comprised of the first 20 Dognition tests. If dogs did not complete the first 20 tests, they would retain a NULL value in the dimension column.

The non-NULL empty string values are more curious. It is not clear where those values would come from.

Question 6: To determine whether there are any features that are common to all dogs that have non-NULL empty strings in the dimension column, write a query that outputs the breed, weight, value in the "exclude" column, first or minimum time stamp in the complete_tests table, last or maximum time stamp in the complete_tests table, and total number of tests completed by each unique DogID that has a non-NULL empty string in the dimension column.

In [14]:

```
%%sql
SELECT d.breed, d.weight, d.exclude, MIN(c.created_at) AS first_time,
       MAX(c.created_at) AS last_time, COUNT(c.created_at) AS num_tests
FROM dogs d
JOIN complete_tests c
ON d.dog_guid = c.dog_guid
WHERE d.dimension = ''
GROUP BY d.dog_guid
LIMIT 20
```

20 rows affected.

Out[14]:

breed	weight	exclude	first_time	last_time	num_tests
Golden Retriever	30	0	2013-05-23 07:06:21	2013-07-02 12:15:18	17
Dachshund	10	1	2014-10-21 18:53:02	2014-10-21 19:10:07	3
Border Collie-Labrador Retriever Mix	50	0	2013-11-16 02:26:15	2013-11-16 02:38:57	4
Belgian Tervuren	70	1	2014-11-10 21:21:06	2014-12-16 01:13:28	13
Pembroke Welsh Corgi	20	1	2014-09-19 17:42:37	2014-09-22 17:58:25	4
Chihuahua	0	1	2014-10-06 00:57:46	2014-10-09 22:55:51	2
Australian Shepherd	50	1	2014-10-06 01:54:49	2014-10-30 02:16:12	14
Mixed	60	1	2014-10-10 01:01:21	2014-10-10 12:33:52	4
Portuguese Water Dog	60	1	2014-10-10 13:22:58	2014-10-10 13:36:17	3
Labrador Retriever	50	1	2014-10-06 15:28:42	2014-10-23 20:24:20	7
Golden Retriever	70	1	2014-10-06 17:27:41	2014-12-10 01:35:43	14
Poodle	50	1	2014-10-09 05:29:07	2014-11-03 03:10:19	16
Other	30	1	2014-10-21 02:05:28	2014-10-21 02:12:10	2
Labrador Retriever	70	1	2014-10-07 02:04:42	2014-10-23 02:06:02	10
West Highland White Terrier	10	1	2014-10-07 13:07:18	2014-10-11 21:57:57	19

Brittany	30	1	2014-10-12 18:09:00	2015-01-04 20:38:26	11
German Shepherd Dog	50	1	2014-10-07 16:58:19	2014-10-08 15:56:42	4
Miniature Schnauzer	10	1	2014-10-16 20:42:50	2014-10-21 17:29:14	6
Shih Tzu-Bichon Frise Mix	10	1	2014-10-08 22:48:34	2015-03-08 18:33:37	17
Mixed	30	1	2014-10-18 14:00:16	2014-10-25 14:18:48	16

A quick inspection of the output from the last query illustrates that almost all of the entries that have non-NULL empty strings in the dimension column also have "exclude" flags of 1, meaning that the entries are meant to be excluded due to factors monitored by the Dognition team. This provides a good argument for excluding the entire category of entries that have non-NULL empty strings in the dimension column from our analyses.

Question 7: Rewrite the query in Question 4 to exclude DogIDs with (1) non-NULL empty strings in the dimension column, (2) NULL values in the dimension column, and (3) values of "1" in the exclude column. NOTES AND HINTS: You cannot use a clause that says d.exclude does not equal 1 to remove rows that have exclude flags, because Dognition clarified that both NULL values and 0 values in the "exclude" column are valid data. A clause that says you should only include values that are not equal to 1 would remove the rows that have NULL values in the exclude column, because NULL values are never included in equals statements (as we learned in the join lessons). In addition, although it should not matter for this query, practice including parentheses with your OR and AND statements that accurately reflect the logic you intend. Your results should return 402 DogIDs in the ace dimension and 626 dogs in the charmer dimension.

