Rwk9hw

Scrabble Words

```r
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------------------- tidyverse
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ---------------------------------------------------------------------- tidyverse_confli
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(dplyr)
```

- 1-1 How many words are there?

```r
scrabble_w <- read_csv("../R_data/words.txt")
```

```
## Parsed with column specification:
## cols(
##   word = col_character()
## )
```

```r
head(scrabble_w)
```

```
## # A tibble: 6 x 1
##   word
##   <chr>
## 1 AA
## 2 AAH
## 3 AAHED
## 4 AAHING
## 5 AAHS
## 6 AAL
```

```
scrabble_w[is.na(scrabble_w)] <- "NA"
scrabble_w %>%
  summarise(across(everything(), ~sum(is.na(.))))
```

```
## # A tibble: 1 x 1
##    word
##   <int>
## 1     0
```

```
nrow(scrabble_w)
```

```
## [1] 276643
```

- 1-2 How many words either begin or end in “X”?

```
scrabble_w %>%
  filter(str_detect(word, "^X") | str_detect(word, "X$")) %>%
  nrow()
```

```
## [1] 885
```

- 1-3 How many words contain all of the vowels?

```
scrabble_w %>%
  filter(str_detect(word, "A") & str_detect(word, "E") & str_detect(word, "I") & str_detect(word, "O") &
  scrabble_w_vowels
  nrow(scrabble_w_vowels)
```

```
## [1] 3476
```

- 1-4 What are the shortest words that contain all of the vowels?

```
scrabble_w_vowels %>%
  mutate(length = str_length(word)) %>%
  arrange(length) %>% # shortest is 7 letters
  filter(length == 7)
```

```
## # A tibble: 5 x 2
##   word    length
##   <chr>    <int>
## 1 DOULEIA      7
## 2 EULOGIA      7
## 3 MIAOUED      7
## 4 MOINEAU      7
## 5 SEQUOIA      7
```

- 1-5 Update the data frame to include a new column of words where you switch the first and last letters of all of the words and a second column to indicate if they are still valid words.

```
scrabble_w %>%
  mutate(switch_word = str_replace_all(word, "^([A-Z])(.*)([A-z])$", "\\3\\2\\1")) %>%
  mutate(still_word = switch_word %in% word) ->
  valid_word_check
head(valid_word_check)
```

```
## # A tibble: 6 x 3
##   word   switch_word still_word
##   <chr>  <chr>       <lgl>
## 1 AA     AA          TRUE
## 2 AAH    HAA         FALSE
## 3 AAHED  DAHEA       FALSE
## 4 AAHING GAHINA      FALSE
## 5 AAHS   SAHA        FALSE
## 6 AAL    LAA         FALSE
```

- 1-6 How many of the words that are still valid words after switching the first and last letters have different first and last letters?

```
valid_word_check %>%
  filter(still_word == TRUE) ->
  still_word_df # still words

still_word_df %>%
  filter(str_detect(word, "^(.)(.*)\\1$")) -> # same first and last letter
  same_FL
still_word_df %>%
  anti_join(same_FL) ->
  diff_FL
```

```
## Joining, by = c("word", "switch_word", "still_word")
```

```
head(diff_FL)
```

```
## # A tibble: 6 x 3
##   word  switch_word still_word
##   <chr> <chr>       <lgl>
## 1 AB    BA          TRUE
## 2 ABO   OBA         TRUE
## 3 AD    DA          TRUE
## 4 ADO   ODA         TRUE
## 5 AE    EA          TRUE
## 6 AH    HA          TRUE
```

```
nrow(diff_FL)
```

```
## [1] 1696
```

- 1-7 What are the longest words that are still words after switching the first and last letters and where the first and last letters are different?

```r
diff_FL %>%
  mutate(length = str_length(word)) %>%
  arrange(desc(length)) %>% # longest is 14 letters
  filter(length == 14)
```

```
## # A tibble: 6 x 4
##   word           switch_word    still_word length
##   <chr>          <chr>          <lgl>       <int>
## 1 DECOMMISSIONER RECOMMISSIONED TRUE           14
## 2 DEMYTHOLOGISER REMYTHOLOGISED TRUE           14
## 3 DEMYTHOLOGIZER REMYTHOLOGIZED TRUE           14
## 4 RECOMMISSIONED DECOMMISSIONER TRUE           14
## 5 REMYTHOLOGISED DEMYTHOLOGISER TRUE           14
## 6 REMYTHOLOGIZED DEMYTHOLOGIZER TRUE           14
```

- 1-8 Scrabble Scores
- 1-8-a

```r
score_word <- function(x){
  low <- c("A","E","I","O","U","D","L","M","N","R","S","T","Y")
  med <- c("B","C","F","G","H","K","P","W", "V")
  high <- c("J","Q","X","Z")
  points <- c(1,4,10)
  sum_score <- (str_count(x, "[AEIOUDLMNRSTY]")*1 + str_count(x, "[BCFGHKPWV]")*4 + str_count(x, "[JQXZ
}
scrabble_w %>%
  mutate(points = score_word(word)) ->
  scrabble_w_scores
head(scrabble_w_scores)
```

```
## # A tibble: 6 x 2
##   word   points
##   <chr>   <dbl>
## 1 AA          2
## 2 AAH         6
## 3 AAHED       8
## 4 AAHING     12
## 5 AAHS        7
## 6 AAL         3
```

- 1-8-b

```r
scrabble_w_scores %>%
  mutate(length = str_length(word)) %>%
  filter(length == 7) %>%
  slice_max(points, n=2)
```

```
## # A tibble: 8 x 3
##   word    points length
##   <chr>    <dbl>  <int>
## 1 FUZZBOX     40      7
## 2 JACUZZI     37      7
## 3 JAZZBOS     37      7
## 4 JAZZING     37      7
## 5 PIZAZZY     37      7
## 6 PZAZZES     37      7
## 7 ZIZZING     37      7
## 8 ZYZZYVA     37      7
```

- 1-8-c

```
# three highest scoring words with no vowels
scrabble_w_scores %>%
  filter(str_detect(word, "^[^AEIOU]+$")) %>%
  slice_max(points, n=3)
```

