Data analysis, predict star ratings from user reviews for yelp dataset

Dimple Sharma

Friday, December 12, 2014

Load and analyze yelp business dataset

Here we load and analyze yelp business dataset

```
#setwd("C:/Dimple/RStats/")
#rm(list=ls())

library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.1.1
```

```
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.1.1
```

```
# Load the yelp business dataset from a csv file
business_data_all <- read.csv("datayelp/yelp_academic_dataset_business.csv",
                    header = T, stringsAsFactors = F, na.strings = c("","NA"))

# Analyze the yelp business dataset
dim(business_data_all)
```

```
## [1] 42153   105
```

```
colnames(business_data_all)
```

```
##   [1] "attributes.Ambience.divey"
##   [2] "attributes.Dietary.Restrictions.vegan"
##   [3] "attributes.Happy.Hour"
##   [4] "hours.Thursday.open"
```

```
##   [5] "attributes.Order.at.Counter"
##   [6] "attributes.Hair.Types.Specialized.In.africanamerican"
##   [7] "attributes.Hair.Types.Specialized.In.kids"
##   [8] "attributes.BYOB"
##   [9] "hours.Friday.open"
##  [10] "categories"
##  [11] "latitude"
##  [12] "attributes.Outdoor.Seating"
##  [13] "attributes.Alcohol"
##  [14] "attributes.Ambience.classy"
##  [15] "attributes.Payment.Types.mastercard"
##  [16] "attributes.Parking.lot"
##  [17] "business_id"
##  [18] "attributes.Ambience.touristy"
##  [19] "attributes.Corkage"
##  [20] "hours.Tuesday.open"
##  [21] "attributes.Good.For.brunch"
##  [22] "attributes.Payment.Types.amex"
##  [23] "name"
##  [24] "hours.Monday.open"
##  [25] "attributes.Waiter.Service"
##  [26] "attributes.Parking.street"
##  [27] "attributes.Ambience.hipster"
##  [28] "attributes.BYOB.Corkage"
##  [29] "attributes.Hair.Types.Specialized.In.straightperms"
##  [30] "attributes.Music.live"
##  [31] "attributes.Dietary.Restrictions.dairy.free"
##  [32] "attributes.Music.background_music"
##  [33] "attributes.Price.Range"
##  [34] "attributes.Good.For.breakfast"
##  [35] "attributes.Parking.garage"
##  [36] "attributes.Music.karaoke"
##  [37] "attributes.Good.For.Dancing"
##  [38] "review_count"
##  [39] "attributes.Hair.Types.Specialized.In.asian"
##  [40] "state"
##  [41] "attributes.Accepts.Credit.Cards"
##  [42] "hours.Friday.close"
##  [43] "attributes.Good.For.lunch"
##  [44] "attributes.Good.For.Kids"
##  [45] "attributes.Parking.valet"
##  [46] "attributes.Take.out"
##  [47] "full_address"
##  [48] "hours.Thursday.close"
##  [49] "attributes.Hair.Types.Specialized.In.coloring"
##  [50] "attributes.Payment.Types.cash_only"
##  [51] "attributes.Good.For.dessert"
##  [52] "attributes.Music.video"
##  [53] "attributes.Dietary.Restrictions.halal"
##  [54] "attributes.Takes.Reservations"
##  [55] "hours.Saturday.open"
##  [56] "attributes.Ages.Allowed"
##  [57] "attributes.Ambience.trendy"
##  [58] "attributes.Delivery"
```

```
##  [59] "hours.Wednesday.close"
##  [60] "attributes.Wi.Fi"
##  [61] "open"
##  [62] "city"
##  [63] "attributes.Payment.Types.discover"
##  [64] "attributes.Wheelchair.Accessible"
##  [65] "attributes.Dietary.Restrictions.gluten.free"
##  [66] "stars"
##  [67] "attributes.Payment.Types.visa"
##  [68] "type"
##  [69] "attributes.Caters"
##  [70] "attributes.Ambience.intimate"
##  [71] "attributes.Music.playlist"
##  [72] "attributes.Good.For.latenight"
##  [73] "attributes.Good.For.dinner"
##  [74] "attributes.Coat.Check"
##  [75] "longitude"
##  [76] "hours.Monday.close"
##  [77] "attributes.Hair.Types.Specialized.In.extensions"
##  [78] "hours.Tuesday.close"
##  [79] "hours.Saturday.close"
##  [80] "attributes.Good.for.Kids"
##  [81] "attributes.Parking.validated"
##  [82] "hours.Sunday.open"
##  [83] "attributes.Accepts.Insurance"
##  [84] "attributes.Music.dj"
##  [85] "attributes.Dietary.Restrictions.soy.free"
##  [86] "attributes.Has.TV"
##  [87] "hours.Sunday.close"
##  [88] "attributes.Ambience.casual"
##  [89] "attributes.By.Appointment.Only"
##  [90] "attributes.Dietary.Restrictions.kosher"
##  [91] "attributes.Dogs.Allowed"
##  [92] "attributes.Drive.Thru"
##  [93] "attributes.Dietary.Restrictions.vegetarian"
##  [94] "hours.Wednesday.open"
##  [95] "attributes.Noise.Level"
##  [96] "attributes.Smoking"
##  [97] "attributes.Attire"
##  [98] "attributes.Hair.Types.Specialized.In.curly"
##  [99] "attributes.Good.For.Groups"
## [100] "neighborhoods"
## [101] "attributes.Open.24.Hours"
## [102] "attributes.Ambience.romantic"
## [103] "attributes.Hair.Types.Specialized.In.perms"
## [104] "attributes.Music.jukebox"
## [105] "attributes.Ambience.upscale"
```

```
# Get sample data
head(business_data_all, 3)
```

```
##   attributes.Ambience.divey attributes.Dietary.Restrictions.vegan
## 1                      <NA>                                  <NA>
## 2                     False                                  <NA>
```

```
## 3                    False                                  <NA>
##   attributes.Happy.Hour hours.Thursday.open attributes.Order.at.Counter
## 1                  <NA>               08:00                        <NA>
## 2                  <NA>                <NA>                        <NA>
## 3                  <NA>               06:00                        <NA>
##   attributes.Hair.Types.Specialized.In.africanamerican
## 1                                                 <NA>
## 2                                                 <NA>
## 3                                                 <NA>
##   attributes.Hair.Types.Specialized.In.kids attributes.BYOB
## 1                                      <NA>            <NA>
## 2                                      <NA>            <NA>
## 3                                      <NA>            <NA>
##   hours.Friday.open                                  categories latitude
## 1             08:00           [u'Doctors', u'Health & Medical']    33.50
## 2              <NA>                             [u'Restaurants']    43.24
## 3             06:00 [u'American (Traditional)', u'Restaurants']    43.25
##   attributes.Outdoor.Seating attributes.Alcohol attributes.Ambience.classy
## 1                       <NA>              <NA>                        <NA>
## 2                      False              none                       False
## 3                      False              <NA>                       False
##   attributes.Payment.Types.mastercard attributes.Parking.lot
## 1                                <NA>                  <NA>
## 2                                <NA>                  True
## 3                                <NA>                  True
##              business_id attributes.Ambience.touristy attributes.Corkage
## 1 vcNAWiLM4dR7D2nwwJ7nCA                         <NA>               <NA>
## 2 JwUE5GmEO-sH1FuwJgKBlQ                        False               <NA>
## 3 uGykseHzyS5xAMWoN6YUqA                        False               <NA>
##   hours.Tuesday.open attributes.Good.For.brunch
## 1              08:00                       <NA>
## 2               <NA>                      False
## 3              06:00                       True
##   attributes.Payment.Types.amex                      name
## 1                          <NA>          Eric Goldberg, MD
## 2                          <NA>       Pine Cone Restaurant
## 3                          <NA> Deforest Family Restaurant
##   hours.Monday.open attributes.Waiter.Service attributes.Parking.street
## 1             08:00                      <NA>                       <NA>
## 2              <NA>                      True                      False
## 3             06:00                      True                      False
##   attributes.Ambience.hipster attributes.BYOB.Corkage
## 1                        <NA>                   <NA>
## 2                       False                   <NA>
## 3                       False                   <NA>
##   attributes.Hair.Types.Specialized.In.straightperms attributes.Music.live
## 1                                               <NA>                  <NA>
## 2                                               <NA>                  <NA>
## 3                                               <NA>                  <NA>
##   attributes.Dietary.Restrictions.dairy.free
## 1                                       <NA>
## 2                                       <NA>
## 3                                       <NA>
##   attributes.Music.background_music attributes.Price.Range
```

```
## 1                              <NA>                     NA
## 2                              <NA>                      1
## 3                              <NA>                      1
##   attributes.Good.For.breakfast attributes.Parking.garage
## 1                          <NA>                      <NA>
## 2                         False                     False
## 3                         False                     False
##   attributes.Music.karaoke attributes.Good.For.Dancing review_count
## 1                     <NA>                        <NA>            7
## 2                     <NA>                        <NA>           26
## 3                     <NA>                        <NA>           16
##   attributes.Hair.Types.Specialized.In.asian state
## 1                                      <NA>    AZ
## 2                                      <NA>    WI
## 3                                      <NA>    WI
##   attributes.Accepts.Credit.Cards hours.Friday.close
## 1                            <NA>              17:00
## 2                            True               <NA>
## 3                            True              22:00
##   attributes.Good.For.lunch attributes.Good.For.Kids
## 1                      <NA>                     <NA>
## 2                      True                     <NA>
## 3                     False                     <NA>
##   attributes.Parking.valet attributes.Take.out
## 1                     <NA>               <NA>
## 2                    False               True
## 3                    False               True
##                                      full_address hours.Thursday.close
## 1 4840 E Indian School Rd\nSte 101\nPhoenix, AZ 85018               17:00
## 2             6162 US Highway 51\nDe Forest, WI 53532                <NA>
## 3                 505 W North St\nDe Forest, WI 53532               22:00
##   attributes.Hair.Types.Specialized.In.coloring
## 1                                          <NA>
## 2                                          <NA>
## 3                                          <NA>
##   attributes.Payment.Types.cash_only attributes.Good.For.dessert
## 1                               <NA>                       <NA>
## 2                               <NA>                      False
## 3                               <NA>                      False
##   attributes.Music.video attributes.Dietary.Restrictions.halal
## 1                   <NA>                                <NA>
## 2                   <NA>                                <NA>
## 3                   <NA>                                <NA>
##   attributes.Takes.Reservations hours.Saturday.open
## 1                          <NA>               <NA>
## 2                         False               <NA>
## 3                         False              06:00
##   attributes.Ages.Allowed attributes.Ambience.trendy attributes.Delivery
## 1                    <NA>                      <NA>                <NA>
## 2                    <NA>                     False               False
## 3                    <NA>                     False               False
##   hours.Wednesday.close attributes.Wi.Fi open      city
## 1                 17:00             <NA> True   Phoenix
## 2                  <NA>             <NA> True De Forest
```

```
## 3                 22:00             <NA> True De Forest
##   attributes.Payment.Types.discover attributes.Wheelchair.Accessible
## 1                              <NA>                              <NA>
## 2                              <NA>                              <NA>
## 3                              <NA>                              <NA>
##   attributes.Dietary.Restrictions.gluten.free stars
## 1                                        <NA>   3.5
## 2                                        <NA>   4.0
## 3                                        <NA>   4.0
##   attributes.Payment.Types.visa     type attributes.Caters
## 1                          <NA> business              <NA>
## 2                          <NA> business             False
## 3                          <NA> business             False
##   attributes.Ambience.intimate attributes.Music.playlist
## 1                         <NA>                      <NA>
## 2                        False                      <NA>
## 3                        False                      <NA>
##   attributes.Good.For.latenight attributes.Good.For.dinner
## 1                          <NA>                       <NA>
## 2                         False                      False
## 3                         False                      False
##   attributes.Coat.Check longitude hours.Monday.close
## 1                  <NA>   -111.98              17:00
## 2                  <NA>    -89.34               <NA>
## 3                  <NA>    -89.35              22:00
##   attributes.Hair.Types.Specialized.In.extensions hours.Tuesday.close
## 1                                            <NA>              17:00
## 2                                            <NA>               <NA>
## 3                                            <NA>              22:00
##   hours.Saturday.close attributes.Good.for.Kids
## 1                 <NA>                    <NA>
## 2                 <NA>                    True
## 3                22:00                    True
##   attributes.Parking.validated hours.Sunday.open
## 1                         <NA>              <NA>
## 2                        False              <NA>
## 3                        False             06:00
##   attributes.Accepts.Insurance attributes.Music.dj
## 1                         <NA>                <NA>
## 2                         <NA>                <NA>
## 3                         <NA>                <NA>
##   attributes.Dietary.Restrictions.soy.free attributes.Has.TV
## 1                                    <NA>              <NA>
## 2                                    <NA>              True
## 3                                    <NA>              True
##   hours.Sunday.close attributes.Ambience.casual
## 1               <NA>                      <NA>
## 2               <NA>                     False
## 3              21:00                      True
##   attributes.By.Appointment.Only attributes.Dietary.Restrictions.kosher
## 1                           True                                   <NA>
## 2                           <NA>                                   <NA>
## 3                           <NA>                                   <NA>
##   attributes.Dogs.Allowed attributes.Drive.Thru
```

```
## 1                    <NA>                  <NA>
## 2                    <NA>                  <NA>
## 3                    <NA>                  <NA>
##   attributes.Dietary.Restrictions.vegetarian hours.Wednesday.open
## 1                                       <NA>                08:00
## 2                                       <NA>                 <NA>
## 3                                       <NA>                06:00
##   attributes.Noise.Level attributes.Smoking attributes.Attire
## 1                   <NA>              <NA>              <NA>
## 2                average              <NA>            casual
## 3                  quiet              <NA>            casual
##   attributes.Hair.Types.Specialized.In.curly attributes.Good.For.Groups
## 1                                       <NA>                       <NA>
## 2                                       <NA>                       True
## 3                                       <NA>                       True
##   neighborhoods attributes.Open.24.Hours attributes.Ambience.romantic
## 1            []                    <NA>                         <NA>
## 2            []                    <NA>                        False
## 3            []                    <NA>                        False
##   attributes.Hair.Types.Specialized.In.perms attributes.Music.jukebox
## 1                                       <NA>                     <NA>
## 2                                       <NA>                     <NA>
## 3                                       <NA>                     <NA>
##   attributes.Ambience.upscale
## 1                        <NA>
## 2                       False
## 3                       False
```

```
tail(business_data_all, 3)
```

```
##       attributes.Ambience.divey attributes.Dietary.Restrictions.vegan
## 42151                      <NA>                                  <NA>
## 42152                     False                                  <NA>
## 42153                      <NA>                                  <NA>
##       attributes.Happy.Hour hours.Thursday.open
## 42151                  <NA>                <NA>
## 42152                  <NA>               11:00
## 42153                  <NA>                <NA>
##       attributes.Order.at.Counter
## 42151                        <NA>
## 42152                        <NA>
## 42153                        <NA>
##       attributes.Hair.Types.Specialized.In.africanamerican
## 42151                                                 <NA>
## 42152                                                 <NA>
## 42153                                                 <NA>
##       attributes.Hair.Types.Specialized.In.kids attributes.BYOB
## 42151                                      <NA>            <NA>
## 42152                                      <NA>            <NA>
## 42153                                      <NA>            <NA>
##       hours.Friday.open
## 42151              <NA>
## 42152             11:00
## 42153              <NA>
```

```
##                                                categories latitude
## 42151                          [u'Yelp Events', u'Local Flavor']    33.46
## 42152         [u'Kosher', u'Italian', u'Pizza', u'Restaurants']    33.53
## 42153 [u'Food', u'Ethnic Food', u'Grocery', u'Specialty Food']    55.94
##       attributes.Outdoor.Seating attributes.Alcohol
## 42151                       <NA>               <NA>
## 42152                      False               none
## 42153                       <NA>               <NA>
##       attributes.Ambience.classy attributes.Payment.Types.mastercard
## 42151                       <NA>                                <NA>
## 42152                      False                                <NA>
## 42153                       <NA>                                <NA>
##       attributes.Parking.lot           business_id
## 42151                   <NA> nYer89hXYAoddMEKTxw7kA
## 42152                   True BMjggIgOghBMEXPo8q7q3w
## 42153                  False BVxlrYWgmi-8TPGMe6CTpg
##       attributes.Ambience.touristy attributes.Corkage hours.Tuesday.open
## 42151                         <NA>               <NA>               <NA>
## 42152                        False               <NA>              11:00
## 42153                         <NA>               <NA>               <NA>
##       attributes.Good.For.brunch attributes.Payment.Types.amex
## 42151                       <NA>                          <NA>
## 42152                      False                          <NA>
## 42153                       <NA>                          <NA>
##                                    name hours.Monday.open
## 42151 Yelp's Secret Cinema: Made In Arizona             <NA>
## 42152       LaBella Pizzeria and Restaurant            11:00
## 42153                  Oriental Supermarket             <NA>
##       attributes.Waiter.Service attributes.Parking.street
## 42151                      <NA>                      <NA>
## 42152                      True                     False
## 42153                      <NA>                     False
##       attributes.Ambience.hipster attributes.BYOB.Corkage
## 42151                        <NA>                   <NA>
## 42152                       False                   <NA>
## 42153                        <NA>                   <NA>
##       attributes.Hair.Types.Specialized.In.straightperms
## 42151                                               <NA>
## 42152                                               <NA>
## 42153                                               <NA>
##       attributes.Music.live attributes.Dietary.Restrictions.dairy.free
## 42151                  <NA>                                       <NA>
## 42152                  <NA>                                       <NA>
## 42153                  <NA>                                       <NA>
##       attributes.Music.background_music attributes.Price.Range
## 42151                              <NA>                    NA
## 42152                              <NA>                     2
## 42153                              <NA>                     2
##       attributes.Good.For.breakfast attributes.Parking.garage
## 42151                          <NA>                      <NA>
## 42152                         False                     False
## 42153                          <NA>                     False
##       attributes.Music.karaoke attributes.Good.For.Dancing review_count
## 42151                     <NA>                        <NA>           17
```

```
## 42152                    <NA>                        <NA>            5
## 42153                    <NA>                        <NA>            7
##       attributes.Hair.Types.Specialized.In.asian state
## 42151                                      <NA>    AZ
## 42152                                      <NA>    AZ
## 42153                                      <NA>   EDH
##       attributes.Accepts.Credit.Cards hours.Friday.close
## 42151                            <NA>              <NA>
## 42152                            True             15:00
## 42153                            True              <NA>
##       attributes.Good.For.lunch attributes.Good.For.Kids
## 42151                      <NA>                     <NA>
## 42152                     False                     <NA>
## 42153                      <NA>                     <NA>
##       attributes.Parking.valet attributes.Take.out
## 42151                     <NA>               <NA>
## 42152                    False               True
## 42153                    False               <NA>
##                                      full_address hours.Thursday.close
## 42151       FilmBar\n815 N 2nd St\nPhoenix, AZ 85004                <NA>
## 42152               6505 N 7th St\nPhoenix, AZ 85014               21:00
## 42153 125 Lauriston Pl\nTollcross\nEdinburgh EH3 9JN                <NA>
##       attributes.Hair.Types.Specialized.In.coloring
## 42151                                          <NA>
## 42152                                          <NA>
## 42153                                          <NA>
##       attributes.Payment.Types.cash_only attributes.Good.For.dessert
## 42151                               <NA>                        <NA>
## 42152                               <NA>                       False
## 42153                               <NA>                        <NA>
##       attributes.Music.video attributes.Dietary.Restrictions.halal
## 42151                   <NA>                                 <NA>
## 42152                   <NA>                                 <NA>
## 42153                   <NA>                                 <NA>
##       attributes.Takes.Reservations hours.Saturday.open
## 42151                          <NA>               <NA>
## 42152                          True              21:30
## 42153                          <NA>               <NA>
##       attributes.Ages.Allowed attributes.Ambience.trendy
## 42151                    <NA>                       <NA>
## 42152                    <NA>                      False
## 42153                    <NA>                       <NA>
##       attributes.Delivery hours.Wednesday.close attributes.Wi.Fi open
## 42151                <NA>                  <NA>             <NA> True
## 42152                True                 21:00               no True
## 42153                <NA>                  <NA>             <NA> True
##           city attributes.Payment.Types.discover
## 42151   Phoenix                             <NA>
## 42152   Phoenix                             <NA>
## 42153 Edinburgh                             <NA>
##       attributes.Wheelchair.Accessible
## 42151                             <NA>
## 42152                             <NA>
## 42153                             <NA>
```

```
##       attributes.Dietary.Restrictions.gluten.free stars
## 42151                                        <NA>   5.0
## 42152                                        <NA>   5.0
## 42153                                        <NA>   3.5
##       attributes.Payment.Types.visa     type attributes.Caters
## 42151                          <NA> business              <NA>
## 42152                          <NA> business              True
## 42153                          <NA> business              <NA>
##       attributes.Ambience.intimate attributes.Music.playlist
## 42151                         <NA>                      <NA>
## 42152                        False                      <NA>
## 42153                         <NA>                      <NA>
##       attributes.Good.For.latenight attributes.Good.For.dinner
## 42151                          <NA>                       <NA>
## 42152                         False                      False
## 42153                          <NA>                       <NA>
##       attributes.Coat.Check longitude hours.Monday.close
## 42151                  <NA>  -112.071              <NA>
## 42152                  <NA>  -112.065             21:00
## 42153                  <NA>    -3.203              <NA>
##       attributes.Hair.Types.Specialized.In.extensions hours.Tuesday.close
## 42151                                            <NA>               <NA>
## 42152                                            <NA>              21:00
## 42153                                            <NA>               <NA>
##       hours.Saturday.close attributes.Good.for.Kids
## 42151                 <NA>                   False
## 42152                00:00                    True
## 42153                 <NA>                    <NA>
##       attributes.Parking.validated hours.Sunday.open
## 42151                         <NA>             <NA>
## 42152                         True            11:00
## 42153                        False             <NA>
##       attributes.Accepts.Insurance attributes.Music.dj
## 42151                         <NA>                <NA>
## 42152                         <NA>                <NA>
## 42153                         <NA>                <NA>
##       attributes.Dietary.Restrictions.soy.free attributes.Has.TV
## 42151                                    <NA>             <NA>
## 42152                                    <NA>             True
## 42153                                    <NA>             <NA>
##       hours.Sunday.close attributes.Ambience.casual
## 42151               <NA>                       <NA>
## 42152              21:00                      False
## 42153               <NA>                       <NA>
##       attributes.By.Appointment.Only
## 42151                           <NA>
## 42152                           <NA>
## 42153                           <NA>
##       attributes.Dietary.Restrictions.kosher attributes.Dogs.Allowed
## 42151                                  <NA>                    <NA>
## 42152                                  <NA>                   False
## 42153                                  <NA>                    <NA>
##       attributes.Drive.Thru attributes.Dietary.Restrictions.vegetarian
## 42151                  <NA>                                       <NA>
```

```
## 42152                  <NA>                                        <NA>
## 42153                  <NA>                                        <NA>
##       hours.Wednesday.open attributes.Noise.Level attributes.Smoking
## 42151                 <NA>                  <NA>               <NA>
## 42152                11:00               average               <NA>
## 42153                 <NA>                  <NA>               <NA>
##       attributes.Attire attributes.Hair.Types.Specialized.In.curly
## 42151              <NA>                                       <NA>
## 42152            casual                                       <NA>
## 42153              <NA>                                       <NA>
##       attributes.Good.For.Groups               neighborhoods
## 42151                       <NA>                          []
## 42152                       True                          []
## 42153                       <NA> [u'Tollcross', u'Old Town']
##       attributes.Open.24.Hours attributes.Ambience.romantic
## 42151                     <NA>                         <NA>
## 42152                     <NA>                        False
## 42153                     <NA>                         <NA>
##       attributes.Hair.Types.Specialized.In.perms attributes.Music.jukebox
## 42151                                       <NA>                     <NA>
## 42152                                       <NA>                     <NA>
## 42153                                       <NA>                     <NA>
##       attributes.Ambience.upscale
## 42151                        <NA>
## 42152                       False
## 42153                        <NA>
```

```
summary(business_data_all)
```

```
##  attributes.Ambience.divey attributes.Dietary.Restrictions.vegan
##  Length:42153              Length:42153
##  Class :character          Class :character
##  Mode  :character          Mode  :character
##
##
##
##
##  attributes.Happy.Hour hours.Thursday.open attributes.Order.at.Counter
##  Length:42153          Length:42153        Length:42153
##  Class :character      Class :character    Class :character
##  Mode  :character      Mode  :character    Mode  :character
##
##
##
##
##  attributes.Hair.Types.Specialized.In.africanamerican
##  Length:42153
##  Class :character
##  Mode  :character
##
##
##
##
##  attributes.Hair.Types.Specialized.In.kids attributes.BYOB
```

```
##  Length:42153                            Length:42153
##  Class :character                        Class :character
##  Mode  :character                        Mode  :character
##
##
##
##
##  hours.Friday.open   categories           latitude
##  Length:42153       Length:42153       Min.   :32.9
##  Class :character   Class :character   1st Qu.:33.5
##  Mode  :character   Mode  :character   Median :33.7
##                                        Mean   :36.5
##                                        3rd Qu.:36.1
##                                        Max.   :56.0
##
##  attributes.Outdoor.Seating attributes.Alcohol attributes.Ambience.classy
##  Length:42153               Length:42153       Length:42153
##  Class :character           Class :character   Class :character
##  Mode  :character           Mode  :character   Mode  :character
##
##
##
##
##  attributes.Payment.Types.mastercard attributes.Parking.lot
##  Length:42153                        Length:42153
##  Class :character                    Class :character
##  Mode  :character                    Mode  :character
##
##
##
##
##  business_id        attributes.Ambience.touristy attributes.Corkage
##  Length:42153       Length:42153                 Length:42153
##  Class :character   Class :character             Class :character
##  Mode  :character   Mode  :character             Mode  :character
##
##
##
##
##  hours.Tuesday.open attributes.Good.For.brunch
##  Length:42153       Length:42153
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
##  attributes.Payment.Types.amex     name           hours.Monday.open
##  Length:42153                  Length:42153       Length:42153
##  Class :character              Class :character   Class :character
##  Mode  :character              Mode  :character   Mode  :character
##
##
##
```

```
##
##  attributes.Waiter.Service attributes.Parking.street
##  Length:42153              Length:42153
##  Class :character          Class :character
##  Mode  :character          Mode  :character
##
##
##
##
##  attributes.Ambience.hipster attributes.BYOB.Corkage
##  Length:42153                Length:42153
##  Class :character            Class :character
##  Mode  :character            Mode  :character
##
##
##
##
##  attributes.Hair.Types.Specialized.In.straightperms attributes.Music.live
##  Length:42153                                       Length:42153
##  Class :character                                   Class :character
##  Mode  :character                                   Mode  :character
##
##
##
##
##  attributes.Dietary.Restrictions.dairy.free
##  Length:42153
##  Class :character
##  Mode  :character
##
##
##
##
##  attributes.Music.background_music attributes.Price.Range
##  Length:42153                      Min.   :1
##  Class :character                  1st Qu.:1
##  Mode  :character                  Median :2
##                                    Mean   :2
##                                    3rd Qu.:2
##                                    Max.   :4
##                                    NA's   :14196
##  attributes.Good.For.breakfast attributes.Parking.garage
##  Length:42153                  Length:42153
##  Class :character              Class :character
##  Mode  :character              Mode  :character
##
##
##
##
##  attributes.Music.karaoke attributes.Good.For.Dancing  review_count
##  Length:42153             Length:42153                Min.   :   3
##  Class :character         Class :character            1st Qu.:   4
##  Mode  :character         Mode  :character            Median :   8
##                                                       Mean   :  29
```

```
##                                                        3rd Qu.:  21
##                                                        Max.   :4084
##
##  attributes.Hair.Types.Specialized.In.asian    state
##  Length:42153                                Length:42153
##  Class :character                            Class :character
##  Mode  :character                            Mode  :character
##
##
##
##
##  attributes.Accepts.Credit.Cards hours.Friday.close
##  Length:42153                    Length:42153
##  Class :character                Class :character
##  Mode  :character                Mode  :character
##
##
##
##
##  attributes.Good.For.lunch attributes.Good.For.Kids
##  Length:42153              Length:42153
##  Class :character          Class :character
##  Mode  :character          Mode  :character
##
##
##
##
##  attributes.Parking.valet attributes.Take.out full_address
##  Length:42153             Length:42153        Length:42153
##  Class :character         Class :character    Class :character
##  Mode  :character         Mode  :character    Mode  :character
##
##
##
##
##  hours.Thursday.close attributes.Hair.Types.Specialized.In.coloring
##  Length:42153         Length:42153
##  Class :character     Class :character
##  Mode  :character     Mode  :character
##
##
##
##
##  attributes.Payment.Types.cash_only attributes.Good.For.dessert
##  Length:42153                       Length:42153
##  Class :character                   Class :character
##  Mode  :character                   Mode  :character
##
##
##
##
##  attributes.Music.video attributes.Dietary.Restrictions.halal
##  Length:42153           Length:42153
##  Class :character       Class :character
```

