Stat 133 HW05: NFL Scores and Weather

Gaston Sanchez

Introduction

This assignment involves performing a simple yet complete data analysis from A to Z. You'll have the opportunity to apply what we've seen so far in the course, and other additional topics.

There are two main parts for this project:

a. Data cleaning and preprocessing
b. Data analysis

You'll have to write two separate R scripts (i.e. two files with `.R` extension) for each part.

Data: NFL Weather

The data for this project comes from nflsavant.com by Daren Willman. It consists of scores and weather conditions of all NFL (National Football League) games from 1960 to 2013. The data is in a csv file located at http://nflsavant.com/dump/weather_20131231.csv.

Download a copy of the file (don't read it directly in R!). You can use the function `download.file()` to get a copy of the file in your computer. Name the file as `raw_nflweather.csv`. This copy will be your **raw** data.

DATA CLEANING AND PREPARATION

Create an R script dedicated to write all the instructions and commands to clean and preprocess the data (the instructions of the analysis will be in a different R script).

First contact

- Import `raw_nflweather.csv` in R (make sure that strings are NOT converted into factors)
- Perform a first inspection of the data using functions like `str(), head(), tail()`, etc.

Weather Information

The data contains columns `temperature`, `wind_chill`, `humidity`, and `wind_mph`. The same information is contained in the column `weather`. Let's use the column `weather` to practice **regular expressions**. Use regex functions (either base R or from `"stringr"`) to:

- remove the percent symbol `%` from the values in column `humidity` and convert such values to numeric format
- extract the temperature values from column `weather`; and create a column `temperature2` with these values (as numeric).
- extract the humidity values from column `weather`; and create a column `humidity2` with these values (as numeric)

- extract the wind speed values from column `weather`; and create a column `wind2` with these values (as numeric)

Make sure that the new columns coincide with the pre-existing ones. For instance, make sure that when running a `summary()` on both `temperature` and `temperature2` you get the same values. Do the same for `humidity` and `humidity2`, and also for `wind_mph` and `wind2`.

Date Information

There is a column `date` that contains dates in format `month/day/year`.

- Create a column `year` that contains the number of the year (as numeric).
- Create a column `monthnum` that contains the number of the month (as numeric)
- Create a column `month` that contains the name of the corresponding month (as factor); e.g. if the month number is 9 then month will be `september`.
- Create a column `decade` that indicates the corresponding decade (as factor) of each played game. Use labels: `1960s`, `1970s`, `1980s`, `1990s`, `2000s`, `2010s`. For instance, all games between 1970 and 1979 will have the associated decade `1970s`.

Scores Information

- Create a column `total_score` that contains the total number of scored points in each game. In other words, the sum of `home_score` and `away_score`
- Create a column `diff_score` that indicates the difference of `home_score` and `away_score`. In other words, the subtraction of `home_score` and `away_score`
- Create a column `home_win` that shows whether `home_score` is greater than `away_score`. This column will have logical values (`TRUE` or `FALSE`)

Basic Exploration

- Inspect variables `home_score`, `away_score`, `temperature`, and `wind_mph` by getting summary statistics (this is just a first inspection making sure there are no "abnormal" values)
- Visually inspect variables `home_score`, `away_score`, `temperature`, `humidity`, and `wind_mph` (these are "quick and basic" plots)
- What team has the maximum home score?
- What team has the maximum away score?
- What is the most common home score?
- What is the most common away score?
- What has been the maximum temperature in a game?
- What was the date of the maximum temperature?
- What has been the minimum temperature in a game?
- What was the date of the minimum temperature?
- How many games have been played with a temperature of 90 degrees or more?
- How many games have been played with a temperature below 0 degrees (do not include 0)?
- What is the most common temperature?
- Make a bar chart with the frequency table of temperatures? Is there anything that catches your attention?

Data files

Let's practice some basic data frame subsetting as well as manipulating names of files. The idea is to subset the dataset into different decades. That is, obtain one data.frame with data from the 1960s, another one from the 1970s, and so on. Select only the following variables (in the displayed order):

1. `id`	9. `date`
2. `home_team`	10. `year`
3. `home_score`	11. `month`
4. `away_team`	12. `decade`
5. `away_score`	13. `temperature`
6. `total_score`	14. `humidity`
7. `diff_score`	15. `wind_mph`
8. `home_win`	

You'll have to export a csv file for each decade:

- `nflweather1960s.csv` (years 1960 - 1969)
- `nflweather1970s.csv` (years 1970 - 1979)
- `nflweather1980s.csv` (years 1980 - 1989)
- `nflweather1990s.csv` (years 1990 - 1999)
- `nflweather2000s.csv` (years 2000 - 2009)
- `nflweather2010s.csv` (years 2010 - 2013)

Write a for loop that generates and saves these files into a folder named `cleandata`.

In addition to the datasets for each decade, you will export a file `nflweather.csv` with all the games (all decades). This will be your **cleaned** data, and the one you'll use for the main analysis part.