In [31]: %%sql

```

SELECT dimension, COUNT(DISTINCT dog_guid) AS num_dogs, AVG(num_tests) AS avgnum_
FROM (SELECT d.dog_guid, d.dimension, COUNT(c.created_at) AS num_tests
      FROM dogs d
      JOIN complete_tests c
        ON d.dog_guid = c.dog_guid
     WHERE (d.dimension IS NOT NULL AND d.dimension != '') AND (d.exclude IS NUL
      GROUP BY d.dog_guid) AS dog_numtests
   GROUP BY dimension
  
```

9 rows affected.

Out[31]:

dimension	num_dogs	avgnum_tests
ace	402	23.5100
charmer	626	23.3594
einstein	109	23.2385
expert	273	23.4249
maverick	245	22.7673
protodog	535	22.9570
renaissance-dog	463	23.0410
socialite	792	23.0997
stargazer	310	22.7968

The results of Question 7 suggest there are not appreciable differences in the number of tests completed by dogs with different Dognition personality dimensions. Although these analyses are not definitive on their own, these results suggest focusing on Dognition personality dimensions will not likely lead to significant insights about how to improve Dognition completion rates.

2. Assess whether dog breeds are related to the number of tests completed

The next variable in the Dognition sPAP we want to investigate is Dog Breed. We will run one analysis with Breed Group and one analysis with Breed Type.

First, determine how many distinct breed groups there are.

Questions 8: Write a query that will output all of the distinct values in the breed_group field.

In [15]:

```
%%sql
SELECT DISTINCT breed_group
FROM dogs
```

9 rows affected.

Out[15]:

breed_group
Sporting
Herding
Toy
Working
None
Hound
Non-Sporting
Terrier

You can see that there are NULL values in the breed_group field. Let's examine the properties of these entries with NULL values to determine whether they should be excluded from our analysis.

Question 9: Write a query that outputs the breed, weight, value in the "exclude" column, first or minimum time stamp in the complete_tests table, last or maximum time stamp in the complete_tests table, and total number of tests completed by each unique DogID that has a NULL value in the breed_group column.

In [16]:

```
%%sql
SELECT d.breed, d.weight, d.exclude, MIN(c.created_at) AS first_time,
       MAX(c.created_at) AS last_time, COUNT(c.created_at) AS num_tests
FROM dogs d
JOIN complete_tests c
ON d.dog_guid = c.dog_guid
WHERE breed_group IS NULL
GROUP BY d.dog_guid
LIMIT 20
```

20 rows affected.

Out[16]:

breed	weight	exclude	first_time	last_time	num_tests
Mixed	50	None	2013-02-05 18:57:05	2013-02-05 22:38:01	20
Shih Tzu-Poodle Mix	0	None	2013-02-05 21:44:38	2013-02-10 03:33:37	20
German Shepherd Dog-Pembroke Welsh Corgi Mix	40	None	2013-02-06 04:45:28	2014-01-06 05:58:13	14
German Shepherd Dog-Nova Scotia Duck Tolling Retriever Mix	30	None	2013-05-17 17:45:46	2013-06-14 23:42:53	11
Mixed	10	None	2013-02-06 04:44:50	2013-02-06 04:48:29	2
Australian Shepherd-German Shepherd Dog Mix	90	None	2013-02-07 05:15:48	2013-12-20 21:03:18	21
Golden Doodle	70	None	2013-02-09 05:49:46	2013-02-09 06:10:11	6
Mixed	30	None	2013-02-10 03:28:12	2013-07-20 02:12:37	28
Mixed	90	1	2014-09-24 15:10:03	2014-09-24 21:23:37	20
Mudi	20	None	2014-10-06 22:21:56	2014-10-06 22:24:02	2
Parson Russell Terrier-Beagle Mix	30	None	2013-02-06 18:07:18	2013-02-06 18:16:13	4
I Don't Know	50	None	2013-02-06 22:14:00	2013-02-06 22:41:28	6
Mixed	70	None	2013-02-10 04:06:03	2015-09-28 17:33:05	45
Chihuahua- Mix	0	None	2013-02-08 04:04:51	2013-02-11 03:35:44	6
Mixed	20	None	2013-02-07 03:00:05	2013-02-07 03:16:21	4

