# Introducing Very Large Data Sets into the Classroom

Dason Kurkiewicz

Department of Statistics, Iowa State University

March 9, 2012

# Outline

- Dason's Introduction
- The GUI
- The Usability Study
- Future Work
- Conclusions

# Outline

- Dason's Introduction
- The GUI
- The Usability Study
- Future Work
- Conclusions

# Outline

- Dason's Introduction
- The GUI
- The Usability Study
- Future Work
- Conclusions

# Outline

- Dason's Introduction
- The GUI
- The Usability Study
- Future Work
- Conclusions

# Outline

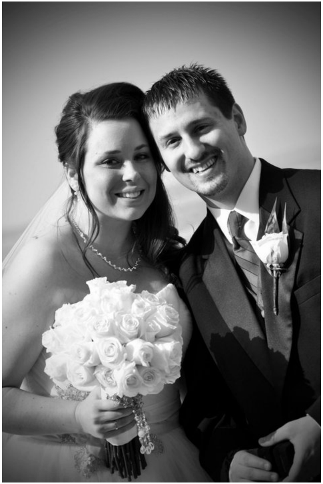
- Dason's Introduction
- The GUI
- The Usability Study
- Future Work
- Conclusions

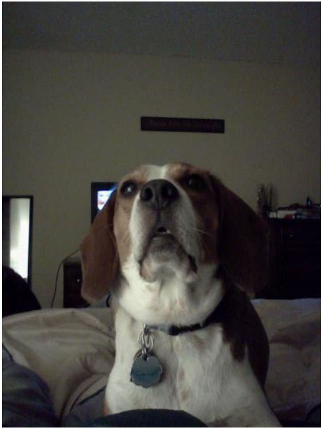
# Outline

- Dason's Introduction
- The GUI
- The Usability Study
- Future Work
- Conclusions

# Introduction

## Meet Dason





My Family

# dbConnectGUI

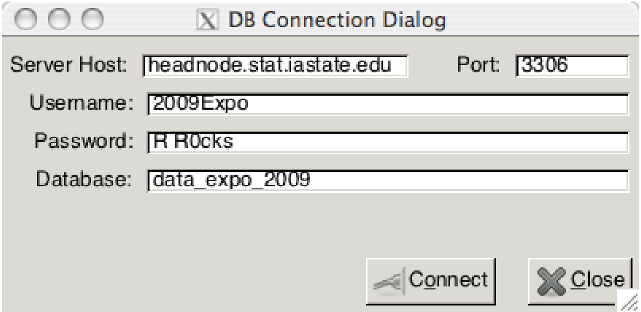
## Installation

```
# install.packages("gWidgetsRGtk2")
# install.packages("DBI")
# install.packages("RMySQL")

install.packages("dbConnectGUI")
library(dbConnectGUI)
# Try it out
dbConnectGUI()
# Or
connect <- getConnection()
dbConnectGUI(connect$con)
```

# dbConnectGUI

## Connection Dialog



Server Host: headnode.stat.iastate.edu

Port: 3306

Username: 2009Expo

Password: R R0cks

Database: data_expo_2009

# dbConnectGUI

## Data Viewer

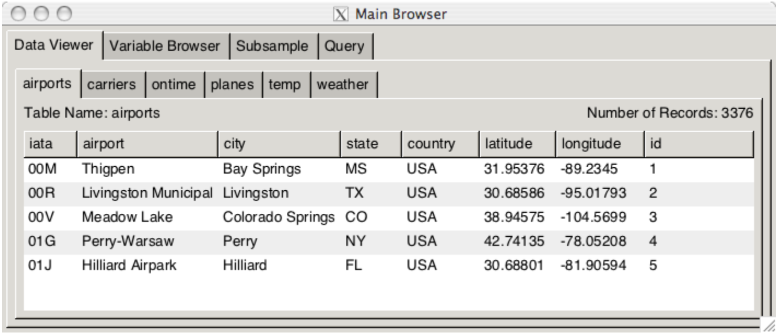


Table Name: airports

Number of Records: 3376

| iata | airport | city | state | country | latitude | longitude | id |
|---|---|---|---|---|---|---|---|
| 00M | Thigpen | Bay Springs | MS | USA | 31.95376 | -89.2345 | 1 |
| 00R | Livingston Municipal | Livingston | TX | USA | 30.68586 | -95.01793 | 2 |
| 00V | Meadow Lake | Colorado Springs | CO | USA | 38.94575 | -104.5699 | 3 |
| 01G | Perry-Warsaw | Perry | NY | USA | 42.74135 | -78.05208 | 4 |
| 01J | Hilliard Airpark | Hilliard | FL | USA | 30.68801 | -81.90594 | 5 |

