**Data Scientist Role Play: Profiling and Analyzing the Yelp Dataset**

**Coursera Worksheet**

Mehrshad Esfahani

**Part 1: Yelp Dataset Profiling and Understanding**

1. Profile the data by finding the total number of records for each of the tables below:

i. Attribute table = 10000

ii. Business table = 10000

iii. Category table = 10000

iv. Checkin table = 10000

v. elite\_years table = 10000

vi. friend table = 10000

vii. hours table = 10000

viii. photo table = 10000

ix. review table = 10000

x. tip table = 10000

xi. user table = 10000

Sample code (including NULL values):

select count(\*) as

total\_records

from attribute;

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

| total\_records |

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

| 10000 |

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

2. Find the total number of distinct records for each of the keys listed below:

i. Business = 44092

ii. Hours = 3614

iii. Category = 3355

iv. Attribute = 1285

v. Review = 40568

vi. Checkin = 697

vii. Photo = 21194

viii. Tip = 16751

ix. User = 18944

x. Friend = 9426

xi. Elite\_years = 2793

Sample code:

select count(distinct name) + count(distinct business\_id)

+ count(distinct value)

as

total\_records

from attribute;

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

| total\_records |

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

| 1285 |

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

3. Are there any columns with null values in the Users table? Indicate "yes," or "no."

Answer: Zero rows in the answer shows that there is no null values in the User table

SQL code used to arrive at answer:

select id

, name, review\_count, yelping\_since, useful, funny, cool, fans, average\_stars

, compliment\_hot, compliment\_more, compliment\_profile, compliment\_cute, compliment\_list, compliment\_note, compliment\_plain, compliment\_cool, compliment\_funny, compliment\_writer, compliment\_photos

from user

where id = NULL or name = NULL or review\_count = NULL or yelping\_since = NULL or useful = NULL or funny = NULL or cool = NULL or fans= NULL or average\_stars= NULL or compliment\_hot= NULL or compliment\_more= NULL or compliment\_profile= NULL or compliment\_cute= NULL or compliment\_list= NULL or compliment\_note= NULL or compliment\_plain = NULL or compliment\_cool= NULL or compliment\_funny= NULL or compliment\_writer= NULL or compliment\_photos= NULL;

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

| id | name | review\_count | yelping\_since | useful | funny | cool | fans | average\_stars | compliment\_hot | compliment\_more | compliment\_profile | compliment\_cute | compliment\_list | compliment\_note | compliment\_plain | compliment\_cool | compliment\_funny | compliment\_writer | compliment\_photos |

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

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

(Zero rows)

4. Find the minimum, maximum, and average value for the following fields:

i. Table: Review, Column: Stars

min: 1 max: 5 avg: 3.7082

ii. Table: Business, Column: Stars

min: 1 max: 5 avg: 3.6549

iii. Table: Tip, Column: Likes

min: 0 max: 2 avg: 0.0144

iv. Table: Checkin, Column: Count

min: 1 max: 53 avg: 1.9414

v. Table: User, Column: Review\_count

min: 0 max: 2000 avg: 24.2995

Sample code:

select min(stars)

,max(stars)

,avg(stars)

from review;

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

| min(stars) | max(stars) | avg(stars) |

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

| 1 | 5 | 3.7082 |

5. List the cities with the most reviews in descending order:

SQL code used to arrive at answer:

select

city

, count(review\_count) as total\_review

from business

group by city

order by total\_review desc;

Copy and Paste the Result Below:

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

| city | total\_review |

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

| Las Vegas | 1561 |

| Phoenix | 1001 |

| Toronto | 985 |

| Scottsdale | 497 |

| Charlotte | 468 |

| Pittsburgh | 353 |

| Montréal | 337 |

| Mesa | 304 |

| Henderson | 274 |

| Tempe | 261 |

| Edinburgh | 239 |

| Chandler | 232 |

| Cleveland | 189 |

| Gilbert | 188 |

| Glendale | 188 |

| Madison | 176 |

| Mississauga | 150 |

| Stuttgart | 141 |

| Peoria | 105 |

| Markham | 80 |

| Champaign | 71 |

| North Las Vegas | 70 |

| North York | 64 |

| Surprise | 60 |

| Richmond Hill | 54 |

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

(Output limit exceeded, 25 of 362 total rows shown)