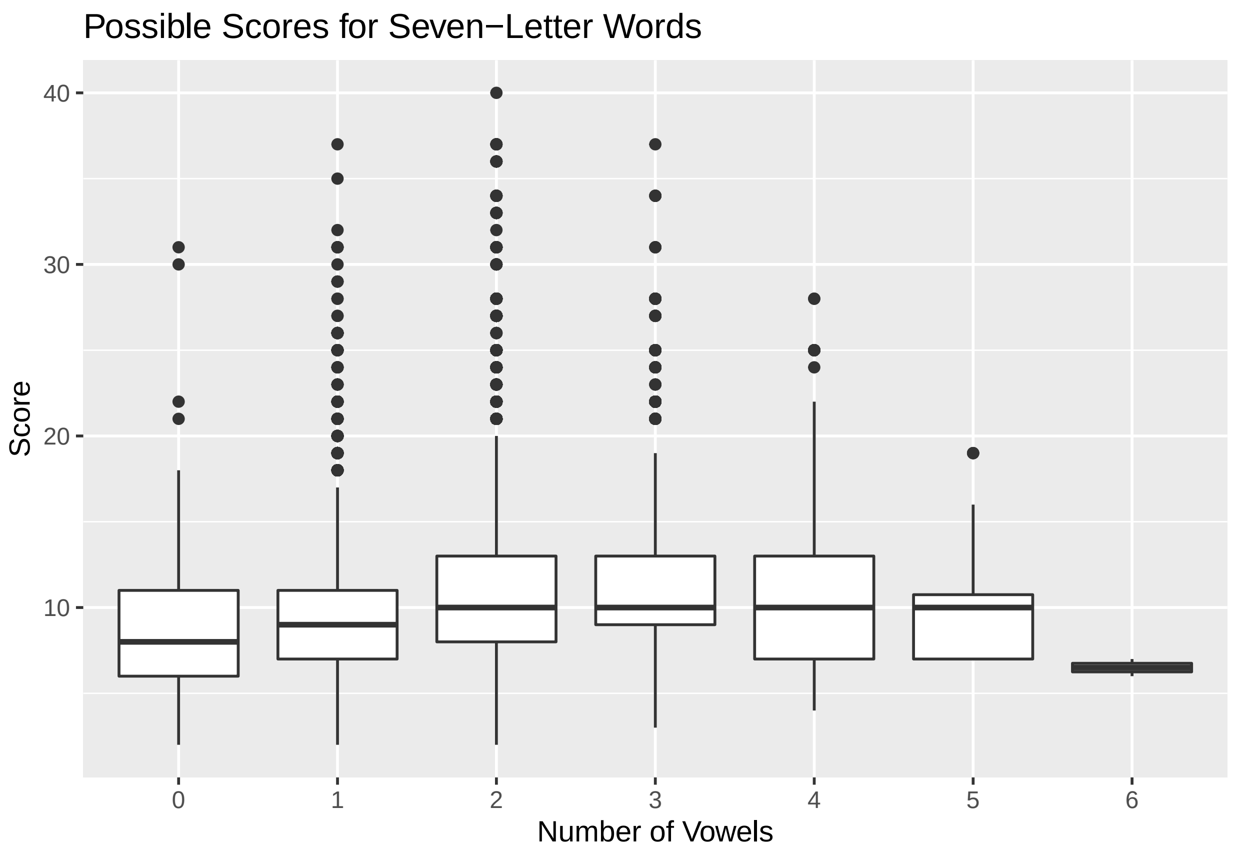
```
## # A tibble: 3 x 2
##   word  points
##   <chr>  <dbl>
## 1 ZZZS      31
## 2 ZZZ       30
## 3 JYNX      22
```

```
# three longest scoring words with no vowels
scrabble_w_scores %>%
  mutate(length = str_length(word)) %>%
  filter(str_detect(word, "^[^AEIOU]+$")) %>%
  arrange(desc(length)) %>%
  slice(1:3)
```

```
## # A tibble: 3 x 3
##   word    points length
##   <chr>    <dbl>  <int>
## 1 GLYCYLS     13      7
## 2 NYMPHLY     13      7
## 3 RHYTHMS     13      7
```

- 1-8-d

```
scrabble_w_scores %>%
  mutate(vowels_count = str_count(word, "[AEIOU]")) %>%
  mutate(length = str_length(word)) %>%
  filter(length <= 7) %>%
  ggplot(aes(x = as.factor(vowels_count),y = points)) +
  geom_boxplot() +
  xlab("Number of Vowels") +
  ylab("Score") +
  ggtitle("Possible Scores for Seven-Letter Words")
```



- 1-8-f Interpret: As the plot shows that 2-4 vowels in a word has approximately same average of score , which is higher than other words with less or more vowels.

Bank Data

- 2-1 show only how many rows there are in the data frame, Show a random sample of 2 percent of the rows.

```
bank_df <- read_csv("../R_data/fed_large_c_bank_ratings.csv")
```

```
## Parsed with column specification:
## cols(
##   name = col_character(),
##   rank = col_double(),
##   charter = col_character(),
##   consolidated_assets = col_double()
## )
```

```
nrow(bank_df)
```

```
## [1] 375
```

```
bank_df %>%
  slice_sample(prop = .02)
```

```
## # A tibble: 7 x 4
##   name                                      rank charter consolidated_assets
##   <chr>                                    <dbl> <chr>                 <dbl>
## 1 BANK7/BANK7 CORP                           753 SMB                     865
## 2 TEXAS CMNTY BK/VISION BSHRS                465 SMB                    1476
## 3 BANK OF NY MELLON/BANK OF NY MELLON CORP    10 SMB                  311387
## 4 DIETERICH BK/PRIME BANC CORP               801 SMB                     806
## 5 MIDWEST BK/WESTERN IL BSHRS               1155 SMB                     522
## 6 FIRST BK/FB CORP                           161 SMB                    6167
## 7 ROLLING HILLS B&T/ANITA BC                1744 SMB                     318
```