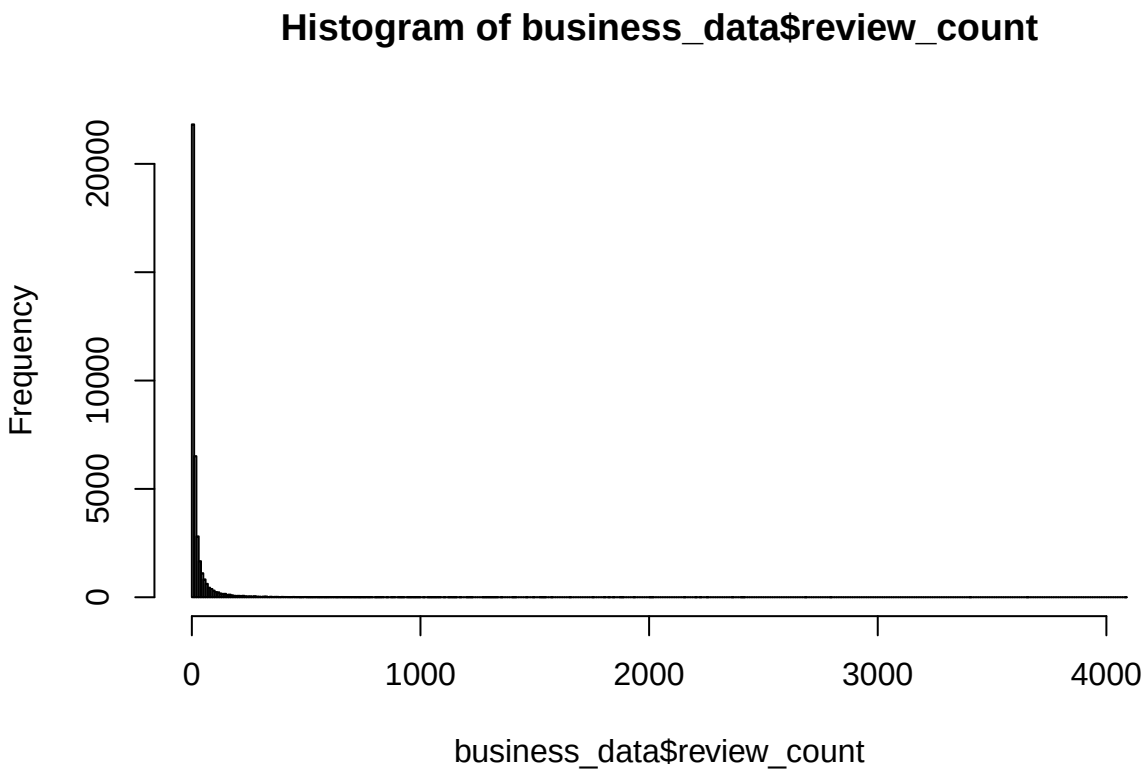
```
##  Mode  :character       Mode  :character
##
##
##
##
##  attributes.Takes.Reservations hours.Saturday.open attributes.Ages.Allowed
##  Length:42153                  Length:42153        Length:42153
##  Class :character              Class :character    Class :character
##  Mode  :character              Mode  :character    Mode  :character
##
##
##
##
##  attributes.Ambience.trendy attributes.Delivery hours.Wednesday.close
##  Length:42153               Length:42153        Length:42153
##  Class :character           Class :character    Class :character
##  Mode  :character           Mode  :character    Mode  :character
##
##
##
##
##  attributes.Wi.Fi       open               city
##  Length:42153       Length:42153       Length:42153
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  attributes.Payment.Types.discover attributes.Wheelchair.Accessible
##  Length:42153                      Length:42153
##  Class :character                  Class :character
##  Mode  :character                  Mode  :character
##
##
##
##
##  attributes.Dietary.Restrictions.gluten.free     stars
##  Length:42153                                   Min.   :1.00
##  Class :character                               1st Qu.:3.00
##  Mode  :character                               Median :3.50
##                                                 Mean   :3.67
##                                                 3rd Qu.:4.50
##                                                 Max.   :5.00
##
##  attributes.Payment.Types.visa     type           attributes.Caters
##  Length:42153                  Length:42153       Length:42153
##  Class :character              Class :character   Class :character
##  Mode  :character              Mode  :character   Mode  :character
##
##
##
##
##  attributes.Ambience.intimate attributes.Music.playlist
```

```
##  Length:42153                Length:42153
##  Class :character            Class :character
##  Mode  :character            Mode  :character
##
##
##
##
##  attributes.Good.For.latenight attributes.Good.For.dinner
##  Length:42153                  Length:42153
##  Class :character              Class :character
##  Mode  :character              Mode  :character
##
##
##
##
##  attributes.Coat.Check   longitude       hours.Monday.close
##  Length:42153          Min.   :-115.37   Length:42153
##  Class :character      1st Qu.:-115.14   Class :character
##  Mode  :character      Median :-112.07   Mode  :character
##                        Mean   :-104.09
##                        3rd Qu.:-111.88
##                        Max.   :  -3.05
##
##  attributes.Hair.Types.Specialized.In.extensions hours.Tuesday.close
##  Length:42153                                    Length:42153
##  Class :character                                Class :character
##  Mode  :character                                Mode  :character
##
##
##
##
##  hours.Saturday.close attributes.Good.for.Kids
##  Length:42153         Length:42153
##  Class :character     Class :character
##  Mode  :character     Mode  :character
##
##
##
##
##  attributes.Parking.validated hours.Sunday.open
##  Length:42153                 Length:42153
##  Class :character             Class :character
##  Mode  :character             Mode  :character
##
##
##
##
##  attributes.Accepts.Insurance attributes.Music.dj
##  Length:42153                 Length:42153
##  Class :character             Class :character
##  Mode  :character             Mode  :character
##
##
##
```

```
##
##  attributes.Dietary.Restrictions.soy.free attributes.Has.TV
##  Length:42153                             Length:42153
##  Class :character                         Class :character
##  Mode  :character                         Mode  :character
##
##
##
##
##  hours.Sunday.close attributes.Ambience.casual
##  Length:42153       Length:42153
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
##  attributes.By.Appointment.Only attributes.Dietary.Restrictions.kosher
##  Length:42153                   Length:42153
##  Class :character               Class :character
##  Mode  :character               Mode  :character
##
##
##
##
##  attributes.Dogs.Allowed attributes.Drive.Thru
##  Length:42153            Length:42153
##  Class :character        Class :character
##  Mode  :character        Mode  :character
##
##
##
##
##  attributes.Dietary.Restrictions.vegetarian hours.Wednesday.open
##  Length:42153                               Length:42153
##  Class :character                           Class :character
##  Mode  :character                           Mode  :character
##
##
##
##
##  attributes.Noise.Level attributes.Smoking attributes.Attire
##  Length:42153           Length:42153       Length:42153
##  Class :character       Class :character   Class :character
##  Mode  :character       Mode  :character   Mode  :character
##
##
##
##
##  attributes.Hair.Types.Specialized.In.curly attributes.Good.For.Groups
##  Length:42153                               Length:42153
##  Class :character                           Class :character
##  Mode  :character                           Mode  :character
##
```

```
##
##
##
##  neighborhoods      attributes.Open.24.Hours attributes.Ambience.romantic
##  Length:42153       Length:42153             Length:42153
##  Class :character   Class :character         Class :character
##  Mode  :character   Mode  :character         Mode  :character
##
##
##
##
##  attributes.Hair.Types.Specialized.In.perms attributes.Music.jukebox
##  Length:42153                               Length:42153
##  Class :character                           Class :character
##  Mode  :character                           Mode  :character
##
##
##
##
##  attributes.Ambience.upscale
##  Length:42153
##  Class :character
##  Mode  :character
##
##
##
##
```

