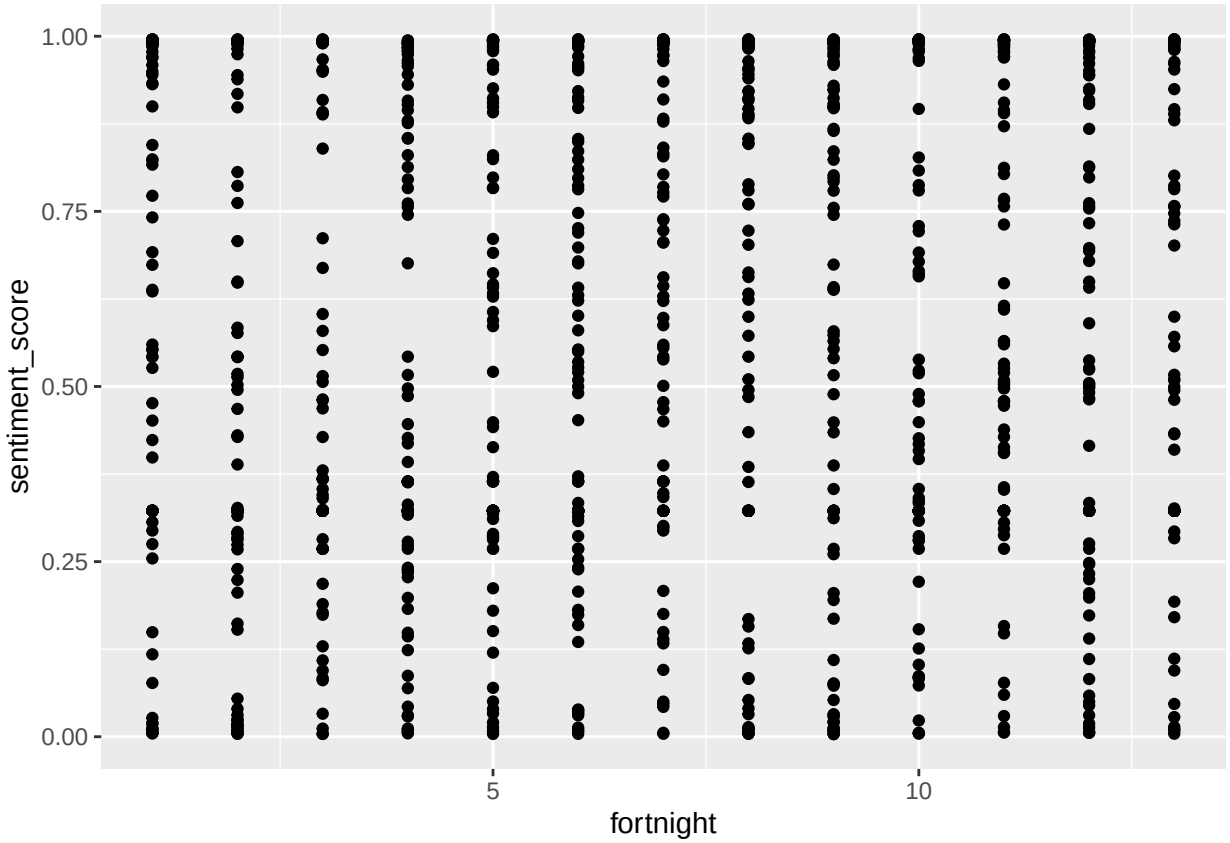
Datafest Data Analysis India

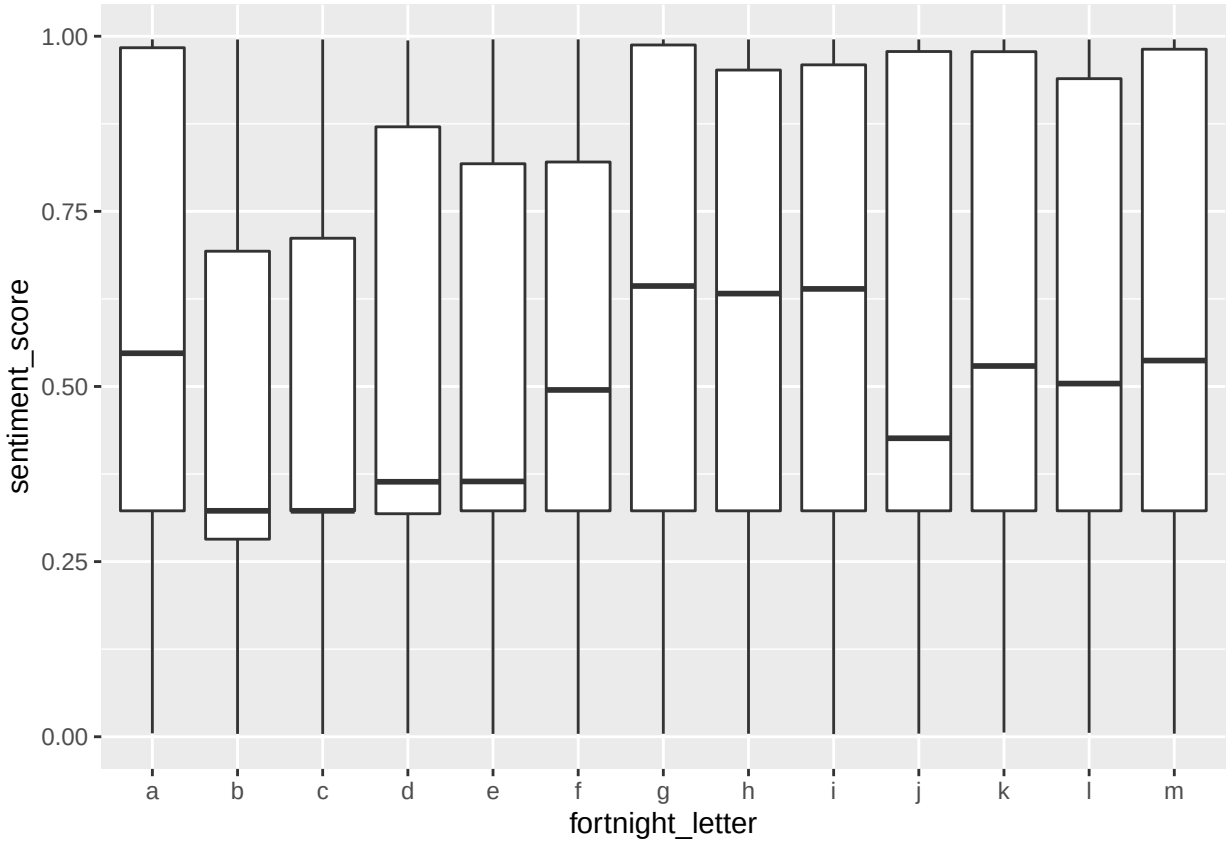
```r
#India analysis import
India_analysis = read_excel("C:\\Users\\gtham\\OneDrive - Pomona College\\A - DATAFEST\\Analysis Datase
```

```r
India_analysis_music <- India_analysis %>%
  filter(video_category == "Music")

India_analysis_travel <- India_analysis %>%
  filter(video_category == "Travel and Events")

India_analysis_people <- India_analysis %>%
  filter(video_category == "People and Blogs")

India_analysis_entertainment <- India_analysis %>%
  filter(video_category == "Entertainment")

India_analysis_news <- India_analysis %>%
  filter(video_category == "News and Politics")

India_analysis_how_to <- India_analysis %>%
  filter(video_category == "How-to and Style")

India_analysis_education <- India_analysis %>%
  filter(video_category == "Education")

India_analysis_science <- India_analysis %>%
  filter(video_category == "Science and Technology")
```

```r
#full India data data summaries
ggplot(India_analysis) +
  geom_point(aes(x = fortnight, y = sentiment_score))
```



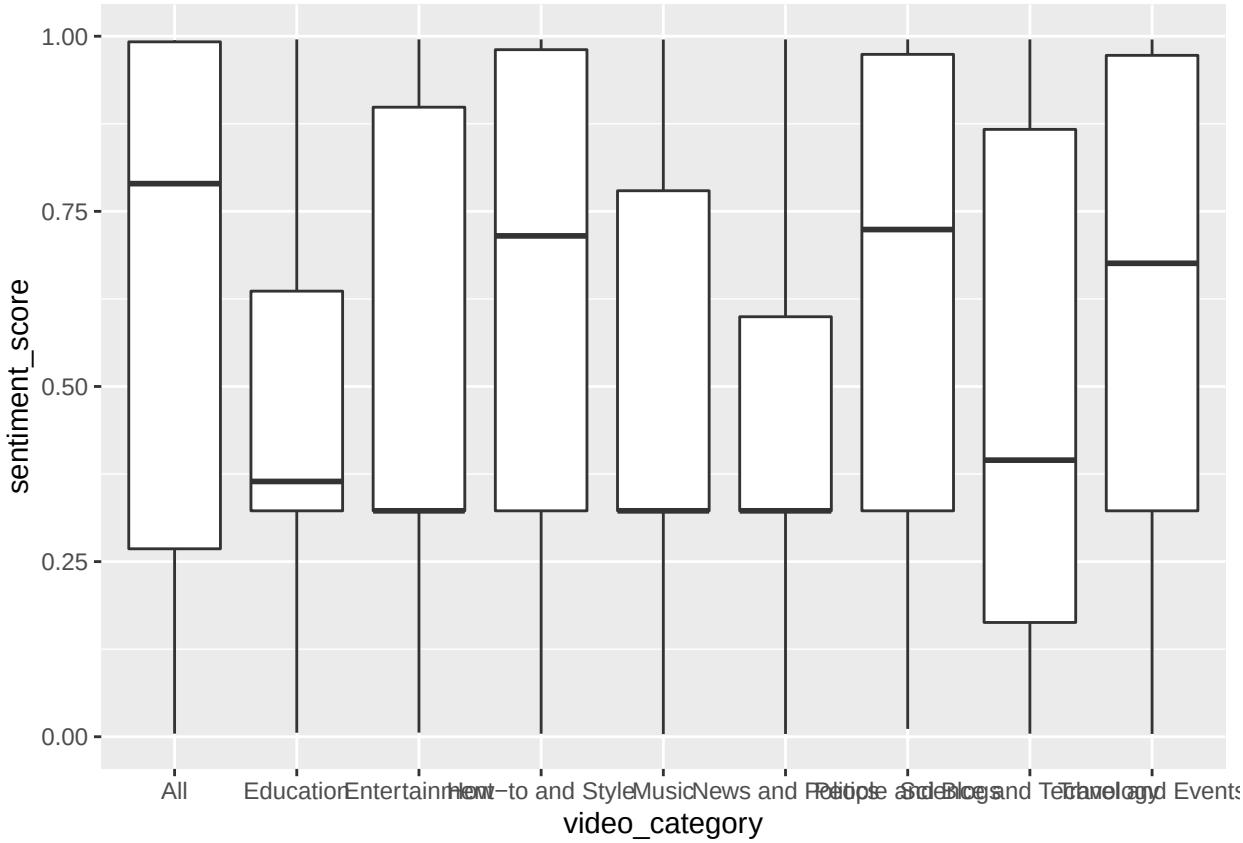

```
ggplot(India_analysis) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




```
India_analysis %>%
  group_by(fortnight) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.580
##  2         2                   0.453
##  3         3                   0.476
##  4         4                   0.505
##  5         5                   0.507
##  6         6                   0.523
##  7         7                   0.623
##  8         8                   0.587
##  9         9                   0.590
## 10        10                   0.548
## 11        11                   0.609
## 12        12                   0.559
## 13        13                   0.598
```

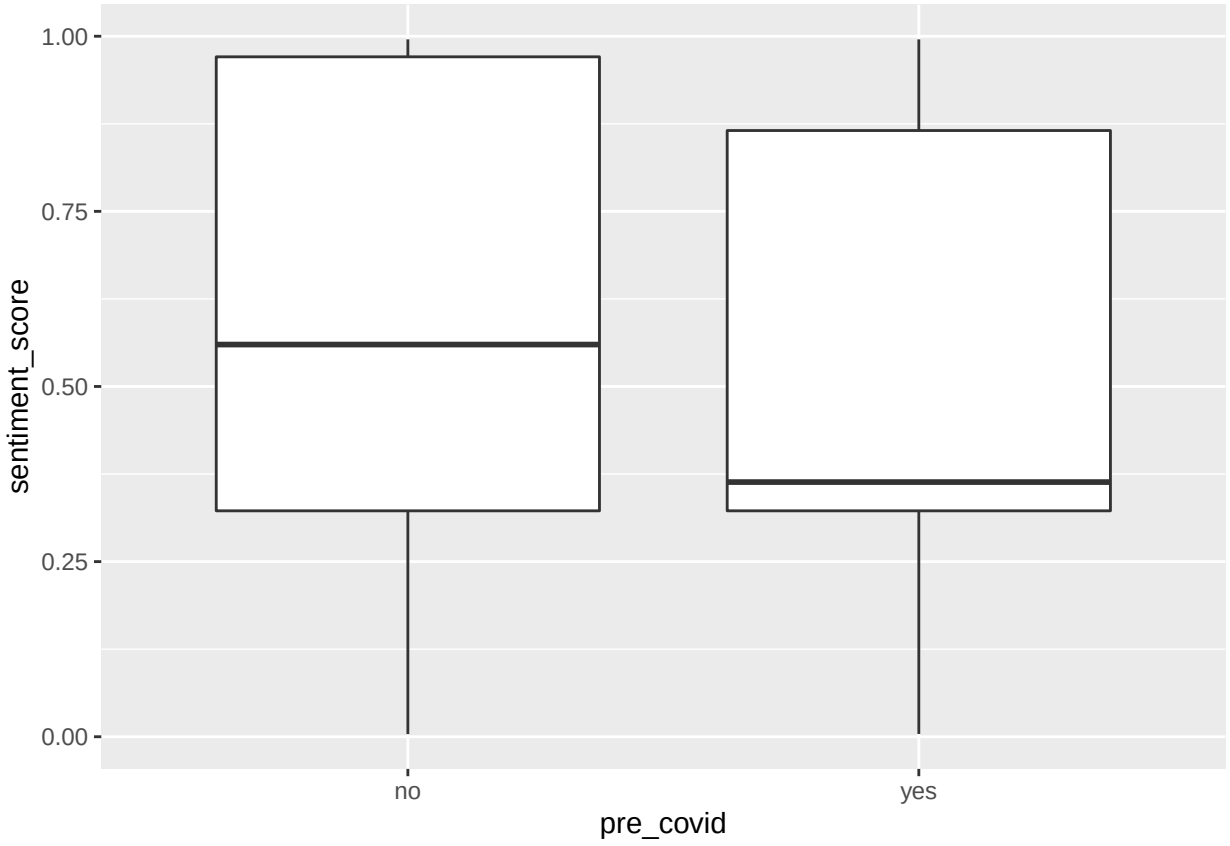
```
ggplot(India_analysis) +
  geom_boxplot(aes(x = video_category, y = sentiment_score))
```




```
India_analysis %>%
  group_by(video_category) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 9 x 2
##   video_category         `mean(sentiment_score)`
##   <chr>                                    <dbl>
## 1 All                                      0.664
## 2 Education                                0.471
## 3 Entertainment                            0.514
## 4 How-to and Style                         0.619
## 5 Music                                    0.477
## 6 News and Politics                        0.450
## 7 People and Blogs                         0.653
## 8 Science and Technology                   0.476
## 9 Travel and Events                        0.629
```

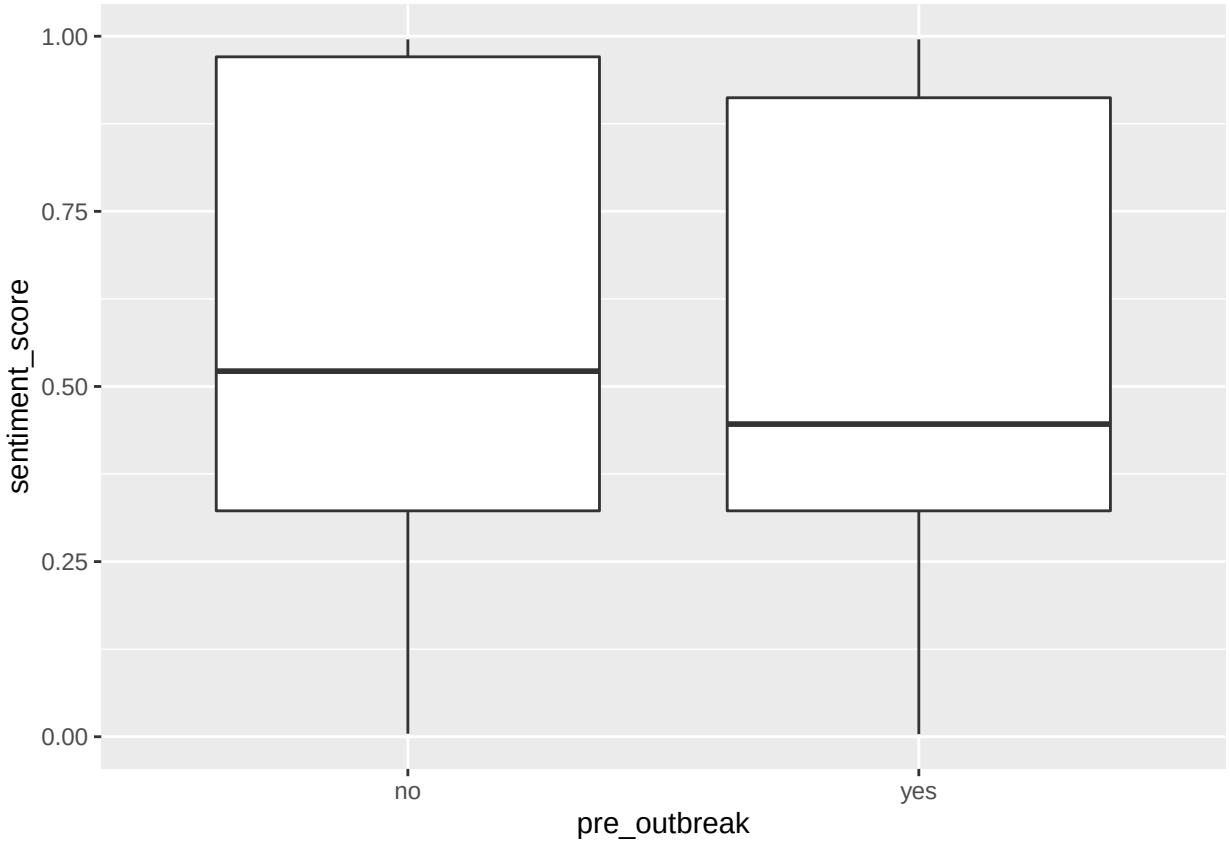
```
ggplot(India_analysis) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




```
India_analysis %>%
  group_by(pre_covid) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.588
## 2 yes                         0.507
```

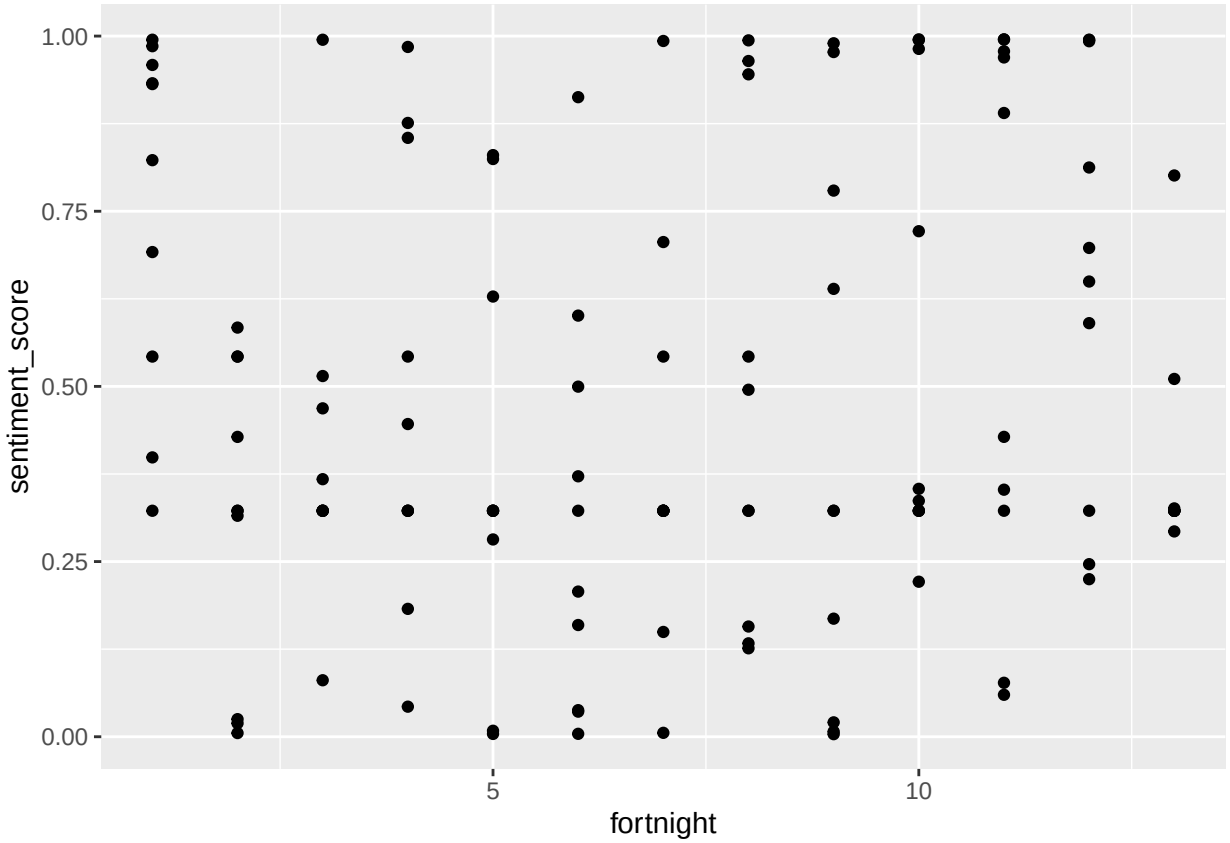
```
ggplot(India_analysis) +
  geom_boxplot(aes(x = pre_outbreak, y = sentiment_score))
```



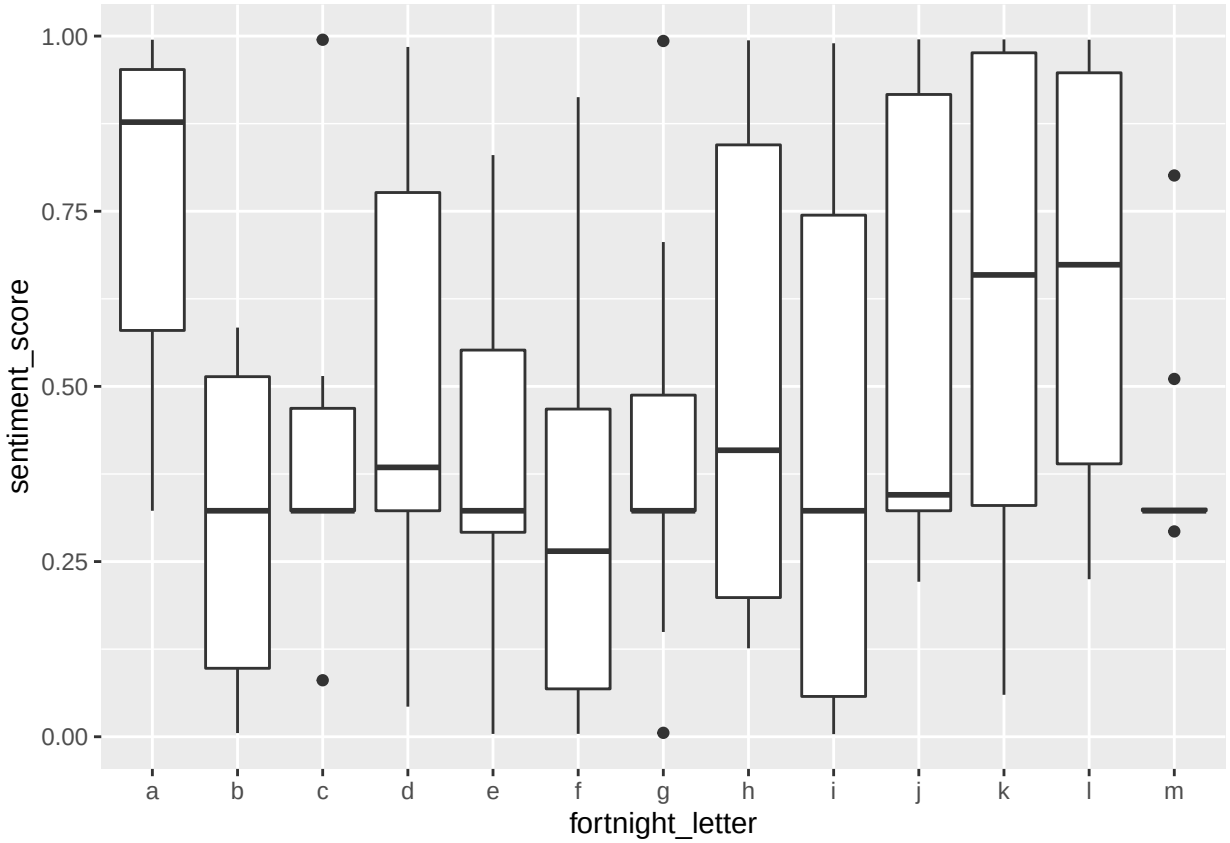

```
India_analysis %>%
  group_by(pre_outbreak) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.589
## 2 yes                            0.539
```

```
#data summary and analysis for music dataset
ggplot(India_analysis_music) +
  geom_point(aes(x = fortnight, y = sentiment_score))
```



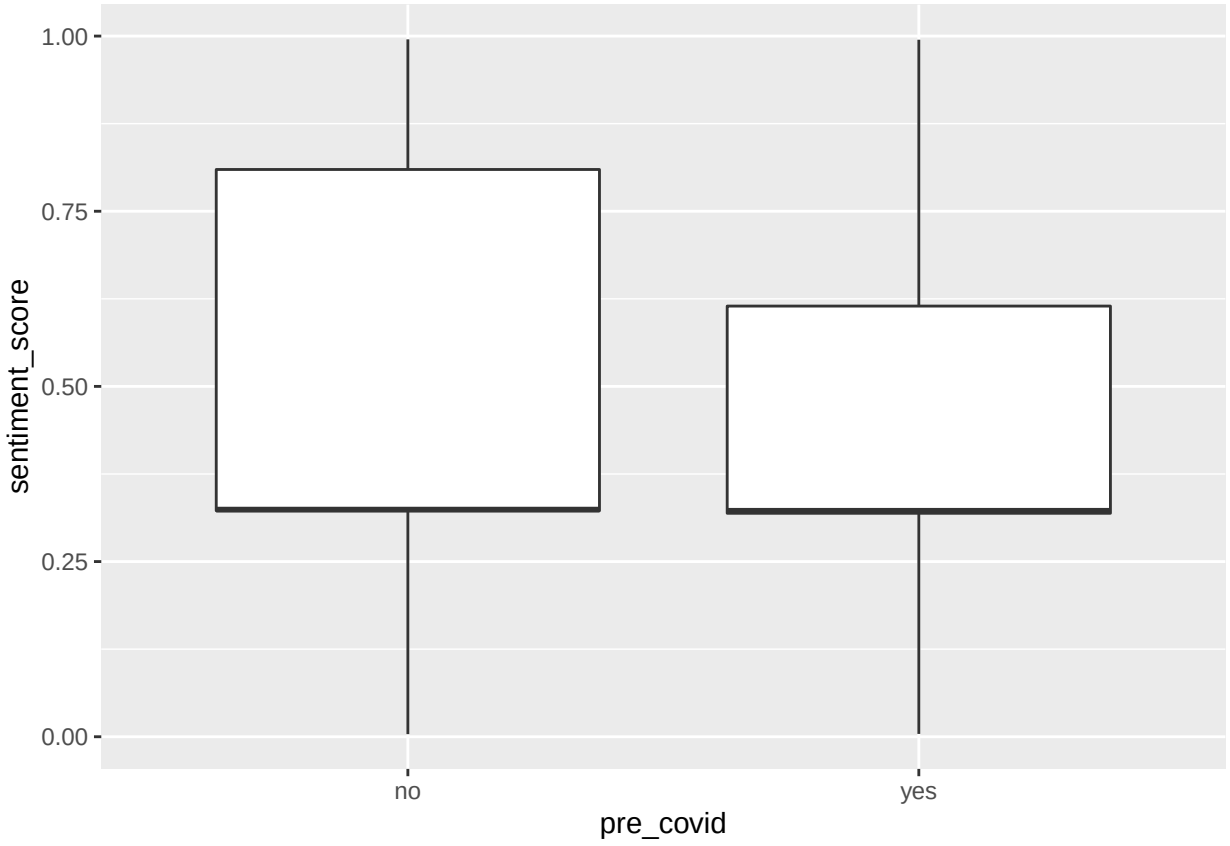

```
ggplot(India_analysis_music) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