- 2-2

```
bank_df %>%
  separate(name,
           into = c("name", "alternate_name"),
           sep = "/",
           extra = "drop") ->
  bank
head(bank)
```

```
## # A tibble: 6 x 5
##   name                 alternate_name           rank charter consolidated_assets
##   <chr>                <chr>                   <dbl> <chr>                 <dbl>
## 1 BANK OF NY MELLON    BANK OF NY MELLON CORP     10 SMB                  311387
## 2 STATE STREET B&TC    STATE STREET CORP          11 SMB                  242148
## 3 GOLDMAN SACHS BK USA GOLDMAN SACHS GROUP THE    12 SMB                  228836
## 4 ALLY BK              ALLY FNCL                  15 SMB                  167492
## 5 NORTHERN TC          NORTHERN TR CORP           20 SMB                  135885
## 6 REGIONS BK           REGIONS FC                 22 SMB                  125641
```

- 2-3 How many bank primary names begin with a digit?

```
bank %>%
  filter(str_detect(name, "^\\d")) %>%
  nrow()
```

```
## [1] 2
```

- 2-4-a How many of the bank primary names have the letters “BANK” in them? “BANKING” counts

```
bank %>%
  filter(str_detect(name, "BANK")) %>%
  nrow()
```

```
## [1] 41
```

- 2-4-b How many of the bank primary names have the stand-alone word “BANK” in them? “BANKING” does not count

```
bank %>%
  filter(str_detect(name, "^BANK\\s") | str_detect(name, "\\sBANK\\s") | str_detect(name, "\\sBANK$")) %
  nrow()
```

```
## [1] 21
```

- 2-5-a

```
bank %>%
  mutate(name = str_replace_all(name, "BK", "BANK")) ->
  bank_newname
head(bank_newname)
```

```
## # A tibble: 6 x 5
##   name                   alternate_name           rank charter consolidated_asse~
##   <chr>                  <chr>                   <dbl> <chr>                <dbl>
## 1 BANK OF NY MELLON      BANK OF NY MELLON CORP     10 SMB                 311387
## 2 STATE STREET B&TC      STATE STREET CORP          11 SMB                 242148
## 3 GOLDMAN SACHS BANK U~  GOLDMAN SACHS GROUP THE    12 SMB                 228836
## 4 ALLY BANK              ALLY FNCL                  15 SMB                 167492
## 5 NORTHERN TC            NORTHERN TR CORP           20 SMB                 135885
## 6 REGIONS BANK           REGIONS FC                 22 SMB                 125641
```

- 2-5-b

```
bank_newname %>%
  mutate(position =
  ifelse(str_detect(name, "^BANK"), "start",
  ifelse(str_detect(name, "BANK$"), "end",
  ifelse(str_detect(name, "\\s(.*)BANK(.*)\\s"), "middle", "none")))) ->
  bank_wposition
head(bank_wposition)
```

```
## # A tibble: 6 x 6
##   name              alternate_name        rank charter consolidated_ass~ position
##   <chr>             <chr>                <dbl> <chr>               <dbl> <chr>
## 1 BANK OF NY MELL~  BANK OF NY MELLON C~    10 SMB                311387 start
## 2 STATE STREET B&~  STATE STREET CORP       11 SMB                242148 none
## 3 GOLDMAN SACHS B~  GOLDMAN SACHS GROUP~    12 SMB                228836 middle
## 4 ALLY BANK         ALLY FNCL               15 SMB                167492 end
## 5 NORTHERN TC       NORTHERN TR CORP        20 SMB                135885 none
## 6 REGIONS BANK      REGIONS FC              22 SMB                125641 end
```

- 2-5-c

```
bank_wposition %>%
  group_by(position) %>%
  summarise(prop = n()/nrow(bank_wposition))
```

