Final Project

Lin Cheng

7/30/2019

Data Import

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Sentiment Code

```
library(dplyr)
library(tidytext)
library(textdata)
test <- Amazon5 %>%
    unnest_tokens(word, review_body) %>%
    count(word, sort = TRUE) %>%
    ungroup() %>%
    inner_join(get_sentiments("afinn"), by = "word") %>%
    group_by(word) %>%
    mutate(contribution = value*n) %>%
    arrange(desc(abs(contribution)))
```

```
test <- Amazon5[1,] %>%
    unnest_tokens(word, review_body)
```

```
test <- Amazon5[1:100,] %>%
    unnest_tokens(word, review_body) %>%
    count(word, sort = TRUE)
```

```
test %>%
    ungroup() %>%
    inner_join(get_sentiments("afinn"), by = "word") %>%
    group_by(word) %>%
    mutate(contribution = value*n) %>%
    arrange(desc(abs(contribution)))
```

```
sentiment  <-  Amazon5 %>%
    unnest_tokens(word, review_body) %>%
    inner_join(get_sentiments("afinn"), by = "word") %>%
    group_by(review_id) %>%
    summarize(sentiment = mean(value), words = n()) #%>%
    #filter(words >= 5)
```

```
sentiment %>% arrange(desc(sentiment))
```

```
## # A tibble: 418,279 x 3
##    review_id      sentiment words
##    <chr>              <dbl> <int>
##  1 R104R4K3XBSM1L         5     1
##  2 R1056S7HZGSBH          5     1
##  3 R10GWDDIL9VDTP         5     1
##  4 R10HOG9H099F4Z         5     1
##  5 R10IVYXUCXSIUV         5     1
##  6 R10NT8Q2ZVQ793         5     1
##  7 R11M5V0EWUYBQ6         5     1
##  8 R11NJ5MJ4NEQJH         5     1
##  9 R123827INJW25H         5     1
## 10 R12H5GP85S9V6H         5     1
## # … with 418,269 more rows
```

1 Statistical Summary

```
summary(Amazon5)
```

```
##  customer_id.x       review_id          star_rating    review_body
##  Min.   :   10291   Length:448511      Min.   :1.000   Length:448511
##  1st Qu.:15034588   Class :character   1st Qu.:4.000   Class :character
##  Median :27877844   Mode  :character   Median :5.000   Mode  :character
##  Mean   :28794823                      Mean   :4.456
##  3rd Qu.:43202158                      3rd Qu.:5.000
##  Max.   :53096401                      Max.   :5.000
##
##   review_date             vine            product_id
##  Min.   :2012-05-03   Length:448511      Length:448511
##  1st Qu.:2013-06-20   Class :character   Class :character
##  Median :2014-05-11   Mode  :character   Mode  :character
##  Mean   :2014-03-31
##  3rd Qu.:2015-01-13
##  Max.   :2015-08-31
##  NA's   :57
##  review_headline    customer_id.y      product_parent
##  Length:448511      Min.   :   10291   Length:448511
##  Class :character   1st Qu.:15034588   Class :character
##  Mode  :character   Median :27877844   Mode  :character
##                     Mean   :28794823
##                     3rd Qu.:43202158
##                     Max.   :53096401
##
##  helpful_votes       total_votes       verified_purchase
##  Min.   :   0.000   Min.   :   0.000   Length:448511
##  1st Qu.:   0.000   1st Qu.:   0.000   Class :character
##  Median :   0.000   Median :   0.000   Mode  :character
##  Mean   :   1.693   Mean   :   2.495
##  3rd Qu.:   1.000   3rd Qu.:   2.000
##  Max.   :2893.000   Max.   :3589.000
##
##  product_title
##  Length:448511
##  Class :character
##  Mode  :character
##
##
##
##
```

```
glimpse(Amazon5)
```

```
## Observations: 448,511
## Variables: 14
## $ customer_id.x     <dbl> 35112398, 20421275, 50211175, 51401494, 26978…
## $ review_id         <chr> "R16XFH1LI30ZSW", "R39IJD7J9NRKC2", "R1WXQMAV…
## $ star_rating       <dbl> 4, 5, 4, 5, 5, 3, 5, 1, 4, 3, 5, 5, 5, 4, 5, …
## $ review_body       <chr> "Though it was not the book I thought it was,…
## $ review_date       <date> 2013-10-04, 2015-07-19, 2014-03-02, 2015-03-…
## $ vine              <chr> "N", "N", "N", "N", "N", "N", "N", "N", "Y", …
## $ product_id        <chr> "0965915905", "0205823149", "0763660531", "09…
## $ review_headline   <chr> "Ordered by mistake, still happy", "Five Star…
## $ customer_id.y     <dbl> 35112398, 20421275, 50211175, 51401494, 26978…
## $ product_parent    <chr> "350345906", "364677111", "954675172", "59705…
## $ helpful_votes     <dbl> 0, 0, 0, 1, 0, 1, 1, 2, 1, 0, 1, 0, 1, 0, 4, …
## $ total_votes       <dbl> 1, 0, 1, 1, 0, 2, 1, 19, 1, 0, 1, 0, 1, 1, 4,…
## $ verified_purchase <chr> "Y", "Y", "Y", "Y", "Y", "Y", "Y", "N", "N", …
## $ product_title     <chr> "The Alamo: An Illustrated History", "Allyn &…
```

