# Task 1 submission: Data cleaning and regression

This is your submission document for [DataCapX Submission 1: Data cleaning and regression](https://courses.edx.org/courses/course-v1:AdelaideX+DataCapX+3T2017/courseware/5a5fa8c2bfd24d40bcfee2054bee384d/89fca51057e344c8b6ca09f5ab98fb5f/2?activate_block_id=block-v1%3AAdelaideX%2BDataCapX%2B3T2017%2Btype%40vertical%2Bblock%409c19fdbe059c469cb24e0d1dc41b9a5a).   
Save this document on your local machine and include all of your work within the relevant part of the assignment. Once you’ve completed every part of Task 1, upload this document via the [Your Response](https://courses.edx.org/courses/course-v1:AdelaideX+DataCapX+3T2017/courseware/5a5fa8c2bfd24d40bcfee2054bee384d/492ce9a548644a67a3e1ccf64267fac2/1?activate_block_id=block-v1%3AAdelaideX%2BDataCapX%2B3T2017%2Btype%40vertical%2Bblock%4038d7bfc39ebe464aa9d2cc60ea74f248) area.

**Note: Don’t forget to save your scripts related to this task and submit them when you submit this document.**

# Checklist

* Have you answered every question?
* Have you shown all of your working, including evidence of your code?
* Have you included all R output in this document to support your answers?
* Have you clearly stated conclusions where required?
* Have you saved your code in a script?

# Task 1 submission: Data cleaning and regression

1. Read in the reddit dataset correctly and provide evidence of your code.   
   [1 point]

reddit <- read.csv("RS\_2017-09\_filtered70.csv", stringsAsFactors=TRUE)

reddit <- as\_tibble(reddit)

reddit

Table . Showing the first 10 items of the dataset loaded as a tibble.

|  |
| --- |
| # A tibble: 139,823 x 69  X adserver\_click\_u… adserver\_imp\_pi… archived author author\_cakeday author\_flair\_cs… author\_flair\_te… author\_id brand\_safe  <int> <lgl> <lgl> <fct> <fct> <fct> <fct> <fct> <fct> <fct>  1 0 NA NA False Podify "" "" "" "t2\_biwh… False  2 1 NA NA False darth… "" "" "Free Republics… "" False  3 2 NA NA False [dele… "" "" "" "" True  4 3 NA NA False [dele… "" "" "" "" True  5 4 NA NA False last-… "" "" "" "" True  6 5 NA NA False calif… "" "" "" "" True  7 6 NA NA False [dele… "" "" "" "" False  8 7 NA NA False alan\_s "" "" "dx 2002 d&amp;… "" False  9 8 NA NA False ThatG… "" "" "" "" True  10 9 NA NA False Steve… "" "" "" "" True  # … with 139,813 more rows, and 59 more variables: contest\_mode <fct>, created\_utc <int>, crosspost\_parent <fct>,  # crosspost\_parent\_list <fct>, disable\_comments <fct>, distinguished <fct>, domain <fct>, domain\_override <fct>, edited <fct>,  # embed\_type <fct>, embed\_url <fct>, gilded <int>, hidden <fct>, hide\_score <fct>, href\_url <fct>, id <fct>, imp\_pixel <lgl>,  # is\_crosspostable <fct>, is\_reddit\_media\_domain <fct>, is\_self <fct>, is\_video <fct>, link\_flair\_css\_class <fct>,  # link\_flair\_text <fct>, locked <fct>, media <fct>, media\_embed <fct>, mobile\_ad\_url <fct>, num\_comments <int>,  # num\_crossposts <int>, original\_link <lgl>, over\_18 <fct>, parent\_whitelist\_status <fct>, permalink <fct>, pinned <fct>,  # post\_hint <fct>, preview <fct>, promoted <fct>, promoted\_by <lgl>, promoted\_display\_name <lgl>, promoted\_url <lgl>,  # retrieved\_on <int>, score <int>, secure\_media <fct>, secure\_media\_embed <fct>, selftext <fct>, spoiler <fct>,  # stickied <fct>, subreddit <fct>, subreddit\_id <fct>, suggested\_sort <fct>, third\_party\_trackers <fct>,  # third\_party\_tracking <lgl>, third\_party\_tracking\_2 <lgl>, thumbnail <fct>, thumbnail\_height <dbl>, thumbnail\_width <dbl>,  # title <fct>, url <fct>, whitelist\_status <fct> |

Include the name/s of any associated code (.r) script file/s related to Step 1 that are included in your .zip file:

* task01.R

1. Remove the first column of the data frame and provide evidence.   
   [1 point]

reddit <- select(reddit, -X)

reddit

Table . Dataset with the first column (X) removed.

|  |
| --- |
| # A tibble: 139,823 x 68  adserver\_click\_u… adserver\_imp\_pix… archived author author\_cakeday author\_flair\_css… author\_flair\_text author\_id brand\_safe  <lgl> <lgl> <fct> <fct> <fct> <fct> <fct> <fct> <fct>  1 NA NA False Podify "" "" "" "t2\_biwh… False  2 NA NA False darthho… "" "" "Free Republics |… "" False  3 NA NA False [delete… "" "" "" "" True  4 NA NA False [delete… "" "" "" "" True  5 NA NA False last-te… "" "" "" "" True  6 NA NA False califor… "" "" "" "" True  7 NA NA False [delete… "" "" "" "" False  8 NA NA False alan\_s "" "" "dx 2002 d&amp;e … "" False  9 NA NA False ThatGuy… "" "" "" "" True  10 NA NA False StevenR… "" "" "" "" True  # … with 139,813 more rows, and 59 more variables: contest\_mode <fct>, created\_utc <int>, crosspost\_parent <fct>,  # crosspost\_parent\_list <fct>, disable\_comments <fct>, distinguished <fct>, domain <fct>, domain\_override <fct>, edited <fct>,  # embed\_type <fct>, embed\_url <fct>, gilded <int>, hidden <fct>, hide\_score <fct>, href\_url <fct>, id <fct>, imp\_pixel <lgl>,  # is\_crosspostable <fct>, is\_reddit\_media\_domain <fct>, is\_self <fct>, is\_video <fct>, link\_flair\_css\_class <fct>,  # link\_flair\_text <fct>, locked <fct>, media <fct>, media\_embed <fct>, mobile\_ad\_url <fct>, num\_comments <int>,  # num\_crossposts <int>, original\_link <lgl>, over\_18 <fct>, parent\_whitelist\_status <fct>, permalink <fct>, pinned <fct>,  # post\_hint <fct>, preview <fct>, promoted <fct>, promoted\_by <lgl>, promoted\_display\_name <lgl>, promoted\_url <lgl>,  # retrieved\_on <int>, score <int>, secure\_media <fct>, secure\_media\_embed <fct>, selftext <fct>, spoiler <fct>,  # stickied <fct>, subreddit <fct>, subreddit\_id <fct>, suggested\_sort <fct>, third\_party\_trackers <fct>,  # third\_party\_tracking <lgl>, third\_party\_tracking\_2 <lgl>, thumbnail <fct>, thumbnail\_height <dbl>, thumbnail\_width <dbl>,  # title <fct>, url <fct>, whitelist\_status <fct> |

