William D. Gizzi

a.

```
> food_df=scan("~/Documents/datascience/ds710fall2017assignment11/food_clean.csv",
+         what = list(
+         "Number of people who voted this review helpful" = numeric(),
+         "Product ID"=character(),
+         "Rating of the product" =numeric(),
+         "Total number of people who rated this review"= numeric(),
+         "Length of review in characters"= numeric(),
+         "Number of exclamation points"= numeric(),
+     "Fraction of people who rated this review helpful" = numeric()
+ ),
+         ,sep=",", skip =1)
```

Read 568454 records

b.

```
> food_df=data.frame(food_df)
```

```
> food_df.header=scan("~/Documents/datascience/ds710fall2017assignment11/food_clean.csv",
what=character(),sep=",", nlines =1)
```

Read 7 items

```
> colnames(food_df)=food_df.header
```

c.

```
> sapply(food_df, is.numeric)
```

All of the columns that needed to be numeric were already numeric because I specified so when I used scan().

d.

```
unrealistic=which(food_df$"Fraction of people who rated this review helpful">1)
```

```
> head(unrealistic)
```

```
[1]  33  34  83 159 214 288
```

```
> food_df$"Fraction of people who rated this review helpful"[unrealistic] <- NA
```

> food_df$"Number of people who voted this review helpful"[unrealistic] <- NA

> food_df$"Total number of people who rated this review"[unrealistic] <- NA

e.

My criteria for unrealistic is any row that has the "Fraction of people who rated this review helpful" greater than 1. This is unrealistic because it is impossible to have more review votes than actual reviews. There were 22,716 unrealistic entries in this dataset.

> length(unrealistic)

[1] 22716

f.

> helpful = which(food_df$"Fraction of people who rated this review helpful">0.5)

> head(helpful)

[1] 1 3 4 9 11 12

> nothelpful = which(food_df$"Fraction of people who rated this review helpful"<=0.5)

> head(nothelpful)

[1] 27 28 32 48 49 50

g.

> helpfulreview = food_df$"Length of review in characters"[helpful]

> nothelpfulreview = food_df$"Length of review in characters"[nothelpful]

> t.test(helpfulreview, nothelpfulreview)

Welch Two Sample t-test

data: helpfulreview and nothelpfulreview

t = 16.9, df = 114550, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

 30.60424 38.63405

sample estimates:

mean of x mean of y

478.5387 443.9196

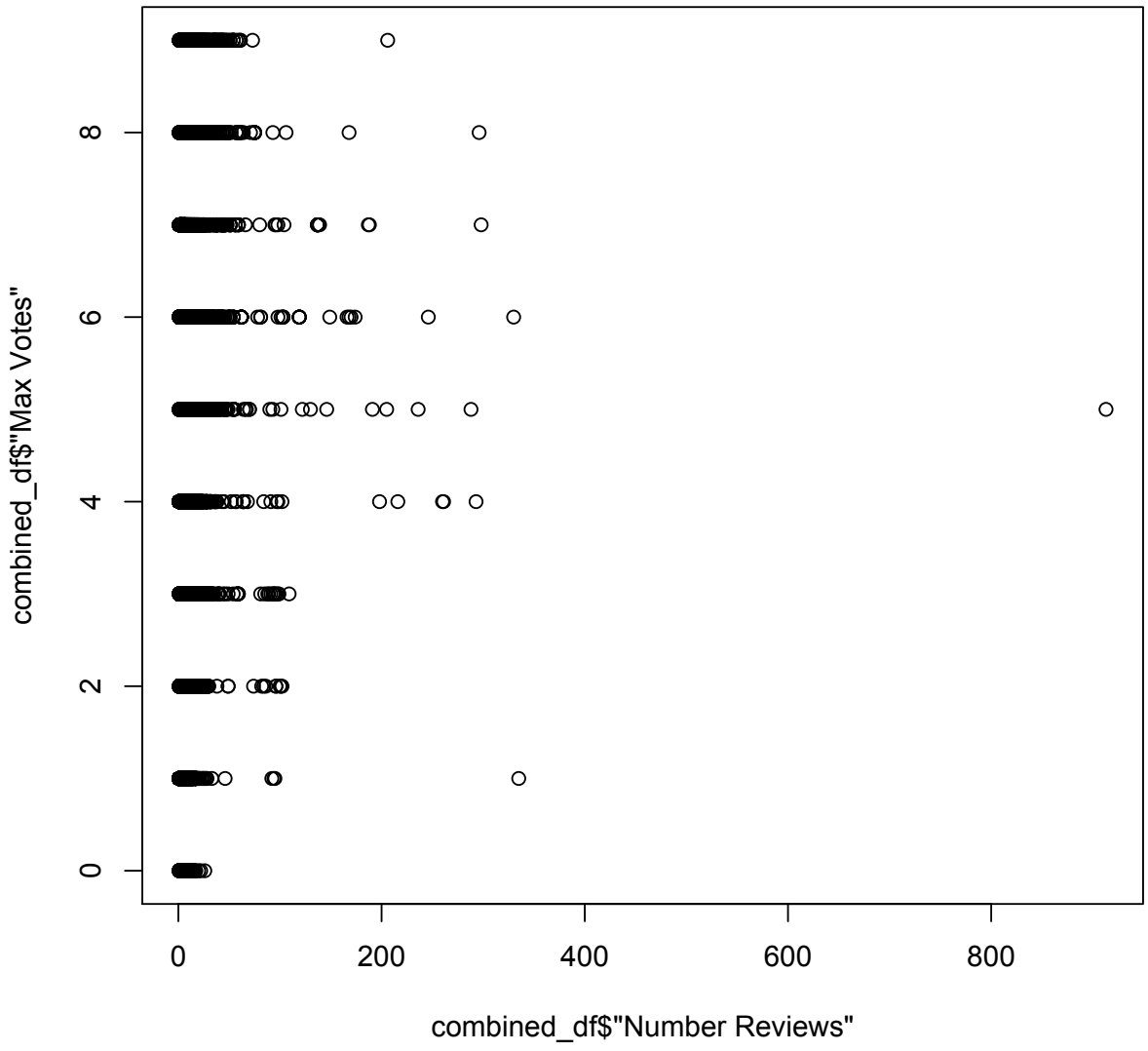
At a p-value of 0.05, it can be concluded that helpful reviews typically have a greater character length than non-helpful reviews.

h.

```
> maxVotes = tapply(food_df$"Number of people who voted this review helpful",food_df$"Product ID",max)

> maxVotes_df <- data.frame(names(maxVotes), as.vector(maxVotes))

> colnames(maxVotes_df) <- c('Product ID', 'Max Votes')

> reviewCount <- table(food_df$"Product ID")

> reviewCount_df <- data.frame(names(reviewCount), as.vector(reviewCount))

> colnames(reviewCount_df) <- c('Product ID', 'Number Reviews')

> combined_df <- merge(reviewCount_df, maxVotes_df, by="Product ID")

> head(combined_df)

    Product ID Number Reviews Max Votes
1  0006641040\n           37        NA
2  141278509X\n            1         1
3  2734888454\n            2         1
4  2841233731\n            1         0
5  7310172001\n          173        NA
6  7310172101\n          173        NA
```

i.

```
> plot(combined_df$"Number Reviews", combined_df$"Max Votes")
```



There doesn't seem to be much of a visible trend between number of reviews and max votes.

j.

```
> maxVotes2_df = combined_df[which(combined_df$"Max Votes" > 0), ]

> head(maxVotes2_df)

     Product ID Number Reviews Max Votes

2   141278509X\n           1         1
```