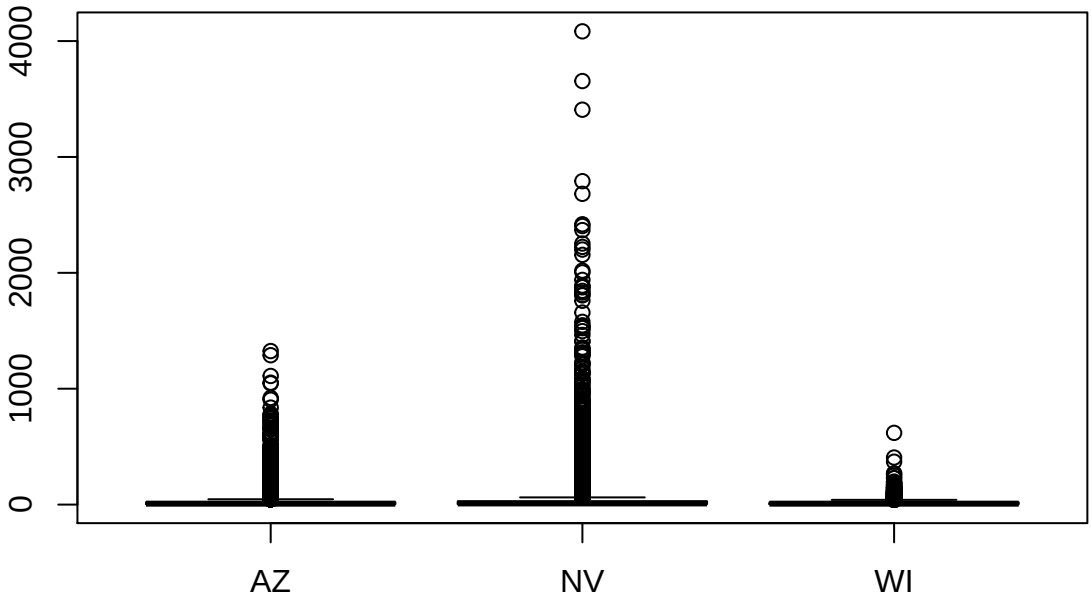
```r
# We are interested in few columns
business_colnames <- c("business_id", "name", "categories", "type","stars",
                       "review_count", "city" ,"state" ,"latitude",
                       "longitude", "neighborhoods")
# Get only the columns which will be used for analysis
business_data_all <- business_data_all[, business_colnames]
business_data <- filter(business_data_all, grepl("AZ|NV|WI", state))

# Remove large objects for memory efficiency
rm(business_data_all)

# A histogram of review counts gives the skewness of the data. Here the data is skewed the right hence
hist(business_data$review_count, breaks = 400)
```



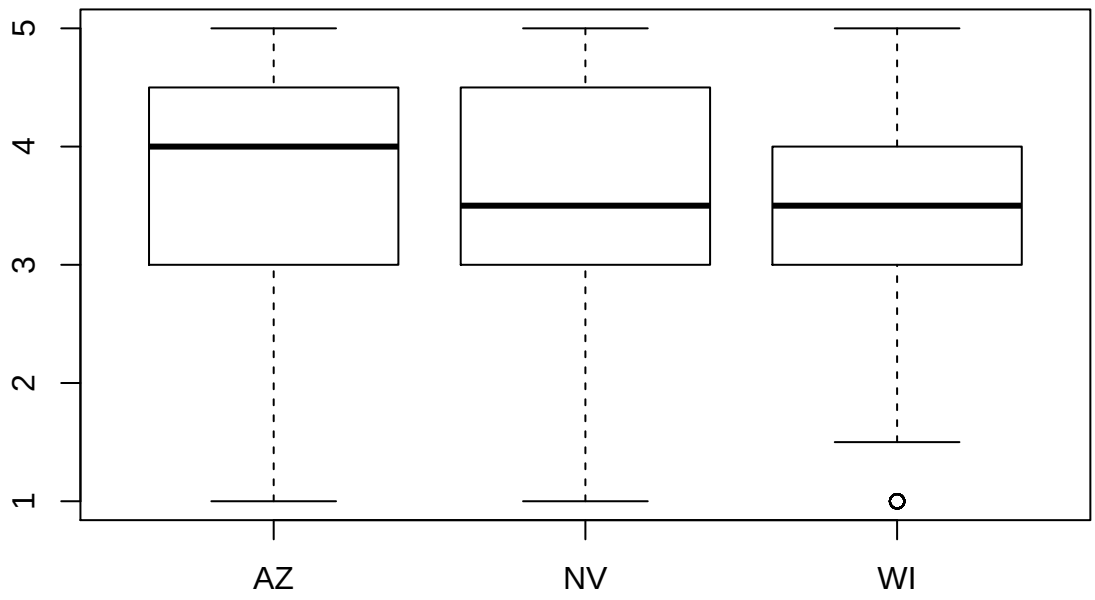

```
# Following is a choropleth graph which shows the number of review counts by state.
library(ggmap)
```

```
## Warning: package 'ggmap' was built under R version 3.1.2
```

```
# Get number of review for each state

# box plot of review count vs states
boxplot(business_data$review_count ~ business_data$state)
```

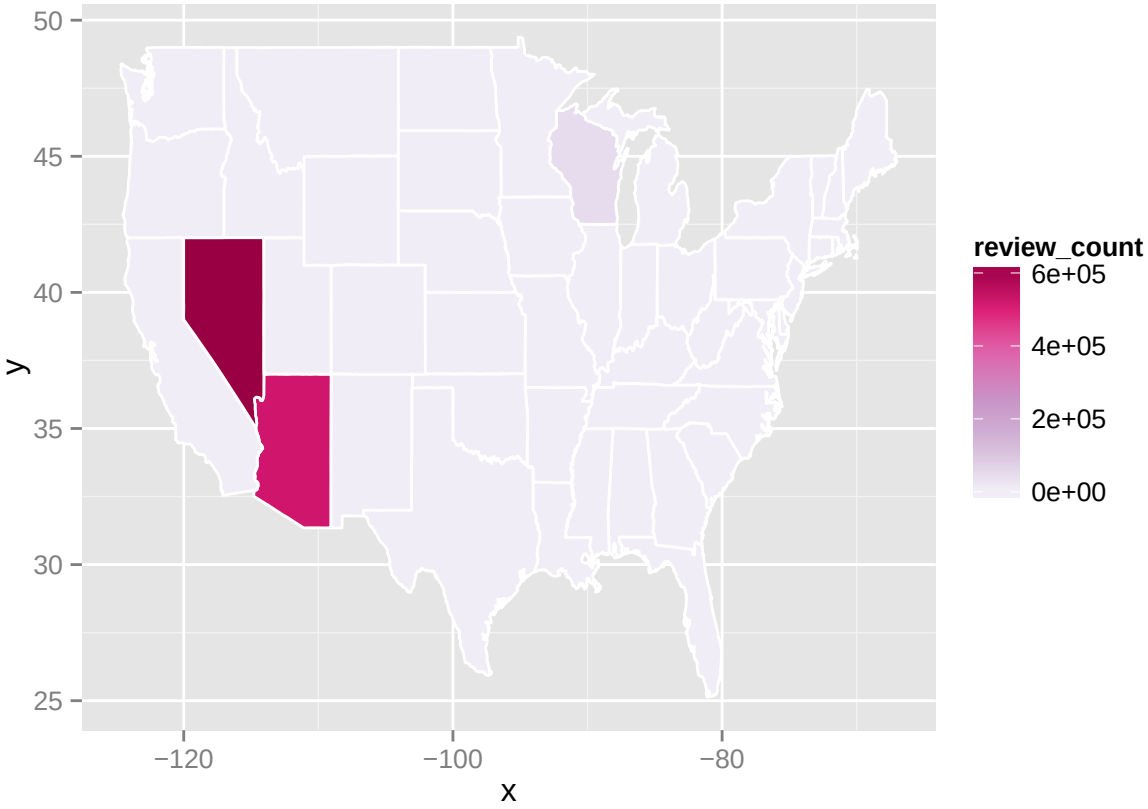



```
# box plot of stars vs states
boxplot(business_data$stars ~ business_data$state)
```

