Bowen Zhang

BA 479

1. What is the total number of documents retrieved by subjects during the experiment?

**Select count(\*) from docs;**

**356**

1. Which document types were involved in these retrievals? (Hint: your query should return four records, one of which is NULL)

**Select doc\_type from docs group by doc\_type;**

**doc\_type**

**NULL**

**activity**

**lesson**

**unit**

1. What does the NULL result from the previous query signify?

**This means there is no specific type of data or it is can be nullable.**

1. Using only the docs table, how many different subjects (usernames) did participate in the experiment? (remember; the query should return but a single(!) number)

**Select count(distinct(username)) from docs;**

**64**

1. Check this result against the subjects table and explain the difference. (Hint: the count from the previous question is different from the number of different usernames in the subjects table. How can that be?)

**Select count(username)from subjects; 63**

**The username in docs table is not in subjects table**

1. What is the 'extra' username in the docs table which is missing from the subjects table? (Hint: SQL query should return only a single record).

**Select docs.username from docs where docs.username not in(select subjects.username from subjects);**

**Jonhson**

1. Update the docs table to set the username of the 'extra' username in the docs table to Bang\_Bang\_Johnson and check the result of question 4 again.

**Update docs Set Username=’Bang\_Bang\_Johnson’ where username=’Johnson’;**

**Select count(distinct(username)) from docs;**

**63**

1. Which participant has the longest username and how many characters does that username have? (use a single query) (Hint: the longest username has 18 characters)

**Select top 1 username, len(username) as Name\_Length order by Name\_Length desc;**

**Dante\_Thunderstone 18**

1. Which two subjects retrieved the most documents and how many documents did each of those subjects retrieve? Note: your query should return only two rows. (Hint: both of these users each retrieved 11 documents)

**Select top 2 username, count(\*) as numberofdoc group by username, order by count(\*) desc;**

**username numberofdoc**

**carrot 11**

**MileyCyrus 11**

1. How many tasks were completed for each of the two experimental conditions? (do this with a single query!!).

**Select exp\_condition ,count(\*) as number from subjects inner join tasks on subjects.username=tasks.username group by exp\_condition;**

**exp\_condition number**

**1 58**

**2 81**

|  |  |  |
| --- | --- | --- |
| **Experimental condition** | **List (1)** | **Map (2)** |
| number of subjects  **Select exp\_condition,count(\*) from subjects group by exp\_condition;** | 28 | 35 |
| number of documents retrieved  **Select exp\_condition, count(\*) from subjects inner join docs on subjects.username=docs.username group by exp\_condition;** | 154 | 202 |
| smallest number of documents retrieved by any subject (you may use several queries for this)  **Select top 1 docs.username, exp\_condition, count(\*) as #documents from subjects inner join docs on subjects.username=docs.username where exp\_condition=1 group by docs.username, exp\_condition order by count(\*);** | 2 | 1 |
| largest number of documents retrieved by any subject (you may use several queries for this)  **Select top 1 docs.username, exp\_condition, count(\*) as #documents from subjects inner join docs on subjects.username=docs.username where exp\_condition=2 group by docs.username, exp\_condition order by count(\*) desc;** | 11 | 11 |
| average number of documents retrieved per subject (do not use SQL for this; just compute from the numbers you already have) | 5.5 | 5.7 |
| sigma (std. dev.) of the number of documents retrieved per subject: the SQL to compute this is as follows:  *select stdev(my\_table.my\_count) from (select count(\*) as my\_count from docs, subjects where subjects.username = docs.username and subjects.exp\_condition = 1 (...or 2) group by subjects.username) as my\_table;* | 2.54587538608658 | 2.52150415293676 |



This not a SQL question, but it is a natural fol This not a SQL question, but it is a natural follow up from what you computed in the previous question (plus it explains why we take stats classes!). From the results in the previous question we may conclude that subjects in the 'Map' condition, on average retrieve more documents than subjects in the 'List' condition. Before we draw that conclusion, however, we should ask ourselves if the apparent difference is likely the result of random effects; i.e., we must check for the statistical significance of the difference. The test that applies here is the t-test for testing equality of means in two samples.low up from what you computed in the previous question (plus it explains why we take stats classes!). From the results in the previous question we may conclude that subjects in the 'Map' condition, on average retrieve more documents than subjects in the 'List' condition. Before we draw that conclusion, however, we should ask ourselves if the apparent difference is likely the result of random effects; *i.e.*, we must check for the statistical significance of the difference. The test that applies here is the t-test for testing equality of means in two samples.