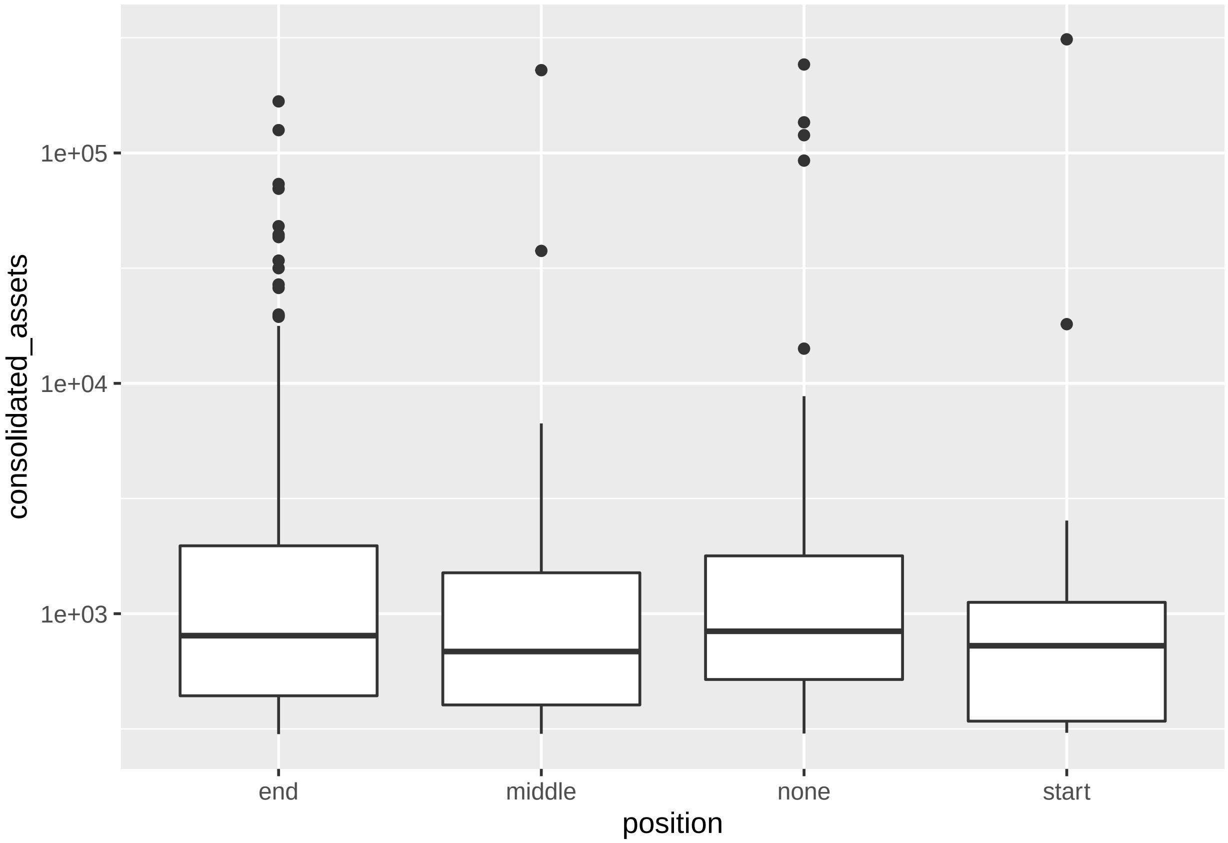
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 4 x 2
##   position  prop
##   <chr>    <dbl>
## 1 end      0.691
## 2 middle   0.107
## 3 none     0.131
## 4 start    0.072
```

- 2-6 Interpret: The position of the word “BANK” doesn’t have significant relationship to the log of total assets.

```
bank_wposition %>%
  ggplot(aes(x = position, y = consolidated_assets)) +
  geom_boxplot() +
  scale_y_log10()
```

```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```


consolidated_assets
1e+05
1e+04
1e+03
end
middle
none
start
position