INFS 3200: Assignment

Name: Peng Yu

Student ID: 46635884

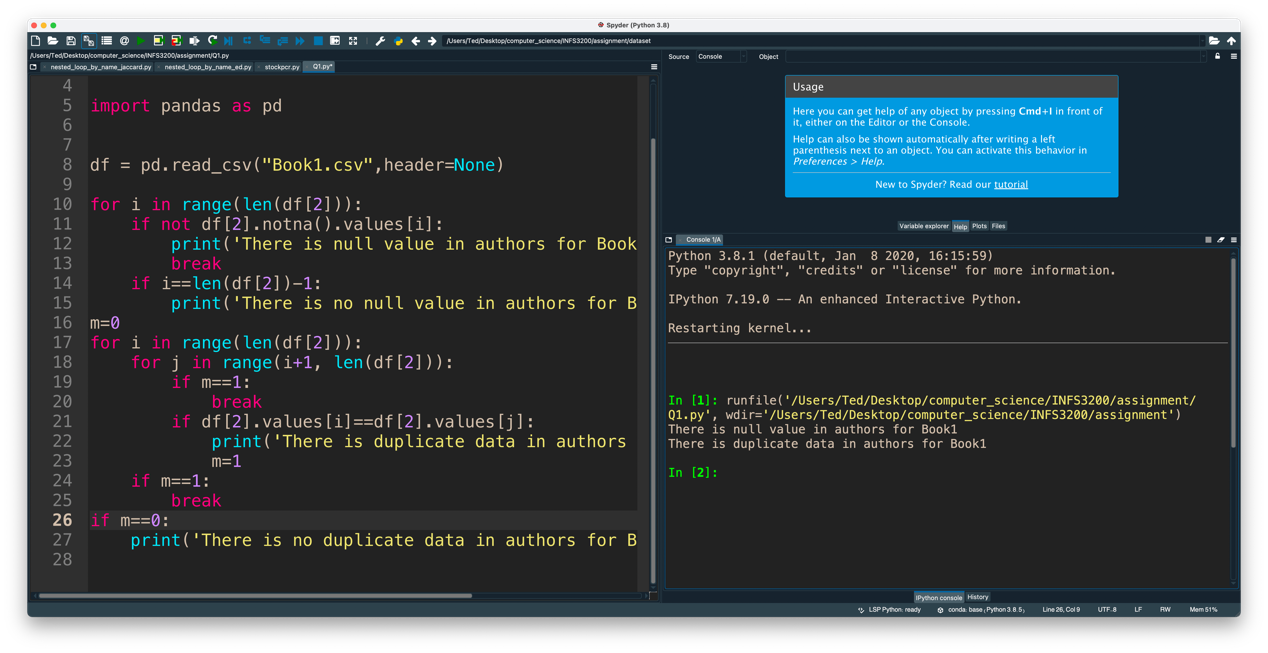
Part 1

Question 1

Answer (1):

The attribute authors cannot be the primary key for Book 1.

Because if an attribute wants to be a primary key, it must have unique and not null. But when we examine the attribute authors, we find that it is neither unique nor not null. Therefore it cannot be the primary key for Book 1.



Answer (2):

Select in DML (Data Manipulation Language) can answering this query.

Code:

SELECT SalesRank, Title, PaperbackPrice, HardcoverPrice, EbookPrice, AudiobookPrice

FROM Book3

WHERE SalesRank < 100

ORDER BY SalesRank asc ;

Question 2

Answer (1):

Vertically Fragmented

We break up Book2 into two fragments, Fragment1 and Fragment2. Each fragment keeps the primary key ID of Book2 is the key for connecting them, and make sure they include all the attributes in Book2

Fragment1: id, publication\_year, language, pages, publisher\_name.

Fragment2: id, book\_title, authors, publication\_month, publication\_day, edition, isbn13, series.

Answer (2):

First, make sure that the id and publication\_day of the inserted data cannot be null.

Second, compare the id of the inserted data with the id of the three fragments one by one to see if there is any duplicate id. If there is, the inserted data cannot be inserted.

Finally, find the publication\_day data of the inserted data, if this value 1 and  10, then insert Fragment1, if this value 11 and 20, then insert Fragment2, if this value 21 and 31, then insert Fragment3.