Note

- 1. There are 448511 observation and 14 variables. There are 4 numeric variables and 10 categorical variables. Though all numerical variables are useful, only a portion is meaningful.
  - Meaningful numerical variables: star_rating, review_date, helpful_votes and total_votes
  - Meaningful categorical variables: review_body, vine, review_headline, verified_purchase and product_title
- 2. The average star rating is high (4.456). Rating is in range 1-5.
- 3. The year of data range from 2012 - 2015.
- 4. There are some outliers in helpful/total votes. For example, helpful votes have a max of 2893 while the mean is 1.693.
- 5. There are 57 missing dates. Other than this, there is no na's for all other variables.
- 6. Review id is unique, while there can be duplicates for customer id (a customer can write multiple reviews).

(a)

```
Amazon5$product_title<-sapply(Amazon5$product_title, factor)
```

```
library(dplyr)
top_2<-Amazon5 %>%
  group_by(product_title) %>%
  summarise(star_rating = mean(star_rating),count=n()) %>%
  arrange(desc(star_rating)) %>%
  filter(count>25) %>%
  top_n(2, star_rating)
bottom_2<-Amazon5 %>%
  group_by(product_title) %>%
  summarise(star_rating = mean(star_rating),count=n()) %>%
  arrange(desc(star_rating)) %>%
  filter(count>25) %>%
  top_n(-2, star_rating)
bind<-bind_rows(top_2,bottom_2)
bind
```

```
## # A tibble: 4 x 3
##   product_title                                      star_rating count
##   <fct>                                                    <dbl> <int>
## 1 Pete the Cat: I Love My White Shoes                       4.98    42
## 2 Carry On, Warrior: Thoughts on Life Unarmed               4.97    35
## 3 To Train Up a Child                                       1.66    58
## 4 It Could Happen To Anyone: Why Battered Women Stay        1.42    65
```

Ordered from highest to lowest (decreasing order)

(b) Statistical Summary (The four books from part(a))

```
library(dplyr)
sentiment_2  <-  Amazon5 %>%
    unnest_tokens(word, review_body) %>%
    inner_join(get_sentiments("afinn"), by = "word") %>%
    group_by(product_title) %>%
    summarize(sentiment = mean(mean(value)), words = n()) #%>%
    #filter(words >= 5)
summary(sentiment_2)
```

```
##
product_title
##  The Alamo: An Illustrated History
:     1
##  Allyn & Bacon Guide to Writing, The, Concise Edition (6th Edition)
:     1
##  Journey
:     1
##  Never In Your Wildest Dreams: A Transformational Story to Tap Into Your Hidden Gifts
to Create a Life of Passion:     1
##  A Passion for the Impossible: The Life of Lilias Trotter
:     1
##  The Detox Prescription: Supercharge Your Health, Strip Away Pounds, and Eliminate th
e Toxins Within             :     1
##  (Other)
:203339
##    sentiment           words
##  Min.   :-5.0000   Min.   :   1.00
##  1st Qu.: 0.9787   1st Qu.:   2.00
##  Median : 1.7778   Median :   5.00
##  Mean   : 1.6240   Mean   :  13.48
##  3rd Qu.: 2.5000   3rd Qu.:  13.00
##  Max.   : 5.0000   Max.   :3962.00
##
```

```
books<-left_join(bind, sentiment_2, by= "product_title")
books[,1:4]
```

```
## # A tibble: 4 x 4
##   product_title                                    star_rating count sentiment
##   <fct>                                                  <dbl> <int>     <dbl>
## 1 Pete the Cat: I Love My White Shoes                     4.98    42     2.09
## 2 Carry On, Warrior: Thoughts on Life Unarmed             4.97    35     1.48
## 3 To Train Up a Child                                     1.66    58    -0.420
## 4 It Could Happen To Anyone: Why Battered Wome…           1.42    65    -0.906
```

