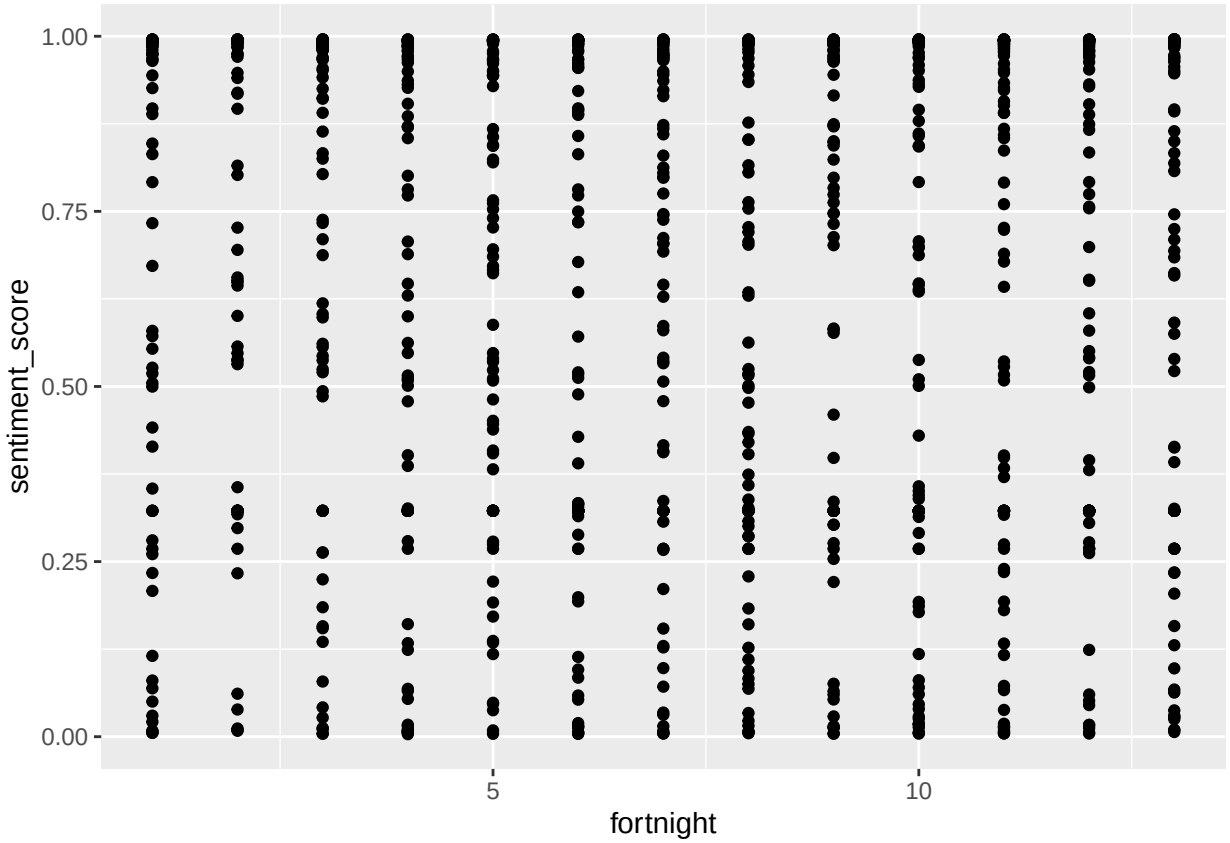
Datafest Data Analysis

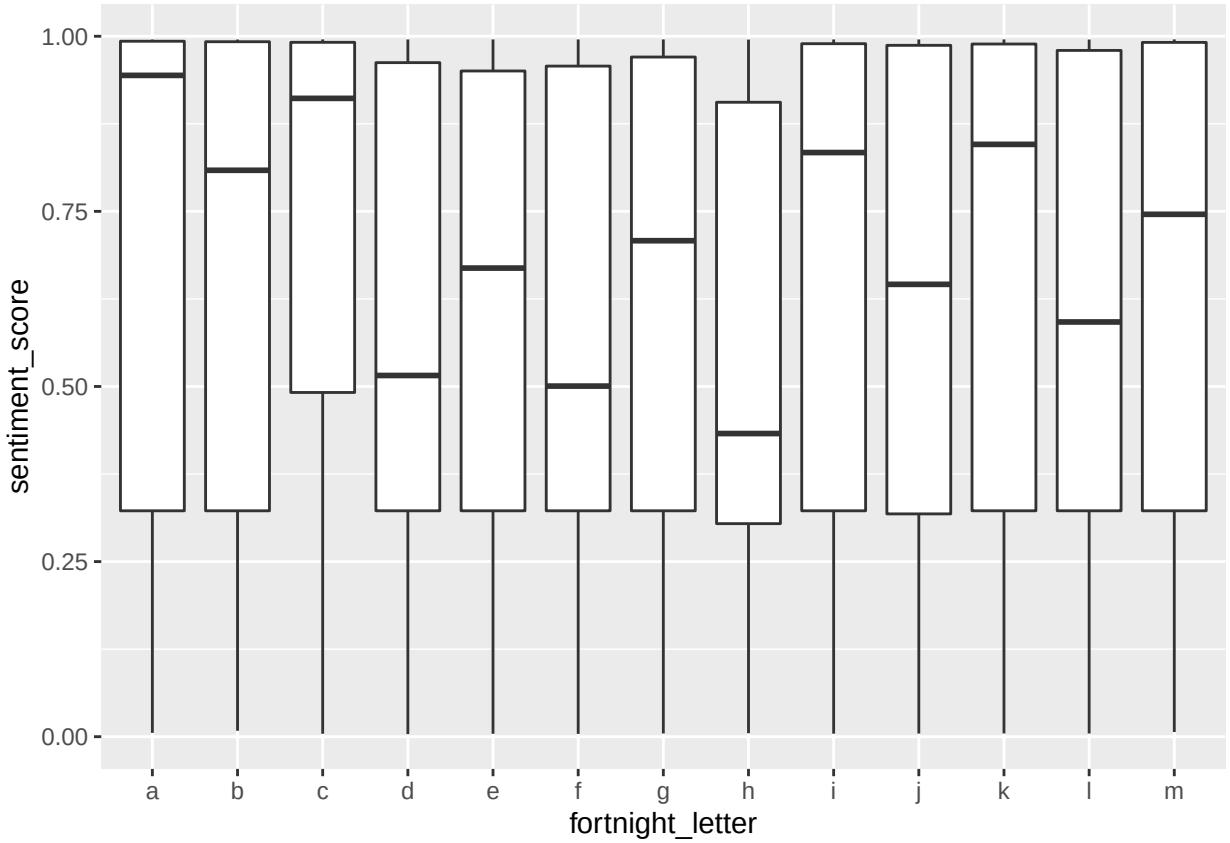
```
#Australia analysis import
Australia_analysis = read_excel("C:\\Users\\gtham\\OneDrive - Pomona College\\A - DATAFEST\\Analysis Da
```

```
Australia_analysis_music <- Australia_analysis %>%
  filter(video_category == "Music")

Australia_analysis_travel <- Australia_analysis %>%
  filter(video_category == "Travel and Events")

Australia_analysis_people <- Australia_analysis %>%
  filter(video_category == "People and Blogs")

Australia_analysis_entertainment <- Australia_analysis %>%
  filter(video_category == "Entertainment")

Australia_analysis_news <- Australia_analysis %>%
  filter(video_category == "News and Politics")

Australia_analysis_how_to <- Australia_analysis %>%
  filter(video_category == "How-to and Style")

Australia_analysis_education <- Australia_analysis %>%
  filter(video_category == "Education")

Australia_analysis_science <- Australia_analysis %>%
  filter(video_category == "Science and Technology")
```

```
#fullAustralia data data summaries
ggplot(Australia_analysis) +
  geom_point(aes(x = fortnight, y = sentiment_score))
```



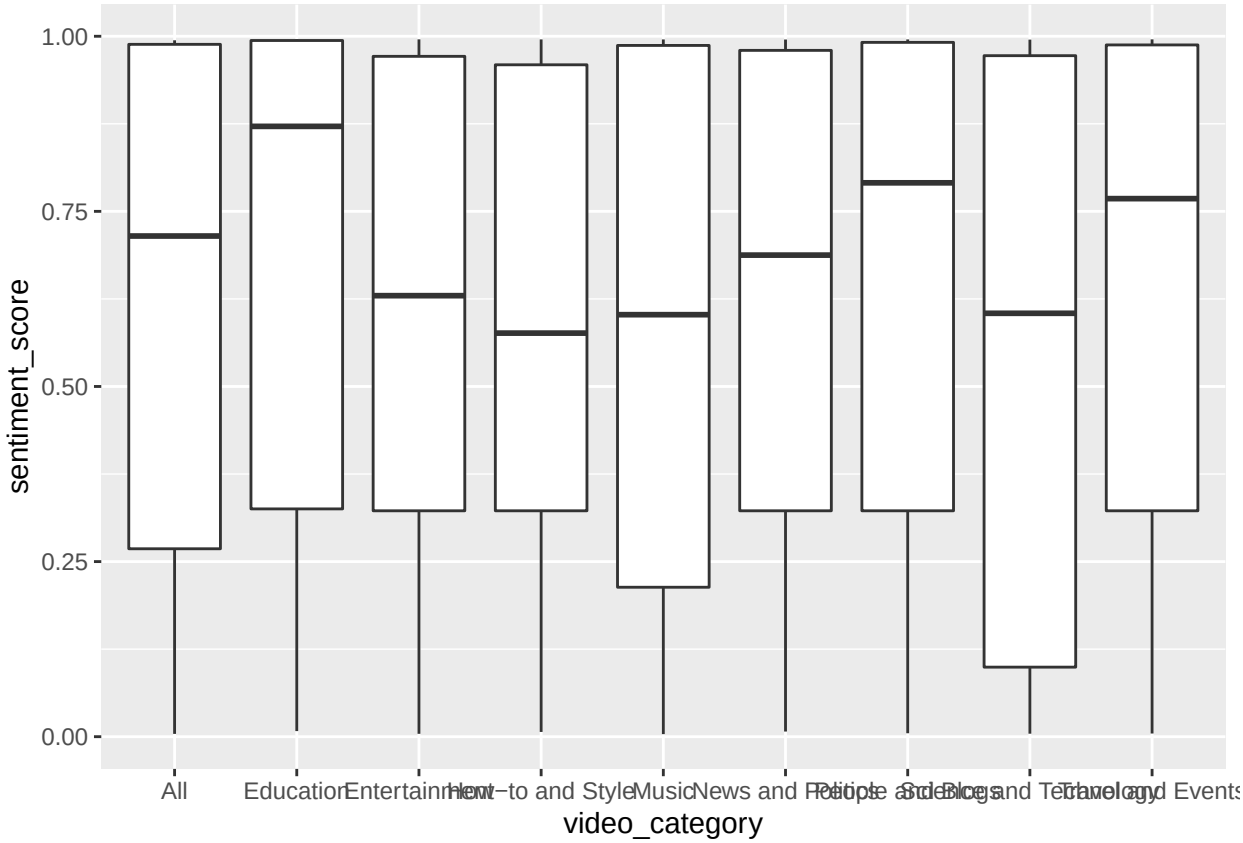

```
ggplot(Australia_analysis) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




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

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.689
##  2         2                   0.683
##  3         3                   0.706
##  4         4                   0.563
##  5         5                   0.618
##  6         6                   0.559
##  7         7                   0.624
##  8         8                   0.528
##  9         9                   0.669
## 10        10                   0.586
## 11        11                   0.642
## 12        12                   0.609
## 13        13                   0.638
```

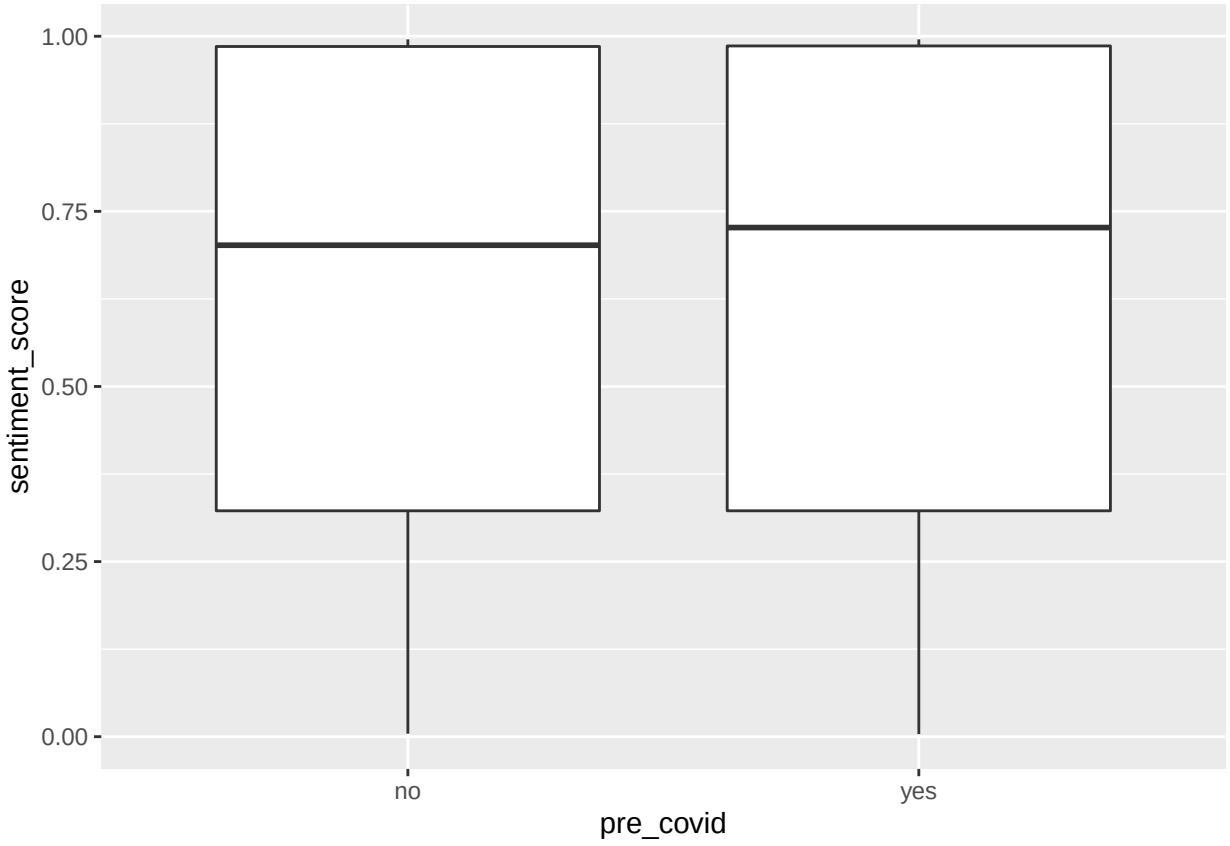
```
ggplot(Australia_analysis) +
  geom_boxplot(aes(x = video_category, y = sentiment_score))
```




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

```
## # A tibble: 9 x 2
##   video_category         `mean(sentiment_score)`
##   <chr>                                    <dbl>
## 1 All                                      0.637
## 2 Education                                0.695
## 3 Entertainment                            0.589
## 4 How-to and Style                         0.599
## 5 Music                                    0.562
## 6 News and Politics                        0.606
## 7 People and Blogs                         0.667
## 8 Science and Technology                   0.547
## 9 Travel and Events                        0.670
```

