# E-Commerce Transaction Data:
## An Online Gift Shop Retailer Case

DATA607 Project 3 Presentation

Cassie Boylan, DH Kim, Alexis Mekueko

# Who Presents What

**DH Kim**: Data Acquisition from Web

- Motivation, Gift shop business, How to Import Excel Data from Web, Data Structure

Cassie Boylan: Revenue Analysis from Invoices Data

- An Online Gift Shop Retailer, Pre-processing Data, Top 10 and Bottom 10 Gift Items Sold, Results of Analysis

Alexis Mekueko: Data Management with Database

- Why Database?: Securing Data, Theoretical E-R Diagram, Normalizing: The Customer Table

Data Acquisition from Web

# Motivation

### Recent crisis of retailer stores

### Linking business-related activities to real-world data

*Virtually every aspect of business is now open to data collection and often instrumented for data collection: operations, manufacturing, **supply-chain management**, customer behaviour, market campaign performance, workflow procedures, and so on (page 1, Data Science for Business)*

# Gift Shop Business and Transaction Data

## An Online Retailer

- This is an UK-based online gift shop retailer selling gift goods to customers (mainly wholesalers) across countries.
- Main items sold include Assorted Color Bird Wind Ornaments, Pink Cheery Lights, Floral Elephant Soft Toy, and so on.

## E-Commerce Invoice Data

- Information on which items are sold, how many, and how much, who buys them, and when and what time are they ordered.
- Data covering from 12/1/2009 to 12/9/2011, which is stored in an Excel file with two separate sheets.

# Importing Excel Data from the Web

Source: The Website of UCI Machine Learning Repo

Packages needed:

```
library(readxl)
library(httr)
```

The GET() and read_excel() functions

```
retailURL <-
    "http://archive.ics.uci.edu//ml//machine-learning-databases//00502//online_retail_II.xlsx"

GET(retailURL, write_disk(tempFileName <- tempfile(fileext = ".xlsx")))

retail_sheet_2009 <- read_excel(tempFileName, sheet = "Year 2009-2010")
retail_sheet_2010 <- read_excel(tempFileName, sheet = "Year 2010-2011")
retaildf <- rbind(retail_sheet_2009, retail_sheet_2010)
```

# Description of Data

## Invoices Data

```
library(tidyverse)
glimpse(retaildf)
```

```
## Rows: 1,067,371
## Columns: 8
## $ Invoice       <chr> "489434", "489434", "489434", "489434", "489434", "48...
## $ StockCode     <chr> "85048", "79323P", "79323W", "22041", "21232", "22064...
## $ Description   <chr> "15CM CHRISTMAS GLASS BALL 20 LIGHTS", "PINK CHERRY L...
## $ Quantity      <dbl> 12, 12, 12, 48, 24, 24, 24, 10, 12, 12, 24, 12, 10, 1...
## $ InvoiceDate   <dttm> 2009-12-01 07:45:00, 2009-12-01 07:45:00, 2009-12-01...
## $ Price         <dbl> 6.95, 6.75, 6.75, 2.10, 1.25, 1.65, 1.25, 5.95, 2.55,...
## $ `Customer ID` <dbl> 13085, 13085, 13085, 13085, 13085, 13085, 13085, 1308...
## $ Country       <chr> "United Kingdom", "United Kingdom", "United Kingdom",...
```

# Summary of Revenue and Transaction by Year

```r
library(stringr)
library(kableExtra)
retaildf$Date <- str_sub(retaildf$InvoiceDate, start=1, end=10)
retaildf$Year <- str_extract(retaildf$InvoiceDate, "\\d{4}")

summaReTr <- retaildf %>%
  filter(!is.na(Description)&!is.na(`Customer ID`)&Quantity > 0&Price > 0) %>%
  mutate(Revenue = round(Quantity*Price, digit = 2)) %>%
  group_by(Year) %>%
  summarise(Revenue = sum(Revenue),
            n_transactions = n_distinct(Invoice),
            n_obs = n(),
            First_Date = min(Date),
            Last_Date = max(Date))
kbl(summaReTr, booktabs = T)
```

| Year | Revenue | n_transactions | n_obs | First_Date | Last_Date |
|------|---------|----------------|-------|------------|-----------|
| 2009 | 686654.2 | 1512 | 30754 | 2009-12-01 | 2009-12-23 |
| 2010 | 8718063.0 | 18325 | 403067 | 2010-01-04 | 2010-12-23 |
| 2011 | 8338712.0 | 17132 | 371728 | 2011-01-04 | 2011-12-09 |

Revenue Analysis from Invoices Data

# An Online Retailer Case

- This is Online Gift Shop Retailer
- Data from 12/1/2009 to 12/9/2011 daily including time information
- The number of annual transactions is about 18,000 (Year 2010)
- The annual revenue is about 8 millions (Year 2010)

# Pre-processing 1

```r
library(dplyr)
library(ggplot2)
library(lubridate)
library(tidyverse)
library(scales)
library(janitor)
library(epiDisplay)

retaildf_2009 <- retaildf %>%
    filter(as.Date.POSIXct(InvoiceDate)=='2009-12-01') %>%
    filter(!is.na(Description) &!is.na(`Customer ID`) & Quantity > 0) %>%
    mutate(Dollar_Total = Quantity * Price) %>%
    group_by(StockCode, Description) %>%
    summarise(Total_Earned = sum(Dollar_Total),
              Total_Sold = sum(Quantity)) %>%
    arrange(desc(Total_Earned)) %>%
    ungroup() %>%
    mutate(Proportion_of_Revenue = scales::percent( Total_Earned/sum(Total_Earned)))
```

# Top 10 Gift Items Sold

```
top10 <- retaildf_2009 %>%
    filter(Total_Earned> 0) %>%
    slice_max(Total_Earned, n=10)
top10$StockCode <- NULL
kbl(top10, booktabs = T) %>%
kable_styling(latex_options = "striped")
```

| Description | Total_Earned | Total_Sold | Proportion_of_Revenue |
|---|---|---|---|
| ASSORTED COLOUR BIRD ORNAMENT | 1919.28 | 1272 | 4% |
| PAPER CHAIN KIT 50'S CHRISTMAS | 998.40 | 368 | 2% |
| PAPER CHAIN KIT RETRO SPOT | 729.95 | 277 | 2% |
| RETRO SPOT TEA SET CERAMIC 11 PC | 727.80 | 164 | 2% |
| WHITE HANGING HEART T-LIGHT HOLDER | 681.35 | 257 | 2% |
| PINK CHERRY LIGHTS | 601.65 | 103 | 1% |
| SCOTTIE DOG HOT WATER BOTTLE | 583.20 | 128 | 1% |
| WHITE CHERRY LIGHTS | 558.60 | 92 | 1% |
| POSTAGE | 505.00 | 15 | 1% |
| FLORAL ELEPHANT SOFT TOY | 398.25 | 105 | 1% |

# Bottom 10 Gift Items Sold

```r
bottom10 <- retaildf_2009 %>%
    filter(Total_Earned > 0) %>%
    slice_min(Total_Earned, n=10)
bottom10$StockCode <- NULL
kbl(bottom10, booktabs = T) %>%
kable_styling(latex_options = "striped")
```

| Description | Total_Earned | Total_Sold | Proportion_of_Revenue |
|-------------|-------------:|-----------:|-----------------------|
| PAPER POCKET TRAVELING FAN | 0.28 | 2 | 0% |
| LOVE POTION MASALA INCENSE | 0.42 | 2 | 0% |
| HAPPY ANNIVERSARY CANDLE LETTERS | 0.42 | 1 | 0% |
| KITCHEN METAL SIGN | 0.55 | 1 | 0% |
| TOILET METAL SIGN | 0.55 | 1 | 0% |
| HEART DECORATION PAINTED ZINC | 0.65 | 1 | 0% |
| DOVE DECORATION PAINTED ZINC | 0.65 | 1 | 0% |
| STAR DECORATION PAINTED ZINC | 0.65 | 1 | 0% |
| RAIN HAT WITH RED SPOTS | 0.84 | 2 | 0% |
| POP ART PEN CASE & PENS | 0.85 | 1 | 0% |
| 12 PENCILS TALL TUBE WOODLAND | 0.85 | 1 | 0% |
| BIRD BOX CHRISTMAS TREE DECORATION | 0.85 | 1 | 0% |
| SET/20 POSIES PAPER NAPKINS | 0.85 | 1 | 0% |
| PANDA AND BUNNIES STICKER SHEET | 0.85 | 1 | 0% |
| PACK 20 ENGLISH ROSE PAPER NAPKINS | 0.85 | 1 | 0% |

# Pre-processing 2

```r
proportion_totals <- retaildf_2009 %>%
    mutate(daily_revenue = case_when(
        between(Total_Earned, 1001, max(Total_Earned)) ~ "> $1000",
        between(Total_Earned, 500, 1000) ~ ">$500",
        between(Total_Earned, 301, 500) ~ ">$300",
        between(Total_Earned, 201, 300) ~ ">$200",
        between(Total_Earned, 101, 200) ~ ">$100",
        between(Total_Earned, 51, 100) ~ ">$50",
        between(Total_Earned, 25, 50) ~ ">=$25",
        between(Total_Earned, 0, 24) ~ "<$25",
        TRUE ~ "loss")) %>%
    filter(daily_revenue != "loss")
```

# Results: Distribution of Revenue by SKU

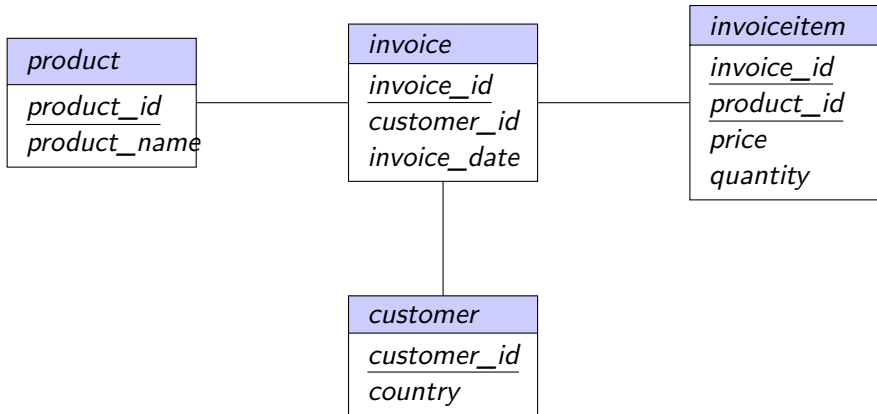**Distribution of Daily Revenue by SKU**
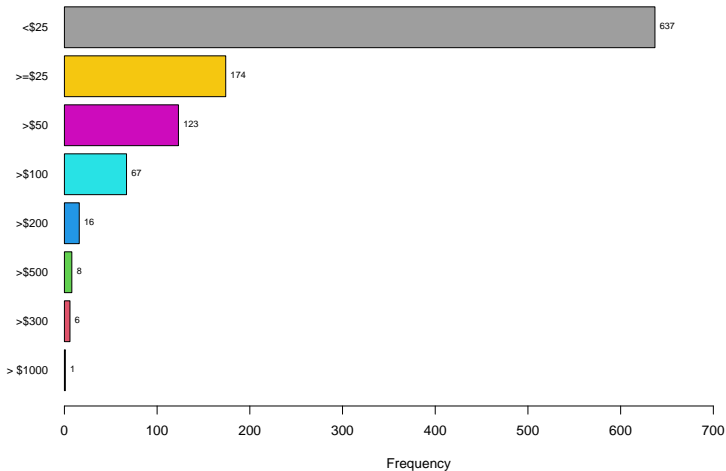
# Data Management with Database

## Why Database: An Example of Securing Data

```
# establishing the connectioon to SQL server to access db
con <- dbConnect(odbc(),
  # server type
  Driver = "SQL Server",
  #server name
  Server = "ATM\\ATMSERVER",
  # this is one of the db I want to import
  Database = "Data607_Project3_db",
  UID = "Alex",
  # password required
  PWD = rstudioapi::askForPassword("Database password"),
  ort = 1433)
}
```

PWD = rstudioapi::askForPassword("Database password")

# E-R Diagram



| product | | invoice | | invoiceitem |
|---|---|---|---|---|

**product**
- *product_id*
- *product_name*

**invoice**
- *invoice_id*
- *customer_id*
- *invoice_date*

**invoiceitem**
- *invoice_id*
- *product_id*
- *price*
- *quantity*

**customer**
- *customer_id*
- *country*

# Normalizing

### A set of normalized tables

Customer, product, invoice, and invoiceitem tables

### An efficient way to storing data

Easy for data maintenance and upgrade

### A secured way to access data

Control access to data with password

# An Example: The customer table

```r
varKeep <- c("Customer ID", "Country")
customerTable <-
  retaildf[unique(retaildf$`Customer ID`), varKeep]
names(customerTable) <- c("CustomerID", "Country")
customerTable <- customerTable %>%
  drop_na(CustomerID) %>%
  arrange(desc(CustomerID))
```