```
India_analysis_music %>%
  group_by(fortnight) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.758
##  2         2                   0.311
##  3         3                   0.413
##  4         4                   0.490
##  5         5                   0.387
##  6         6                   0.315
##  7         7                   0.401
##  8         8                   0.500
##  9         9                   0.423
## 10        10                   0.557
## 11        11                   0.607
## 12        12                   0.653
## 13        13                   0.387
```

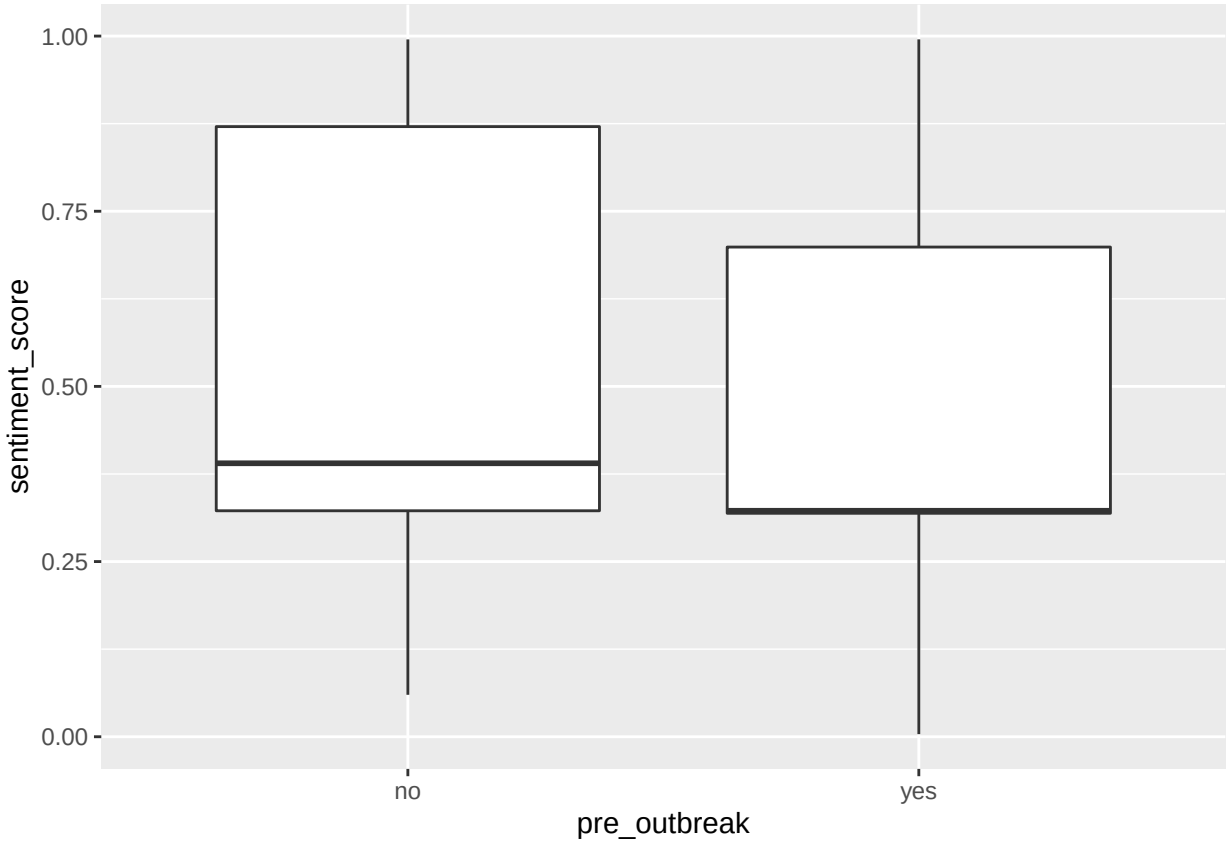
```
ggplot(India_analysis_music) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




```
India_analysis_music %>%
  group_by(pre_covid) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.504
## 2 yes                         0.446
```

```
ggplot(India_analysis_music) +
  geom_boxplot(aes(x = pre_outbreak, y = sentiment_score))
```




```
India_analysis_music %>%
  group_by(pre_outbreak) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.549
## 2 yes                            0.456
```

```
#two proportion z-test for music dataset
#null hypothesis: the true proportion of positive sentiment music videos published precovid and postcov

count(India_analysis_music, pre_covid == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_covid == "yes"`     n
##   <lgl>                <int>
## 1 FALSE                   70
## 2 TRUE                    59
```

```
m_num_precovid = 59
m_num_postcovid = 70
m_num = 129

India_analysis_music %>%
  filter(pre_covid == "yes") %>%
  count(sentiment_score > 0.5)
```

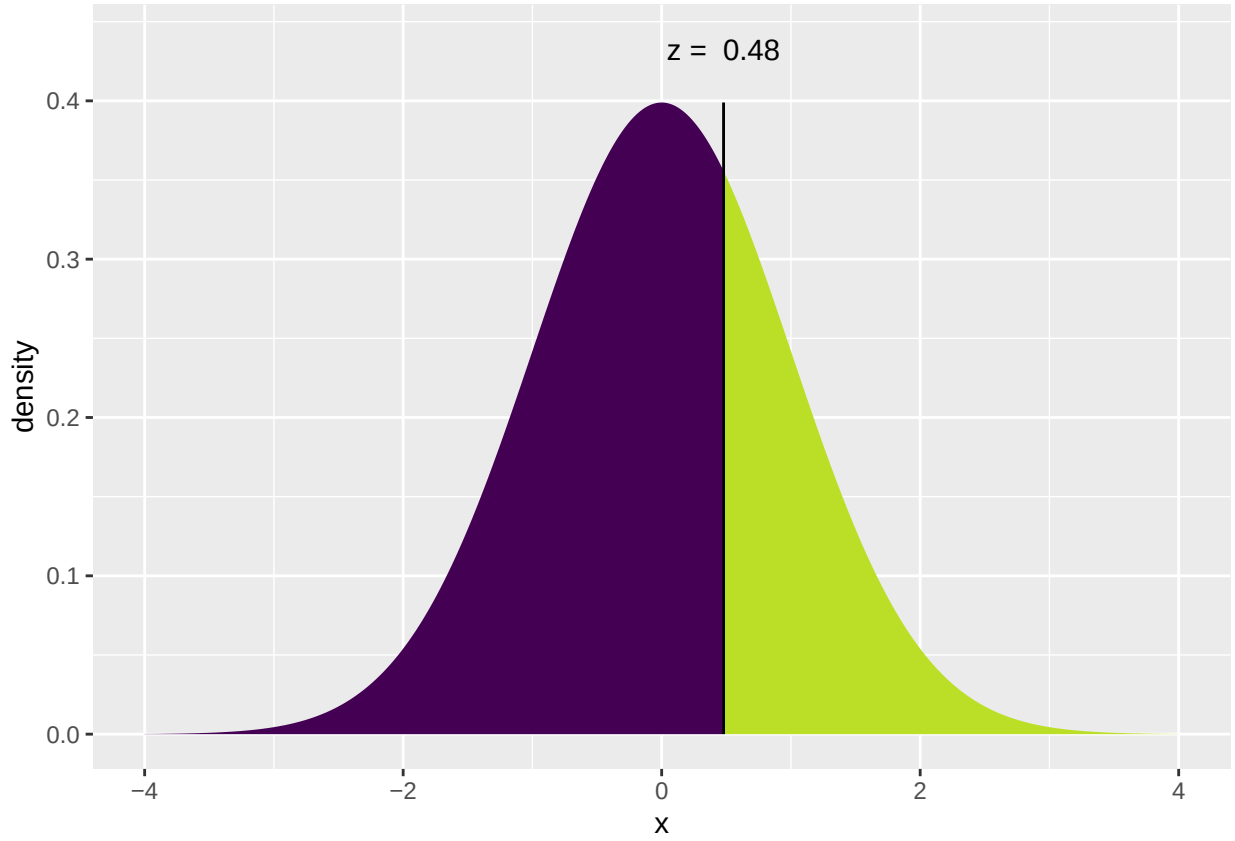
```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      37
## 2 TRUE                       22
```

```
p_hat_1_m_pos = 22/59

India_analysis_music %>%
  filter(pre_covid == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      41
## 2 TRUE                       29
```

```
p_hat_2_m_pos = 29/70

p_hat_m_pos = (22+29)/(59+70)

sd <- sqrt((((p_hat_m_pos)*(1-p_hat_m_pos))/59)+(((p_hat_m_pos)*(1-p_hat_m_pos))/70))
z_score <- ((p_hat_2_m_pos-p_hat_1_m_pos)-0)/sd

#p-value
2* (1-xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= 0.4792) = P(Z <= 0.4792) = 0.6841
##  P(X >  0.4792) = P(Z >  0.4792) = 0.3159
##
```




```
## [1] 0.6318225
```

```
#outbreak music
count(India_analysis_music, pre_outbreak == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_outbreak == "yes"`     n
##   <lgl>                   <int>
## 1 FALSE                      30
## 2 TRUE                       99
```

```
m_num_preoutbreak = 99
m_num_postoutbreak = 30
m_num = 129

India_analysis_music %>%
  filter(pre_outbreak == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      62
## 2 TRUE                       37
```

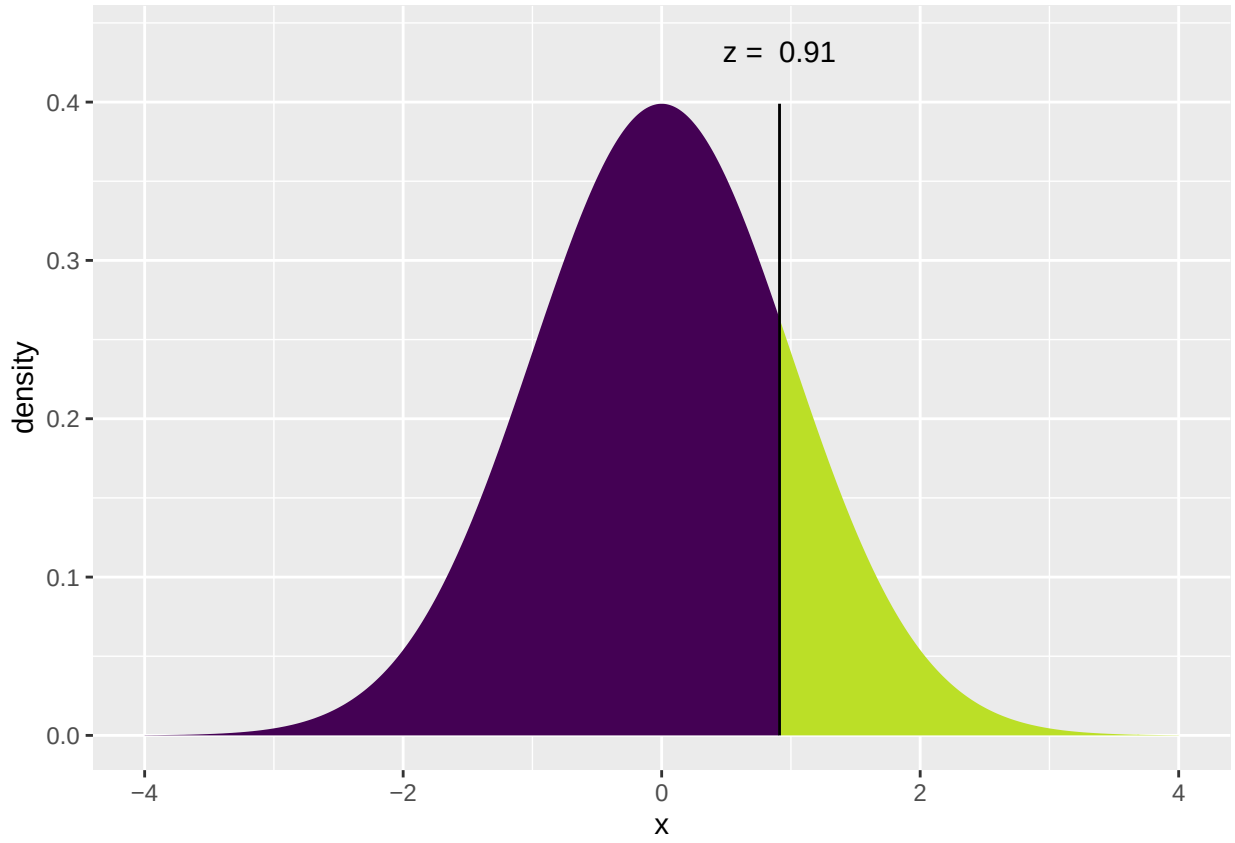
```
p_hat_1_m_pos = 37/99

India_analysis_music %>%
```

```
  filter(pre_outbreak == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      16
## 2 TRUE                       14
```

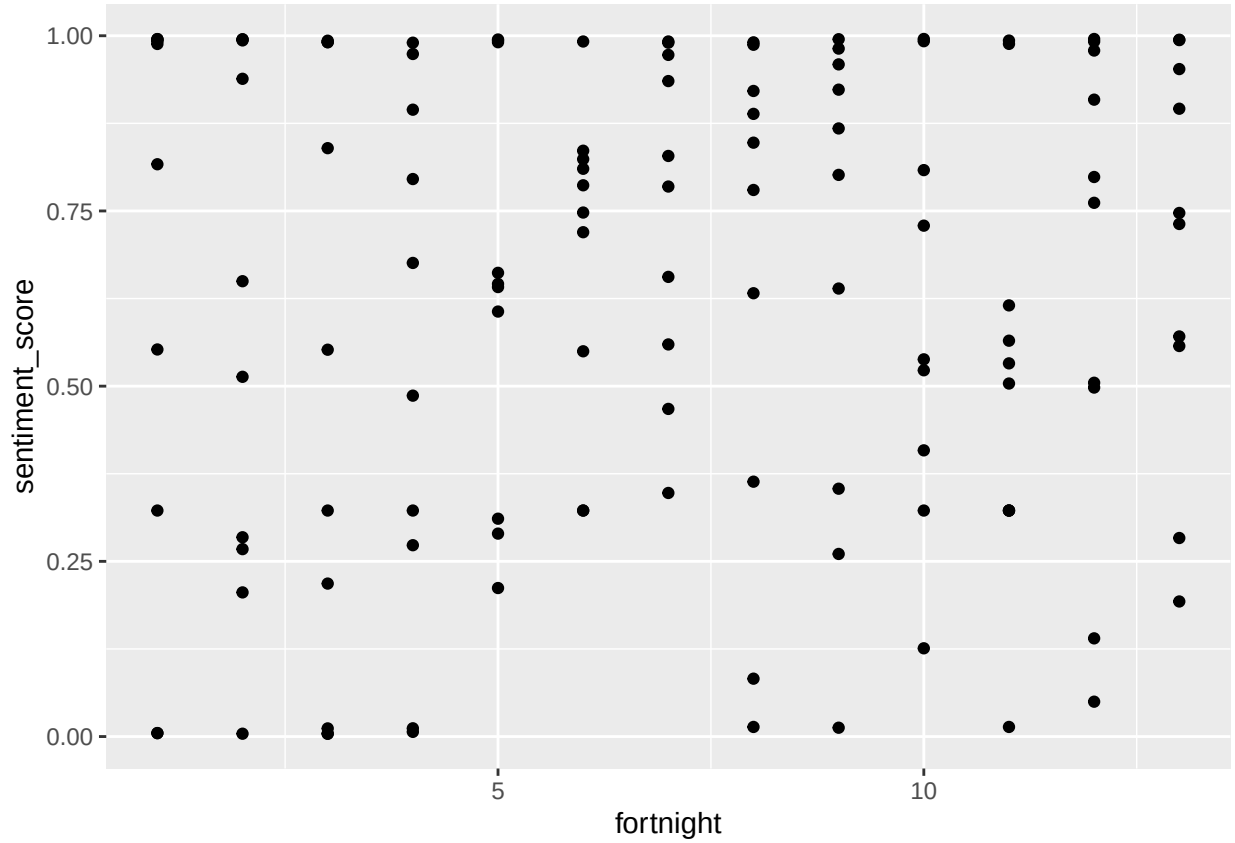
```
p_hat_2_m_pos = 14/30

p_hat_m_pos = (37+14)/(99+30)

sd <- sqrt((((p_hat_m_pos)*(1-p_hat_m_pos))/99)+(((p_hat_m_pos)*(1-p_hat_m_pos))/30))
z_score <- ((p_hat_2_m_pos-p_hat_1_m_pos)-0)/sd

#p-value
2* (1-xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= 0.912) = P(Z <= 0.912) = 0.8191
##  P(X >  0.912) = P(Z >  0.912) = 0.1809
##
```

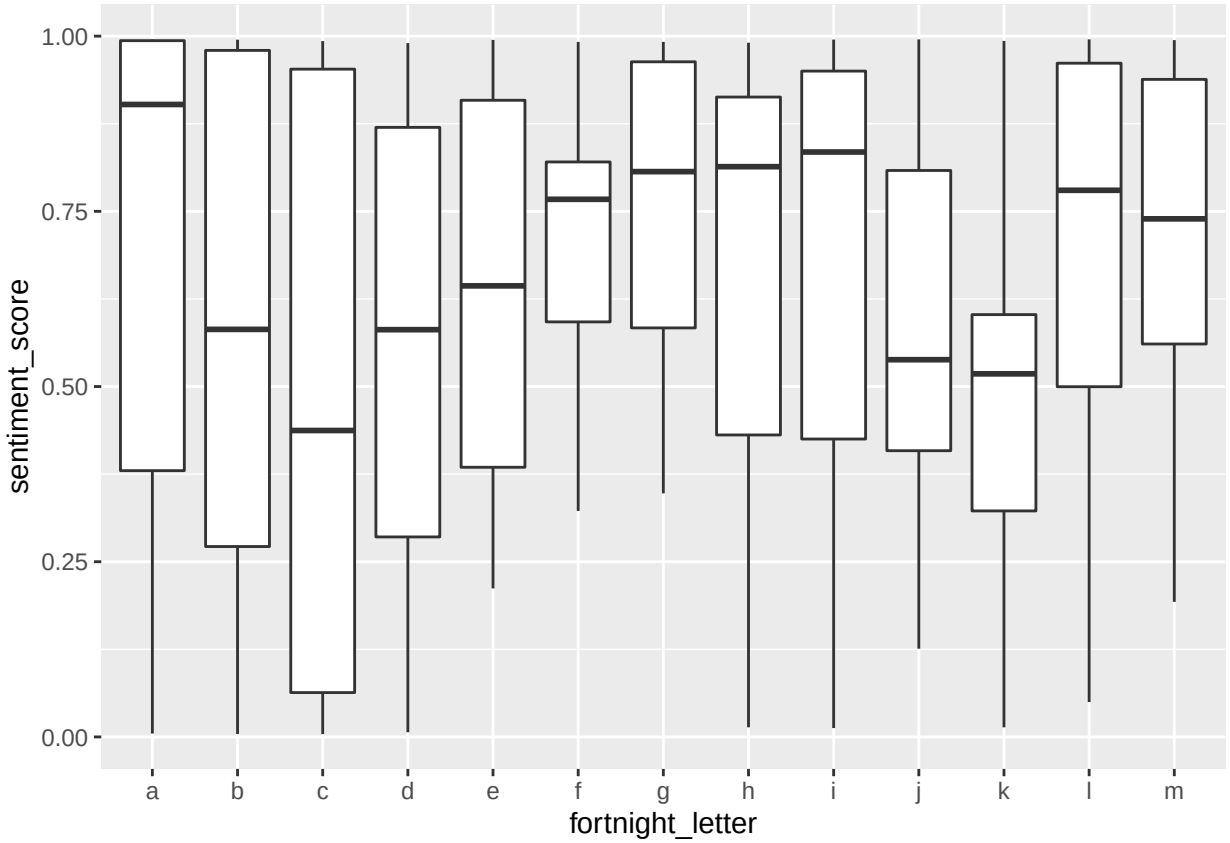


```
## [1] 0.3617704
```

```
#data summary travel
ggplot(India_analysis_travel) +
  geom_point(aes(x = fortnight, y = sentiment_score))
```



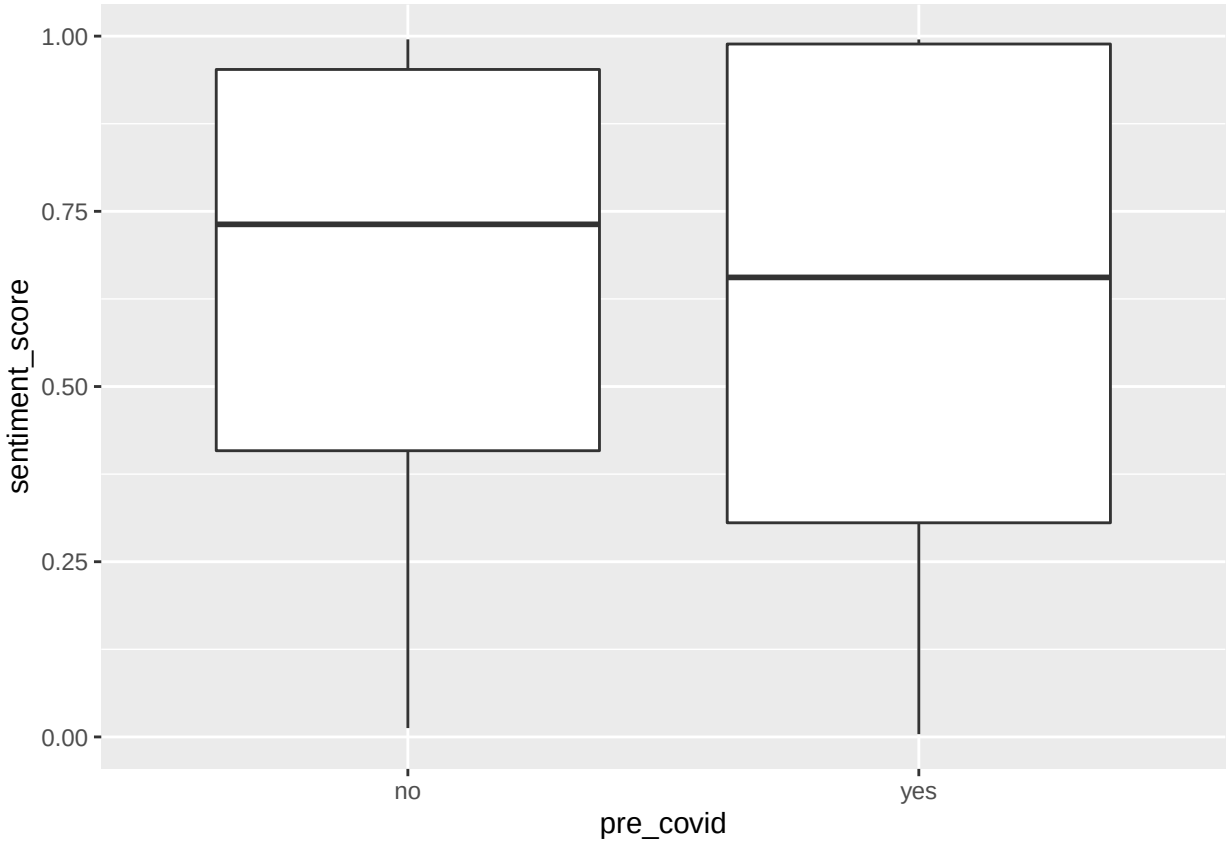

```
ggplot(India_analysis_travel) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




```
India_analysis_travel %>%
  group_by(fortnight) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.667
##  2         2                   0.585
##  3         3                   0.493
##  4         4                   0.543
##  5         5                   0.635
##  6         6                   0.691
##  7         7                   0.753
##  8         8                   0.651
##  9         9                   0.679
## 10        10                   0.605
## 11        11                   0.518
## 12        12                   0.663
## 13        13                   0.692
```

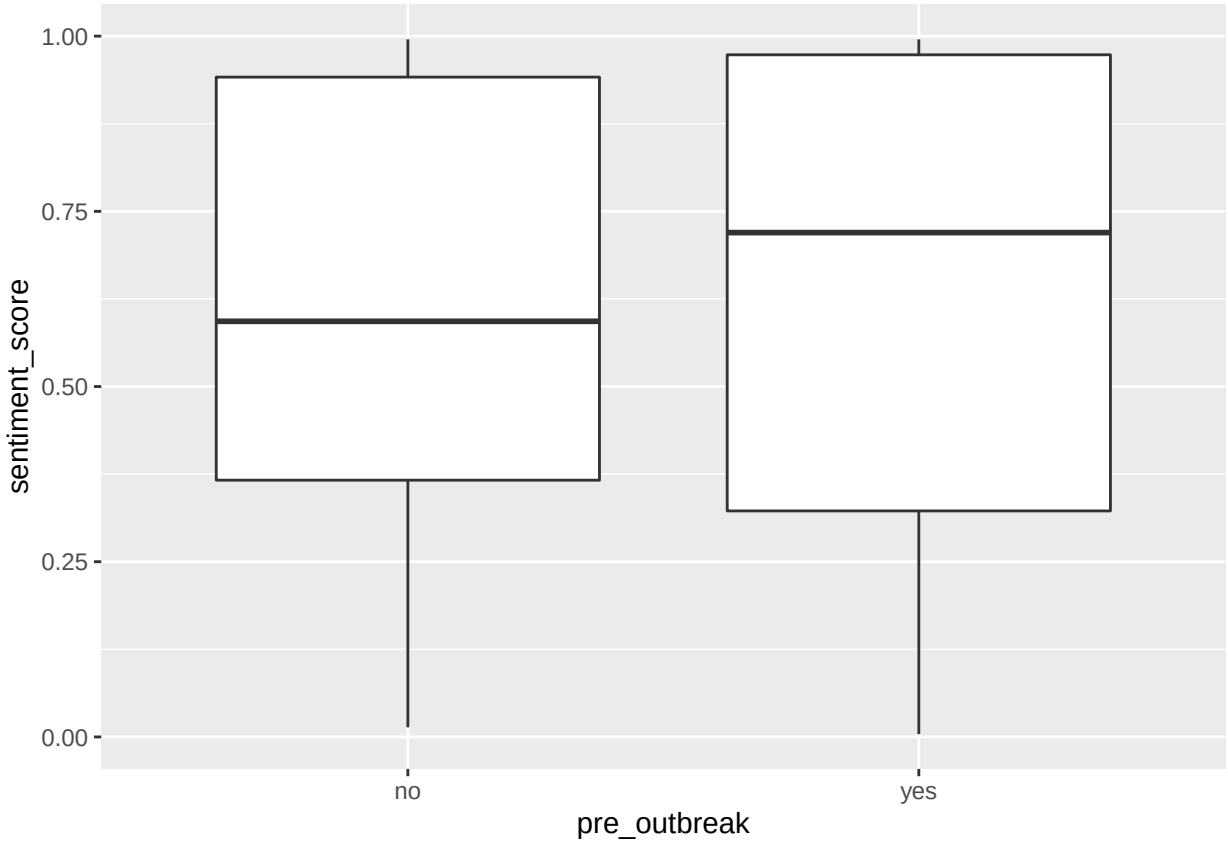
```
ggplot(India_analysis_travel) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




```
India_analysis_travel %>%
  group_by(pre_covid) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.652
## 2 yes                         0.602
```

```
ggplot(India_analysis_travel) +
  geom_boxplot(aes(x = pre_outbreak, y = sentiment_score))
```




```
India_analysis_travel %>%
  group_by(pre_outbreak) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.624
## 2 yes                            0.630
```

```
#precovid travel
#null hypothesis: the true proportion of positive sentiment travel videos published precovid and postco

count(India_analysis_travel, pre_covid == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_covid == "yes"`     n
##   <lgl>                <int>
## 1 FALSE                   69
## 2 TRUE                    60
```

```
t_num_precovid = 60
t_num_postcovid = 69
t_num = 129

India_analysis_travel %>%
  filter(pre_covid == "yes") %>%
  count(sentiment_score > 0.5)
```