# dbConnectGUI

## Variable Browser

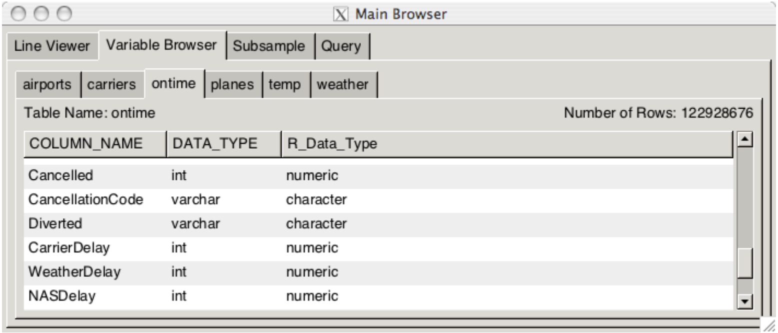


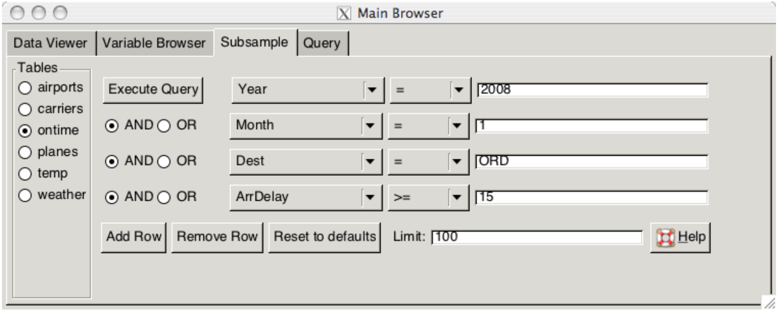
Table Name: ontime

Number of Rows: 122928676

| COLUMN_NAME | DATA_TYPE | R_Data_Type |
|---|---|---|
| Cancelled | int | numeric |
| CancellationCode | varchar | character |
| Diverted | varchar | character |
| CarrierDelay | int | numeric |
| WeatherDelay | int | numeric |
| NASDelay | int | numeric |

# dbConnectGUI

## Subsample



Tables
○ airports
○ carriers
● ontime
○ planes
○ temp
○ weather

Year = 2008

● AND ○ OR Month = 1

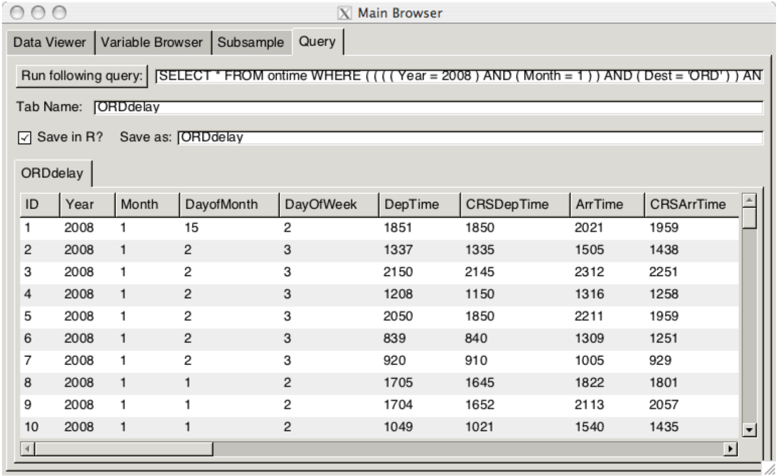
● AND ○ OR Dest = ORD

● AND ○ OR ArrDelay >= 15

Limit: 100

# dbConnectGUI

## Query



SELECT * FROM ontime WHERE ( ( ( ( Year = 2008 ) AND ( Month = 1 ) ) AND ( Dest = 'ORD' ) ) AN