**P value and statistical significance:   
  The two-tailed P value equals 0.4608   
  By conventional criteria, this difference is considered to be not statistically significant.**  
**Confidence interval:   
  The mean of Group One minus Group Two equals -0.200000   
  95% confidence interval of this difference: From -0.732722 to 0.332722   
  
Intermediate values used in calculations:   
  t = 0.7384   
  df = 354   
  standard error of difference = 0.271**

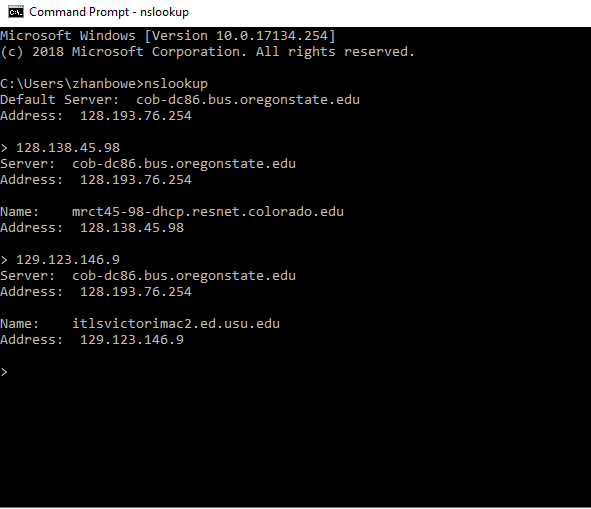
**It is difference and not statistical significance in means in 0.05 critical level, because P-value 0.4608>0.05.**

1. If we consider the two left-most bytes of an IP address to indicate the organization hosting the address, how many different *host\_ip*s in our experiment a If we consider the two left-most bytes of an IP address to indicate the organization hosting the address, how many different host\_ips in our experiment are associated with the 128.138.\* network? (Hint: the count should be 33)

**Select count(distinct(host\_ip)) from exp\_data where host\_ip like ‘128.138%’;**

**33**

1. Not a SQL question: Which organization is associated with the 128.138 addresses? How about the 129.123 ones? Hint: find a few host\_ips in the exp\_data table which are associated with these organizations and then do a reverse DNS lookup (on Windows: nslookup command).



**128.138: University of Colorado Boulder**

**129.123 :Utah State University**

1. List, in order from earliest to latest, the different dates (your SQL should return dates only, not times of day!) during which experimental data were collected and the number of exp\_data records collected on each of those dates. Again, use a single query! (Hint: your list of dates should have 13 dates)

**Select convert(date, timestamp) as Data, count(username) as number from exp\_data group by convert(date, timestamp)**

**order by convert(date, timestamp) ;**

Data number

2014-09-11 978

2014-11-07 63

2014-11-08 42

2014-11-24 242

2014-11-25 5

2015-04-28 79

2015-04-30 68

2015-05-01 37

2015-05-28 353

2015-06-08 1669

2015-06-14 437

2015-07-09 326

2015-08-17 905

1. What is the daily minimum, maximum and average number of exp\_data records collected? Again, do not read or compute these numbers manually from the results in the previous question. Use SQL to do it. Hint: perhaps the easiest way to do this is in two steps: first make a new table from the results of your previous query; one which contains the totals for each day, and then query that new table for the minimum, maximum and average. You can make a new table from a query using the select ... into new-table-name from ... syntax. (Hint: if the result for your average is 400, you should realize that that is not quite correct!)

**Average: 400.307692**

**Select avg(convert(decimal, My\_table.number)) from**

**(Select convert(date, timestamp) as Data, count(username) as number from exp\_data group by convert(date,timestamp))**

**as My\_table;**

**Min: 5**

**Select Min(convert(decimal, My\_table.number)) from**

**(Select convert(date, timestamp) as Data, count(username) as number from exp\_data group by convert(date,timestamp))**

**as My\_table;**

**Max: 1669**

**Select MAX(convert(decimal, My\_table.number)) from**

**(Select convert(date, timestamp) as Data, count(username) as number from exp\_data group by convert(date,timestamp))**

**as My\_table;**

1. How many days have passed between the first record being collected and the last? Note: you must write a SQL query which computes this number. Do not compute it yourself from the min and max timestamps you computed in the previous question.

**340**

**Select datediff(day,min(exp\_data.timestamp), max(exp\_data.timestamp)) from exp\_data ;**