Tidy_Tuesday_Oct22

Set up working space

```
rm(list = ls())
```

```
library(ggplot2)
library(tidyr)
```

Get the data

```
horror_movies <- readr::read_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/
```

```
## Parsed with column specification:
## cols(
##   title = col_character(),
##   genres = col_character(),
##   release_date = col_character(),
##   release_country = col_character(),
##   movie_rating = col_character(),
##   review_rating = col_double(),
##   movie_run_time = col_character(),
##   plot = col_character(),
##   cast = col_character(),
##   language = col_character(),
##   filming_locations = col_character(),
##   budget = col_character()
## )
```

```
df<-horror_movies
```

```
head(df)
```

```
## # A tibble: 6 x 12
##   title genres release_date release_country movie_rating review_rating
##   <chr> <chr>  <chr>        <chr>           <chr>                <dbl>
## 1 Gut ~ Drama~ 26-Oct-12    USA             <NA>                   3.9
## 2 The ~ Horror 13-Jan-17    USA             <NA>                  NA
## 3 Slee~ Horror 21-Oct-17    Canada          <NA>                  NA
## 4 Trea~ Comed~ 23-Apr-13    USA             NOT RATED              3.7
## 5 Infi~ Crime~ 10-Apr-15    USA             <NA>                   5.8
## 6 In E~ Horro~ 2017         UK              <NA>                  NA
## # ... with 6 more variables: movie_run_time <chr>, plot <chr>, cast <chr>,
## #   language <chr>, filming_locations <chr>, budget <chr>
```

Does review rating correpond to budget spent?

data wrangling

```
# subset data so only get values that we have
df.sub<-df[complete.cases(df$review_rating),]
df.sub<-df.sub[complete.cases(df.sub$budget),]
head(df.sub)
```

```
## # A tibble: 6 x 12
##   title genres release_date release_country movie_rating review_rating
##   <chr> <chr>  <chr>        <chr>           <chr>                <dbl>
```

```
## 1 Rise~ Adven~ 1-May-12     USA             NOT RATED              3.6
## 2 Sexy~ Drama~ 21-Mar-17    USA             <NA>                   5.9
## 3 Circ~ Actio~ 13-Jan-17    USA             <NA>                   6
## 4 Zomb~ Horror 23-Mar-15    UK              NOT RATED              2.7
## 5 Devi~ Horror 16-Sep-14    UK              <NA>                   3.4
## 6 Befo~ Horror 8-Jun-13     Japan           NOT RATED              4.7
## # ... with 6 more variables: movie_run_time <chr>, plot <chr>, cast <chr>,
## #   language <chr>, filming_locations <chr>, budget <chr>
```

```
#budgets are in dolars, euros and pounds, lets only look at values in $
dfsub2<-df.sub[grep("\\$",df.sub$budget),]
head(dfsub2); dim(dfsub2) # 847 movies
```

```
## # A tibble: 6 x 12
##   title genres release_date release_country movie_rating review_rating
##   <chr> <chr>  <chr>        <chr>           <chr>                <dbl>
## 1 Rise~ Adven~ 1-May-12     USA             NOT RATED              3.6
## 2 Circ~ Actio~ 13-Jan-17    USA             <NA>                   6
## 3 Devi~ Horror 16-Sep-14    UK              <NA>                   3.4
## 4 Appa~ Fanta~ 5-May-15     USA             NOT RATED              4
## 5 2: V~ Horror 1-Oct-12     USA             <NA>                   4.5
## 6 Her ~ Horror 19-Apr-13    USA             NOT RATED              5.4
## # ... with 6 more variables: movie_run_time <chr>, plot <chr>, cast <chr>,
## #   language <chr>, filming_locations <chr>, budget <chr>
```

```
## [1] 847  12
```

```
dfsub2$dollars<-gsub('\\$','',dfsub2$budget) # use reexpressions to reformat values
dfsub2$dollars<-as.numeric(gsub(',','',dfsub2$dollars))
range(dfsub2$dollars)
```

```
## [1] 1.0e+02 1.9e+08
```

```
#split filiming locations
head(dfsub2)
```

```
## # A tibble: 6 x 13
##   title genres release_date release_country movie_rating review_rating
##   <chr> <chr>  <chr>        <chr>           <chr>                <dbl>
## 1 Rise~ Adven~ 1-May-12     USA             NOT RATED              3.6
## 2 Circ~ Actio~ 13-Jan-17    USA             <NA>                   6
## 3 Devi~ Horror 16-Sep-14    UK              <NA>                   3.4
## 4 Appa~ Fanta~ 5-May-15     USA             NOT RATED              4
## 5 2: V~ Horror 1-Oct-12     USA             <NA>                   4.5
## 6 Her ~ Horror 19-Apr-13    USA             NOT RATED              5.4
## # ... with 7 more variables: movie_run_time <chr>, plot <chr>, cast <chr>,
## #   language <chr>, filming_locations <chr>, budget <chr>, dollars <dbl>
```