Include the name/s of any associated code (.r) script file/s related to Step 2 that are included in your .zip file:

* task01.R

1. Correctly identify and remove the columns that contain no data. Provide a list of the variables removed and evidence of your code.   
   [3 points]

Identify Columns with no data

reddit.is\_empty <- reddit %>%

  apply(

    MARGIN=2,

    FUN=function(x) all(is.na(x))) %>%

  data.frame

colnames(reddit.is\_empty)[1] <- "empty"

reddit.is\_empty %>%

  filter(empty == TRUE)

Table . Columns identified as having no values other than NA.

|  |
| --- |
| empty  adserver\_click\_url TRUE  adserver\_imp\_pixel TRUE  imp\_pixel TRUE  original\_link TRUE  promoted\_by TRUE  promoted\_display\_name TRUE  promoted\_url TRUE  third\_party\_tracking TRUE  third\_party\_tracking\_2 TRUE |

Remove Empty Columns

reddit <- reddit[, !sapply(reddit, function(x) all(is.na(x))), drop=FALSE]

reddit

After removing the first column there remained 68 columns, when removing the 9 columns as identified as empty, the dataset correctly shows only 59 columns remaining.

|  |
| --- |
| # A tibble: 139,823 x 59  archived author author\_cakeday author\_flair\_css\_c… author\_flair\_text author\_id brand\_safe contest\_mode created\_utc  <fct> <fct> <fct> <fct> <fct> <fct> <fct> <fct> <int>  1 False Podify "" "" "" "t2\_biwhg… False False 1504224000  2 False darthholo "" "" "Free Republics | 3" "" False False 1504224013  3 False [deleted] "" "" "" "" True False 1504224028  4 False [deleted] "" "" "" "" True False 1504224046  5 False last-term-… "" "" "" "" True False 1504224063  6 False california… "" "" "" "" True False 1504224081  7 False [deleted] "" "" "" "" False False 1504224102  8 False alan\_s "" "" "dx 2002 d&amp;e 2000m… "" False False 1504224119  9 False ThatGuyWit… "" "" "" "" True False 1504224137  10 False StevenRam95 "" "" "" "" True False 1504224157  # … with 139,813 more rows, and 50 more variables: crosspost\_parent <fct>, crosspost\_parent\_list <fct>, disable\_comments <fct>,  # distinguished <fct>, domain <fct>, domain\_override <fct>, edited <fct>, embed\_type <fct>, embed\_url <fct>, gilded <int>,  # hidden <fct>, hide\_score <fct>, href\_url <fct>, id <fct>, is\_crosspostable <fct>, is\_reddit\_media\_domain <fct>,  # is\_self <fct>, is\_video <fct>, link\_flair\_css\_class <fct>, link\_flair\_text <fct>, locked <fct>, media <fct>,  # media\_embed <fct>, mobile\_ad\_url <fct>, num\_comments <int>, num\_crossposts <int>, over\_18 <fct>,  # parent\_whitelist\_status <fct>, permalink <fct>, pinned <fct>, post\_hint <fct>, preview <fct>, promoted <fct>,  # retrieved\_on <int>, score <int>, secure\_media <fct>, secure\_media\_embed <fct>, selftext <fct>, spoiler <fct>,  # stickied <fct>, subreddit <fct>, subreddit\_id <fct>, suggested\_sort <fct>, third\_party\_trackers <fct>, thumbnail <fct>,  # thumbnail\_height <dbl>, thumbnail\_width <dbl>, title <fct>, url <fct>, whitelist\_status <fct> |

Include the name/s of any associated code (.r) script file/s related to Step 3 that are included in your .zip file:

* task01.R

1. Your code correctly identified and removed the columns that contained only one value. Provide a list of the variables removed and evidence of your code.   
   [3 point]

Identify columns with a single value

reddit %>%

  select\_if(~n\_distinct(.) == 1) %>%

  head(1)

Table . Columns with a single value, and what the value is - "False" in all the cases.

|  |
| --- |
| # A tibble: 1 x 3  hidden is\_crosspostable pinned  <fct> <fct> <fct>  1 False False False |

Remove columns with a single value

reddit <- reddit %>%

  select\_if(~n\_distinct(.) > 1)

reddit

Table . Dataset after removing the 3 columns with single values.