Chihuahua-Dachshund Mix	10	None	2014-06-23 20:46:28	2014-06-23 20:50:48	3
Mixed	70	None	2013-02-19 15:46:44	2013-02-19 15:51:52	2
English Cocker Spaniel-Cocker Spaniel Mix	20	None	2013-02-16 16:29:48	2013-02-28 19:30:51	20
Cavalier King Charles Spaniel-Bichon Frise Mix	10	None	2013-02-08 21:48:11	2013-02-08 22:21:21	7
Poodle-Cocker Spaniel Mix	20	None	2013-02-07 23:03:17	2013-02-08 18:05:43	10

There are a lot of these entries and there is no obvious feature that is common to all of them, so at present, we do not have a good reason to exclude them from our analysis. Therefore, let's move on to question 10 now....

Question 10: Adapt the query in Question 7 to examine the relationship between breed_group and number of tests completed. Exclude DogIDs with values of "1" in the exclude column. Your results should return 1774 DogIDs in the Herding breed group.

In [32]:

```
%%sql
SELECT breed_group, COUNT(DISTINCT dog_guid) AS num_dogs, AVG(num_tests) AS avgnumtests
FROM (SELECT d.dog_guid, d.breed_group, COUNT(c.created_at) AS num_tests
      FROM dogs d
      JOIN complete_tests c
      ON d.dog_guid = c.dog_guid
      WHERE d.exclude IS NULL OR d.exclude = 0
      GROUP BY d.dog_guid) AS dog_numtests
GROUP BY breed_group
```

9 rows affected.

Out[32]:

breed_group	num_dogs	avgnum_tests
None	8564	10.2251
	179	19.7542
Herding	1774	11.2469
Hound	564	10.0603
Non-Sporting	964	10.0197
Sporting	2470	10.9915
Terrier	780	9.9333
Toy	1041	8.7157
Working	865	10.2358

The results show there are non-NULL entries of empty strings in breed_group column again.

Ignoring them for now, Herding and Sporting breed_groups complete the most tests, while Toy breed groups complete the least tests. This suggests that one avenue an analyst might want to explore further is whether it is worth it to target marketing or certain types of Dogognition tests to dog owners with dogs in the Herding and Sporting breed_groups. Later in this lesson we will discuss whether using a median instead of an average to summarize the number of completed tests might affect this potential course of action.

Question 11: Adapt the query in Question 10 to only report results for Sporting, Hound, Herding, and Working breed_groups using an IN clause.

In [33]:

```
%%sql
SELECT breed_group, COUNT(DISTINCT dog_guid) AS num_dogs, AVG(num_tests) AS avgnumtests
FROM (SELECT d.dog_guid, d.breed_group, COUNT(c.created_at) AS num_tests
      FROM dogs d
      JOIN complete_tests c
      ON d.dog_guid = c.dog_guid
      WHERE d.exclude IS NULL OR d.exclude = 0
      GROUP BY d.dog_guid) AS dog_numtests
GROUP BY breed_group
HAVING breed_group IN ('Sporting', 'Hound', 'Herding', 'Working')
```

4 rows affected.

Out[33]:

breed_group	num_dogs	avgnum_tests
Herding	1774	11.2469
Hound	564	10.0603
Sporting	2470	10.9915
Working	865	10.2358

Next, let's examine the relationship between breed_type and number of completed tests.

Question 12: Begin by writing a query that will output all of the distinct values in the breed_type field.