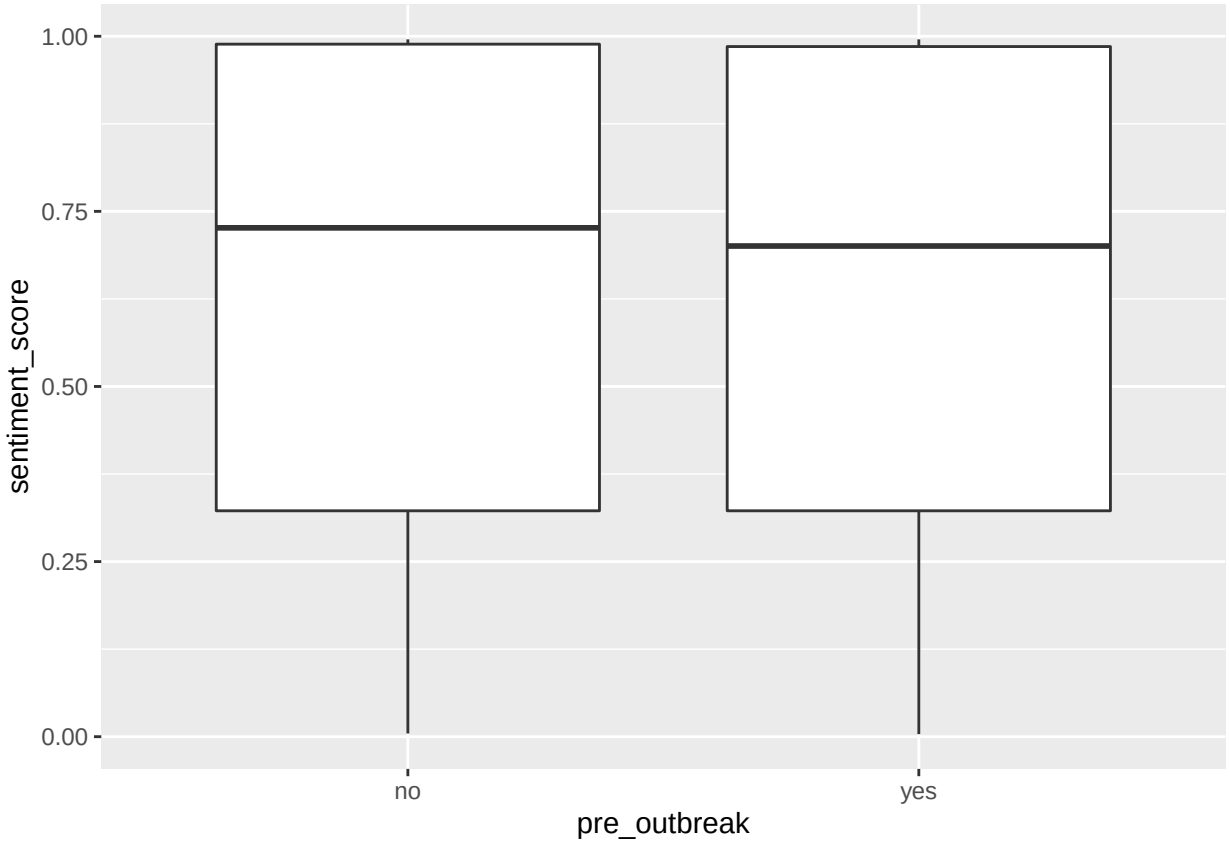
```
ggplot(Australia_analysis) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




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

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.614
## 2 yes                         0.634
```

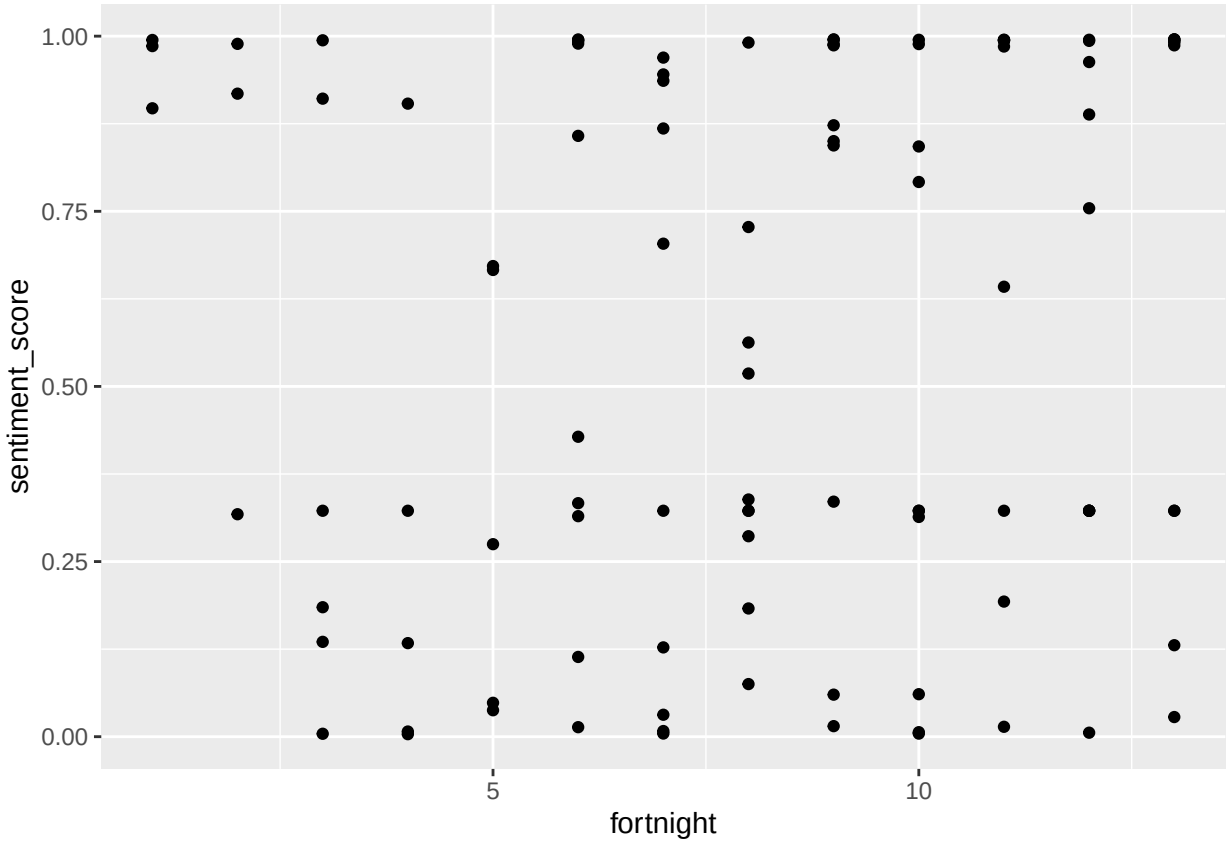
```
ggplot(Australia_analysis) +
  geom_boxplot(aes(x = pre_outbreak, y = sentiment_score))
```



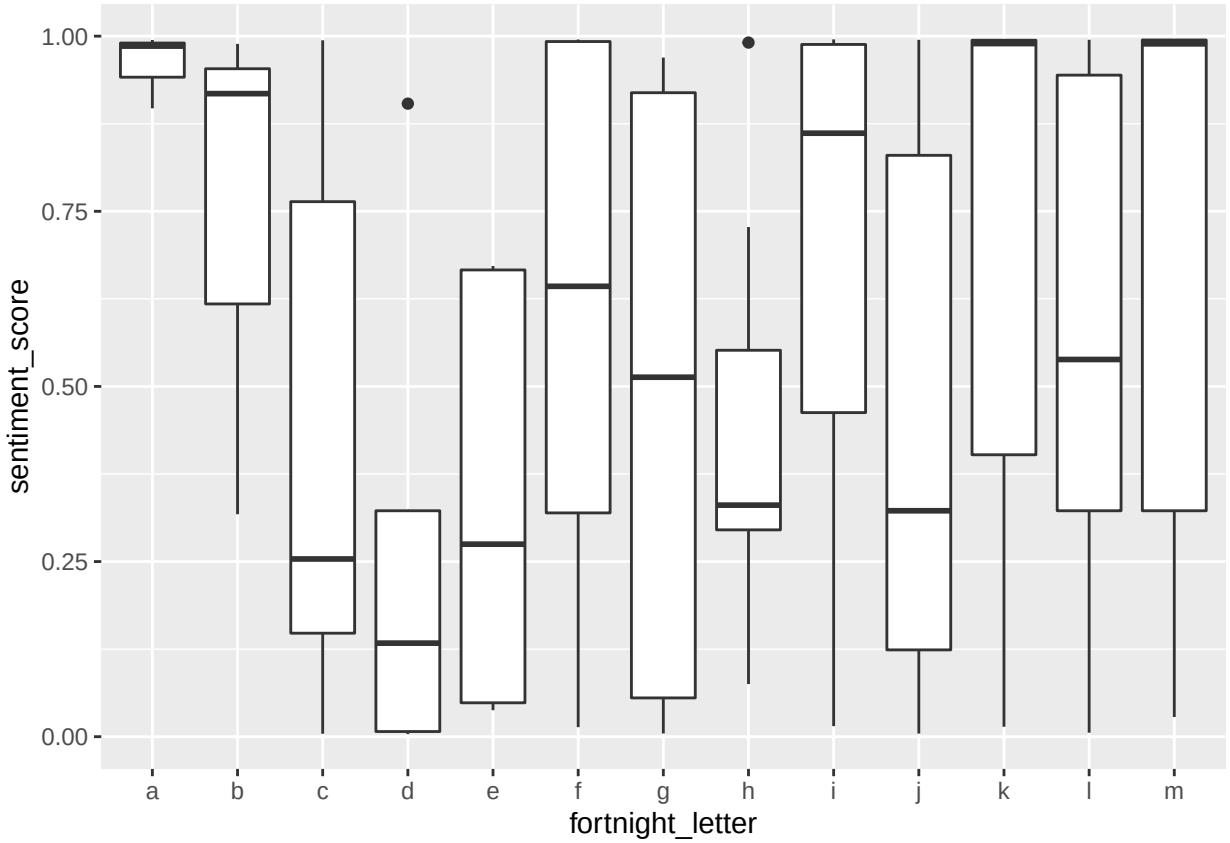

```
Australia_analysis %>%
  group_by(pre_outbreak) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.630
## 2 yes                            0.620
```

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



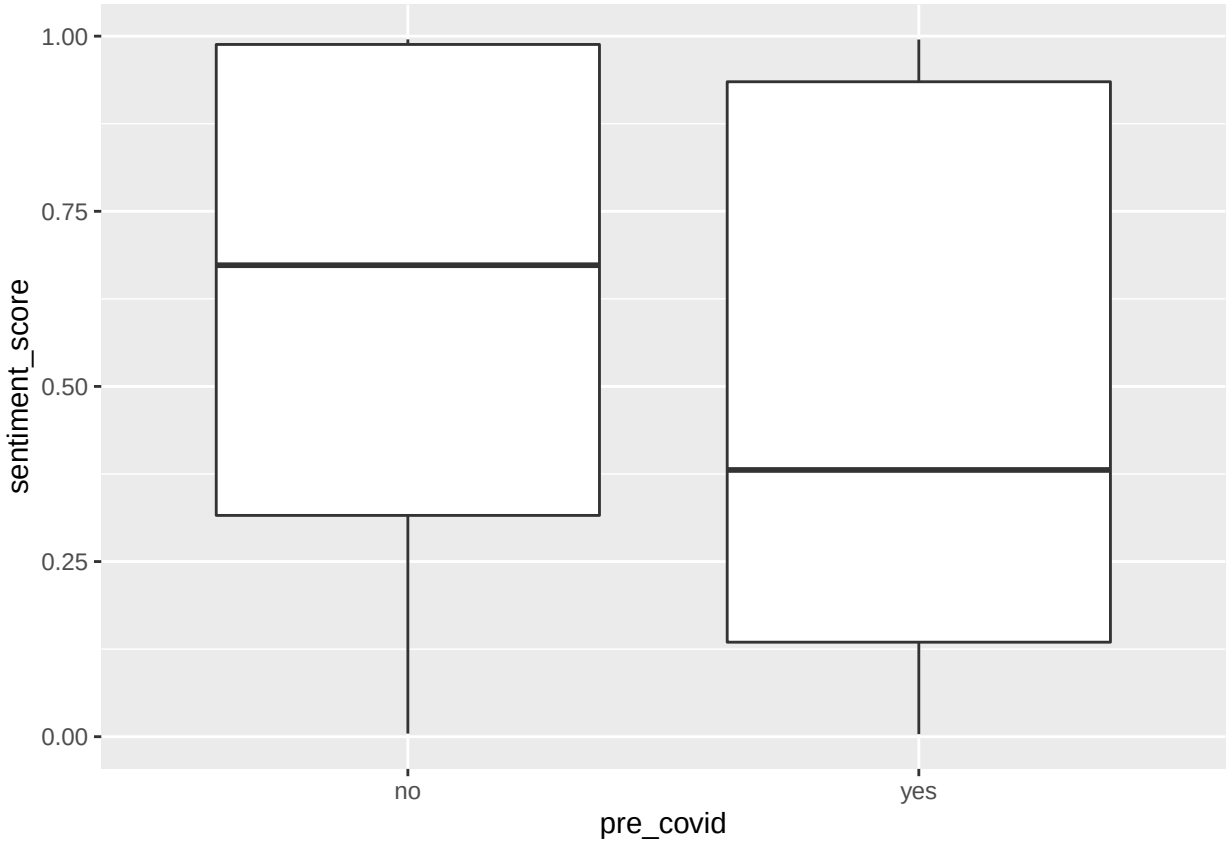

```
ggplot(Australia_analysis_music) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




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

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.959
##  2         2                   0.741
##  3         3                   0.425
##  4         4                   0.274
##  5         5                   0.340
##  6         6                   0.603
##  7         7                   0.492
##  8         8                   0.433
##  9         9                   0.694
## 10        10                   0.465
## 11        11                   0.713
## 12        12                   0.589
## 13        13                   0.676
```

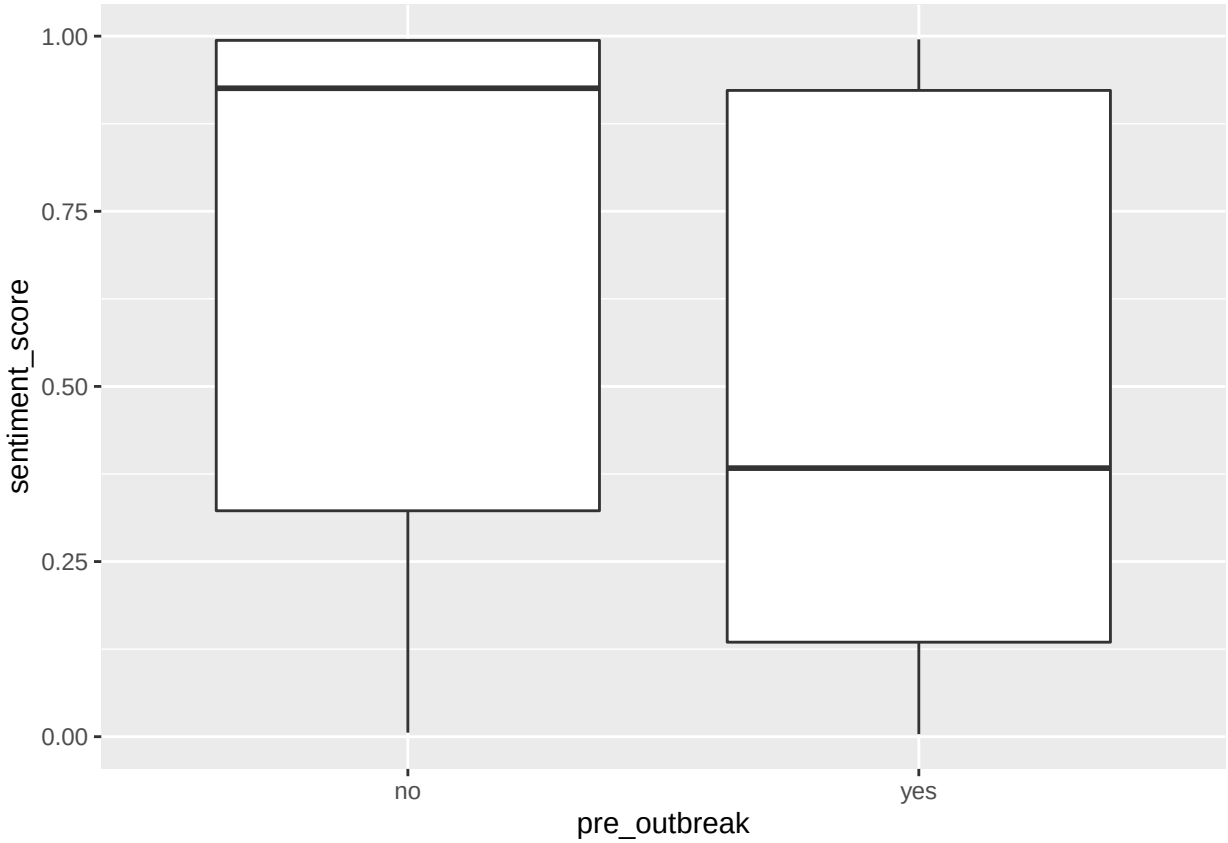
```
ggplot(Australia_analysis_music) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




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

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.580
## 2 yes                         0.524
```