|  |
| --- |
| # A tibble: 139,823 x 56  archived author author\_cakeday author\_flair\_css\_c… author\_flair\_text author\_id brand\_safe contest\_mode created\_utc  <fct> <fct> <fct> <fct> <fct> <fct> <fct> <fct> <int>  1 False Podify "" "" "" "t2\_biwhg… False False 1504224000  2 False darthholo "" "" "Free Republics | 3" "" False False 1504224013  3 False [deleted] "" "" "" "" True False 1504224028  4 False [deleted] "" "" "" "" True False 1504224046  5 False last-term-… "" "" "" "" True False 1504224063  6 False california… "" "" "" "" True False 1504224081  7 False [deleted] "" "" "" "" False False 1504224102  8 False alan\_s "" "" "dx 2002 d&amp;e 2000m… "" False False 1504224119  9 False ThatGuyWit… "" "" "" "" True False 1504224137  10 False StevenRam95 "" "" "" "" True False 1504224157  # … with 139,813 more rows, and 47 more variables: crosspost\_parent <fct>, crosspost\_parent\_list <fct>, disable\_comments <fct>,  # distinguished <fct>, domain <fct>, domain\_override <fct>, edited <fct>, embed\_type <fct>, embed\_url <fct>, gilded <int>,  # hide\_score <fct>, href\_url <fct>, id <fct>, is\_reddit\_media\_domain <fct>, is\_self <fct>, is\_video <fct>,  # link\_flair\_css\_class <fct>, link\_flair\_text <fct>, locked <fct>, media <fct>, media\_embed <fct>, mobile\_ad\_url <fct>,  # num\_comments <int>, num\_crossposts <int>, over\_18 <fct>, parent\_whitelist\_status <fct>, permalink <fct>, post\_hint <fct>,  # preview <fct>, promoted <fct>, retrieved\_on <int>, score <int>, secure\_media <fct>, secure\_media\_embed <fct>,  # selftext <fct>, spoiler <fct>, stickied <fct>, subreddit <fct>, subreddit\_id <fct>, suggested\_sort <fct>,  # third\_party\_trackers <fct>, thumbnail <fct>, thumbnail\_height <dbl>, thumbnail\_width <dbl>, title <fct>, url <fct>,  # whitelist\_status <fct> |

Include the name/s of any associated code (.r) script file/s related to Step 4 that are included in your .zip file:

* task01.R

1. Convert the variables *created\_utc* and *retrieved\_on* to unordered factors containing the weekdays. Provide tables of the frequencies for the two new variables and evidence of your code.   
   [4 points]

Convert the two columns to abbreviated week day names

reddit <- reddit %>%

  mutate(created\_utc = as.POSIXct(created\_utc, origin = "1970-01-01"))  %>%

  mutate(retrieved\_on = as.POSIXct(retrieved\_on, origin = "1970-01-01")) %>%

  mutate(created\_utc = wday(created\_utc, label=TRUE, abbr=TRUE))  %>%

  mutate(retrieved\_on = wday(retrieved\_on, label=TRUE, abbr=TRUE))

Convert the columns to factors

reddit$created\_utc <- factor(reddit$created\_utc)

reddit$retrieved\_on <- factor(reddit$retrieved\_on)

Summarize the columns

reddit %>%

  select(created\_utc) %>%

  group\_by(created\_utc) %>%

  summarise(freq=n())

Table . Weekday frequencies for created\_utc.

|  |
| --- |
| # A tibble: 7 x 2  created\_utc freq  <ord> <int>  1 Sun 17260  2 Mon 19026  3 Tue 19648  4 Wed 19609  5 Thu 19479  6 Fri 23853  7 Sat 20948 |

reddit %>%

  select(retrieved\_on) %>%

  group\_by(retrieved\_on) %>%

  summarise(freq=n())

Table . Weekday frequencies for retrieved\_on.

|  |
| --- |
| # A tibble: 3 x 2  retrieved\_on freq  <ord> <int>  1 Mon 50335  2 Tue 31146  3 Fri 58342 |

Include the name/s of any associated code (.r) script file/s related to Step 5 that are included in your .zip file:

* task01.R

1. Your code successfully converts the titles of the reddit posts to an incidence matrix of the words that appear in at least 500 posts. Provide a list of the words that appear in at least 500 posts and incorporate the incidence matrix into the reddit data frame. Provided evidence of your code.  
   [10 points]

Create the Incidence Matrix with words occurring in at least 500 titles

# clean the titles for processing

docs <- Corpus(VectorSource(pull(reddit, title)))

docs <- tm\_map(docs, content\_transformer(tolower))

docs <- tm\_map(docs, removeNumbers)

docs <- tm\_map(docs, removeWords, stopwords("english"))

docs <- tm\_map(docs, removePunctuation)

docs <- tm\_map(docs, content\_transformer(function(x) gsub(x, pattern = "–", replacement = " ")))

docs <- tm\_map(docs, stripWhitespace)

docs <- tm\_map(docs, stemDocument)

# get the document term matrix and convert it to an incidence matrix

dtm <- TermDocumentMatrix(docs)

dtm$v <-rep(1, length(dtm$v))

# find terms that appears at least in 500 titles

ft <- findFreqTerms(dtm, lowfreq=500)

# filter the document term matrix to only frequent terms

dtm <- dtm[Terms(dtm)%in%ft,]

*\* Please note, in the world list I found "–" listed as a frequent word and have removed it as an additional step.*

Convert the Incident Matrix to an ordinary matrix and show the frequent words

# convert the dtm to a matrix

m <- as.matrix(dtm)

v <- sort(rowSums(m),decreasing=TRUE)

d <- data.frame(word = names(v),freq=v)

# display the word frequencies

pull(d, word)

Table . Stem Words appearing in at least 500 titles.

|  |
| --- |
| [1] "new" "get" "look" "can" "game" "like" "just" "help" "time" "one" "need"  [12] "world" "will" "make" "want" "first" "best" "anyon" "key" "day" "know" "use"  [23] "now" "peopl" "video" "trump" "amp" "play" "post" "question" "good" "year" "say"  [34] "work" "think" "guy" "free" "love" "take" "live" "got" "show" "week" "back"  [45] "today" "friend" "watch" "way" "girl" "offer" "nation" "pleas" "come" "thing" "start"  [56] "xbox" "tri" "see" "top" "find" "busi" "someon" "man" "black" "made" "buy"  [67] "right" "white" "call" "feel" "found" "ever" "septemb" "still" "hurrican" "last" "player"  [78] "onlin" "life" "kill" "give" "old" "realli" "end" "thought" "chang" "final" "stori"  [89] "set" "two" "fuck" "big" "team" "home" "sport" "polit" "updat" "reddit" "open"  [100] "car" "movi" "someth" "night" "build" "hous" "littl" "win" "talk" "much" "irma"  [111] "school" "long" "run" "season" "idea" "state" "happen" "great" "full" "even" "high"  [122] "part" "stream" "war" "fun" "news" "north" "hour" "let" "real" "song" "music"  [133] "discuss" "month" "advic" "die" "tell" "never" "els" "ask" "citi" "blue" "name"  [144] "test" "review" "sell" "keep" "lfm" "level" "stop" "next" "power" "plan" "check"  [155] "trade" "better" "bad" "raid" "face" "place" "fire" "support" "servic" "store" "card"  [166] "person" "thread" "move" "tip" "dog" "request" "everi" "spoiler" "price" "star" "report"  [177] "red" "case" "anoth" "korea" "deal" "offici" "job" "remov" "differ" |

