Assignment 9.1

Scott Breitbach

7/27/2020

## Assignment 9.1: Final Project Step 2: Cleaning Your Data and Exploratory Data Analysis

### Data importing and cleaning steps are explained in the text and in the Github exercises. (Tell me why you are doing the data cleaning activities that you perform). Follow a logical process.

With the Husker game data, some of the teams had numbers in front of them, presumably indicating their ranking if they were in the top 25 going into that game. I had to remove the numbers, characters, and spaces in order to unify the team names. I also had to convert the dates and times to the appropriate formats using lubridate.

I will change some of the information into more useful formats. For example the away games have an “@” sign indicating they are at the opposing team’s field. This column will be relabeled “Home”; the “@”s will be converted to 0’s and the “NA”s converted to 1’s. The Win/Loss column will be relabeled “Win” and the “W”s converted to 1’s and “L”s to 0’s. The remaining columns will be removed.

Some of the police data was already compiled from 2013 to 2020, some from 2017 to 2020, and some was just available by year so I had to merge data sets in order to make them all 2013 to 2019 (I removed 2020 data because football season has not started). While there are a few exceptions, the vast majority of the games are in September, October, and November so I will use these three months and remove all other data.

Some of the newer data sets also had more columns. Since I am really only interested in number of events, I am paring them down to the essentials: dates, times, and the overall category (I don’t plan to use the categories, but I feel they may come in useful later). Some of the rows didn’t parse, but there were a small number of them (for the incident reports, there were ~136 out of >250,000 rows) so I am accepting that risk as minor.

### With a clean dataset, show what the final data set looks like. However, do not print off a data frame with 200+ rows; show me the data in the most condensed form possible.

## [1] "Game Data"

## # A tibble: 6 x 7  
## Date Time Day School Home Opponent Win  
## <date> <time> <ord> <chr> <dbl> <chr> <dbl>  
## 1 2013-09-07 18:00 Sat Nebraska 1 Southern Mississippi 1  
## 2 2013-09-14 12:00 Sat Nebraska 1 UCLA 0  
## 3 2013-09-21 15:30 Sat Nebraska 1 South Dakota State 1  
## 4 2013-10-05 12:00 Sat Nebraska 1 Illinois 1  
## 5 2013-10-12 12:00 Sat Nebraska 0 Purdue 1  
## 6 2013-10-26 12:00 Sat Nebraska 0 Minnesota 0

## [1] "Arrests and Citations"

## # A tibble: 6 x 4  
## Charge Date Time Day   
## <chr> <date> <time> <ord>  
## 1 ATTEMPT THEFT BY SHOPLIFTING $0-$500 2018-11-21 15:03 Wed   
## 2 VIOLATE PROTECTION ORDER 2018-10-17 12:00 Wed   
## 3 IDENTITY THEFT $5000/MORE 2018-09-11 00:00 Tue   
## 4 IDENTITY THEFT $5000/MORE 2018-09-11 00:00 Tue   
## 5 LEAVE ACCIDENT-FAIL TO FURNISH INFO/1ST 2018-11-22 21:24 Thu   
## 6 DRIVE DURING SUSPENSION/BEFORE REINSTATED 2018-11-22 21:24 Thu

## [1] "Incident Reports"

## # A tibble: 6 x 4  
## Type Date Time Day   
## <chr> <date> <time> <ord>  
## 1 SEX OFF 2013-11-01 00:01 Fri   
## 2 SEX OFF 2013-11-01 00:01 Fri   
## 3 SUSP ITEM 2013-09-11 08:00 Wed   
## 4 STALKING 2013-11-12 07:00 Tue   
## 5 SELL NARCO 2013-09-19 20:35 Thu   
## 6 SUSP ITEM 2013-10-07 12:00 Mon

## [1] "Traffic Crashes"

## # A tibble: 6 x 12  
## Type Action Pedestrian Bike Motorcycle Moped Train Truck Bus Date   
## <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <date>   
## 1 INJU~ REAR ~ NO NO YES NO NO NO NO 2013-11-16  
## 2 INJU~ DRIVE~ NO NO YES NO NO YES NO 2013-11-29  
## 3 INJU~ OTHER NO YES YES NO NO NO NO 2013-10-10  
## 4 INJU~ RIGHT~ NO YES NO NO NO NO NO 2013-09-02  
## 5 INJU~ OTHER YES NO YES NO NO NO NO 2013-10-01  
## 6 INJU~ DRIVE~ NO YES NO NO NO NO NO 2013-09-02  
## # ... with 2 more variables: Time <time>, Day <ord>

## [1] "Traffic Stops"

## # A tibble: 6 x 4  
## Reason Date Time Day   
## <dbl> <date> <time> <ord>  
## 1 1 2013-09-02 03:07 Mon   
## 2 1 2013-09-01 02:30 Sun   
## 3 2 2013-09-01 02:36 Sun   
## 4 1 2013-09-02 03:50 Mon   
## 5 1 2013-09-02 18:25 Mon   
## 6 1 2013-09-03 08:11 Tue

### What do you not know how to do right now that you need to learn to import and cleanup your dataset?

Importing csv files was pretty straightforward using the readr library and the rbind() function to combine consecutive data sets. Cleaning was a bit more challenging. Most of the functions I could not remember off of the top of my head and I had to look them up. This ranged from removing extra numbers, characters, and spaces from the “School” and “Opponent” columns, to renaming columns and formatting dates, to removing columns and converting factors to 1s and 0s (including some NAs).

I still need to figure out how to remove rows within certain date ranges, but since I am keeping only three of the months I think I should be able to create a “Month” column and essentially filter out (or in) the appropriate months.

### Discuss how you plan to uncover new information in the data that is not self-evident.

I plan to uncover information primarily through the use of different types of charts in order to look for patterns in the data that may not be immediately obvious.

### What are different ways you could look at this data to answer the questions you want to answer?

I plan to take advantage of the fact that all of my data contains date-time information. This should allow me to looks for patterns based on different date-time groupings, like day of the week or time of day as well as which weekends the Huskers had home or away games and who they played against.

### Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain

Apart from merging data sets from multiple years, I think most of the slicing and dicing I’m going to encounter will revolve around working with dates and times, especially as they relate to the dates and times of the Husker games. At a minimum, I will need to create a column for day of the week.

In addition, I kept type of incidents listed in the police reports, which may allow me to look for correlations with certain types of incidents instead of just number of incidents in general.

### How could you summarize your data to answer key questions?

I think primarily the use of charts will be the most effective way to summarize my data and show any correlations between Husker game timing and crime and traffic incidents. I may then be able to use these correlations to predict likelihood of incidences occurring around future games.

### What types of plots and tables will help you to illustrate the findings to your questions? Ensure that all graph plots have axis titles, legend if necessary, scales are appropriate, appropriate geoms used, etc.).

I think the best way to illustrate my findings will be through the use of bar charts and histograms in order to compare the variable of interest to a baseline or to other, related variables.

### What do you not know how to do right now that you need to learn to answer your questions?

I need to figure out how to categories date and time ranges into both Husker home game weekends and morning vs evening games and Friday vs Saturday games, as well as before and after Husker games. I also need to figure out how to calculate the number of incidents occurring within a given time frame.

### Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

If there is any effect, I think I should be able to use linear or multiple regression in order to predict the number of incidents based on the different criteria I’m looking at, such as day of the week, home vs away game, and opposing team.