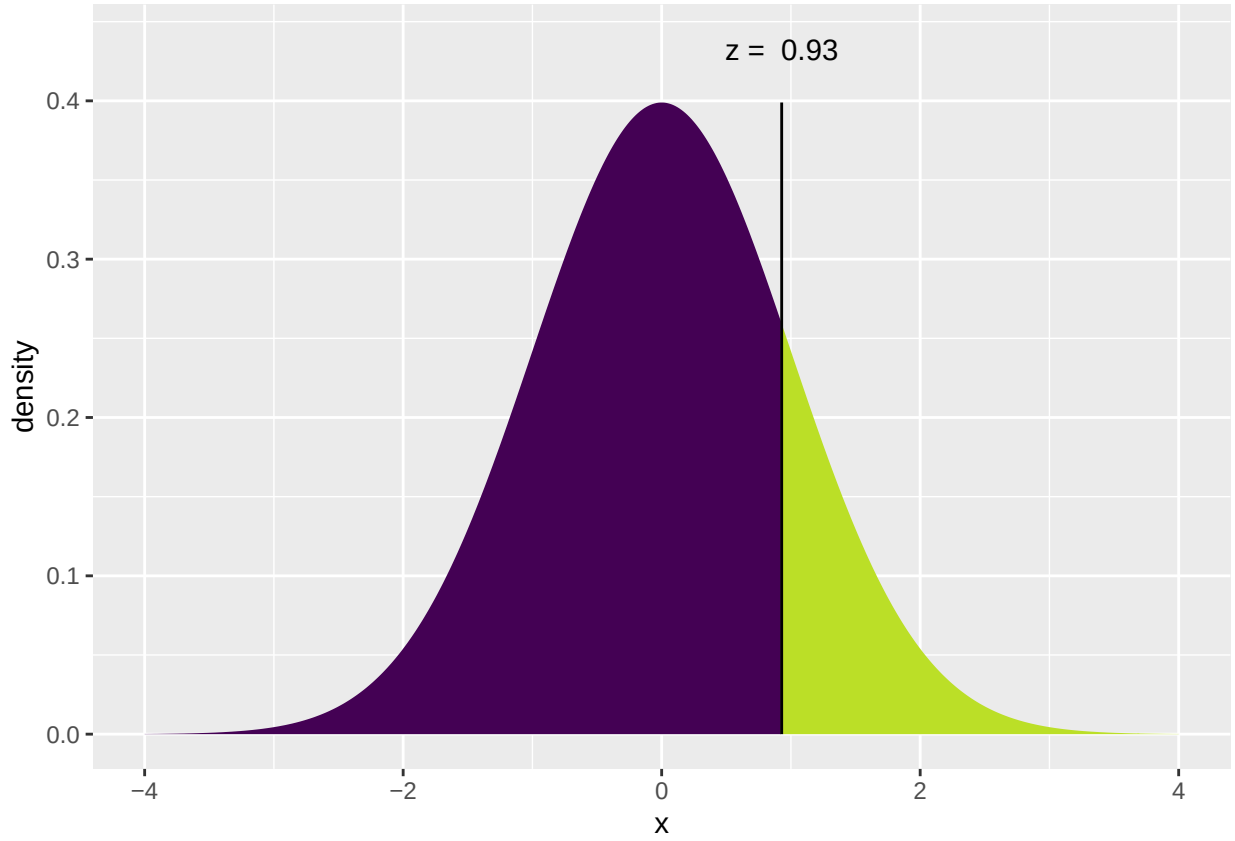
```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      22
## 2 TRUE                       38
```

```
p_hat_1_t_pos = 38/60

India_analysis_travel %>%
  filter(pre_covid == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      20
## 2 TRUE                       49
```

```
p_hat_2_t_pos = 49/69

p_hat_t_pos = (38+49)/(60+69)

sd <- sqrt((((p_hat_t_pos)*(1-p_hat_t_pos))/60)+(((p_hat_t_pos)*(1-p_hat_t_pos))/69))
z_score <- ((p_hat_2_t_pos-p_hat_1_t_pos)-0)/sd

#p-value
2* (1-xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= 0.9286) = P(Z <= 0.9286) = 0.8235
##  P(X >  0.9286) = P(Z >  0.9286) = 0.1765
##
```




```
## [1] 0.3530866
```

```
#outbreak travel
count(India_analysis_travel, pre_outbreak == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_outbreak == "yes"`     n
##   <lgl>                   <int>
## 1 FALSE                      30
## 2 TRUE                       99
```

```
t_num_preoutbreak = 99
t_num_postoutbreak = 30
t_num = 129

India_analysis_travel %>%
  filter(pre_outbreak == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      33
## 2 TRUE                       66
```

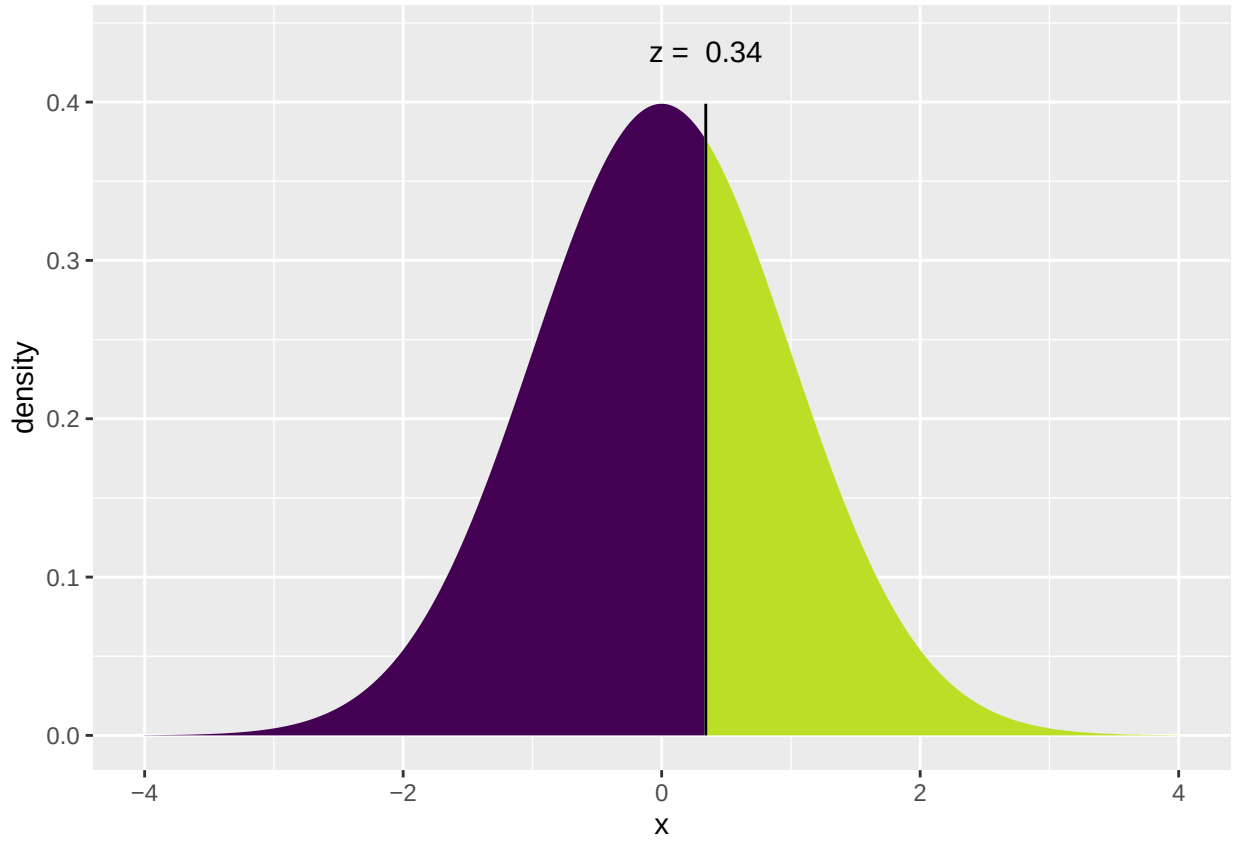
```
p_hat_1_t_pos = 66/99

India_analysis_travel %>%
```

```
  filter(pre_outbreak == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                       9
## 2 TRUE                       21
```

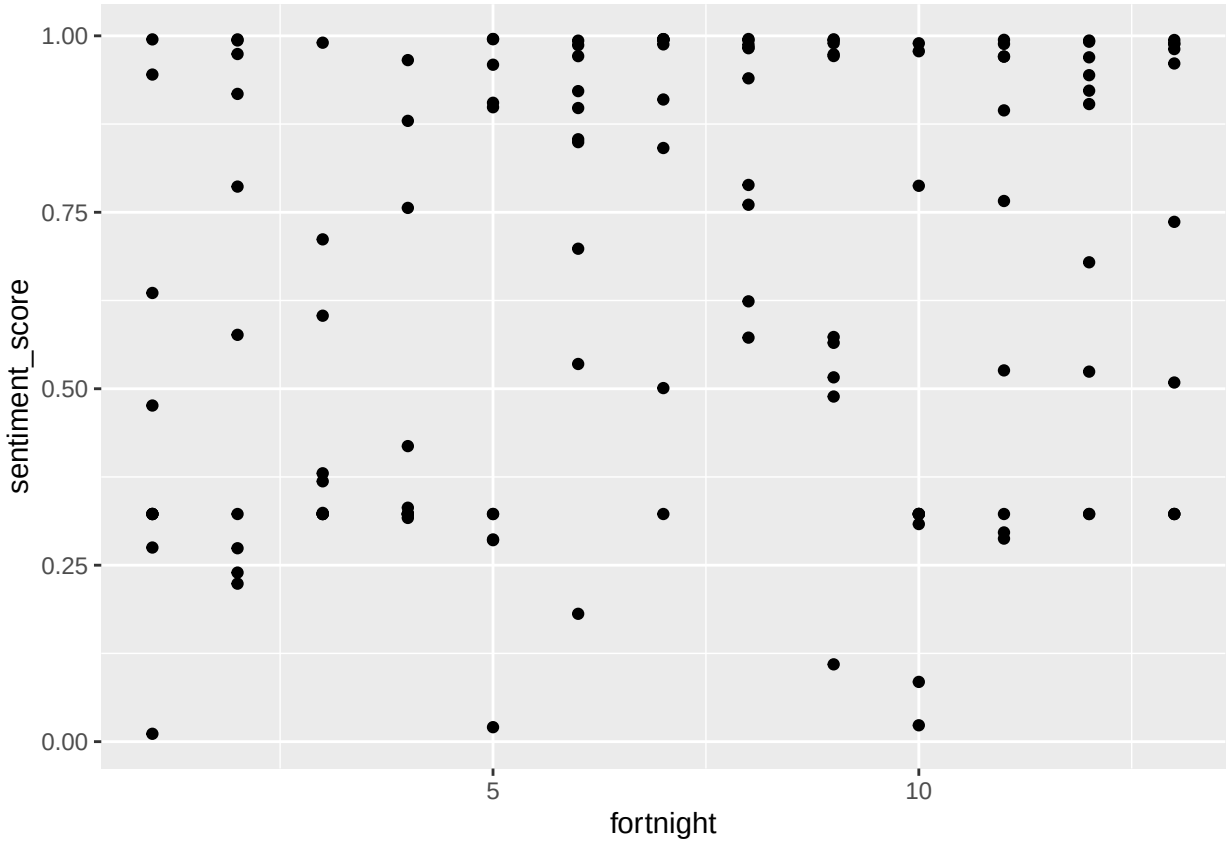
```
p_hat_2_t_pos = 21/30

p_hat_t_pos = (66+21)/(99+30)

sd <- sqrt((((p_hat_t_pos)*(1-p_hat_t_pos))/99)+(((p_hat_t_pos)*(1-p_hat_t_pos))/30))
z_score <- ((p_hat_2_t_pos-p_hat_1_t_pos)-0)/sd

#p-value
2* (1-xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= 0.3413) = P(Z <= 0.3413) = 0.6336
##  P(X >  0.3413) = P(Z >  0.3413) = 0.3664
##
```

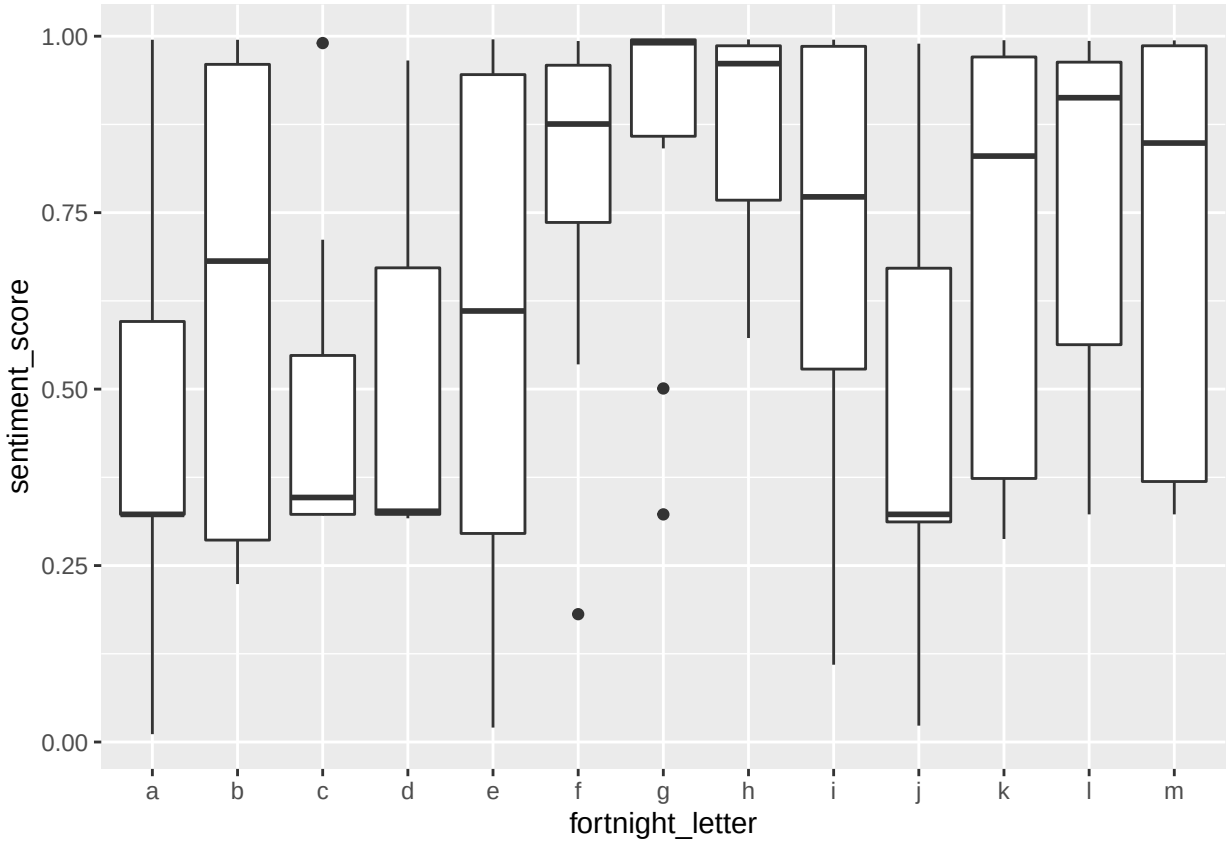


```
## [1] 0.7328593
```

```
#data summary people and blogs
ggplot(India_analysis_people) +
  geom_point(aes(x = fortnight, y = sentiment_score))
```



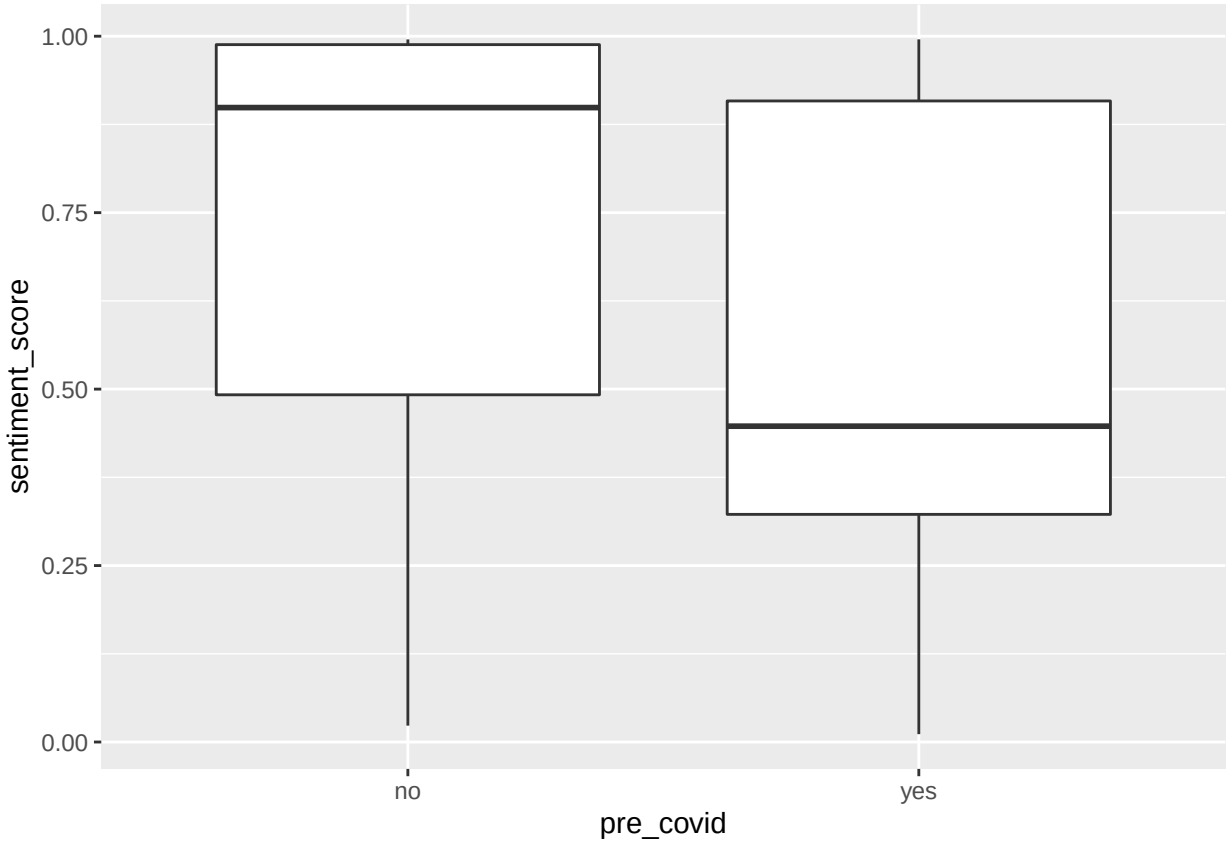

```
ggplot(India_analysis_people) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




```
India_analysis_people %>%
  group_by(fortnight) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.463
##  2         2                   0.630
##  3         3                   0.467
##  4         4                   0.496
##  5         5                   0.599
##  6         6                   0.789
##  7         7                   0.854
##  8         8                   0.863
##  9         9                   0.718
## 10        10                   0.446
## 11        11                   0.702
## 12        12                   0.757
## 13        13                   0.713
```

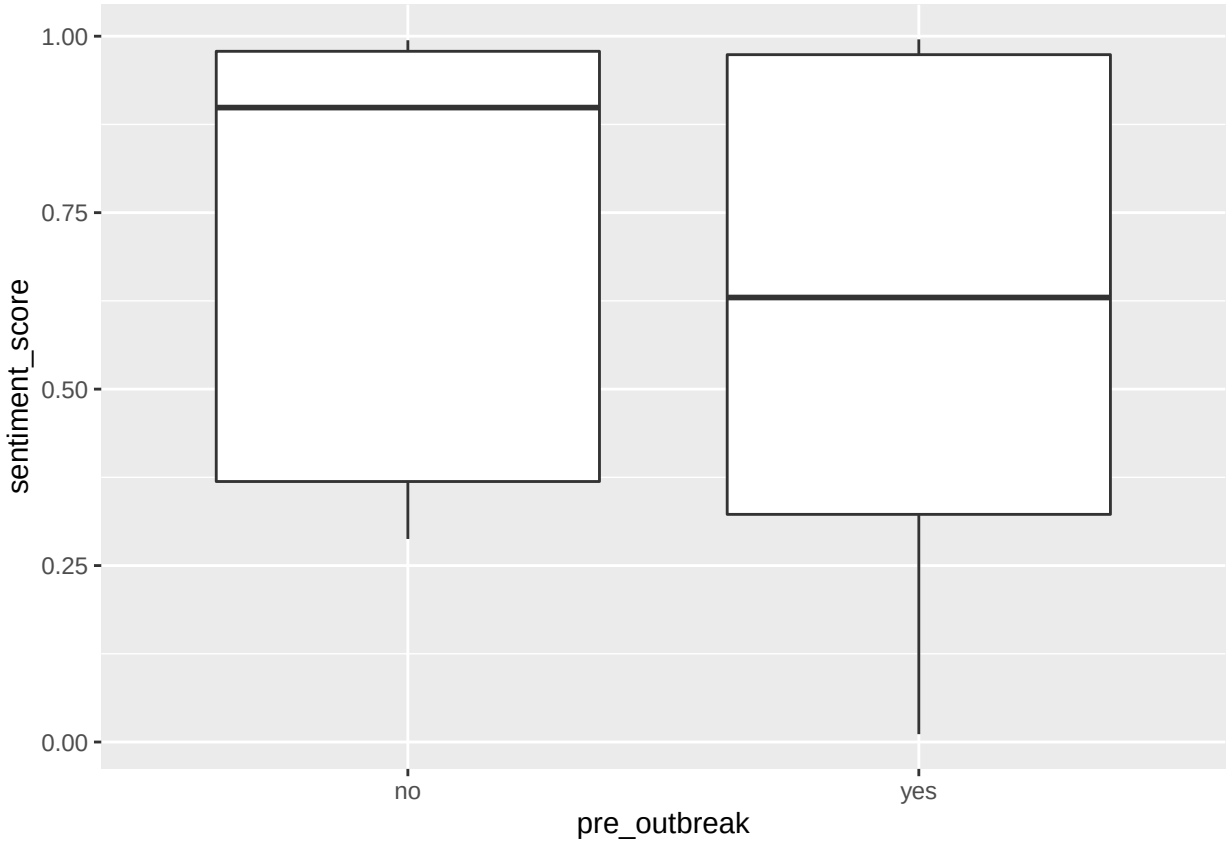
```
ggplot(India_analysis_people) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




```
India_analysis_people %>%
  group_by(pre_covid) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.722
## 2 yes                         0.574
```

```
ggplot(India_analysis_people) +
  geom_boxplot(aes(x = pre_outbreak, y = sentiment_score))
```




```
India_analysis_people %>%
  group_by(pre_outbreak) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.724
## 2 yes                            0.632
```

```
#precovid people
count(India_analysis_people, pre_covid == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_covid == "yes"`     n
##   <lgl>                <int>
## 1 FALSE                   70
## 2 TRUE                    60
```

```
p_num_precovid = 60
p_num_postcovid = 70
p_num = 130

India_analysis_people %>%
  filter(pre_covid == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
```

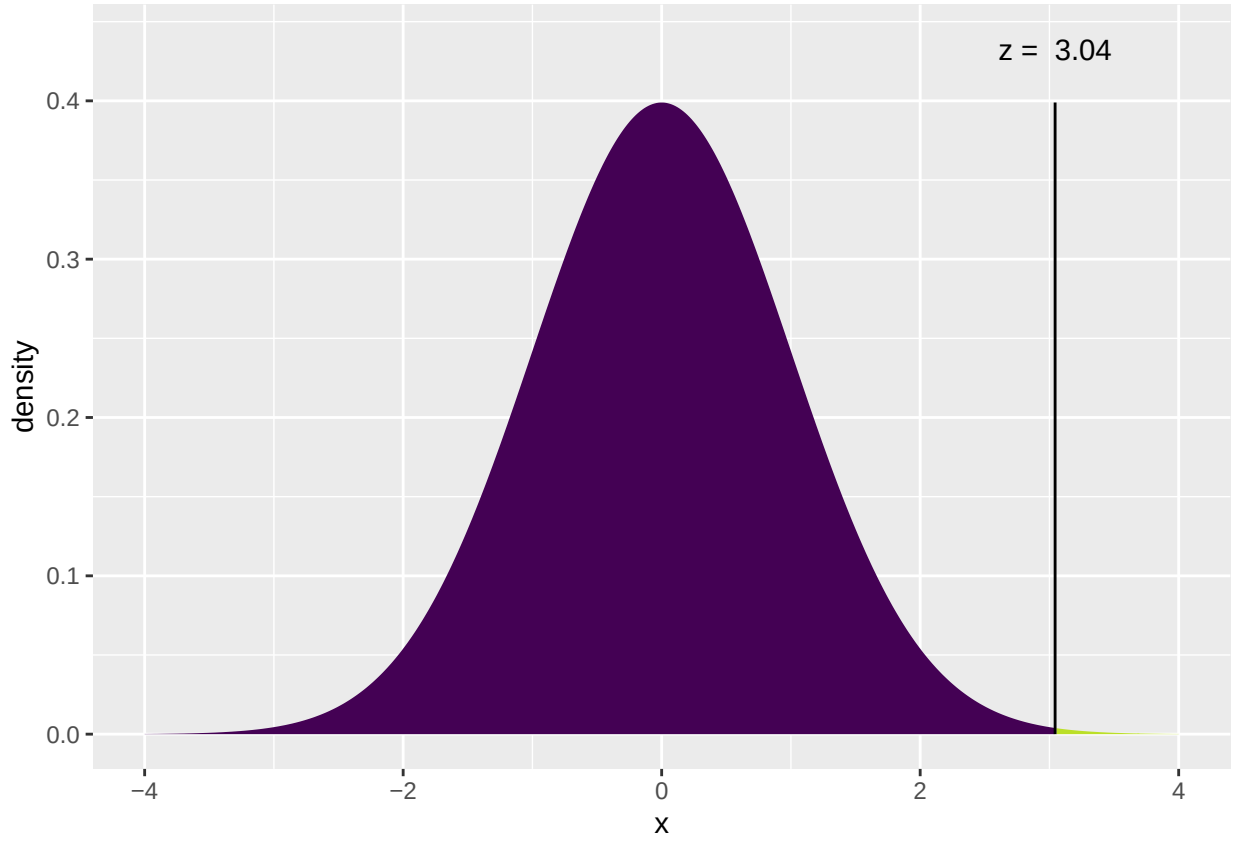
```
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      31
## 2 TRUE                       29
```

```
p_hat_1_p_pos = 29/60

India_analysis_people %>%
  filter(pre_covid == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      18
## 2 TRUE                       52
```

```
p_hat_2_p_pos = 52/70

p_hat_p_pos = (29+52)/(60+70)

sd <- sqrt((((p_hat_p_pos)*(1-p_hat_p_pos))/60)+(((p_hat_p_pos)*(1-p_hat_p_pos))/70))
z_score <- ((p_hat_2_p_pos-p_hat_1_p_pos)-0)/sd

#p-value
2* (1-xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= 3.044) = P(Z <= 3.044) = 0.9988
##  P(X >  3.044) = P(Z >  3.044) = 0.001168
##
```




```
## [1] 0.002335186
```

```
#outbreak people
count(India_analysis_people, pre_outbreak == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_outbreak == "yes"`     n
##   <lgl>                   <int>
## 1 FALSE                      30
## 2 TRUE                      100
```

```
p_num_preoutbreak = 100
p_num_postoutbreak = 30
p_num = 130

India_analysis_people %>%
  filter(pre_outbreak == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      41
## 2 TRUE                       59
```

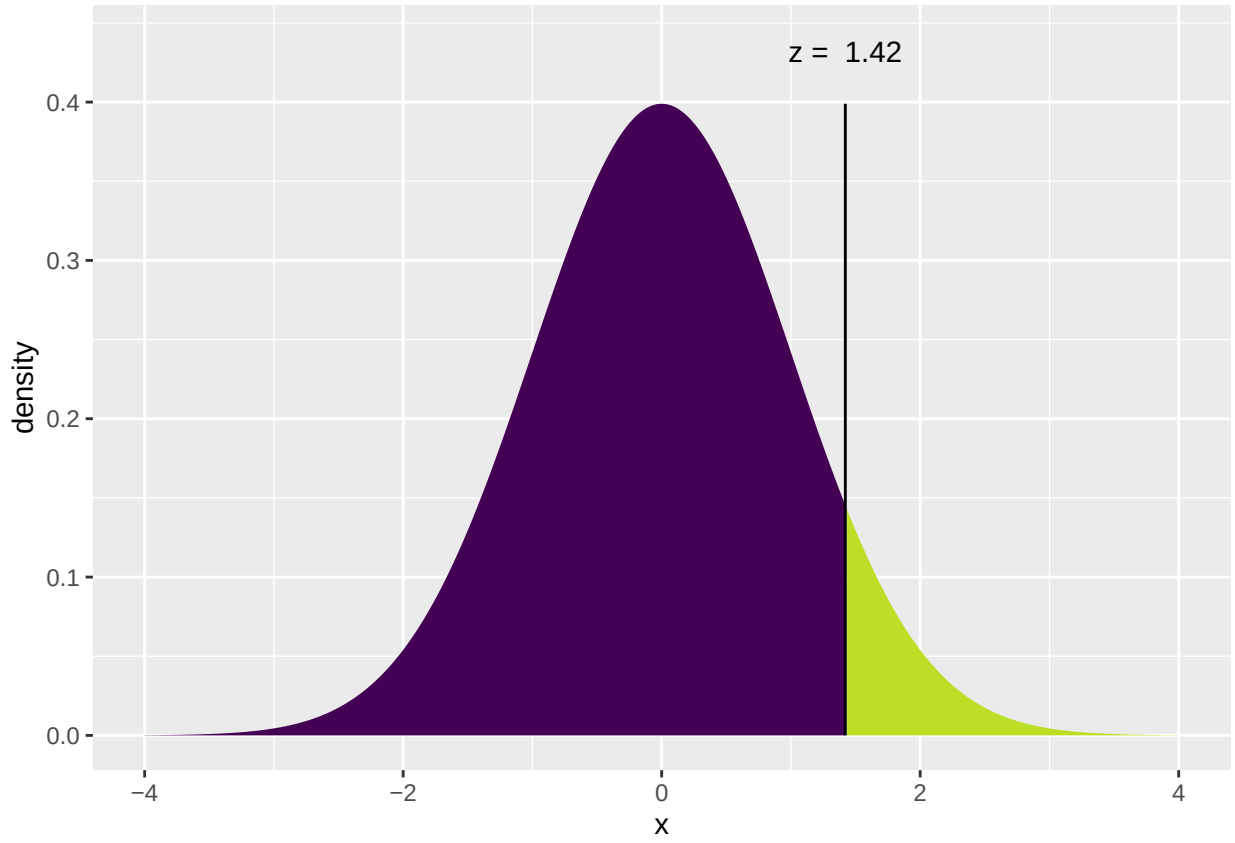
```
p_hat_1_p_pos = 59/100

India_analysis_people %>%
```

```
  filter(pre_outbreak == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                       8
## 2 TRUE                       22
```

