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:

```r
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:

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

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

```r
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)
```

```r
## '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)
```

```r
##      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:

```r
hist(salesData$Temperature, xlab='Temperature',
     main='Histogram of Weekly Average Temperature 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:

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

We can look at a table summary of the values

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

```r
# # FALSE  TRUE
# 391909 29661
```
Or a plot, such as a bar chart

```r
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:

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

```
## [7] 20/01/2012 20/01/2012 20/01/2012 20/01/2012 20/01/2012
```

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**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Type</th>
<th>Units</th>
<th>Comments</th>
<th>Missing Value Code</th>
<th>Graph 1</th>
<th>Graph 2</th>
<th>Graph 3</th>
<th>Suspicious Data</th>
<th>Numeric Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dept</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Date</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Temperature</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fuel Price</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weekly Sales</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MarkDown1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MarkDown2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MarkDown3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MarkDown4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MarkDown5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CPI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unemployment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IsHoliday</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>