­Peer Assessment 2: Summary

* **Log:** 15 hrs
* Lectures example: <http://www.rpubs.com/rdpeng/13396>
* Try to make formatting phenomenal. Some RMD are too dense.

**Research**

(1) <https://rpubs.com/cbradbury/repres-peerassessment2>

* Excessive heat and tornado cause most fatalities; tornado causes most injuries in the United States from 1995 to 2011.
* Based on the histograms above, we find that flood and hurricane/typhoon cause most property damage; drought and flood causes most crop damage in the United States from 1995 to 2011.
* Used column graphs with two panels using ggplot for both questions.

(2) <https://rpubs.com/suelynnk/120251>

* most harmful weather event to population health is Tornado. Caused the highest fatalities and injuries across US.
* -The weather event caused the greatest economic consequences. They are flood, drought, tornado and typhoon. Across the United States, flood, tornado and typhoon have caused the greatest damage to properties. While drought and flood was the reason that caused the greatest damage to the crops.
* Used simple column graphs with two panels using base package barplot().

(3) <http://rstudio-pubs-static.s3.amazonaws.com/18671_b8cb6c37db084d5cb3cb8a8af27a7839.html>

* Tornadoes cause the highest number of total injuries and fatalities in the U.S. Excessive heat, occurs less often than tornadoes, still causes a relatively high number of injuries and fatalites per event.
* From the economic perspective; damage to property is higher than the damage to crops. The most damaging weather event is flood when considering both property and crops or when analysing only property. For crops; the worst weather event is drought.
* Used a horizontal bar graph with 1 panel with fatilities + injuries combined in one bar. Did this for both.

**Assignment instructions** (1)

**Introduction**

Explore the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database. This tracks characteristics of major storms and weather events in the US, including when and where they occur, as well as estimates of any fatalities, injuries, and property damage.

**Data**

* [Storm Data](https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2) [47Mb] There is also some documentation where you will find how some of the variables are constructed/defined.
* National Weather Service [Storm Data Documentation](https://d396qusza40orc.cloudfront.net/repdata%2Fpeer2_doc%2Fpd01016005curr.pdf)
* National Climatic Data Center Storm Events [FAQ](https://d396qusza40orc.cloudfront.net/repdata%2Fpeer2_doc%2FNCDC%20Storm%20Events-FAQ%20Page.pdf)

The events in the database start in the year 1950 and end in November 2011. In the earlier years of the database there are generally fewer events recorded, most likely due to a lack of good records. More recent years should be considered more complete.

**Assignment:**

The basic goal of this assignment is to explore the NOAA Storm Database and answer some basic questions about severe weather events. You must use the database to answer the questions below and show the code for your entire analysis. Your analysis can consist of tables, figures, or other summaries. You may use any R package you want to support your analysis.

**Questions**

Your data analysis must address the following questions:

1. Across the United States, which types of events (as indicated in the EVTYPE variable) are most harmful with respect to population health?
2. Across the United States, which types of events have the greatest economic consequences?

**Requirements**

For this assignment you will need to use the knitr package and R to publish your completed analysis document to RPubs. You can also use RStudio to edit/write your analysis.

**Document layout**

* Your document should be written in English. Your document should have a title that **briefly** summarizes your data analysis, Immediately after the title, there should be a **synopsis** which describes and summarizes your analysis in at **most 10 complete sentences**. There should be a section titled **Results** in which your results are presented. You may have other sections in your analysis, but Data Processing and Results are **required**.
* There should be a section titled **Data Processing** which describes (in words and code) how the data were loaded into R and processed for analysis. In particular, your analysis *must* start from the raw CSV file containing the data. You cannot do any preprocessing outside the document. If preprocessing is time-consuming you may consider using the cache = TRUE option for certain code chunks.
* The analysis document must have **at least one figure containing a plot**, and **no more than three figures**. Figures may have multiple plots in them (i.e. panel plots). You must show all your code for the work in your analysis document.

**Publishing Your Analysis and submitting your assignment**

You will need to publish your analysis on [RPubs.com](http://rpubs.com/). Read assignment to find out how if I need too. To submit the assignment, please submit the URL from RPubs that points to your full report for this assignment

**Important information about Project 2** (2)

* ~~make sure the data file was downloaded correctly. The compressed file (bz2 file) must be 46.8 MB (49,177,144 bytes). If the size is not correct, then try to redownload it. There are 902297 records on 37 variables if you get anything else then you need to check your reading step before you go farther.~~
* ~~If you opened the 'CSV' file outside the R you may find a different number of records. For example, in Notepad++ the number of rows will be 1,415,170 while in Microsoft excel that will be 903,870. The reason for this is the presence of carriage returns characters in some column (LF, CR & EOL). All these numbers are wrong!~~
* ~~You have the right to start your analysis from the 'bz2' file or from the "CSV' one. The decompressed one (csv) is tenfold the 'bz2' file, but R will read it almost twice faster. Nonetheless, the decompression itself will take some times too. So, decompress or not decompress that is the question? ;)~~

~~If you go with the decompression option, you can use 'bunzip2' function from '~~[~~R.utils~~](https://cran.r-project.org/web/packages/R.utils/index.html)~~' library. In that case make sure you set 'remove' option to FALSE, otherwise, your original data file will be deleted! It is useful to check the 'skip' option too~~