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




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

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.659
## 2 yes                            0.522
```

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

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

```
## # A tibble: 2 x 2
##   `pre_covid == "yes"`     n
##   <lgl>                <int>
## 1 FALSE                   70
## 2 TRUE                    32
```

```
m_num_precovid = 32
m_num_postcovid = 70
m_num = 102

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

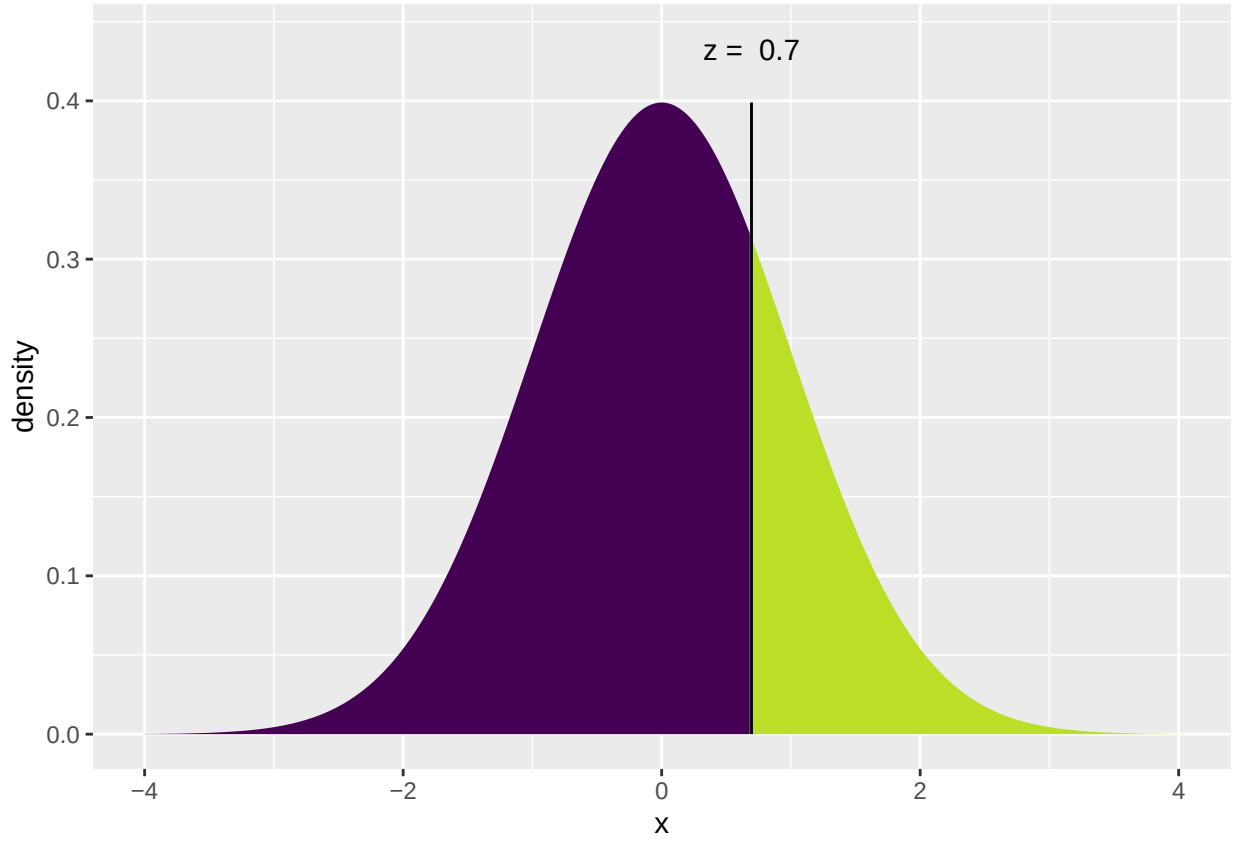
```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      17
## 2 TRUE                       15
```

```
p_hat_1_m_pos = 15/32

Australia_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                      32
## 2 TRUE                       38
```

```
p_hat_2_m_pos = 38/70

p_hat_m_pos = (15+38)/(32+70)

sd <- sqrt((((p_hat_m_pos)*(1-p_hat_m_pos))/32)+(((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 <= 0.6951) = P(Z <= 0.6951) = 0.7565
##  P(X >  0.6951) = P(Z >  0.6951) = 0.2435
##
```




```
## [1] 0.4869917
```

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

```
## # A tibble: 2 x 2
##   `pre_outbreak == "yes"`     n
##   <lgl>                   <int>
## 1 FALSE                      30
## 2 TRUE                       72
```

```
m_num_preoutbreak = 72
m_num_postoutbreak = 30
m_num = 102

Australia_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                      37
## 2 TRUE                       35
```

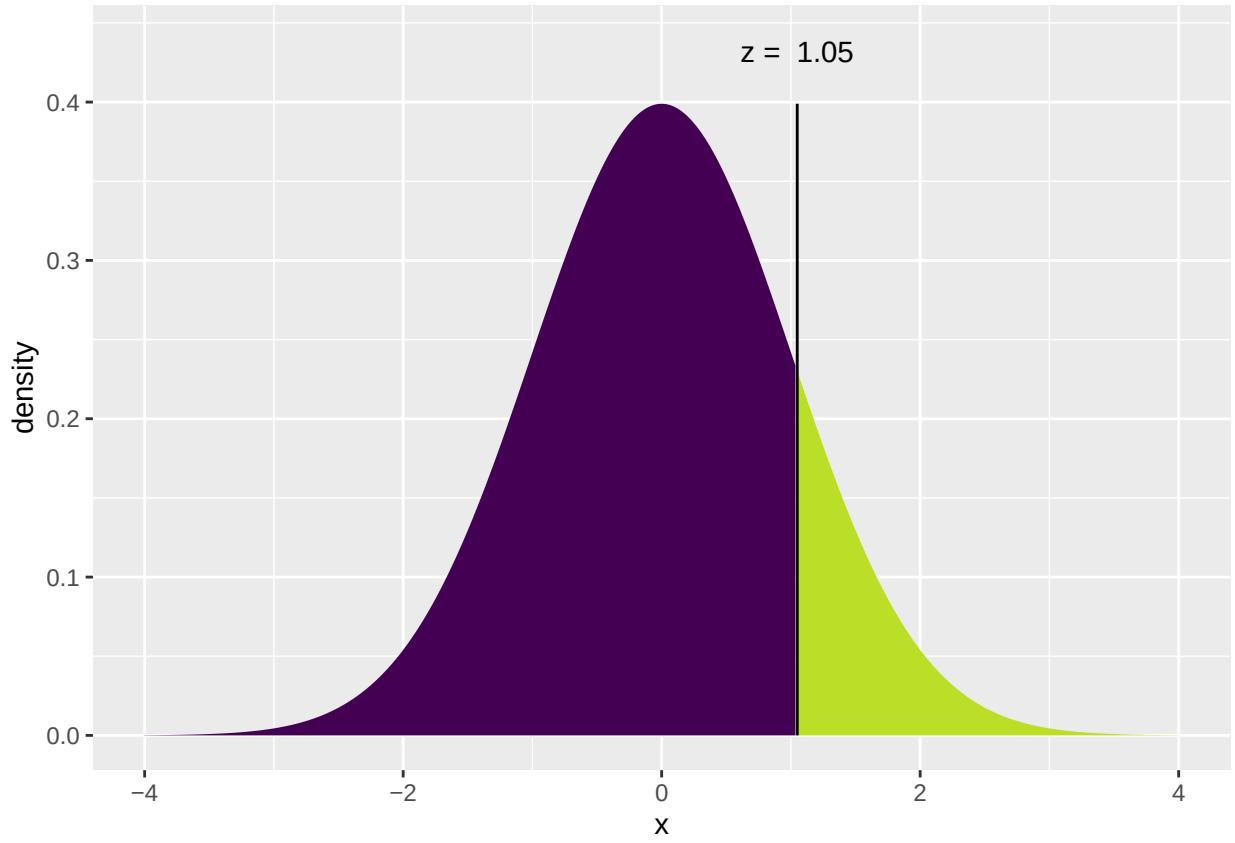
```
p_hat_1_m_pos = 35/72

Australia_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                      12
## 2 TRUE                       18
```

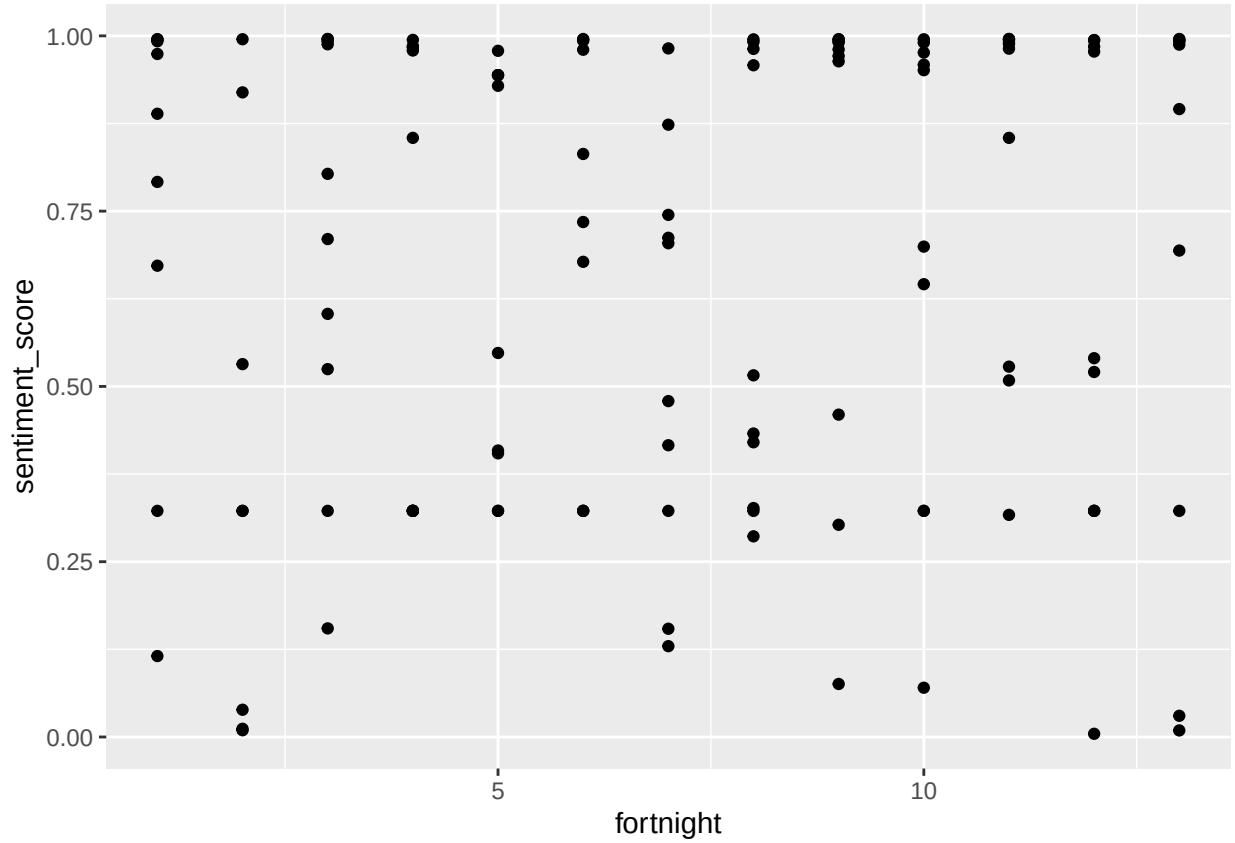
```
p_hat_2_m_pos = 18/30

p_hat_m_pos = (35+18)/(72+30)

sd <- sqrt((((p_hat_m_pos)*(1-p_hat_m_pos))/72)+(((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 <= 1.049) = P(Z <= 1.049) = 0.8529
##  P(X >  1.049) = P(Z >  1.049) = 0.1471
##
```