Tab Name: ORDdelay

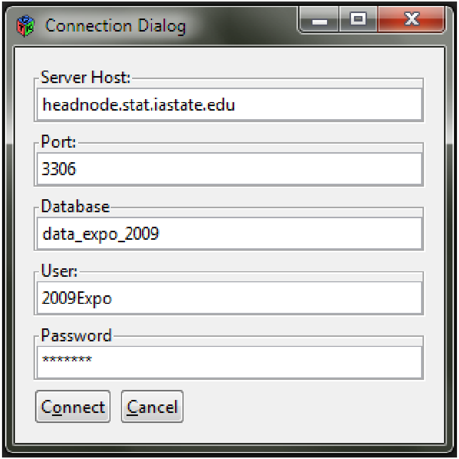
☑ Save in R? Save as: ORDdelay

ORDdelay

| ID | Year | Month | DayofMonth | DayOfWeek | DepTime | CRSDepTime | ArrTime | CRSArrTime |
|---|---|---|---|---|---|---|---|---|
| 1 | 2008 | 1 | 15 | 2 | 1851 | 1850 | 2021 | 1959 |
| 2 | 2008 | 1 | 2 | 3 | 1337 | 1335 | 1505 | 1438 |
| 3 | 2008 | 1 | 2 | 3 | 2150 | 2145 | 2312 | 2251 |
| 4 | 2008 | 1 | 2 | 3 | 1208 | 1150 | 1316 | 1258 |
| 5 | 2008 | 1 | 2 | 3 | 2050 | 1850 | 2211 | 1959 |
| 6 | 2008 | 1 | 2 | 3 | 839 | 840 | 1309 | 1251 |
| 7 | 2008 | 1 | 2 | 3 | 920 | 910 | 1005 | 929 |
| 8 | 2008 | 1 | 1 | 2 | 1705 | 1645 | 1822 | 1801 |
| 9 | 2008 | 1 | 1 | 2 | 1704 | 1652 | 2113 | 2057 |
| 10 | 2008 | 1 | 1 | 2 | 1049 | 1021 | 1540 | 1435 |

# dbConnectGUI

## Changes



Server Host: headnode.stat.iastate.edu

Port: 3306

Database data_expo_2009

User: 2009Expo

Password ********

# Learning Objectives

- Familiarizing students with databases and exploring the utility of databases for effcient data storage.
- Obtaining access to data stored in databases.
- Searching databases for specifc data records of interest as well as learning how to extract such data records for further statistical analysis in R or a different software package.
- Providing students with a first gentle and more paced introduction to the SQL language. The GUI can help facilitate the learning process and potentially improve student's attitude toward learning the database language SQL.

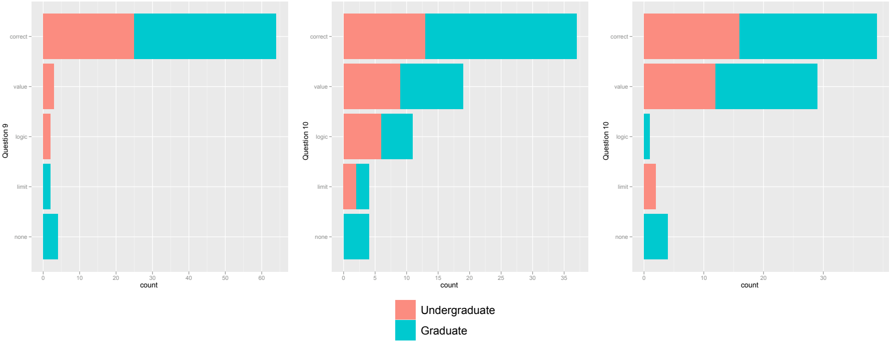
# Usability Study

## Student Overview

| | 400-level | | 500-level | |
|---|---|---|---|---|
| | Statistics | non-Statistics | Statistics | non-Statistics |
| Undergraduate | 23 | 6 | 1 | 0 |
| Graduate | 1 | 8 | 21 | 15 |

Table: Overview of students participating in the usability study by course, status, and area.