6. Find the distribution of star ratings to the business in the following cities:

i. Avon

SQL code used to arrive at answer:

select

name

, stars

, review\_count

from business

where city = 'Avon';

Copy and Paste the Resulting Table Below (2 columns - star rating and count):

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

| name | stars | review\_count |

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

| Helen & Kal's | 2.5 | 3 |

| Marc's | 4.0 | 4 |

| Hoban Pest Control | 5.0 | 3 |

| Light Salon & Spa | 3.5 | 7 |

| Portrait Innovations | 1.5 | 10 |

| Winking Lizard Tavern | 3.5 | 31 |

| Dervish Mediterranean & Turkish Grill | 4.5 | 31 |

| Mulligans Pub and Grill | 3.5 | 50 |

| Mr. Handyman of Cleveland's Northwest Suburbs | 2.5 | 3 |

| Cambria hotel & suites Avon - Cleveland | 4.0 | 17 |

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

ii. Beachwood

SQL code used to arrive at answer:

select

name

, stars

, review\_count

from business

where city = 'Beachwood';

Copy and Paste the Resulting Table Below (2 columns - star rating and count):

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

| name | stars | review\_count |

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

| Maltz Museum of Jewish Heritage | 3.0 | 8 |

| Charley's Grilled Subs | 3.0 | 3 |

| Sixth & Pine | 4.5 | 14 |

| Beechmont Country Club | 5.0 | 6 |

| Hyde Park Prime Steakhouse | 4.0 | 69 |

| Origins | 4.5 | 3 |

| Fyodor Bridal Atelier | 5.0 | 4 |

| College Planning Network | 2.0 | 8 |

| Lucky Brand Jeans | 3.5 | 3 |

| American Eagle Outfitters | 3.5 | 3 |

| Shaker Women's Wellness | 5.0 | 6 |

| Avis Rent A Car | 2.5 | 3 |

| Cleveland Acupuncture | 5.0 | 3 |

| Studio Mz | 5.0 | 4 |

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

7. Find the top 3 users based on their total number of reviews:

SQL code used to arrive at answer:

select

select

name

, id

, review\_count

from user

order by review\_count desc;

Copy and Paste the Result Below:

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

| name | id | review\_count |

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

| Gerald | -G7Zkl1wIWBBmD0KRy\_sCw | 2000 |

| Sara | -3s52C4zL\_DHRK0ULG6qtg | 1629 |

| Yuri | -8lbUNlXVSoXqaRRiHiSNg | 1339 |

8. Does posing more reviews correlate with more fans?

Please explain your findings and interpretation of the results:

As table below illustrates, posing more reviews does not necessarily correlate with more fans. For example, although, Gerald has posed the most reviews, he has fewer fans in comparison with Mimi. Therefore, sorting the users in descending order based on their total number of reviews does not sort the fans in the same order, meaning that there is not a correlation between the total number of reviews and number of fans.

select

name

, id

, review\_count

, fans

from user

order by review\_count desc;