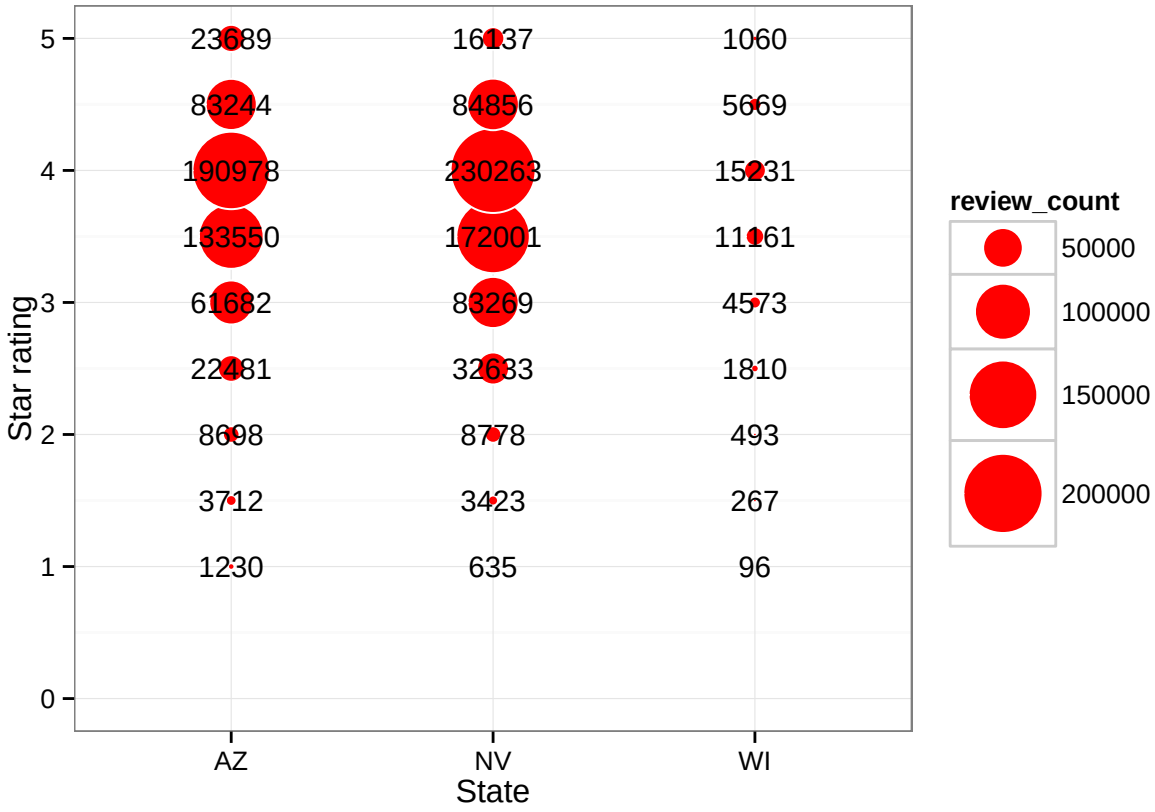



```
#Get a base map ready
world_map <- borders("world", colour = "grey10", fill = "grey10")

# 1. Maps data frame
states_map <- map_data("state")

# 2. Data frame with plotting values
#state.name <- c("Arizona", "Nevda", "Wisconson")
#review_count <- c(100,200,300)
state_name <- tolower(state.name)
state_name <- as.character(state_name)
review_count <- rep(0,50)
data <- data.frame(state_name, review_count)

# Get total review count for the three states
review_by_state <- aggregate( review_count ~ state, data = business_data, FUN = sum)

data[data$state_name %in% c("arizona"), 2] <- review_by_state[review_by_state == "AZ", 2]
data[data$state_name %in% c("nevada"), 2] <- review_by_state[review_by_state == "NV", 2]
data[data$state_name %in% c("wisconsin"), 2] <- review_by_state[review_by_state == "WI", 2]

#col = c("#EDF8B1", "#1D91C0", "#225EA8", "#5A005A")
mapcolours <- c("#F1EEF6", "#D4B9DA", "#C994C7", "#DF65B0", "#DD1C77", "#980043")
mp <- ggplot(data, aes(map_id = state_name)) + geom_map(aes(fill = review_count), color = "white", map=
mp <- mp + expand_limits(x = states_map$long, y = states_map$lat)
mp <- mp + scale_fill_gradientn(colours = mapcolours, limits=c(0,max(data$review_count)))
```

```
mp
```




```
# Bubble chart of review count, stars, and state
data_review_stars_state <- aggregate(review_count ~ state + stars, data=business_data, FUN = sum)
# Bubble chart using ggplot
gph_state_stars <- ggplot(data_review_stars_state, aes(x=state, y=stars, size=review_count, label=revie
gph_state_stars <- gph_state_stars + geom_point(colour="white", fill="red", shape=21)+ scale_size_area(
gph_state_stars <- gph_state_stars + scale_x_discrete (name="State")
gph_state_stars <- gph_state_stars + scale_y_continuous(name="Star rating", limits=c(0,max(data_review_
gph_state_stars <- gph_state_stars + geom_text(size = 4) + theme_bw()
gph_state_stars
```



Get categories of business

```
# Business categories in business dataset
business_categories <- unlist(lapply(business_data$categories, FUN = function(x)
                        strsplit(x, ",")))
names(business_categories) <- NULL

# Split categoreis and remove punctuations and blanks
business_categories <- gsub("u\'", "\'", business_categories)
business_categories <- gsub("[[:punct:]]", "", business_categories)
business_categories <- gsub("[[:blank:]]", "", business_categories)
head(business_categories, 10)
```

```
##  [1] "Doctors"              "HealthMedical"        "Restaurants"
##  [4] "AmericanTraditional"  "Restaurants"          "Food"
##  [7] "IceCreamFrozenYogurt" "FastFood"             "Restaurants"
## [10] "Chinese"
```

```
categories <- unique(business_categories)
n_cat <- length(categories)
paste("The dataset contains ", n_cat, " unique categories")
```

```
## [1] "The dataset contains  707  unique categories"
```

Doctors, HealthMedical, Restaurants, AmericanTraditional, Food, IceCreamFrozenYogurt, FastFood, Chinese, TelevisionStations, MassMedia, HomeServices, HeatingAirConditioningHVAC, Libraries, PublicServicesGovernment, Veterinarians, Pets, Bars, Nightlife, Lounges, HotelsTravel, BedBreakfast, EventPlanningServices, Hotels, ActiveLife, Bowling, Pizza, PartyEventPlanning, Caterers, AsianFusion, VenuesEventSpaces, Mexican, AutoRepair, Automotive, AutoPartsSupplies, CarDealers, ArtsEntertainment, StadiumsArenas, PetServices, Pilates, Yoga, WeightLossCenters, FitnessInstruction, Sandwiches, ChickenWings, Shopping, ArtsCrafts, CardsStationery, ArtSupplies, FlowersGifts, KnittingSupplies, HobbyShops, uWomensClothing, ShoppingCenters, Fashion, Beer, WineSpirits, CoffeeTea, Bakeries, BreakfastBrunch, AmericanNew, PetBoardingPetSitting, Diners, Dentists, GeneralDentistry, Burgers, CajunCreole, AmateurSportsTeams, HardwareStores, HomeGarden, Barbers, BeautySpas, HairSalons, DaySpas, Massage, Tires, OilChangeStations, Drugstores, ConvenienceStores, CosmeticsBeautySupply, LightingFixturesEquipment, Electronics, MusicalInstrumentsTeachers, ShoeStores, uMensClothing, Optometrists, CarWash, SkinCare, Italian, Donuts, Accessories, RealEstate, Apartments, LandscapeArchitects, NurseriesGardening, LocalServices, SewingAlterations, Breweries, ShippingCenters, PrintingServices, Notaries, Delis, VideosVideoGameRental, Books, Mags, MusicVideo, Golf, Gyms, Trainers, DepartmentStores, RaceTracks, Towing, GunsAmmo, CheeseShops, SpecialtyFood, BanksCreditUnions, FinancialServices, MortgageBrokers, MusicDVDs, uChildrensClothing, MobilePhones, ArtGalleries, SportsBars, , FabricStores, PerformingArts, BabyGearFurniture, EyewearOpticians, ShoeRepair, Seafood, DryCleaningLaundry, TobaccoShops, Lingerie, Adult, Taxis, AirportShuttles, Transportation, CarRental, Appliances, Grocery, DiveBars, Japanese, SportingGoods, Bikes, TexMex, BikeRentals, HomeDecor, Antiques, FurnitureStores, HomeCleaning, Contractors, Pubs, Buffets, Florists, RecreationCenters, Bridal, Cafes, MeatShops, KitchenBath, Lebanese, Mediterranean, MiddleEastern, NailSalons, MusicVenues, PetGroomers, PetTraining, PetStores, ThriftStores, InternetCafes, BotanicalGardens, Indian, Steakhouses, InternetServiceProviders, ProfessionalServices, Airports, Tanning, VinylRecords, LaserEyeSurgeryLasik, Barbeque, Greek, Accountants, Soup, Salad, Museums, TruckRental, SelfStorage, WeddingPlanning, SportsWear, Hiking, Parks, GasServiceStations, PayrollServices, TaxServices, Mattresses, BuildingSupplies, CarStereoInstallation, HotDogs, DiscountStore, CarpetCleaning, Caribbean, CommunityServiceNonProfit, Cinema, BodyShops, WindshieldInstallationRepair, AutoGlassServices, GlutenFree, Plumbing, AnimalShelters, DanceClubs, Thai, Midwives, Jewelry, TeaRooms, FruitsVeggies, EthnicFood, Watches, Churches, ReligiousOrganizations, ElementarySchools, MiddleSchoolsHighSchools, Education, SushiBars, Neurologist, Southern, RadioStations, SpecialtySchools, VocationalTechnicalSchool, MedicalCenters, Cheesesteaks, WineBars, Cardiologists, Hospitals, PhotographyStoresServices, GayBars, CareerCounseling, Bookstores, SoulFood, AdultEntertainment, FamilyPractice, LaserHairRemoval, MedicalSpas, HairRemoval, ObstetriciansGynecologists, AppliancesRepair, FishChips, LatinAmerican, Salvadoran, Flooring, Movers, Uniforms, Costumes, Orthopedists, Pakistani, Carpeting, Karaoke, UrgentCare, ElectronicsRepair, Zoos, LandmarksHistoricalBuildings, ChildCareDayCare, OutletStores, WholesaleStores, Preschools, NewspapersMagazines, AdultEducation, LifeCoach, CosmeticDentists, PawnShops, CookingSchools, British, HairStylists, ComfortFood, CandyStores, JuiceBarsSmoothies, Acupuncture, CollegesUniversities, Vegetarian, ComicBooks, Dermatologists, CosmeticSurgeons, TattooRemoval, Electricians, SolarInstallation, Lawyers, EstatePlanningLaw, CounselingMentalHealth, OfficeEquipment, PartySupplies, PartyEquipmentRentals, Desserts, JazzBlues, ToyStores, Used, VintageConsignment, MaternityWear, OperaBallet, Chiropractors, WindowsInstallation, Kosher, French, CarpetInstallation, Photographers, SessionPhotography, PropertyManagement, SpeechTherapists, OccupationalTherapy, PhysicalTherapy, Bagels, OutdoorGear, Computers, Polish, SportsMedicine, ITServicesComputerRepair, FormalWear, GlassMirrors, HealthMarkets, Creperies, Orthodontists, Waxing, Landscaping, Ophthalmologists, uMensHairSalons, GiftShops, GraphicDesign, Festivals, AmusementParks, AutoDetailing, OralSurgeons, Framing, LeatherGoods, SocialClubs, PoolHalls, ComedyClubs, Vegan, DanceStudios, Hawaiian, EmploymentAgencies, InternalMedicine, HotTubPool, JewelryRepair, Tours, Cuban, Allergists, Gastropubs, SecuritySystems, WatchRepair, KeysLocksmiths, SprayTanning, SkatingRinks, FurnitureReupholstery, PoolCleaners, Pediatricians, ReligiousSchools, FuneralServicesCemeteries, GarageDoorServices, PestControl, Luggage, MotorcycleRepair, MotorcycleDealers, HairExtensions, MassageTherapy, Nutritionists, WheelRimRepair, FinancialAdvising, VapeShops, Russian, EarNoseThroat, LaboratoryTesting, DiagnosticServices, CommercialRealEstate, SkiResorts, TravelServices, Insurance, HomeTheatreInstallation, GolfEquipment, Tattoo, Piercing, MakeupArtists, Paintball, BespokeClothing, Pretzels, HomeHealthCare, PediatricDentists, Irish, Wigs, SwimmingLessonsSchools, FarmersMarket, ShadesBlinds, Gymnastics, SummerCamps, BoatRepair, Boating, BoatDealers, Airlines, Firewood, Parking, PrintMedia, InsulationInstallation, HorseRacing, Gardeners, BusinessLaw,

BankruptcyLaw, GoKarts, PublicTransportation, Arcades, MiniGolf, InteriorDesign, MartialArts, Roofing, CouriersDeliveryServices, FencesGates, HorsebackRiding, LocalFlavor, Playgrounds, ShavedIce, Fondue, Radiologists, Periodontists, DJs, TicketSales, Resorts, DoItYourselfFood, Diving, Podiatrists, Urologists, WindowWashing, Vietnamese, Psychiatrists, Korean, Rheumatologists, ScubaDiving, Endodontists, Arabian, TestPreparation, RealEstateServices, AutoCustomization, GunRifleRanges, SeafoodMarkets, TreeServices, FleaMarkets, RVRental, BikeRepairMaintenance, Peruvian, DepartmentsofMotorVehicles, HomeWindowTinting, Gelato, Halal, EyelashService, SwimmingPools, RealEstateAgents, TelevisionServiceProviders, DimSum, Painters, Gastroenterologist, Signmaking, DrivingSchools, Hats, KidsActivities, Archery, MountainBiking, Propane, Handyman, SportsClubs, Tennis, Advertising, RaftingKayaking, VacationRentals, SharedOfficeSpaces, PoolHotTubService, Casinos, EmbroideryCrochet, ScreenPrintingTShirtPrinting, DiagnosticImaging, TutoringCenters, RetirementHomes, MotorcycleRental, HotAirBalloons, Officiants, ChocolatiersShops, CosmetologySchools, FoodDeliveryServices, Mongolian, NaturopathicHolistic, MassageSchools, PermanentMakeup, Marketing, CriminalDefenseLaw, DUILaw, RegistrationServices, ProfessionalSportsTeams, PersianIranian, HookahBars, Endocrinologists, Aquariums, PsychicsAstrologers, RehabilitationCenter, German, FlightInstruction, Fertility, Pharmacy, Oncologist, SmogCheckStations, OfficeCleaning, DivorceFamilyLaw, Orthotics, Filipino, Limos, Utilities, Cantonese, Taiwanese, BlowDryOutServices, RVDealers, BartendingSchools, Hostels, Swimwear, Surfing, ScreenPrinting, Climbing, HorseBoarding, ArtClasses, Audiologist, VideoFilmProduction, PublicRelations, Argentine, BuddhistTemples, NannyServices, TrophyShops, CocktailBars, PersonalShopping, HimalayanNepalese, Moroccan, Falafel, Ethiopian, African, EducationalServices, Indonesian, Turkish, Afghan, Videographers, WebDesign, CPRClasses, HerbsSpices, HomeInspectors, HealthRetreats, LaserTag, CulturalCenter, RecyclingCenter, OsteopathicPhysicians, Soccer, ArtSchools, BoatCharters, Fishing, Campgrounds, HomeOrganization, FoodStands, ChimneySweeps, AquariumServices, LeisureCenters, GeneralLitigation, Lakes, Surgeons, Wineries, Cabinetry, AuctionHouses, HinduTemples, TapasSmallPlates, TalentAgencies, Basque, Spanish, Tubing, MedicalSupplies, HearingAidProviders, RVParks, DrywallInstallationRepair, BailBondsmen, IslandPub, PersonalInjuryLaw, DanceSchools, CheckCashingPaydayLoans, TapasBars, Skydiving, Brazilian, GuestHouses, Hospice, PersonalChefs, Magicians, Laotian, Szechuan, Beaches, Belgian, DogParks, BootCamps, PianoBars, PoliceDepartments, Buses, HotPot, SkateParks, FoodTrucks, DiscGolf, LiveRawFood, PostOffices, DamageRestoration, JunkRemovalHauling, DogWalkers, Reiki, Bistros, TanningBeds, Reflexology, Synagogues, RecordingRehearsalStudios, Rugs, ImmigrationLaw, FirstAidClasses, SepticServices, Irrigation, Investing, Courthouses, TraditionalChineseMedicine, Brasseries, Boxing, SpecialEducation, DoorSalesInstallation, BubbleTea, PlusSizeFashion, Bartenders, StreetVendors, Malaysian, Singaporean, CyclingClasses, Butcher, ChampagneBars, Burmese, PersonalAssistants, Cafeteria, Scandinavian, RockClimbing, FoodCourt, ModernEuropean, BarreClasses, PhotoBoothRentals, Doulas, LactationServices, GoldBuyers, ShreddingServices, RealEstateLaw, Mosques, EventPhotography, Cupcakes, WaterDelivery, DayCamps, MobilePhoneRepair, LanguageSchools, SurfShop, SecurityServices, PartyBusRentals, SoftwareDevelopment, Ramen, Ukrainian, HomeStaging, PrivateInvestigation, Paddleboarding, BrewingSupplies, VacationRentalAgents, PressureWashers, HypnosisHypnotherapy, FencingClubs, MotorcycleGear, Shanghainese, Distilleries, Anesthesiologists, Cambodian, PetAdoption, Architects, Venezuelan, Psychologists, TrampolineParks, Matchmakers, MasonryConcrete, PrivateTutors, Colombian, DataRecovery, Musicians, MedicalTransportation, CSA, CannabisClinics, Clowns, YelpEvents, Dominican, BattingCages, BeerBar, JetSkis, ATVRentalsTours, TaiChi, CookingClasses, CountryDanceHalls, CollegeCounseling, Pulmonologist, AutoLoanProviders, Australian, Flowers, EmploymentLaw, MeditationCenters, Rolfing, UniversityHousing, ChallengeCourses, WalkinClinics, Portuguese, CarShareServices

