# 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")

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> <chr> <chr> <chr> <chr> <chr> <chr> <chr>  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 <chr>, created\_utc <int>, crosspost\_parent <chr>,  # crosspost\_parent\_list <chr>, disable\_comments <chr>, distinguished <chr>, domain <chr>, domain\_override <chr>, edited <chr>,  # embed\_type <chr>, embed\_url <chr>, gilded <int>, hidden <chr>, hide\_score <chr>, href\_url <chr>, id <chr>, imp\_pixel <lgl>,  # is\_crosspostable <chr>, is\_reddit\_media\_domain <chr>, is\_self <chr>, is\_video <chr>, link\_flair\_css\_class <chr>,  # link\_flair\_text <chr>, locked <chr>, media <chr>, media\_embed <chr>, mobile\_ad\_url <chr>, num\_comments <int>,  # num\_crossposts <int>, original\_link <lgl>, over\_18 <chr>, parent\_whitelist\_status <chr>, permalink <chr>, pinned <chr>,  # post\_hint <chr>, preview <chr>, promoted <chr>, promoted\_by <lgl>, promoted\_display\_name <lgl>, promoted\_url <lgl>,  # retrieved\_on <int>, score <int>, secure\_media <chr>, secure\_media\_embed <chr>, selftext <chr>, spoiler <chr>,  # stickied <chr>, subreddit <chr>, subreddit\_id <chr>, suggested\_sort <chr>, third\_party\_trackers <chr>,  # third\_party\_tracking <lgl>, third\_party\_tracking\_2 <lgl>, thumbnail <chr>, thumbnail\_height <dbl>, thumbnail\_width <dbl>,  # title <chr>, url <chr>, whitelist\_status <chr> |

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

* task01.Rmd

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> <chr> <chr> <chr> <chr> <chr> <chr> <chr>  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 <chr>, created\_utc <int>, crosspost\_parent <chr>,  # crosspost\_parent\_list <chr>, disable\_comments <chr>, distinguished <chr>, domain <chr>, domain\_override <chr>, edited <chr>,  # embed\_type <chr>, embed\_url <chr>, gilded <int>, hidden <chr>, hide\_score <chr>, href\_url <chr>, id <chr>, imp\_pixel <lgl>,  # is\_crosspostable <chr>, is\_reddit\_media\_domain <chr>, is\_self <chr>, is\_video <chr>, link\_flair\_css\_class <chr>,  # link\_flair\_text <chr>, locked <chr>, media <chr>, media\_embed <chr>, mobile\_ad\_url <chr>, num\_comments <int>,  # num\_crossposts <int>, original\_link <lgl>, over\_18 <chr>, parent\_whitelist\_status <chr>, permalink <chr>, pinned <chr>,  # post\_hint <chr>, preview <chr>, promoted <chr>, promoted\_by <lgl>, promoted\_display\_name <lgl>, promoted\_url <lgl>,  # retrieved\_on <int>, score <int>, secure\_media <chr>, secure\_media\_embed <chr>, selftext <chr>, spoiler <chr>,  # stickied <chr>, subreddit <chr>, subreddit\_id <chr>, suggested\_sort <chr>, third\_party\_trackers <chr>,  # third\_party\_tracking <lgl>, third\_party\_tracking\_2 <lgl>, thumbnail <chr>, thumbnail\_height <dbl>, thumbnail\_width <dbl>,  # title <chr>, url <chr>, whitelist\_status <chr> |

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

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  <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <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 <chr>, crosspost\_parent\_list <chr>, disable\_comments <chr>,  # distinguished <chr>, domain <chr>, domain\_override <chr>, edited <chr>, embed\_type <chr>, embed\_url <chr>, gilded <int>,  # hidden <chr>, hide\_score <chr>, href\_url <chr>, id <chr>, is\_crosspostable <chr>, is\_reddit\_media\_domain <chr>,  # is\_self <chr>, is\_video <chr>, link\_flair\_css\_class <chr>, link\_flair\_text <chr>, locked <chr>, media <chr>,  # media\_embed <chr>, mobile\_ad\_url <chr>, num\_comments <int>, num\_crossposts <int>, over\_18 <chr>,  # parent\_whitelist\_status <chr>, permalink <chr>, pinned <chr>, post\_hint <chr>, preview <chr>, promoted <chr>,  # retrieved\_on <int>, score <int>, secure\_media <chr>, secure\_media\_embed <chr>, selftext <chr>, spoiler <chr>,  # stickied <chr>, subreddit <chr>, subreddit\_id <chr>, suggested\_sort <chr>, third\_party\_trackers <chr>, thumbnail <chr>,  # thumbnail\_height <dbl>, thumbnail\_width <dbl>, title <chr>, url <chr>, whitelist\_status <chr> |

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

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  <chr> <chr> <chr>  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  <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <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 <chr>, crosspost\_parent\_list <chr>, disable\_comments <chr>,  # distinguished <chr>, domain <chr>, domain\_override <chr>, edited <chr>, embed\_type <chr>, embed\_url <chr>, gilded <int>,  # hide\_score <chr>, href\_url <chr>, id <chr>, is\_reddit\_media\_domain <chr>, is\_self <chr>, is\_video <chr>,  # link\_flair\_css\_class <chr>, link\_flair\_text <chr>, locked <chr>, media <chr>, media\_embed <chr>, mobile\_ad\_url <chr>,  # num\_comments <int>, num\_crossposts <int>, over\_18 <chr>, parent\_whitelist\_status <chr>, permalink <chr>, post\_hint <chr>,  # preview <chr>, promoted <chr>, retrieved\_on <int>, score <int>, secure\_media <chr>, secure\_media\_embed <chr>,  # selftext <chr>, spoiler <chr>, stickied <chr>, subreddit <chr>, subreddit\_id <chr>, suggested\_sort <chr>,  # third\_party\_trackers <chr>, thumbnail <chr>, thumbnail\_height <dbl>, thumbnail\_width <dbl>, title <chr>, url <chr>,  # whitelist\_status <chr> |

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

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:

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]

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

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]

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

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]

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

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]

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

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]

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

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]

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

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]

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

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]

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

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]

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

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]

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

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:

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