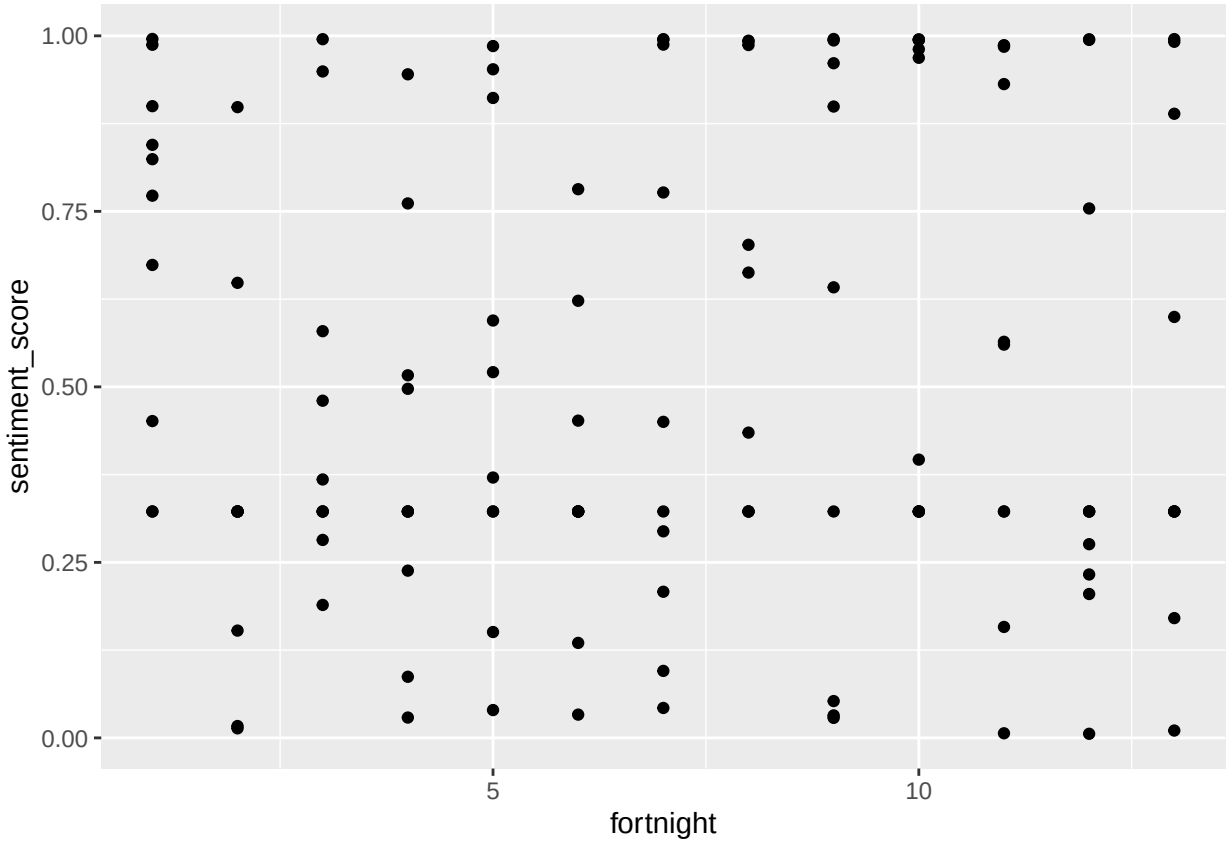
```
p_hat_2_p_pos = 22/30

p_hat_p_pos = (59+22)/(100+30)

sd <- sqrt((((p_hat_p_pos)*(1-p_hat_p_pos))/100)+(((p_hat_p_pos)*(1-p_hat_p_pos))/30))
z_score <- ((p_hat_2_p_pos-p_hat_1_p_pos)-0)/sd

#p-value
2* (1-xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= 1.421) = P(Z <= 1.421) = 0.9223
##  P(X >  1.421) = P(Z >  1.421) = 0.07768
##
```

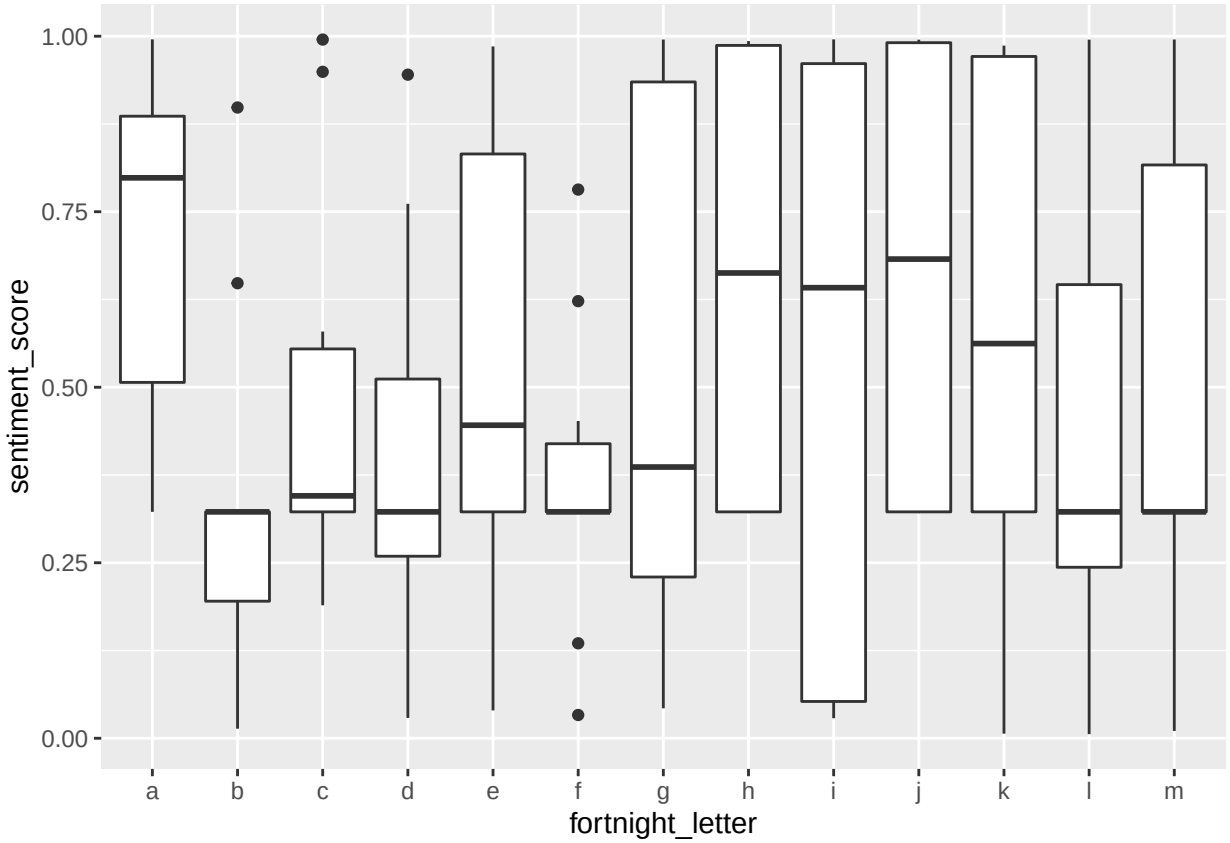


```
## [1] 0.1553692
```

```
#data summary entertainment
ggplot(India_analysis_entertainment) +
  geom_point(aes(x = fortnight, y = sentiment_score))
```



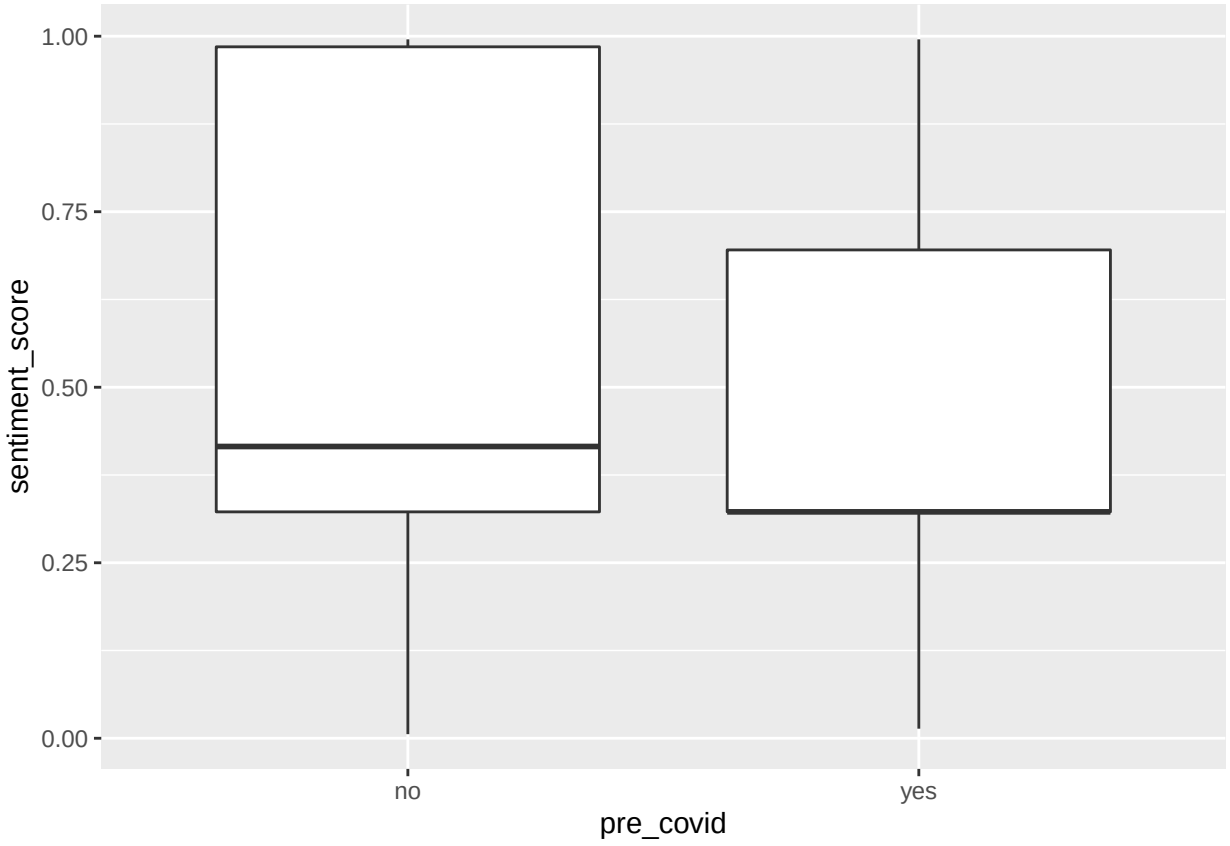

```
ggplot(India_analysis_entertainment) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




```
India_analysis_entertainment %>%
  group_by(fortnight) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.709
##  2         2                   0.334
##  3         3                   0.481
##  4         4                   0.404
##  5         5                   0.517
##  6         6                   0.364
##  7         7                   0.517
##  8         8                   0.638
##  9         9                   0.547
## 10        10                   0.662
## 11        11                   0.582
## 12        12                   0.443
## 13        13                   0.495
```

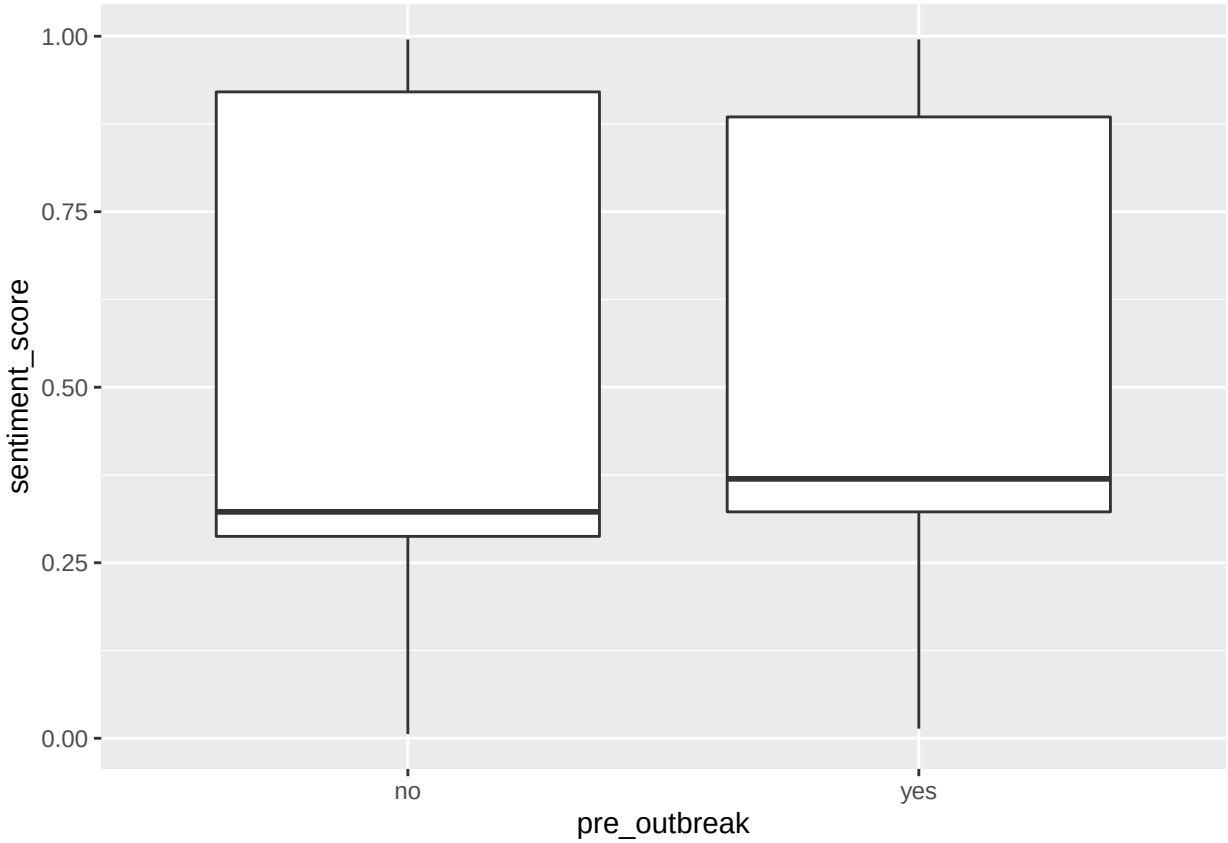
```
ggplot(India_analysis_entertainment) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




```
India_analysis_entertainment %>%
  group_by(pre_covid) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.554
## 2 yes                         0.468
```

```
ggplot(India_analysis_entertainment) +
  geom_boxplot(aes(x = pre_outbreak, y = sentiment_score))
```




```
India_analysis_entertainment %>%
  group_by(pre_outbreak) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.507
## 2 yes                            0.516
```

```
#precovid entertainment
count(India_analysis_entertainment, pre_covid == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_covid == "yes"`     n
##   <lgl>                <int>
## 1 FALSE                   68
## 2 TRUE                    60
```

```
num_precovid = 60
num_postcovid = 68
num = 130

India_analysis_entertainment %>%
  filter(pre_covid == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
```

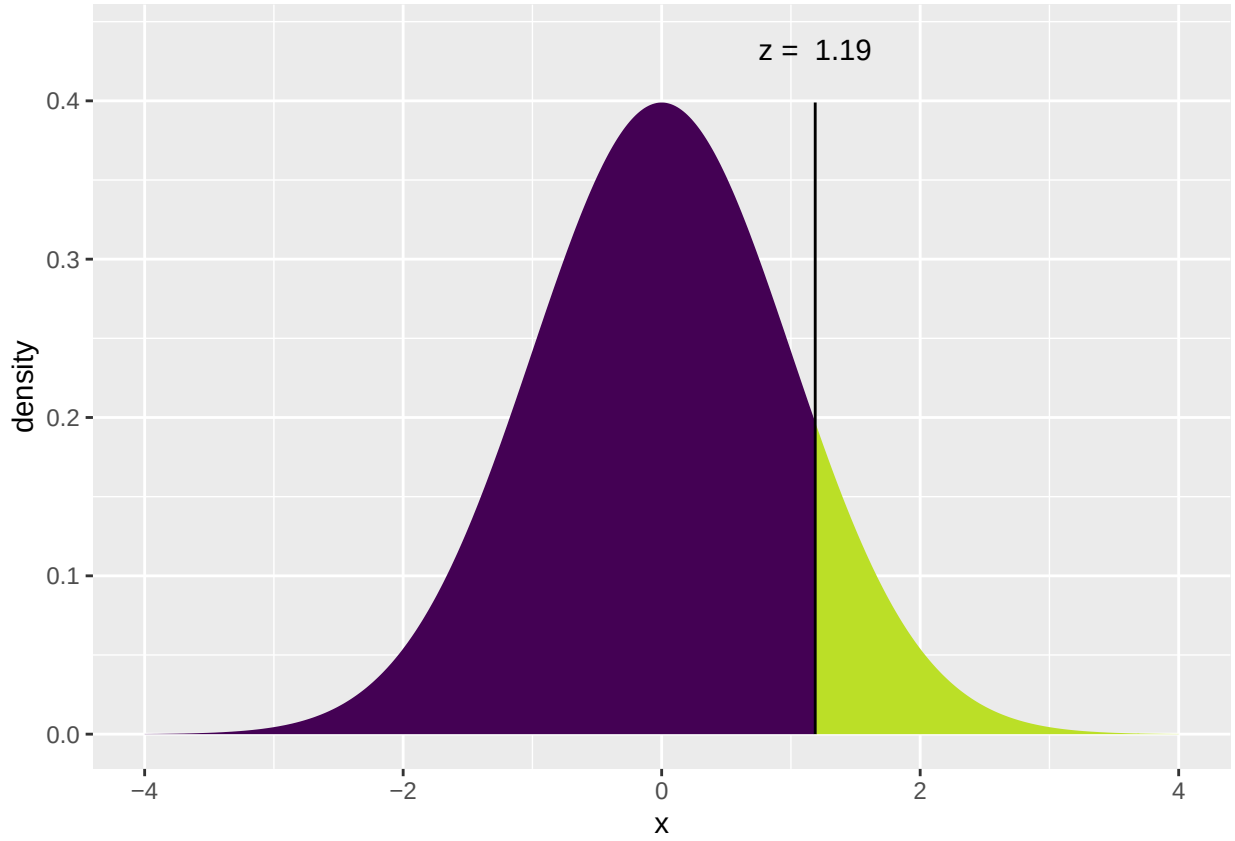
```
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      38
## 2 TRUE                       22
```

```
#proportion of positive sentiment videos precovid from sample
p_hat1 = 22/60

India_analysis_entertainment %>%
  filter(pre_covid == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      36
## 2 TRUE                       32
```

```
#proportion of positive sentiment videos postcovid from sample
p_hat2 = 32/68

p_hat = (22+32)/(60+68)

sd <- sqrt((((p_hat)*(1-p_hat))/60)+(((p_hat)*(1-p_hat))/68))
z_score <- ((p_hat2-p_hat1)-0)/sd

#p-value
2* (1-xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= 1.188) = P(Z <= 1.188) = 0.8826
##  P(X >  1.188) = P(Z >  1.188) = 0.1174
##
```




```
## [1] 0.2348209
```

```
#outbreak entertainment
count(India_analysis_entertainment, pre_outbreak == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_outbreak == "yes"`     n
##   <lgl>                   <int>
## 1 FALSE                      30
## 2 TRUE                       98
```

```
num_preoutbreak = 98
num_postoutbreak = 30
num = 128

India_analysis_entertainment %>%
  filter(pre_outbreak == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      57
## 2 TRUE                       41
```

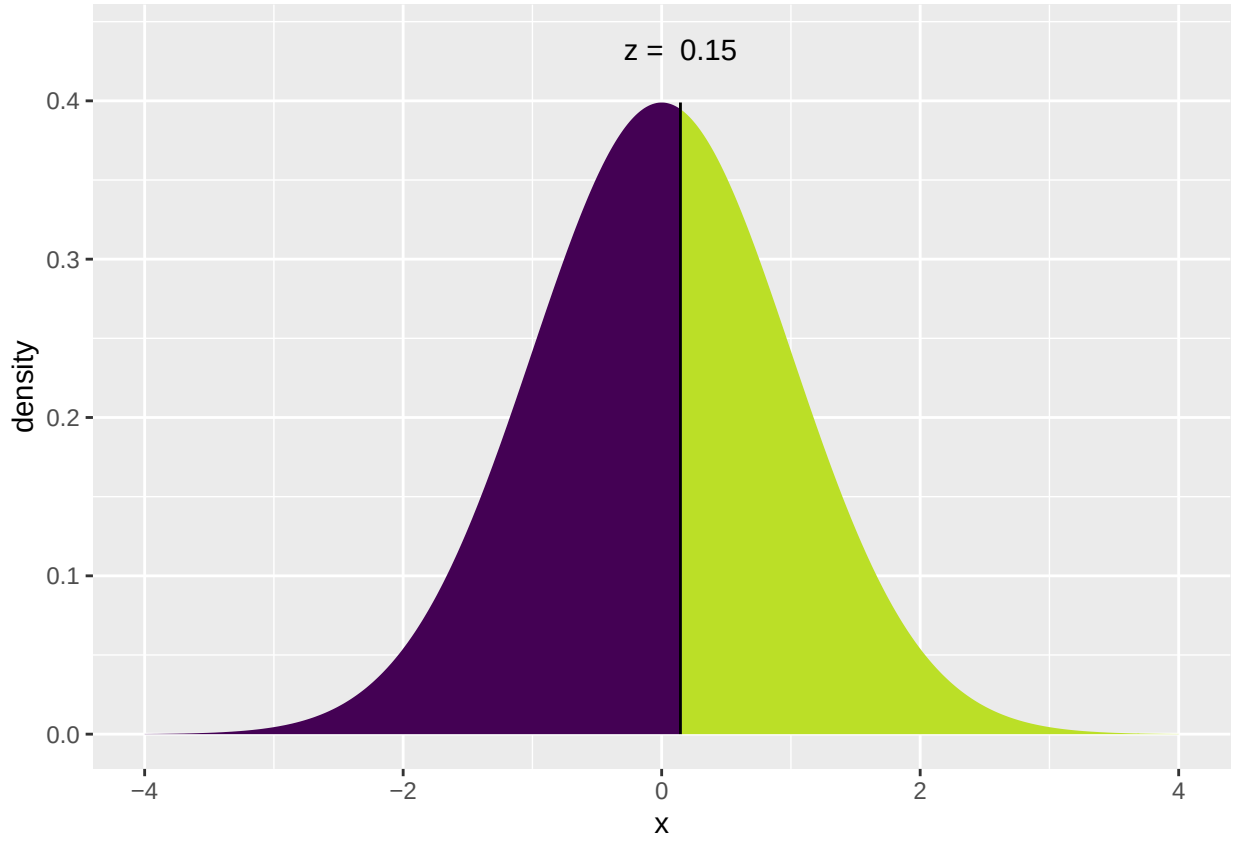
```
#proportion of positive sentiment videos preoutbreak from sample
p_hat1 = 41/98
```

```r
India_analysis_entertainment %>%
  filter(pre_outbreak == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      17
## 2 TRUE                       13
```

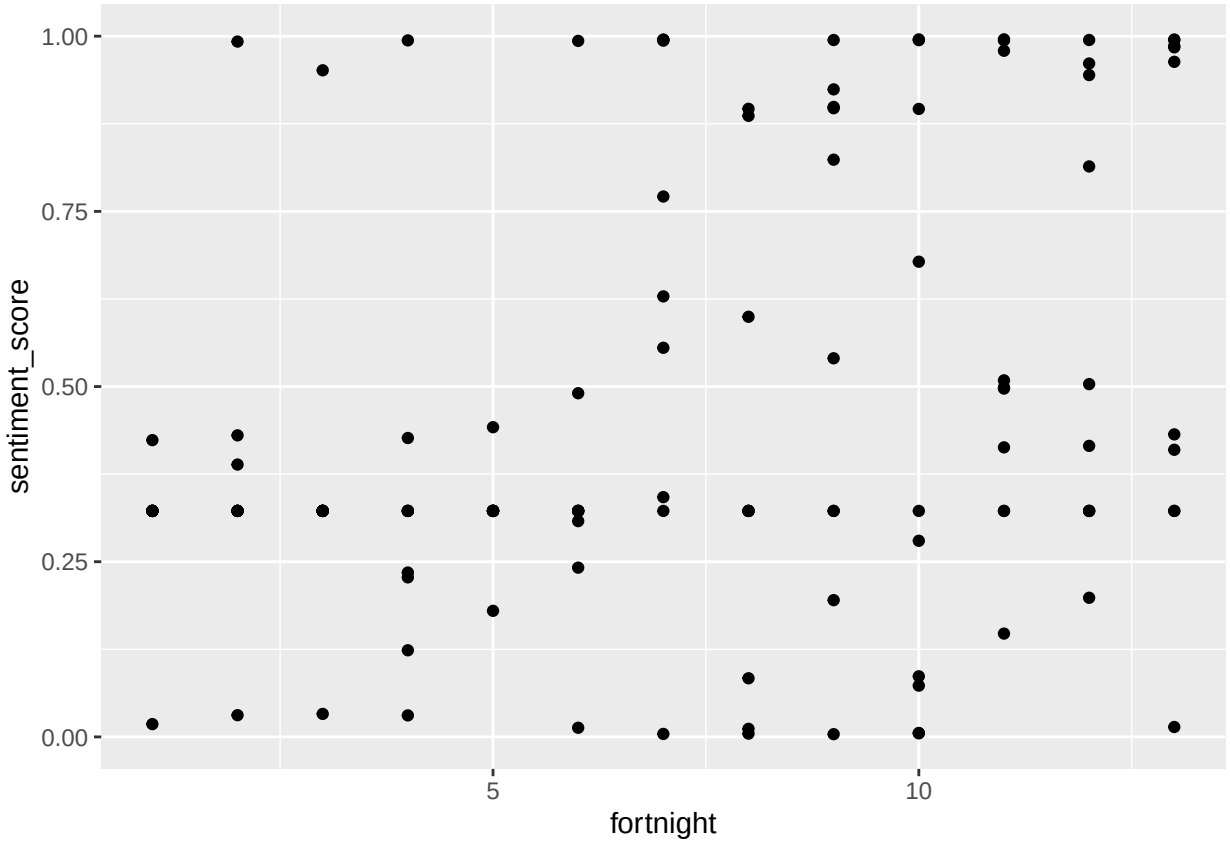
```r
#proportion of positive sentiment videos postoutbreak from sample
p_hat2 = 13/30

p_hat = (41+13)/(98+30)

sd <- sqrt((((p_hat)*(1-p_hat))/98)+(((p_hat)*(1-p_hat))/30))
z_score <- ((p_hat2-p_hat1)-0)/sd

#p-value
2* (1-xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= 0.1452) = P(Z <= 0.1452) = 0.5577
##  P(X >  0.1452) = P(Z >  0.1452) = 0.4423
##
```