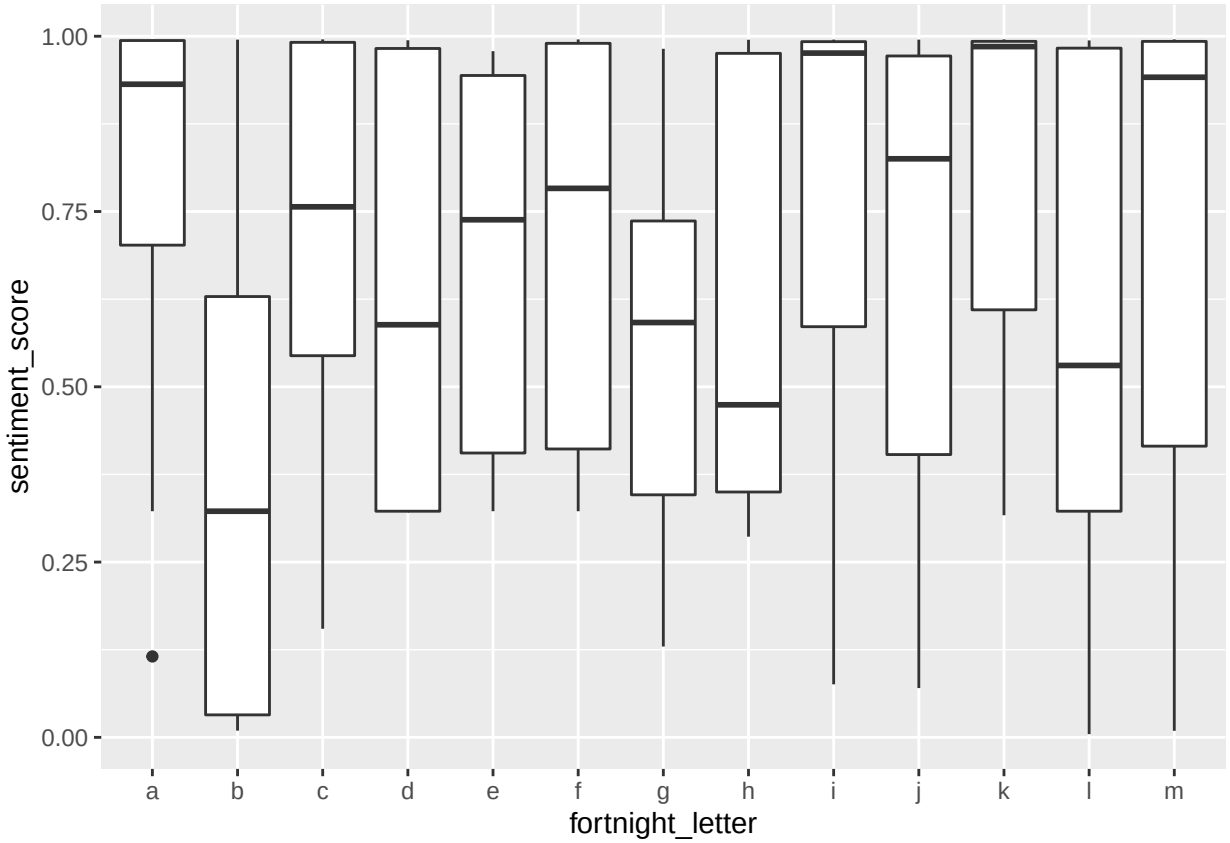


```
## [1] 0.2941816
```

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



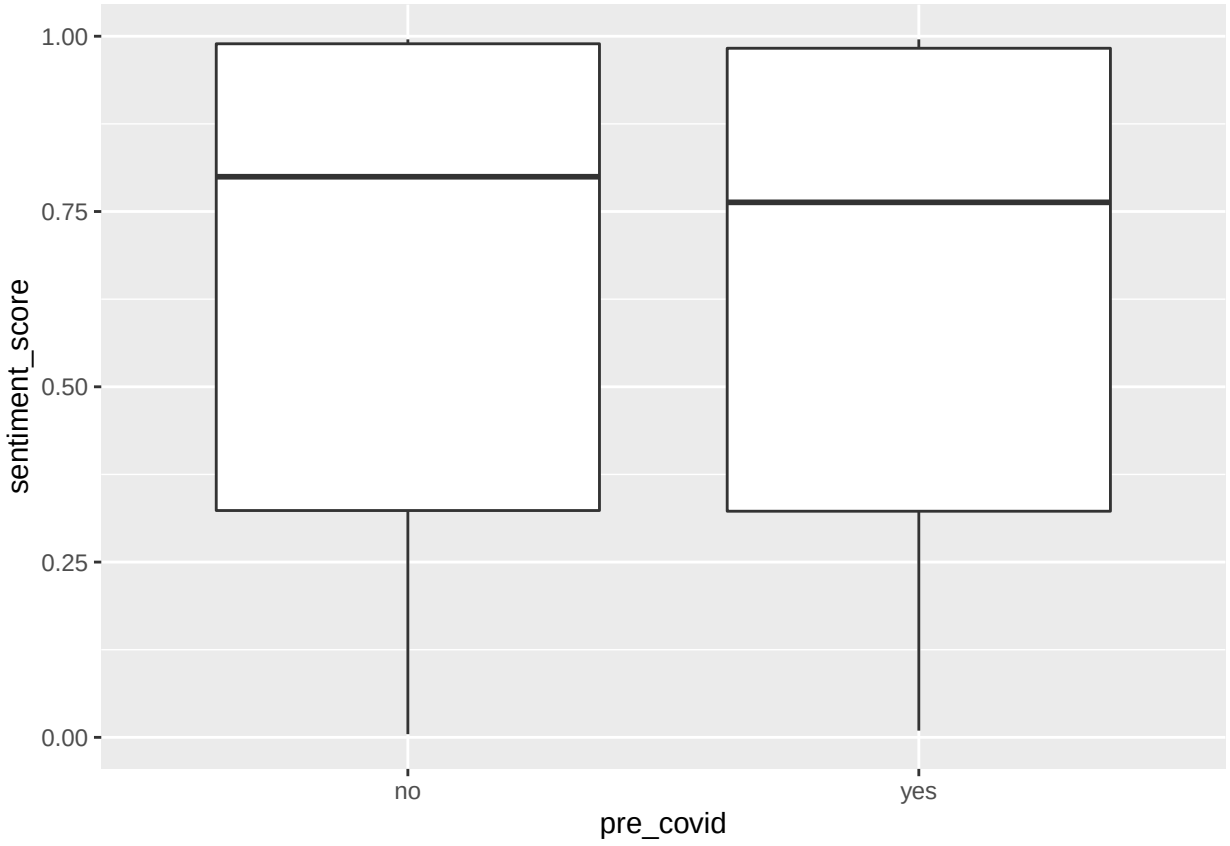

```
ggplot(Australia_analysis_travel) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




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

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.774
##  2         2                   0.394
##  3         3                   0.709
##  4         4                   0.641
##  5         5                   0.675
##  6         6                   0.717
##  7         7                   0.552
##  8         8                   0.623
##  9         9                   0.773
## 10        10                   0.693
## 11        11                   0.815
## 12        12                   0.598
## 13        13                   0.691
```

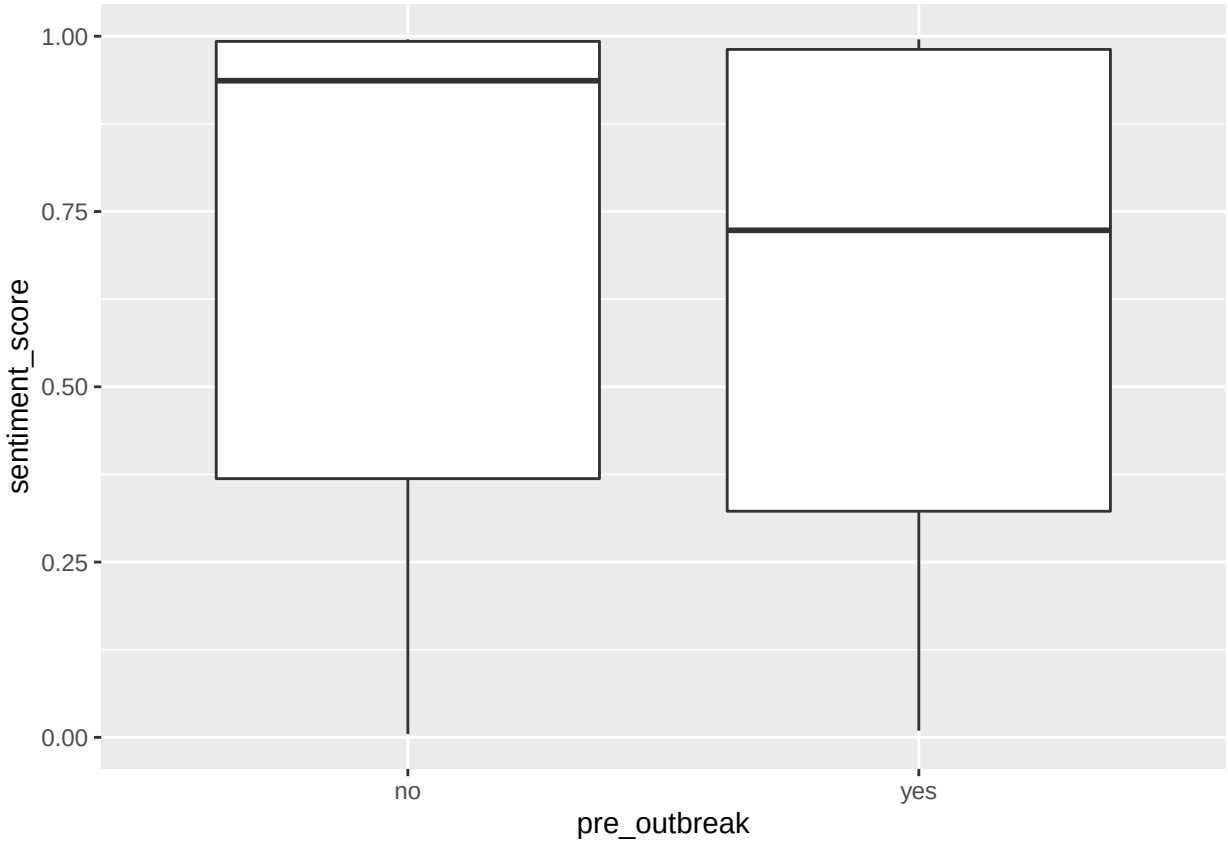
```
ggplot(Australia_analysis_travel) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




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

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.678
## 2 yes                         0.661
```

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




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

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.702
## 2 yes                            0.660
```

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

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

```
## # A tibble: 2 x 2
##   `pre_covid == "yes"`     n
##   <lgl>                <int>
## 1 FALSE                   70
## 2 TRUE                    58
```

```
t_num_precovid = 58
t_num_postcovid = 70
t_num = 128

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

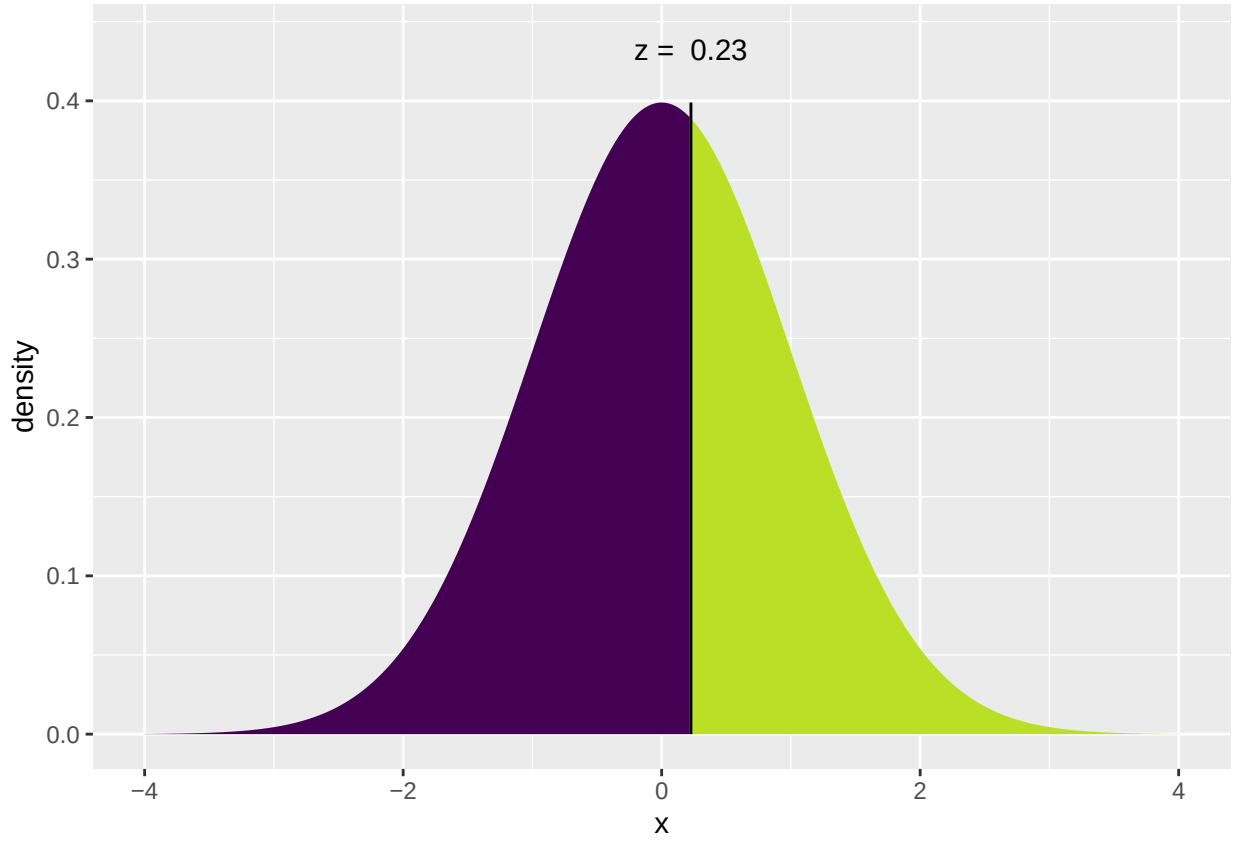
```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      21
## 2 TRUE                       37
```

```
p_hat_1_t_pos = 37/58

Australia_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                      24
## 2 TRUE                       46
```

```
p_hat_2_t_pos = 46/70

p_hat_t_pos = (37+46)/(58+70)

sd <- sqrt((((p_hat_t_pos)*(1-p_hat_t_pos))/58)+(((p_hat_t_pos)*(1-p_hat_t_pos))/70))
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 <= 0.2266) = P(Z <= 0.2266) = 0.5896
##  P(X >  0.2266) = P(Z >  0.2266) = 0.4104
##
```




```
## [1] 0.8207221
```

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

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

```
t_num_preoutbreak = 98
t_num_postoutbreak = 30
t_num = 128

Australia_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                      37
## 2 TRUE                       61
```

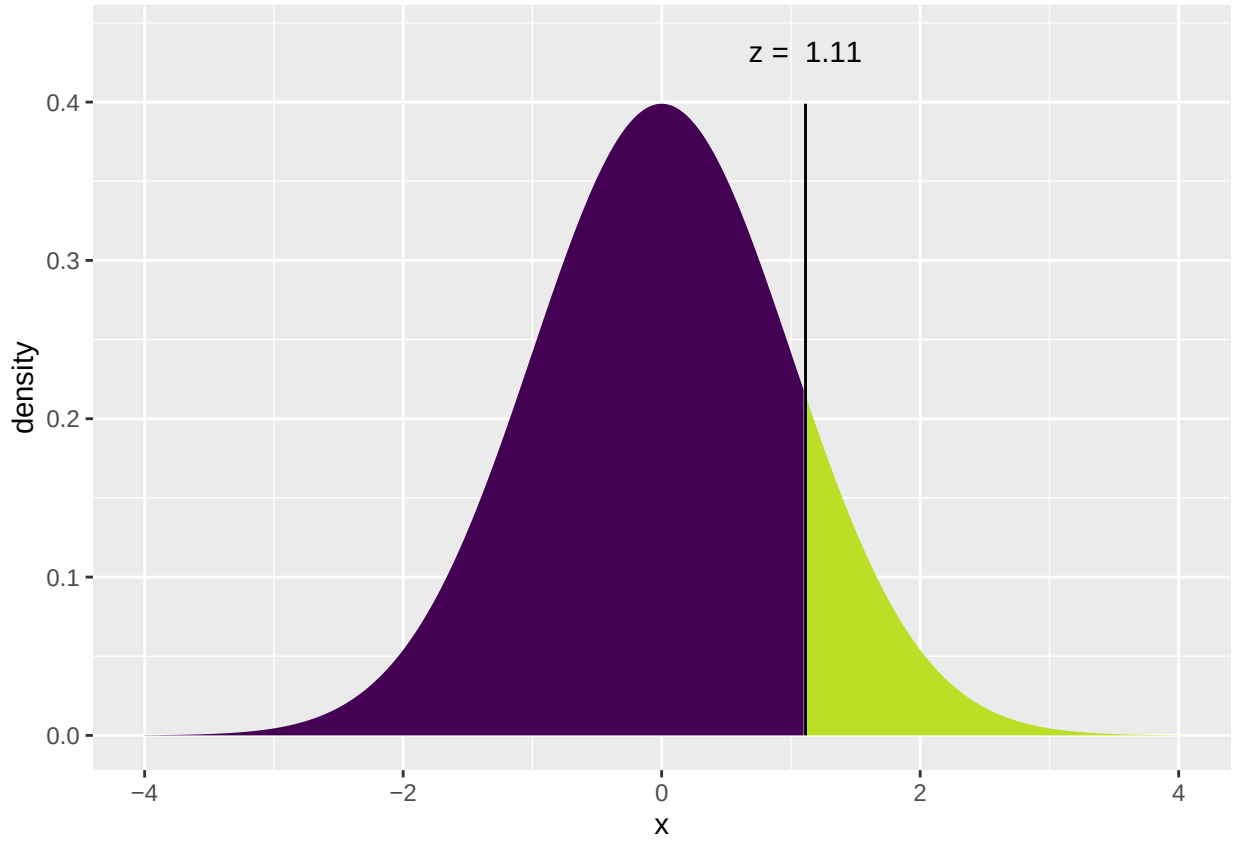
```
p_hat_1_t_pos = 61/98

Australia_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                       8
## 2 TRUE                       22
```