Incorporate the Incident Matrix in the reddit data frame and remove the title column

# transpose the matrix and add a column prefix

m <- t(m)

colnames(m) <- paste("title", colnames(m), sep = "\_")

#add the incident matrix data to the reddit data frame

reddit <- bind\_cols(reddit, as\_tibble(m), .name\_repair="unique")

# remove the title column

reddit <- select(reddit, -title)

colnames(reddit)

Table . Reddit data frame columns.

|  |
| --- |
| [1] "archived" "author" "author\_cakeday" "author\_flair\_css\_class"  [5] "author\_flair\_text" "author\_id" "brand\_safe" "contest\_mode"  [9] "created\_utc" "crosspost\_parent" "crosspost\_parent\_list" "disable\_comments"  [13] "distinguished" "domain" "domain\_override" "edited"  [17] "embed\_type" "embed\_url" "gilded" "hide\_score"  [21] "href\_url" "id" "is\_reddit\_media\_domain" "is\_self"  [25] "is\_video" "link\_flair\_css\_class" "link\_flair\_text" "locked"  [29] "media" "media\_embed" "mobile\_ad\_url" "num\_comments"  [33] "num\_crossposts" "over\_18" "parent\_whitelist\_status" "permalink"  [37] "post\_hint" "preview" "promoted" "retrieved\_on"  [41] "score" "secure\_media" "secure\_media\_embed" "selftext"  [45] "spoiler" "stickied" "subreddit" "subreddit\_id"  [49] "suggested\_sort" "third\_party\_trackers" "thumbnail" "thumbnail\_height"  [53] "thumbnail\_width" "url" "whitelist\_status" "title\_busi"  [57] "title\_know" "title\_like" "title\_new" "title\_stream"  [61] "title\_final" "title\_full" "title\_can" "title\_red"  [65] "title\_use" "title\_will" "title\_made" "title\_high"  [69] "title\_free" "title\_game" "title\_top" "title\_key"  [73] "title\_person" "title\_just" "title\_peopl" "title\_report"  [77] "title\_citi" "title\_day" "title\_car" "title\_get"  [81] "title\_say" "title\_world" "title\_got" "title\_end"  [85] "title\_night" "title\_season" "title\_spoiler" "title\_make"  [89] "title\_song" "title\_offer" "title\_year" "title\_septemb"  [93] "title\_back" "title\_part" "title\_two" "title\_call"  [97] "title\_last" "title\_let" "title\_think" "title\_team"  [101] "title\_idea" "title\_one" "title\_polit" "title\_name"  [105] "title\_today" "title\_talk" "title\_work" "title\_chang"  [109] "title\_post" "title\_want" "title\_level" "title\_video"  [113] "title\_offici" "title\_white" "title\_time" "title\_fire"  [117] "title\_littl" "title\_bad" "title\_player" "title\_look"  [121] "title\_man" "title\_hour" "title\_stop" "title\_play"  [125] "title\_home" "title\_win" "title\_onlin" "title\_thing"  [129] "title\_ask" "title\_happen" "title\_question" "title\_reddit"  [133] "title\_ever" "title\_month" "title\_deal" "title\_long"  [137] "title\_support" "title\_keep" "title\_see" "title\_black"  [141] "title\_life" "title\_take" "title\_tri" "title\_hurrican"  [145] "title\_nation" "title\_girl" "title\_school" "title\_remov"  [149] "title\_need" "title\_trade" "title\_come" "title\_set"  [153] "title\_love" "title\_feel" "title\_give" "title\_trump"  [157] "title\_never" "title\_xbox" "title\_even" "title\_someth"  [161] "title\_sport" "title\_buy" "title\_great" "title\_realli"  [165] "title\_next" "title\_live" "title\_help" "title\_watch"  [169] "title\_amp" "title\_found" "title\_thread" "title\_face"  [173] "title\_die" "title\_tell" "title\_war" "title\_price"  [177] "title\_test" "title\_guy" "title\_first" "title\_good"  [181] "title\_anyon" "title\_els" "title\_someon" "title\_review"  [185] "title\_run" "title\_everi" "title\_hous" "title\_check"  [189] "title\_fuck" "title\_pleas" "title\_sell" "title\_open"  [193] "title\_thought" "title\_way" "title\_dog" "title\_kill"  [197] "title\_place" "title\_move" "title\_servic" "title\_blue"  [201] "title\_updat" "title\_store" "title\_discuss" "title\_start"  [205] "title\_better" "title\_friend" "title\_old" "title\_now"  [209] "title\_right" "title\_differ" "title\_movi" "title\_still"  [213] "title\_find" "title\_fun" "title\_plan" "title\_music"  [217] "title\_week" "title\_much" "title\_news" "title\_big"  [221] "title\_build" "title\_real" "title\_show" "title\_raid"  [225] "title\_state" "title\_best" "title\_stori" "title\_request"  [229] "title\_card" "title\_job" "title\_case" "title\_advic"  [233] "title\_star" "title\_lfm" "title\_korea" "title\_power"  [237] "title\_anoth" "title\_tip" "title\_north" "title\_irma" |

Include the name/s of any associated code (.r) script file/s related to Step 6 that are included in your .zip file:

* task01.R

1. Your code correctly identifies all factors where at least one level occurs less than 30 times and recodes them by amalgamating levels appropriately. Provide a list of the factors that were recoded and provide evidence of your code.   
   [4 points]

Identify factors to recode

factor\_cols <- reddit %>%

  select\_if(is.factor) %>%

  gather(name, value) %>%

  count(name, value) %>%

  filter(n < 30) %>%

  group\_by(name) %>%

  summarise() %>%

  pull(name)

factor\_cols

Table . Factors where at least one level occurs less than 30 times.