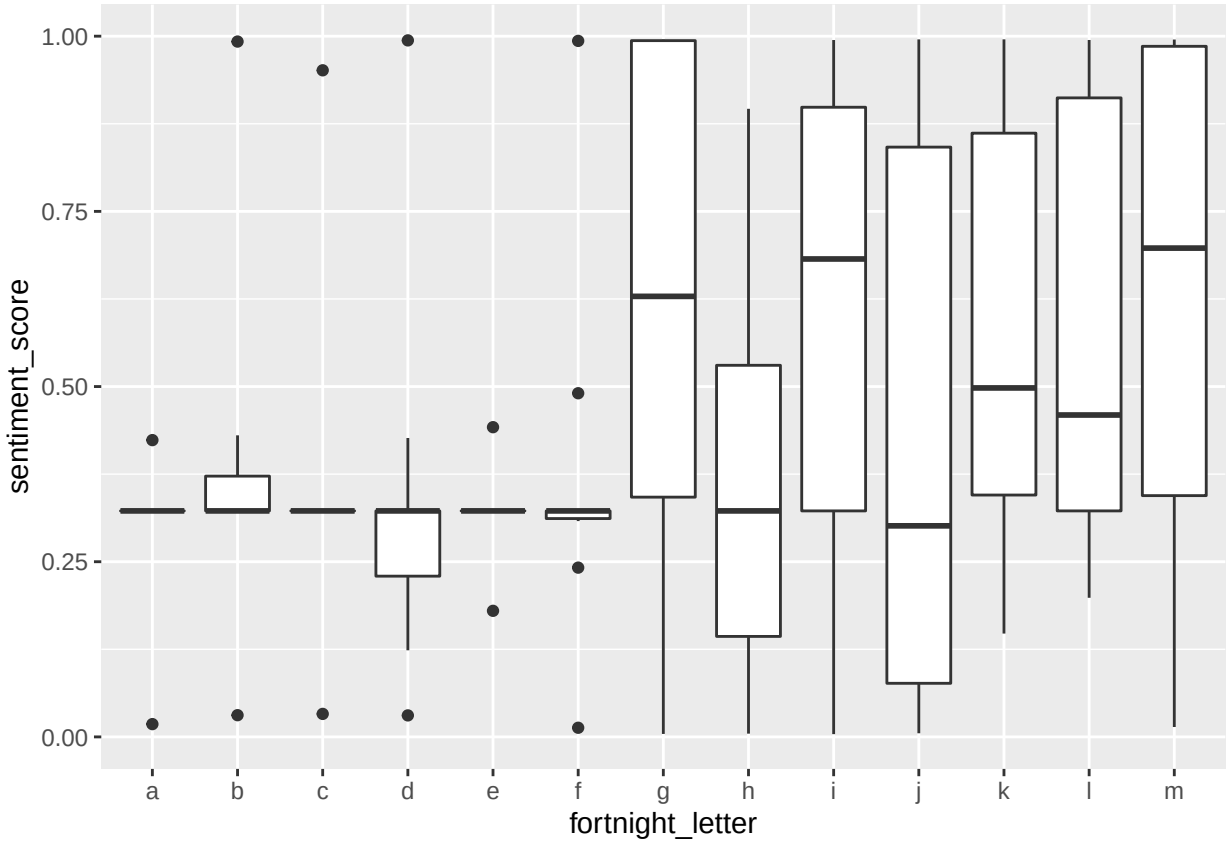



```
## [1] 0.8845253
```

```
#data summary news and politics
ggplot(India_analysis_news) +
  geom_point(aes(x = fortnight, y = sentiment_score))
```



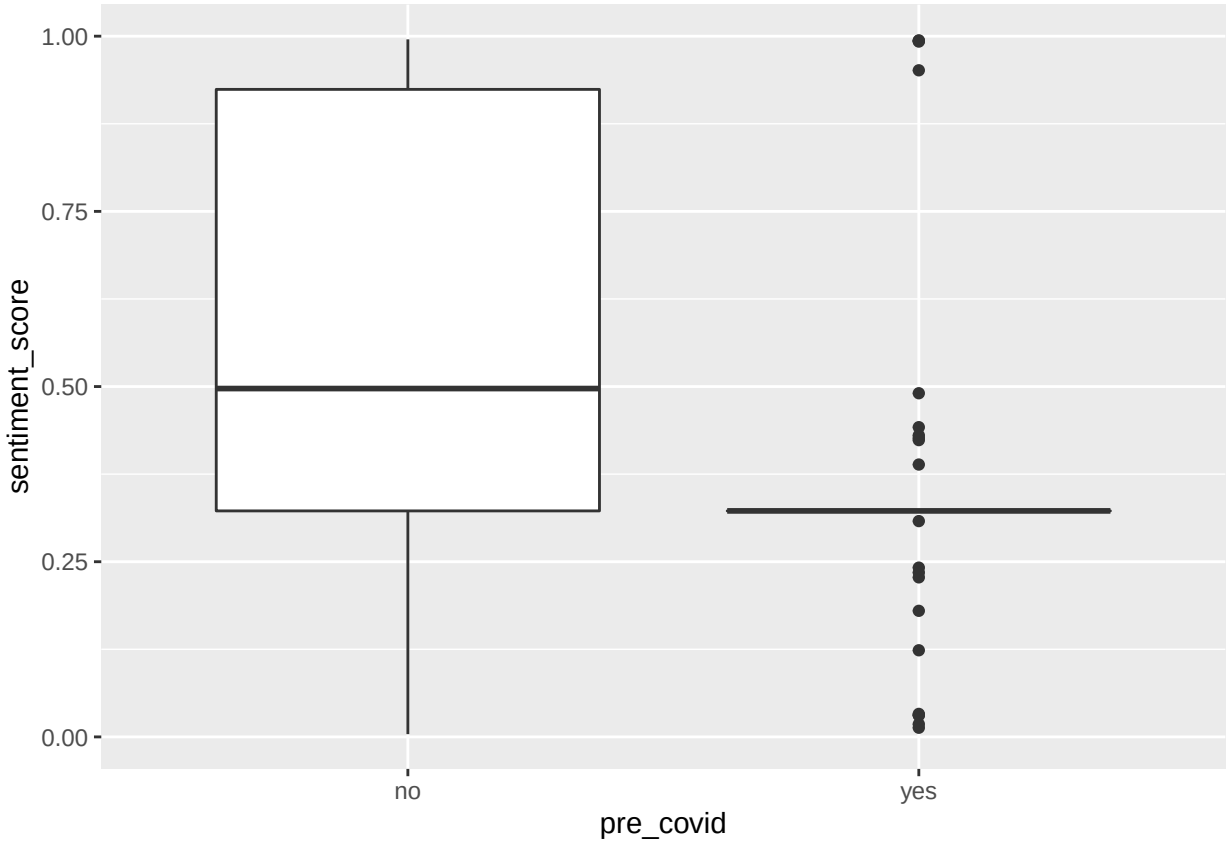

```
ggplot(India_analysis_news) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




```
India_analysis_news %>%
  group_by(fortnight) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.302
##  2         2                   0.378
##  3         3                   0.356
##  4         4                   0.333
##  5         5                   0.320
##  6         6                   0.366
##  7         7                   0.623
##  8         8                   0.377
##  9         9                   0.592
## 10        10                   0.434
## 11        11                   0.568
## 12        12                   0.580
## 13        13                   0.642
```

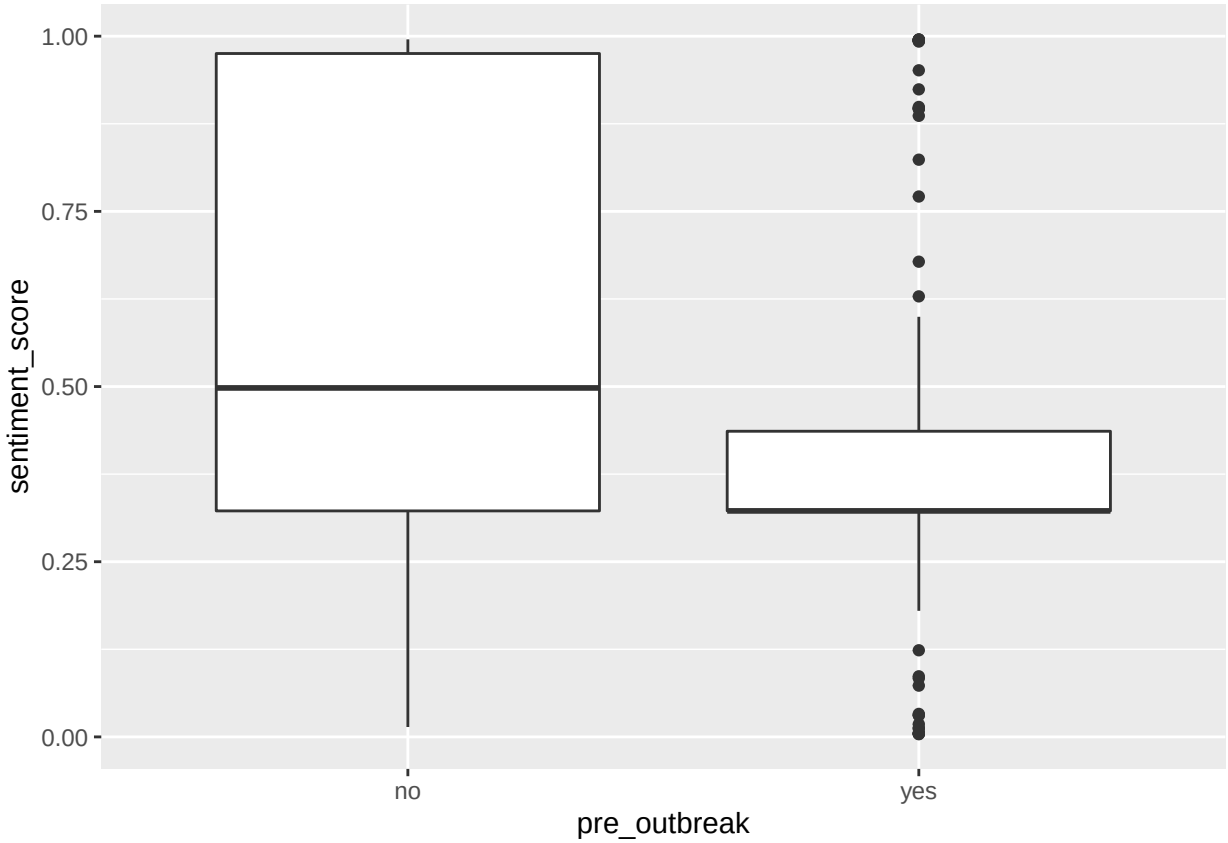
```
ggplot(India_analysis_news) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




```
India_analysis_news %>%
  group_by(pre_covid) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.544
## 2 yes                         0.343
```

```
ggplot(India_analysis_news) +
  geom_boxplot(aes(x = pre_outbreak, y = sentiment_score))
```




```
India_analysis_news %>%
  group_by(pre_outbreak) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.597
## 2 yes                            0.406
```

```
#precovid news
count(India_analysis_news, pre_covid == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_covid == "yes"`     n
##   <lgl>                <int>
## 1 FALSE                   69
## 2 TRUE                    60
```

```
num_precovid = 60
num_postcovid = 69
num = 129

India_analysis_news %>%
  filter(pre_covid == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
```

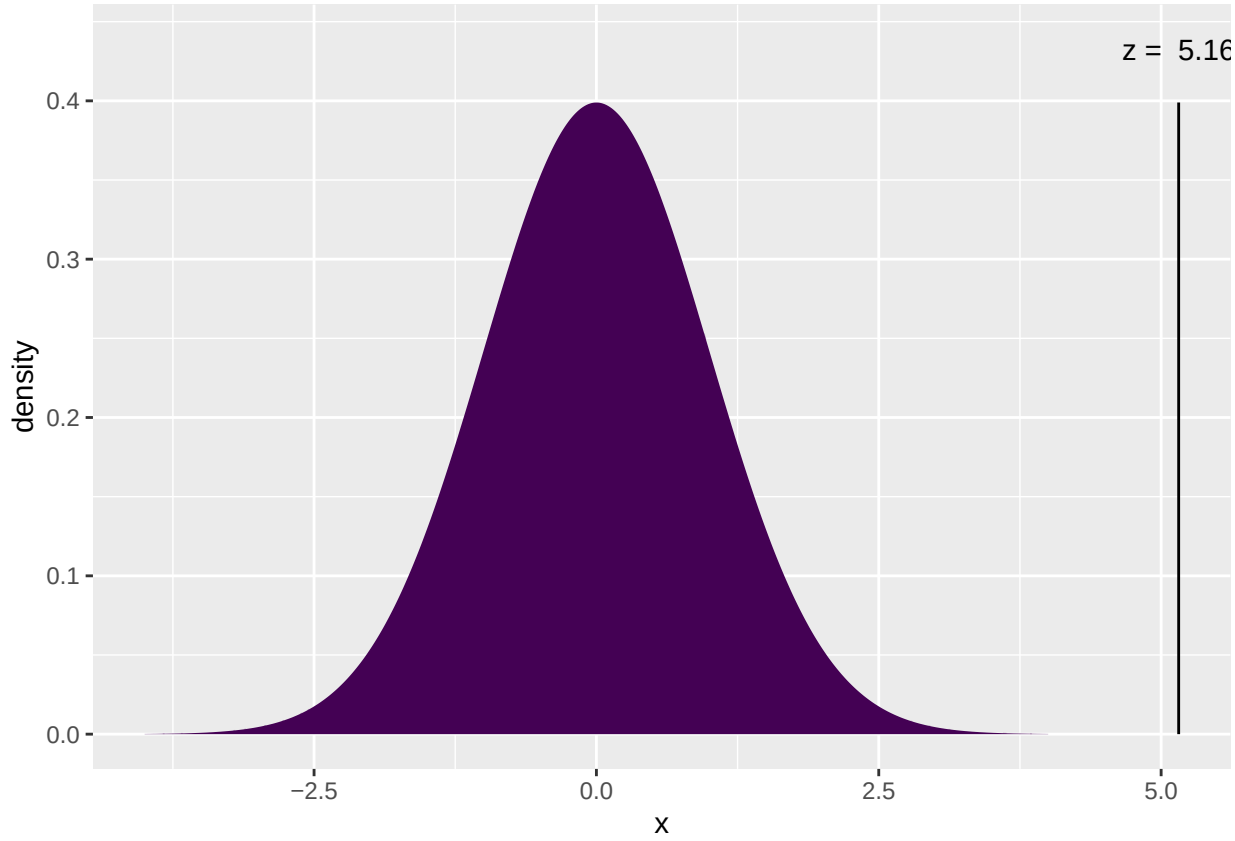
```
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      56
## 2 TRUE                        4
```

```
#proportion of positive sentiment videos precovid from sample
p_hat1 = 4/60

India_analysis_news %>%
  filter(pre_covid == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      36
## 2 TRUE                       33
```

```
#proportion of positive sentiment videos postcovid from sample
p_hat2 = 33/69

p_hat = (4+33)/(60+69)

sd <- sqrt((((p_hat)*(1-p_hat))/60)+(((p_hat)*(1-p_hat))/69))
z_score <- ((p_hat2-p_hat1)-0)/sd

#p-value
2* (1-xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= 5.155) = P(Z <= 5.155) = 1
##  P(X >  5.155) = P(Z >  5.155) = 1.265e-07
##
```




```
## [1] 2.529757e-07
```

```
#outbreak news
count(India_analysis_news, pre_outbreak == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_outbreak == "yes"`     n
##   <lgl>                   <int>
## 1 FALSE                      30
## 2 TRUE                       99
```

```
num_preoutbreak = 99
num_postoutbreak = 30
num = 129

India_analysis_news %>%
  filter(pre_outbreak == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      76
## 2 TRUE                       23
```

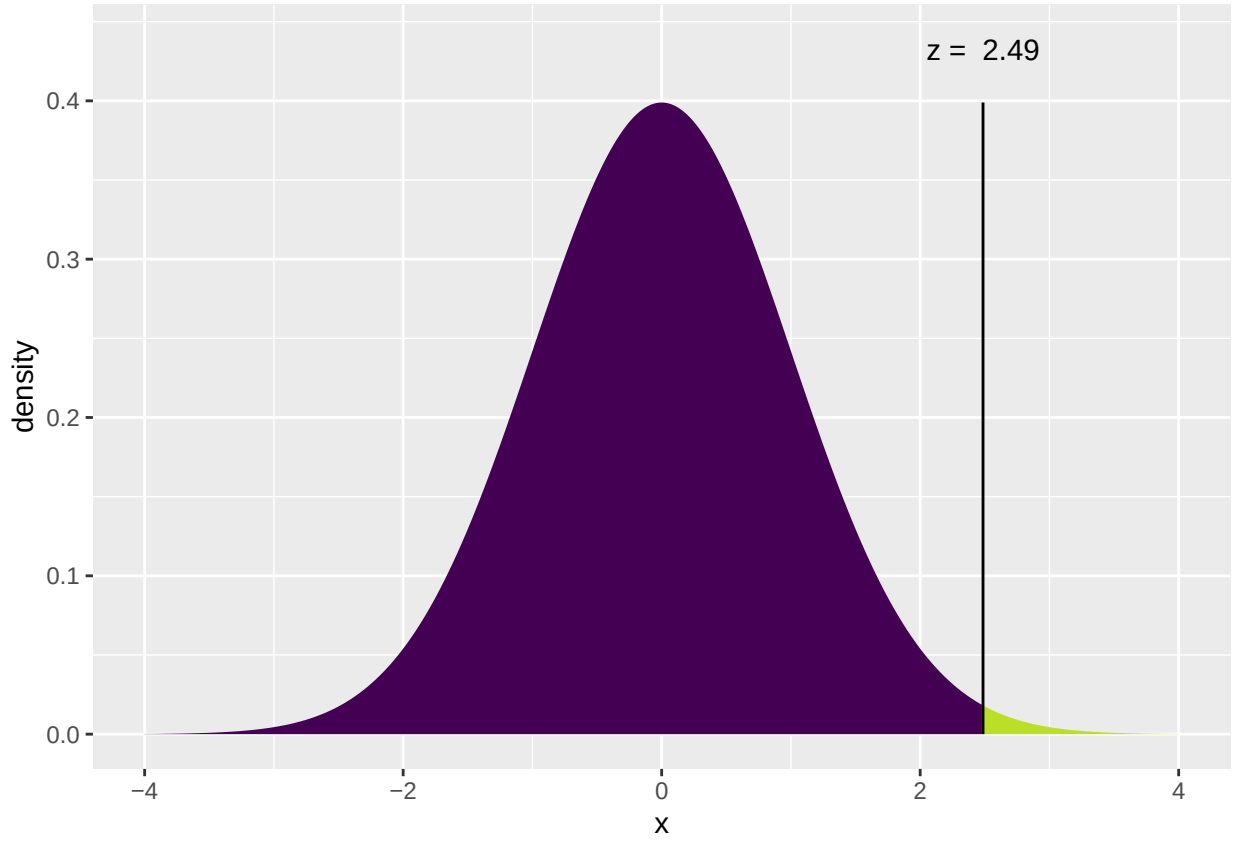
```
#proportion of positive sentiment videos preoutbreak from sample
p_hat1 = 23/99
```

```
India_analysis_news %>%
  filter(pre_outbreak == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      16
## 2 TRUE                       14
```

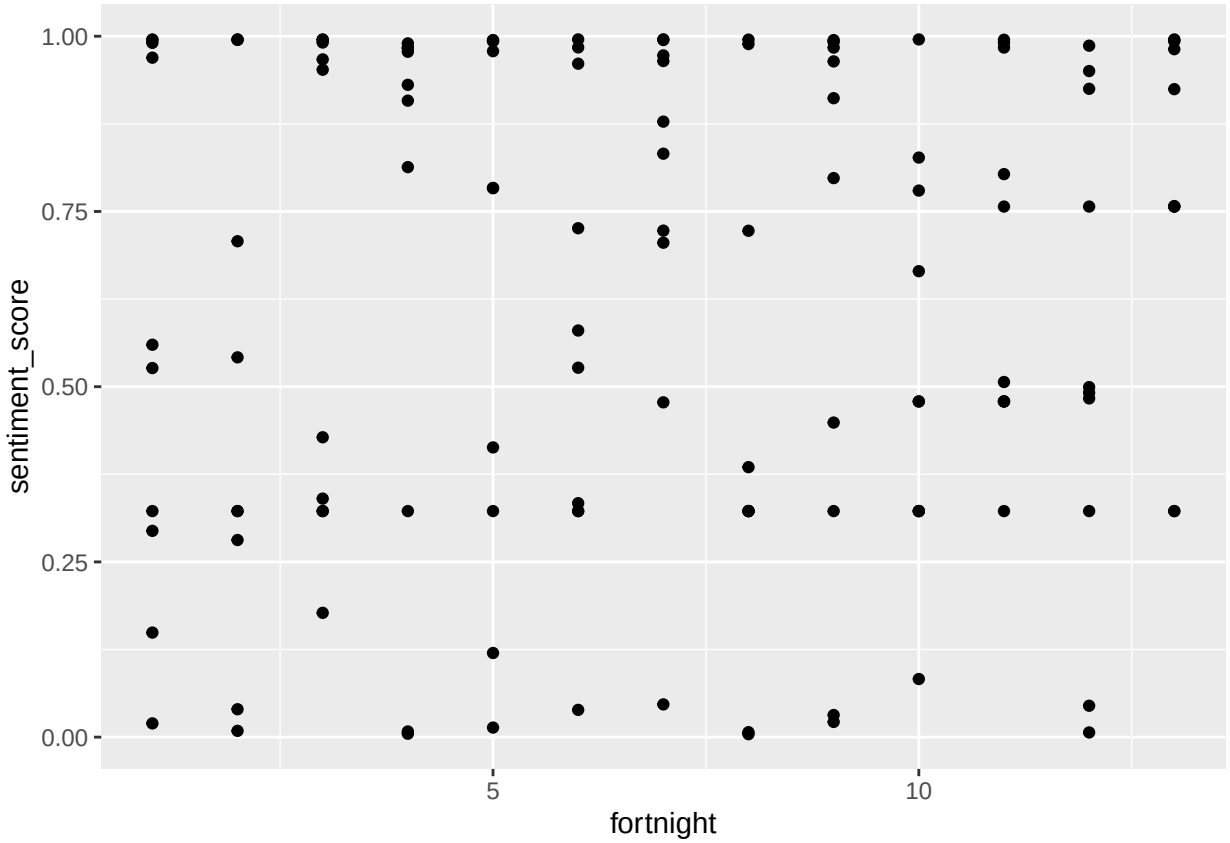
```
#proportion of positive sentiment videos postoutbreak from sample
p_hat2 = 14/30

p_hat = (23+14)/(99+30)

sd <- sqrt((((p_hat)*(1-p_hat))/99)+(((p_hat)*(1-p_hat))/30))
z_score <- ((p_hat2-p_hat1)-0)/sd

#p-value
2* (1-xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= 2.486) = P(Z <= 2.486) = 0.9935
##  P(X >  2.486) = P(Z >  2.486) = 0.006456
##
```

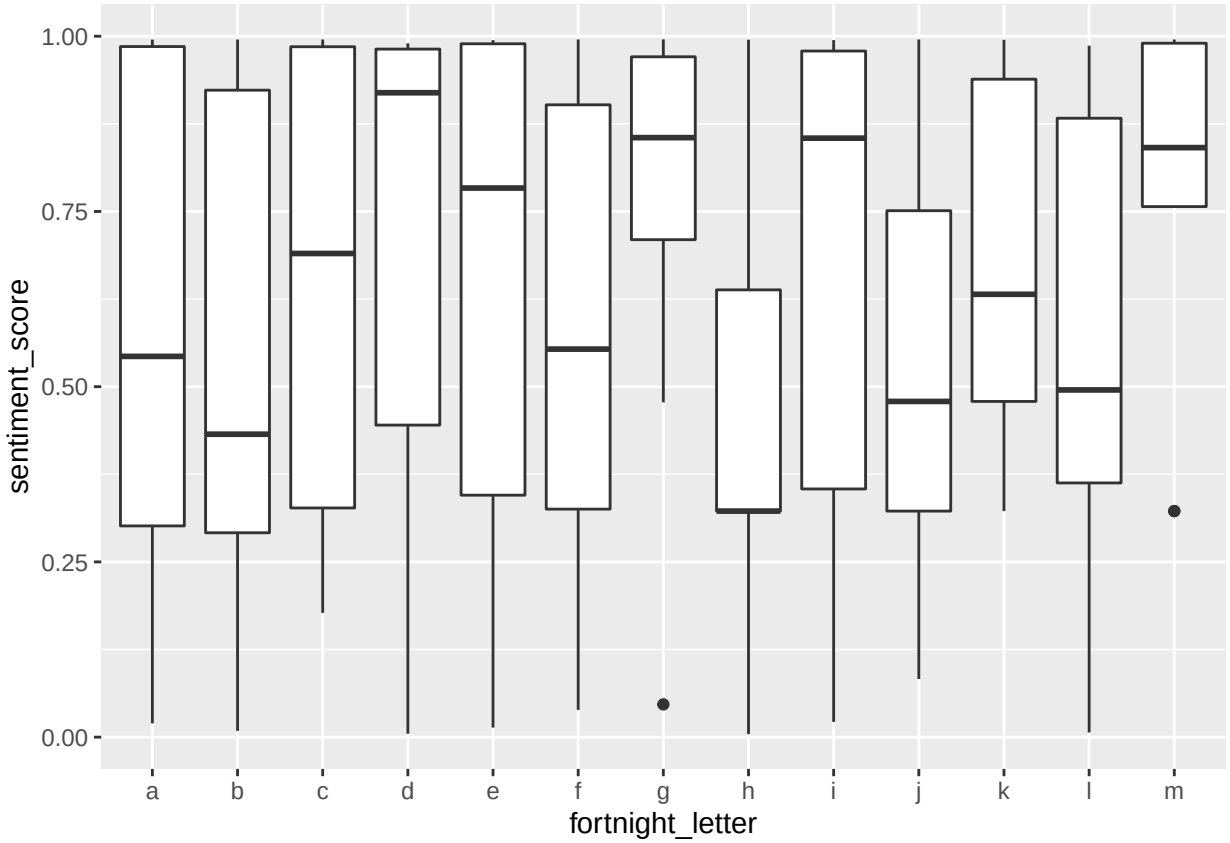



```
## [1] 0.01291257
```

```
#data summary how-to and style
ggplot(India_analysis_how_to) +
  geom_point(aes(x = fortnight, y = sentiment_score))
```



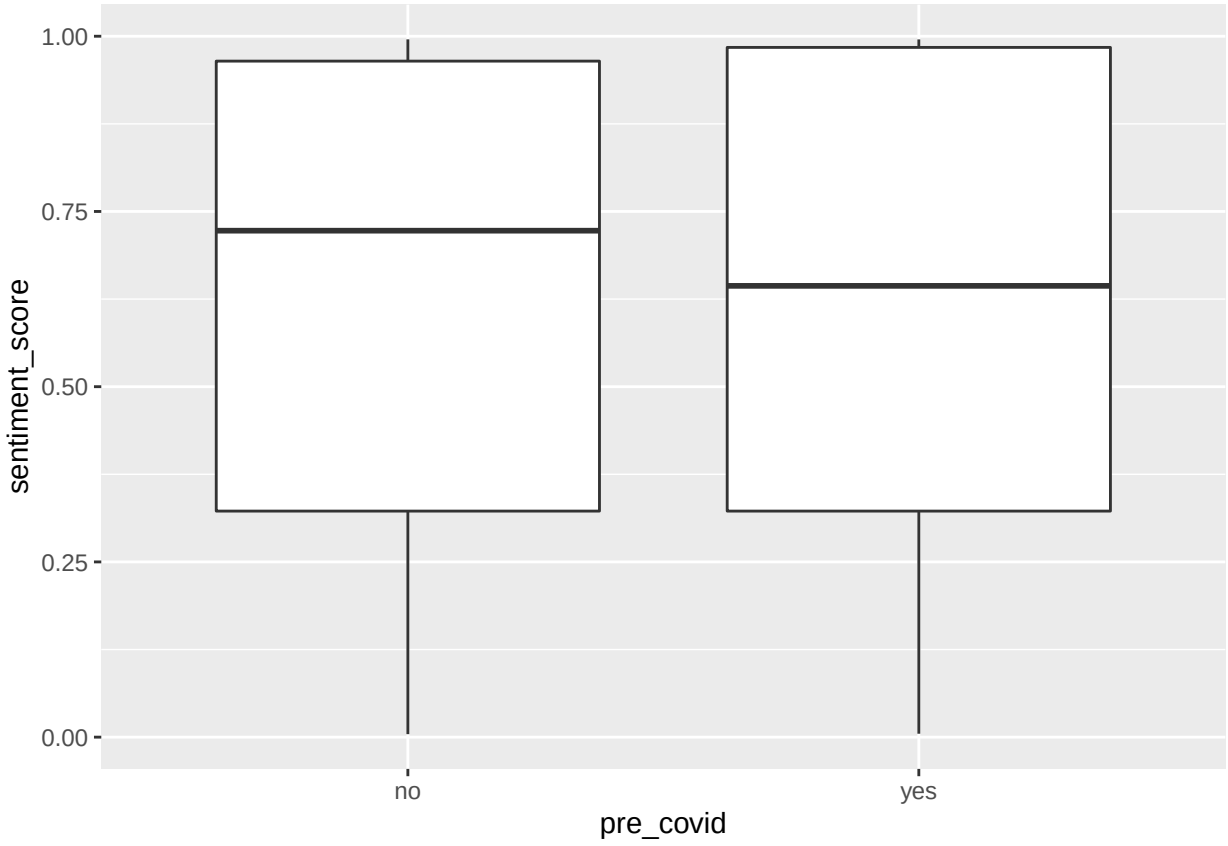

```
ggplot(India_analysis_how_to) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




