Practicum_Two

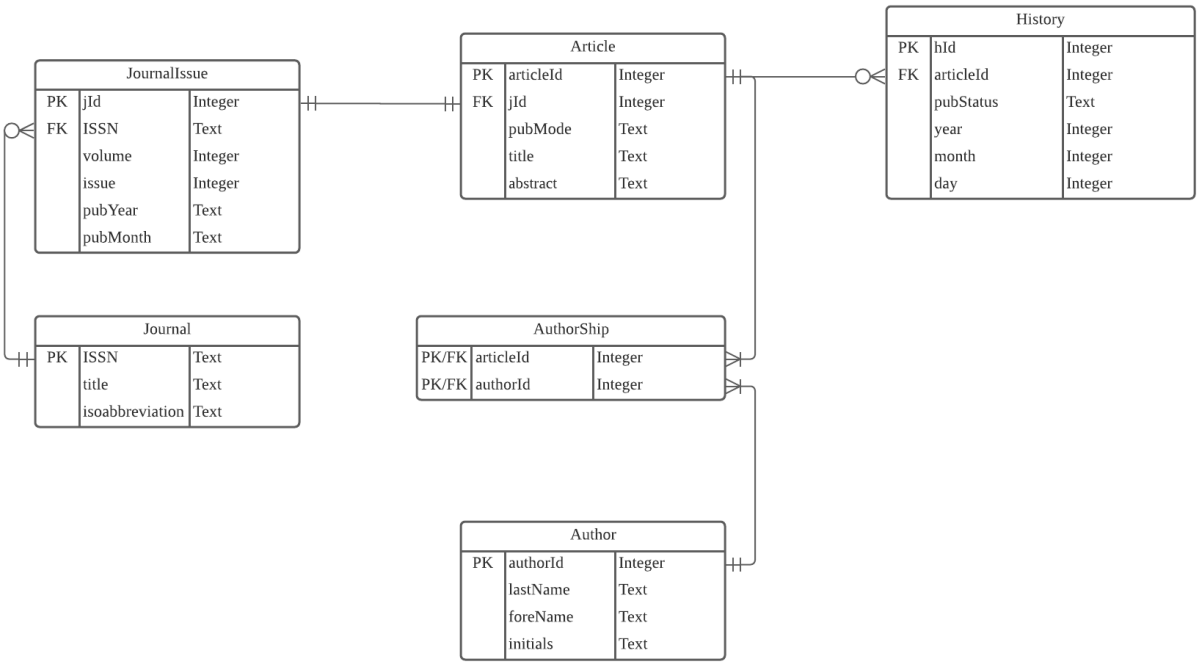
Introduction

Group member

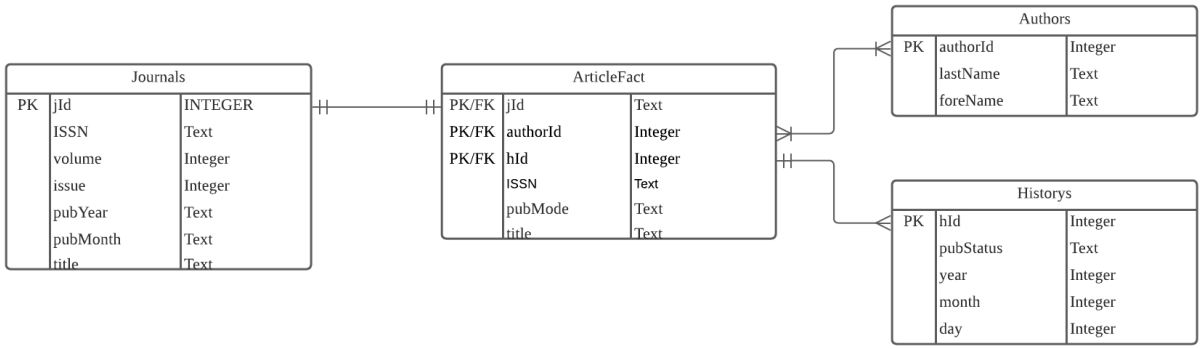
Bingyan Li, li.bingy@northeastern.edu
Pu Liu, liu.pu1@northeastern.edu
Sicong Ye ye.sic@northeastern.edu
Yuecheng Shi, shi.yuec@northeastern.edu

Schema

Part 1



Part 2



Part1 - Normalized database from XML

Connect to database

```
library(RSQLite)

dbFile = "./Practicum2.sqlite"

dbcon <- dbConnect(RSQLite::SQLite(), dbFile)
```

Create table Author

```
DROP TABLE IF EXISTS Author;
```

```
CREATE TABLE Author (
  authorId INTEGER PRIMARY KEY,
  lastName TEXT,
  foreName TEXT,
  initials TEXT
);
```

Create table Journal

```
DROP TABLE IF EXISTS Journal;
```

```
CREATE TABLE Journal (
  ISSN TEXT PRIMARY KEY,
  title TEXT NOT NULL,
  isoabbreviation TEXT NOT NULL
);
```

Create table JournalIssue

```
DROP TABLE IF EXISTS JournalIssue;
```

```
CREATE TABLE JournalIssue (
  jId INTEGER PRIMARY KEY,
  ISSN INTEGER REFERENCES Journal(ISSN) ON DELETE CASCADE,
```

```
  volume INTEGER NOT NULL,
  issue INTEGER NOT NULL,
  pubYear TEXT NOT NULL,
  pubMonth TEXT NOT NULL
);
```

Create table AuthorShip

```
DROP TABLE IF EXISTS AuthorShip;
```

```
CREATE TABLE AuthorShip (
  articleId INTEGER,
  authorId INTEGER,
  PRIMARY KEY (articleId, authorId)
)
```

Create table Article

```
DROP TABLE IF EXISTS Article;
```

```
CREATE TABLE Article (
  articleId INTEGER PRIMARY KEY,
  jId INTEGER REFERENCES JournalIssue(jId) ON DELETE CASCADE,
  pubMode TEXT NOT NULL,
  title TEXT NOT NULL,
  abstract TEXT NOT NULL
)
```

Create table History

```
DROP TABLE IF EXISTS History;
```

```
CREATE TABLE History (
  hId INTEGER PRIMARY KEY,
  articleId INTEGER REFERENCES Article(articleId) ON DELETE CASCADE,
  pubStatus TEXT NOT NULL,
  year INTEGER NOT NULL,
  month INTEGER NOT NULL,
  day INTEGER NOT NULL
)
```

Loading Libraries

```
library(XML)
library(xslt)
```

```
## Loading required package: xml2
```

Load XML data to database

Read XML file

```
xmlPath <- "./pubmed_sample.xml"
xmlDoc = read_xml(xmlPath)
```

Load Journal table

```
# Extract Journal Data
xslPath = "./journal.xsl"
xslDoc = read_xml(xslPath,package="xslt")
journal_raw = xml_xslt(xmlDoc,xslDoc)
journals = unique(xmlToDataFrame(XML::xmlParse(journal_raw)))
dbWriteTable(dbcon,name = "Journal", value = journals, append = TRUE)
```

Load Author table

```
# Extract Author Data
xslPath = "./author.xsl"
xslDoc = read_xml(xslPath,package="xslt")
author_raw = xml_xslt(xmlDoc,xslDoc)
authors = unique(xmlToDataFrame(XML::xmlParse(author_raw)))
authors = cbind(authorId=rownames(authors),authors)
dbWriteTable(dbcon,name = "Author", value = authors, append = TRUE)
```

```
SELECT * FROM Author;
```

Table 1: Displaying records 1 - 10

authorId	lastName	foreName	initials
1	Kuo	Cassie	C
2	Edwards	Alison	A
3	Mazumdar	Madhu	M
4	Memtsoudis	Stavros G	SG
5	Stundner	Ottokar	O
6	Kirksey	Meghan	M
7	Chiu	Ya Lin	YL
9	Poultsides	Lazaros	L
10	Gerner	Peter	P
12	Gupta	Ajay	A

Load JournalIssue table

```
# Extract JournalIssue Data
xslPath = "./journalIssue.xsl"
xslDoc = read_xml(xslPath,package="xslt")
journalIssue_raw = xml_xslt(xmlDoc,xslDoc)
journalIssues = unique(xmlToDataFrame(XML::xmlParse(journalIssue_raw)))
journalIssues = cbind(jId=rownames(journalIssues),journalIssues)
medDate <- journalIssues$medlineDate
for (i in 1 : nrow(journalIssues)) {
  if (medDate[i] != '') {
  pubDate <- strsplit(medDate[i],split = ' ')
```

```
  journalIssues$pubYear[i] = pubDate[[1]][1]
  journalIssues$pubMonth[i] = pubDate[[1]][2]
  }
}
journalIssues <- subset(journalIssues, select = -c(7))
dbWriteTable(dbcon,name = "JournalIssue", value = journalIssues, append = TRUE)
```

```
SELECT * FROM JournalIssue;
```

Table 2: Displaying records 1 - 10

jId	ISSN	volume	issue	pubYear	pubMonth
1	1556-3316	8	2	2012	Jul
2	1545-7206	54	2	2013	Mar-Apr
3	1524-4628	43	11	2012	Nov
4	1532-8651	37	6	2012	Nov-Dec
5	1532-2688	22	1	2013	Jan
6	1528-1132	471	1	2013	Jan
7	1532-8406	27	10	2012	Dec
8	1528-1175	117	1	2012	Jul
9	1432-1998	42	8	2012	Aug
10	1530-0358	55	4	2012	Apr

Load Article table

```
# Extract Article Data
xslPath = "./Article.xsl"
xslDoc = read_xml(xslPath,package="xslt")
article_raw = xml_xslt(xmlDoc,xslDoc)
article = unique(xmlToDataFrame(XML::xmlParse(article_raw)))
rownames(article) <- NULL
jIds = c()
for(i in 1:nrow(article)){
  temp = article[i,]
  jIds = append(jIds,journalIssues[which(temp$ISSN==journalIssues$ISSN
                        & temp$volume==journalIssues$volume
                        & temp$issue == journalIssues$issue),]$jId)
}
article = cbind(articleId=rownames(article),jId=jIds,article)
article = subset(article, select = 1:5)
dbWriteTable(dbcon,name = "Article", value = article, append = TRUE)
```

Load AuthorShip table

```
# Extract authorship Data
xslPath = "./AuthorShip.xsl"
xslDoc = read_xml(xslPath,package="xslt")
authorship_raw = xml_xslt(xmlDoc,xslDoc)
authorship = unique(xmlToDataFrame(XML::xmlParse(authorship_raw)))
rownames(authorship) <- NULL
authorIds = c()
articleIds = c()
```

```
for(i in 1:nrow(authorship)){
  temp = authorship[i,]
  articleIds = append(articleIds, article[which(temp$title==article$title),]$articleId)
  authorIds = append(authorIds,authors[which(temp$LastName==authors$lastName
                      & temp$ForeName==authors$foreName),]$authorId)
}
authorship = cbind(articleId=articleIds,authorId=authorIds,authorship)
authorship = subset(authorship, select = 1:2)
dbWriteTable(dbcon,name = "Authorship", value = authorship, append = TRUE)
```

```
SELECT * FROM AuthorShip;
```

Table 3: Displaying records 1 - 10

articleId	authorId
1	1
1	2
1	3
1	4
2	5
2	6
2	7
2	3
2	9
2	10

Load History table

```
# Extract History Data
xslPath = "./history.xsl"
xslDoc = read_xml(xslPath,package="xslt")
history_raw = xml_xslt(xmlDoc,xslDoc)
historys = unique(xmlToDataFrame(XML::xmlParse(history_raw)))
rownames(historys) <- NULL
articleIds = c()
for(i in 1:nrow(historys)){
  temp = historys[i,]
  articleIds = append(articleIds,article[which(temp$title==article$title),]$articleId)
}
historys = cbind(hId=rownames(historys),articleId=articleIds,historys)
historys = subset(historys, select = 1:6)
dbWriteTable(dbcon,name = "History", value = historys, append = TRUE)
```

```
SELECT * FROM History;
```

Table 4: Displaying records 1 - 10

hId	articleId	pubStatus	year	month	day
1	1	received	2012	1	15
2	1	accepted	2012	4	16
3	1	epublish	2012	6	20
4	1	entrez	2013	7	23

hId	articleId	pubStatus	year	month	day
5	1	pubmed	2013	7	23
6	1	medline	2013	7	23
7	2	received	2012	7	16
8	2	revised	2012	8	17
9	2	accepted	2012	8	20
10	2	aheadofprint	2012	11	27

Part2 - Data warehouse

```
sqlCmd <- "SELECT ji.jId AS jId, j.ISSN AS ISSN, j.title AS title,
              ji.volume AS volume, ji.issue AS issue,
              ji.pubYear As pubYear, ji.pubMonth As pubMonth
            FROM Journal AS j
            JOIN JournalIssue AS ji ON j.ISSN = ji.ISSN"
journal = dbGetQuery(dbcon, sqlCmd)
```

```
sqlCmd <- "SELECT authorID, lastName, foreName
            FROM Author"
author = dbGetQuery(dbcon, sqlCmd)
```

```
sqlCmd <- "SELECT hId, pubStatus, year, month, day
            FROM History"
history = dbGetQuery(dbcon, sqlCmd)
```

```
sqlCmd <- "SELECT JournalIssue.jId AS jId, AuthorShip.authorId AS authorId,
            History.hId AS hId, JournalIssue.ISSN AS ISSN,
            Article.pubMode AS pubMode, Article.title AS title
          FROM Article
          JOIN AuthorShip ON Article.articleId = AuthorShip.articleId
          JOIN JournalIssue ON Article.jId = JournalIssue.jId
          JOIN History ON Article.articleId = History.articleId"
articleFact = dbGetQuery(dbcon,sqlCmd)
```

```
library(RSQLite)
library(ggplot2)

dbFile = "./Practicum2_Part2.sqlite"

dbcon1 <- dbConnect(RSQLite::SQLite(), dbFile)
```

Create table Journals

```
DROP TABLE IF EXISTS Journals;
```

```
CREATE TABLE Journals (
  jId INTEGER PRIMARY KEY,
  ISSN TEXT NOT NULL,
  volume INTEGER NOT NULL,
  issue INTEGER NOT NULL,
  pubYear TEXT NOT NULL,
  pubMonth TEXT NOT NULL,
  title TEXT NOT NULL
```

```
);
```

Create table Author

```
DROP TABLE IF EXISTS Authors;
```

```
CREATE TABLE Authors (
  authorId INTEGER PRIMARY KEY,
  lastName TEXT,
  foreName TEXT
);
```

Create table History

```
DROP TABLE IF EXISTS Historys;
```

```
CREATE TABLE Historys (
  hId INTEGER PRIMARY KEY,
  pubStatus TEXT NOT NULL,
  year INTEGER NOT NULL,
  month INTEGER NOT NULL,
  day INTEGER NOT NULL
)
```

Create table ArticleFact

```
DROP TABLE IF EXISTS ArticleFact;
```

```
CREATE TABLE ArticleFact (
  jId TEXT NOT NULL,
  authorId INTEGER NOT NULL,
  hId INTEGER NOT NULL,
  pubMode TEXT,
  title TEXT,
  PRIMARY KEY (jId, authorId, hId),
  FOREIGN KEY (jId) REFERENCES Journals (jId),
  FOREIGN KEY (authorId) REFERENCES Authors (authorId),
  FOREIGN KEY (hId) REFERENCES Historys (hId)
);
```

```
dbWriteTable(dbcon1, "Journals", journal, overwrite = T, row.names = F)
dbWriteTable(dbcon1, "Authors", author, overwrite = T, row.names = F)
dbWriteTable(dbcon1, "Historys", history, overwrite = T, row.names = F)
dbWriteTable(dbcon1, "ArticleFact", articleFact, overwrite = T, row.names = F)
```

Examine Database

```
SELECT * FROM Journals
```

Table 5: Displaying records 1 - 10

jId	ISSN	title	volume	issue	pubYear	pubMonth
1	1556-3316	HSS journal : the musculoskeletal journal of Hospital for Special Surgery	8	2	2012	Jul
2	1545-7206	Psychosomatics	54	2	2013	Mar-Apr
3	1524-4628	Stroke; a journal of cerebral circulation	43	11	2012	Nov
13	1532-8651	Regional anesthesia and pain medicine	37	1	2012	Jan-Feb
4	1532-8651	Regional anesthesia and pain medicine	37	6	2012	Nov-Dec
5	1532-2688	Seizure : the journal of the British Epilepsy Association	22	1	2013	Jan
6	1528-1132	Clinical orthopaedics and related research	471	1	2013	Jan
14	1532-8406	The Journal of arthroplasty	27	6	2012	Jun
7	1532-8406	The Journal of arthroplasty	27	10	2012	Dec
8	1528-1175	Anesthesiology	117	1	2012	Jul

```
SELECT * FROM Authors
```

Table 6: Displaying records 1 - 10

authorId	lastName	foreName
1	Kuo	Cassie
2	Edwards	Alison
3	Mazumdar	Madhu
4	Memtsoudis	Stavros G
5	Stundner	Ottokar
6	Kirksey	Meghan
7	Chiu	Ya Lin
9	Poultsides	Lazaros
10	Gerner	Peter
12	Gupta	Ajay

```
SELECT * FROM Historys
```

Table 7: Displaying records 1 - 10

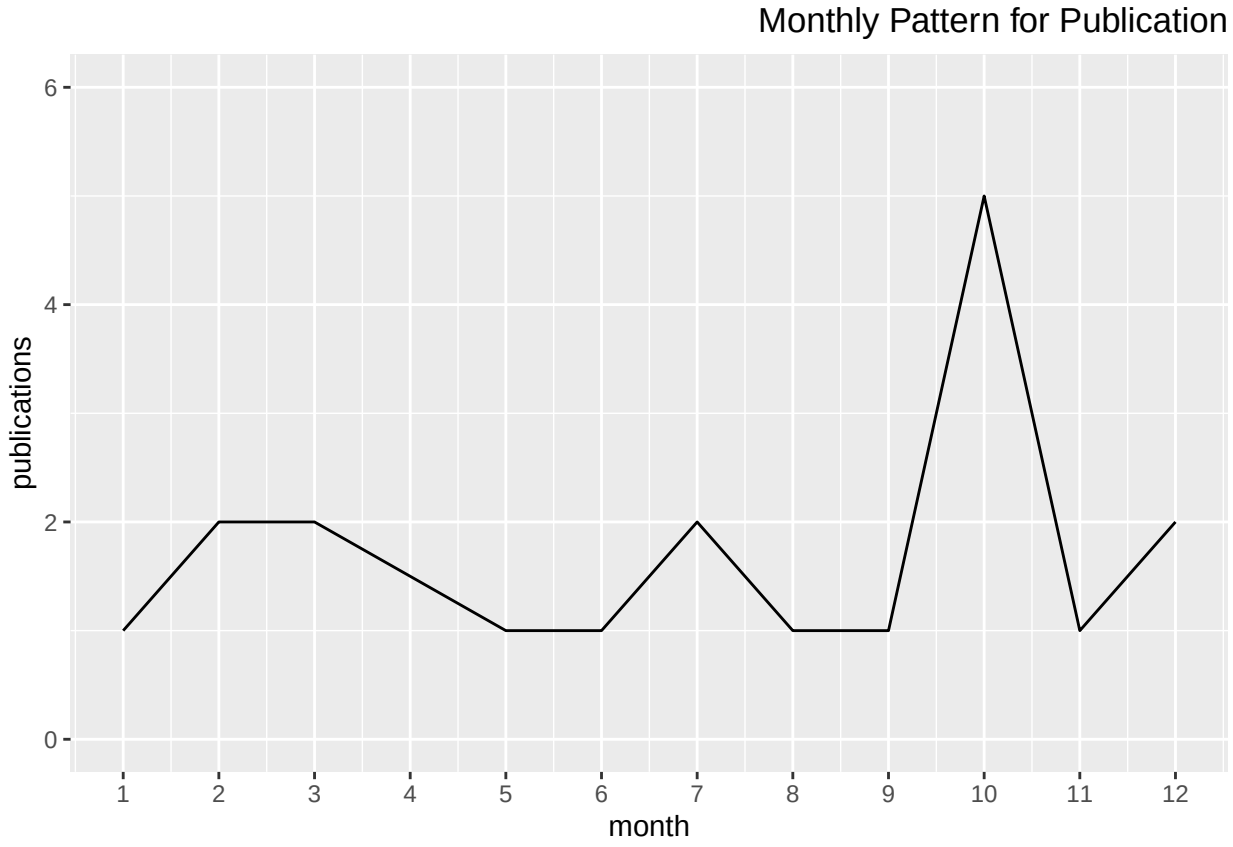
hId	pubStatus	year	month	day
1	received	2012	1	15
2	accepted	2012	4	16
3	epublish	2012	6	20
4	entrez	2013	7	23
5	pubmed	2013	7	23
6	medline	2013	7	23

hId	pubStatus	year	month	day
7	received	2012	7	16
8	revised	2012	8	17
9	accepted	2012	8	20
10	aheadofprint	2012	11	27

Part3 - Data mining

Monthly publication

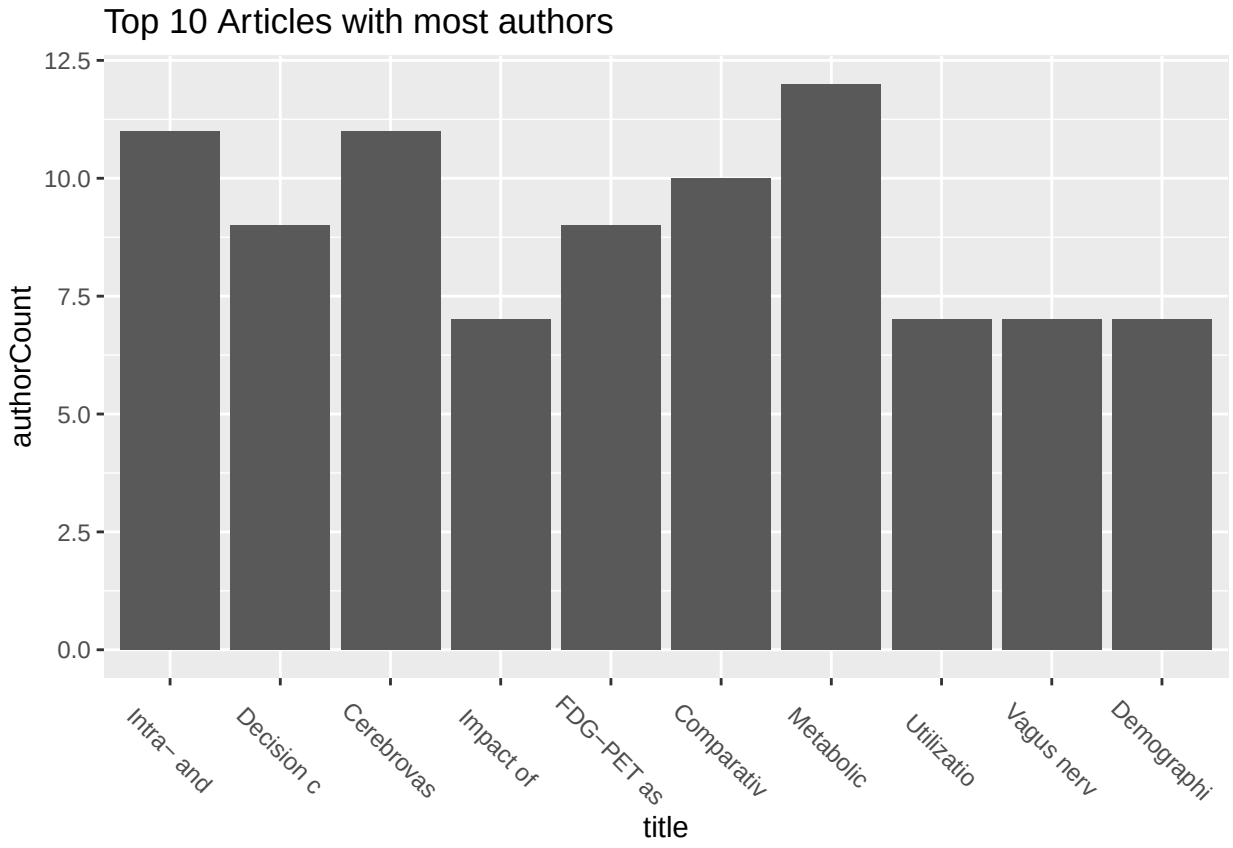
```
patternData <- dbGetQuery(dbcon, "SELECT month, count(1) as hid from History
                          WHERE pubstatus = 'pubmed' GROUP BY month
                          ORDER BY month DESC")
month <- patternData[,1]
publications <- patternData[,2]

ggplot(patternData, aes(x=month, y=publications)) + geom_line()  + ylim(0,6) + labs(title="Monthly Patt
theme(plot.title = element_text(hjust = 1)) +  scale_x_continuous(breaks=c(1,2,3,4,5,6,7,8,9,10,11,12))
```



Top 10 Articles with most authors

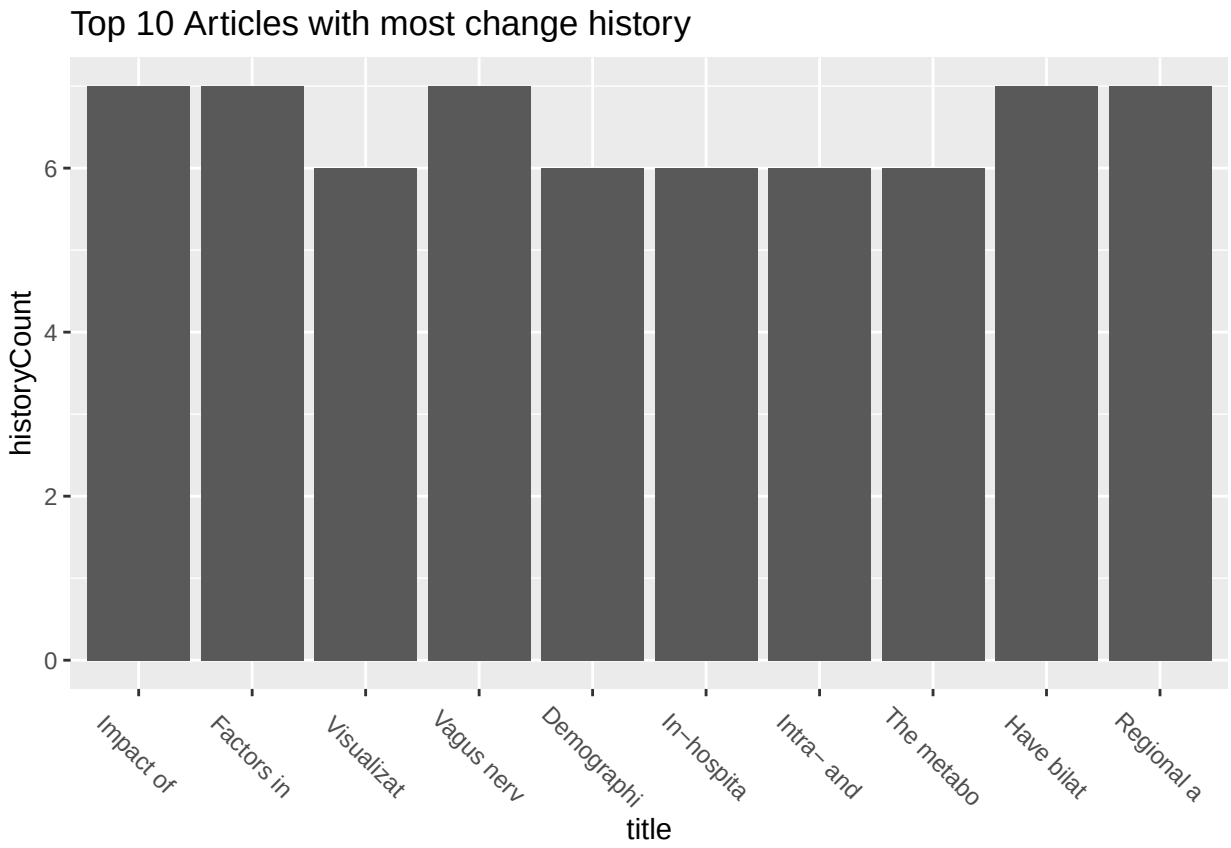
```
authorCntData = dbGetQuery(dbcon1,"SELECT COUNT(DISTINCT(authorId)) AS authorCount, title
                          FROM ArticleFact GROUP BY jId
                          ORDER BY authorCount DESC LIMIT 10;")
ggplot(authorCntData, aes(x = title, y = authorCount)) +
  geom_bar(stat = 'identity') +
```

```
  theme(axis.text.x = element_text(angle = -45, hjust = 0.5, vjust = 0.5)) +
  scale_x_discrete(labels=substr(authorCntData$title,0,10)) +
  labels(title="Top 10 Articles with most authors")
```



Top 10 Articles with most change history

```
historyCntData = dbGetQuery(dbcon1,"SELECT COUNT(DISTINCT(hId)) AS historyCount, title
                            FROM ArticleFact GROUP BY jId
                            ORDER BY historyCount DESC LIMIT 10;")
ggplot(historyCntData, aes(x = title, y = historyCount)) +
  geom_bar(stat = 'identity') +
  theme(axis.text.x = element_text(angle = -45, hjust = 0.5, vjust = 0.5)) +
  scale_x_discrete(labels=substr(historyCntData$title,0,10)) +
  labels(title="Top 10 Articles with most change history")
```




```
dbDisconnect(dbcon)
dbDisconnect(dbcon1)
```