Author: Soesilo Wijono (https://github.com/flyingdisc/RepData_PeerAssessment2)
Date: 9 Febrary 2015

How To Handle Exponent Value of PROPDMGEXP and CROPDMGEXP

Reproducible Research Project 2, Coursera, Johns Hopkins University

U.S. National Oceanic and Atmospheric Administration's (NOAA) Storm Database

Data repository:
Storm Data [47Mb]
Documentation:
National Weather Service Storm Data Documentation
National Climatic Data Center Storm Events FAQ

There is confusion on how to handle exponent value of PROPDMGEXP and CROPDMGEXP columns of the database. Due to lack of official information in the NOAA website.

This is an attempt to compare downloaded database with the online version, to find conclusion what is meaning of each value actually.

This analysis is inspired by a post made by **David Hood**, himself is CTA in the Data Science Specialization courses.
At the end of this report, there is most accurate analysis done by Eddie Song.

Note: EXP = exponent
These are possible values of CROPDMGEXP and PROPDMGEXP:

- H,h,K,k,M,m,B,b,+,-,?,0,1,2,3,4,5,6,7,8, and blank-character

  - H,h = hundreds = 100
  - K,k = kilos = thousands = 1,000
  - M,m = millions = 1,000,000
  - B,b = billions = 1,000,000,000
  - (+) = 1
  - (-) = 0
  - (?) = 0
  - black/empty character = 0
  - numeric 0..8 = 10

Proof:

After downloading the database.
Compare [storm data from this link] to the [StormData.csv].

In the R Studio, first read the data.

```
data <- read.csv("StormData.csv", sep=",", header=TRUE)
```

COMPARISONS:

(1.a) For numeric "3",

```
number <- data[data$PROPDMGEXP == "3",]
number[(number$EVTYPE == "THUNDERSTORM WINDS") & (number$STATE == "MO"),
       c("BGN_DATE", "BGN_TIME", "END_DATE", "STATE", "COUNTYNAME",
```

```
        "EVTYPE", "PROPDMG", "PROPDMGEXP")]

#        BGN_DATE BGN_TIME END_DATE STATE COUNTYNAME             EVTYPE PROPDMG  PROPDMGEXP
#214375 5/16/1995     1750             MO     SHELBY THUNDERSTORM WINDS      20           3
```

From NOAA link,

- Select State/Area = "Missouri" (MO)
- Select County = "All"
- Select Begin Date = End Date = "05/16/1995"
- Select Event Type = "Thunderstorm Wind"
- Click "Search"

Result:

```
"Shelbyville, SHELBY CO, MO, 05/16/1995, 17:50, Thunderstorm Wind, 0.20K, 0.00K"
```

Found, PrD (property damage) = 0.20K = 200,
While PROPDMG = 20,
Conclusion: (exp 3) is == (10)

(1.b) For numeric "5",

```
number <- data[data$PROPDMGEXP == "5",]
number[number$EVTYPE == "TORNADO",
       c("BGN_DATE", "BGN_TIME", "END_DATE", "STATE", "COUNTYNAME",
         "EVTYPE", "PROPDMG", "PROPDMGEXP")]

#        BGN_DATE  BGN_TIME END_DATE STATE COUNTYNAME  EVTYPE PROPDMG  PROPDMGEXP
#198635 5/27/1995      1715             IL     GREENE TORNADO    14.0           5
#199072 5/18/1995      1137             IL     MONROE TORNADO    88.0           5
#241111 5/17/1995      0055             TX     PARMER TORNADO     0.2           5
```

From NOAA link,

- Select State/Area = "Illinois", (IL)
- Select County = "All"
- Select Begin Date = End Date = "05/18/1995"
- Select Event Type = "Tornado"
- Click "Search"

Result:

```
"Waterloo, MONROE CO, IL, 05/18/1995, 11:37, Tornado, 0.88K, 0.00K"
```

Found, PrD (property damage) = 0.88K = 880,
While PROPDMG = 88,
Conclusion: (exp 5) is == (10)

(1.c) For numeric "0",

```
number <- data[data$PROPDMGEXP == "0",]
number[(number$EVTYPE == "TORNADO") & (number$STATE == "IA"),
       c("BGN_DATE", "BGN_TIME", "END_DATE", "STATE", "COUNTYNAME",
         "EVTYPE", "PROPDMG", "PROPDMGEXP")]

#        BGN_DATE  BGN_TIME END_DATE STATE          COUNTYNAME  EVTYPE PROPDMG PROPDMGEXP
#201982 10/6/1994      1814             IA SHELBY AND CRAWFORD TORNADO      50          0
```

From NOAA link,

- Select State/Area = "Iowa" (IA)
- Select County = "All"
- Select Begin Date = End Date = "10/06/1994"

- Select Event Type = "Tornado"
- Click "Search"

Result:

```
"Manilla to, SHELBY AND CRAWFORD CO., IA, 10/06/1994, 18:14, Tornado, 0.50K, 3.00K"
```

Found, PrD (property damage) = 0.50K = 500,
While PROPDMG = 50,
Conclusion: (exp 0) is == (10)

Overall conclusion for all numeric, exp 0, 1, 2, 3, 4, 5, 6, 7, 8, **they are multiplier of 10**.