Linear regression analysis

```
col_classes <- c(rep("NULL",3),"numeric","character",
                 rep("numeric",2), rep("NULL",2), rep("numeric",2))

review_data <- read.csv("datayelp/yelp_academic_dataset_review.csv",
                colClasses = col_classes,
```

```
               header = T, stringsAsFactors = F)

# Summerize the sentiment score and user votes in a new dataset
#
business_review <- review_data %>%
                   group_by(business_id) %>%
                   summarise( total_score = sum(scores),
                              avg_score = mean(scores),
                              min_score = min(scores),
                              max_score = max(scores),
                              total_votes_cool = sum(votes.cool),
                              total_votes_funny = sum(votes.funny),
                              total_votes_useful = sum(votes.useful))

data <- select(business_data, business_id, review_count, stars)

# Left outer join yelp business data with user reviews for business
data2 <- merge(data, business_review, by.x = "business_id", by.y = "business_id",
               all.x = TRUE)

summary(data2)
```

```
##  business_id         review_count      stars       total_score
##  Length:38882       Min.   :   3   Min.   :1.00   Min.   : -493
##  Class :character   1st Qu.:   4   1st Qu.:3.00   1st Qu.:   11
##  Mode  :character   Median :   9   Median :3.50   Median :   28
##                     Mean   :  31   Mean   :3.66   Mean   :  126
##                     3rd Qu.:  23   3rd Qu.:4.50   3rd Qu.:   82
##                     Max.   :4084   Max.   :5.00   Max.   :23056
##                                                   NA's   :22
##    avg_score        min_score        max_score     total_votes_cool
##  Min.   :-15.00   Min.   :-55.00   Min.   :-11.0   Min.   :   0.0
##  1st Qu.:  2.25   1st Qu.: -4.00   1st Qu.:  6.0   1st Qu.:   1.0
##  Median :  3.67   Median : -2.00   Median : 10.0   Median :   3.0
##  Mean   :  3.65   Mean   : -2.17   Mean   : 11.6   Mean   :  18.6
##  3rd Qu.:  5.00   3rd Qu.:  0.00   3rd Qu.: 15.0   3rd Qu.:  13.0
##  Max.   : 22.12   Max.   : 21.00   Max.   : 77.0   Max.   :2301.0
##  NA's   :22       NA's   :22       NA's   :22      NA's   :22
##  total_votes_funny total_votes_useful
##  Min.   :   0      Min.   :   0
##  1st Qu.:   0      1st Qu.:   3
##  Median :   3      Median :   8
##  Mean   :  15      Mean   :  32
##  3rd Qu.:  10      3rd Qu.:  24
##  Max.   :2633      Max.   :3580
##  NA's   :22        NA's   :22
```

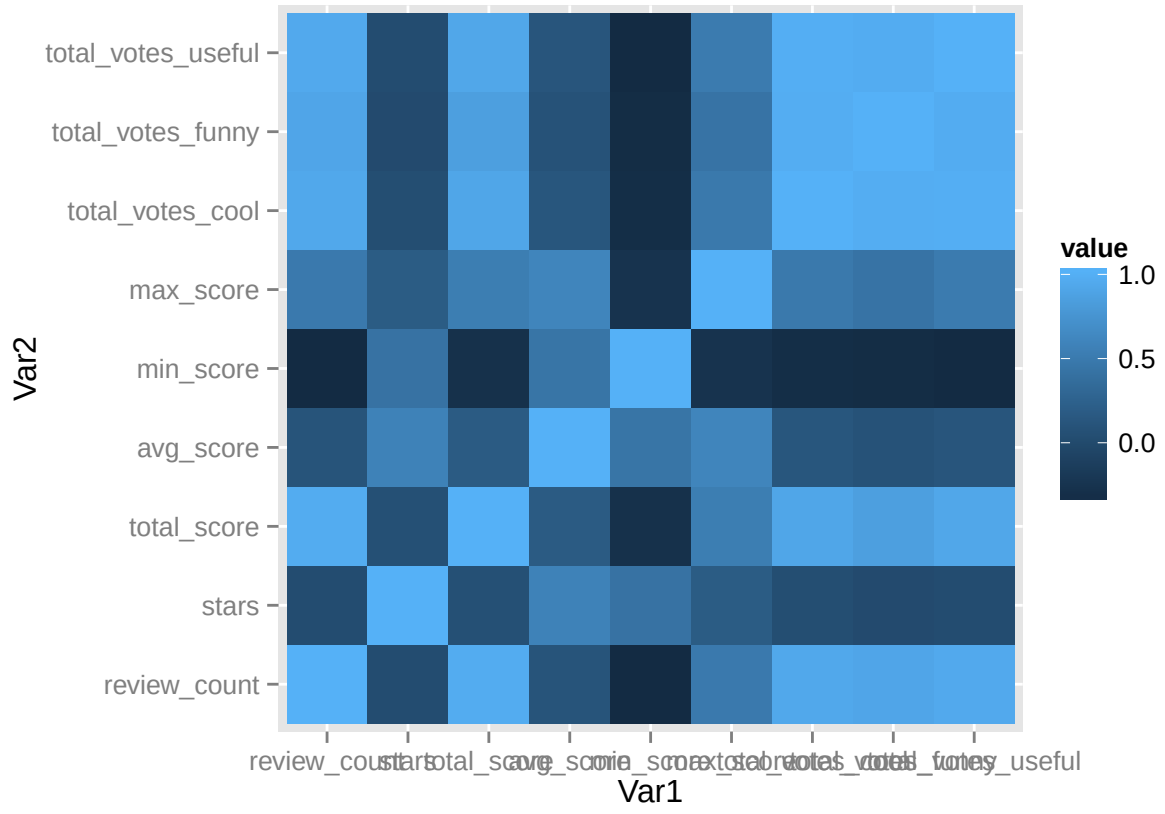
```
# After merging two dataset we can see that some business do not have
# sentiment score and votes so we will remove those business from the analysis
data_analyze <- na.omit(data2)

# Final data with sentiment score, votes, and review count.
dim(data_analyze)
```

```
## [1] 38860    10
```

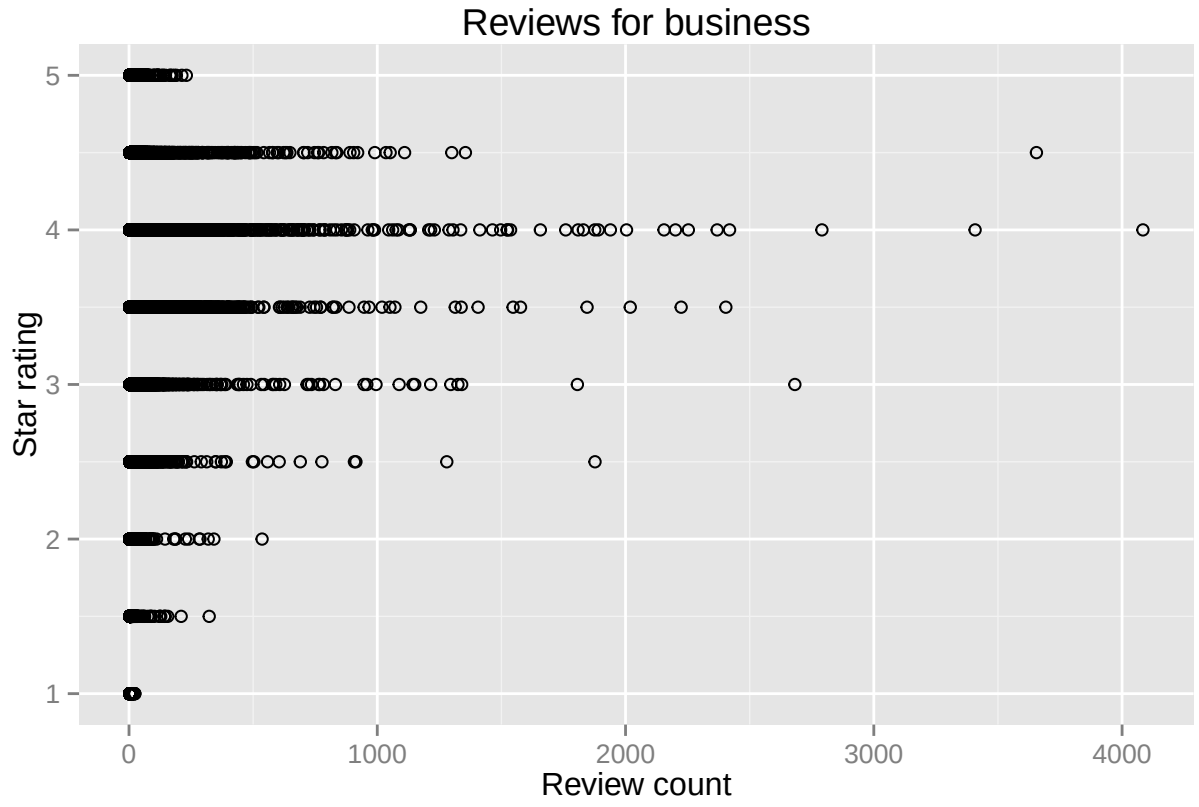
```
summary(data_analyze)
```

```
##  business_id         review_count      stars       total_score
##  Length:38860       Min.   :   3   Min.   :1.00   Min.   : -493
##  Class :character   1st Qu.:   4   1st Qu.:3.00   1st Qu.:   11
##  Mode  :character   Median :   9   Median :3.50   Median :   28
##                     Mean   :  31   Mean   :3.66   Mean   :  126
##                     3rd Qu.:  23   3rd Qu.:4.50   3rd Qu.:   82
##                     Max.   :4084   Max.   :5.00   Max.   :23056
##    avg_score        min_score        max_score     total_votes_cool
##  Min.   :-15.00   Min.   :-55.00   Min.   :-11.0   Min.   :   0.0
##  1st Qu.:  2.25   1st Qu.: -4.00   1st Qu.:  6.0   1st Qu.:   1.0
##  Median :  3.67   Median : -2.00   Median : 10.0   Median :   3.0
##  Mean   :  3.65   Mean   : -2.17   Mean   : 11.6   Mean   :  18.6
##  3rd Qu.:  5.00   3rd Qu.:  0.00   3rd Qu.: 15.0   3rd Qu.:  13.0
##  Max.   : 22.12   Max.   : 21.00   Max.   : 77.0   Max.   :2301.0
##  total_votes_funny total_votes_useful
##  Min.   :   0      Min.   :   0
##  1st Qu.:   0      1st Qu.:   3
##  Median :   3      Median :   8
##  Mean   :  15      Mean   :  32
##  3rd Qu.:  10      3rd Qu.:  24
##  Max.   :2633      Max.   :3580
```

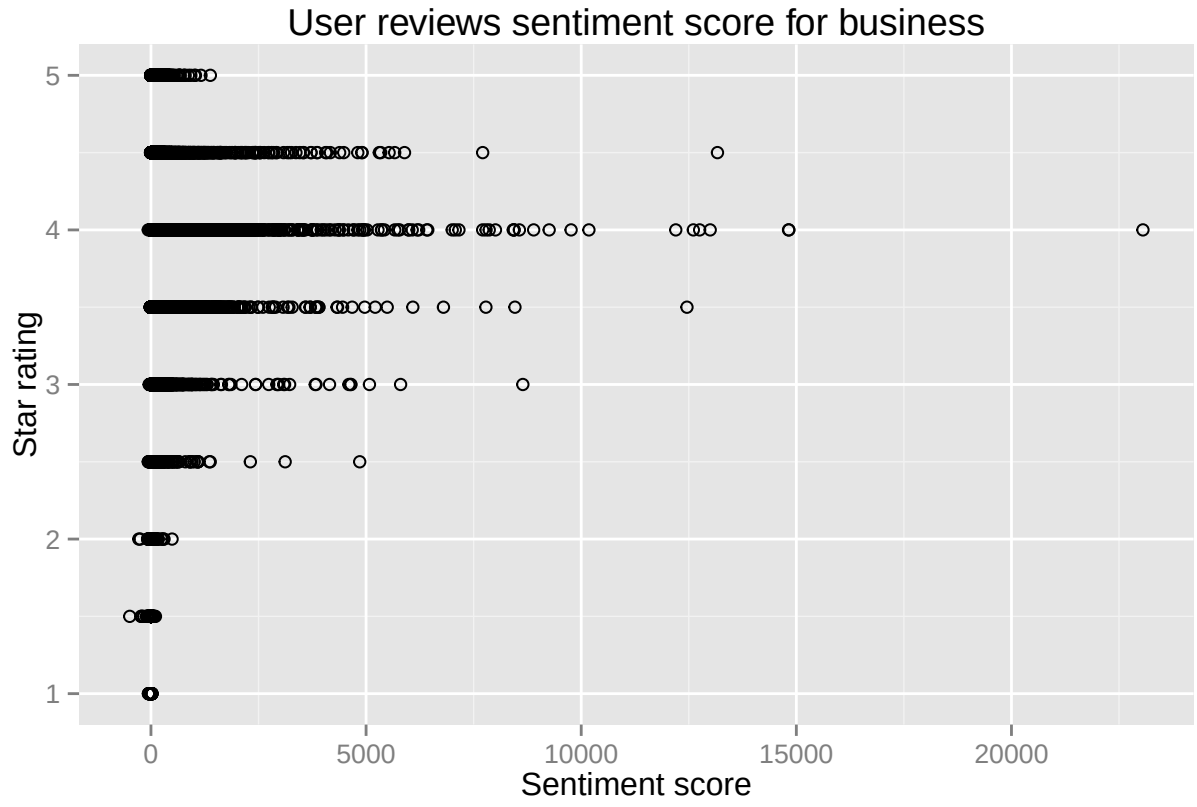
```
cor_variable <- cor(data_analyze[,-1])

library(reshape2)
library(ggplot2)
cor_data <- melt(cor_variable)

cor_heatmap <- ggplot(data = cor_data, aes(x = Var1, y = Var2, fill = value))
cor_heatmap <- cor_heatmap + geom_tile()
cor_heatmap
```




```
ggplot(data_analyze, aes(x=review_count, y=stars)) +
    geom_point(shape=1) +    # Use hollow circles
    labs(title="Reviews for business", x = "Review count", y = "Star rating")
```




```
    #geom_smooth(method=lm,   # Add linear regression line
    #            se=FALSE)

ggplot(data_analyze, aes(x=total_score, y=stars)) +
    geom_point(shape=1) +    # Use hollow circles
    labs(title="User reviews sentiment score for business", x = "Sentiment score",
         y = "Star rating")
```




```
#boxplot(data_analyze[,c(8,9,10)], las = 2)
#boxplot(data_analyze[,c(2,4)], las = 2)
#boxplot(data_analyze[,-1], las = 2)
```