DATA ANALYSIS

Use a new R script dedicated to the analysis part of this assignment. The dataset that you have to use is the one from the file `nflweather.csv` (*the clean data*). In this part you'll have to focus on two main research questions:

1. Does playing at home really have an advantage for the home team?
2. Has the total number of scored points per game have changed over time?

Note that these are general open questions. You'll have to think about what is needed to provide an answer to each question. This is where you'll put your analytical skills in practice.

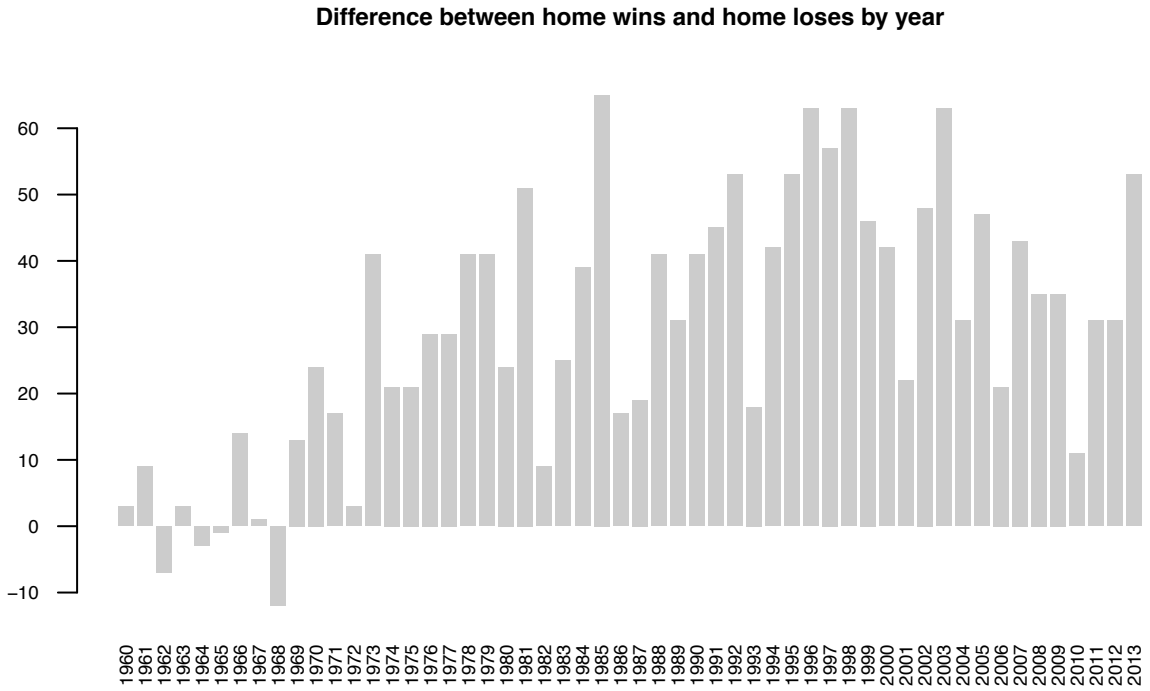
1) Playing at home has an advantage for the home-team?

Think about it: in any sport, the team that plays at home seems to have an advantage over the visiting team. The team that plays at home know their field better than anyone else. They are also used to the local weather conditions. Playing at home also means having the support of the fans. But as you know, playing at home does not guarantee a victory.

Your task is to provide evidence in favor of the claim that *"playing at home does give the home team an advantage over the visiting team"*. For this, you should work with aggregated data by year.

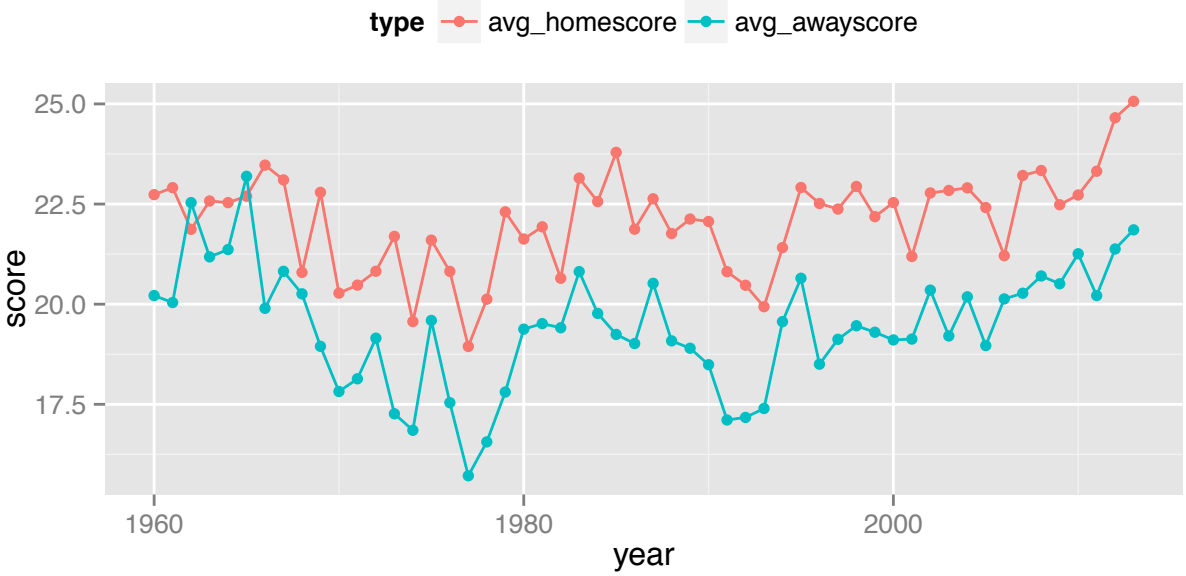
Your analysis should include—but it shouldn't be limited to—the following two figures:

Number of "More Home Victories" per year One way to provide supporting evidence that playing at home has an advantage, is by using the data in column `home_win`. Overall, there should be more games won by the home-team than those won by the visiting team. Aggregating the difference of home victories and away victories by year allows us to get a bar chart like this:



Except for a few years in the 1960s (1962, 1964, 1965, and 1968), the rest of the seasons confirm that more games are won by the home-team rather than the visiting team.

Average score points per year



Another way to provide supporting evidence that confirms that playing at home is advantageous for the home-team is to use game average scores per year. As you can tell from the previous plot, the average of `home_score` is almost always above the average of `away_score`.

What other type of analysis and/or visualization(s) can you think of in order to answer the first research question?

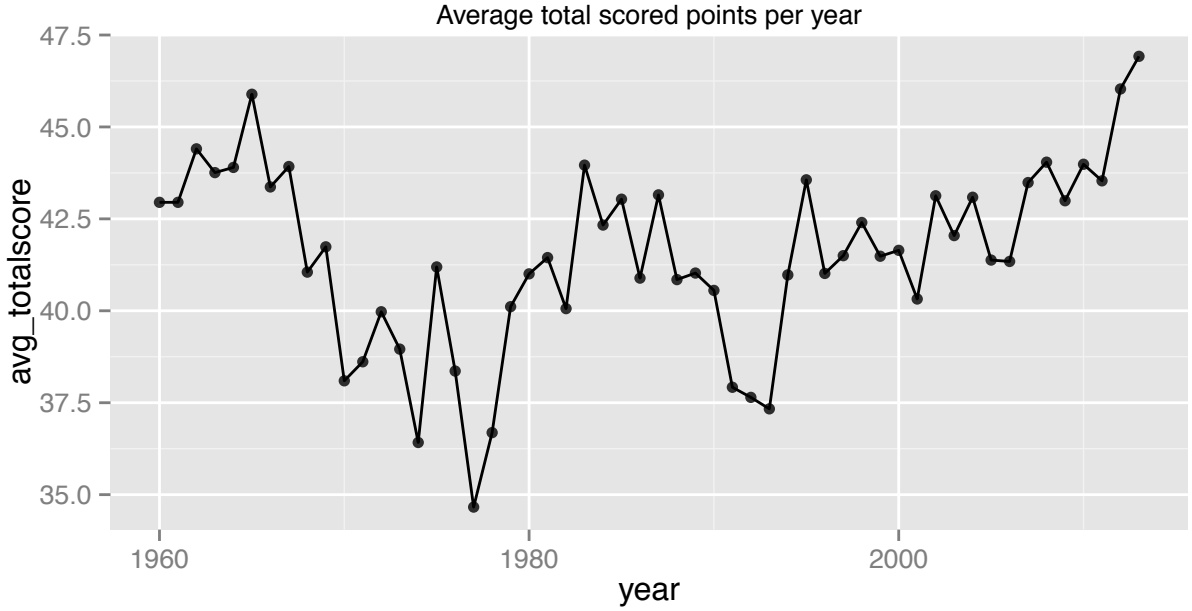
---

2) Has the total number of scored points per game changed over time?

The second research question is: Has there been any change in the total number of scored points per game over time?

Over the years, there have been modifications of NFL game rules. There have also been changes in gear, equipment, training methods, game strategies, diets, etc. Also the NFL business has evolved, along with the consumption patterns from the general public. What can you say about the scored points per game? Are there more points scored *now* than in the *past*? Or has the number of scored points remained more or less constant?

One way to answer this question is by aggregating the total number of scored points per year, and taking the average. The following plot shows the average total scored points per year. Your analysis should include—but it shouldn't be limited to—the following figure:



As you can tell, the total scored points have indeed changed over time (they haven't remained constant). However, the changes seem to have a random behavior (like in a stock market). Some years have a low average while other years have a high average.

Try to replicate the figure above (it doesn't have to be necessarily obtained in `"ggplot2"`); feel free to customize your plot with colors, sizes, other visual elements, legends, etc.

What other type of analysis and/or visualization(s) can you think of in order to answer the second research question?