|  |
| --- |
| [1] "archived" "author" "author\_flair\_css\_class" "author\_flair\_text" "author\_id"  [6] "crosspost\_parent" "crosspost\_parent\_list" "domain" "domain\_override" "edited"  [11] "embed\_type" "embed\_url" "hide\_score" "href\_url" "id"  [16] "link\_flair\_css\_class" "link\_flair\_text" "media" "media\_embed" "mobile\_ad\_url"  [21] "permalink" "preview" "secure\_media" "secure\_media\_embed" "selftext"  [26] "subreddit" "subreddit\_id" "suggested\_sort" "third\_party\_trackers" "thumbnail"  [31] "url" |

Recode Identified Factors

reddit[factor\_cols] <- sapply(reddit[factor\_cols], function(x) fct\_lump\_min(f=x, min=30, other\_level="other"))

reddit <- reddit %>% mutate\_at(all\_of(factor\_cols), as.factor)

Include the name/s of any associated code (.r) script file/s related to Step 7 that are included in your .zip file:

* task01.R

1. Your code correctly identifies all factors with more than 100 levels and recodes them by amalgamating levels appropriately. Provide a list of the factors that were recoded and evidence of your code.   
   [4 points]

Identify all factors with more than 100 levels

many\_factor <- reddit %>%

  select\_if(is.factor) %>%

  sapply(nlevels) %>%

  data.frame() %>%

  rownames\_to\_column() %>%

  setNames(c("name", "n")) %>%

  filter(n > 100) %>%

  pull(name)

many\_factor

Table . Factors with more than 100 levels.

|  |
| --- |
| [1] "author" "author\_flair\_css\_class" "author\_flair\_text" "author\_id" "crosspost\_parent"  [6] "crosspost\_parent\_list" "domain" "edited" "href\_url" "id"  [11] "link\_flair\_css\_class" "link\_flair\_text" "media" "media\_embed" "mobile\_ad\_url"  [16] "permalink" "preview" "secure\_media" "secure\_media\_embed" "selftext"  [21] "subreddit" "subreddit\_id" "thumbnail" "url" |

Amalgamate Levels

reddit[many\_factor] <- sapply(reddit[many\_factor], function(x) fct\_lump\_n(f=x, n=100, other\_level="other"))

reddit <- reddit %>% mutate\_at(all\_of(many\_factor), as.factor)

Include the name/s of any associated code (.r) script file/s related to Step 8 that are included in your .zip file:

* task01.R

1. Your code correctly identifies and removes all factors that cannot be suitably recoded so that each level occurs at least 30 times. Provide a list of the factors that were removed and evidence of your code.   
   [3 points]

Identify factors that can't be successfully recoded

removal\_factors <- reddit %>%

  select\_if(is.factor) %>%

  gather(name, value) %>%

  count(name, value) %>%

  filter(n < 30) %>%

  group\_by(name) %>%

  summarise() %>%

  pull(name)

removal\_factors

Table . Remaining factors where at least one level occurs less than 30 times.

|  |
| --- |
| [1] "archived" "author\_flair\_css\_class" "author\_flair\_text" "author\_id" "crosspost\_parent"  [6] "crosspost\_parent\_list" "domain\_override" "edited" "embed\_type" "embed\_url"  [11] "hide\_score" "href\_url" "id" "media" "media\_embed"  [16] "mobile\_ad\_url" "permalink" "preview" "secure\_media" "secure\_media\_embed"  [21] "selftext" "suggested\_sort" "third\_party\_trackers" "thumbnail" "url" |

Remove factors not successfully recoded

reddit <- reddit %>%

  select(-all\_of(removal\_factors))

Identify Factors with only one level

uniform\_factors <- reddit %>%

  select\_if(is.factor) %>%

  sapply(nlevels) %>%

  data.frame() %>%

  rownames\_to\_column() %>%

  setNames(c("name", "n")) %>%

  filter(n == 1) %>%

  pull(name)

uniform\_factors

Table . Uniform factors.

|  |
| --- |
| [1] "id" "permalink" "url" |

Remove Uniform Factors

reddit <- reddit %>%

  select(-all\_of(uniform\_factors))

Include the name/s of any associated code (.r) script file/s related to Step 9 that are included in your .zip file:

* task01.R

1. Your code correctly identifies and removes all columns with an excessive number of missing values. Provide a list of the variables that were removed and evidence of your code.   
   [3 points]

Identify columns with excessive missing values

na\_factors <- reddit %>%

  sapply(function(x) sum(length(which(is.na(x))))) %>%

  data.frame() %>%

  rownames\_to\_column() %>%

  setNames(c("name", "n")) %>%

  filter(n > 20) %>%

  pull(name)

na\_factors

Table . Columns with more than 20 missing values.

|  |
| --- |
| [1] "link\_flair\_text" "thumbnail\_height" "thumbnail\_width" |

Remove Identified Columns from the data set

reddit <- reddit %>%

  select(-all\_of(na\_factors))

Include the name/s of any associated code (.r) script file/s related to Step 10 that are included in your .zip file:

* task01.R

1. Provide a list of the remaining factors and the number of levels in the reddit data frame at the end of the data cleaning steps.  
   [1 point]

reddit %>%

  select\_if(is.factor) %>%

  sapply(nlevels) %>%

  data.frame() %>%

  rownames\_to\_column() %>%

  setNames(c("factor", "levels"))

Table . Remaining factors.

|  |
| --- |
| factor levels  1 author 104  2 author\_cakeday 2  3 author\_flair\_css\_class 35  4 author\_flair\_text 28  5 author\_id 2  6 brand\_safe 2  7 contest\_mode 2  8 created\_utc 7  9 crosspost\_parent 2  10 crosspost\_parent\_list 2  11 disable\_comments 3  12 distinguished 2  13 domain 100  14 edited 2  15 href\_url 2  16 is\_reddit\_media\_domain 2  17 is\_self 2  18 is\_video 2  19 link\_flair\_css\_class 102  20 locked 2  21 media 2  22 media\_embed 2  23 mobile\_ad\_url 2  24 over\_18 2  25 parent\_whitelist\_status 8  26 post\_hint 7  27 preview 26  28 promoted 2  29 retrieved\_on 3  30 secure\_media 2  31 secure\_media\_embed 2  32 selftext 4  33 spoiler 2  34 stickied 2  35 subreddit 100  36 subreddit\_id 100  37 thumbnail 6  38 whitelist\_status 8 |