* ~~'read.csv' and 'read.table' both can read 'bz2' file directly. Also, '~~[~~fread~~](https://gist.github.com/wush978/93c0f96b68f529678e2d)~~' (from 'data.tabel' package) can do good work but its best performance will be on Linux/Unix OS. Go with the fastest one in your system.~~
* You can use "Sys.time()" before and after running function and subtract that to get the total time it takes. Whatever you choose, make sure you go with the best performance combination. Reading data step is 'heavy' and optimizing it will make a big difference
* ~~Even if you optimize data reading, 'knitr' will load that step every time you run your script, that will waste considerable amounts of time. As mentioned in the instructions, you need to use '~~[~~cache~~](http://yihui.name/knitr/demo/cache/)~~'.~~[~~Parameterized Reports~~](http://rmarkdown.rstudio.com/developer_parameterized_reports.html)~~provide more information about the available options. For the benefit of 'cache' keep data reading step in its own chunk.~~
* Most variable names are self-explanatory but few are vague and difficult to find in the [Strom Data Documentation](https://d396qusza40orc.cloudfront.net/repdata%2Fpeer2_doc%2Fpd01016005curr.pdf) or in [FAQ](https://d396qusza40orc.cloudfront.net/repdata%2Fpeer2_doc%2FNCDC%20Storm%20Events-FAQ%20Page.pdf), I'll give more details for those ones.
  + The 'CROPDMGEXP' is the exponent values for 'CROPDMG' (crop damage). In the same way, 'PROPDMGEXP' is the exponent values for 'PROPDMG' (property damage). You should use both to get the total values for crops and property damage. (B or b = Billion, M or m = Million, K or k = Thousand, H or h = Hundred). The number from one to ten represent the power of ten (10^The number). The symbols "-", "+" and "?" refers to less than, greater than and low certainty. You have the option to ignore these three symbols altogether. Also, there is a nice work entitled: [How To Handle Exponent Value of PROPDMGEXP and CROPDMGEXP](https://rstudio-pubs-static.s3.amazonaws.com/58957_37b6723ee52b455990e149edde45e5b6.html). It discusses that issue in more depth.
  + **~~WFO~~**~~=~~[~~Weather Forecast Office~~](https://en.wikipedia.org/wiki/List_of_National_Weather_Service_Weather_Forecast_Offices)~~,~~**~~F~~**~~= Fujita tornado intensity scale (~~[~~F-Scale~~](https://en.wikipedia.org/wiki/Fujita_scale)~~),~~**~~MAG~~**~~= Magnitude, or Strength, of the event. It is required by NOAA for Wind and Hail events if it is known. Wind Events are in KNOTS. Hail is in INCHES and TENTHS without the decimal (one and one-half are150).~~
  + **~~STATE\_\_~~**~~=~~[~~State FIPS number~~](https://en.wikipedia.org/wiki/FIPS_county_code)~~, you can download a full list of FIPS in excel sheet from~~[~~here~~](http://www.census.gov/2010census/xls/fips_codes_website.xls)~~.~~**~~LENGTH~~**~~= Path length of a tornado (in miles and tenths of miles).~~**~~WIDTH~~**~~= Path width of a tornado, in yards. According to NOAA documentation the length and width used for all tornadoes, including each member of~~[~~families of tornadoes~~](https://en.wikipedia.org/wiki/Tornado_family)~~, or for all segments of multi-segmented tornadoes.~~
* ~~According to~~[~~NOAA~~](http://www.ncdc.noaa.gov/stormevents/details.jsp?type=eventtype)~~the data recording start from Jan. 1950. At that time they recorded one event type, tornado. They add more events gradually and only from Jan. 1996 they start recording all events type. Since our objective is comparing the effects of different weather events, do we need to include all years, even it has only single event type?~~
* ~~The official events type are 48. However, if you use 'unique' function on 'EVTYPE' column you will get near one thousand events! All that is just typo. The regular expression ('~~[~~regex~~](https://stat.ethz.ch/R-manual/R-devel/library/base/html/regex.html)~~', grepl, regexpr, gregexpr) and 'tolower', and 'toupper' functions can be a great help here. But cleaning up all that mess can take several days.~~

~~However, reducing the size of data first will make a great difference here. Subset your data first, think of years, total value, frequency ... anything else that can be used to reduce the number of rows. Also, there are 37 variables, you do not need all of that for your analysis, remove as much as you can to make the analysis faster.~~

* ~~Speaking about data subset, can you rely on 'quantile' to get the top (some) percentage? or can we just get a random sample from the data for our analysis?~~
* ~~Back to 'EVTYPE' column, one way to fix the typo is loading the official list of storm event types and manually mapping all (or most) list of unique recorded events with a typo. This way is the most painful one. Another way is using '~~[~~match~~](https://stat.ethz.ch/R-manual/R-devel/library/base/html/match.html)~~' (from 'base' package) or even better you may use 'amatch' function from '~~[~~stringdist~~](https://cran.r-project.org/web/packages/stringdist/index.html)~~' package for Approximate String Matching. In that case, you will need to experiments with 'maxDist' option to get good accuracy.~~

**References**

1. Assignment Instructions
2. <https://www.coursera.org/learn/reproducible-research/discussions/weeks/4/threads/IdtP_JHzEeaePQ71AQUtYw>