```
India_analysis_how_to %>%
  group_by(fortnight) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.582
##  2         2                   0.521
##  3         3                   0.649
##  4         4                   0.692
##  5         5                   0.640
##  6         6                   0.579
##  7         7                   0.759
##  8         8                   0.439
##  9         9                   0.647
## 10        10                   0.528
## 11        11                   0.680
## 12        12                   0.547
## 13        13                   0.781
```

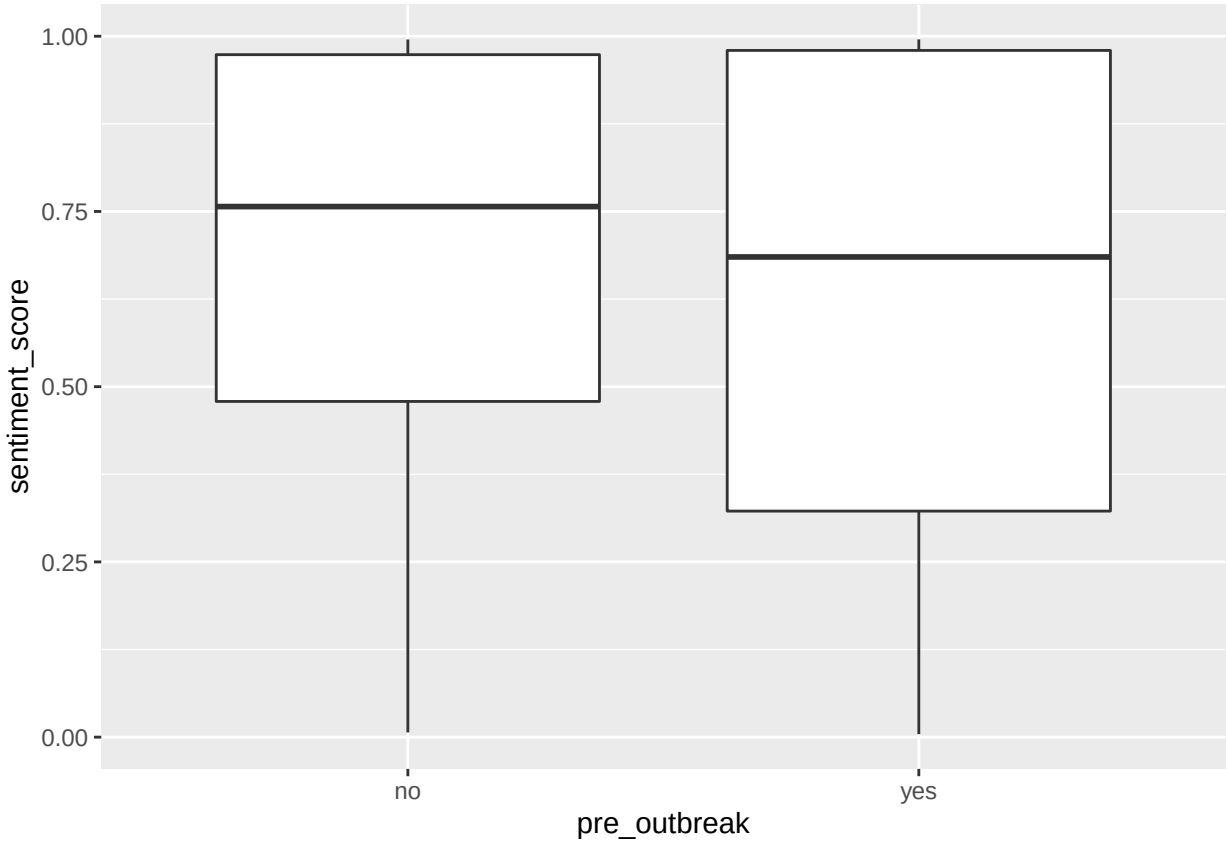
```
ggplot(India_analysis_how_to) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




```
India_analysis_how_to %>%
  group_by(pre_covid) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.626
## 2 yes                         0.611
```

```
ggplot(India_analysis_how_to) +
  geom_boxplot(aes(x = pre_outbreak, y = sentiment_score))
```




```
India_analysis_how_to %>%
  group_by(pre_outbreak) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.669
## 2 yes                            0.604
```

```
#precovid how-to
count(India_analysis_how_to, pre_covid == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_covid == "yes"`     n
##   <lgl>                <int>
## 1 FALSE                   70
## 2 TRUE                    60
```

```
num_precovid = 60
num_postcovid = 70
num = 130

India_analysis_how_to %>%
  filter(pre_covid == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
```

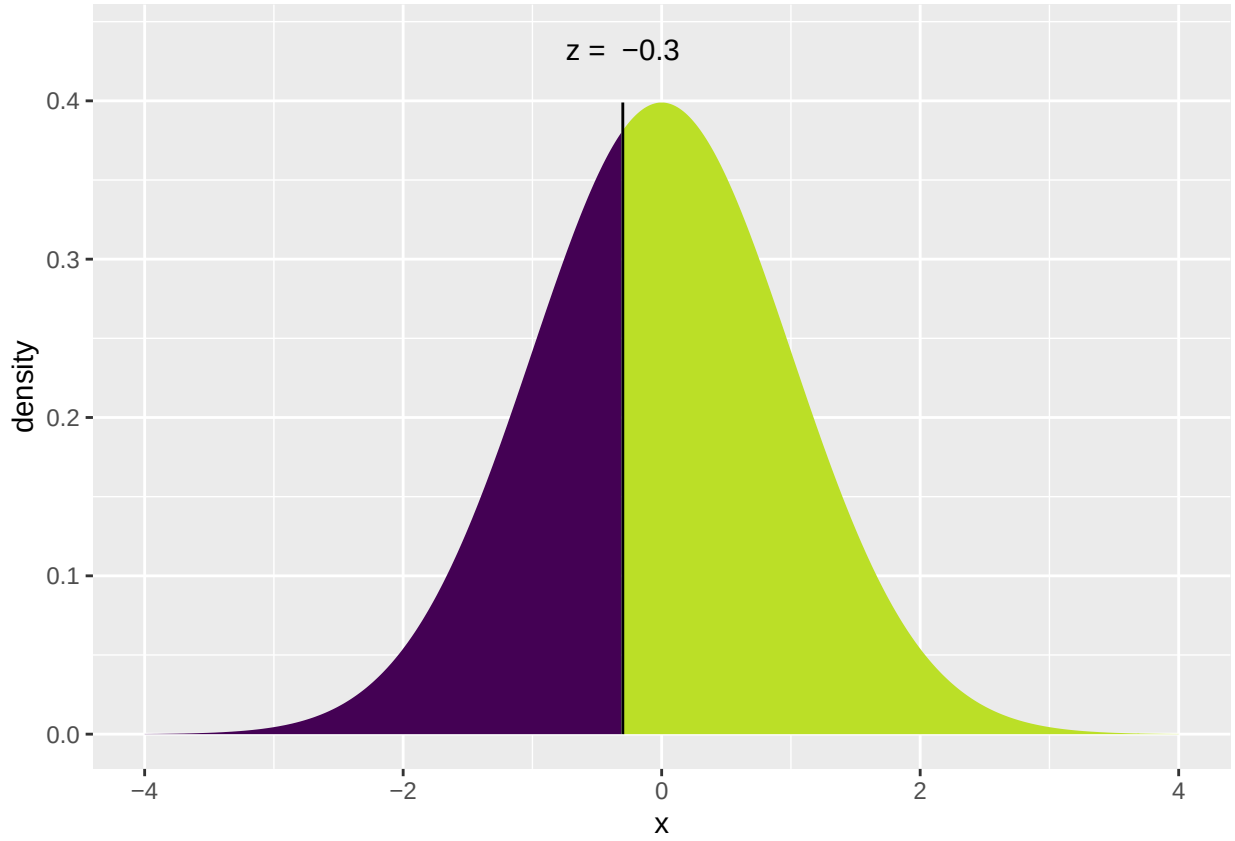
```
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      25
## 2 TRUE                       35
```

```
#proportion of positive sentiment videos precovid from sample
p_hat1 = 35/60

India_analysis_how_to %>%
  filter(pre_covid == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      31
## 2 TRUE                       39
```

```
#proportion of positive sentiment videos postcovid from sample
p_hat2 = 39/70

p_hat = (35+39)/(60+70)

sd <- sqrt((((p_hat)*(1-p_hat))/60)+(((p_hat)*(1-p_hat))/70))
z_score <- ((p_hat2-p_hat1)-0)/sd

#p-value
2* (xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= -0.3006) = P(Z <= -0.3006) = 0.3818
##  P(X >  -0.3006) = P(Z >  -0.3006) = 0.6182
##
```




```
## [1] 0.763698
```

```
#outbreak how-to
count(India_analysis_how_to, pre_outbreak == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_outbreak == "yes"`     n
##   <lgl>                   <int>
## 1 FALSE                      30
## 2 TRUE                      100
```

```
num_preoutbreak = 100
num_postoutbreak = 30
num = 130

India_analysis_how_to %>%
  filter(pre_outbreak == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      44
## 2 TRUE                       56
```

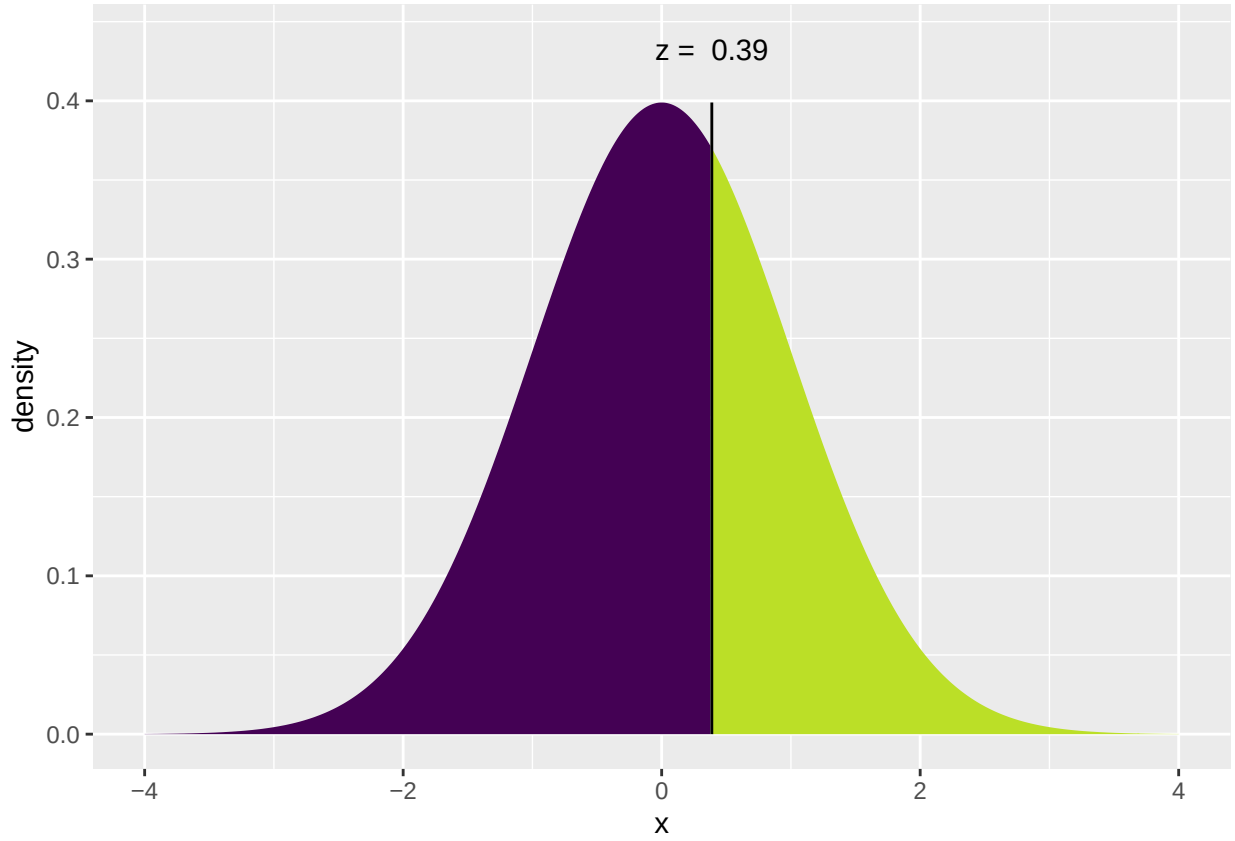
```
#proportion of positive sentiment videos preoutbreak from sample
p_hat1 = 56/100
```

```
India_analysis_how_to %>%
  filter(pre_outbreak == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      12
## 2 TRUE                       18
```

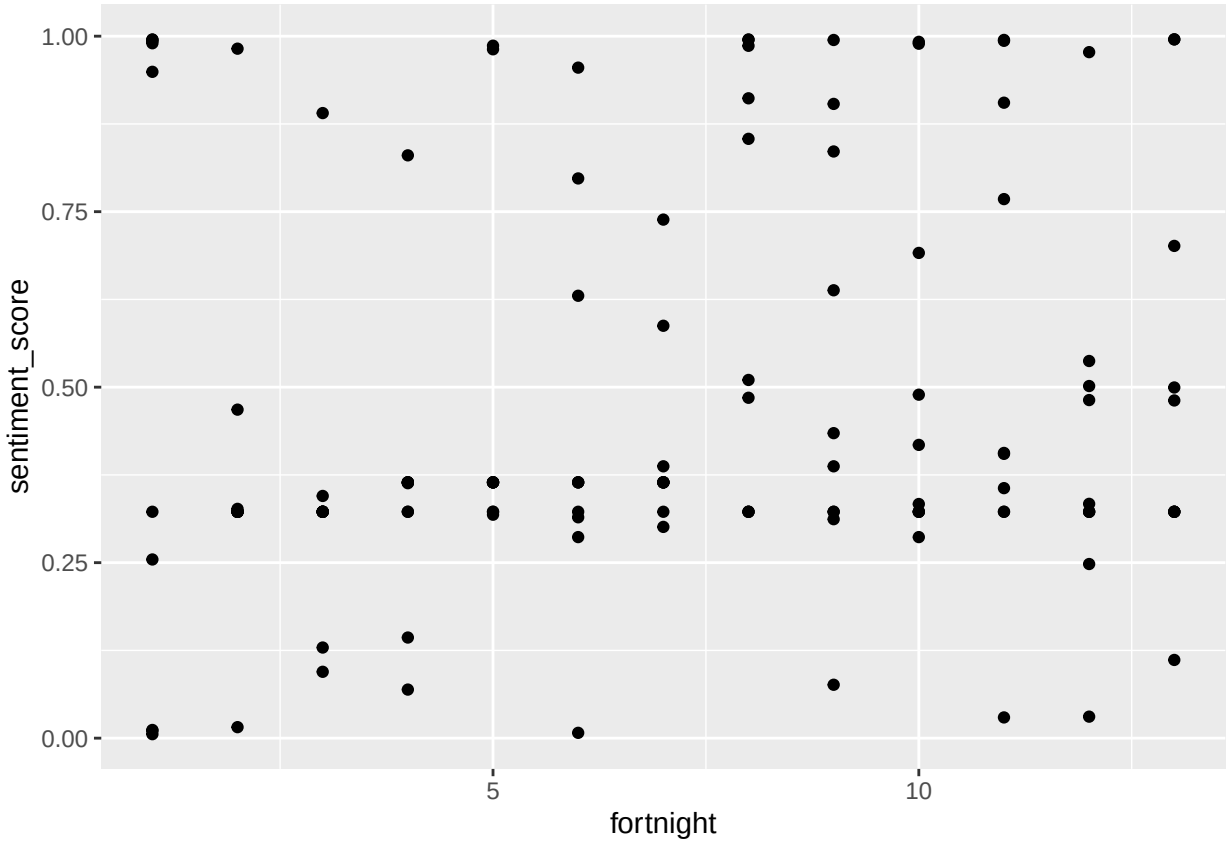
```
#proportion of positive sentiment videos postoutbreak from sample
p_hat2 = 18/30

p_hat = (56+18)/(100+30)

sd <- sqrt((((p_hat)*(1-p_hat))/100)+(((p_hat)*(1-p_hat))/30))
z_score <- ((p_hat2-p_hat1)-0)/sd

#p-value
2* (1-xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= 0.388) = P(Z <= 0.388) = 0.651
##  P(X >  0.388) = P(Z >  0.388) = 0.349
##
```

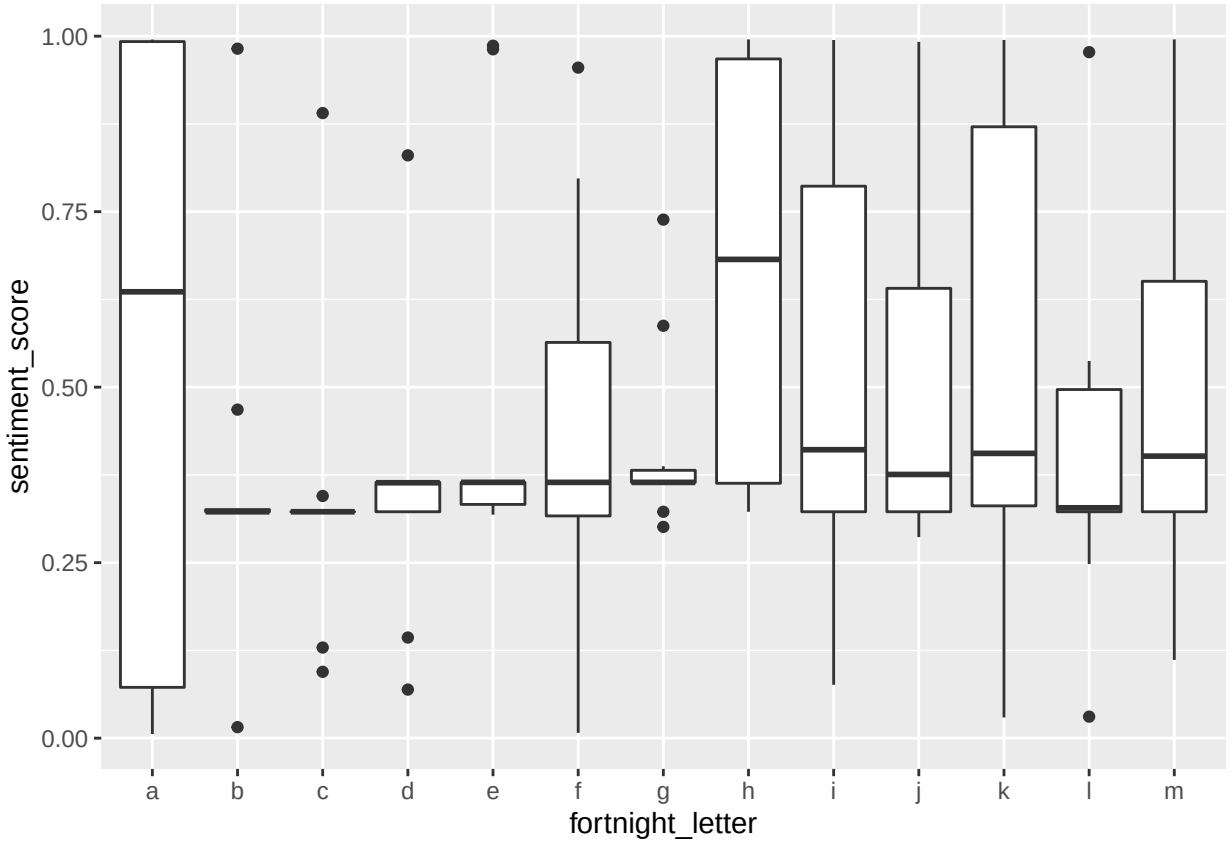



```
## [1] 0.6979825
```

```
#data summary education
ggplot(India_analysis_education) +
  geom_point(aes(x = fortnight, y = sentiment_score))
```



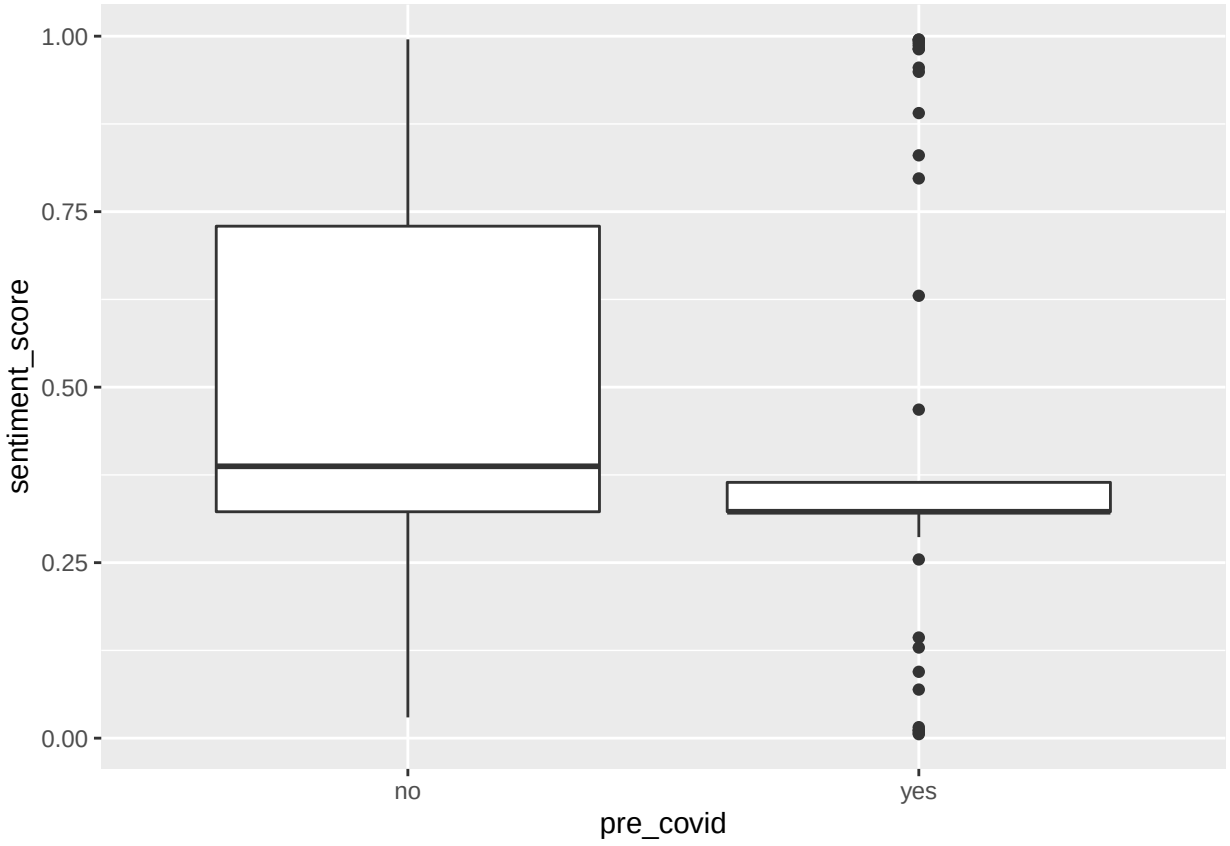

```
ggplot(India_analysis_education) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




```
India_analysis_education %>%
  group_by(fortnight) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.553
##  2         2                   0.373
##  3         3                   0.339
##  4         4                   0.351
##  5         5                   0.475
##  6         6                   0.441
##  7         7                   0.416
##  8         8                   0.670
##  9         9                   0.523
## 10        10                   0.517
## 11        11                   0.550
## 12        12                   0.408
## 13        13                   0.507
```

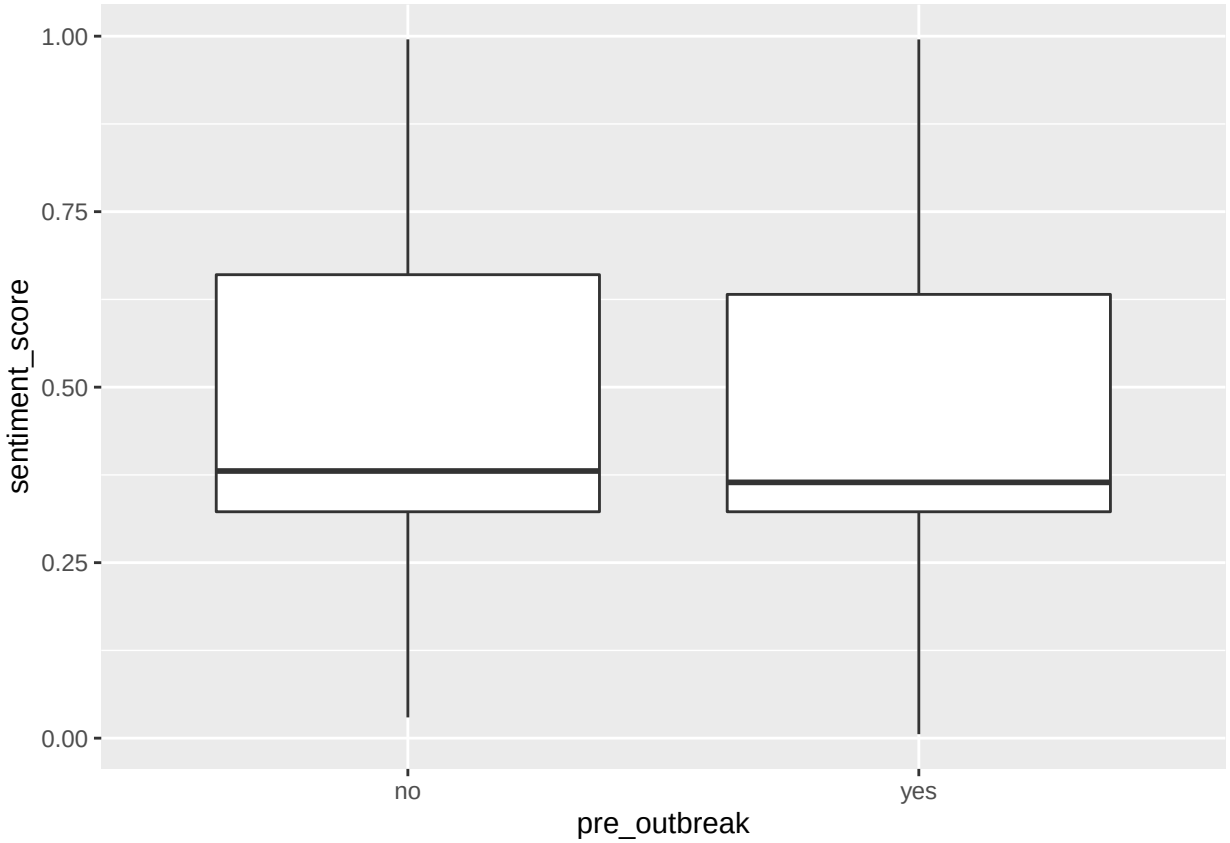
```
ggplot(India_analysis_education) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




```
India_analysis_education %>%
  group_by(pre_covid) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.513
## 2 yes                         0.422
```

```
ggplot(India_analysis_education) +
  geom_boxplot(aes(x = pre_outbreak, y = sentiment_score))
```




```
India_analysis_education %>%
  group_by(pre_outbreak) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.489
## 2 yes                            0.466
```

```
#precovid education
count(India_analysis_education, pre_covid == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_covid == "yes"`     n
##   <lgl>                <int>
## 1 FALSE                   70
## 2 TRUE                    60
```

```
num_precovid = 60
num_postcovid = 70
num = 130

India_analysis_education %>%
  filter(pre_covid == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
```

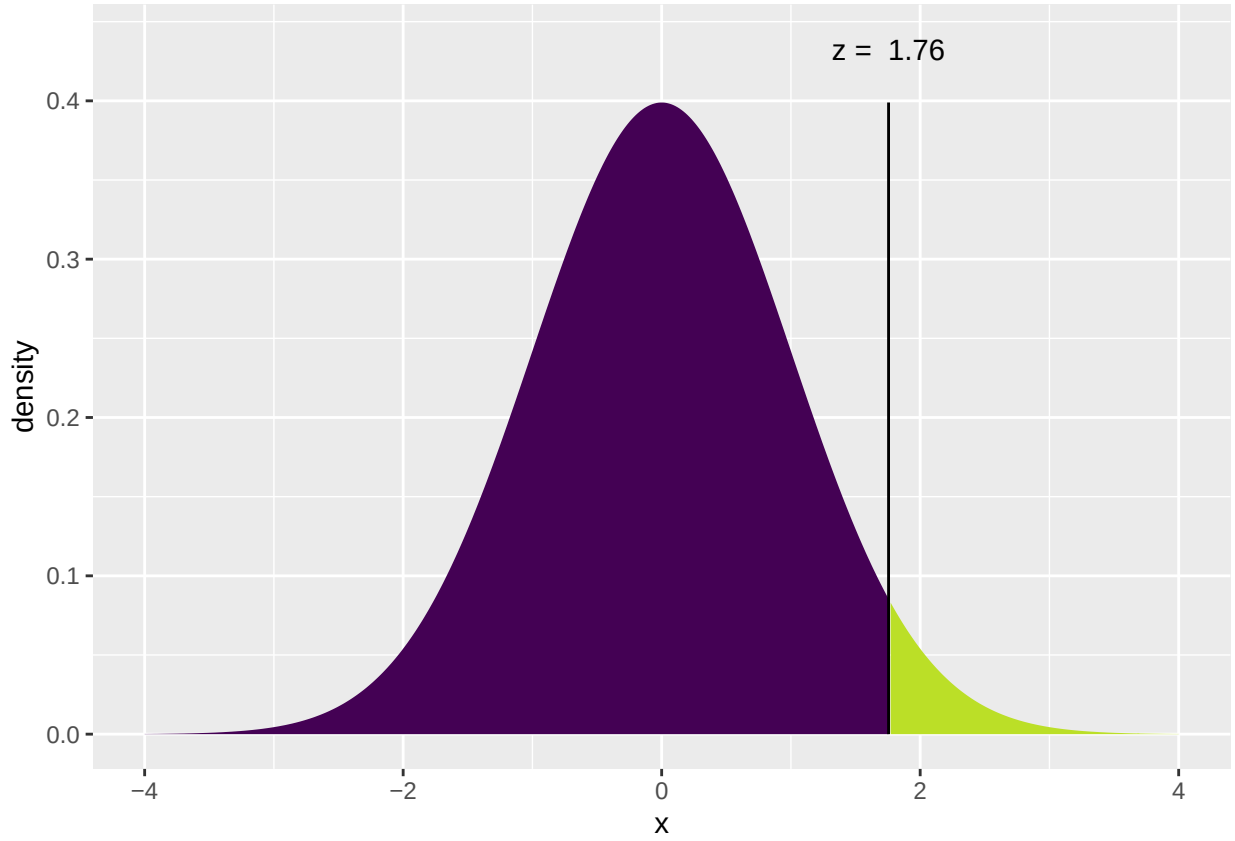
```
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      47
## 2 TRUE                       13
```

```
#proportion of positive sentiment videos precovid from sample
p_hat1 = 13/60

India_analysis_education %>%
  filter(pre_covid == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      45
## 2 TRUE                       25
```

```
#proportion of positive sentiment videos postcovid from sample
p_hat2 = 25/70

p_hat = (13+25)/(60+70)

sd <- sqrt((((p_hat)*(1-p_hat))/60)+(((p_hat)*(1-p_hat))/70))
z_score <- ((p_hat2-p_hat1)-0)/sd

#p-value
2* (1-xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= 1.756) = P(Z <= 1.756) = 0.9604
##  P(X >  1.756) = P(Z >  1.756) = 0.03958
##
```




```
## [1] 0.07916519
```

```
#outbreak education
count(India_analysis_education, pre_outbreak == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_outbreak == "yes"`     n
##   <lgl>                   <int>
## 1 FALSE                      30
## 2 TRUE                      100
```

```
num_preoutbreak = 100
num_postoutbreak = 30
num = 130

India_analysis_education %>%
  filter(pre_outbreak == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      72
## 2 TRUE                       28
```