Linear regression with sentiment score, review count, and votes

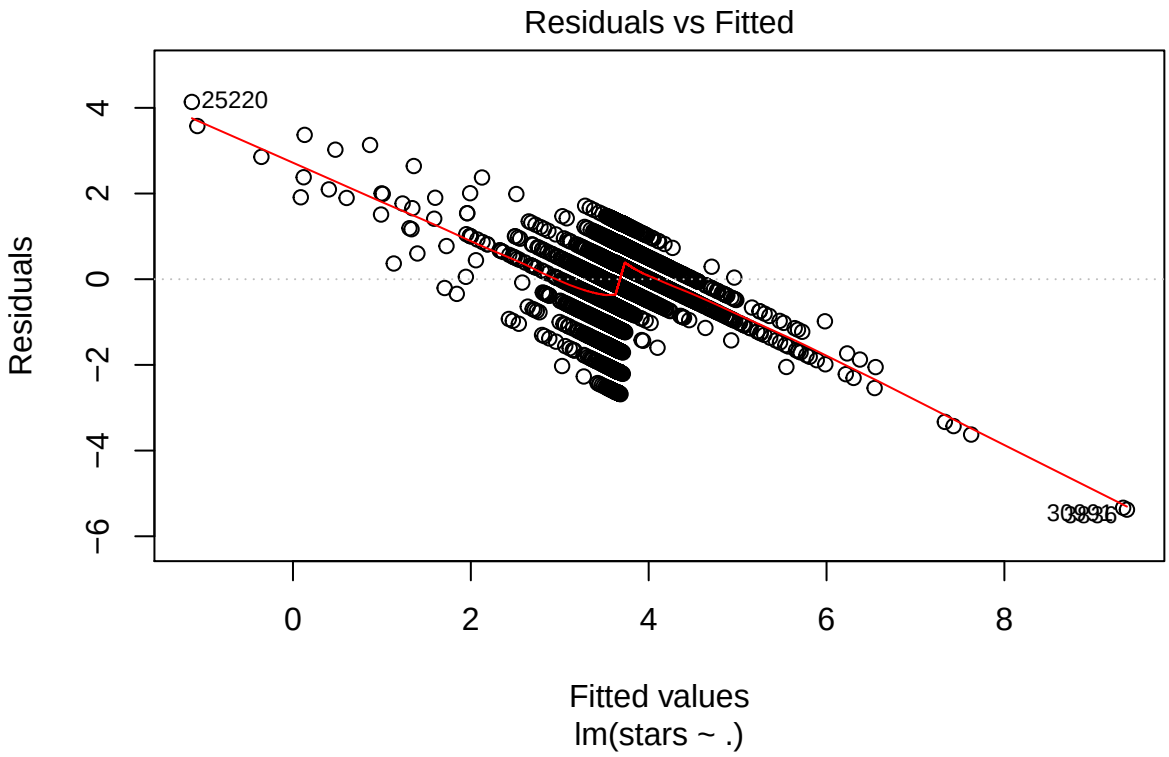
We create a multiple linear regression model with review counts, total score, total votes funny, total votes cool, total votes useful.

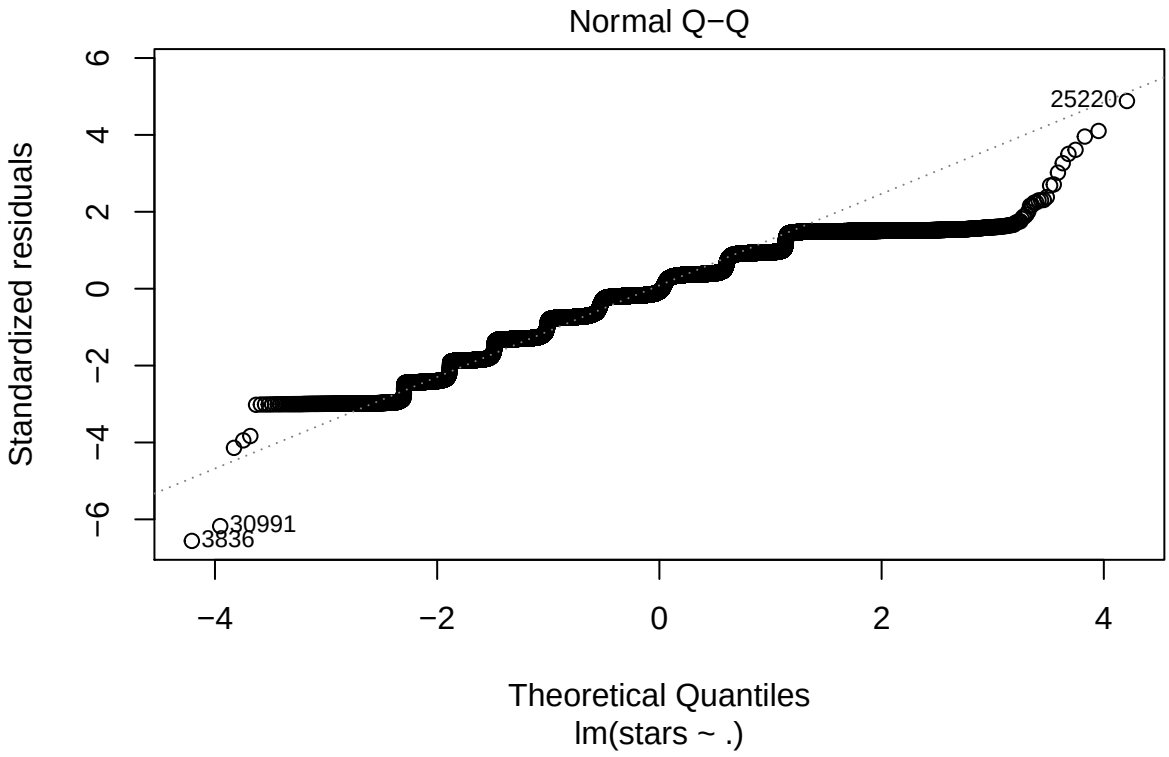
```
# select columns
data_analyze_2 <- select(data_analyze, stars, review_count, total_score, total_votes_funny,
                         total_votes_cool, total_votes_useful)

# Create linear regression model
fit5 <- lm( stars ~ ., data = data_analyze_2)
summary(fit5)
```

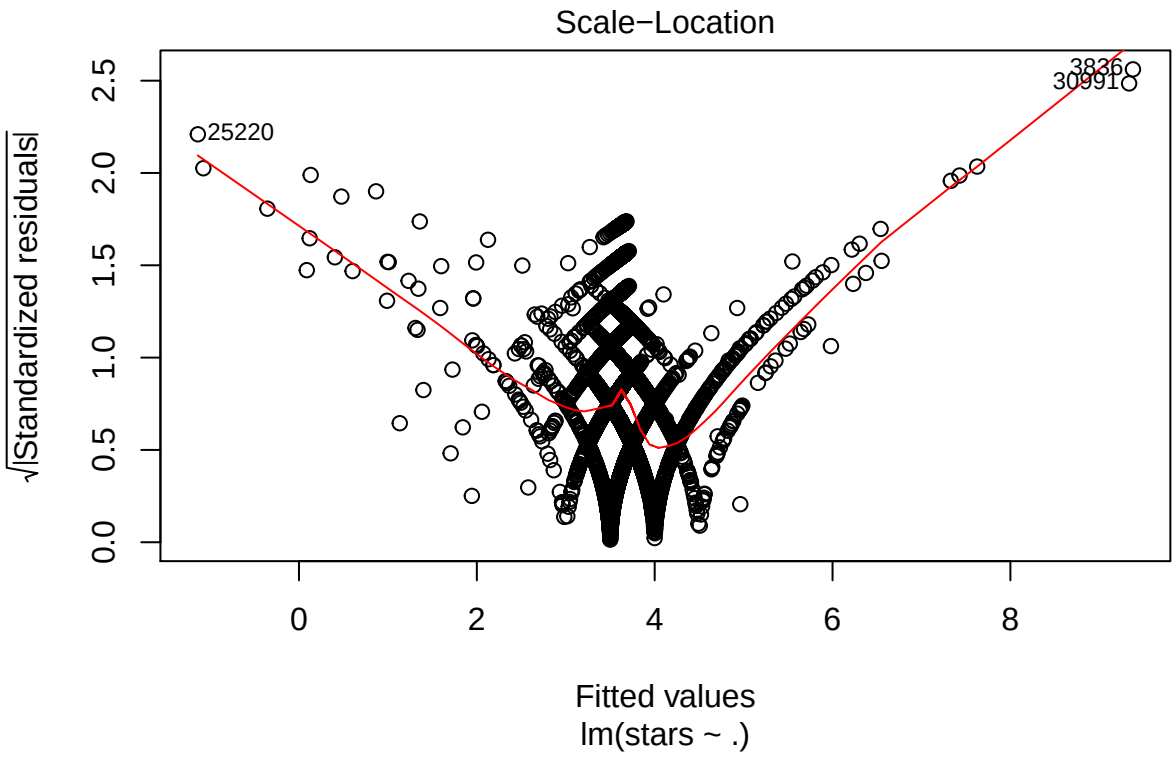
```
##
## Call:
## lm(formula = stars ~ ., data = data_analyze_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

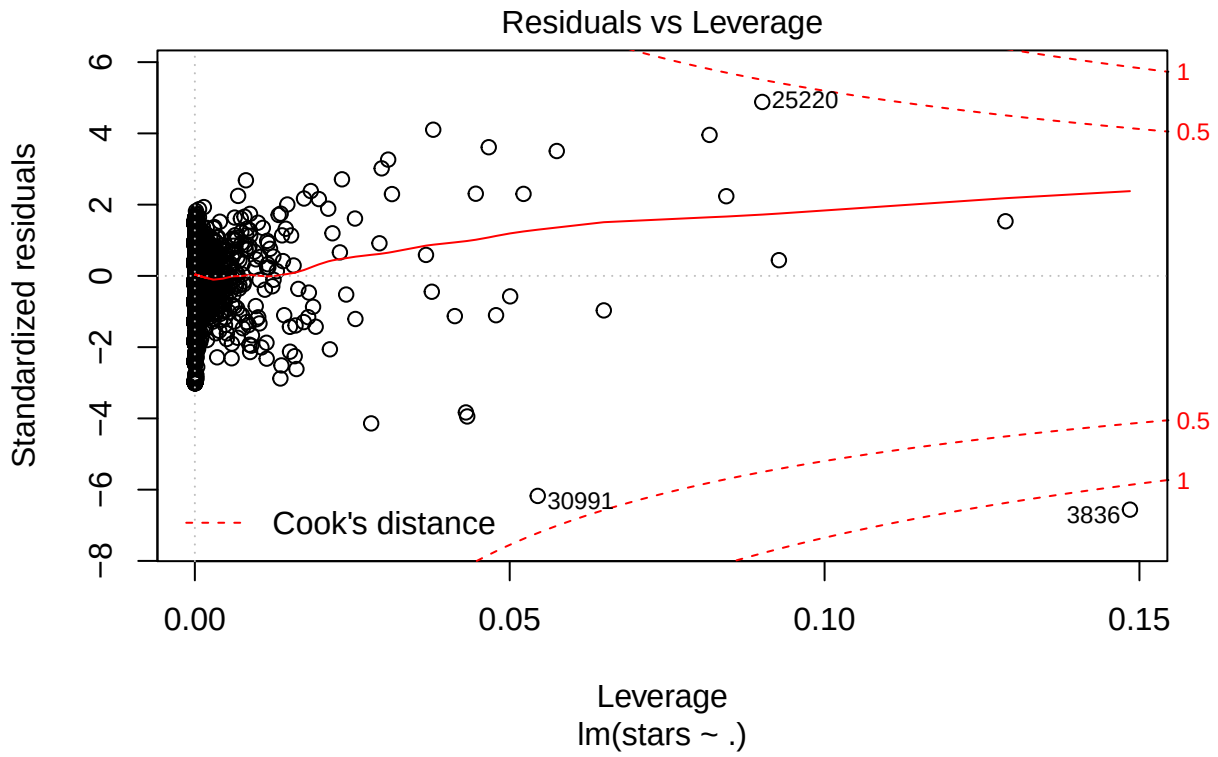
```
## -5.378 -0.638 -0.068  0.790  4.137
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.68e+00   4.86e-03  756.61   <2e-16 ***
## review_count       -3.60e-03   2.00e-04  -18.01   <2e-16 ***
## total_score         8.30e-04   3.98e-05   20.86   <2e-16 ***
## total_votes_funny  -3.08e-03   3.71e-04   -8.31   <2e-16 ***
## total_votes_cool    7.42e-03   4.09e-04   18.12   <2e-16 ***
## total_votes_useful -2.98e-03   2.52e-04  -11.85   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.888 on 38854 degrees of freedom
## Multiple R-squared:  0.0309, Adjusted R-squared:  0.0308
## F-statistic:  248 on 5 and 38854 DF,  p-value: <2e-16
```

```
plot(fit5)
```


Residuals vs Fitted
Residuals
Fitted values
lm(stars ~ .)
25220
30991
3836
4
2
0
−2
−4
−6
0
2
4
6
8


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(stars ~ .)
25220
30991
3836
6
4
2
0
−2
−4
−6
−4
−2
0
2
4


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
25220
3836
30991
0
2
4
6
8
Fitted values
lm(stars ~ .)