Part 2

Question 3

Answer:

The Day, Publisher and Language are dimension columns.

The Sales is a fact column.

Question 4

Answer (1):

The advantages of them are that they have a highly compressed structure, making them fast to read and their structure makes it possible for the system to combine multiple indexes together for fast access to the underlying table.

Compressed indexes, like bitmap indexes, represent a trade-off between CPU usage and disk space usage. A compressed structure is faster to read from disk but takes additional CPU cycles to decompress for access - an uncompressed structure imposes a lower CPU load but requires more bandwidth to read in a short time.

One belief concerning bitmap indexes is that they are only suitable for indexing low-cardinality data. This is not necessarily true, and bitmap indexes can be used very successfully for indexing columns with many thousands of different values.

Bitmap indexes apply to scenarios

1. Build on columns with high-value repeatability.

2. Specific types of queries such as COUNT, OR, AND logical operations because only bitwise operations are required.

3. It is suitable for OLAP scenarios such as Ad-hoc query and multi-dimensional analysis.

Answer (2):

Create bitmap indices for “Publisher”.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Day | 07/15/1984 | 05/05/1990 | 16/04/1995 | 12/11/2000 | 04/03/2004 | 05/01/2008 | 11/19/2012 | 08/06/2014 |
| Row | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| AAAI Press | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| Springer International Publishing | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| Springer London | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| IEEE Computer Society Press | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |

Create bitmap indices for “Language”.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Day | 07/15/1984 | 05/05/1990 | 06/04/1995 | 12/11/2000 | 04/03/2004 | 05/01/2008 | 11/19/2012 | 08/06/2014 |
| Row | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| English | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| Spanish | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |

Answer (3):

First, we find out the bitmap indices for “Spanish” and “AAAI Press”.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Day | 07/15/1984 | 05/05/1990 | 06/04/1995 | 12/11/2000 | 04/03/2004 | 05/01/2008 | 11/19/2012 | 08/06/2014 |
| Row | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| AAAI Press | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| Spanish | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |

Second, we intersect “Spanish” and “AAAI Press” and get the result as shown in the figure below.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Day | 07/15/1984 | 05/05/1990 | 06/04/1995 | 12/11/2000 | 04/03/2004 | 05/01/2008 | 11/19/2012 | 08/06/2014 |
| Row | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Spanish AAAI Press | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |

So we get that Spanish AAAI Press is equal to 1 and the row is 5 and the day is 04/03/2004. Therefore the total sales of “Spanish” books published by “AAAI Press” is 2.

|  |  |  |  |
| --- | --- | --- | --- |
| Day | Publisher | Language | Sales |
| 04/03/2004 | AAAI Press | Spanish | 2 |

Part 3

Question 5

Answers (1):

Book1

(id, title, authors, pubyear, pubmonth, pubday, edition, publisher, isbn13, language, series, pages)

Book2

(id, book\_title, authors, publication\_year, publication\_month, publication\_day, edition, publisher\_name, isbn13, language, series, pages)

Book3

(ID, Title, Author1, Author2, Author3, Publisher, ISBN13, Date, Pages, ProductDimensions, SalesRank, RatingsCount, RatingValue, PaperbackPrice, HardcoverPrice, EbookPrice, AudiobookPrice)

Book4

(ID, Title, UsedPrice, NewPrice, Author, ISBN10, ISBN13, Publisher, Publication\_Date, Pages, Dimensions)

The common attributes of Book1, Book2, Book3, and Book4 are found out and taken as a global conceptual schema.

Book Global View

(ID, Title, Author, ISBN13, Publisher, Publication\_Date, Pages)

Answers (2):

First, the ID in Book1, Book2, and Book3 is INT types (example: 1992), and the ID in Book4 is CHARACTER types (example: HC1992). If these tables are integrated, it may cause structural heterogeneity.

Solutions:

We will add an attribute (int\_id) to Book4 to make its INT type and have a 1:1 relationship with book4.ID.

Second, the year in Book3 and Book4 is represented by four digits (example: 2008), but the year in Book1 and Book2 is represented by two digits (example: 13). In Book3, dates are represented with numbers (example: 06/30/2008), while in Book4, dates are represented with letters (example: Jun.17th.2008). If these tables are integrated, it may cause semantic heterogeneity.

Solutions:

We will add an attribute (date) to Book1 with the format dd/mm/yyyy, and simply change the information for book1.pubday, pubmonth, pubday (Example, dd: 9 changed to 09, mm: 3 changed to 03, and yyyy: 15 changed to 2015) , and then fill in dd, mm, yyyy.

Book2 is same as Book1.

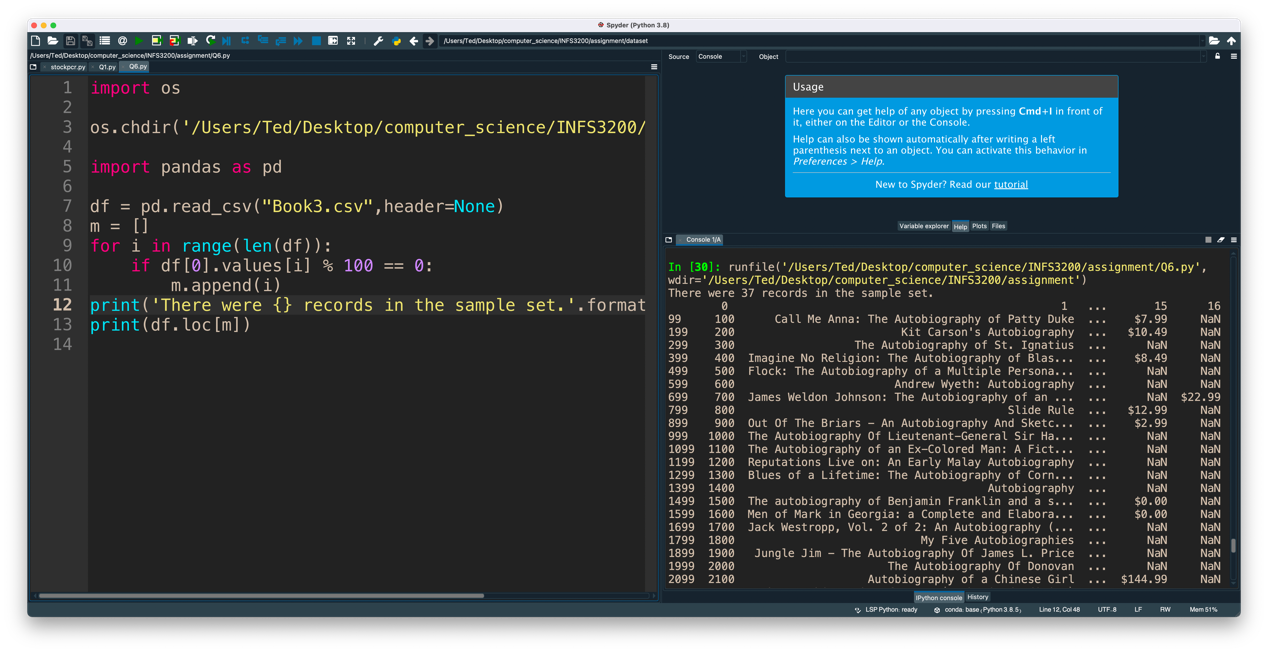
We will add an attribute (date) to Book4 with the format dd/mm/yyyy, convert the format of book4.publication\_date (Example: Jun.17th.2008 -> 06/17/2008), and fill in the date.

Part 4

Question 6

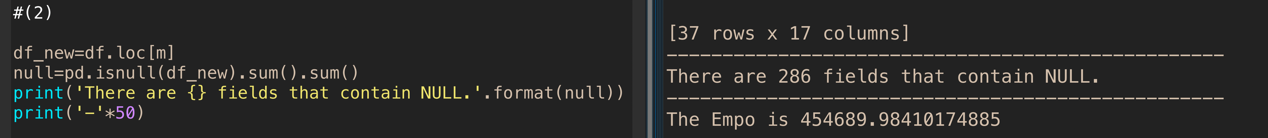
Answers (1):

There were 37 records in the sample set.



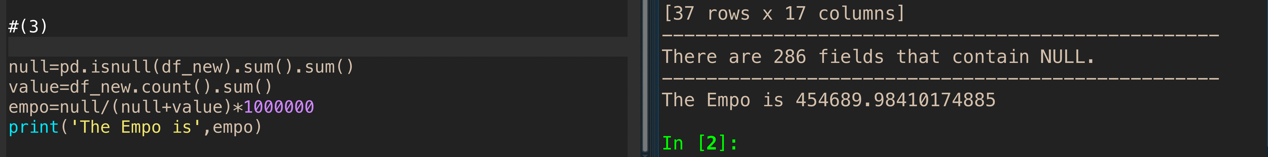
Answers (2):

There are 286 fields that contain NULL.



Answers (3):

The Empo is 454689.98410174885.



Question 7

Answers (1):

Edit Distance = 0.4137931034482759

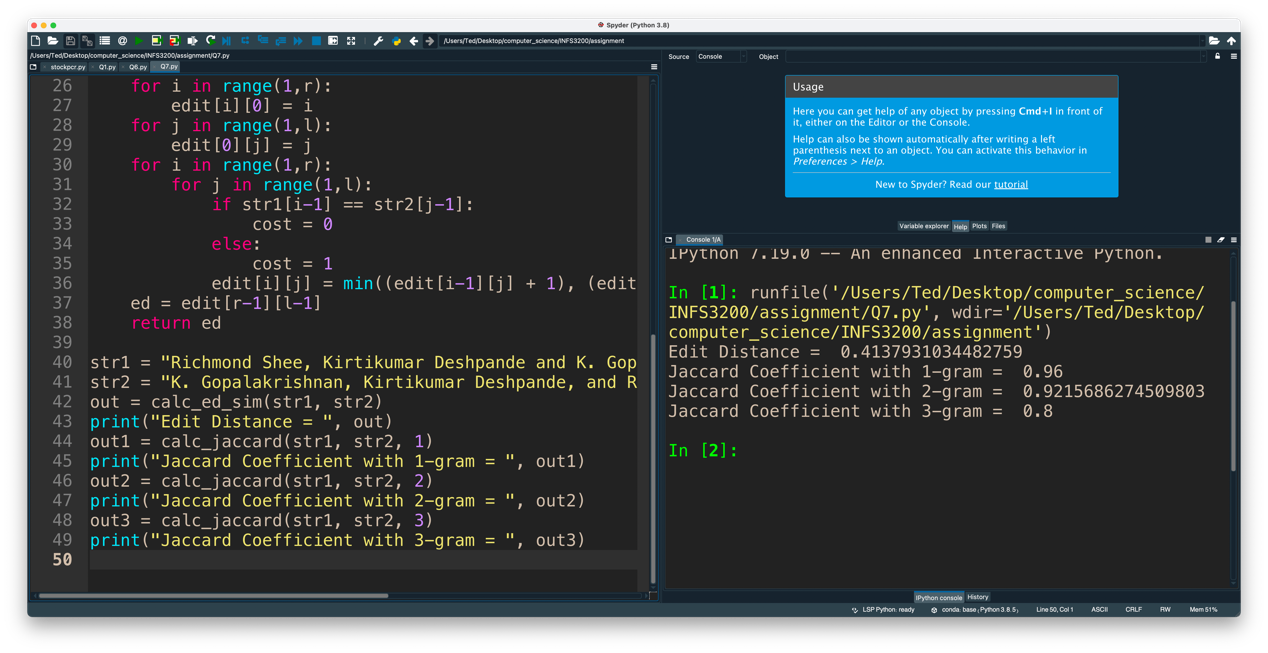
Jaccard Coefficient with 1-gram = 0.96

Jaccard Coefficient with 2-gram = 0.9215686274509803

Jaccard Coefficient with 3-gram = 0.8

Through the calculation of Edit distance and Jaccard coefficient, it can be found that the similarity of the Jaccard coefficient is higher.

Edit distance is more suitable for the comparison between words, and Jaccard coefficient is more suitable for the comparison between sentences.



Answers (2):

Precision= 0.18450826805918188, Recall= 0.9137931034482759, Fmeasure= 0.3070238957277335