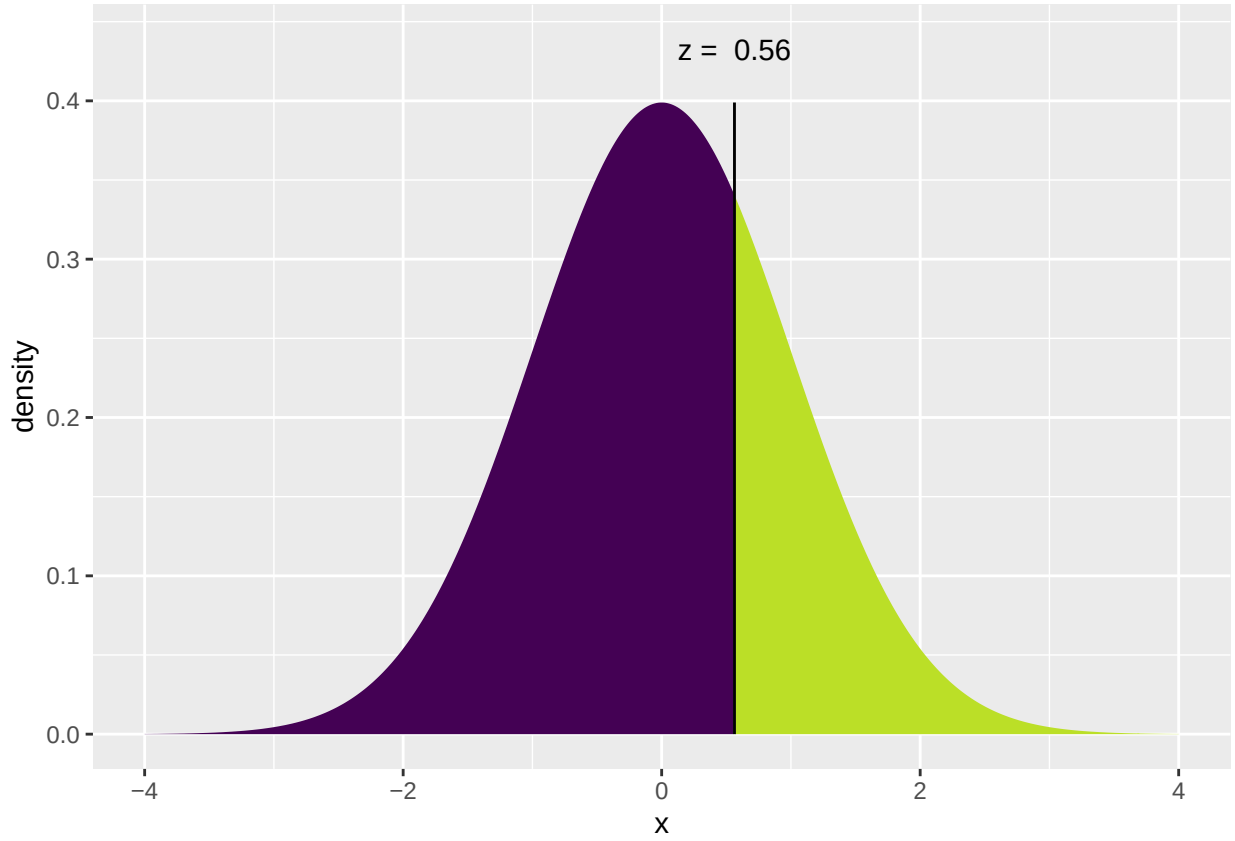
```
#proportion of positive sentiment videos preoutbreak from sample
p_hat1 = 28/100
```

```
India_analysis_education %>%
  filter(pre_outbreak == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      20
## 2 TRUE                       10
```

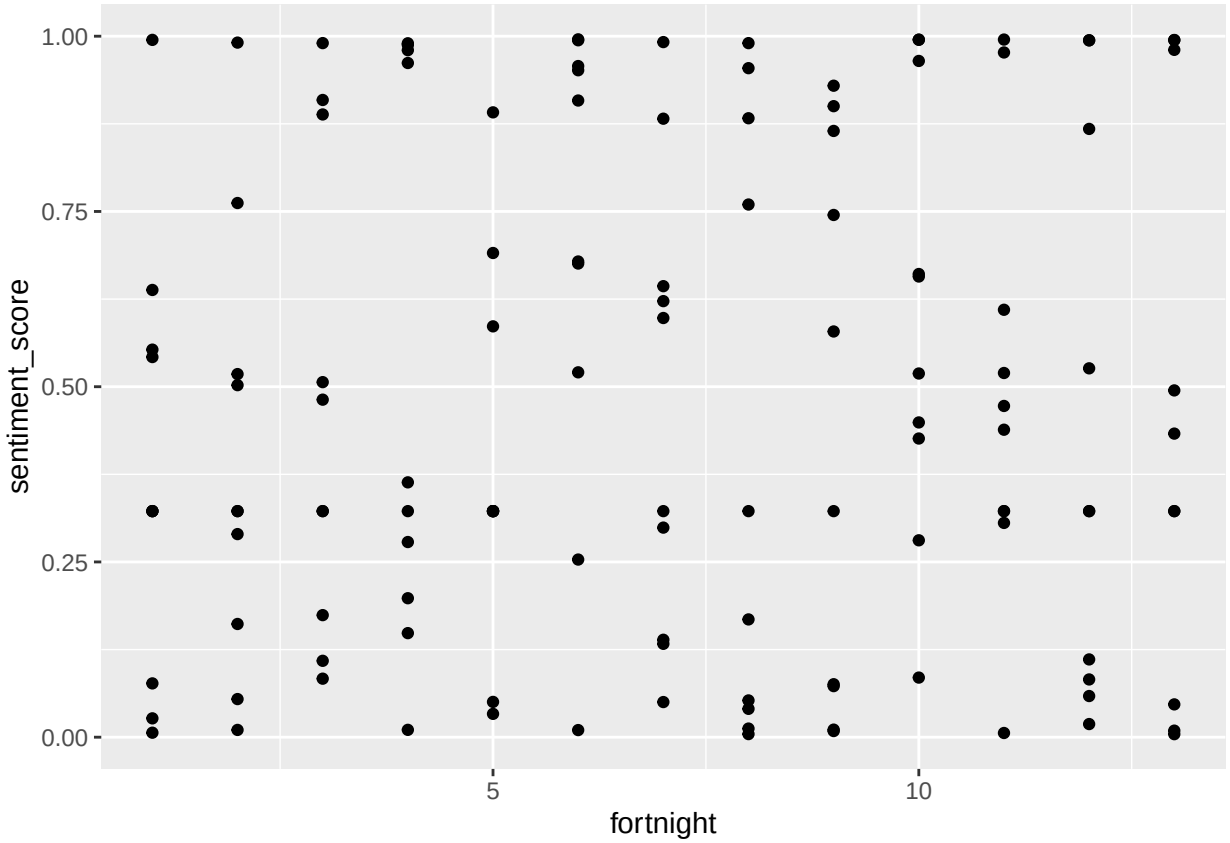
```
#proportion of positive sentiment videos postoutbreak from sample
p_hat2 = 10/30

p_hat = (28+10)/(100+30)

sd <- sqrt((((p_hat)*(1-p_hat))/100)+(((p_hat)*(1-p_hat))/30))
z_score <- ((p_hat2-p_hat1)-0)/sd

#p-value
2* (1-xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= 0.5633) = P(Z <= 0.5633) = 0.7134
##  P(X >  0.5633) = P(Z >  0.5633) = 0.2866
##
```

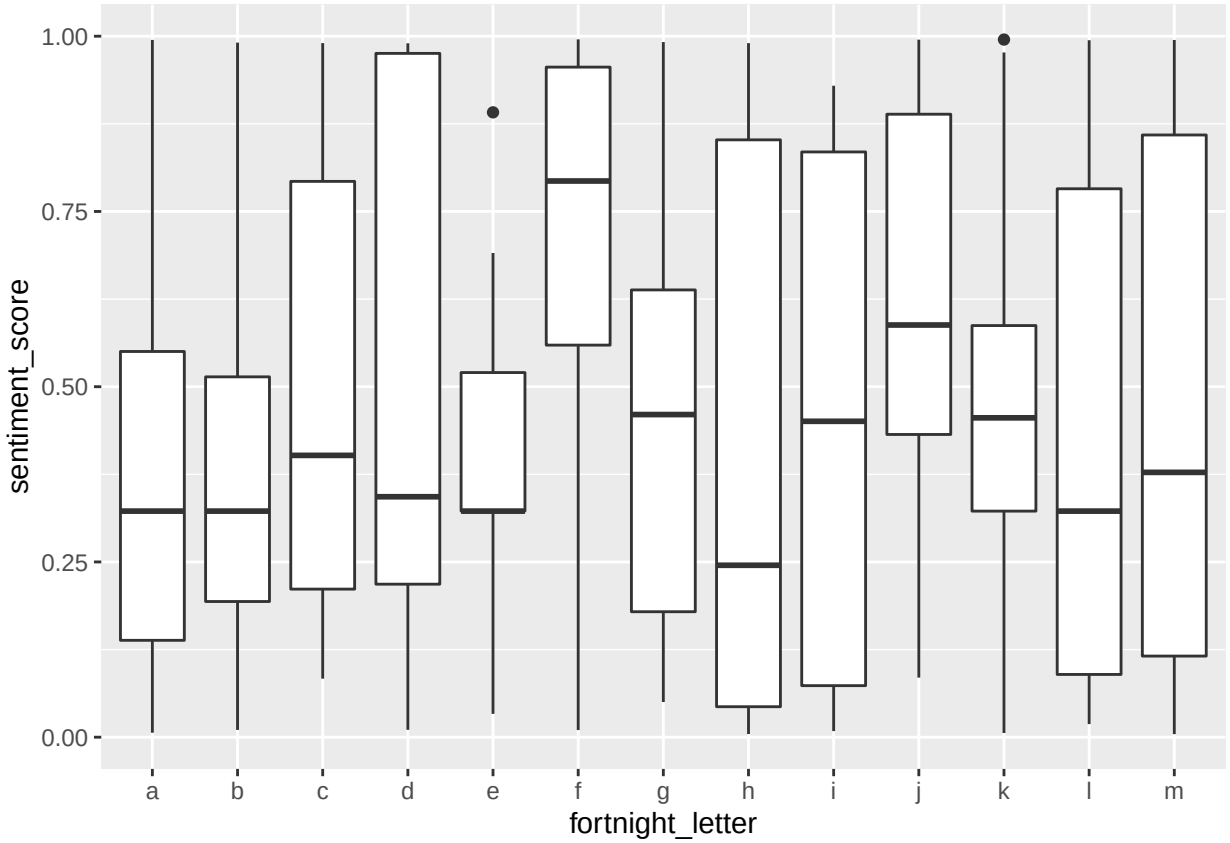



```
## [1] 0.5732257
```

```
#data summary science and technology
ggplot(India_analysis_science) +
  geom_point(aes(x = fortnight, y = sentiment_score))
```



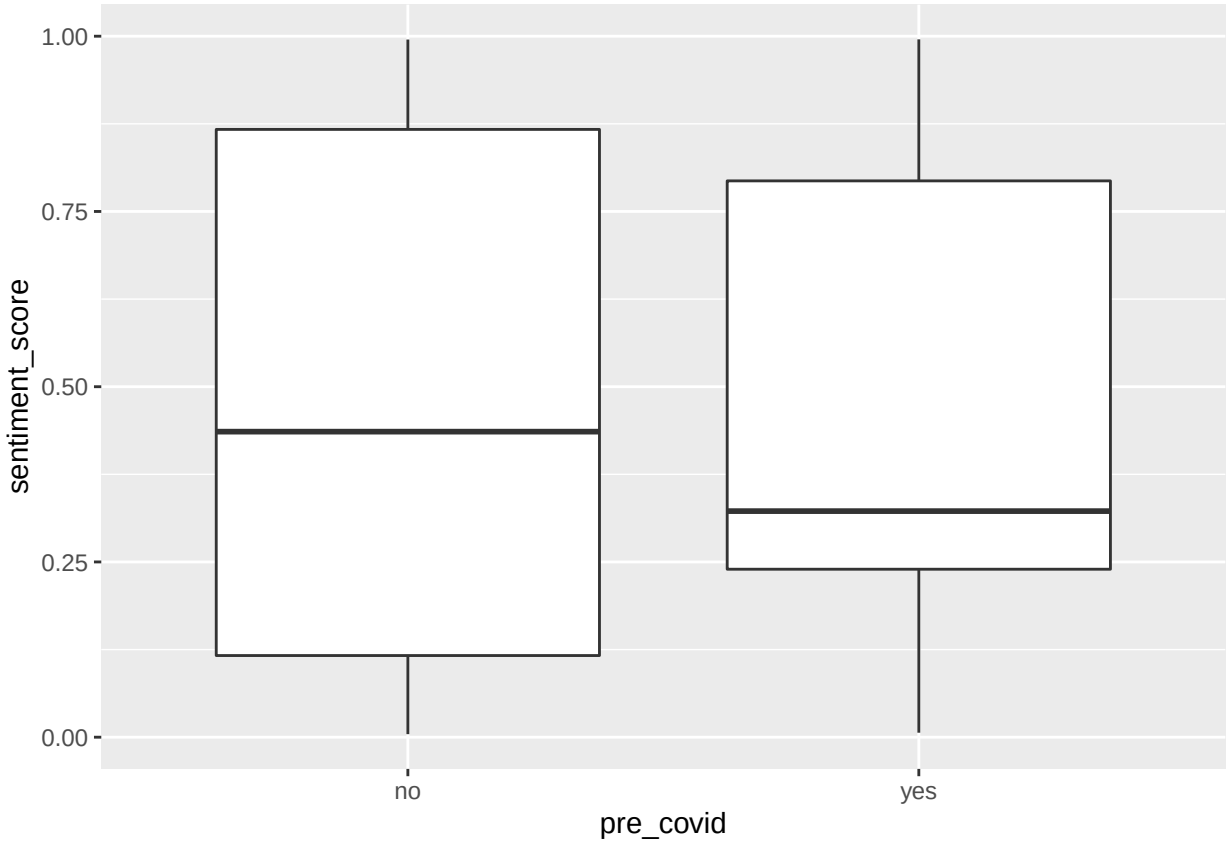

```
ggplot(India_analysis_science) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




```
India_analysis_science %>%
  group_by(fortnight) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.381
##  2         2                   0.393
##  3         3                   0.479
##  4         4                   0.524
##  5         5                   0.386
##  6         6                   0.695
##  7         7                   0.468
##  8         8                   0.419
##  9         9                   0.451
## 10        10                   0.603
## 11        11                   0.497
## 12        12                   0.430
## 13        13                   0.460
```

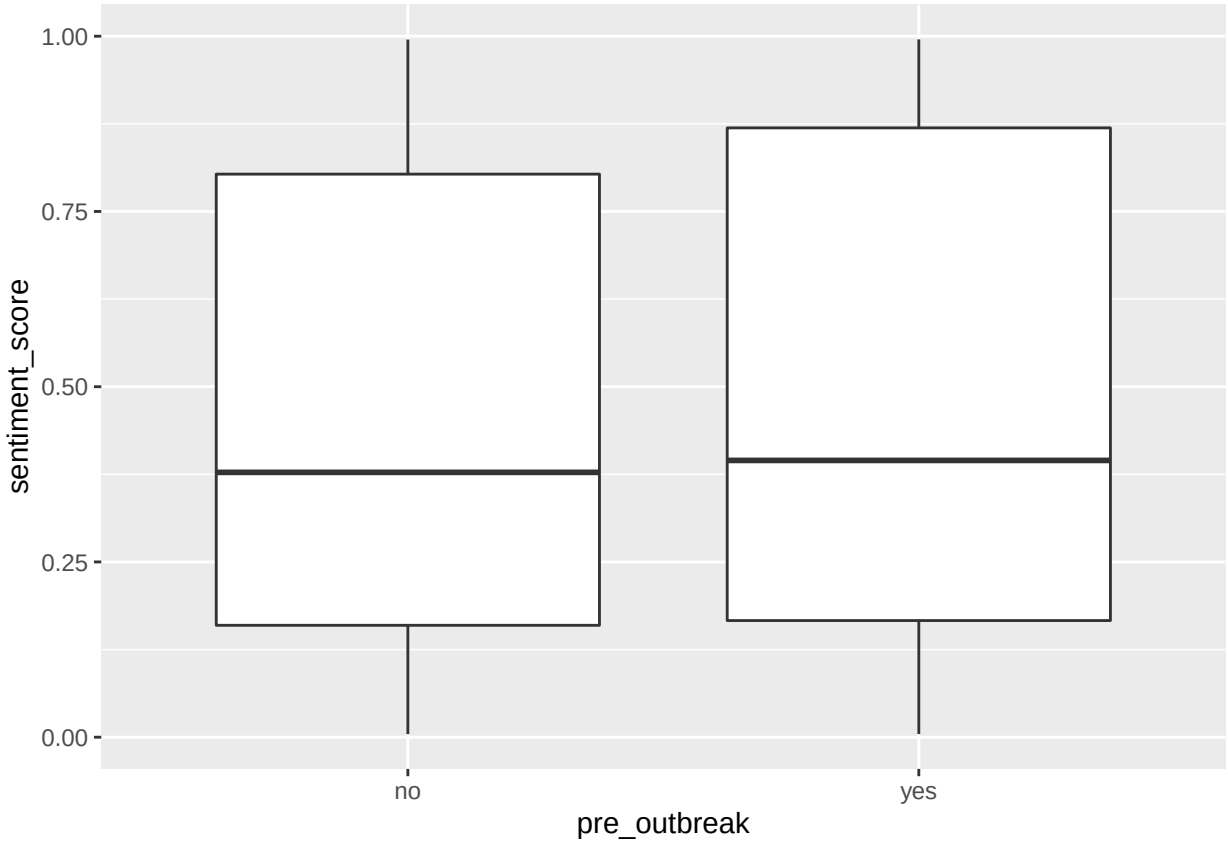
```
ggplot(India_analysis_science) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




```
India_analysis_science %>%
  group_by(pre_covid) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.475
## 2 yes                         0.476
```

```
ggplot(India_analysis_science) +
  geom_boxplot(aes(x = pre_outbreak, y = sentiment_score))
```




```
India_analysis_science %>%
  group_by(pre_outbreak) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.462
## 2 yes                            0.480
```

```
#precovid scitech
count(India_analysis_science, pre_covid == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_covid == "yes"`     n
##   <lgl>                <int>
## 1 FALSE                   70
## 2 TRUE                    60
```

```
num_precovid = 60
num_postcovid = 70
num = 130

India_analysis_science %>%
  filter(pre_covid == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
```

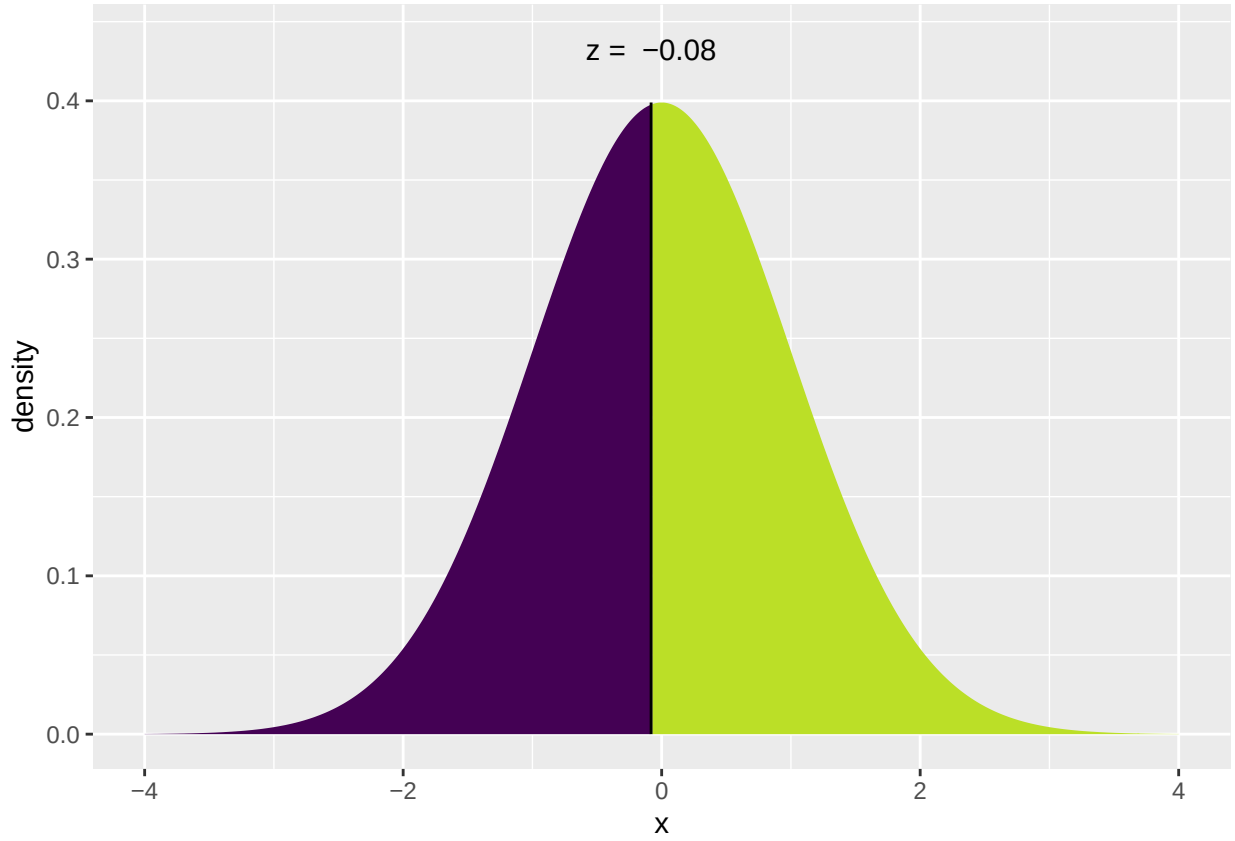
```
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      33
## 2 TRUE                       27
```

```
#proportion of positive sentiment videos precovid from sample
p_hat1 = 27/60

India_analysis_science %>%
  filter(pre_covid == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      39
## 2 TRUE                       31
```

```
#proportion of positive sentiment videos postcovid from sample
p_hat2 = 31/70

p_hat = (27+31)/(60+70)

sd <- sqrt((((p_hat)*(1-p_hat))/60)+(((p_hat)*(1-p_hat))/70))
z_score <- ((p_hat2-p_hat1)-0)/sd

#p-value
2* (xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= -0.08167) = P(Z <= -0.08167) = 0.4675
##  P(X >  -0.08167) = P(Z >  -0.08167) = 0.5325
##
```




```
## [1] 0.9349053
```

```
#outbreak scitech
count(India_analysis_science, pre_outbreak == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_outbreak == "yes"`     n
##   <lgl>                   <int>
## 1 FALSE                      30
## 2 TRUE                      100
```

```
num_preoutbreak = 100
num_postoutbreak = 30
num = 130

India_analysis_science %>%
  filter(pre_outbreak == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      53
## 2 TRUE                       47
```

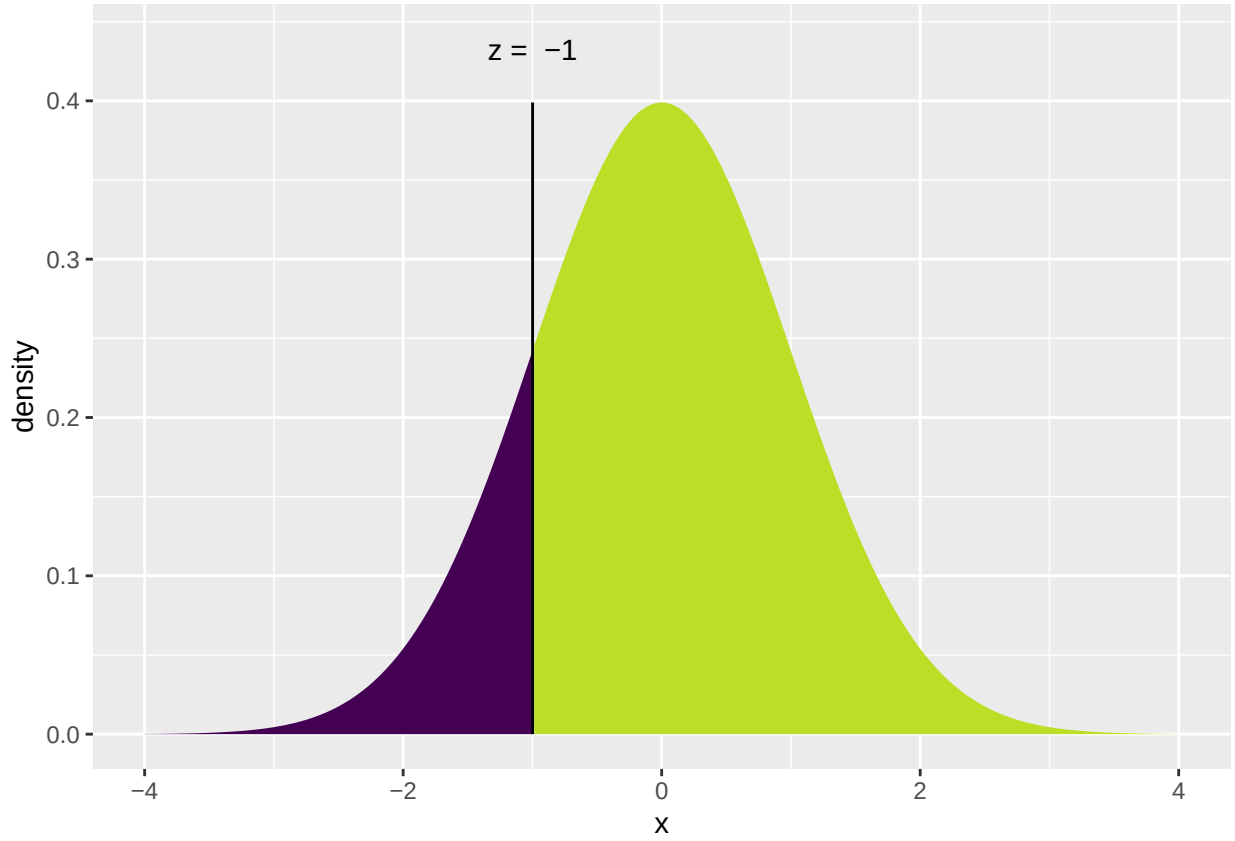
```
#proportion of positive sentiment videos preoutbreak from sample
p_hat1 = 47/100
```

```
India_analysis_science %>%
  filter(pre_outbreak == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      19
## 2 TRUE                       11
```

```
#proportion of positive sentiment videos postoutbreak from sample
p_hat2 = 11/30

p_hat = (47+11)/(100+30)

sd <- sqrt((((p_hat)*(1-p_hat))/100)+(((p_hat)*(1-p_hat))/30))
z_score <- ((p_hat2-p_hat1)-0)/sd

#p-value
2* (xpnorm(z_score, 0, 1))
```