```
# Does the varibales have an impact on star ratings.
aov_model5 <- aov( stars ~ . , data_analyze_2)
aov_model5
```

```
## Call:
##    aov(formula = stars ~ ., data = data_analyze_2)
##
## Terms:
##                 review_count total_score total_votes_funny
## Sum of Squares            16         692                 2
## Deg. of Freedom            1           1                 1
##                 total_votes_cool total_votes_useful Residuals
## Sum of Squares               157                 111     30669
## Deg. of Freedom                1                   1     38854
##
## Residual standard error: 0.8884
## Estimated effects may be unbalanced
```

```
summary(aov_model5)
```

```
##                      Df Sum Sq Mean Sq F value  Pr(>F)
## review_count          1     16      16   20.77 5.2e-06 ***
## total_score           1    692     692  876.71 < 2e-16 ***
## total_votes_funny     1      2       2    2.24    0.13
## total_votes_cool      1    157     157  198.54 < 2e-16 ***
```

```
## total_votes_useful     1    111     111  140.31 < 2e-16 ***
## Residuals          38854  30669       1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

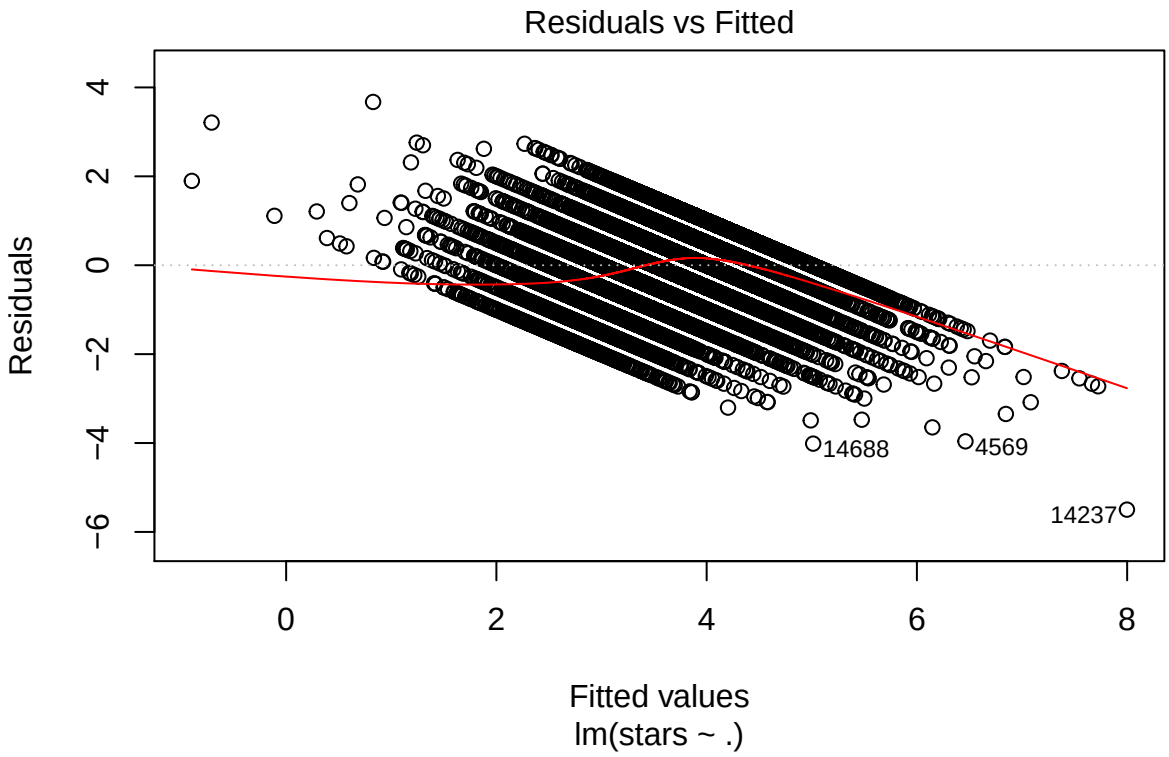
Multiple Linear regression with multiple variables

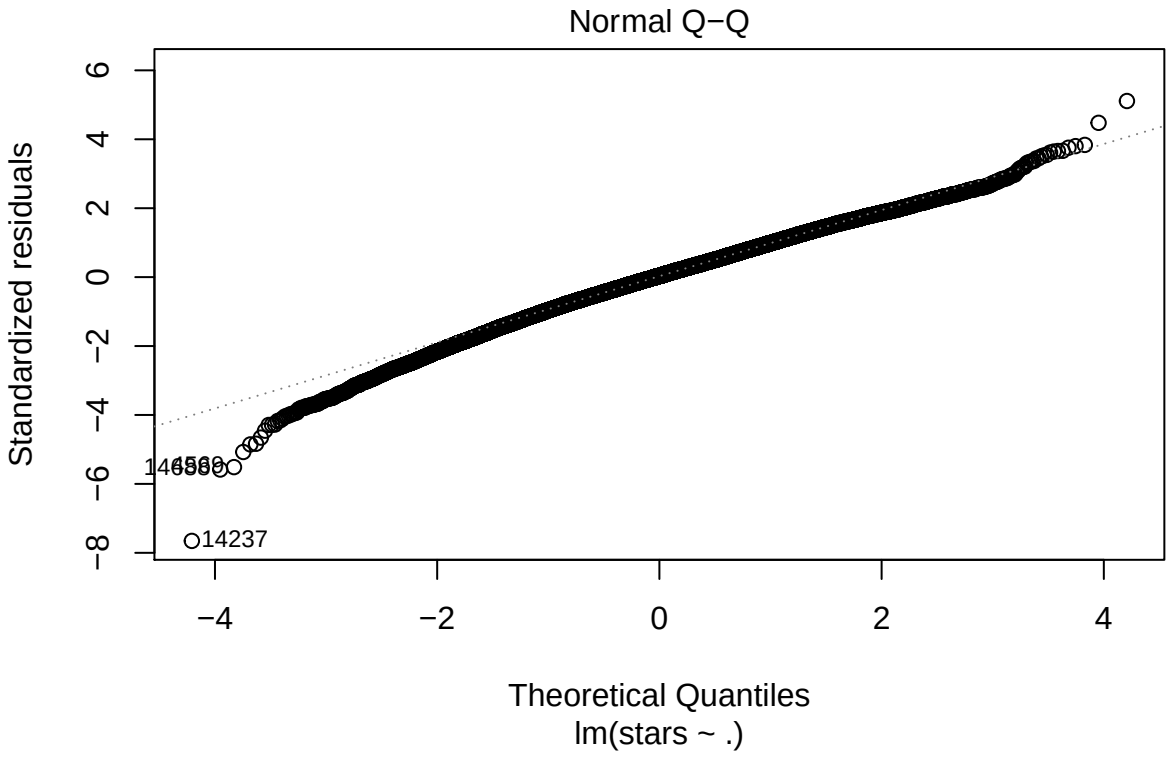
This linear regression modeling includes more features that can make better predictions. This model will quantify a relationship between a star rating and predictor variables which will include total review counts, total sentiment score, average sentiment score, minimum sentiment score, maximum sentiment score, number of cool votes, number of funny votes, and number of useful votes.

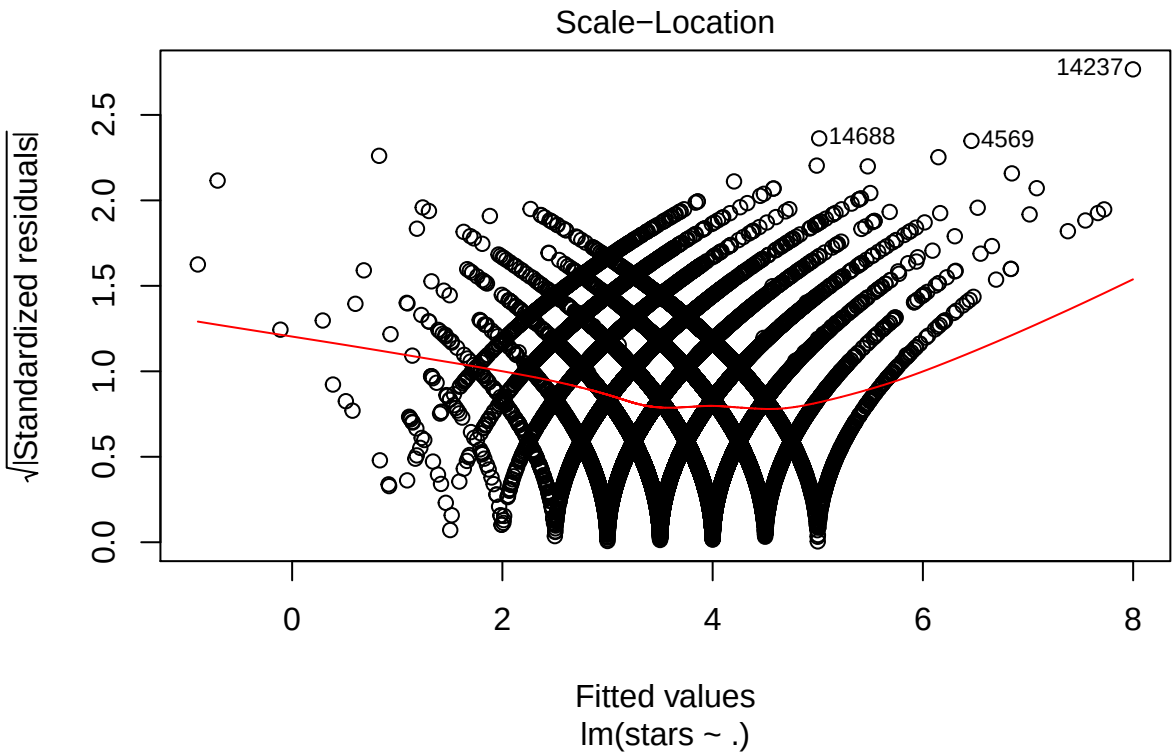
```
# Create a linear regression model which will include multiple predictors such as
# total review counts, total sentiment score, average sentiment score, minimum sentiment score, maximum
# sentiment score, number of cool votes, number of funny votes, and number of useful votes
fit <- lm(stars ~ ., data = data_analyze[, -1])
summary(fit)
```

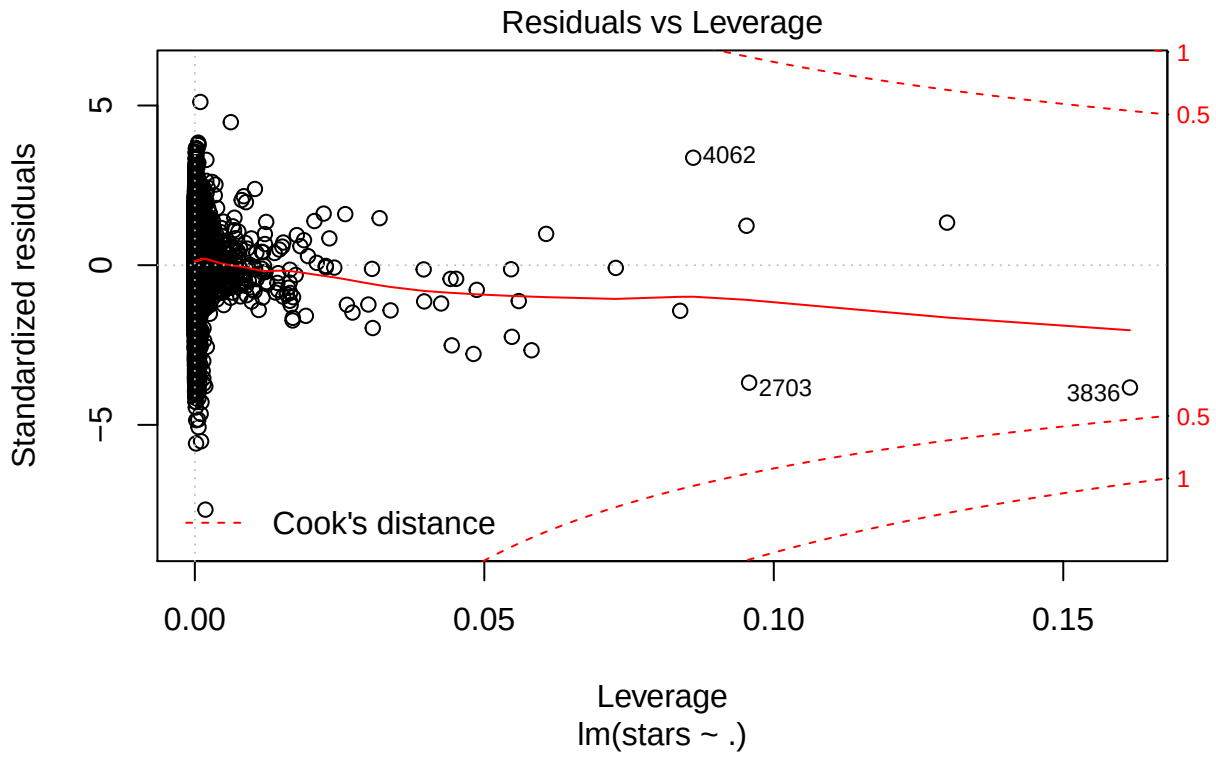
```
##
## Call:
## lm(formula = stars ~ ., data = data_analyze[, -1])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.498 -0.446  0.026  0.485  3.673
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.14e+00   8.50e-03  369.25  < 2e-16 ***
## review_count        8.40e-04   1.65e-04    5.08  3.7e-07 ***
## total_score        -1.08e-04   3.32e-05   -3.27   0.0011 **
## avg_score           2.22e-01   2.81e-03   79.26  < 2e-16 ***
## min_score           3.03e-02   1.41e-03   21.56  < 2e-16 ***
## max_score          -2.12e-02   8.67e-04  -24.48  < 2e-16 ***
## total_votes_cool    4.68e-03   3.33e-04   14.05  < 2e-16 ***
## total_votes_funny  -3.70e-03   3.03e-04  -12.20  < 2e-16 ***
## total_votes_useful -4.47e-04   2.11e-04   -2.12   0.0340 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.719 on 38851 degrees of freedom
## Multiple R-squared:  0.365,  Adjusted R-squared:  0.365
## F-statistic: 2.8e+03 on 8 and 38851 DF,  p-value: <2e-16
```

```
plot(fit)
```


Residuals vs Fitted
Residuals
Fitted values
lm(stars ~ .)
14688
4569
14237


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(stars ~ .)
14237


Scale−Location
14237
14688
4569
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
0
2
4
6
8
Fitted values
lm(stars ~ .)




```
# Check if the coefficients are significant
aov_model1 <- aov( stars ~ . , data_analyze[,-1])
aov_model1
```

```
## Call:
##    aov(formula = stars ~ ., data = data_analyze[, -1])
##
## Terms:
##                 review_count total_score avg_score min_score max_score
## Sum of Squares            16         692      9299      1145       291
## Deg. of Freedom            1           1         1         1         1
##                 total_votes_cool total_votes_funny total_votes_useful
## Sum of Squares                20                98                 2
## Deg. of Freedom                1                 1                 1
##                 Residuals
## Sum of Squares      20082
## Deg. of Freedom     38851
##
## Residual standard error: 0.719
## Estimated effects may be unbalanced
```

```
summary(aov_model1)
```

```
##                      Df Sum Sq Mean Sq F value  Pr(>F)
## review_count          1     16      16    31.7 1.8e-08 ***
```

```
## total_score            1    692     692  1338.8 < 2e-16 ***
## avg_score              1   9299    9299 17990.1 < 2e-16 ***
## min_score              1   1145    1145  2215.6 < 2e-16 ***
## max_score              1    291     291   563.0 < 2e-16 ***
## total_votes_cool       1     20      20    39.1 4.0e-10 ***
## total_votes_funny      1     98      98   189.1 < 2e-16 ***
## total_votes_useful     1      2       2     4.5   0.034 *
## Residuals          38851  20082       1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multiple Regression with subset

We create a third model that will take a subset of the data, we find a fit for a regression with subset of the data. In this model we take a subset of the data to focus on business that have 100 or more reviews.

This model will quantify a relationship between a star rating and predictor variables which will include total review counts, total sentiment score, average sentiment score, minimum sentiment score, maximum sentiment score, number of cool votes, number of funny votes, and number of useful votes.