---

(2) For (+),

```
plus <- data[data$PROPDMGEXP == "+",]
plus[plus$EVTYPE == "TORNADO", c("BGN_DATE", "END_DATE", "STATE", "EVTYPE", "PROPDMG", "PROPDMGEXP")]

#       BGN_DATE  END_DATE  STATE  EVTYPE  PROPDMG PROPDMGEXP
#216802 6/5/1995  6/5/1995     NV TORNADO      60          +
```

From NOAA link,

- Select State/Area = "Nevada" (NV)
- Select County = "All"
- Select Begin Date = End Date = "06/05/1995"
- Select Event Type = "Tornado"
- Click "Search"

Found, PrD (property damage) = 0.06K = 60.
So it's consistent with the StormData.csv, PROPDMG=60.
Conclusion: (+) is == multiplier of (1)

---

(3) For (-),

The same way, there is only one data,

```
minus <- data[data$CROPDMGEXP == "-",]
dim(minus)
# [1]  0 37

minus <- data[data$PROPDMGEXP == "-",]
dim(minus)
# [1]  1 37

minus[minus$EVTYPE == "HIGH WIND", c("BGN_DATE", "END_DATE", "STATE",
                      "EVTYPE", "PROPDMG", "PROPDMGEXP")]

#         BGN_DATE    END_DATE  STATE    EVTYPE PROPDMG PROPDMGEXP
#229327 12/12/1995  12/12/1995     OR HIGH WIND      15          -
```

But, searching on the NOAA link, on the same date period, there is no data (null).
Temporary conclusion: Rows with ***DMGEXP == (-) are omitted, or multiplier of 0.

---

(4) For (?)

All CROPDMG and PROPDMG values == 0, so it doesn't matter whatever our choice.

(5) For () blank character and CROPDMG != 0,

```
empty <- data[(data$CROPDMGEXP == "") & (data$CROPDMG != 0),]
empty[empty$EVTYPE == "HAIL",
      c("BGN_DATE", "BGN_TIME", "END_DATE", "STATE", "COUNTYNAME",
        "EVTYPE", "CROPDMG", "CROPDMGEXP")]

#       BGN_DATE  BGN_TIME END_DATE STATE COUNTYNAME EVTYPE CROPDMG CROPDMGEXP
#221857 7/4/1994      0400             ND   STUTSMAN   HAIL       3
```

From NOAA link,

- Select State/Area = "North Dakota", (ND)
- Select County = "All"
- Select Begin Date = End Date = "07/04/1994"
- Select Event Type = "Hail"
- Click "Search"

Found, for the County=Stutsman and Time=04:00, CrD (crop damage) = 0.00
Conclusion: empty-character () is == multiplier of 0.

More accurate analysis done by Eddie Song, which is basically agree with the comparison steps above. His post is quoted here with a bit of modification, credit goes to Eddie Song,

The version history of the database:
Version 2.0 (Apr 3, 2012): Major architecture and system changes required the complete rebuild of the Storm Events Database in 2012. A new, stricter database structure is now used, but significant work is needed to reload the original formats of data and fix data inconsistencies.

While the StormEvent.csv data ends in November 2011. The search above is pulling from a database newer than downloaded database (1950-2011). This is newest data for 1995 after major update with complete data cleaning, from http://www.ncdc.noaa.gov/stormevents/ftp.jsp (updated on 2014 Sep 16, apparently).

```
data <- read.csv(gzfile("StormEvents_details-ftp_v1.0_d1995_c20140916.csv.gz"))
```

Run through the examples in the scripts above,

```
> data[data$CZ_NAME=="SHELBY" & data$BEGIN_TIME==1750,]$DAMAGE_PROPERTY
[1] 203

> data[data$CZ_NAME=="MONROE" & data$BEGIN_TIME==1137,]$DAMAGE_PROPERTY
[1] 885

##The third one omitted because I didn't download the 1994 data, which only had one malformed record

> data[data$BEGIN_YEARMONTH==199506 & data$BEGIN_DAY==5 & data$STATE=="NEVADA",]$DAMAGE_PROPERTY
[1] 60

> data[data$BEGIN_YEARMONTH==199512 & data$BEGIN_DAY==12 & data$STATE=="OREGON",]$DAMAGE_PROPERTY
factor(0) ##no result
```

In these new tables (post-2012), it looks like PROPDMGEXP is gone, and DAMAGE_PROPERTY contains the data that was previously separated into PROPDMG and PROPDMGEXP.

```
&gt; head(data$DAMAGE_PROPERTY)
[1] 0   1K  .5K 0   .5K 2K
```

That explains why "20","3" is now corrected into "203", and so on. Looks like the "+" and "-" were entered by under-trained staff to indicate that the damage is more than x or less than x.

So the conclusion is that the malformed rows are indeed caused by improper handling, and they were fixed during the 2012 update. My approach was to **ignore** all of them because these amounts are negligible when we are talking about sums that go into the billions.

Note: Looks like all data with EXP numeric values only happened in the range of 1994-1995 period which is very small compared to overall data.