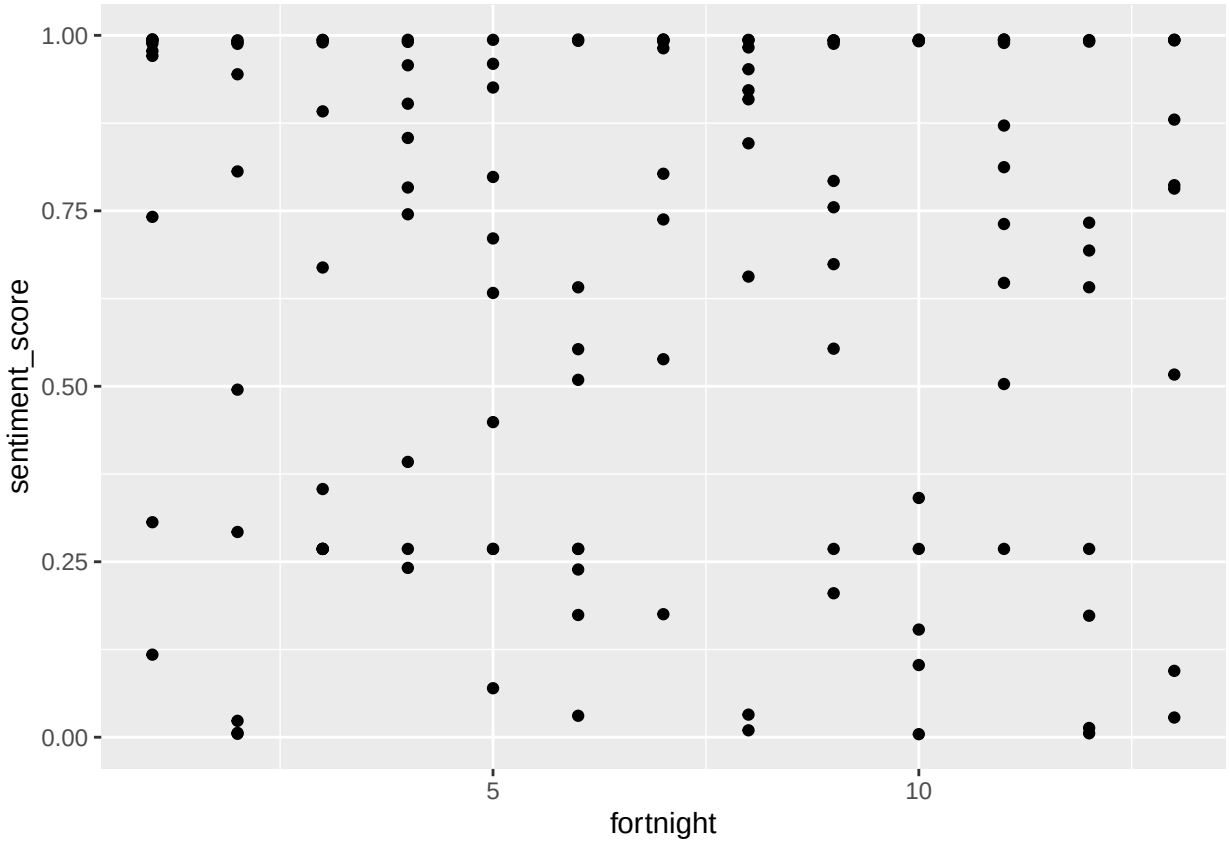
```
##
## If X ~ N(0, 1), then
##  P(X <= -0.9986) = P(Z <= -0.9986) = 0.159
##  P(X >  -0.9986) = P(Z >  -0.9986) = 0.841
##
```



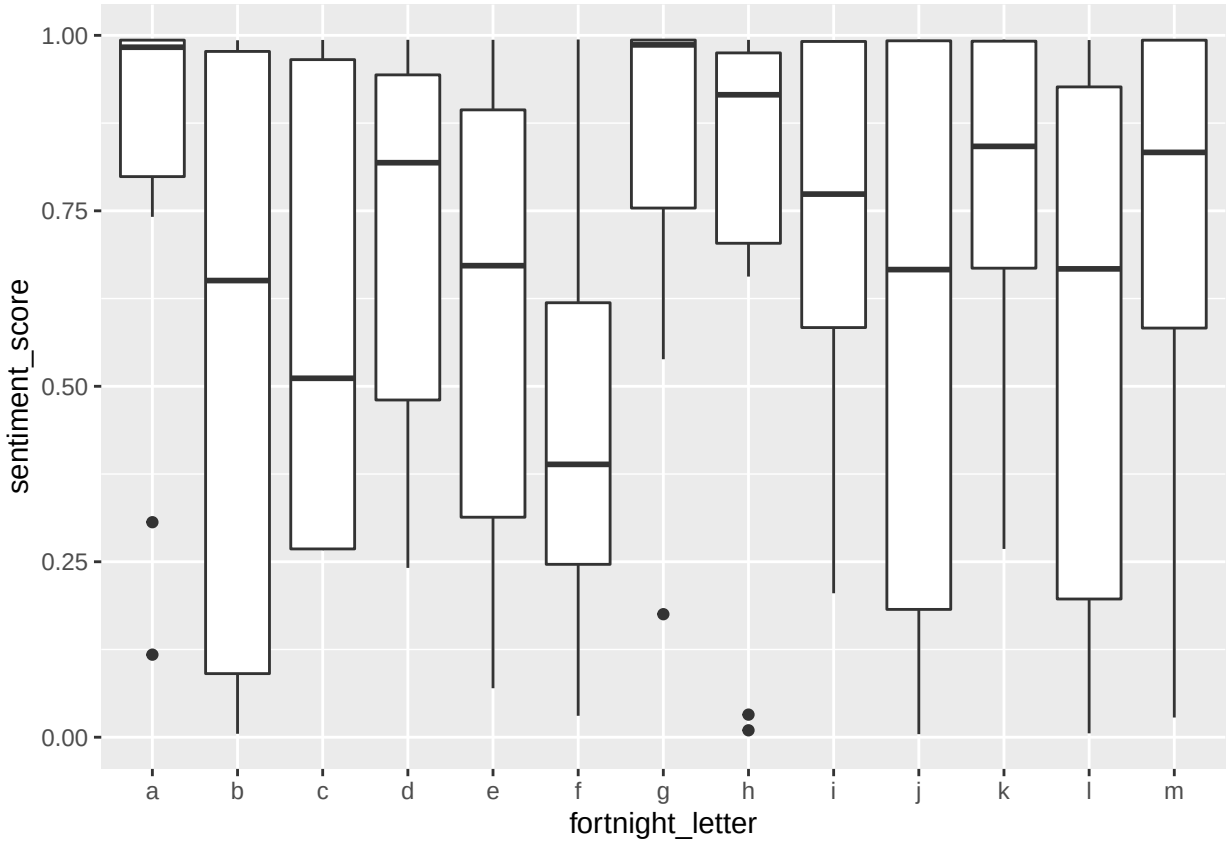

```
## [1] 0.3179875
```

```
#Youtube API All Categories
India_analysis_all <- India_analysis %>%
  filter(video_category == "All")
```

```
#data summary all categories
ggplot(India_analysis_all) +
  geom_point(aes(x = fortnight, y = sentiment_score))
```



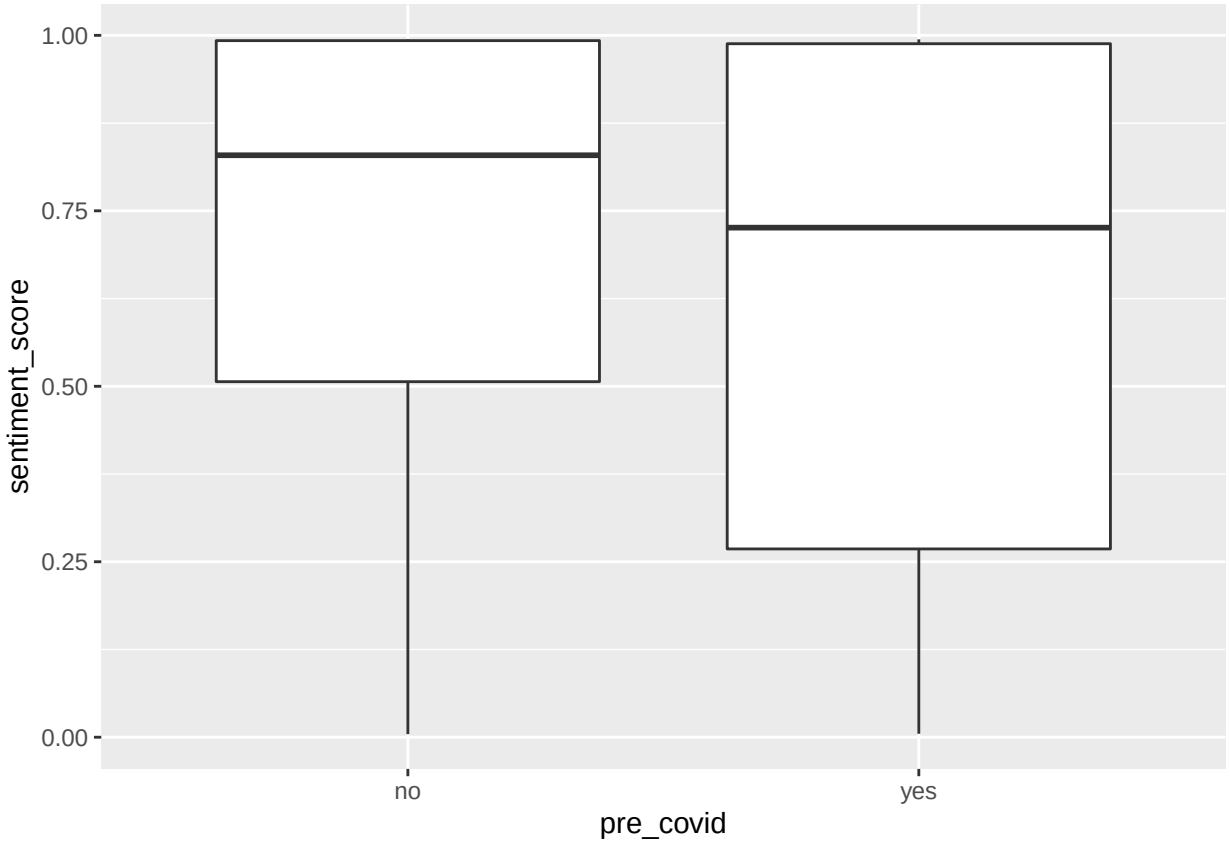

```
ggplot(India_analysis_all) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




```
India_analysis_all %>%
  group_by(fortnight) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.808
##  2         2                   0.554
##  3         3                   0.596
##  4         4                   0.713
##  5         5                   0.608
##  6         6                   0.467
##  7         7                   0.820
##  8         8                   0.730
##  9         9                   0.721
## 10        10                   0.583
## 11        11                   0.780
## 12        12                   0.550
## 13        13                   0.706
```

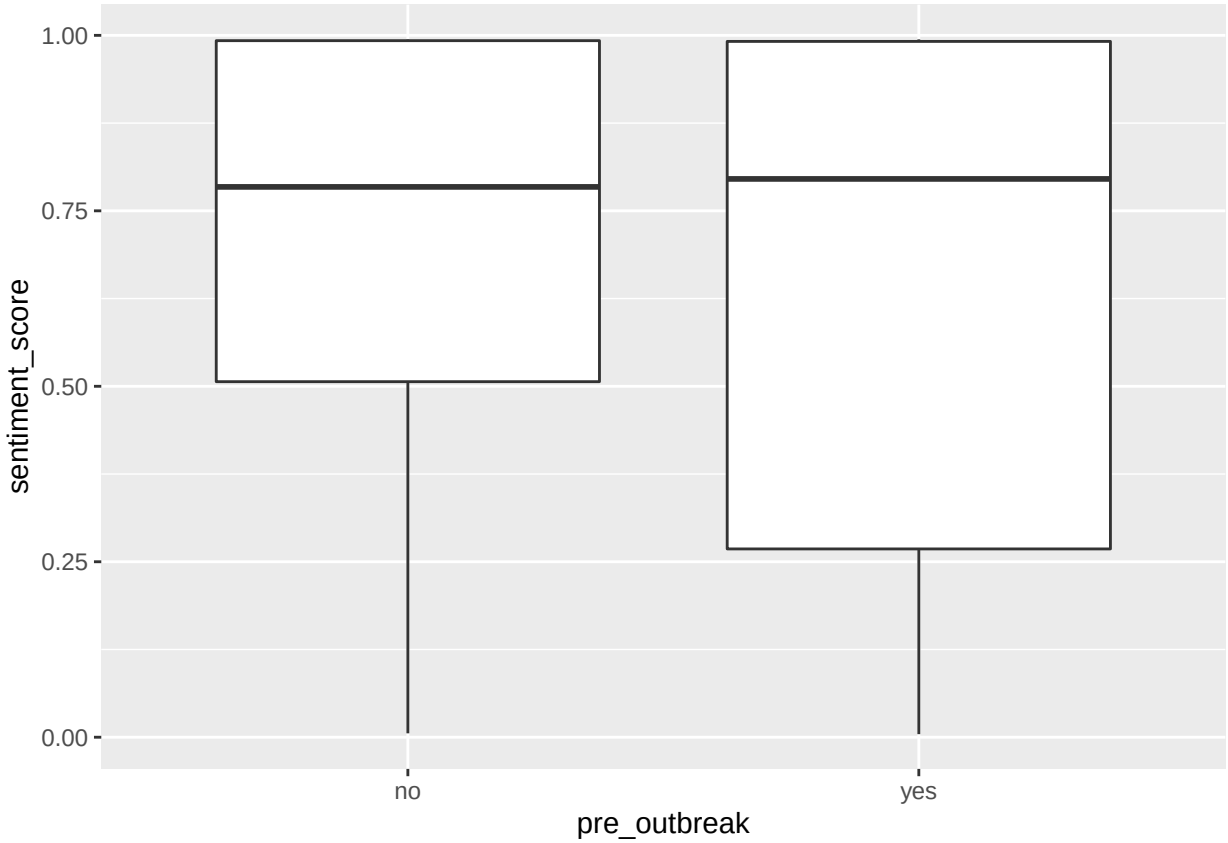
```
ggplot(India_analysis_all) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




```
India_analysis_all %>%
  group_by(pre_covid) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.699
## 2 yes                         0.624
```

```
ggplot(India_analysis_all) +
  geom_boxplot(aes(x = pre_outbreak, y = sentiment_score))
```




```
India_analysis_all %>%
  group_by(pre_outbreak) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.679
## 2 yes                            0.660
```

```
#precovid all categories
count(India_analysis_all, pre_covid == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_covid == "yes"`     n
##   <lgl>                <int>
## 1 FALSE                   70
## 2 TRUE                    60
```

```
num_precovid = 60
num_postcovid = 70
num = 130

India_analysis_all %>%
  filter(pre_covid == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
```

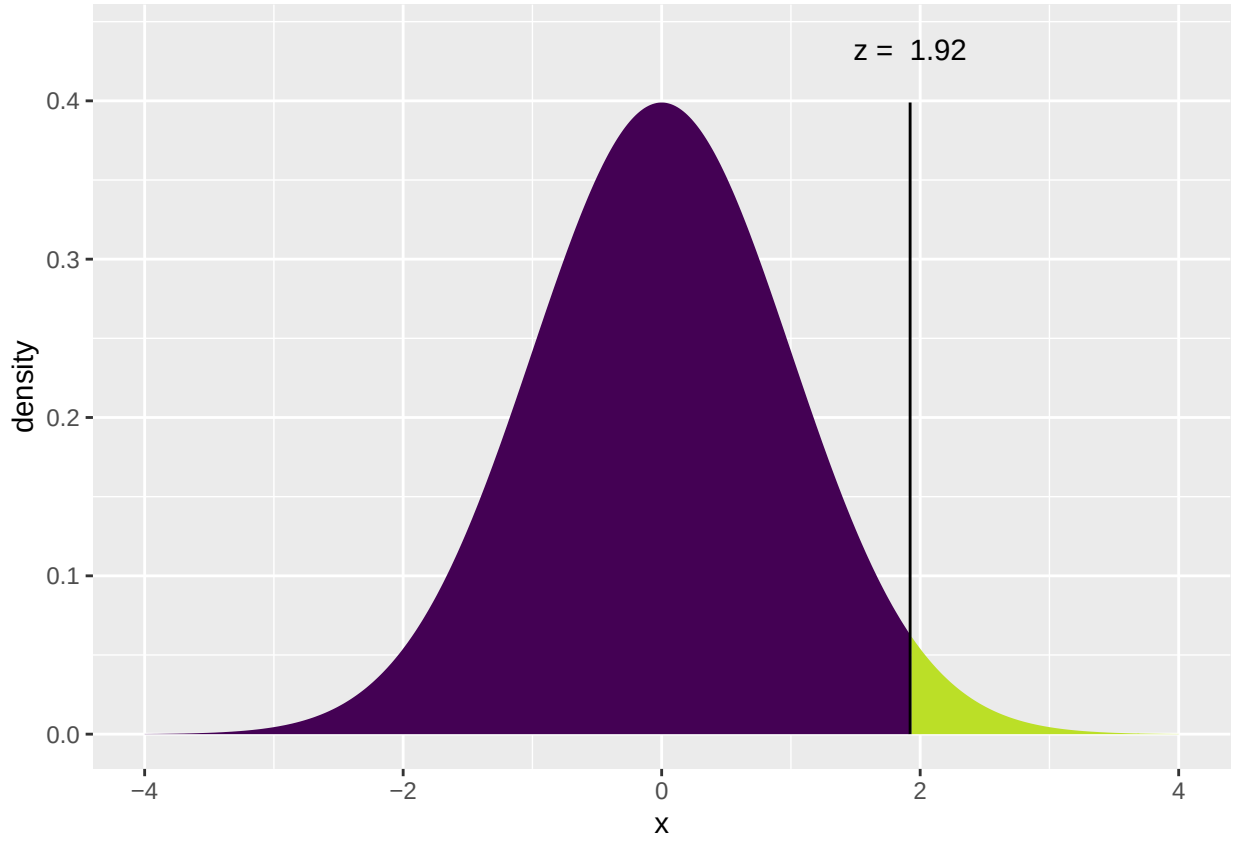
```
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      24
## 2 TRUE                       36
```

```
#proportion of positive sentiment videos precovid from sample
p_hat1 = 36/60

India_analysis_all %>%
  filter(pre_covid == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      17
## 2 TRUE                       53
```

```
#proportion of positive sentiment videos postcovid from sample
p_hat2 = 53/70

p_hat = (36+53)/(60+70)

sd <- sqrt((((p_hat)*(1-p_hat))/60)+(((p_hat)*(1-p_hat))/70))
z_score <- ((p_hat2-p_hat1)-0)/sd

#p-value
2* (1-xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= 1.922) = P(Z <= 1.922) = 0.9727
##  P(X >  1.922) = P(Z >  1.922) = 0.02729
##
```




```
## [1] 0.05457756
```

```
#outbreak all categories
count(India_analysis_all, pre_outbreak == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_outbreak == "yes"`     n
##   <lgl>                   <int>
## 1 FALSE                      30
## 2 TRUE                      100
```

```
num_preoutbreak = 100
num_postoutbreak = 30
num = 130

India_analysis_all %>%
  filter(pre_outbreak == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      34
## 2 TRUE                       66
```

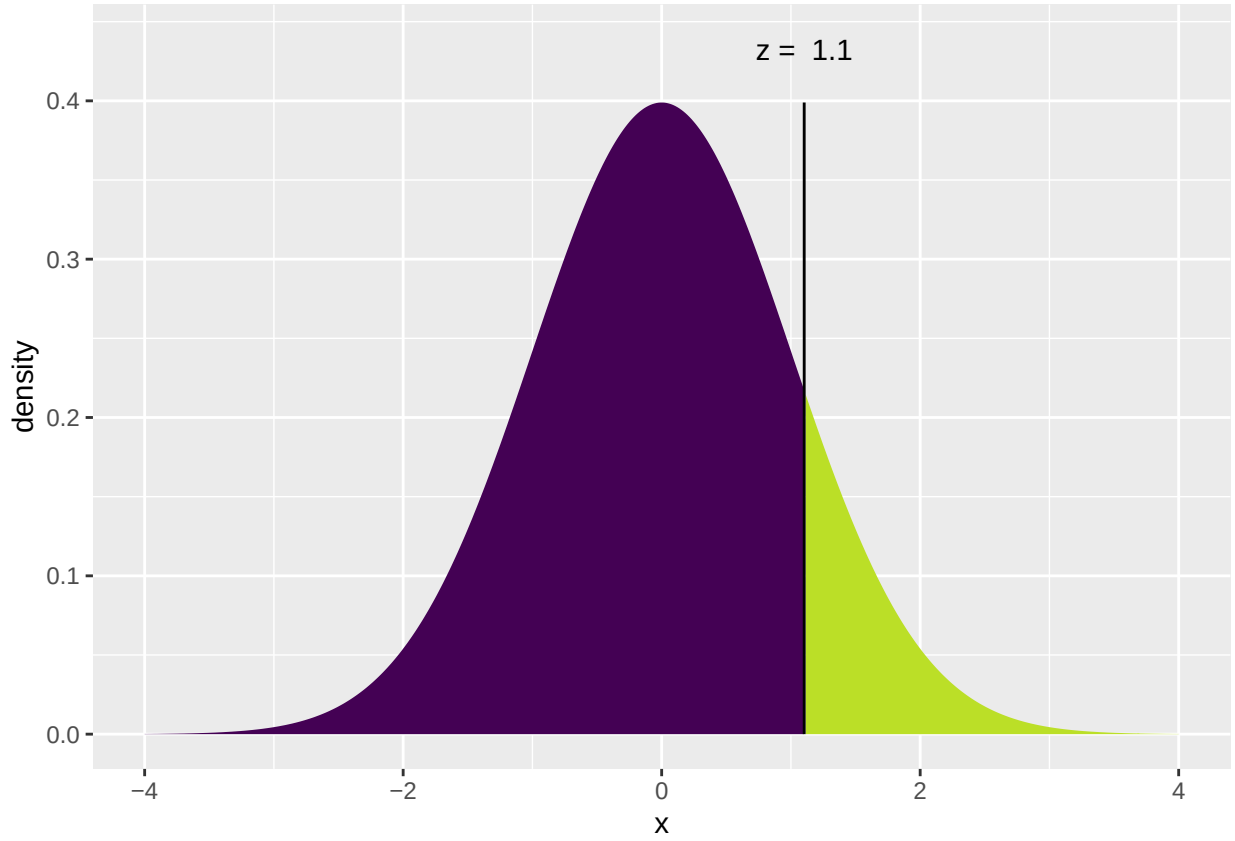
```
#proportion of positive sentiment videos preoutbreak from sample
p_hat1 = 66/100
```

```r
India_analysis_all %>%
  filter(pre_outbreak == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                       7
## 2 TRUE                       23
```

```r
#proportion of positive sentiment videos postoutbreak from sample
p_hat2 = 23/30

p_hat = (66+23)/(100+30)

sd <- sqrt((((p_hat)*(1-p_hat))/100)+(((p_hat)*(1-p_hat))/30))
z_score <- ((p_hat2-p_hat1)-0)/sd

#p-value
2* (1-xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= 1.103) = P(Z <= 1.103) = 0.8649
##  P(X >  1.103) = P(Z >  1.103) = 0.1351
##
```




```
## [1] 0.270139
```

```
#Two independent samples t-tests; Comparing two independent means

#pre_covid music
t.test(sentiment_score ~ pre_covid, alternative = "two.sided", data = India_analysis_music)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = 1.0409, df = 125.64, p-value = 0.2999
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.0520926  0.1676980
## sample estimates:
##  mean in group no mean in group yes
##         0.5039303         0.4461277
```

```
#pre_outbreak music
t.test(sentiment_score ~ pre_outbreak, alternative = "two.sided", data = India_analysis_music)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = 1.4129, df = 47.938, p-value = 0.1641
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##  -0.03923456  0.22469084
## sample estimates:
##  mean in group no mean in group yes
##         0.5486569         0.4559288
```

```
#pre_covid travel and events
t.test(sentiment_score ~ pre_covid, alternative = "two.sided", data = India_analysis_travel)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = 0.84984, df = 118.47, p-value = 0.3971
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.06666182  0.16690022
## sample estimates:
##  mean in group no mean in group yes
##         0.6522514         0.6021322
```

```
#pre_outbreak travel and events
t.test(sentiment_score ~ pre_outbreak, alternative = "two.sided", data = India_analysis_travel)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = -0.092431, df = 50.999, p-value = 0.9267
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1397963  0.1274902
## sample estimates:
##  mean in group no mean in group yes
##         0.6242180         0.6303711
```

```
#pre_covid people and blogs
t.test(sentiment_score ~ pre_covid, alternative = "two.sided", data = India_analysis_people)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = 2.7397, df = 123.76, p-value = 0.00706
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.04099899 0.25445225
## sample estimates:
##  mean in group no mean in group yes
##         0.7216792         0.5739535
```

```
#pre_outbreak people and blogs
t.test(sentiment_score ~ pre_outbreak, alternative = "two.sided", data = India_analysis_people)
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  sentiment_score by pre_outbreak
## t = 1.484, df = 51.983, p-value = 0.1438
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.03222438  0.21521685
## sample estimates:
##  mean in group no mean in group yes
##         0.7238798         0.6323836
```

```
#pre_covid entertainment
t.test(sentiment_score ~ pre_covid, alternative = "two.sided", data = India_analysis_entertainment)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = 1.4894, df = 125.47, p-value = 0.1389
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.02808331  0.19892515
## sample estimates:
##  mean in group no mean in group yes
##         0.5537231         0.4683022
```

```
#pre_outbreak entertainment
t.test(sentiment_score ~ pre_outbreak, alternative = "two.sided", data = India_analysis_entertainment)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = -0.12774, df = 45.354, p-value = 0.8989
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1537038  0.1353658
## sample estimates:
##  mean in group no mean in group yes
##          0.506662          0.515831
```

```
#pre_covid news and politics
t.test(sentiment_score ~ pre_covid, alternative = "two.sided", data = India_analysis_news)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = 4.0934, df = 110.38, p-value = 8.119e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1039827 0.2991390
## sample estimates:
##  mean in group no mean in group yes
##         0.5440929         0.3425321
```

```
#pre_outbreak news and politics
t.test(sentiment_score ~ pre_outbreak, alternative = "two.sided", data = India_analysis_news)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = 2.8921, df = 43.487, p-value = 0.005954
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.0577894 0.3237589
## sample estimates:
##  mean in group no mean in group yes
##         0.5967517         0.4059776
```

```
#pre_covid how-to and style
t.test(sentiment_score ~ pre_covid, alternative = "two.sided", data = India_analysis_how_to)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = 0.2463, df = 119.91, p-value = 0.8059
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1062167  0.1363966
## sample estimates:
##  mean in group no mean in group yes
##         0.6256418         0.6105519
```

```
#pre_outbreak how-to and style
t.test(sentiment_score ~ pre_outbreak, alternative = "two.sided", data = India_analysis_how_to)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = 0.99737, df = 55.567, p-value = 0.3229
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.06590638  0.19656067
## sample estimates:
##  mean in group no mean in group yes
##         0.6689289         0.6036017
```

```
#pre_covid education
t.test(sentiment_score ~ pre_covid, alternative = "two.sided", data = India_analysis_education)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = 1.8198, df = 123.45, p-value = 0.07121
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.007984541  0.190110330
## sample estimates:
##  mean in group no mean in group yes
##         0.5130397         0.4219768
```

```r
#pre_outbreak education
t.test(sentiment_score ~ pre_outbreak, alternative = "two.sided", data = India_analysis_education)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = 0.37547, df = 46.77, p-value = 0.709
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.09918976  0.14470380
## sample estimates:
##  mean in group no mean in group yes
##         0.4885161         0.4657591
```

```r
#pre_covid science and technology
t.test(sentiment_score ~ pre_covid, alternative = "two.sided", data = India_analysis_science)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = -0.01415, df = 126.8, p-value = 0.9887
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1213397  0.1196167
## sample estimates:
##  mean in group no mean in group yes
##         0.4754371         0.4762986
```

```r
#pre_outbreak science and technology
t.test(sentiment_score ~ pre_outbreak, alternative = "two.sided", data = India_analysis_science)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = -0.23755, df = 46.334, p-value = 0.8133
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1661403  0.1310600
## sample estimates:
##  mean in group no mean in group yes
##         0.4623423         0.4798824
```

```r
#pre_covid all categories
t.test(sentiment_score ~ pre_covid, alternative = "two.sided", data = India_analysis_all)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = 1.2103, df = 124.81, p-value = 0.2285
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.0472999  0.1962048
```

```
## sample estimates:
##  mean in group no mean in group yes
##         0.6987750         0.6243226
```

```
#pre_outbreak categories
t.test(sentiment_score ~ pre_outbreak, alternative = "two.sided", data = India_analysis_all)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = 0.26003, df = 48.192, p-value = 0.796
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1272904  0.1651091
## sample estimates:
##  mean in group no mean in group yes
##         0.6789580         0.6600486
```