```
# There are many business that have less review.
summary(data_analyze$review_count)
```

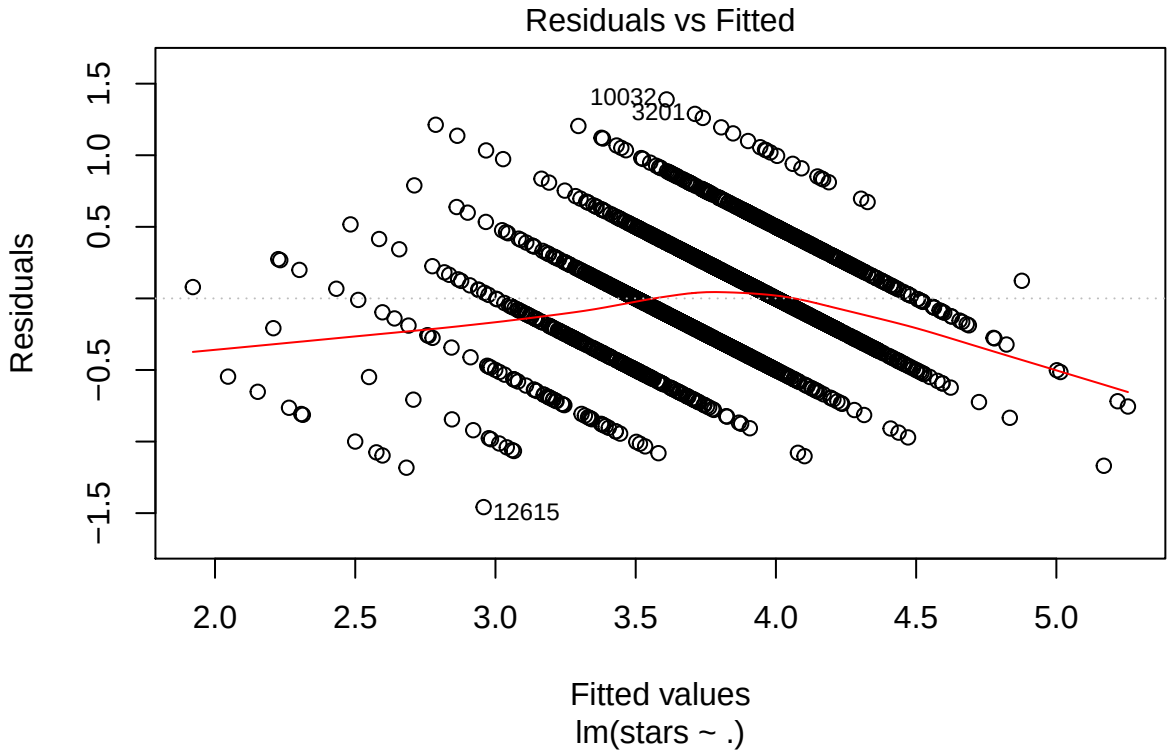
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       3       4       9      31      23    4080
```

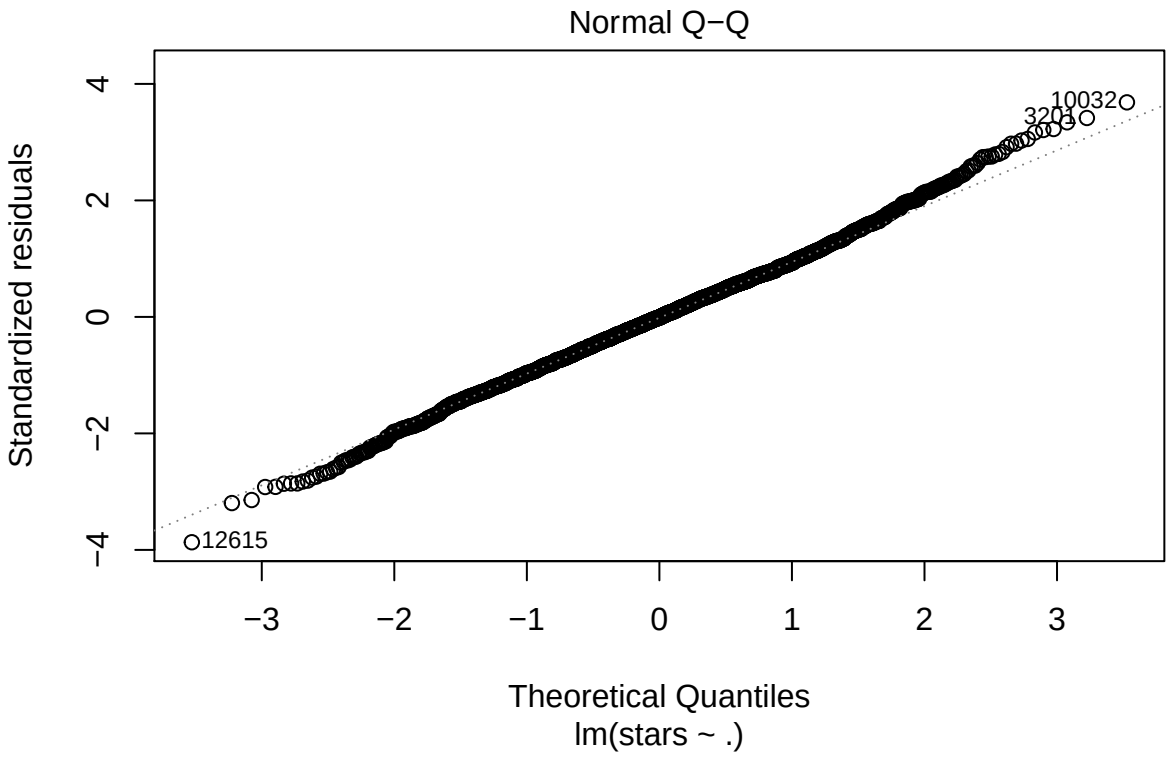
```
# We focus only on business that have 100 or more review
# The second model, use the same columns to predict rating for businesses
# that have more than 100 reviews.
fit2 <- lm(stars ~ ., data = data_analyze[data_analyze$review_count > 100, -c( 1)])
summary(fit2)
```

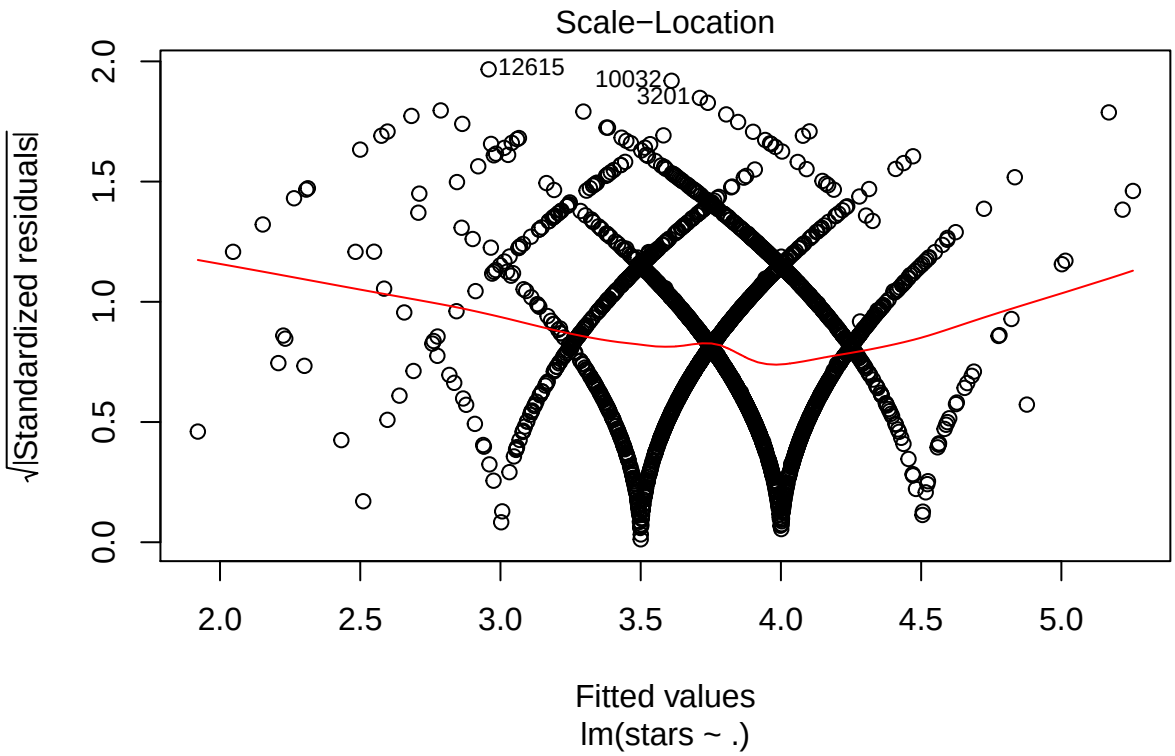
```
##
## Call:
## lm(formula = stars ~ ., data = data_analyze[data_analyze$review_count >
##     100, -c(1)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4578 -0.2495 -0.0035  0.2383  1.3899
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.22e+00   4.36e-02   73.87  < 2e-16 ***
## review_count        1.02e-04   1.13e-04    0.90   0.3668
## total_score        -6.32e-05   2.33e-05   -2.72   0.0066 **
## avg_score           1.96e-01   8.81e-03   22.29  < 2e-16 ***
## min_score           2.68e-02   2.28e-03   11.74  < 2e-16 ***
## max_score          -1.05e-02   1.14e-03   -9.21  < 2e-16 ***
## total_votes_cool    3.13e-03   2.03e-04   15.43  < 2e-16 ***
## total_votes_funny  -1.35e-03   1.86e-04   -7.27  4.7e-13 ***
## total_votes_useful -7.90e-04   1.28e-04   -6.15  8.8e-10 ***
```

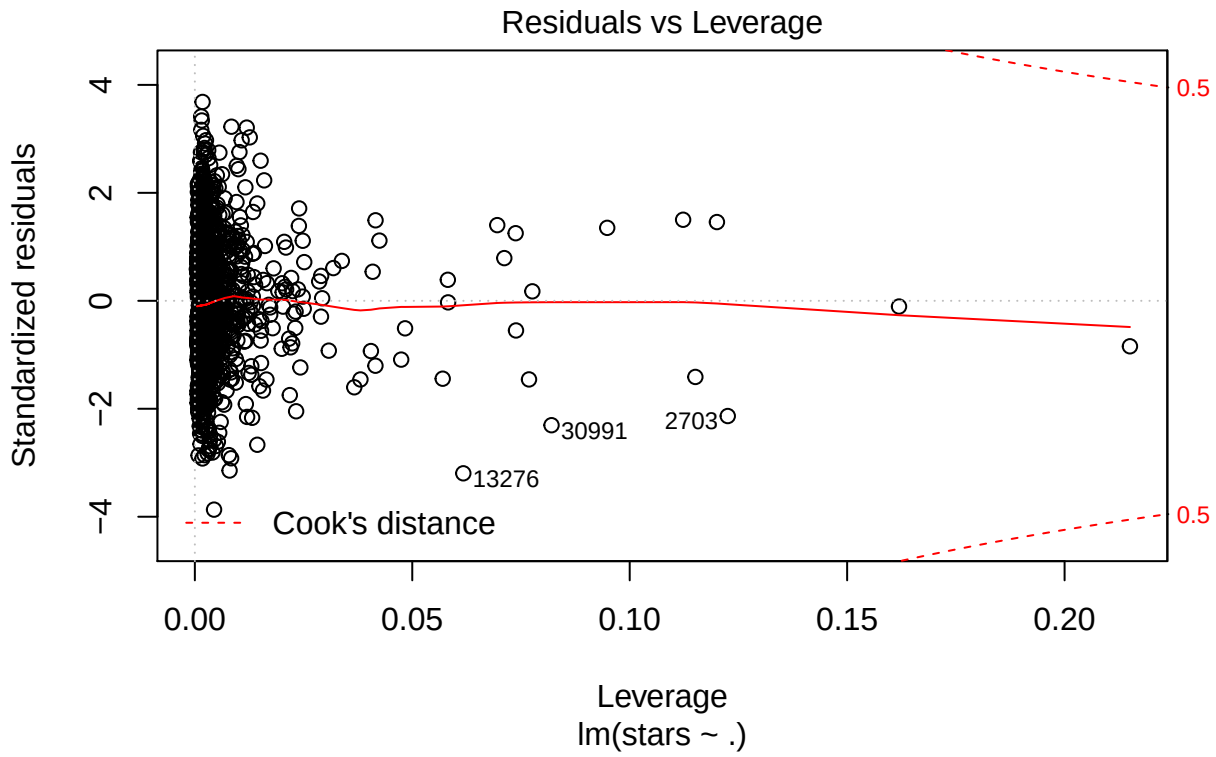
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.378 on 2376 degrees of freedom
## Multiple R-squared:  0.464,  Adjusted R-squared:  0.462
## F-statistic:  257 on 8 and 2376 DF,  p-value: <2e-16
```

```
plot(fit2)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(stars ~ .)
10032
3201
12615
−4
−2
0
2
4
−3
−2
−1
0
1
2
3


Scale−Location
12615
10032
3201
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Fitted values
lm(stars ~ .)




```
# Check how significant are the coefficients?
aov_model2 <- aov( stars ~ . , data_analyze[data_analyze$review_count > 100,-c( 1)])
aov_model2
```

```
## Call:
##    aov(formula = stars ~ ., data = data_analyze[data_analyze$review_count >
##     100, -c(1)])
##
## Terms:
##                 review_count total_score avg_score min_score max_score
## Sum of Squares           0.3        96.3     105.4      40.8      13.9
## Deg. of Freedom            1           1         1         1         1
##                 total_votes_cool total_votes_funny total_votes_useful
## Sum of Squares              16.3              15.1                5.4
## Deg. of Freedom                1                 1                  1
##                 Residuals
## Sum of Squares      338.9
## Deg. of Freedom      2376
##
## Residual standard error: 0.3777
## Estimated effects may be unbalanced
```

```
summary(aov_model2)
```

```
##                      Df Sum Sq Mean Sq F value  Pr(>F)
```

```
## review_count          1      0     0.3    1.87    0.17
## total_score           1     96    96.3  675.24 < 2e-16 ***
## avg_score             1    105   105.4  738.75 < 2e-16 ***
## min_score             1     41    40.8  285.73 < 2e-16 ***
## max_score             1     14    13.9   97.77 < 2e-16 ***
## total_votes_cool      1     16    16.3  114.23 < 2e-16 ***
## total_votes_funny     1     15    15.1  105.99 < 2e-16 ***
## total_votes_useful    1      5     5.4   37.88 8.8e-10 ***
## Residuals          2376    339     0.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```