Database Management Practicum 2

GROUP 4:

YANG HE

YIDAN ZHU

YIXING CHEN

ZHENG ZHENG

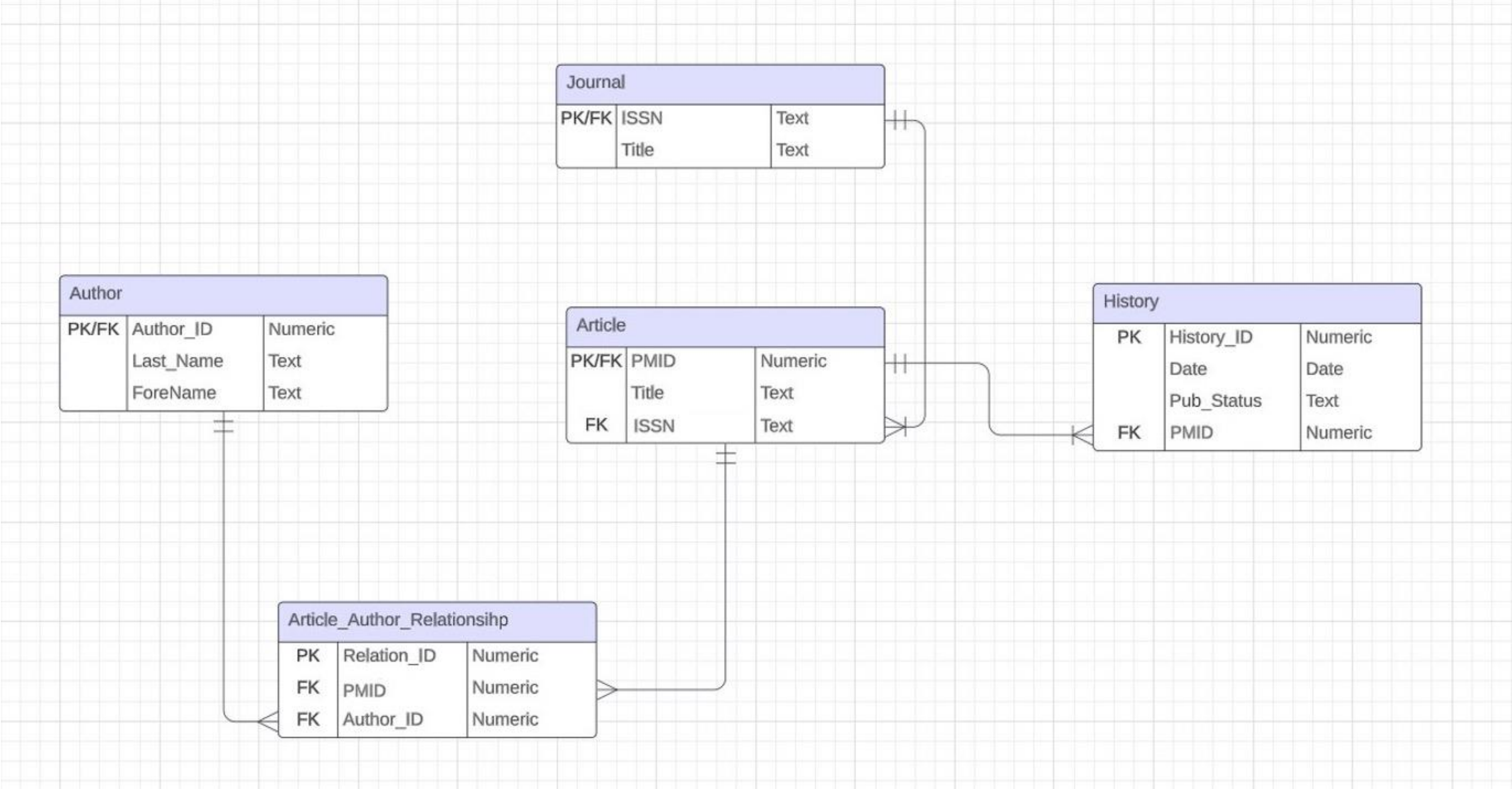
Content

Part1 Normalized Relational schema and XML data loading

Part2 Analysis database & Summary Table

Part3 Data Mining

Part 1.1 &1.2 Normalized Relational Schema



Journal		
PK/FK	ISSN	Text
	Title	Text

Author		
PK/FK	Author_ID	Numeric
	Last_Name	Text
	ForeName	Text

Article		
PK/FK	PMID	Numeric
	Title	Text
FK	ISSN	Text

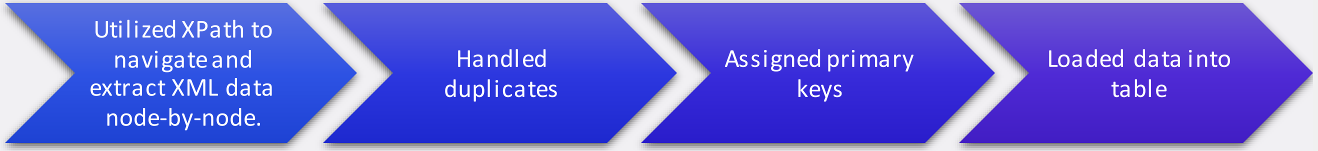
History		
PK	History_ID	Numeric
	Date	Date
	Pub_Status	Text
FK	PMID	Numeric

Article_Author_Relationsihp		
PK	Relation_ID	Numeric
FK	PMID	Numeric
FK	Author_ID	Numeric

Part 1.3 XML Data Loading – Create table

Journal	*ISSN* (Primary Key) Title
Article	*PMID* (Primary Key) Title ISSN (Foreign Key referencing Journal's ISSN)
Author	*Author_ID* (Primary Key) LastName ForeName
Author_Article_Relationship	*Relation_ID* (Primary Key) PMID (Foreign Key referencing Article's PMID) Author_ID (Foreign Key referencing Author's Author_ID)
History	*History_ID* (Primary Key) History_Date Pub_Status(received, accepted, epublish, entrez, pubmed, medline) PMID (Foreign Key referencing Article's PMID)

Part 1.3 XML Data Loading – Load data

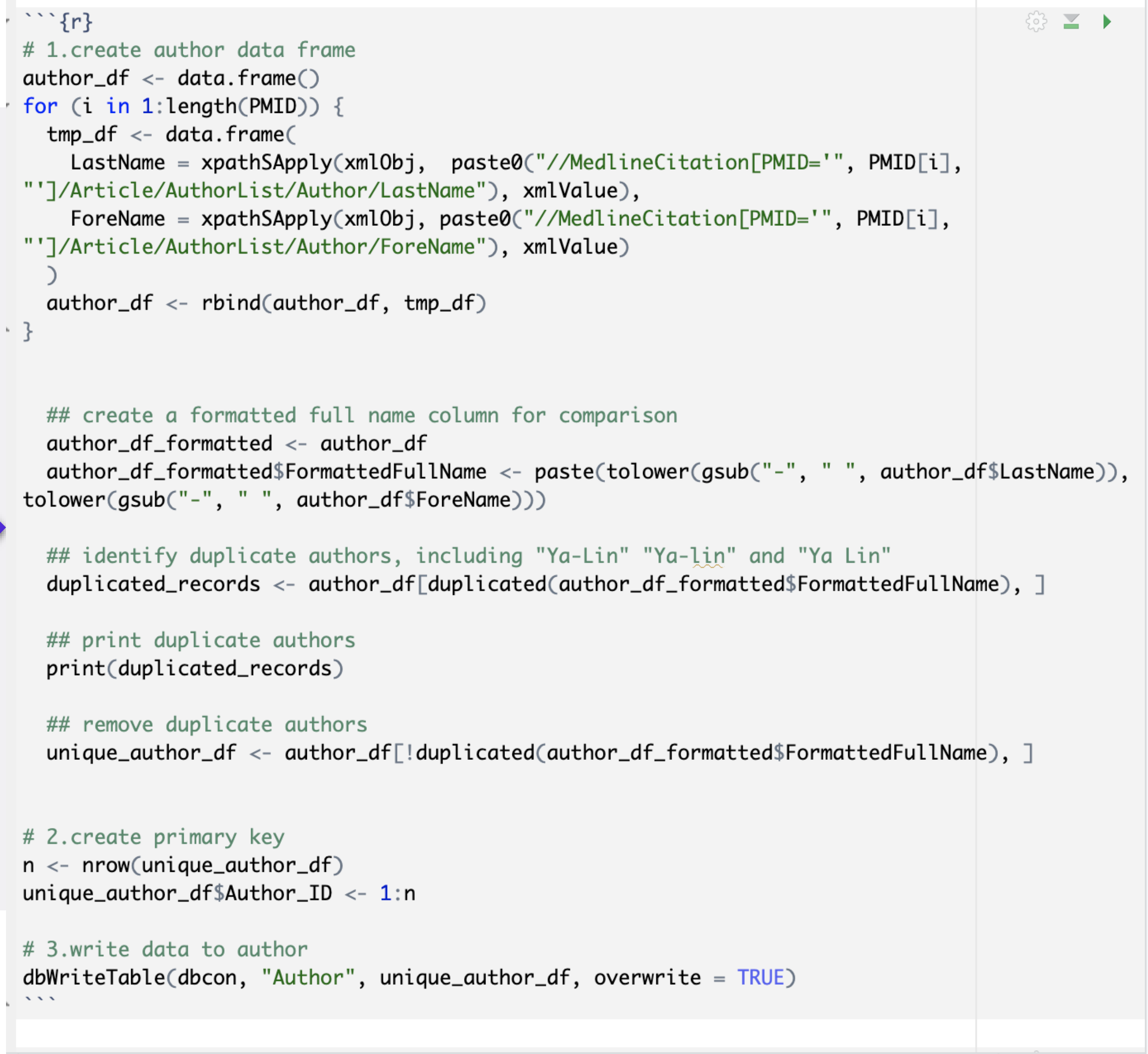


Utilized XPath to navigate and extract XML data node-by-node.

Handled duplicates

Assigned primary keys

Loaded data into table



~~~
```{r}
# 1.create author data frame
author_df <- data.frame()
for (i in 1:length(PMID)) {
  tmp_df <- data.frame(
    LastName = xpathSApply(xmlObj,  paste0("//MedlineCitation[PMID='", PMID[i],
"']/Article/AuthorList/Author/LastName"), xmlValue),
    ForeName = xpathSApply(xmlObj, paste0("//MedlineCitation[PMID='", PMID[i],
"']/Article/AuthorList/Author/ForeName"), xmlValue)
  )
  author_df <- rbind(author_df, tmp_df)
}


  ## create a formatted full name column for comparison
  author_df_formatted <- author_df
  author_df_formatted$FormattedFullName <- paste(tolower(gsub("-", " ", author_df$LastName)),
tolower(gsub("-", " ", author_df$ForeName)))

  ## identify duplicate authors, including "Ya-Lin" "Ya-lin" and "Ya Lin"
  duplicated_records <- author_df[duplicated(author_df_formatted$FormattedFullName), ]

  ## print duplicate authors
  print(duplicated_records)

  ## remove duplicate authors
  unique_author_df <- author_df[!duplicated(author_df_formatted$FormattedFullName), ]


# 2.create primary key
n <- nrow(unique_author_df)
unique_author_df$Author_ID <- 1:n

# 3.write data to author
dbWriteTable(dbcon, "Author", unique_author_df, overwrite = TRUE)
```
~~~

Part 1.3 XML Data Loading - Query data

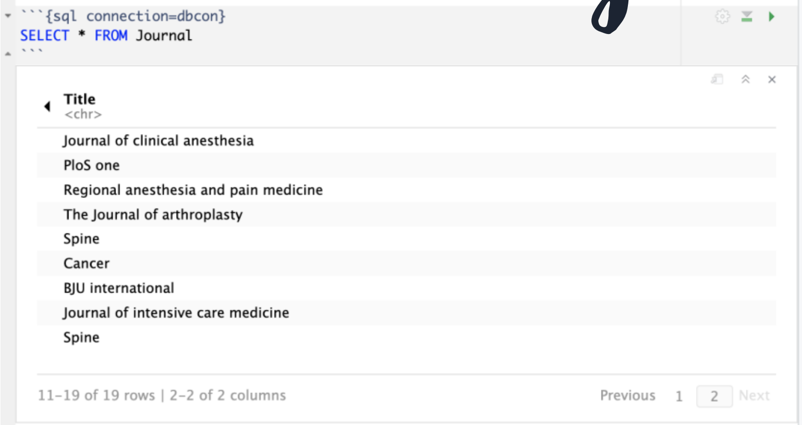
Journal:
19 records

Article:
19 records

Author:
82 records

Author_Article_Relationship:
136 records

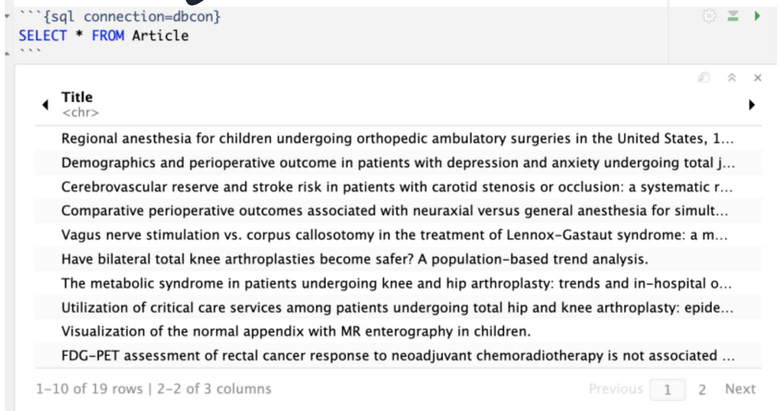
History:
94 records



~~~
```{sql connection=dbcon}
SELECT * FROM Journal
```
~~~

Title <chr>
Journal of clinical anesthesia
PloS one
Regional anesthesia and pain medicine
The Journal of arthroplasty
Spine
Cancer
BJU international
Journal of intensive care medicine
Spine

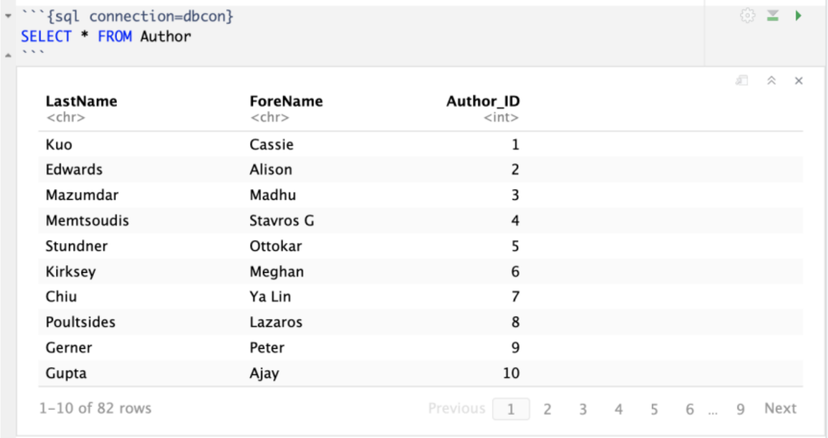
11–19 of 19 rows | 2–2 of 2 columns   Previous 1 2 Next



~~~
```{sql connection=dbcon}
SELECT * FROM Article
```
~~~

Title <chr>
Regional anesthesia for children undergoing orthopedic ambulatory surgeries in the United States, 1...
Demographics and perioperative outcome in patients with depression and anxiety undergoing total j...
Cerebrovascular reserve and stroke risk in patients with carotid stenosis or occlusion: a systematic r...
Comparative perioperative outcomes associated with neuraxial versus general anesthesia for simult...
Vagus nerve stimulation vs. corpus callosotomy in the treatment of Lennox–Gastaut syndrome: a m...
Have bilateral total knee arthroplasties become safer? A population–based trend analysis.
The metabolic syndrome in patients undergoing knee and hip arthroplasty: trends and in–hospital o...
Utilization of critical care services among patients undergoing total hip and knee arthroplasty: epide...
Visualization of the normal appendix with MR enterography in children.
FDG–PET assessment of rectal cancer response to neoadjuvant chemoradiotherapy is not associated ...

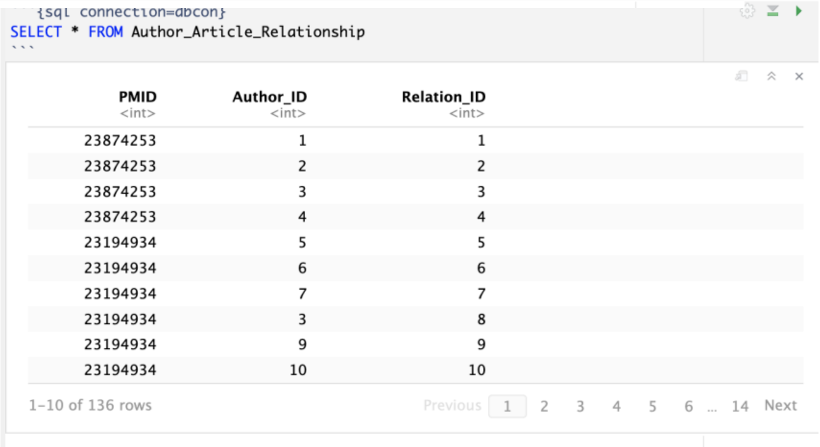
1–10 of 19 rows | 2–2 of 3 columns   Previous 1 2 Next



~~~
```{sql connection=dbcon}
SELECT * FROM Author
```
~~~

LastName <chr>	ForeName <chr>	Author_ID <int>
Kuo	Cassie	1
Edwards	Alison	2
Mazumdar	Madhu	3
Memtsoudis	Stavros G	4
Stundner	Ottokar	5
Kirksey	Meghan	6
Chiu	Ya Lin	7
Poultsides	Lazaros	8
Gerner	Peter	9
Gupta	Ajay	10

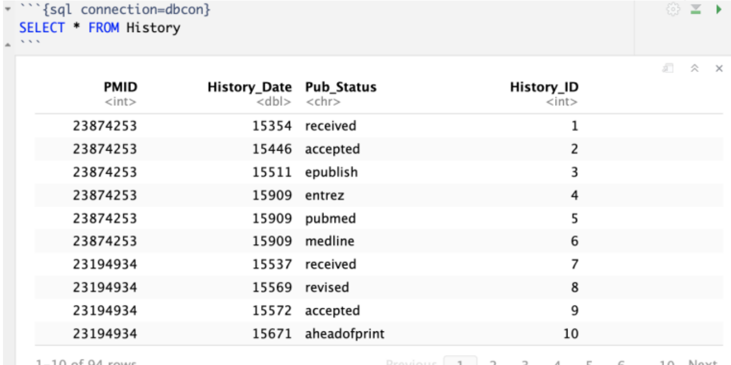
1–10 of 82 rows   Previous 1 2 3 4 5 6 ... 9 Next



~~~
```{sql connection=dbcon}
SELECT * FROM Author_Article_Relationship
```
~~~

PMID <int>	Author_ID <int>	Relation_ID <int>
23874253	1	1
23874253	2	2
23874253	3	3
23874253	4	4
23194934	5	5
23194934	6	6
23194934	7	7
23194934	3	8
23194934	9	9
23194934	10	10

1–10 of 136 rows   Previous 1 2 3 4 5 6 ... 14 Next



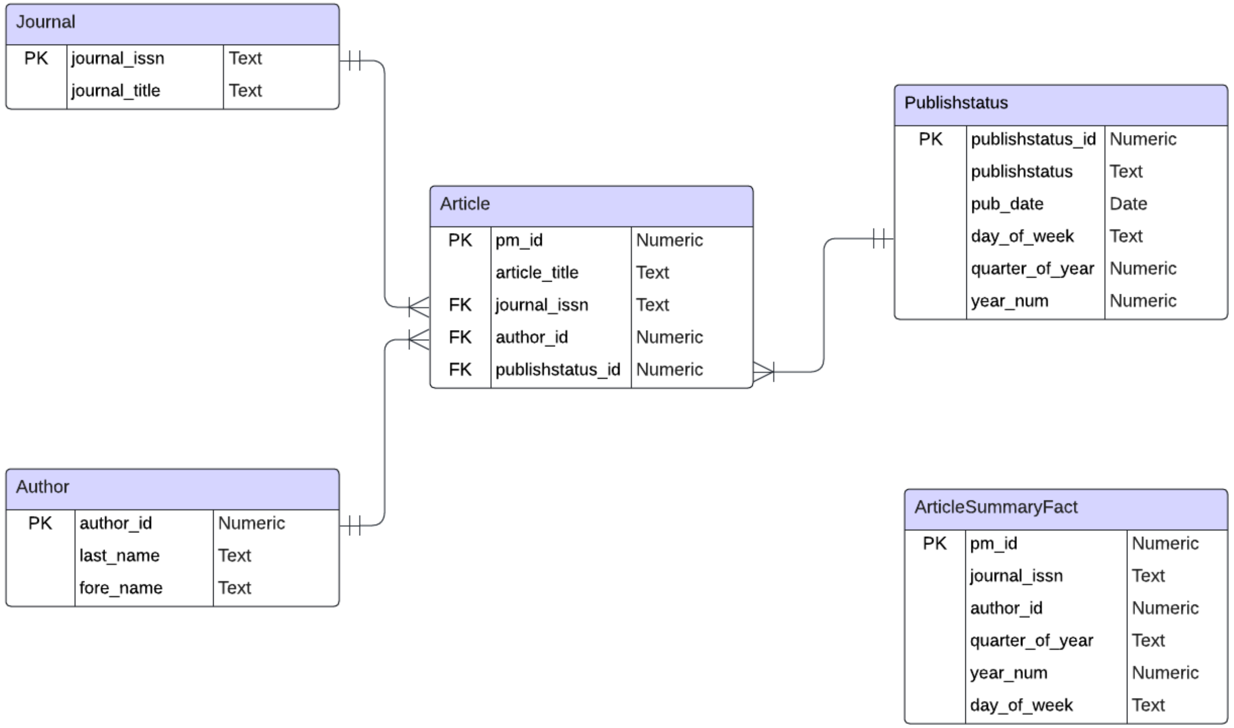
~~~
```{sql connection=dbcon}
SELECT * FROM History
```
~~~

PMID <int>	History_Date <dbl>	Pub_Status <chr>	History_ID <int>
23874253	15354	received	1
23874253	15446	accepted	2
23874253	15511	epublish	3
23874253	15909	entrez	4
23874253	15909	pubmed	5
23874253	15909	medline	6
23194934	15537	received	7
23194934	15569	revised	8
23194934	15572	accepted	9
23194934	15671	aheadofprint	10

1–10 of 94 rows   Previous 1 2 3 4 5 6 ... 10 Next

Part 2.1 Analysis Database

-Design star schema(reduce the use of join)



Journal

PK	journal_issn	Text
	journal_title	Text

Article

PK	pm_id	Numeric
	article_title	Text
FK	journal_issn	Text
FK	author_id	Numeric
FK	publishstatus_id	Numeric

Publishstatus

PK	publishstatus_id	Numeric
	publishstatus	Text
	pub_date	Date
	day_of_week	Text
	quarter_of_year	Numeric
	year_num	Numeric

Author

PK	author_id	Numeric
	last_name	Text
	fore_name	Text

ArticleSummaryFact

PK	pm_id	Numeric
	journal_issn	Text
	author_id	Numeric
	quarter_of_year	Text
	year_num	Numeric
	day_of_week	Text

Transaction fact tables:

- Article

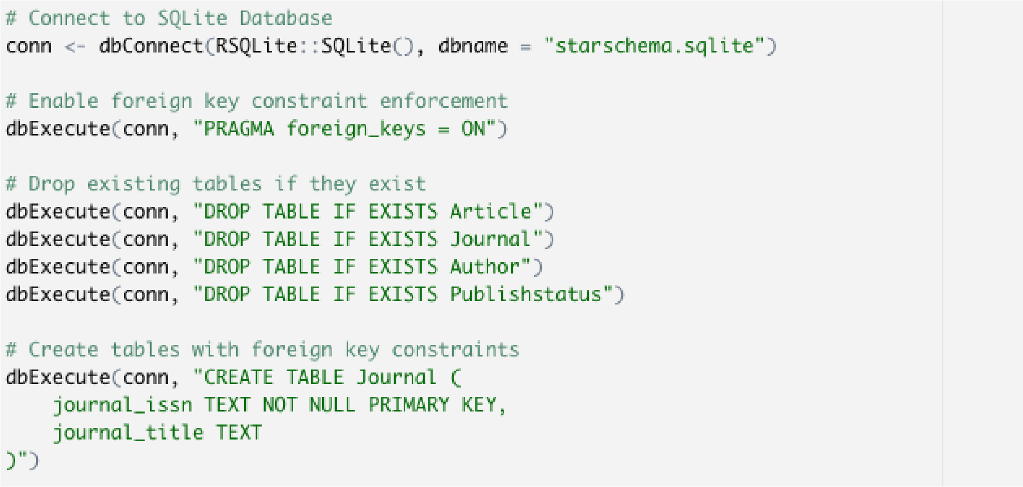
Dimension tables:

- Jornal
- Author
- Publishstatus

Part 2.1 Analysis Database

-code part 1

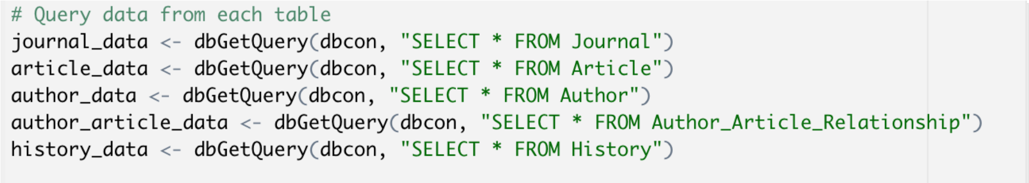
Build database



```
# Connect to SQLite Database
conn <- dbConnect(RSQLite::SQLite(), dbname = "starschema.sqlite")

# Enable foreign key constraint enforcement
dbExecute(conn, "PRAGMA foreign_keys = ON")

# Drop existing tables if they exist
dbExecute(conn, "DROP TABLE IF EXISTS Article")
dbExecute(conn, "DROP TABLE IF EXISTS Journal")
dbExecute(conn, "DROP TABLE IF EXISTS Author")
dbExecute(conn, "DROP TABLE IF EXISTS Publishstatus")

# Create tables with foreign key constraints
dbExecute(conn, "CREATE TABLE Journal (
    journal_issn TEXT NOT NULL PRIMARY KEY,
    journal_title TEXT
)")
```

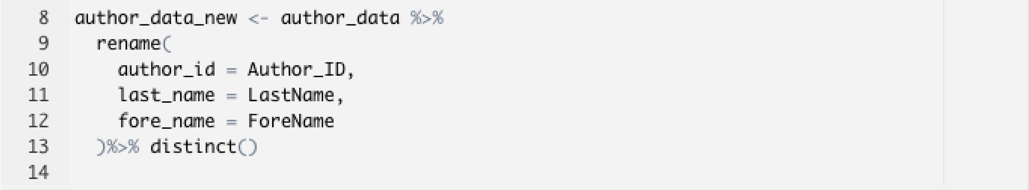
Drop TABLE IF EXISTS ...

Retrieve data from transaction data base Simple data processing for dimension



```
# Query data from each table
journal_data <- dbGetQuery(dbcon, "SELECT * FROM Journal")
article_data <- dbGetQuery(dbcon, "SELECT * FROM Article")
author_data <- dbGetQuery(dbcon, "SELECT * FROM Author")
author_article_data <- dbGetQuery(dbcon, "SELECT * FROM Author_Article_Relationship")
history_data <- dbGetQuery(dbcon, "SELECT * FROM History")
```

dbGetQuery



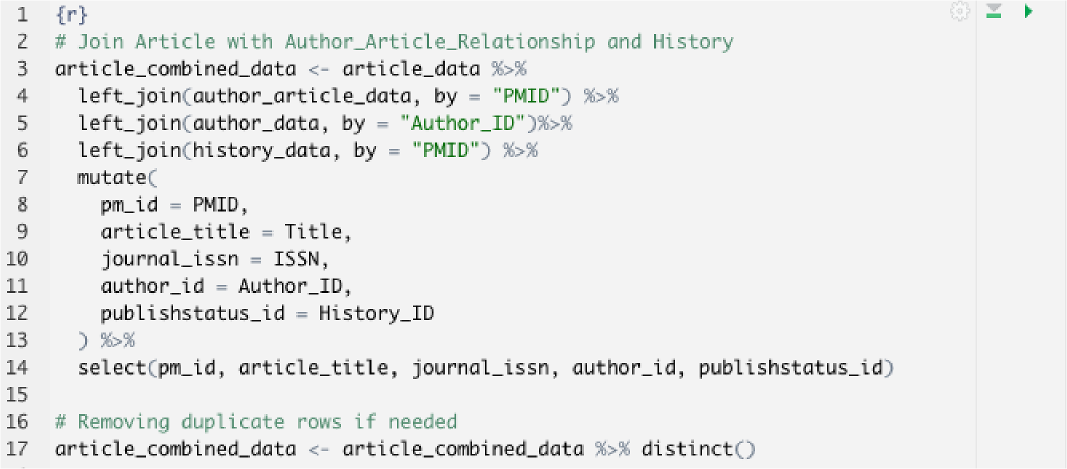
```
author_data_new <- author_data %>%
  rename(
    author_id = Author_ID,
    last_name = LastName,
    fore_name = ForeName
  )%>% distinct()

```

Rename columns based on the new database

Part 2.1 Analysis Database

-code part 2

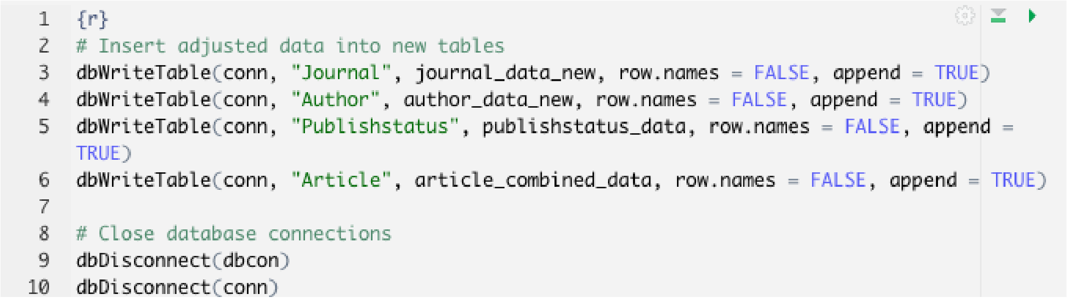
Join tables to create the Fact table



```r
{r}
# Join Article with Author_Article_Relationship and History
article_combined_data <- article_data %>%
  left_join(author_article_data, by = "PMID") %>%
  left_join(author_data, by = "Author_ID")%>%
  left_join(history_data, by = "PMID") %>%
  mutate(
    pm_id = PMID,
    article_title = Title,
    journal_issn = ISSN,
    author_id = Author_ID,
    publishstatus_id = History_ID
  ) %>%
  select(pm_id, article_title, journal_issn, author_id, publishstatus_id)

# Removing duplicate rows if needed
article_combined_data <- article_combined_data %>% distinct()
```

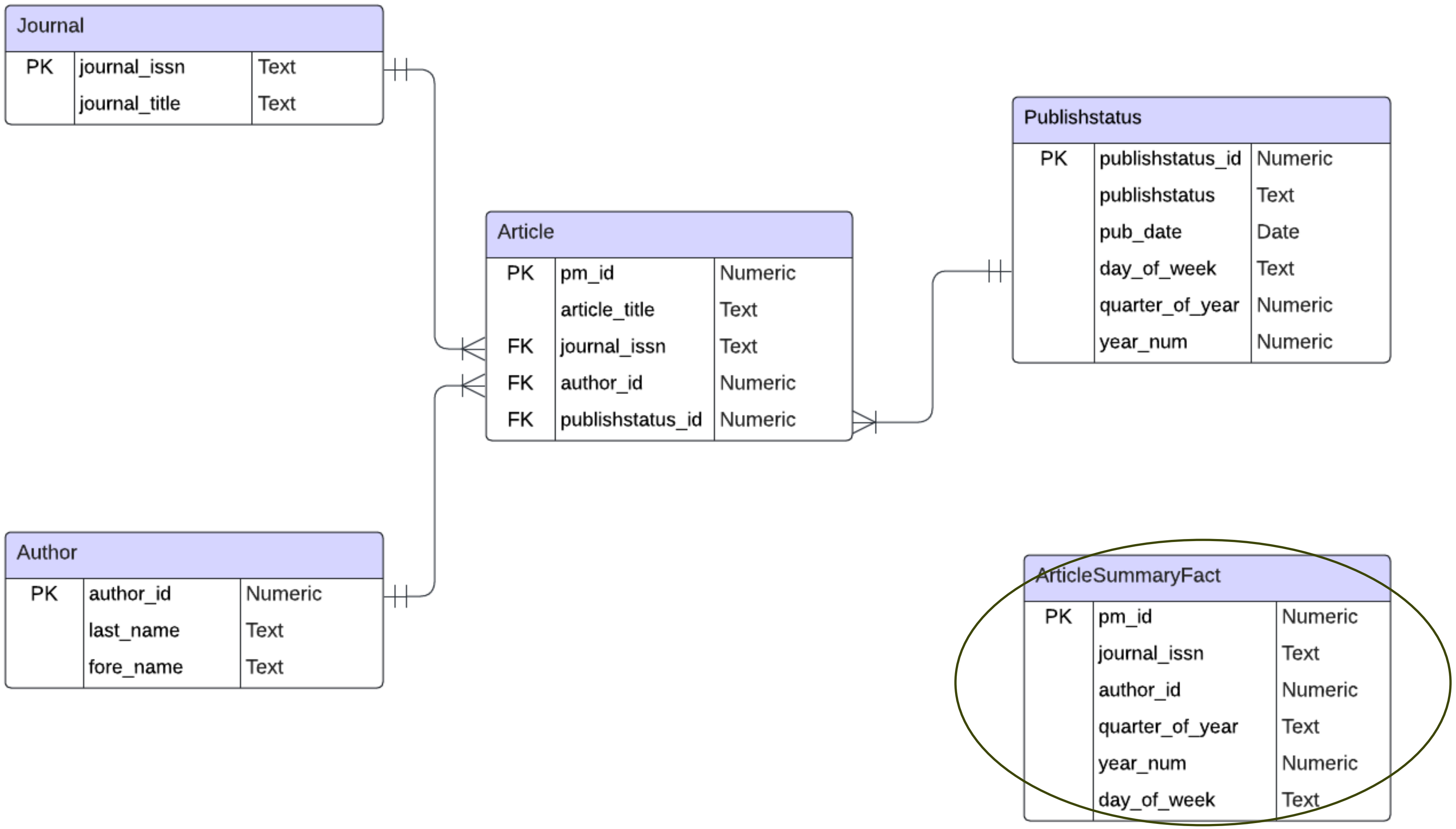
Here I use (dplyr) library

Write data into new database



```r
{r}
# Insert adjusted data into new tables
dbWriteTable(conn, "Journal", journal_data_new, row.names = FALSE, append = TRUE)
dbWriteTable(conn, "Author", author_data_new, row.names = FALSE, append = TRUE)
dbWriteTable(conn, "Publishstatus", publishstatus_data, row.names = FALSE, append = TRUE)
dbWriteTable(conn, "Article", article_combined_data, row.names = FALSE, append = TRUE)

# Close database connections
dbDisconnect(dbcon)
dbDisconnect(conn)
```

Part 2.2 Create a summary fact table



Journal

PK	journal_issn	Text
	journal_title	Text

Article

PK	pm_id	Numeric
	article_title	Text
FK	journal_issn	Text
FK	author_id	Numeric
FK	publishstatus_id	Numeric

Publishstatus

PK	publishstatus_id	Numeric
	publishstatus	Text
	pub_date	Date
	day_of_week	Text
	quarter_of_year	Numeric
	year_num	Numeric

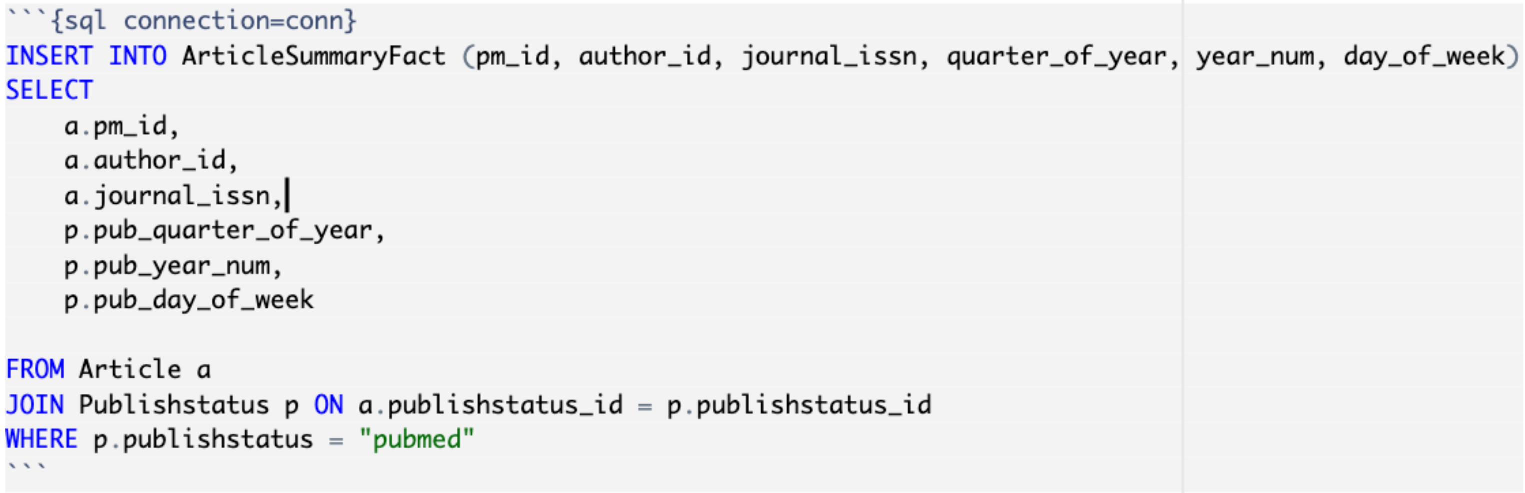
Author

PK	author_id	Numeric
	last_name	Text
	fore_name	Text

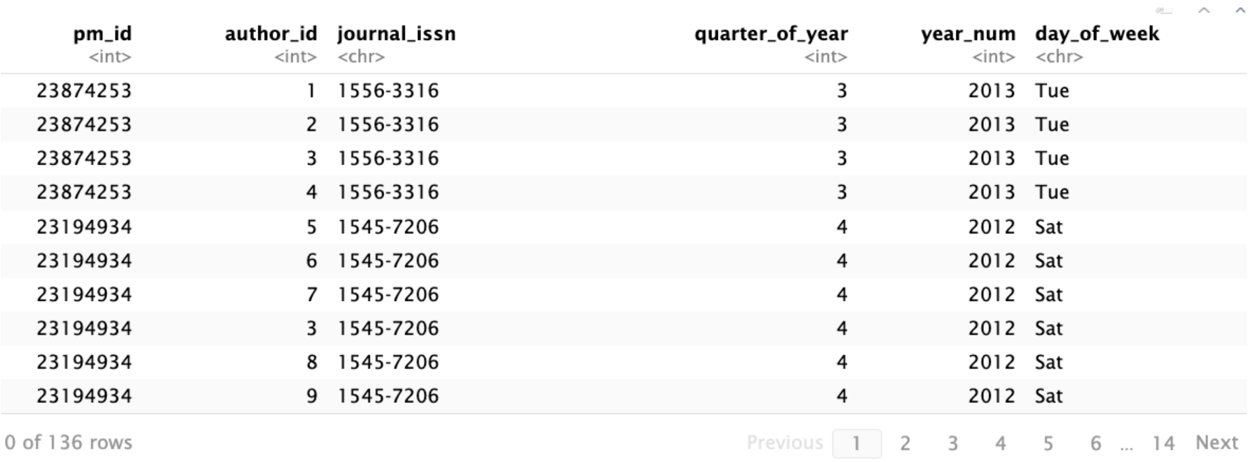
ArticleSummaryFact

PK	pm_id	Numeric
	journal_issn	Text
	author_id	Numeric
	quarter_of_year	Text
	year_num	Numeric
	day_of_week	Text

Part 2.2 Load the information



~~~~
```{sql connection=conn}
INSERT INTO ArticleSummaryFact (pm_id, author_id, journal_issn, quarter_of_year, year_num, day_of_week)
SELECT
    a.pm_id,
    a.author_id,
    a.journal_issn,
    p.pub_quarter_of_year,
    p.pub_year_num,
    p.pub_day_of_week

FROM Article a
JOIN Publishstatus p ON a.publishstatus_id = p.publishstatus_id
WHERE p.publishstatus = "pubmed"
```
~~~~

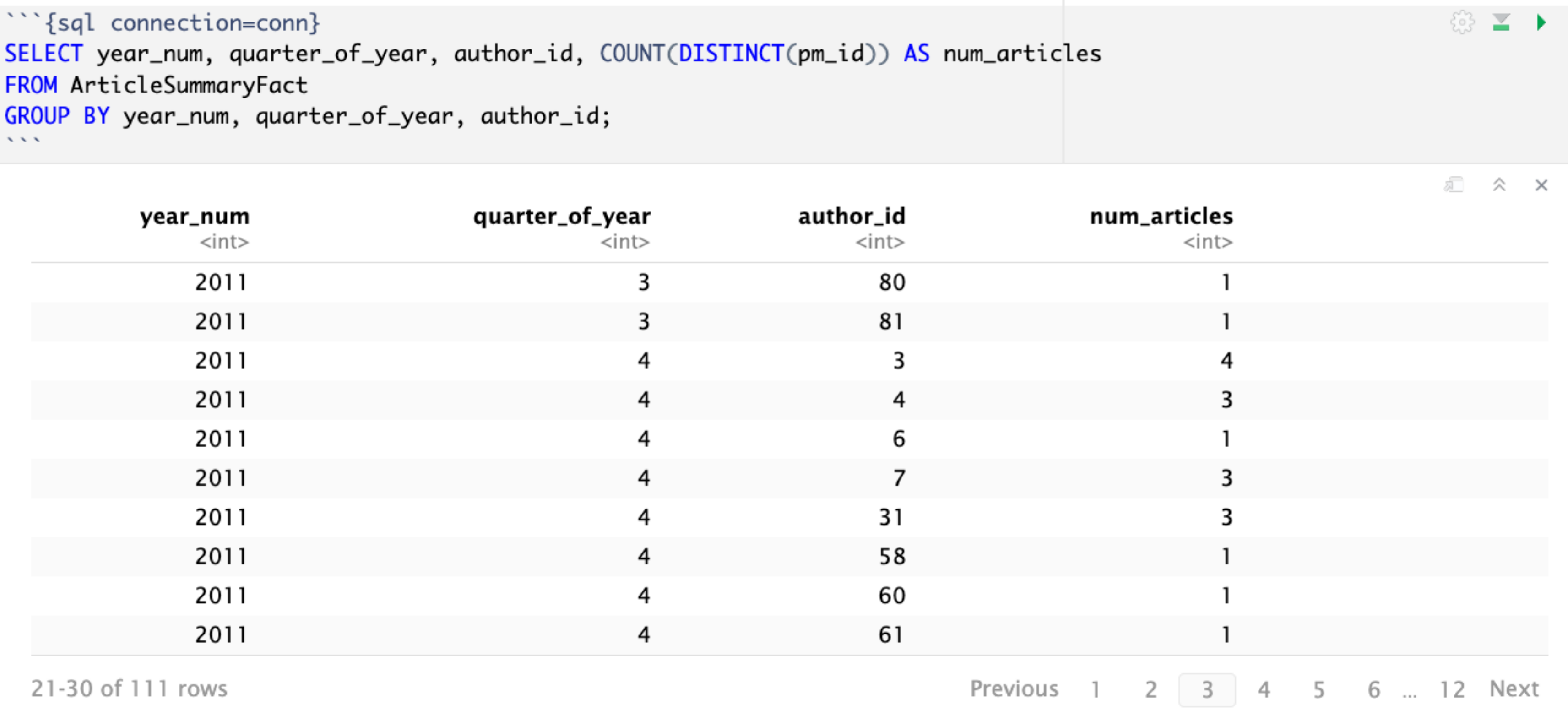


pm_id	author_id	journal_issn	quarter_of_year	year_num	day_of_week
<int>	<int>	<chr>	<int>	<int>	<chr>
23874253	1	1556-3316	3	2013	Tue
23874253	2	1556-3316	3	2013	Tue
23874253	3	1556-3316	3	2013	Tue
23874253	4	1556-3316	3	2013	Tue
23194934	5	1545-7206	4	2012	Sat
23194934	6	1545-7206	4	2012	Sat
23194934	7	1545-7206	4	2012	Sat
23194934	3	1545-7206	4	2012	Sat
23194934	8	1545-7206	4	2012	Sat
23194934	9	1545-7206	4	2012	Sat

0 of 136 rows

Previous 1 2 3 4 5 6 ... 14 Next

Part 2.2 number of articles per time period by author

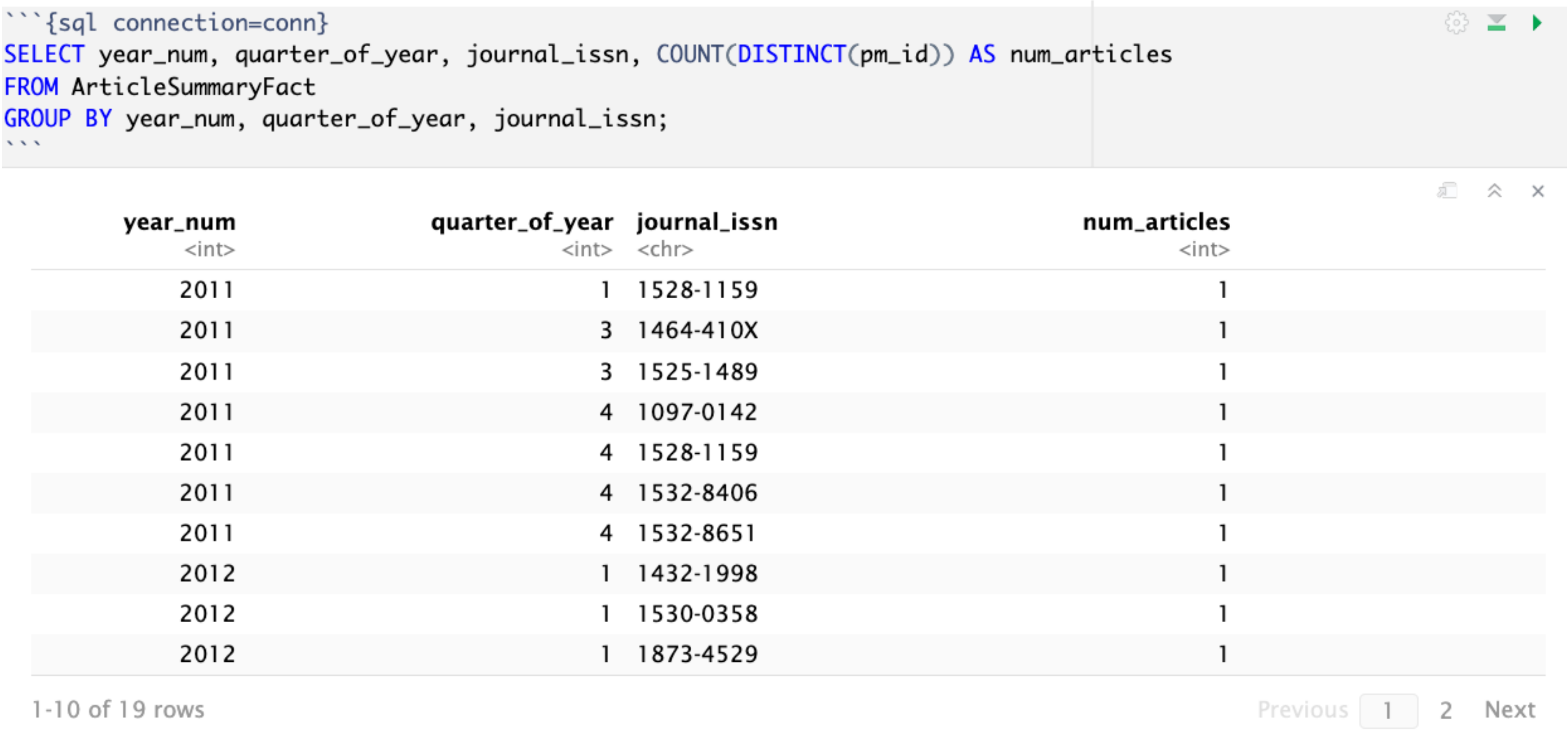


~~~
```{sql connection=conn}
SELECT year_num, quarter_of_year, author_id, COUNT(DISTINCT(pm_id)) AS num_articles
FROM ArticleSummaryFact
GROUP BY year_num, quarter_of_year, author_id;
```
~~~

year_num	quarter_of_year	author_id	num_articles
<int>	<int>	<int>	<int>
2011	3	80	1
2011	3	81	1
2011	4	3	4
2011	4	4	3
2011	4	6	1
2011	4	7	3
2011	4	31	3
2011	4	58	1
2011	4	60	1
2011	4	61	1

21-30 of 111 rows

Part 2.2 number of articles per time period by journal

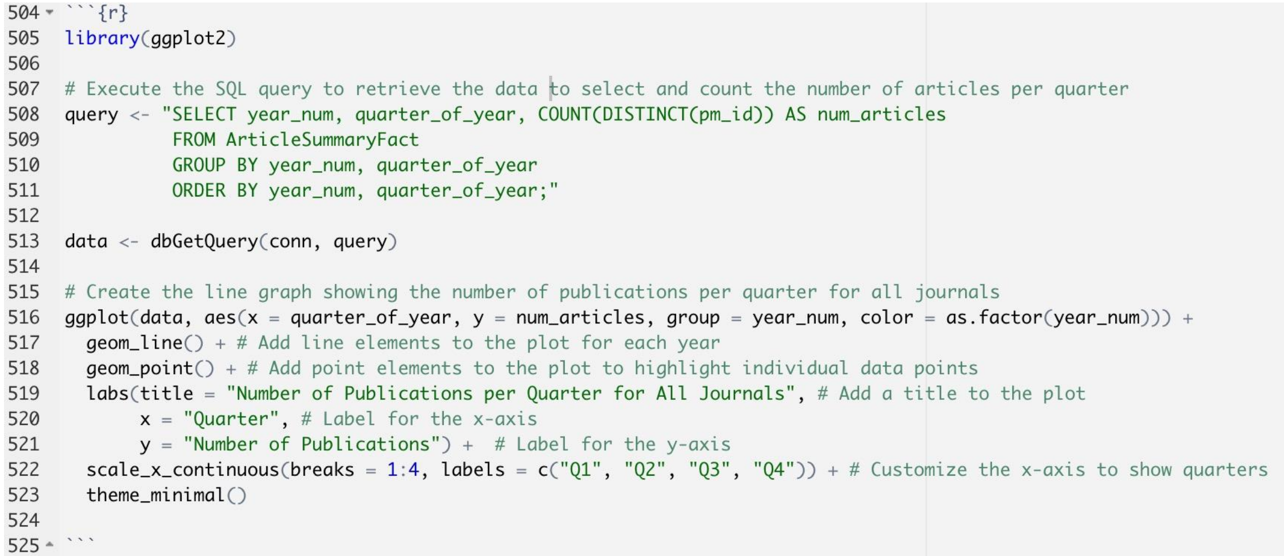


~~~
```{sql connection=conn}
SELECT year_num, quarter_of_year, journal_issn, COUNT(DISTINCT(pm_id)) AS num_articles
FROM ArticleSummaryFact
GROUP BY year_num, quarter_of_year, journal_issn;
```
~~~

year_num	quarter_of_year	journal_issn	num_articles
<int>	<int>	<chr>	<int>
2011	1	1528-1159	1
2011	3	1464-410X	1
2011	3	1525-1489	1
2011	4	1097-0142	1
2011	4	1528-1159	1
2011	4	1532-8406	1
2011	4	1532-8651	1
2012	1	1432-1998	1
2012	1	1530-0358	1
2012	1	1873-4529	1

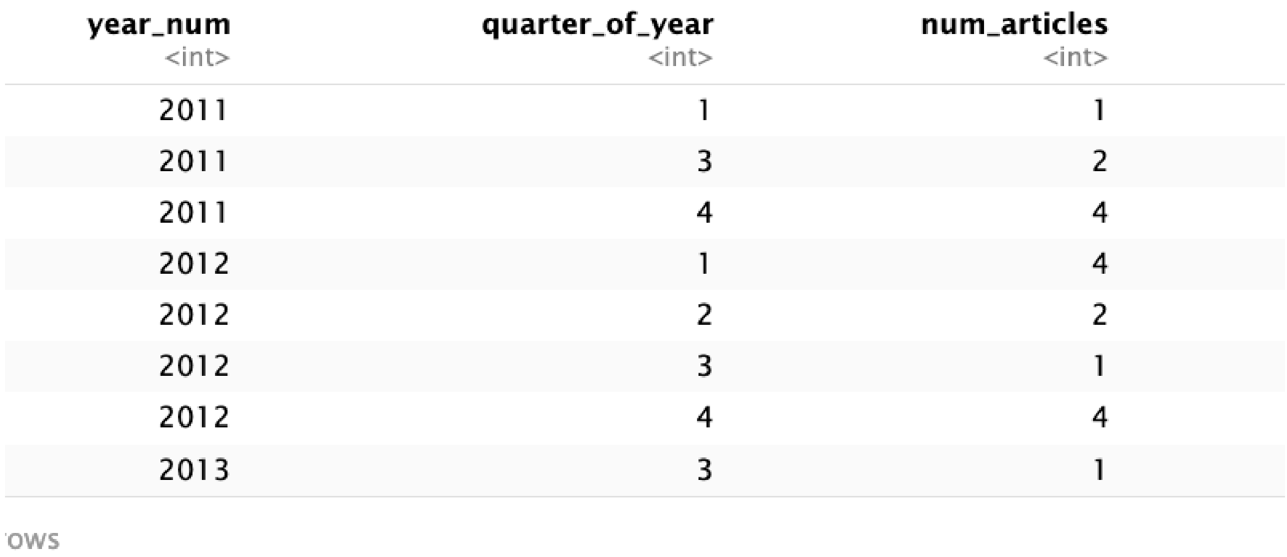
1-10 of 19 rows

Part 3-Explore and Mine Data



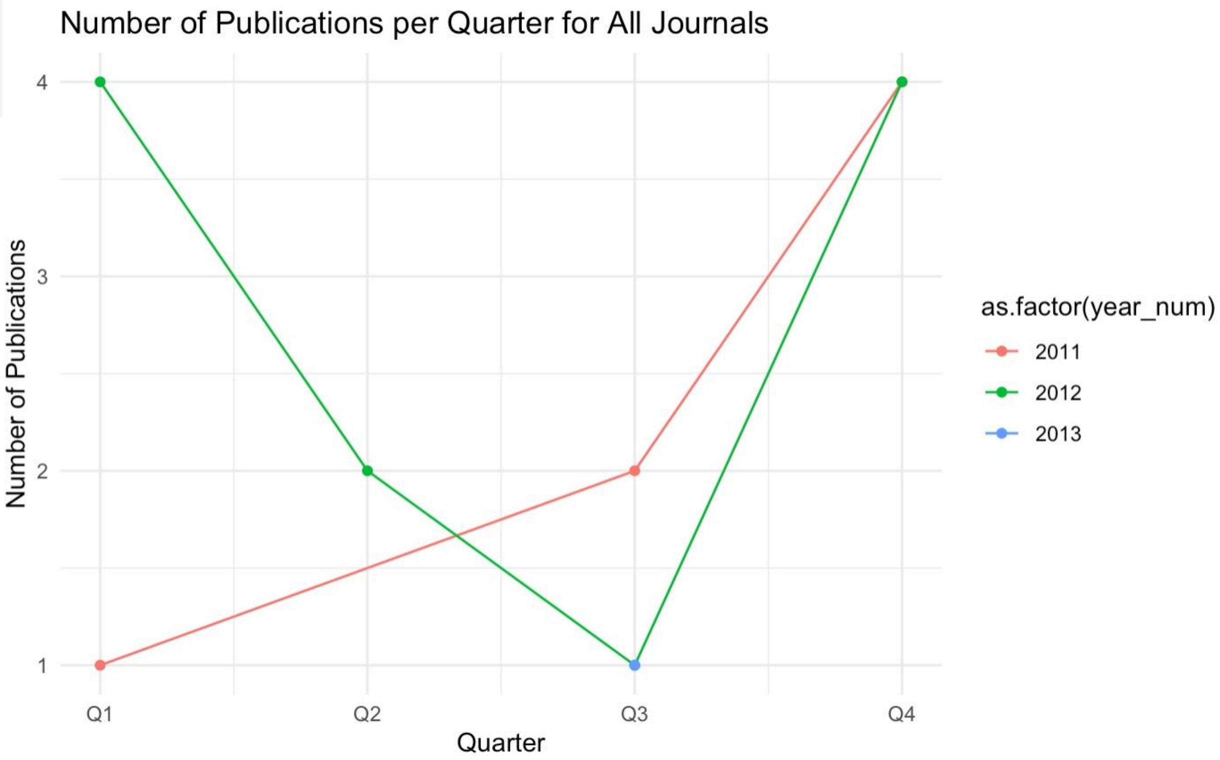
~~~
```{r}
library(ggplot2)

# Execute the SQL query to retrieve the data to select and count the number of articles per quarter
query <- "SELECT year_num, quarter_of_year, COUNT(DISTINCT(pm_id)) AS num_articles
          FROM ArticleSummaryFact
          GROUP BY year_num, quarter_of_year
          ORDER BY year_num, quarter_of_year;"

data <- dbGetQuery(conn, query)

# Create the line graph showing the number of publications per quarter for all journals
ggplot(data, aes(x = quarter_of_year, y = num_articles, group = year_num, color = as.factor(year_num))) +
  geom_line() + # Add line elements to the plot for each year
  geom_point() + # Add point elements to the plot to highlight individual data points
  labs(title = "Number of Publications per Quarter for All Journals", # Add a title to the plot
       x = "Quarter", # Label for the x-axis
       y = "Number of Publications") +  # Label for the y-axis
  scale_x_continuous(breaks = 1:4, labels = c("Q1", "Q2", "Q3", "Q4")) + # Customize the x-axis to show quarters
  theme_minimal()

```
~~~

A line graph that shows the number of publications for all journals each quarter

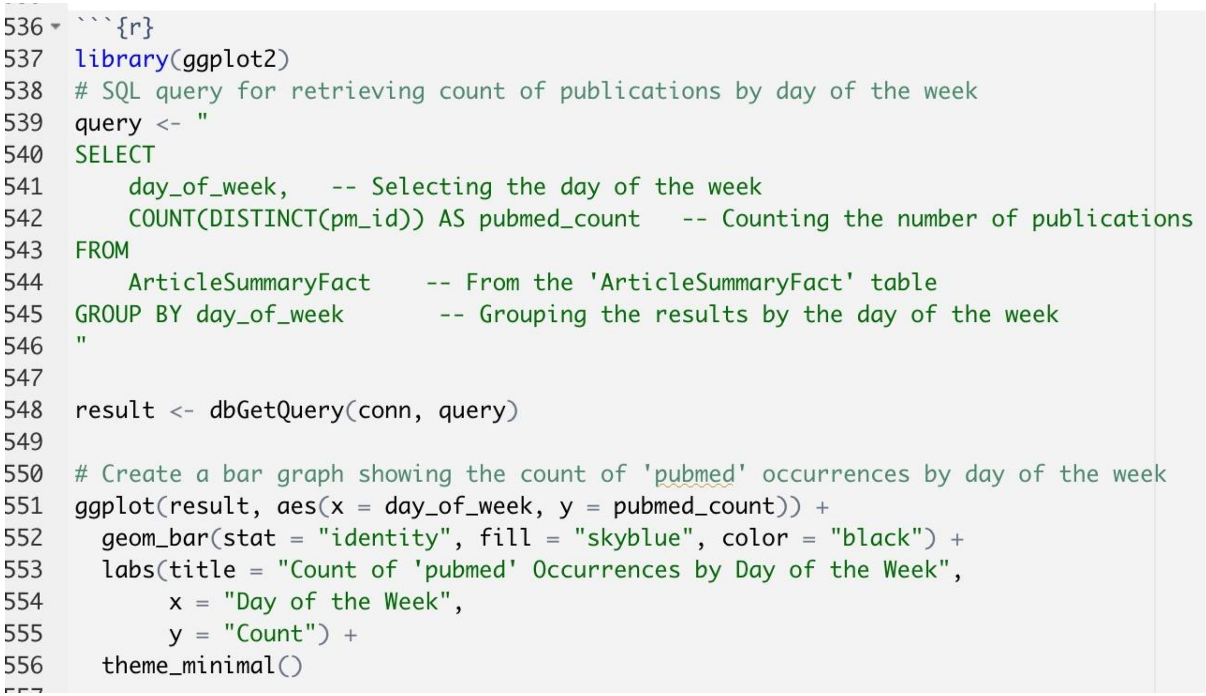


year_num <int>	quarter_of_year <int>	num_articles <int>
2011	1	1
2011	3	2
2011	4	4
2012	1	4
2012	2	2
2012	3	1
2012	4	4
2013	3	1

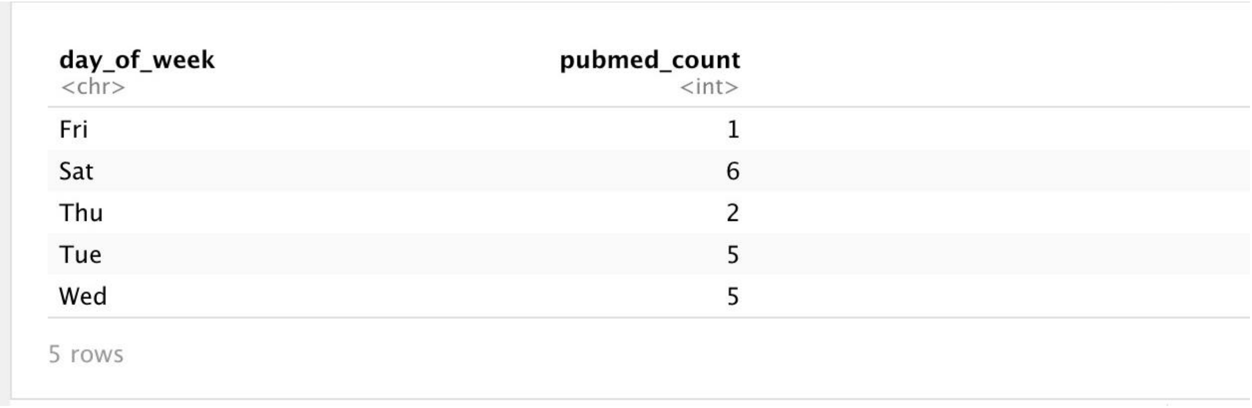
ows



Part 3-Explore and Mine Data



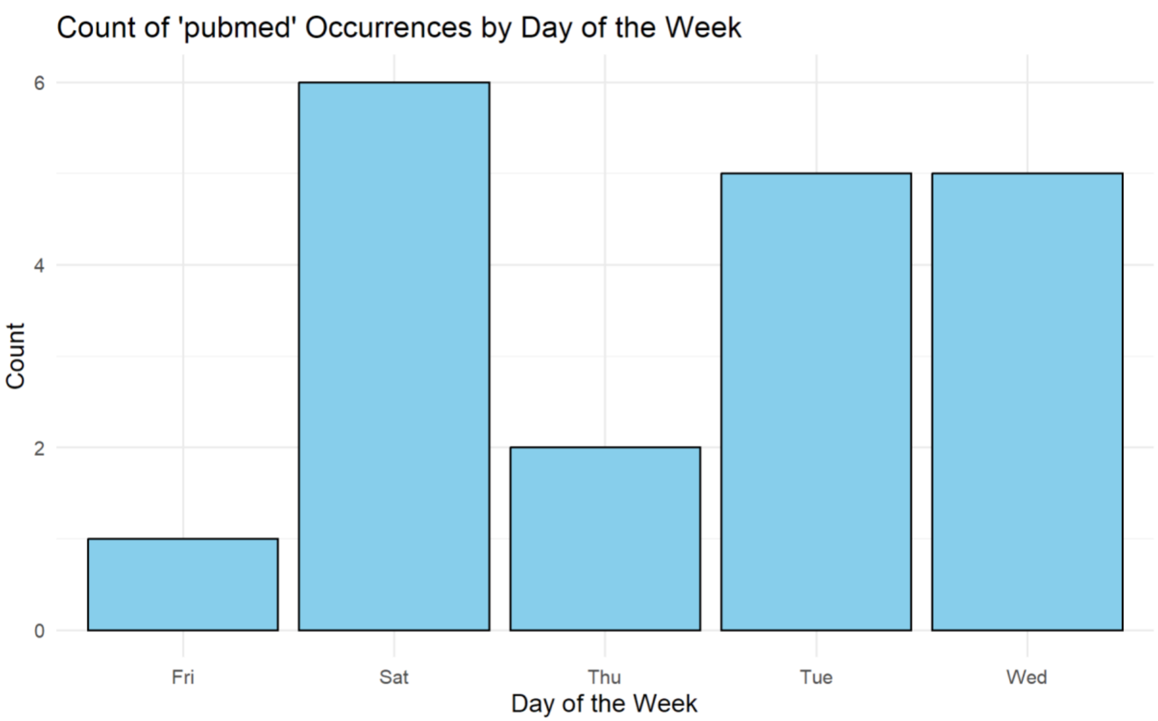
~~~
```{r}
library(ggplot2)
# SQL query for retrieving count of publications by day of the week
query <- "
SELECT
    day_of_week,    -- Selecting the day of the week
    COUNT(DISTINCT(pm_id)) AS pubmed_count   -- Counting the number of publications
FROM
    ArticleSummaryFact    -- From the 'ArticleSummaryFact' table
GROUP BY day_of_week       -- Grouping the results by the day of the week
"

result <- dbGetQuery(conn, query)

# Create a bar graph showing the count of 'pubmed' occurrences by day of the week
ggplot(result, aes(x = day_of_week, y = pubmed_count)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "black") +
  labs(title = "Count of 'pubmed' Occurrences by Day of the Week",
       x = "Day of the Week",
       y = "Count") +
  theme_minimal()
~~~



day_of_week <chr>	pubmed_count <int>
Fri	1
Sat	6
Thu	2
Tue	5
Wed	5

5 rows

Count of 'pubmed' Occurrences by Day of the Week



Thank You