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

| name | id | review\_count | fans |

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

| Gerald | -G7Zkl1wIWBBmD0KRy\_sCw | 2000 | 253 |

| Sara | -3s52C4zL\_DHRK0ULG6qtg | 1629 | 50 |

| Yuri | -8lbUNlXVSoXqaRRiHiSNg | 1339 | 76 |

| .Hon | -K2Tcgh2EKX6e6HqqIrBIQ | 1246 | 101 |

| William | -FZBTkAZEXoP7CYvRV2ZwQ | 1215 | 126 |

| Harald | --2vR0DIsmQ6WfcSzKWigw | 1153 | 311 |

| eric | -gokwePdbXjfS0iF7NsUGA | 1116 | 16 |

| Roanna | -DFCC64NXgqrxlO8aLU5rg | 1039 | 104 |

| Mimi | -8EnCioUmDygAbsYZmTeRQ | 968 | 497 |

| Christine | -0IiMAZI2SsQ7VmyzJjokQ | 930 | 173 |

| Ed | -fUARDNuXAfrOn4WLSZLgA | 904 | 38 |

| Nicole | -hKniZN2OdshWLHYuj21jQ | 864 | 43 |

| Fran | -9da1xk7zgnnfO1uTVYGkA | 862 | 124 |

| Mark | -B-QEUESGWHPE\_889WJaeg | 861 | 115 |

| Christina | -kLVfaJytOJY2-QdQoCcNQ | 842 | 85 |

| Dominic | -kO6984fXByyZm3\_6z2JYg | 836 | 37 |

| Lissa | -lh59ko3dxChBSZ9U7LfUw | 834 | 120 |

| Lisa | -g3XIcCb2b-BD0QBCcq2Sw | 813 | 159 |

| Alison | -l9giG8TSDBG1jnUBUXp5w | 775 | 61 |

| Sui | -dw8f7FLaUmWR7bfJ\_Yf0w | 754 | 78 |

| Tim | -AaBjWJYiQxXkCMDlXfPGw | 702 | 35 |

| L | -jt1ACMiZljnBFvS6RRvnA | 696 | 10 |

| Angela | -IgKkE8JvYNWeGu8ze4P8Q | 694 | 101 |

| Crissy | -hxUwfo3cMnLTv-CAaP69A | 676 | 25 |

| Lyn | -H6cTbVxeIRYR-atxdielQ | 675 | 45 |

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

(Output limit exceeded, 25 of 10000 total rows shown)

9. Are there more reviews with the word "love" or with the word "hate" in them?

Answer:

As the tables below show there are more reviews with the word “love” in them compared to the word “hate”.

SQL code used to arrive at answer:

select

count (\*)

from review

where text like '%love%';

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

| count (\*) |

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

| 1780 |

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

select

count (\*)

from review

where text like '%hate%';

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

| count (\*) |

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

| 232 |

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

10. Find the top 10 users with the most fans:

SQL code used to arrive at answer:

select

name

, id

, fans

from user

order by fans desc;

Copy and Paste the Result Below:

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

| name | id | fans |

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

| Amy | -9I98YbNQnLdAmcYfb324Q | 503 |

| Mimi | -8EnCioUmDygAbsYZmTeRQ | 497 |

| Harald | --2vR0DIsmQ6WfcSzKWigw | 311 |

| Gerald | -G7Zkl1wIWBBmD0KRy\_sCw | 253 |

| Christine | -0IiMAZI2SsQ7VmyzJjokQ | 173 |

| Lisa | -g3XIcCb2b-BD0QBCcq2Sw | 159 |

| Cat | -9bbDysuiWeo2VShFJJtcw | 133 |

| William | -FZBTkAZEXoP7CYvRV2ZwQ | 126 |

| Fran | -9da1xk7zgnnfO1uTVYGkA | 124 |

| Lissa | -lh59ko3dxChBSZ9U7LfUw | 120 |

11. Is there a strong correlation between having a high number of fans and being listed as "useful" or "funny?"

SQL code used to arrive at answer:

select

name

, id

, fans

, useful

, funny

from user

order by fans desc;

Copy and Paste the Result Below:

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

| name | id | fans | useful | funny |

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

| Amy | -9I98YbNQnLdAmcYfb324Q | 503 | 3226 | 2554 |

| Mimi | -8EnCioUmDygAbsYZmTeRQ | 497 | 257 | 138 |

| Harald | --2vR0DIsmQ6WfcSzKWigw | 311 | 122921 | 122419 |

| Gerald | -G7Zkl1wIWBBmD0KRy\_sCw | 253 | 17524 | 2324 |

| Christine | -0IiMAZI2SsQ7VmyzJjokQ | 173 | 4834 | 6646 |

| Lisa | -g3XIcCb2b-BD0QBCcq2Sw | 159 | 48 | 13 |

| Cat | -9bbDysuiWeo2VShFJJtcw | 133 | 1062 | 672 |

| William | -FZBTkAZEXoP7CYvRV2ZwQ | 126 | 9363 | 9361 |

| Fran | -9da1xk7zgnnfO1uTVYGkA | 124 | 9851 | 7606 |

| Lissa | -lh59ko3dxChBSZ9U7LfUw | 120 | 455 | 150 |

| Mark | -B-QEUESGWHPE\_889WJaeg | 115 | 4008 | 570 |

| Tiffany | -DmqnhW4Omr3YhmnigaqHg | 111 | 1366 | 984 |

| bernice | -cv9PPT7IHux7XUc9dOpkg | 105 | 120 | 112 |

| Roanna | -DFCC64NXgqrxlO8aLU5rg | 104 | 2995 | 1188 |

| Angela | -IgKkE8JvYNWeGu8ze4P8Q | 101 | 158 | 164 |

| .Hon | -K2Tcgh2EKX6e6HqqIrBIQ | 101 | 7850 | 5851 |

| Ben | -4viTt9UC44lWCFJwleMNQ | 96 | 1180 | 1155 |

| Linda | -3i9bhfvrM3F1wsC9XIB8g | 89 | 3177 | 2736 |

| Christina | -kLVfaJytOJY2-QdQoCcNQ | 85 | 158 | 34 |

| Jessica | -ePh4Prox7ZXnEBNGKyUEA | 84 | 2161 | 2091 |

| Greg | -4BEUkLvHQntN6qPfKJP2w | 81 | 820 | 753 |

| Nieves | -C-l8EHSLXtZZVfUAUhsPA | 80 | 1091 | 774 |

| Sui | -dw8f7FLaUmWR7bfJ\_Yf0w | 78 | 9 | 18 |

| Yuri | -8lbUNlXVSoXqaRRiHiSNg | 76 | 1166 | 220 |

| Nicole | -0zEEaDFIjABtPQni0XlHA | 73 | 13 | 10 |

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

(Output limit exceeded, 25 of 10000 total rows shown)

Please explain your findings and interpretation of the results:

Based on the table above sorting the users based on their number of fans doesn’t show descending or ascending trend in “useful” or “funny” columns. Therefore, there shouldn’t be a strong correlation between having a high number of fans and being listed as "useful" or "funny.

**Part 2: Inferences and Analysis**

1. Pick one city and category of your choice and group the businesses in that city

or category by their overall star rating. Compare the businesses with 2-3 stars to

the businesses with 4-5 stars and answer the following questions. Include your code.

City: Mesa Category: Food

i. Do the two groups you chose to analyze have a different distribution of hours?

Yes

ii. Do the two groups you chose to analyze have a different number of reviews?

Yes

iii. Are you able to infer anything from the location data provided between these two groups? Explain.

Based on the results, we can see that there seems to be a correlation between the location of the business and their rating. The business that are probably located in the same neighbor have close rating. Also they have similar working hours. Moreover, the business that have longer working hours usually have higher rating.

SQL code used for analysis:

select

business.name

, business.city

, category.category

, business.stars

, hours.hours

, business.review\_count

, business.postal\_code

from (business inner join category on business.id = category.business\_id) inner join hours on hours.business\_id = category.business\_id

where business.city = 'Mesa'

group by business.stars;

2. Group business based on the ones that are open and the ones that are closed. What

differences can you find between the ones that are still open and the ones that are

closed? List at least two differences and the SQL code you used to arrive at your

answer.

i. Difference 1:

The business that are still open have higher rating.

ii. Difference 2:

The business that are still open have more reviews.

iii. Difference 3:

The business that are still open have longer working hours.

SQL code used for analysis:

select

business.name

, business.is\_open

, category.category

, business.stars

, hours.hours

, business.review\_count

, business.postal\_code

from (business inner join category on business.id = category.business\_id) inner join hours on hours.business\_id = category.business\_id

where business.city = 'Mesa'

group by business.is\_open;

3. For this last part of your analysis, you are going to choose the type of analysis you

want to conduct on the Yelp dataset and are going to prepare the data for analysis.

Ideas for analysis include: Parsing out keywords and business attributes for sentiment

analysis, clustering businesses to find commonalities or anomalies between them,

predicting the overall star rating for a business, predicting the number of fans a

user will have, and so on. These are just a few examples to get you started, so feel

free to be creative and come up with your own problem you want to solve. Provide

answers, in-line, to all of the following:

i. Indicate the type of analysis you chose to do:

Finding correlation between the likes with the given rates and using “like” in the reviews.

ii. Write 1-2 brief paragraphs on the type of data you will need for your analysis and why you chose that data:

I need two sources of data (tables). First, I join these two tables based on users and business. Then I sort them based on rating to see if there is a correlation between the number of stars and likes.