Include the name/s of any associated code (.r) script file/s related to Step 11 that are included in your .zip file:

* task01.R

1. Correctly identify the numerical predictor variables. Correctly plot the score against each of the numerical predictor variables, with and without the transformation. Clearly state your conclusion about whether the transformations should be applied.   
   [5 points]

Identify the numerical predictor variables

numeric\_columns <- reddit %>%

  select(where(is.numeric)) %>%

  select(-starts\_with("title\_")) %>%

  names()

numeric\_columns

Table . Identified numeric predictor values and the score value.

|  |
| --- |
| [1] "gilded" "num\_comments" "num\_crossposts" "score" |

Plot the unchanged numeric values

# get the numeric column values to plot

numeric\_data <- reddit %>%

  select(all\_of(numeric\_columns)) %>%

  gather(key="column", value="value", -score)

# plot the unchanged numeric values against score

ggplot(numeric\_data, aes(x=value, y=score)) +

  stat\_bin\_hex(bins=10) +

  scale\_fill\_continuous(low="#B6D191", high="darkgreen") +

  geom\_smooth(method='lm', color="darkgray") +

  facet\_wrap(~column, scales="free")

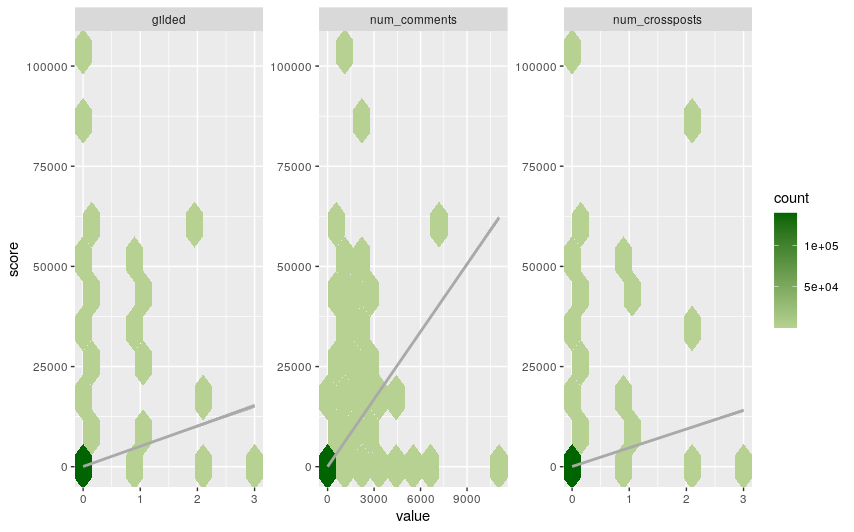


Figure . Scatter plots of the unchanged numeric values.

Transform and plot the numeric values

# transform the numeric values

scale <- function(x) log(1 + x)

scaled\_data <- numeric\_data %>%

  mutate\_if(is.numeric, scale)

# plot the scaled numeric values against score

ggplot(scaled\_data, aes(x=value, y=score)) +

  stat\_bin\_hex(bins=10) +

  scale\_fill\_continuous(low="#B6D191", high="darkgreen") +

  geom\_smooth(method='lm', color="darkgray") +

  facet\_wrap(~column, scales="free")

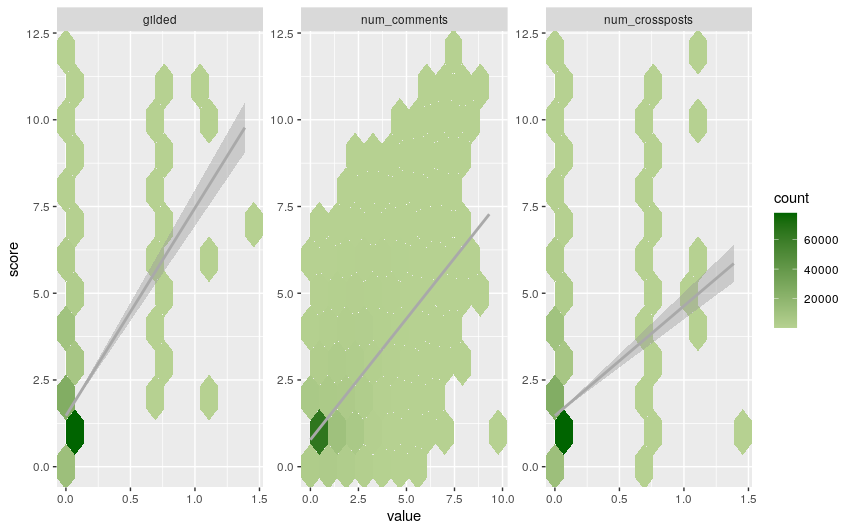


Figure . Scatter plots of the scaled numeric values.

Conclusion

Although far from perfect, the scatter plots clearly show that when the data is scaled it is more evenly spread around the best fit line. The log transformation will thus be applied to the dataset.

# transform the numeric values in the data set

reddit <- reddit %>% mutate\_at(all\_of(numeric\_columns), scale)

Include the name/s of any associated code (.r) script file/s related to Step 12 that are included in your .zip file:

* task01.R

1. Correctly fit the multiple linear regression model to the data. Obtain relevant diagnostic plots and state your conclusion about the model assumptions. Provide a concise summary of the model fit and evidence of your code.  
   [6 points]

Fit the linear regression model and create a basic summary

# fit a linear regression using all the columns in the data frame

reddit.lm <- lm(score~.,data=reddit)

# basic model summary

glance(reddit.lm) %>% gather()

Table . Basic model summary.

|  |
| --- |
| # A tibble: 12 x 2  key value  <chr> <dbl>  1 r.squared 0.584  2 adj.r.squared 0.581  3 sigma 0.923  4 statistic 285.  5 p.value 0  6 df 684  7 logLik -186894.  8 AIC 375161.  9 BIC 381917.  10 deviance 118605.  11 df.residual 139138  12 nobs 139823 |

Create diagnostic plots

layout(matrix(c(1,2,3,4),2,2))

plot(reddit.lm)