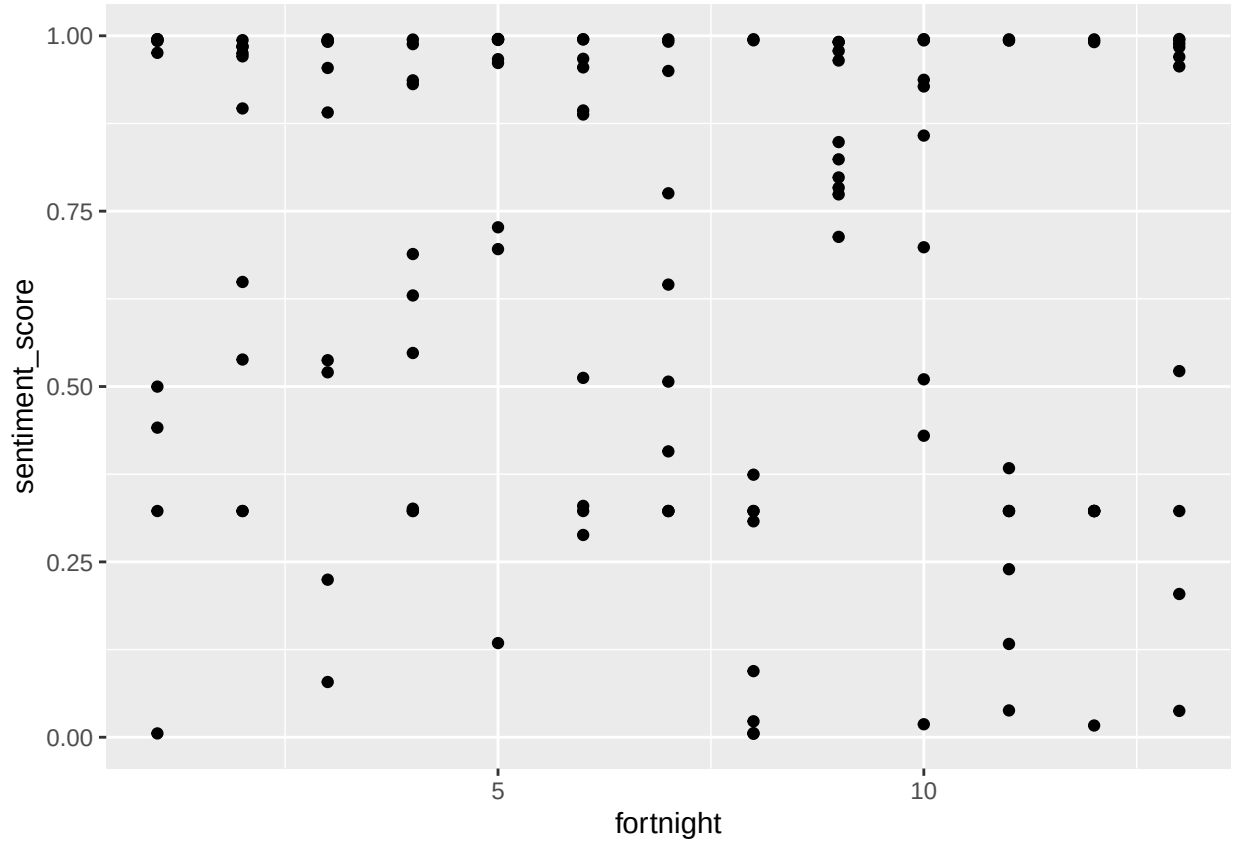
```
p_hat_2_t_pos = 22/30

p_hat_t_pos = (61+22)/(98+30)

sd <- sqrt((((p_hat_t_pos)*(1-p_hat_t_pos))/98)+(((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 <= 1.113) = P(Z <= 1.113) = 0.8672
##  P(X >  1.113) = P(Z >  1.113) = 0.1328
##
```

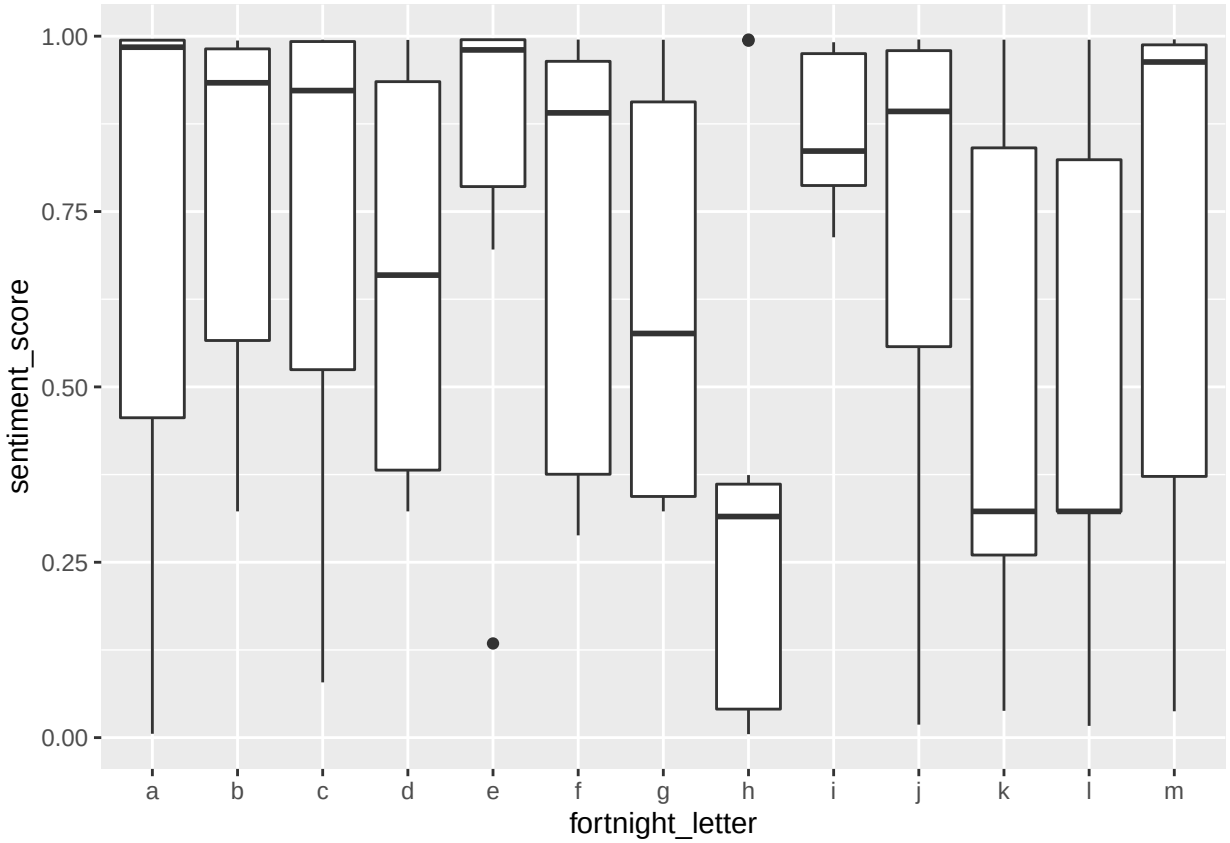


```
## [1] 0.2656991
```

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



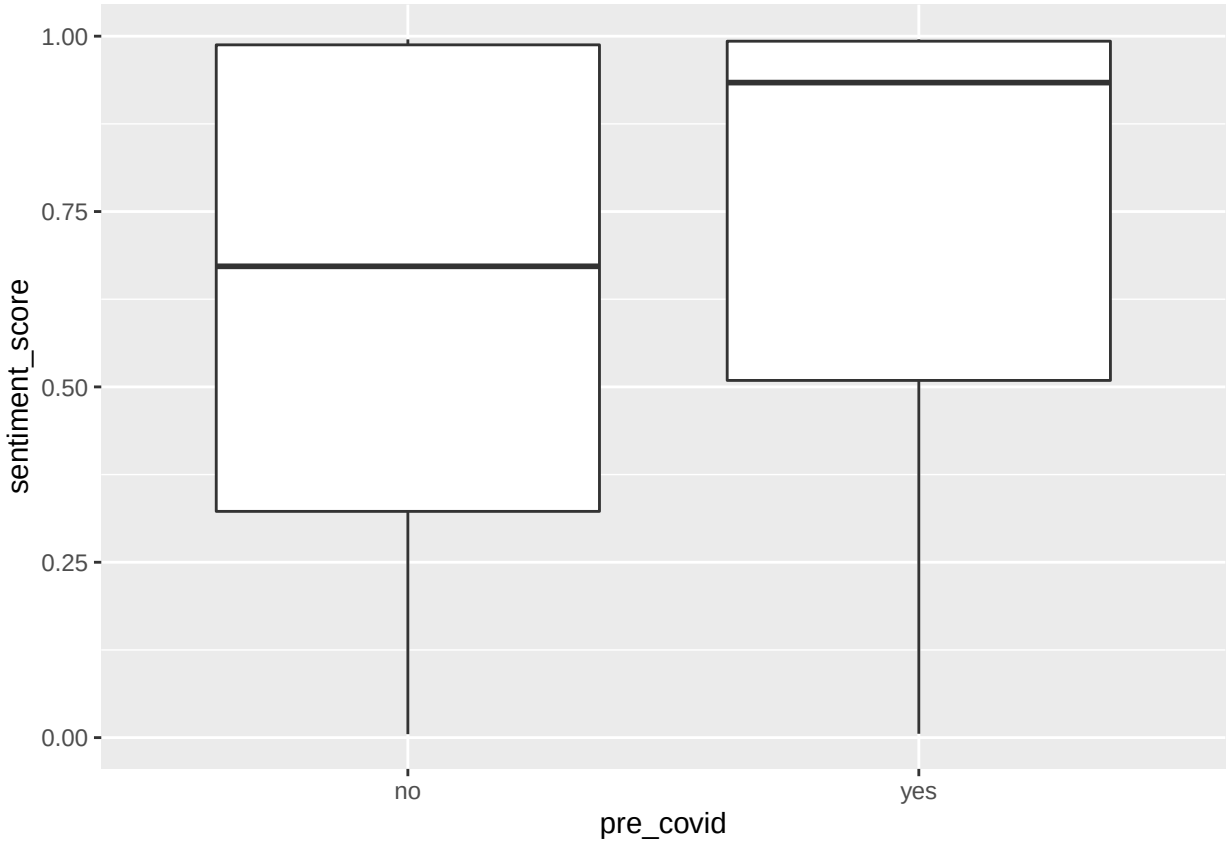

```
ggplot(Australia_analysis_people) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




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

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.722
##  2         2                   0.764
##  3         3                   0.718
##  4         4                   0.669
##  5         5                   0.846
##  6         6                   0.715
##  7         7                   0.624
##  8         8                   0.344
##  9         9                   0.867
## 10        10                   0.736
## 11        11                   0.474
## 12        12                   0.493
## 13        13                   0.697
```

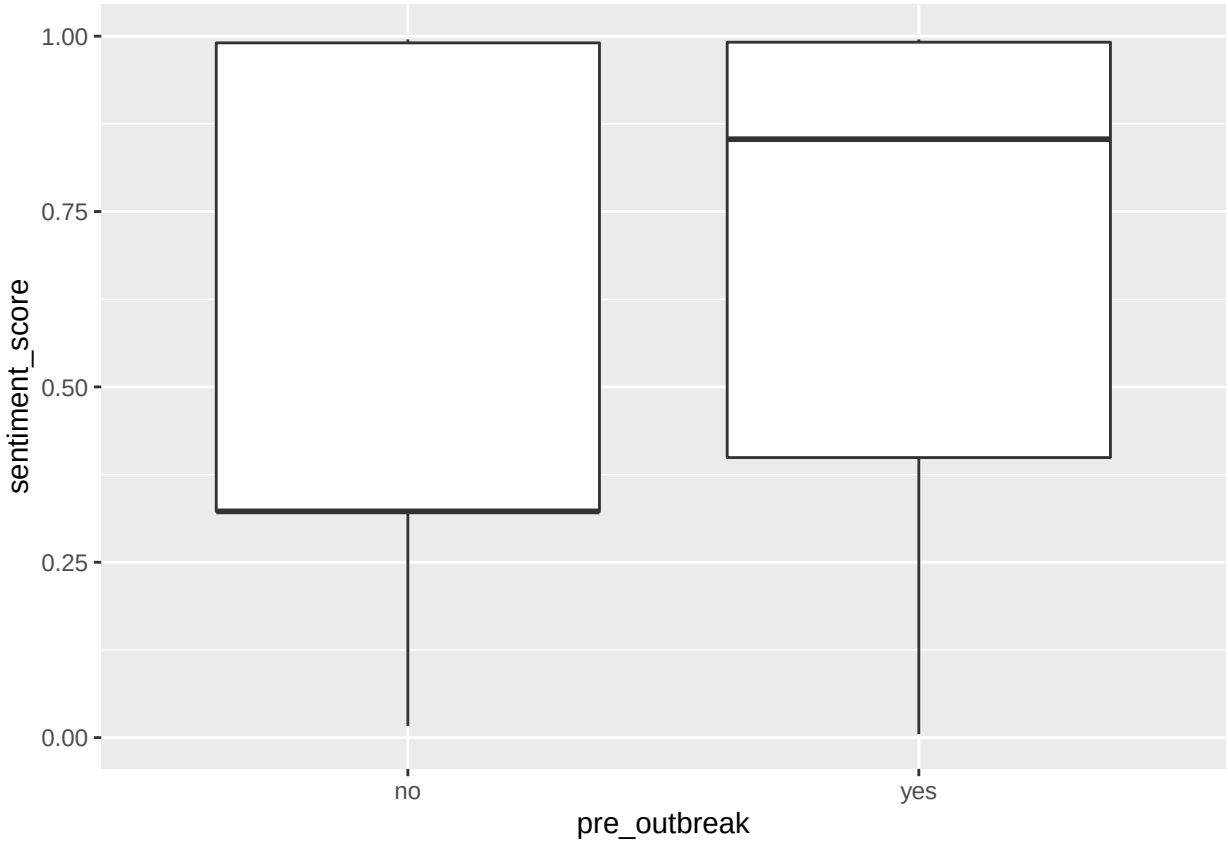
```
ggplot(Australia_analysis_people) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




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

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.605
## 2 yes                         0.739
```

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




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

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.555
## 2 yes                            0.700
```

```
#precovid people
count(Australia_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

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

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

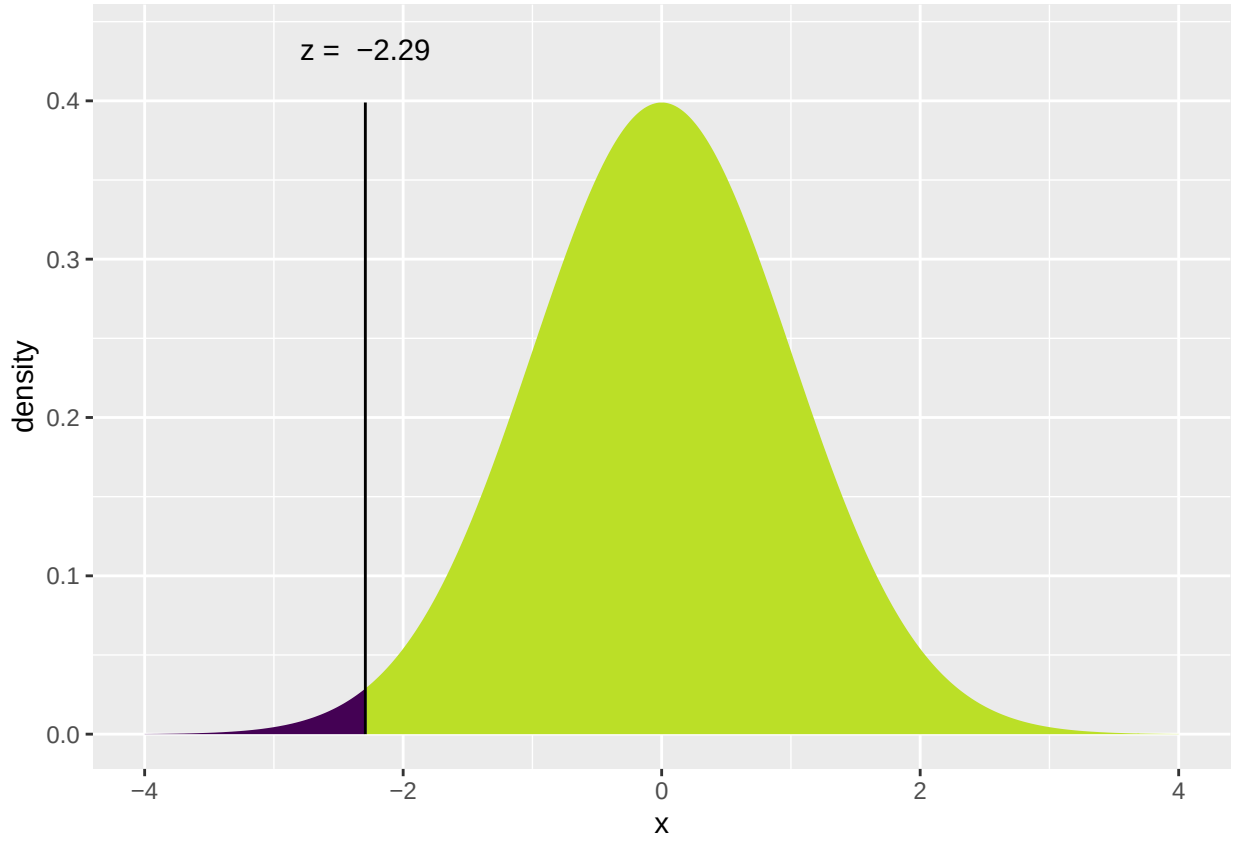
```
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      15
## 2 TRUE                       45
```

```
p_hat_1_p_pos = 45/60

Australia_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                      31
## 2 TRUE                       39
```

```
p_hat_2_p_pos = 39/70

p_hat_p_pos = (45+39)/(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* (xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= -2.293) = P(Z <= -2.293) = 0.01094
##  P(X >  -2.293) = P(Z >  -2.293) = 0.9891
##
```




```
## [1] 0.02187555
```

```
#outbreak people
count(Australia_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

Australia_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                      29
## 2 TRUE                       71
```

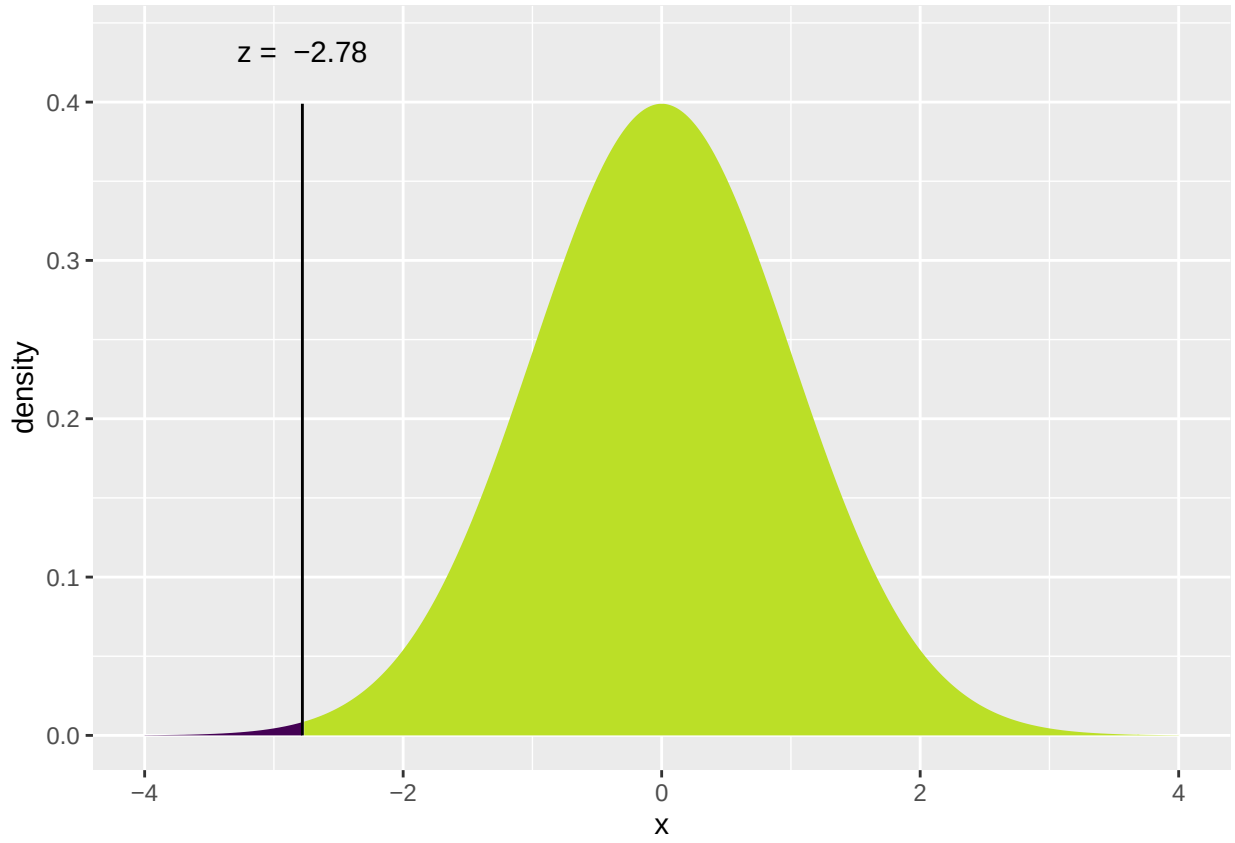
```
p_hat_1_p_pos =71/100

Australia_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                      17
## 2 TRUE                       13
```

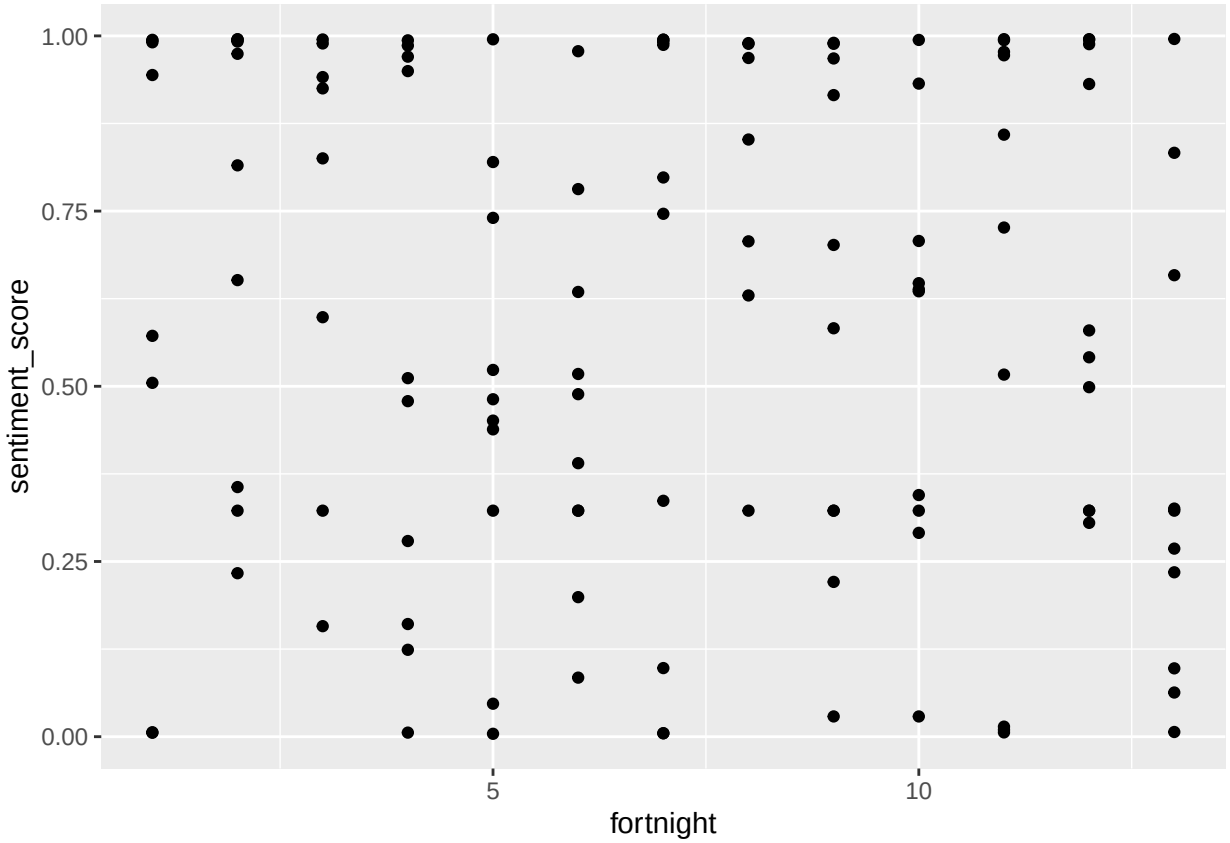
```
p_hat_2_p_pos = 13/30

p_hat_p_pos = (71+13)/(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* (xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= -2.78) = P(Z <= -2.78) = 0.002722
##  P(X >  -2.78) = P(Z >  -2.78) = 0.9973
##
```

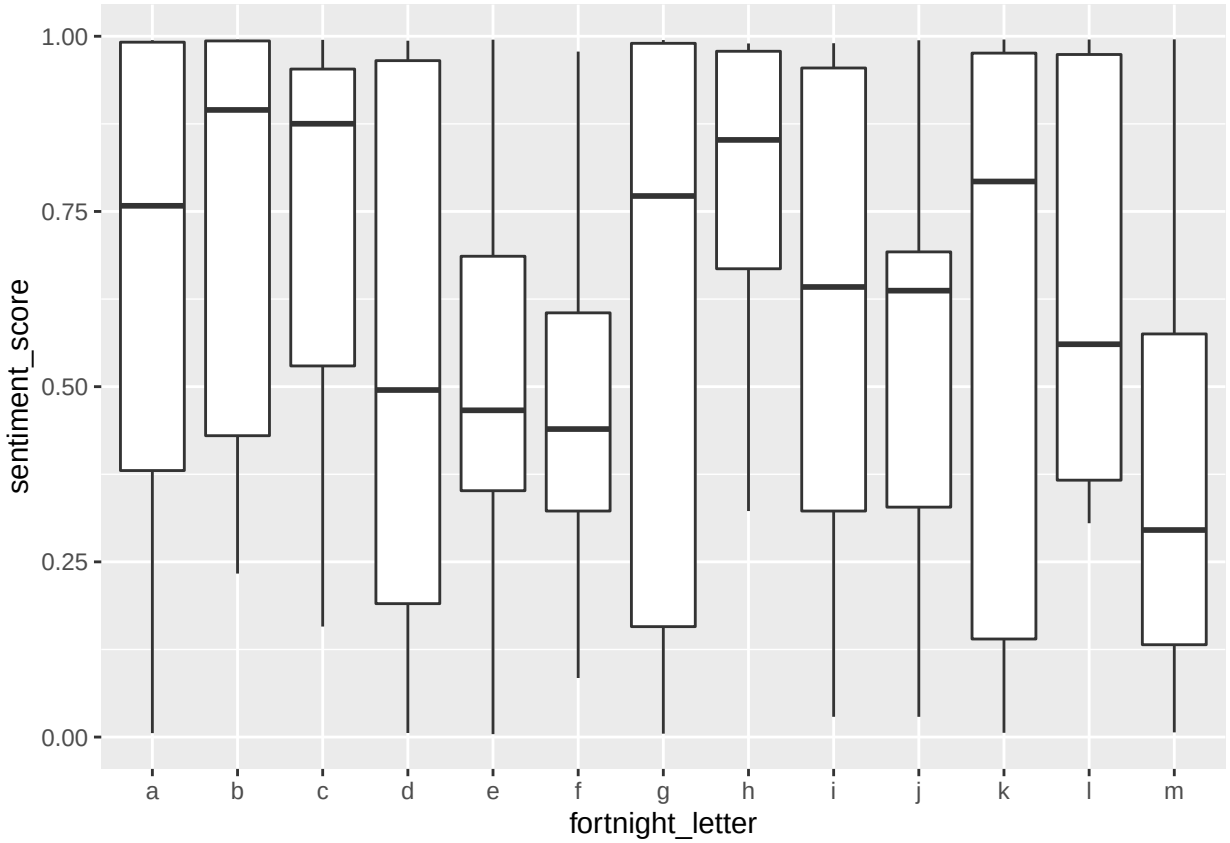


```
## [1] 0.005443847
```

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



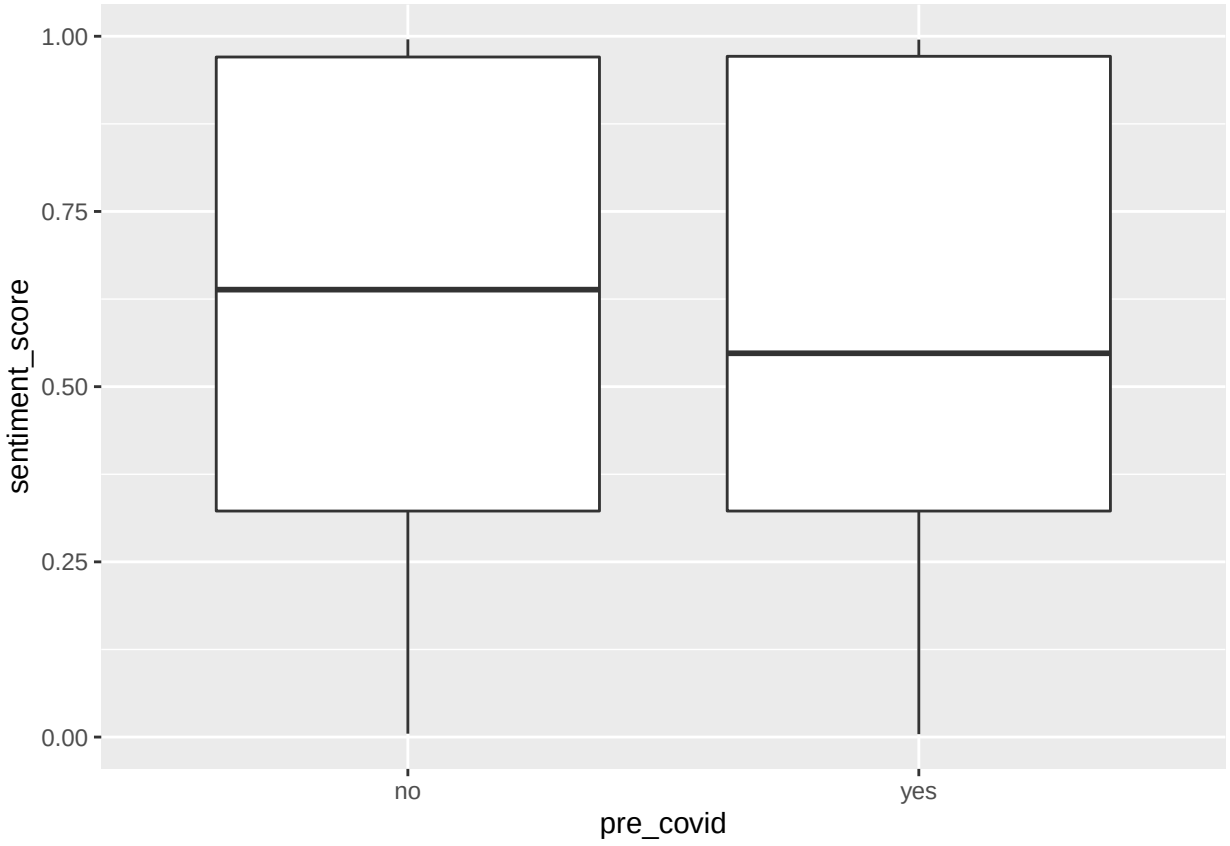

```
ggplot(Australia_analysis_entertainment) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




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

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.626
##  2         2                   0.733
##  3         3                   0.719
##  4         4                   0.546
##  5         5                   0.482
##  6         6                   0.472
##  7         7                   0.596
##  8         8                   0.780
##  9         9                   0.604
## 10        10                   0.554
## 11        11                   0.607
## 12        12                   0.648
## 13        13                   0.381
```

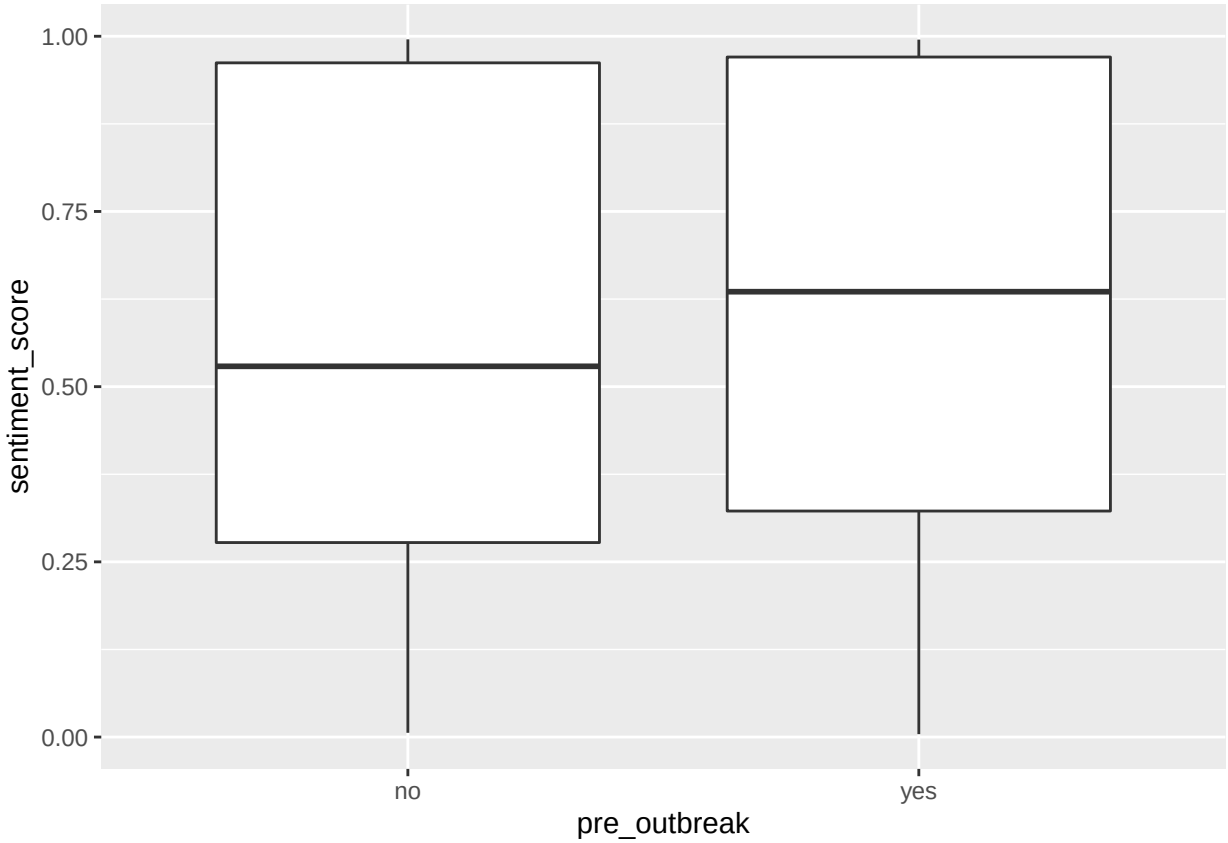
```
ggplot(Australia_analysis_entertainment) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




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

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.587
## 2 yes                         0.591
```

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




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

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.545
## 2 yes                            0.603
```

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

```
## # A tibble: 2 x 2
##   `pre_covid == "yes"`     n
##   <lgl>                <int>
## 1 FALSE                   67
## 2 TRUE                    56
```

```
num_precovid = 56
num_postcovid = 67
num = 123

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

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

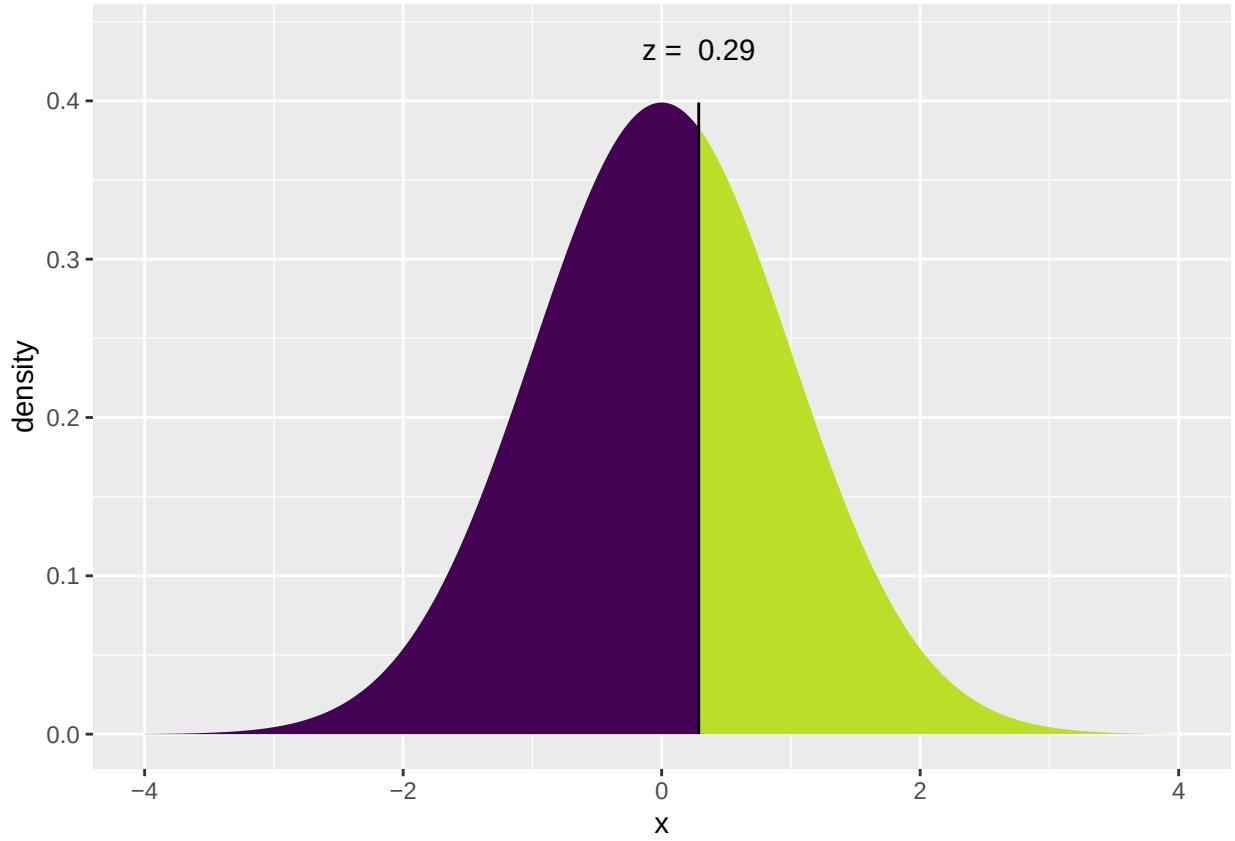
```
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      24
## 2 TRUE                       32
```

```
#proportion of positive sentiment videos precovid from sample
p_hat1 = 32/56

Australia_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                      27
## 2 TRUE                       40
```

```
#proportion of positive sentiment videos postcovid from sample
p_hat2 = 40/67

p_hat = (32+40)/(56+67)

sd <- sqrt((((p_hat)*(1-p_hat))/56)+(((p_hat)*(1-p_hat))/67))
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 <= 0.2868) = P(Z <= 0.2868) = 0.6129
##  P(X >  0.2868) = P(Z >  0.2868) = 0.3871
##
```




```
## [1] 0.7742342
```

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

```
## # A tibble: 2 x 2
##   `pre_outbreak == "yes"`     n
##   <lgl>                   <int>
## 1 FALSE                      30
## 2 TRUE                       93
```

```
num_preoutbreak = 93
num_postoutbreak = 30
num = 130

Australia_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                      37
## 2 TRUE                       56
```

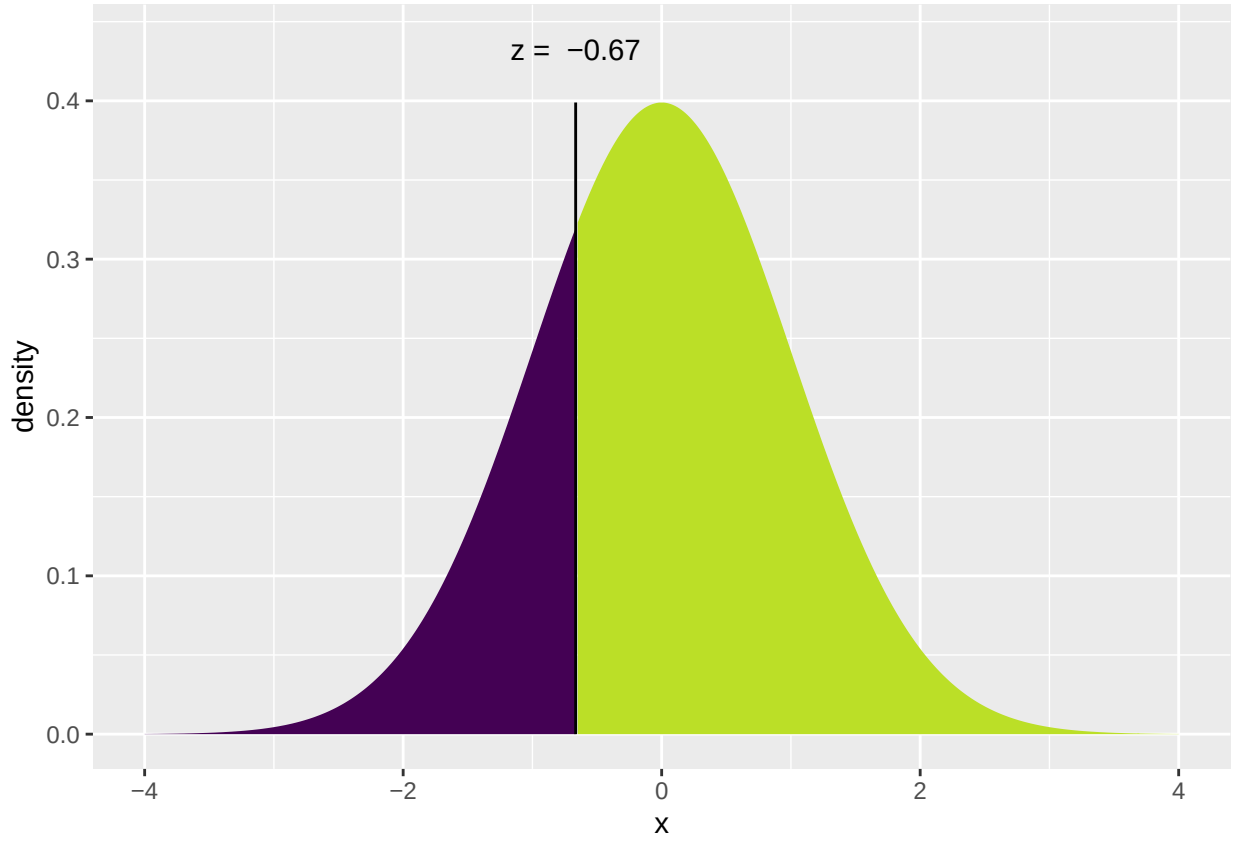
```
#proportion of positive sentiment videos preoutbreak from sample
p_hat1 = 56/93
```

```
Australia_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                      14
## 2 TRUE                       16
```

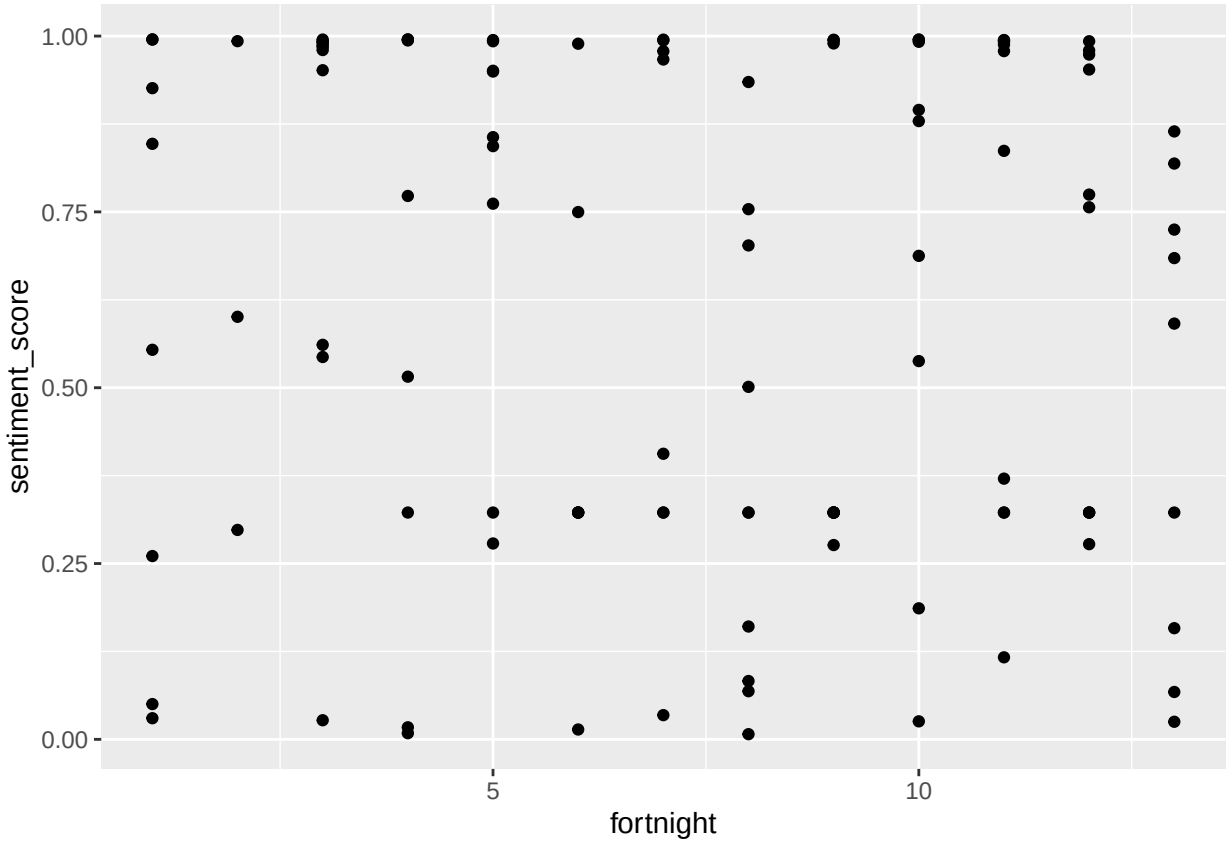
```
#proportion of positive sentiment videos postoutbreak from sample
p_hat2 = 16/30

p_hat = (56+16)/(93+30)

sd <- sqrt((((p_hat)*(1-p_hat))/93)+(((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 <= -0.6653) = P(Z <= -0.6653) = 0.2529
##  P(X >  -0.6653) = P(Z >  -0.6653) = 0.7471
##
```