# Usability Study



# Usability Study

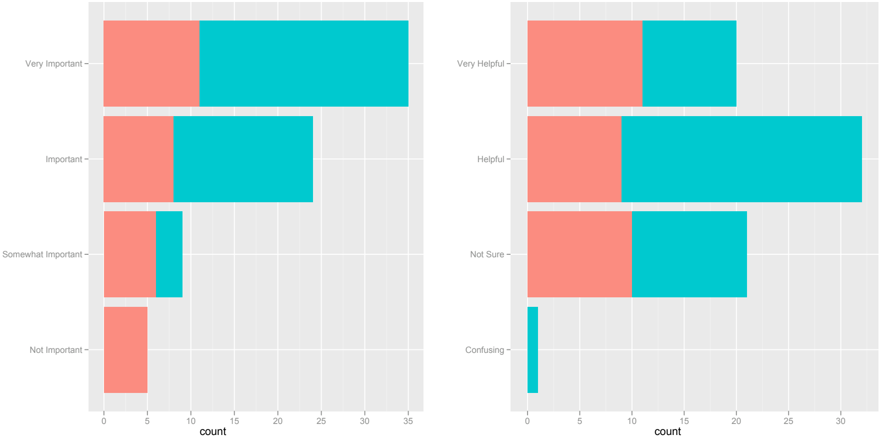


Figure: User feedback: perceived importance of task (left) and usefulness of the GUI to help with the task (right). Undergraduates are red and Graduates are Blue.

# Usability Study

## Students' opinions on task difficulty

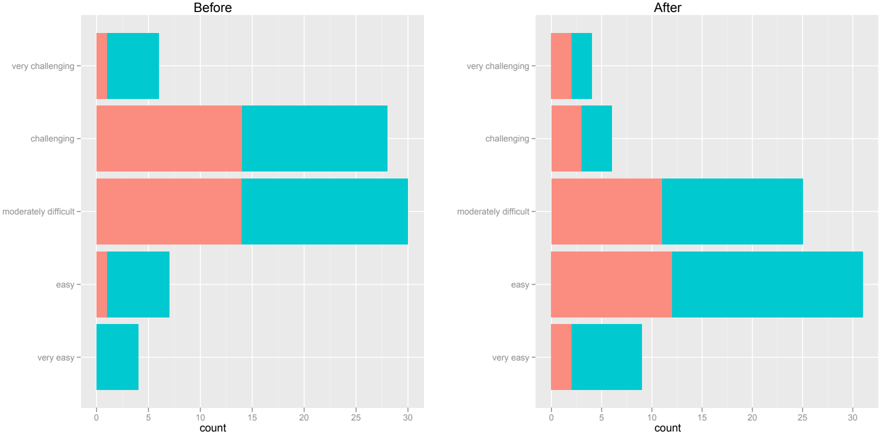
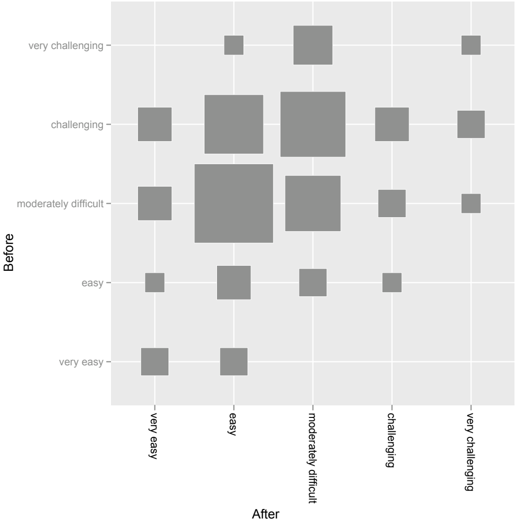


Figure: User feedback: anticipated degree of difficulty is greater than experienced difficulty of the task. Undergraduates are Red and Graduates are Blue.

# Paired Comparison of Before/After Difficulty



# Future Work

- Efficient Random Stratified Samples
- Joins
- Tutorial

# Future Work

- Efficient Random Stratified Samples
- Joins
- Tutorial

# Future Work

- Efficient Random Stratified Samples
- Joins
- Tutorial

# Future Work

- Efficient Random Stratified Samples
- Joins
- Tutorial

# Closing Remarks

- Big data is here to stay
- GUIs provide a level of comfort to some
- The usability study shows there is benefit to using the GUI
- Will keep working to make this better

# Closing Remarks

- Big data is here to stay
- GUIs provide a level of comfort to some
- The usability study shows there is benefit to using the GUI
- Will keep working to make this better

# Closing Remarks

- Big data is here to stay
- GUIs provide a level of comfort to some
- The usability study shows there is benefit to using the GUI
- Will keep working to make this better

# Closing Remarks

- Big data is here to stay
- GUIs provide a level of comfort to some
- The usability study shows there is benefit to using the GUI
- Will keep working to make this better

# Closing Remarks

- Big data is here to stay
- GUIs provide a level of comfort to some
- The usability study shows there is benefit to using the GUI
- Will keep working to make this better

# The End

Questions?