Visualize results

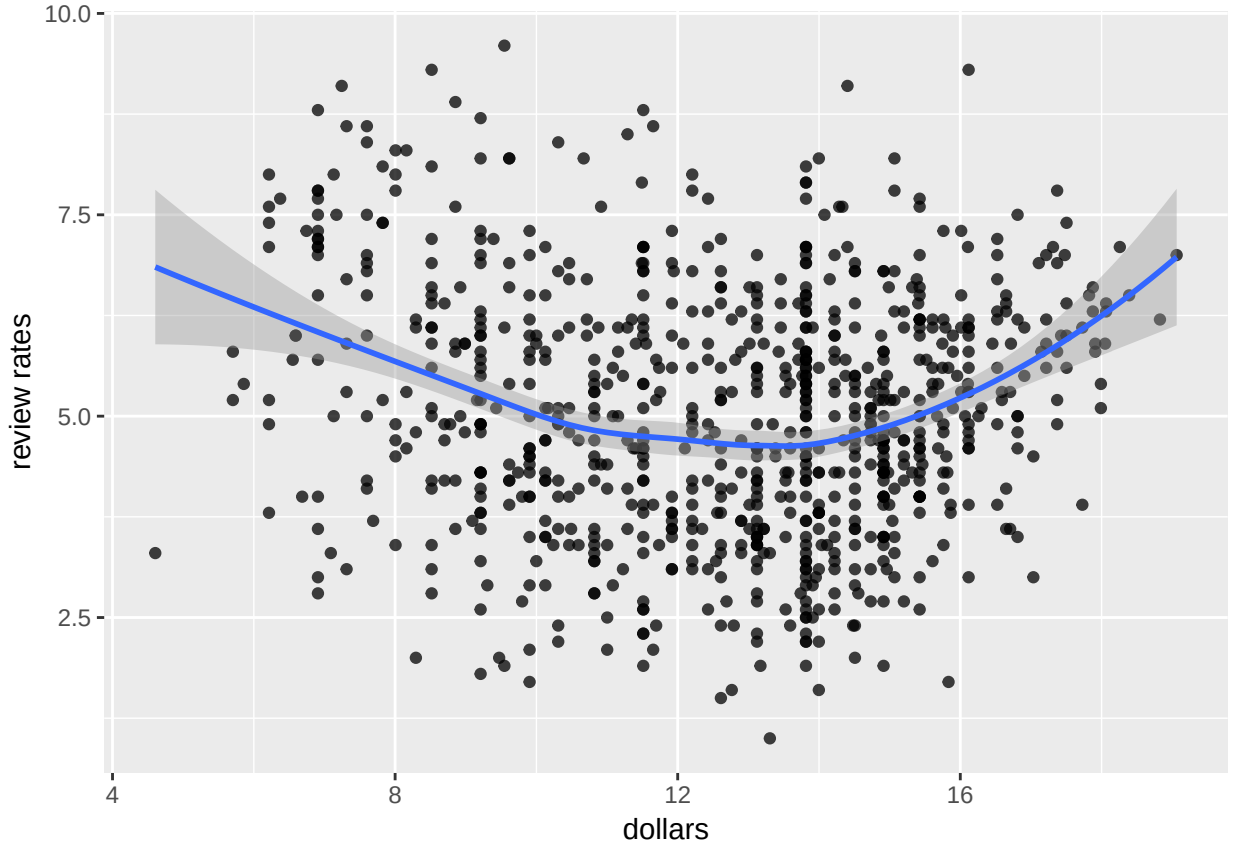
```
head(dfsub2)
```

```
## # A tibble: 6 x 13
##   title genres release_date release_country movie_rating review_rating
##   <chr> <chr>  <chr>        <chr>           <chr>                <dbl>
## 1 Rise~ Adven~ 1-May-12     USA             NOT RATED              3.6
```

```
## 2 Circ~ Actio~ 13-Jan-17    USA             <NA>                  6
## 3 Devi~ Horror 16-Sep-14    UK              <NA>                  3.4
## 4 Appa~ Fanta~ 5-May-15     USA             NOT RATED             4
## 5 2: V~ Horror 1-Oct-12     USA             <NA>                  4.5
## 6 Her ~ Horror 19-Apr-13    USA             NOT RATED             5.4
## # ... with 7 more variables: movie_run_time <chr>, plot <chr>, cast <chr>,
## #   language <chr>, filming_locations <chr>, budget <chr>, dollars <dbl>
```

```
ggplot(dfsub2, aes(x=log(dollars), y=review_rating)) + geom_point(alpha=0.75) + geom_smooth(method = 'a
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
ggsave('my_first_tidyTuesday.pdf')
```

```
## Saving 6.5 x 4.5 in image
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```