In [34]:

```
%%sql
SELECT DISTINCT breed_type
FROM dogs
```

4 rows affected.

Out[34]:

breed_type
Pure Breed
Mixed Breed/ Other/ I Don't Know
Cross Breed
Popular Hybrid

Question 13: Adapt the query in Question 7 to examine the relationship between breed_type

and number of tests completed. Exclude DogIDs with values of "1" in the exclude column. Your results should return 8865 DogIDs in the Pure Breed group.

In [35]:

```
%%sql
SELECT breed_type, COUNT(DISTINCT dog_guid) AS num_dogs, AVG(num_tests) AS avgnum_
FROM (SELECT d.dog_guid, d.breed_type, COUNT(c.created_at) AS num_tests
      FROM dogs d
      JOIN complete_tests c
        ON d.dog_guid = c.dog_guid
     WHERE d.exclude IS NULL OR d.exclude = 0
      GROUP BY d.dog_guid) AS dog_numtests
   GROUP BY breed_type
```

4 rows affected.

Out[35]:

breed_type	num_dogs	avgnum_tests
Cross Breed	2884	10.6009
Mixed Breed/ Other/ I Don't Know	4818	10.2688
Popular Hybrid	634	10.8423
Pure Breed	8865	10.4107

There does not appear to be an appreciable difference between number of tests completed by dogs of different breed types.

3. Assess whether dog breeds and neutering are related to the number of tests completed

To explore the results we found above a little further, let's run some queries that relabel the breed_types according to "Pure_Breed" and "Not_Pure_Breed".

Question 14: For each unique DogID, output its dog_guid, breed_type, number of completed tests, and use a CASE statement to include an extra column with a string that reads "Pure_Breed" whenever breed_type equals 'Pure Breed' and "Not_Pure_Breed" whenever breed_type equals anything else. LIMIT your output to 50 rows for troubleshooting.

In [38]: %%sql

```

SELECT d.dog_guid, d.breed_type AS breed_type, count(c.created_at) AS num_tests,
CASE d.breed_type
    WHEN 'Pure Breed' THEN 'pure_breed'
    ELSE 'not_pure_breed'
END AS pure_breed
FROM dogs d
JOIN complete_tests c
ON d.dog_guid=c.dog_guid
GROUP BY d.dog_guid
LIMIT 10

```

10 rows affected.

Out[38]:

dog_guid	breed_type	num_tests	pure_breed
fd27b272-7144-11e5-ba71-058fbc01cf0b	Pure Breed	21	pure_breed
fd27b5ba-7144-11e5-ba71-058fbc01cf0b	Pure Breed	20	pure_breed
fd27b6b4-7144-11e5-ba71-058fbc01cf0b	Pure Breed	2	pure_breed
fd27b79a-7144-11e5-ba71-058fbc01cf0b	Pure Breed	11	pure_breed
fd27b86c-7144-11e5-ba71-058fbc01cf0b	Pure Breed	31	pure_breed
fd27b948-7144-11e5-ba71-058fbc01cf0b	Pure Breed	20	pure_breed
fd27ba1a-7144-11e5-ba71-058fbc01cf0b	Pure Breed	27	pure_breed
fd27bbbe-7144-11e5-ba71-058fbc01cf0b	Mixed Breed/ Other/ I Don't Know	20	not_pure_breed
fd27c1c2-7144-11e5-ba71-058fbc01cf0b	Pure Breed	20	pure_breed
fd27c5be-7144-11e5-ba71-058fbc01cf0b	Cross Breed	20	not_pure_breed

Question 15: Adapt your queries from Questions 7 and 14 to examine the relationship between breed_type and number of tests completed by Pure_Breed dogs and non_Pure_Breed dogs. Your results should return 8336 DogIDs in the Not_Pure_Breed group.