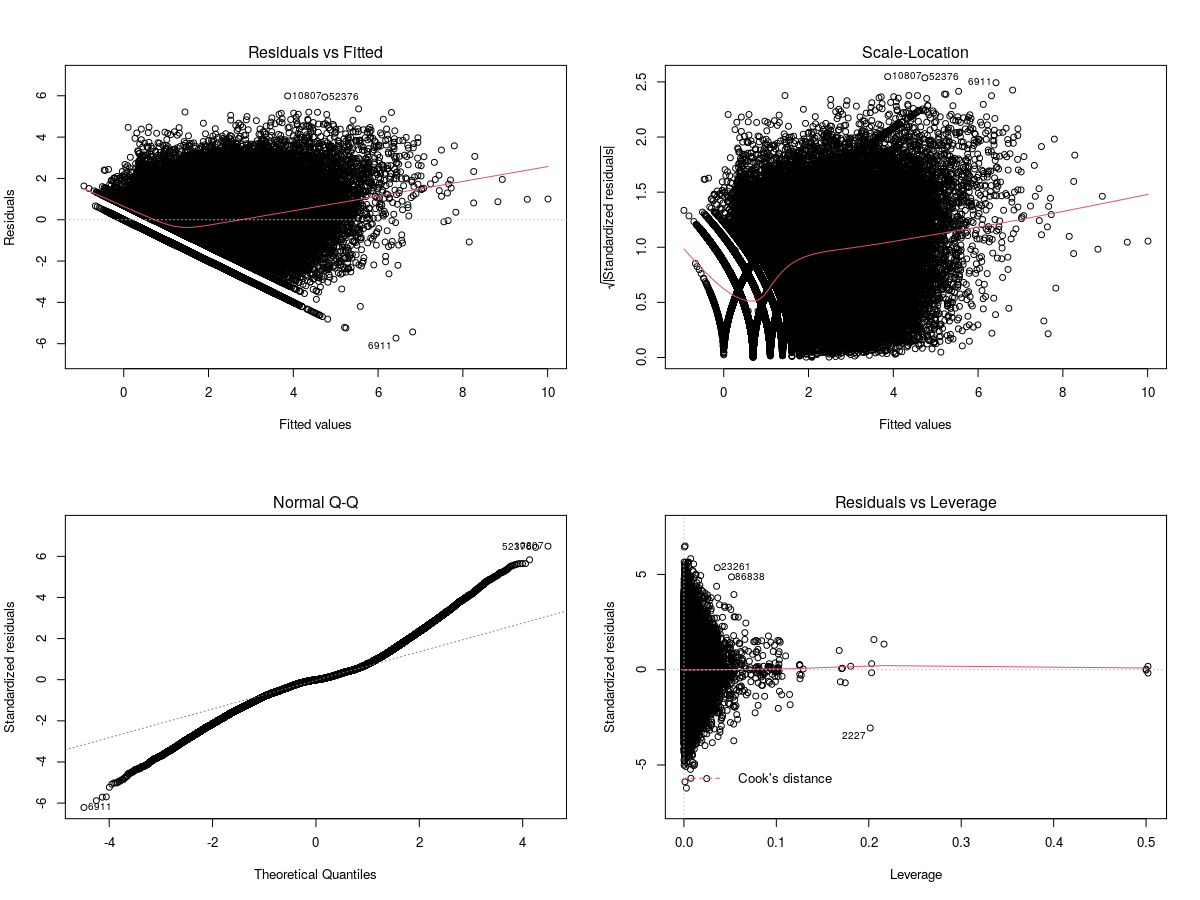


Figure . Model diagnostic plots.

Conclusion

* **R-squared** – In the [AnalyticsX](https://learning.edx.org/course/course-v1:AdelaideX+AnalyticsX+1T2021/block-v1:AdelaideX+AnalyticsX+1T2021+type@sequential+block@7c4a7fce77314b78b64b633aa30df0e5/block-v1:AdelaideX+AnalyticsX+1T2021+type@vertical+block@2b6d8b8a8f2d4828a354c99bb00b2f96) course this value were used to determine how well the model fits the data based on the closer R-squared is to 1 the closer the data points are to the best fit line. This model has an R-squared value of 0.584 which is past the midpoint, but not great.
* **Linearity** – Looking at the *Residual vs Fitted* plot, it is a tough call to make if the assumption of linearity is correct. While it looks like a straight line might be a best way to fit the data, there is slight curve in the line, and there is a hint that there might be a funnel shape pattern. The data is also very dense making it hard to see clear patterns emerging. As an aside when drawing scatter plots and a fit line using the “gam” method specifically the *score vs num\_comments* showed that the line had a sharp downward bend on the right hand side.
* **Constant spread** – On the *Scale-Location* plot it does look that there is more or less am even spread of data points as we move from left to right. Note the strong curve though on the left hand side which might warrant further investigation.
* **Normality** – The Normal Q-Q curve again presents us with a challenge where the data looks like it might be sort of normally distributed, but not quite. To me the plot rather indicates that the distribution has “fat tails”.

*I found deciding if this model is a good fit extremely challenging even after consulting multiple sources (some references below).In my mind it actually looks like a poor model but I am not sure at all. Part of the problem is that there doesn’t seem to be clear answers (or at least unambiguous indicators with good rules of thumb to follow) and it seems more like an art than a science (I asked my wife with a MSc in statistics and she wasn’t much help either, but I blame that on perhaps her being a manger for too long). For example I would think R2 would be a good indicator to look at, but in one of the pages I read it told me that even if this value is low the model can still be a useful for prediction, so that didn’t help much. The graphs also seemed very ambiguous to me and I again could not find a reference giving me something like cut-off ranges to consider and something the like. It seems like for extreme cases it is easy to see when things go bad, but for everything in the middle it is very hard to decide.*