3 2734888454\n 2 1

9 B00002N8SM\n 38 4

25 B00005344V\n 17 8

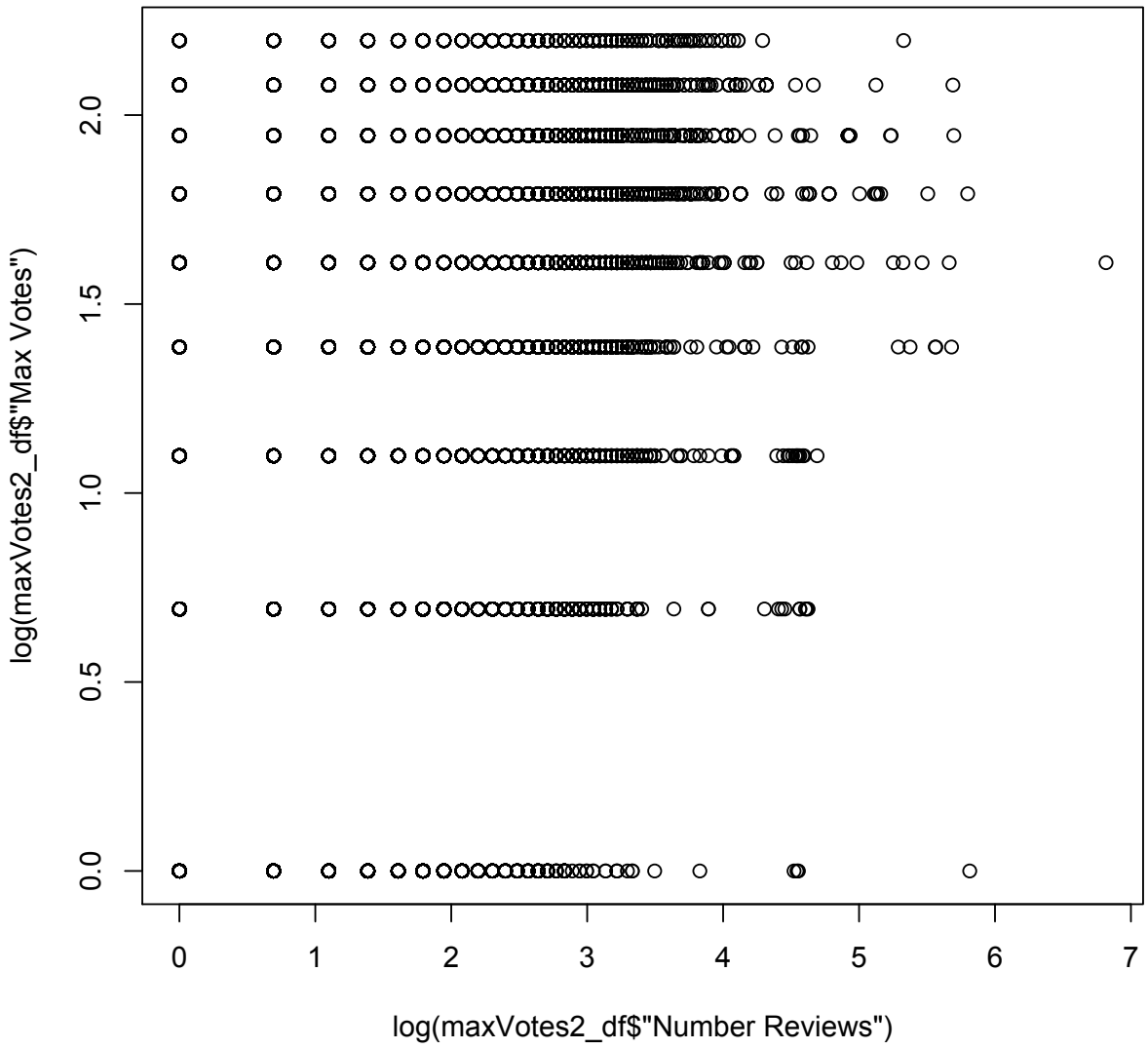
27 B0000537KC\n 16 3

34 B00005OMWO\n 4 2

k.

> plot(log(maxVotes2_df$"Number Reviews"), log(maxVotes2_df$"Max Votes"))

GRAPH IS ON FOLLOWING PAGE



Above, there appears to be a positive correlation between number of reviews and max votes. Generally, as the number of reviews increases, the max votes increases as well. Thus, we can infer that the untransformed number of reviews and untransformed max votes share a similar positive correlation.