###Note * The column named ‘sentiment’ is the average sentiment score for each book. * The books with higher average ratings have higher average sentiment scores related to the words written in the reviews. * It seems like the lower rating books have more reviews than the higher rating ones. This suggests an idea that maybe customers are more willing to inform others of a product they dislike instead of a product that they like.

2 Visualization (similar to Lab 2)

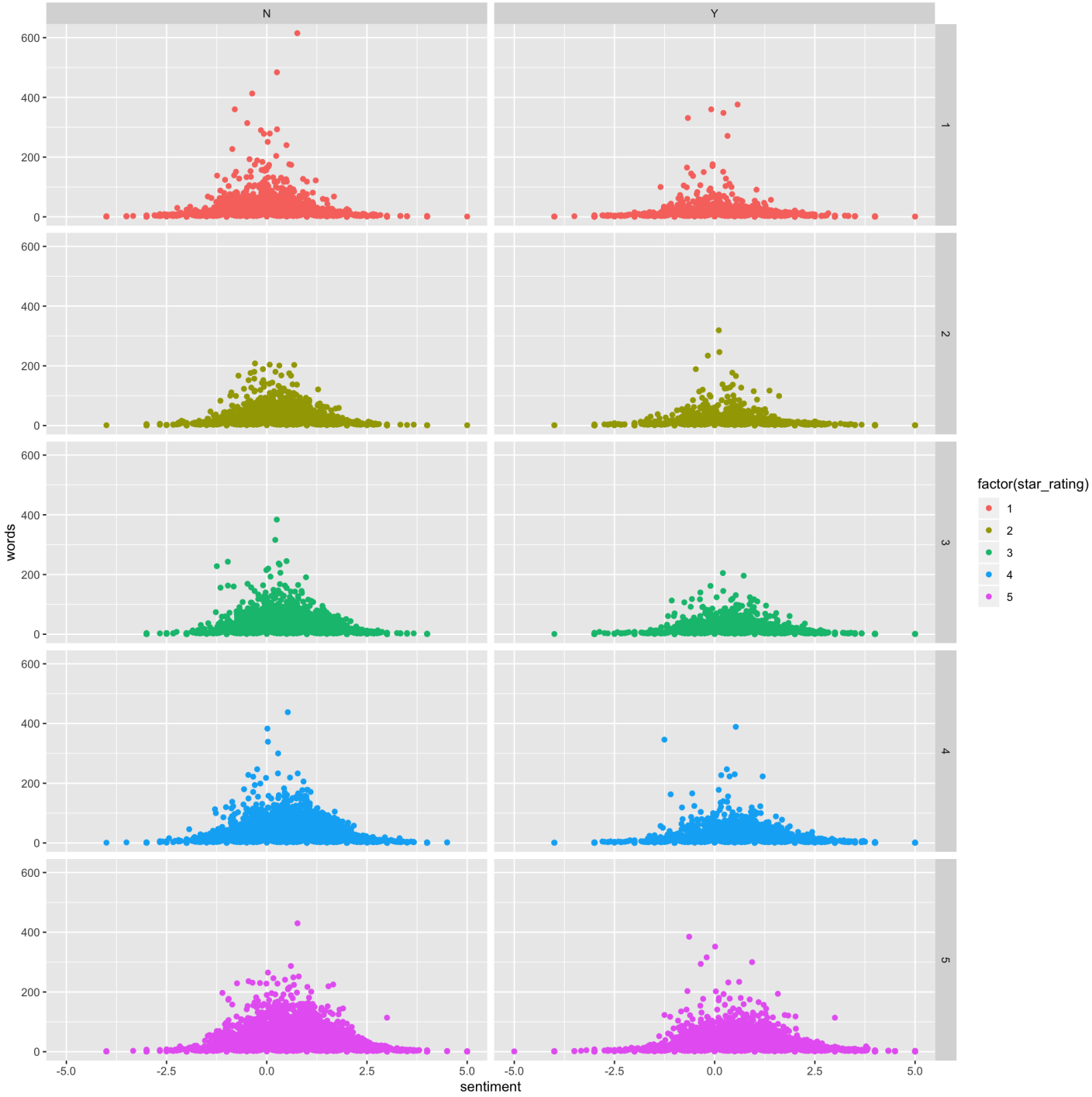
```
library(ggplot2)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

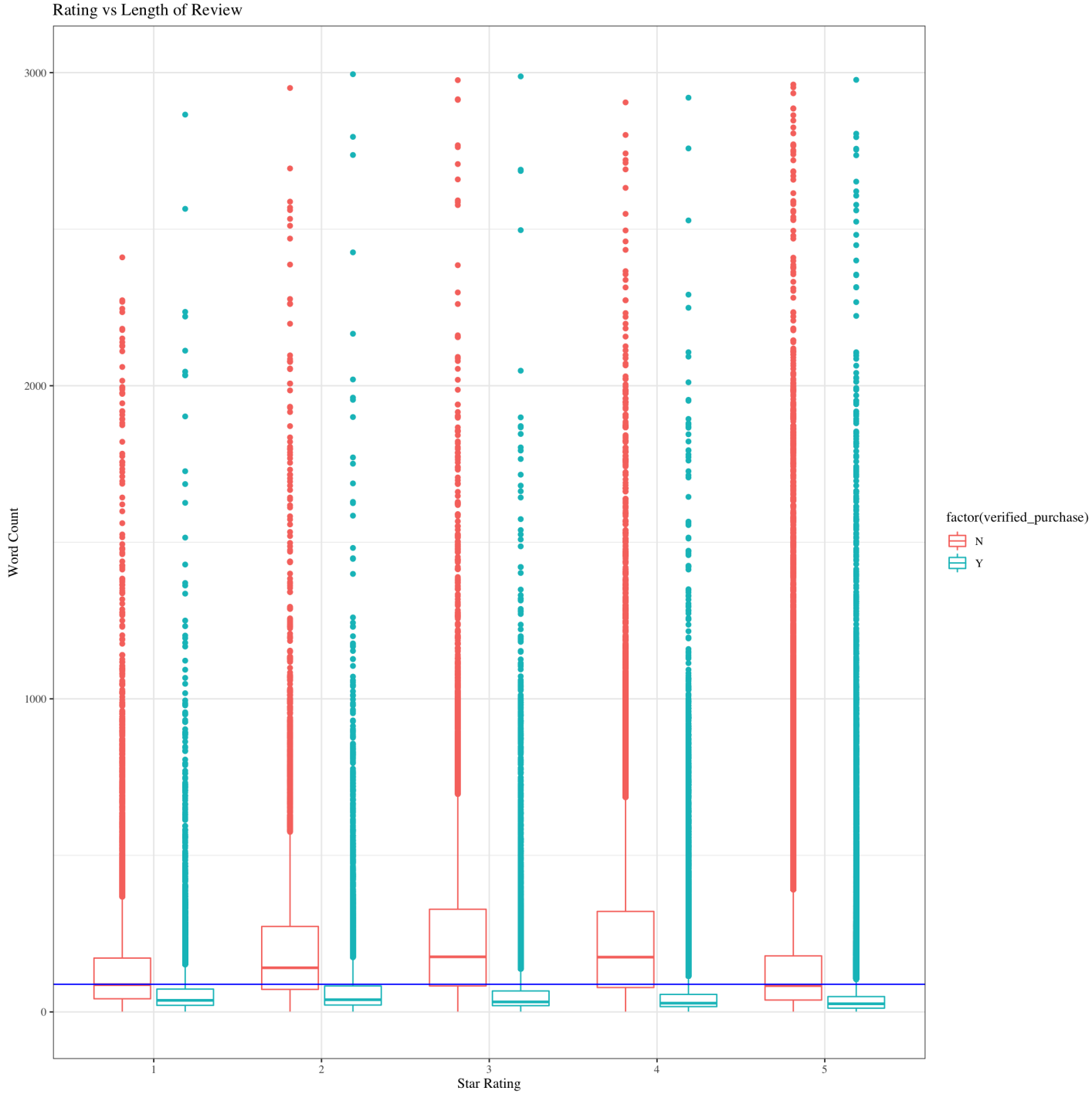
```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(dplyr)
Amazon_2 <- left_join(Amazon5,sentiment, by="review_id")
g <- ggplot(Amazon_2, aes(y=words,x=sentiment,color=factor(star_rating)))+
  geom_point() +
  facet_grid(star_rating ~ verified_purchase)
g
```