* [*https://learning.edx.org/course/course-v1:AdelaideX+AnalyticsX+1T2021/block-v1:AdelaideX+AnalyticsX+1T2021+type@sequential+block@4152bddc80774e3f82e23d5781d9388a/block-v1:AdelaideX+AnalyticsX+1T2021+type@vertical+block@992eb63b7f9e425eb2e7357abd28d8bf*](https://learning.edx.org/course/course-v1:AdelaideX+AnalyticsX+1T2021/block-v1:AdelaideX+AnalyticsX+1T2021+type@sequential+block@4152bddc80774e3f82e23d5781d9388a/block-v1:AdelaideX+AnalyticsX+1T2021+type@vertical+block@992eb63b7f9e425eb2e7357abd28d8bf)
* [*https://blog.learningtree.com/interpret-q-q-plot/*](https://blog.learningtree.com/interpret-q-q-plot/)
* [*https://data.library.virginia.edu/diagnostic-plots/*](https://data.library.virginia.edu/diagnostic-plots/)
* [*https://stats.stackexchange.com/questions/329167/linear-regression-with-seasonality*](https://stats.stackexchange.com/questions/329167/linear-regression-with-seasonality)
* [*https://www.statmethods.net/stats/regression.html*](https://www.statmethods.net/stats/regression.html)
* [*https://towardsdatascience.com/linear-regression-analysis-in-r-fdd59295d4a8*](https://towardsdatascience.com/linear-regression-analysis-in-r-fdd59295d4a8)
* [*http://varianceexplained.org/r/broom-intro/*](http://varianceexplained.org/r/broom-intro/)
* [*https://towardsdatascience.com/q-q-plots-explained-5aa8495426c0*](https://towardsdatascience.com/q-q-plots-explained-5aa8495426c0)

Include the name/s of any associated code (.r) script file/s related to Step 13 that are included in your .zip file:

* task01.R

1. Correctly identify all factors affected by aliasing in the regression model and update the model to remove those factors. Provide a list of the factors removed from the regression model and evidence of your code.  
   [3 points]

Identify variables to remove

# get a list of the factor names

factor\_names <- reddit %>%

  select\_if(is.factor) %>%

  names()

# function to match aliased coefficient back to original factor

find\_match <- function(value, list) {

  for (i in nchar(value):1) {

    current\_value = substr(value, 1, i)

    if (current\_value %in% list) {

      return(current\_value)

    }

  }

  return(NA)

}

# find coefficients with no value

term\_list <- names(coef(reddit.lm))[is.na(coef(reddit.lm))]

# match coefficients labels to factors

term\_list <- as\_tibble(term\_list)

term\_list$variable <- mapply(find\_match, term\_list$value, MoreArgs=list(list=factor\_names))

# get a list of the redundant variables

redundant\_variables <- unique(term\_list$variable)

redundant\_variables

Table . Predictor variables to remove.

|  |
| --- |
| [1] "author\_flair\_text" "crosspost\_parent\_list" "disable\_comments" "domain"  [5] "href\_url" "parent\_whitelist\_status" "preview" "promoted"  [9] "secure\_media\_embed" "subreddit" "subreddit\_id" "whitelist\_status" |

*\* Thinking about it, I am not sure certain that the subreddit predictor should be removed, especially when I ran the basic summary again on the updated model and it looked even worse than before. This left me wondering if something is going wrong.*

Remove redundant variables

update\_formula <- as.formula(

  paste(". ~ . -",

  paste(redundant\_variables, collapse=" -"), sep=""))

reddit.lm <- update(reddit.lm, update\_formula)

Include the name/s of any associated code (.r) script file/s related to Step 14 that are included in your .zip file:

* task01.R

1. Apply an appropriate statistical criterion to determine whether the variables *author\_cakeday* and *contest\_mode* should be included in the regression model. Clearly state your conclusion and provide evidence of your code.  
   [2 points]

# unchanged model

reddit.lm %>% glance %>% gather

# without author\_cakeday

update(reddit.lm, . ~ . -author\_cakeday)  %>% glance %>% gather

# without contest\_mode

update(reddit.lm, . ~ . -contest\_mode)  %>% glance %>% gather

# without author\_cakeday and contest\_mode

update(reddit.lm, . ~ . -author\_cakeday -contest\_mode)  %>% glance %>% gather

Table . Model comparison.

|  |  |  |  |
| --- | --- | --- | --- |
| reddit.lm | -author\_cakeday | -contest\_mode | -author\_cakeday  -contest\_mode |
| # A tibble: 12 x 2  key value  <chr> <dbl>  1 r.squared 0.551  2 adj.r.squared 0.549  3 sigma 0.958  4 statistic 366.  5 p.value 0  6 df 466  7 logLik -192211.  8 AIC 385358.  9 BIC 389967.  10 deviance 127977.  11 df.residual 139356  12 nobs 139823 | # A tibble: 12 x 2  key value  <chr> <dbl>  1 r.squared 0.551  2 adj.r.squared 0.549  3 sigma 0.958  4 statistic 367.  5 p.value 0  6 df 465  7 logLik -192212.  8 AIC 385358.  9 BIC 389957.  10 deviance 127978.  11 df.residual 139357  12 nobs 139823 | # A tibble: 12 x 2  key value  <chr> <dbl>  1 r.squared 0.551  2 adj.r.squared 0.549  3 sigma 0.958  4 statistic 367.  5 p.value 0  6 df 465  7 logLik -192211.  8 AIC 385356.  9 BIC 389955.  10 deviance 127977.  11 df.residual 139357  12 nobs 139823 | # A tibble: 12 x 2  key value  <chr> <dbl>  1 r.squared 0.551  2 adj.r.squared 0.549  3 sigma 0.958  4 statistic 368.  5 p.value 0  6 df 464  7 logLik -192212.  8 AIC 385356.  9 BIC 389945.  10 deviance 127978.  11 df.residual 139358  12 nobs 139823 |

Conclusion

* When measuring the model performance in terms of adj.r.squared removal of the variables individually or both have virtually no effect.
* Removal of the variables shows a very small improvement when looking at the AIC & BIC values, this does however look insignificant.
* The two variables will be removed from the model as their presence have almost no effect and it slightly decreases the model complexity.

reddit.lm <- update(reddit.lm, . ~ . -author\_cakeday -contest\_mode)

Include the name/s of any associated code (.r) script file/s related to Step 15 that are included in your .zip file:

* task01.R

1. Correctly obtain predictions and prediction intervals for the score in the two specified cases, part (a) and part (b). Provide evidence of your code.   
   [7 points]

Include the name/s of any associated code (.r) script file/s related to Step 16 that are included in your .zip file:

* task01.R

Total points possible for Task 1: Data cleaning and regression 60