The reason I chose this analysis and thus, the data sets is that psychologists have shown that how people think about something can completely change even after a few minutes and they think that how people think just after occurrence of an event is a better representative for the quality of that event compared to what they say after thinking about it. Because tip table is related to the occurrence of the event (shopping) and they write a review after hours or even days, comparing these two tables can help us to explore the validity what psychologists claim. As the result shows there is a slight correlation between the number of likes and stars, but this correlation is not strong. So what psychologists claim seems to be fairly valid.

iii. Output of your finished dataset:

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

| stars | likes |

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

| 3 | 2 |

| 5 | 2 |

| 5 | 1 |

| 5 | 1 |

| 5 | 1 |

| 5 | 1 |

| 5 | 1 |

| 5 | 1 |

| 5 | 1 |

| 5 | 1 |

| 3 | 1 |

| 4 | 1 |

| 4 | 1 |

| 4 | 1 |

| 4 | 1 |

| 4 | 1 |

| 4 | 1 |

| 4 | 1 |

| 4 | 1 |

| 4 | 1 |

| 4 | 1 |

| 4 | 1 |

| 4 | 1 |

| 4 | 1 |

| 4 | 1 |

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

(Output limit exceeded, 25 of 1227 total rows shown)

iv. Provide the SQL code you used to create your final dataset:

select

review.stars

, tip.likes

from review inner join tip on review.user\_id = tip.user\_id

order by tip.likes desc;

**Answers**

Part 1, Question 1: Total Number of Records.

The answer for each of the tables should indicate 10,000. I.e. the answer should look similar to the following:

1. Attribute = 10,000 records
2. Business = 10,000 records
3. Category = 10,000 records
4. Checkin = 10,000 records
5. elite\_years = 10,000 records
6. friend = 10,000 records
7. hours = 10,000 records
8. photo = 10,000 records
9. review = 10,000 records
10. tip = 10,000 records
11. user = 10,000 records

Part 1, Question 2: Total Number of Distinct Records by Key.

The answers should be as follows:

1. Business = 10,000
2. Hours = 1562
3. Category = 2643
4. Attribute = 1115
5. Review = 10,000
6. Checkin = 493
7. Photo = 10,000
8. Tip = 537
9. User = 10,000
10. Friend = 11
11. Elite\_years = 2780

Part 1, Question 3: Null values in Users table?

The answer should be "no." There are no null values for this table. The

Part 1, Question 4: Mix, Max, and Avg.

The learner's answers should reflect the following correct answers:

**i. Table: Review, Column: Stars**

*min: 1 max: 5 avg: 3.7082 (okay if rounded)*

**ii. Table: Business, Column: Stars**

*min: 1 max: 5 avg: 3.6549 (okay if rounded)*

**iii. Table: Tip, Column: Likes**

*min: 0 max: 2 avg: 0.0144 (okay if rounded)*

**iv. Table: Checkin, Column: Count**

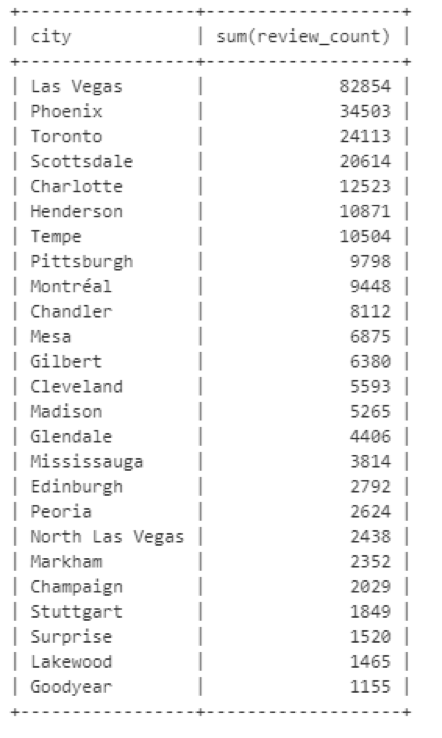
*min: 1 max: 53 avg: 1.9414 (okay if rounded)*