In [43]: %%sql

```

SELECT pure_breed, COUNT(DISTINCT dog_guid) AS num_dogs, AVG(num_tests) AS avgnum
FROM (SELECT d.dog_guid, count(c.created_at) AS num_tests,
CASE d.breed_type
    WHEN 'Pure Breed' THEN 'pure_breed'
    ELSE 'not_pure_breed'
END AS pure_breed
FROM dogs d
JOIN complete_tests c
ON d.dog_guid=c.dog_guid
WHERE d.exclude IS NULL OR d.exclude = 0
GROUP BY d.dog_guid) AS dog_numtests
GROUP BY pure_breed

```

2 rows affected.

Out[43]:

pure_breed	num_dogs	avgnum_tests
not_pure_breed	8336	10.4273
pure_breed	8865	10.4107

Question 16: Adapt your query from Question 15 to examine the relationship between breed_type, whether or not a dog was neutered (indicated in the dog_fixed field), and number of tests completed by Pure_Breed dogs and non_Pure_Breed dogs. There are DogIDs with null values in the dog_fixed column, so your results should have 6 rows, and the average number of tests completed by non-pure-breeds who are neutered is 10.5681.

In [46]: %%sql

```

SELECT pure_breed, neutered, COUNT(DISTINCT dog_guid) AS num_dogs, AVG(num_tests)
FROM (SELECT d.dog_guid, count(c.created_at) AS num_tests, d.dog_fixed AS neutered
      CASE d.breed_type
        WHEN 'Pure Breed' THEN 'pure_breed'
        ELSE 'not_pure_breed'
      END AS pure_breed
     FROM dogs d
    JOIN complete_tests c
      ON d.dog_guid=c.dog_guid
     WHERE d.exclude IS NULL OR d.exclude = 0
   GROUP BY d.dog_guid) AS dog_numtests
  GROUP BY pure_breed, neutered
  
```

6 rows affected.

Out[46]:

pure_breed	neutered	num_dogs	avgnum_tests
not_pure_breed	None	97	9.9897
not_pure_breed	0	592	8.6807
not_pure_breed	1	7647	10.5681
pure_breed	None	135	8.2815
pure_breed	0	1687	9.3788
pure_breed	1	7043	10.6987

These results suggest that although a dog's breed_type doesn't seem to have a strong relationship with how many tests a dog completed, neutered dogs, on average, seem to finish 1-2 more tests than non-neutered dogs. It may be fruitful to explore further whether this effect is consistent across different segments of dogs broken up according to other variables. If the effects are consistent, the next step would be to seek evidence that could clarify whether neutered dogs are finishing more tests due to traits that arise when a dog is neutered, or instead, whether owners who are more likely to neuter their dogs have traits that make it more likely they will want to complete more tests.

4. Other dog features that might be related to the number of tests completed, and a note about using averages as summary metrics

Two other dog features included in our sPAP were speed of game completion and previous behavioral training. Examining the relationship between the speed of game completion and number of games completed is best achieved through creating a scatter plot with a best fit line and/or running a statistical regression analysis. It is possible to achieve the statistical regression analysis through very advanced SQL queries, but the strategy that would be required is outside the scope of this course. Therefore, I would recommend exporting relevant data to a program like Tableau, R, or Matlab in order to assess the relationship between the speed of game completion and number of games completed.

Unfortunately, there is no field available in the Dognition data that is relevant to a dog's previous behavioral training, so more data would need to be collected to examine whether previous behavioral training is related to the number of Dognition tests completed.

One last issue I would like to address in this lesson is the issue of whether an average is a good summary to use to represent the values of a certain group. Average calculations are very sensitive to extreme values, or outliers, in the data. This video provides a nice demonstration of how sensitive averages can be:

<http://www.statisticstutorials.com/topics/outliereffects/>
[\(http://www.statisticstutorials.com/topics/outliereffects/\)](http://www.statisticstutorials.com/topics/outliereffects/)

Ideally, you would summarize the data in a group using a median calculation when you either don't know the distribution of values in your data or you already know that outliers are present (the definition of median is covered in the video above). Unfortunately, medians are more computationally intensive than averages, and there is no pre-made function that allows you to calculate medians using SQL. If you wanted to calculate the median, you would need to use an advanced strategy such as the ones described here:

<https://www.periscopedata.com/blog/medians-in-sql.html>
[\(https://www.periscopedata.com/blog/medians-in-sql.html\)](https://www.periscopedata.com/blog/medians-in-sql.html)

Despite the fact there is no simple way to calculate medians using SQL, there is a way to get a hint about whether average values are likely to be wildly misleading. As described in the first video (<http://www.statisticstutorials.com/topics/outliereffects/>
[\(http://www.statisticstutorials.com/topics/outliereffects/\)](http://www.statisticstutorials.com/topics/outliereffects/)), strong outliers lead to large standard deviation values. Fortunately, we CAN calculate standard deviations in SQL easily using the STDEV function. Therefore, it is good practice to include standard deviation columns with your outputs so that you have an idea whether the average values outputted by your queries are trustworthy. Whenever standard deviations are a significant portion of the average values of a field, and certainly when standard deviations are larger than the average values of a field, it's a good idea to export your data to a program that can handle more sophisticated statistical analyses before you interpret any results too strongly.

Let's practice including standard deviations in our queries and interpreting their values.

Question 17: Adapt your query from Question 7 to include a column with the standard deviation for the number of tests completed by each Dognition personality dimension.

In [47]: %%sql

```

SELECT dimension, COUNT(DISTINCT dog_guid) AS num_dogs, AVG(num_tests) AS avgnum_
FROM (SELECT d.dog_guid, d.dimension, COUNT(c.created_at) AS num_tests
      FROM dogs d
      JOIN complete_tests c
        ON d.dog_guid = c.dog_guid
     WHERE (d.dimension IS NOT NULL AND d.dimension != '') AND (d.exclude IS NUL
      GROUP BY d.dog_guid) AS dog_numtests
   GROUP BY dimension
  
```

9 rows affected.

Out[47]:

dimension	num_dogs	avgnum_tests	stdnum_tests
ace	402	23.5100	5.4896
charmer	626	23.3594	5.1919
einstein	109	23.2385	5.3155
expert	273	23.4249	4.7589
maverick	245	22.7673	4.7353
protodog	535	22.9570	5.3742
renaissance-dog	463	23.0410	4.9508
socialite	792	23.0997	4.9748
stargazer	310	22.7968	4.8254

The standard deviations are all around 20-25% of the average values of each personality dimension, and they are not appreciably different across the personality dimensions, so the average values are likely fairly trustworthy. Let's try calculating the standard deviation of a different measurement.

Question 18: Write a query that calculates the average amount of time it took each dog breed_type to complete all of the tests in the exam_answers table. Exclude negative durations from the calculation, and include a column that calculates the standard deviation of durations for each breed_type group:

In [48]:

```
%%sql
SELECT d.breed_type, AVG(TIMESTAMPDIFF(minute,e.start_time,e.end_time)) AS avg_duration
      STDDEV(TIMESTAMPDIFF(minute,e.start_time,e.end_time)) AS std_duration
  FROM dogs d
  JOIN exam_answers e
    ON d.dog_guid=e.dog_guid
 WHERE TIMESTAMPDIFF(minute,e.start_time,e.end_time) > 0
 GROUP BY breed_type
```

4 rows affected.

Out[48]:

breed_type	avg_duration	std_duration
Cross Breed	11810.3230	59113.4558
Mixed Breed/ Other/ I Don't Know	9145.1575	48748.6268
Popular Hybrid	7734.0763	45577.6582
Pure Breed	12311.2558	60997.3543

This time many of the standard deviations have larger magnitudes than the average duration values. This suggests there are outliers in the data that are significantly impacting the reported average values, so the average values are not likely trustworthy. These data should be exported to another program for more sophisticated statistical analysis.

In the next lesson, we will write queries that assess the relationship between testing circumstances and the number of tests completed. Until then, feel free to practice any additional queries you would like to below!

In []: