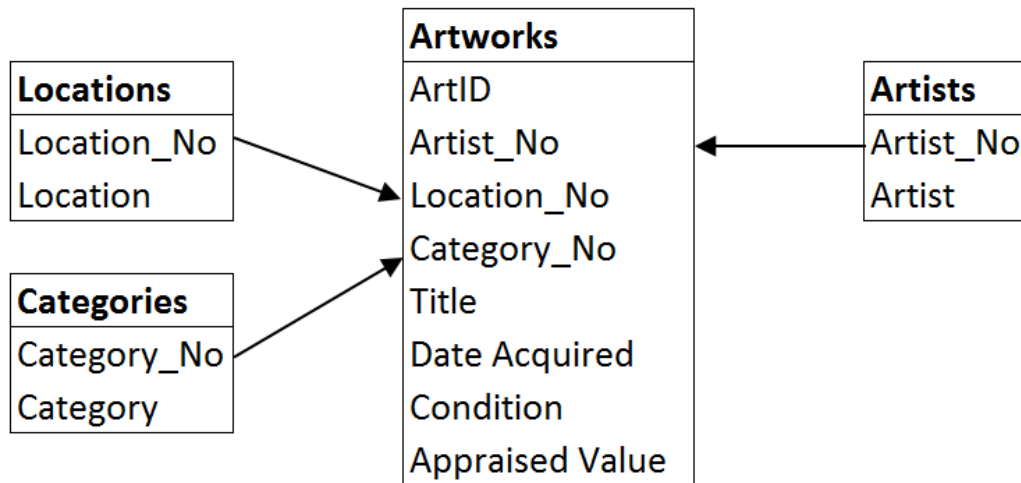




Case Study 2
Simple ETL Using R

1. Introduction

You are to build a simple ETL process for the Art Database. The output of this process is a MS Excel Spreadsheet that will be used for additional analysis in Case 3. The Art Database is composed of the following tables and relationships.



We would like a report that lists down all artworks as well as their corresponding location, category and artist.



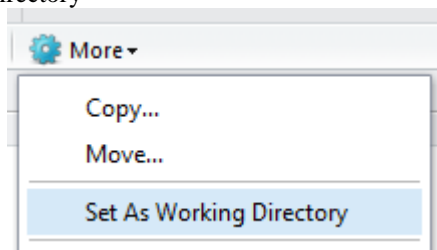
2. Preliminaries

We first download and copy the Art SQLite Database file into your My Documents folder.

2.1. Setting up R Studio

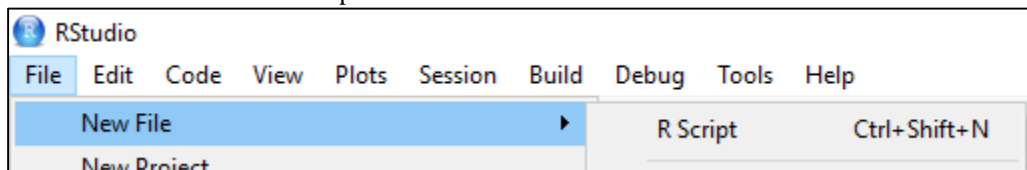
- 2.1.1. Open your My Documents folder and create a New Folder named “BIModule1 Cases”
- 2.1.2. Copy the Art.sqlite database and paste in inside your “My Documents\BIModule1 Cases” folder.
- 2.1.3. Open R Studio from your Programs Menu

- 2.1.4. In the middle right corner of the R Studio window click on the three dots “...”
- 2.1.5. Search for My Documents then Select the BIModule1 Cases Folder. Click on Ok.
- 2.1.6. Click on More -> Set Working Directory



2.2. Extracting Data


- 2.2.1. Click on File->New File – R Script >

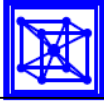


- 2.2.2. Type the following lines of code in the new script window:

```
> library(sqldf)
> db <- dbConnect(SQLite(), dbname="Art.sqlite")
> rs = dbSendQuery(db, "SELECT * FROM artworks")
> artworks = fetch(rs, n=-1)
> rs = dbSendQuery(db, "SELECT * FROM locations")
> locations = fetch(rs, n=-1)
> rs = dbSendQuery(db, "SELECT * FROM categories")
> categories = fetch(rs, n=-1)
> rs = dbSendQuery(db, "SELECT * FROM artists")
> artists = fetch(rs, n=-1)
> dbDisconnect(db)
```



- 2.2.3. Highlight all lines of code from then Click on .
- 2.2.4. As a Result, the Data was extracted from the Database and Displayed in the Environment Tab:



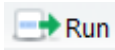
Environment		History
Import Dataset Clear Refresh		List
Global Environment		
Data		
artists	81 obs. of 2 variables	
artworks	115 obs. of 8 variables	
categories	4 obs. of 2 variables	
location	5 obs. of 2 variables	
Values		
con	Class 'RODBC' atomic [1:1] 10	
connectionstring	"Driver={Microsoft Access Driver (*.mdb, *..."	
filedirectory	"C:/Users/Eugene Rex/Documents/BIModule1 C..."	

2.3. Transforming Data

2.3.1. Type the following lines of code:

- #Transforming Data
- artworks = merge(x=artworks,y=artists,by=c("Artist_No"))
- artworks = merge(x=artworks,y=categories,by=c("Category_No"))
- artworks = merge(x=artworks,y=locations,by=c("Location_No"))
- artworks = artworks[,c("ArtID","Artist",
"Title","Date.Acquired","Category","Condition","Location","Appraised.V
alue")]
- artworks\$age <- as.numeric((Sys.Date() -
as.Date(artworks\$Date.Acquired, "%m/%d/%y"))/365)

2.3.2. Highlight all these lines of code from line, then Click on



2.3.3. As a Result, the artworks data was merged with the actual names of the Artists, Categories, and Locations. Additionally, an Age variable per Artwork was calculated based on the Current Date.

2.4. Loading Data

2.4.1. Type the following lines of code:

- #Loading Data
- write.csv(artworks, "artworks.csv")

2.4.2. Highlight all these lines of code from, then Click on



2.4.3. As a Result, the artworks data is saved as a CSV file.

2.4.4. To open the file, go to your "My Documents\BIModule1 Cases" folder. Open the artworks.csv file.

2.4.5. Resulting Dataset would look like this:

