Introductory Data Cleaning and Presentation for Data Science

Karen Starin, Assistant Professor, Columbus State Community College
Andrew Kerr, Instructor, Columbus State Community College

Abstract

There is a significant and growing demand for data-savvy professionals at all levels of the workforce in the emerging field of data science. Methods of data cleaning, dealing with missing values, and introductory exploratory data analysis will be presented at a level appropriate for students of two year colleges.

Students will need to have a thorough understanding of the data, and can look at features of the data set to provide a summary to data scientists, as well as provided a cleaner data file for analyses.

For an example, consider an anonymized data set containing information from 45 retail stores. There is a lot of information here, so let's take a look at some of the steps a student might take in the process of summarizing the data.

First, we need to read in the data, which are stored in a CSV file named retailSalesData.csv:

```
setwd(file.path('S:','AMATYC','Presentation 2017'))
salesData = read.csv('retailSalesData.csv')
```

A quick look at the data shows that the data were read in:

```
head(salesData)
```

```
##   Store Dept Type   Size     Date Temperature Fuel_Price Weekly_Sales
## 1     1   49    A 151315 1/4/2011       59.17      3.524     13167.85
## 2     1   26    A 151315 1/4/2011       59.17      3.524      5946.53
## 3     1   81    A 151315 1/4/2011       59.17      3.524     28545.23
## 4     1   34    A 151315 1/4/2011       59.17      3.524      9949.54
## 5     1   59    A 151315 1/4/2011       59.17      3.524       316.86
## 6     1   30    A 151315 1/4/2011       59.17      3.524      3897.48
##   MarkDown1 MarkDown2 MarkDown3 MarkDown4 MarkDown5      CPI Unemployment
## 1        NA        NA        NA        NA        NA 214.8372        7.682
## 2        NA        NA        NA        NA        NA 214.8372        7.682
## 3        NA        NA        NA        NA        NA 214.8372        7.682
## 4        NA        NA        NA        NA        NA 214.8372        7.682
## 5        NA        NA        NA        NA        NA 214.8372        7.682
## 6        NA        NA        NA        NA        NA 214.8372        7.682
##   IsHoliday
## 1     FALSE
## 2     FALSE
## 3     FALSE
## 4     FALSE
## 5     FALSE
## 6     FALSE
```

One of the things we can look at is the number of observations and the number of variables:

```
dim(salesData)
```

```
## [1] 421570     16
```

This shows that there are 421570 observations with 16 variables. Summaries of each of the variables can be helpful. Such summaries can include the type of variable (numeric or categorical), the distribution of each

variable, such as a histogram for numeric variables or a table/bar chart for categorical variables, and the number of missing values that are in each variable.

In this data set, there are several variables:

```
str(salesData)
```

```
## 'data.frame':    421570 obs. of  16 variables:
##  $ Store       : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Dept        : int  49 26 81 34 59 30 7 85 8 28 ...
##  $ Type        : Factor w/ 3 levels "A","B","C": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Size        : int  151315 151315 151315 151315 151315 151315 151315 151315 151315 151315 ...
##  $ Date        : Factor w/ 143 levels "1/10/2010","1/4/2011",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Temperature : num  59.2 59.2 59.2 59.2 59.2 ...
##  $ Fuel_Price  : num  3.52 3.52 3.52 3.52 3.52 ...
##  $ Weekly_Sales: num  13168 5947 28545 9950 317 ...
##  $ MarkDown1   : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ MarkDown2   : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ MarkDown3   : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ MarkDown4   : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ MarkDown5   : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ CPI         : num  215 215 215 215 215 ...
##  $ Unemployment: num  7.68 7.68 7.68 7.68 7.68 ...
##  $ IsHoliday   : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

Oftentimes, data will come with a reference to what each variable represents. We can see that the variables size, temperature, fuel price, weekly sales, markdown1 - markdown5, CPI, and unemployment are all numeric variables. The variables store, department, type, and IsHoliday are categorical, and there is a date variable included. We can look at some summaries of the dataset for two of the variables: temperature and IsHoliday.

Temperature is considered a numeric variable, so we can look at both summary statistics as well as a histogram of the data:

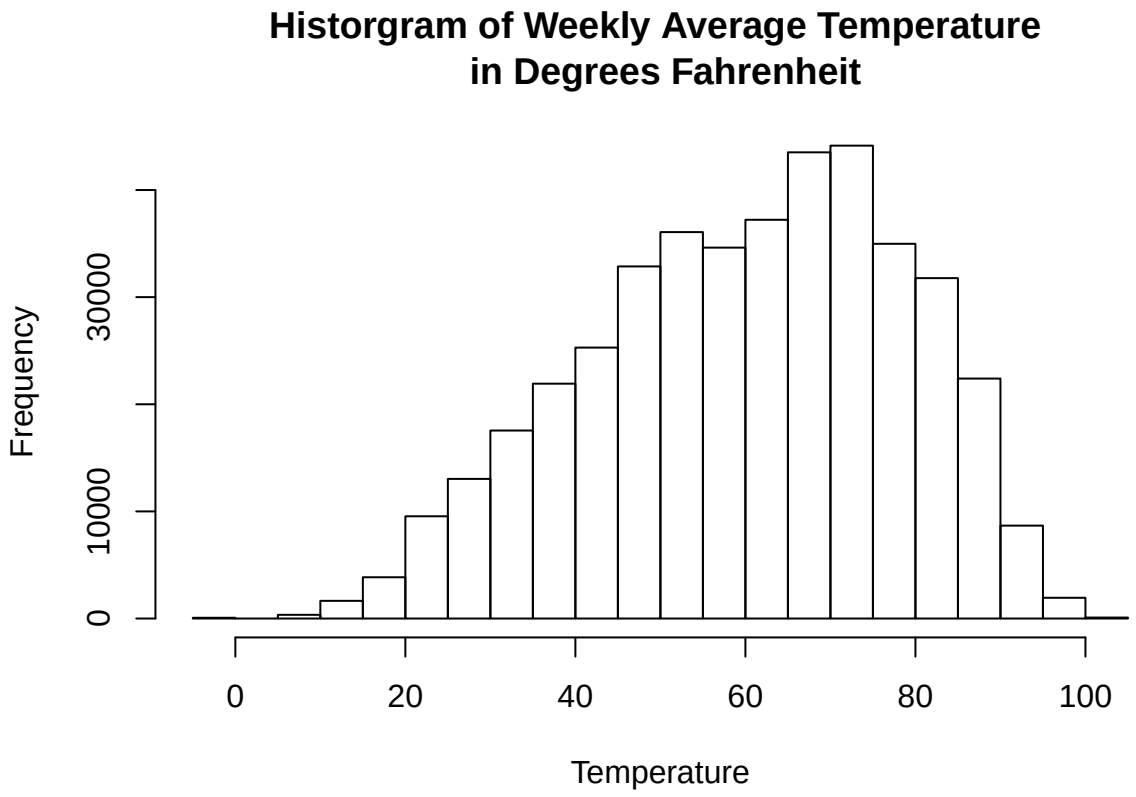
```
summary(salesData$Temperature)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -2.06   46.68   62.09   60.09   74.28  100.10
```

The values of the summary statistics lead us to conclude that the units are degree Fahrenheit

Looking at a histogram of the data:

```
hist(salesData$Temperature,xlab='Temperature',
     main='Historgram of Weekly Average Temperature \n in Degrees Fahrenheit')
```



There tends to be higher temperatures more than lower temperatures, which depending on the location of the retail stores, could be assessed for accuracy.

We can also take a look at the IsHoliday variable, which indicates whether or not a holiday takes place in a given week. This is a categorical variable, so the summaries will differ from numeric variables, such as temperature.

Depending on the software used, it may be helpful to recode IsHoliday with 1s and 0s:

```
salesData$IsHolidayIND = ifelse(salesData$IsHoliday == 'FALSE',0,1)
```

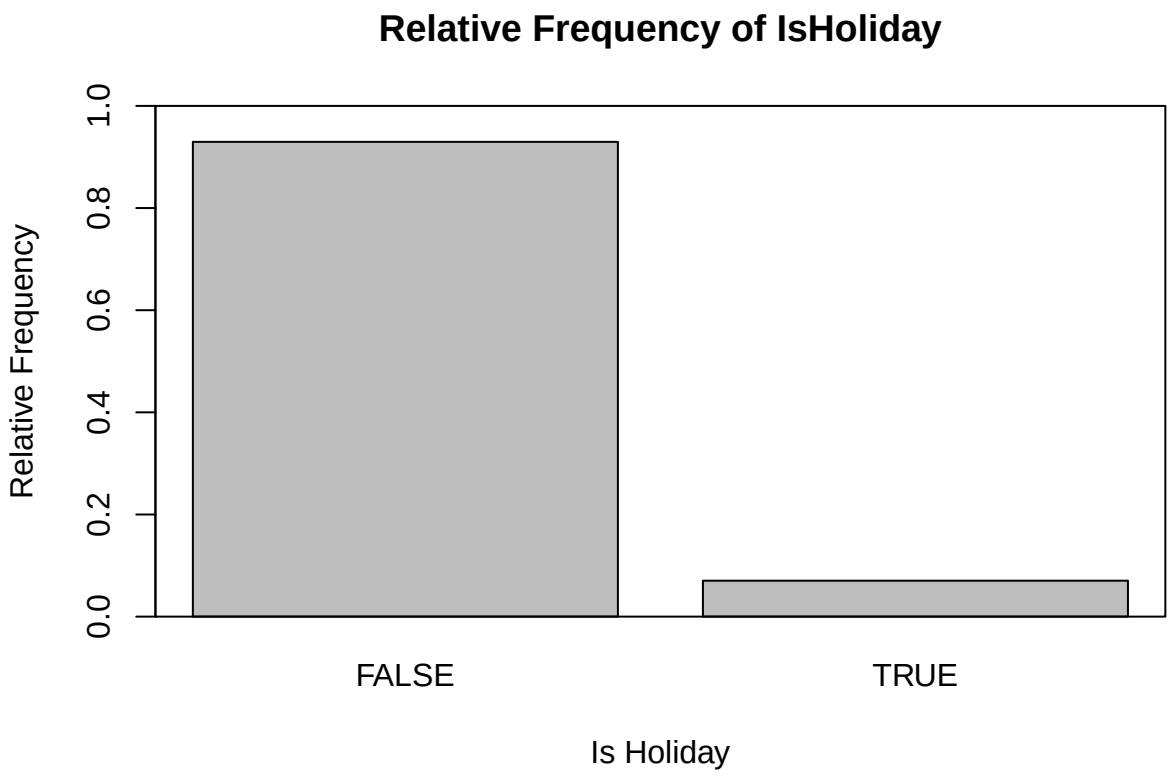
We can look at a table summary of the values

```
table(salesData$IsHoliday)
```

```
##
##  FALSE   TRUE
## 391909  29661
```

Or a plot, such as a bar chart

```
barplot(table(salesData$IsHoliday)/nrow(salesData),xlab='Is Holiday',ylab='Relative Frequency',
        main='Relative Frequency of IsHoliday',ylim=c(0,1))
box()
```



Another bit of interest is the Date variable, which gives the week the data were collected. If we look at some of the values in this variable, we see something that initially looks odd:

```
salesData$Date[96255:96265]
```

```
##  [1] 19/11/2010 19/11/2010 19/11/2010 19/11/2010 20/01/2012 20/01/2012
##  [7] 20/01/2012 20/01/2012 20/01/2012 20/01/2012 20/01/2012
## 143 Levels: 1/10/2010 1/4/2011 1/6/2012 1/7/2011 10/12/2010 ... 9/9/2011
```

These dates are not in the mm/dd/yyyy format with which most students are familiar, but appear to be in dd/mm/yyyy format. So, for example 19/11/2010 refers to November 19, 2010, and 01/02/2011 would refer to February 1, 2011.

These are just a few examples of notes that students could make in an initial run through of a data set, noting summaries and issues as they appear in data preparation.

Variable Report Retail Sales Name: _____

Variable	Type	Units	Comments	Missing Value Code	Graph 1	Graph 2	Graph 3	Suspicious Data	Numeric Code
Store									
Dept									
Type									
Size									
Date									
Temperature									
Fuel Price									
Weekly Sales									
MarkDown1									
MarkDown2									
MarkDown3									
MarkDown4									
MarkDown5									
CPI									
Unemployment									
IsHoliday									