```
## Warning: Removed 30232 rows containing missing values (geom_point).
```



Note * horizontal facet + N: unverified purchase Y: verified purchase * vertical facet + 1-5: star_rating * This graph contains four variables: words, sentiment, star_rating, and verified_purchase. * The main point of the graph is show that the reviews written by customers with #verified# purchase, suprisingly, has a much higher word count than the reviews written by customers with #unverified# purchase. This suggests that people who have actually bought the products are less likely to write a lengthy review, and that there might be fake reviews which deviates the star rating from its true value. This could also affect review contents that real customers perceive. * In addition, we can see that most of the reviews have sentiment scores center around the middle (sentiment score=0). This suggests that customers like to give a neutral statement about the product, which is find but do not actually give any useful information.



Note * This is a complementary graph to the above one, showing the word count for each star rating, and with a third variable – verified_purchase. It clealy shows that the number of words are higher in unverified purchase

- The horizontal blue line is an indication of the mean word count (overall). It clearly shows that the box (representing lower and upper quartile values, and the median value) for verified purchases are below the blue line while box for unverified purchases are at most time at or above the blue line

3 Insight

- The two key factors a company like Amazon strives for are customer satisfaction and sales. Which part to focus on all depends on the company's value. Thus, in deiciding the proportion of each type of book it sells, we have

- 1. Satisfaction: Focus on the star_rating and sentiment score of reviews.
- 2. Sales/Popularity: While some books might have very high star rating, it might not be popular and thus not a lot of people would buy them. On the other hand, some books with moderate reviews might be quite popular to the general public (broader audience). Therefore, it would be too superficial to simply look at the star rating because the ultimate goal of Amazon is to achieve higher total revenue.

i Hypothesis

- My hypothesis is the the average word count for (verified)reviews and (unverified)reviews are significantly different from each other, indicating there are some factor contributing to such differences which the company to take notice of. To compare two means, I will be using the ANOVA test

ii Methodology: how I am going to do it

- ANOVA
- null hypothese: the word count for verified purchase and un-verified purchase are the same
- alternative: not equal appropriate
- Test the difference between two means, in this case they will be the average word count for verified purchase and unverified purchase. Using the ANOVA test, analyze variance and then make infereences about the mean. The null hypothesis would be that the two mean are equal. If the ANOVA test reject the null hypothesis, then we would get that the two mean are different from each other. In that case, it can support my hypothesis that the word count actually differs for reviews written by customers with verified purchase and unverified purchase.

iii Pseudocode: how I think the code should look

- Two means:
  - 1. mean word count for verified purchase
  - 2. mean word count for un-verified purchase *null hypothesis : mean 1 = mean 2* alternative hypotheise: mean 1 != mean 2
- Extract the data subset from Amazon5S which contains the variables star_rating, review_body, and verified_purchase.
- Drop the na's to drop any rows containing na's (though there is no missing value in this data set)
- Using dplyr, add a column (mutate) named 'word_count' which takes in Amazon5S$review_body, split the strings into separate words, and output the length of each review (use the sapply function to quickly does this).
- Group the data into two parts: verified and unverified | Calculate the average word count for verified purchase and unverified purchase.
- Use the function aov() with the calculated means as argument, save it as 'Amazon_aov'.
- summary(Amazon_aov)
- Look at the F value and p value. If the p-value is less than 0.05, then we reject the null hypothesis and accept the alternative hypothese. On the other hand, if the p-value is higher than 0.05, then we accept the null hypothesis. Since there are only two means, there is no need for a post HOC test.

iv R Code

```
library(tidyr)
#Amazon_1<-select(Amazon5,review_id,review_body,star_rating,verified_purchase)
#Amazon_1<-drop_na(Amazon_1)
#Amazon_1<-dplyr::mutate(Amazon_1, word_count = sapply(strsplit(Amazon_1$review_body, "
 "), length))
Amazon_aov<-aov(word_count ~ star_rating, data = Amazon_1)
summary(Amazon_aov)
```

```
##                 Df    Sum Sq   Mean Sq F value Pr(>F)
## star_rating      1 1.585e+08 158452297    5346 <2e-16 ***
## Residuals   448499 1.329e+10     29639
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The p-value (<2e-16) is less than the significance level 0.05, so we can conclude that there are significance differences between the two means.