**v. Table: User, Column: Review\_count**

*min: 0 max: 2000 avg: 24.2995 (okay if rounded)*

Part 1, Question 5: Cities with Most Reviews in Descending Order

The learner provides the SQL code used along with the correct results which should be the following:



Part 1, Question 6: Distribution of Star Ratings

The learner should include both the SQL code to arrive at their answers for both Avon and Beachwood, and those answers should be in the form of the following distribution tables:

Avon

| **Star Rating** | **Count** |
| --- | --- |
| 0 | 0 |
| 1 | 0 |
| 1.5 | 1 |
| 2 | 0 |
| 2.5 | 2 |
| 3 | 1 |
| 3.5 | 2 |
| 4 | 2 |
| 4.5 | 1 |
| 5 | 1 |

Beachwood

|  |  |
| --- | --- |
| **Star Rating** | **Count** |

Part 1, Question 7: Top 3 users

The learner provided the code used to get their answer and the top 3 users should appear in the following order in their results:

* Gerald, 2000
* Sara, 1629
* Yuri, 1339
* Part 1, Question 8: Review/Fan Correlation
* The learner attempts to answer the question based on their findings. It doesn't matter whether the interpretation appears to make sense based on the data or not.

Part 1, Question 9: Love or Hate

The learner should provide the following answer as well as providing the SQL code they used to get there:

There are **more reviews with the word love** in them than the word hate.

Part 1, Question10: Top 10 users with most fans

The learner provided the code used to get their answer and the top 10 users should appear in the following order in their results:

1. Amy, 503
2. Mimi, 497
3. Harald, 311
4. Gerald, 253
5. Christine, 173
6. Lisa, 159
7. Cat, 133
8. William, 126
9. Fran, 124
10. Lissa, 120
11. Part 1, Question 11: Funny or Useful/Fan Correlation
12. The learner attempts to answer the question based on their findings. It doesn't matter whether the interpretation appears to make sense based on the data or not.

Part 2, Question 1: Grouping by Star Rating



4 pts

The learner provided the code they used for analysis, and answered each question. They also provided an inference that makes sense based on location data, and backed it up with an explanation.



3 pts

The learner attempted the question. They provided the code used for analysis and answered each question. They attempt an inference, but it doesn't appear to make logical sense based on their explanation.



1 pt

The learner attempted the question. They provided the code used for analysis, though it doesn't appear correct based on the analysis requested (doesn't attempt to look at distribution of hours or reviews) and/or an inference is missing.

Part 2, Question 2: Grouping by Open/Closed Status



4 pts

The learner provided the code they used for analysis, and provided a minimum of 2 differences they saw in the data.



2 pts

The learner provided the code they used for analysis, but only provided 1 difference they saw in the data.



1 pt

The learner provided the code they used for analysis, but it was done incorrectly so as to not indicate true differences between the groups.

Part 2, Question 3: Custom Analysis



6 pts

The learner did **all** of the following:

* Indicated the type of analysis
* Wrote a short explanation (1-2 paragraphs) indicating the type of data and why they chose to use that data
* An output of the final dataset
* The SQL code they used to get to the dataset



5 pts

The learner did all, **but one** of the following:

* Indicated the type of analysis
* Wrote a short explanation (1-2 paragraphs) indicating the type of data and why they chose to use that data
* An output of the final dataset
* The SQL code they used to get to the dataset



4 pts

The learner did all, **but two** of the following:

* Indicated the type of analysis
* Wrote a short explanation (1-2 paragraphs) indicating the type of data and why they chose to use that data
* An output of the final dataset
* The SQL code they used to get to the dataset



0 pts

The learner provided anything less than indicated in the last point. I.e. they ***didn't do three or more****of the following:*

* Indicated the type of analysis
* Wrote a short explanation (1-2 paragraphs) indicating the type of data and why they chose to use that data
* An output of the final dataset
* The SQL code they used to get to the dataset