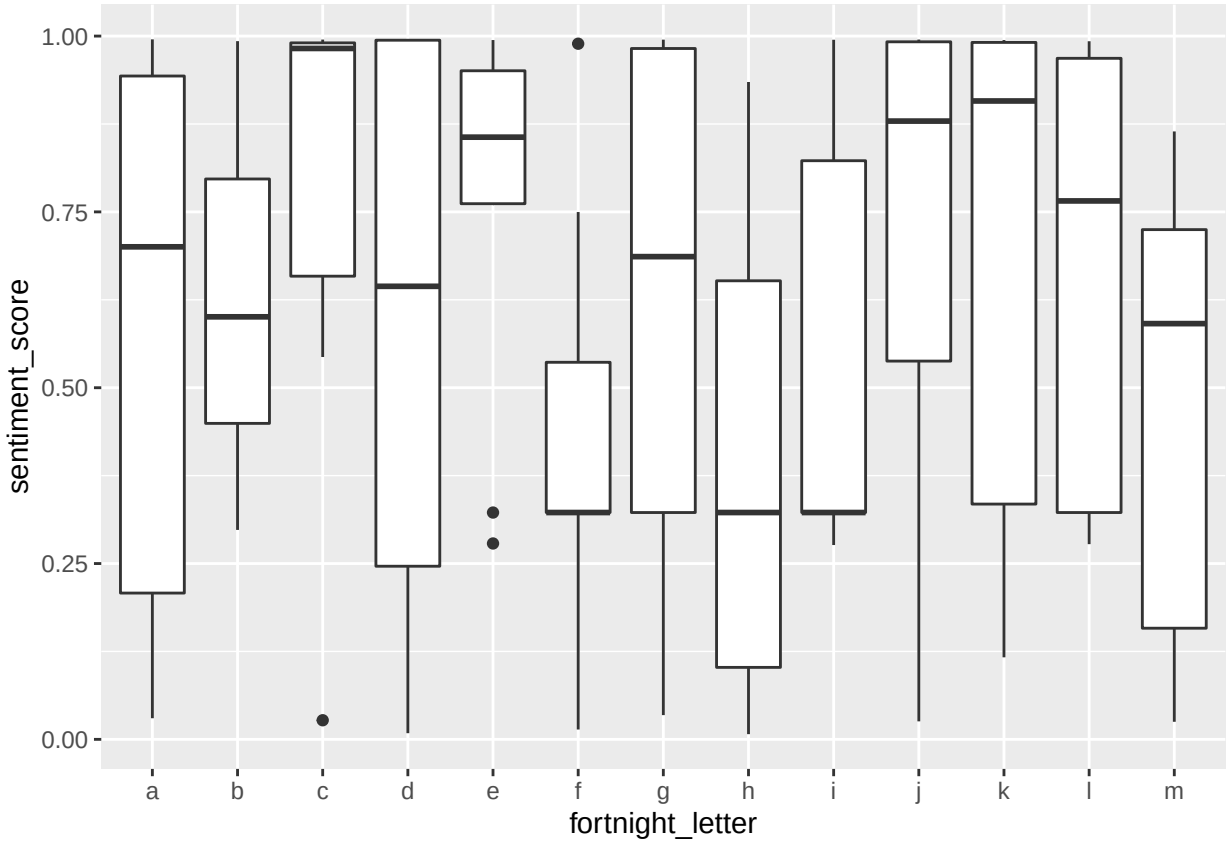



```
## [1] 0.5058755
```

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



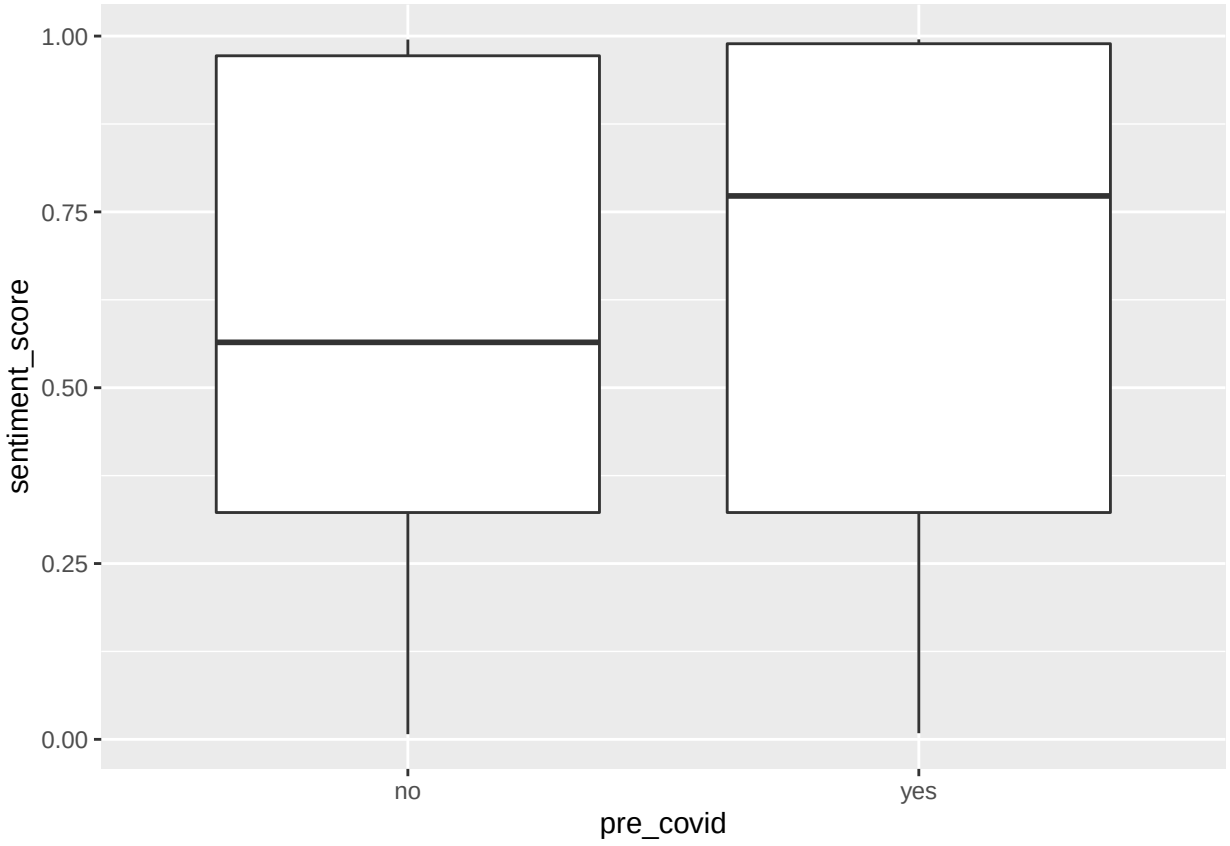

```
ggplot(Australia_analysis_news) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




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

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.582
##  2         2                   0.630
##  3         3                   0.801
##  4         4                   0.578
##  5         5                   0.772
##  6         6                   0.435
##  7         7                   0.627
##  8         8                   0.386
##  9         9                   0.519
## 10        10                   0.688
## 11        11                   0.692
## 12        12                   0.667
## 13        13                   0.473
```

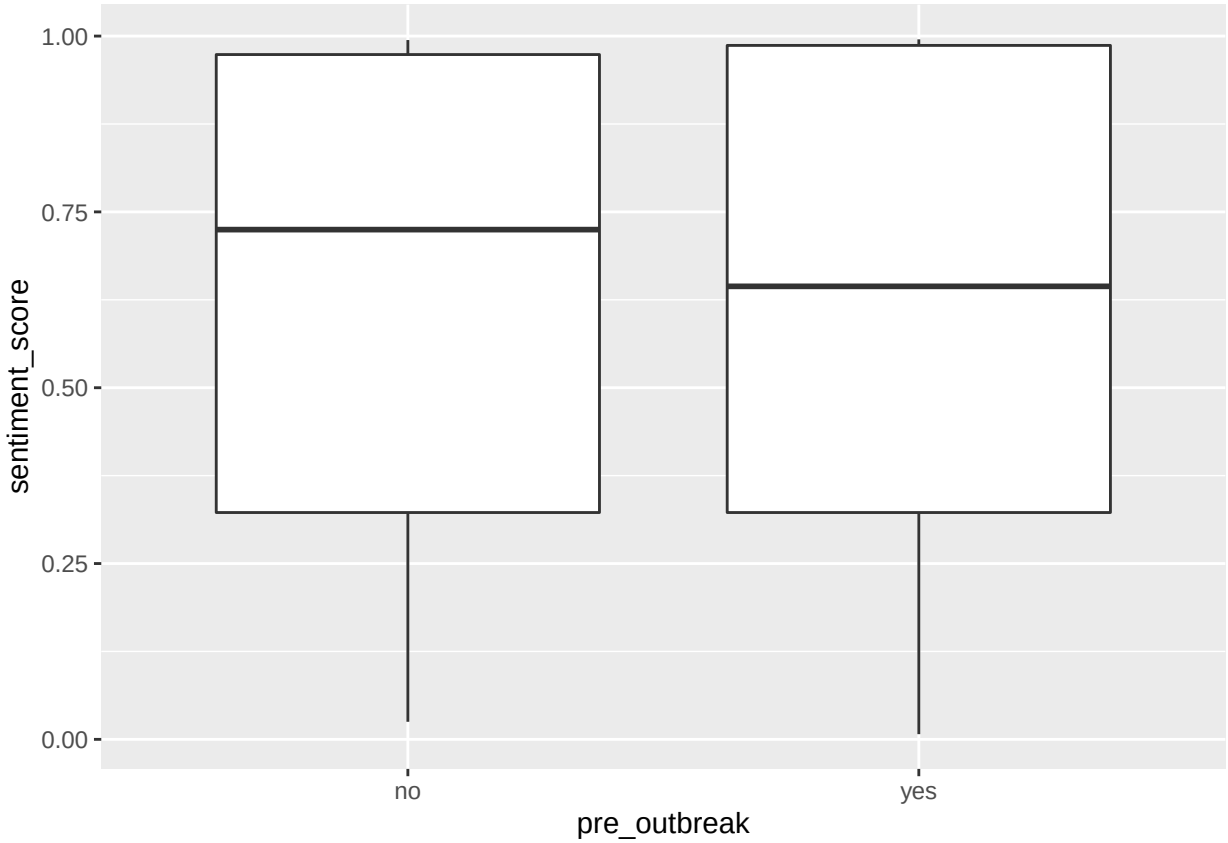
```
ggplot(Australia_analysis_news) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




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

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.577
## 2 yes                         0.648
```

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




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

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.615
## 2 yes                            0.603
```

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

```
## # A tibble: 2 x 2
##   `pre_covid == "yes"`     n
##   <lgl>                <int>
## 1 FALSE                   66
## 2 TRUE                    45
```

```
num_precovid = 45
num_postcovid = 66
num = 111

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

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

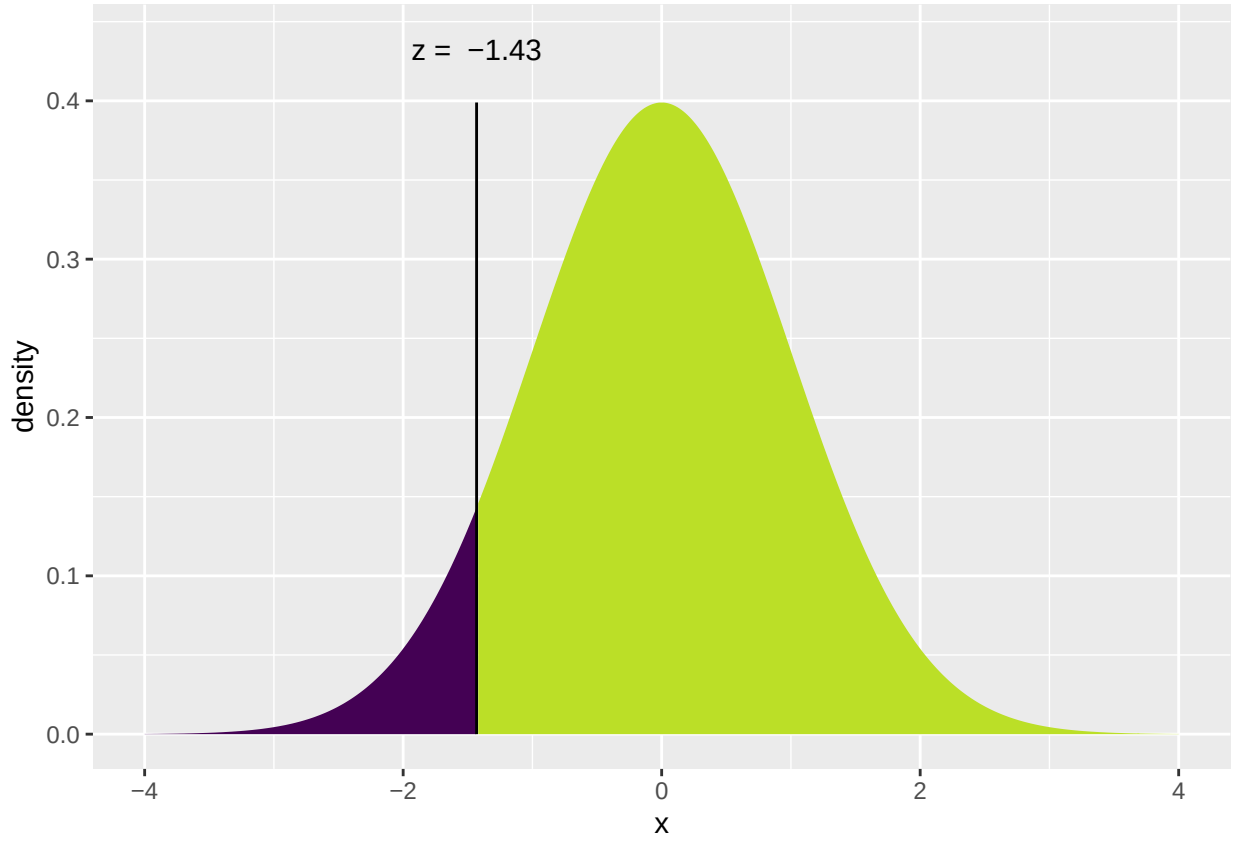
```
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      15
## 2 TRUE                       30
```

```
#proportion of positive sentiment videos precovid from sample
p_hat1 = 30/45

Australia_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                      31
## 2 TRUE                       35
```

```
#proportion of positive sentiment videos postcovid from sample
p_hat2 = 35/66

p_hat = (30+35)/(45+66)

sd <- sqrt((((p_hat)*(1-p_hat))/45)+(((p_hat)*(1-p_hat))/66))
z_score <- ((p_hat2-p_hat1)-0)/sd

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

```
##
## If X ~ N(0, 1), then
##  P(X <= -1.432) = P(Z <= -1.432) = 0.07609
##  P(X >  -1.432) = P(Z >  -1.432) = 0.9239
##
```




```
## [1] 0.1521819
```

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

```
## # A tibble: 2 x 2
##   `pre_outbreak == "yes"`     n
##   <lgl>                   <int>
## 1 FALSE                      29
## 2 TRUE                       82
```

```
num_preoutbreak = 82
num_postoutbreak = 29
num = 111

Australia_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                      34
## 2 TRUE                       48
```

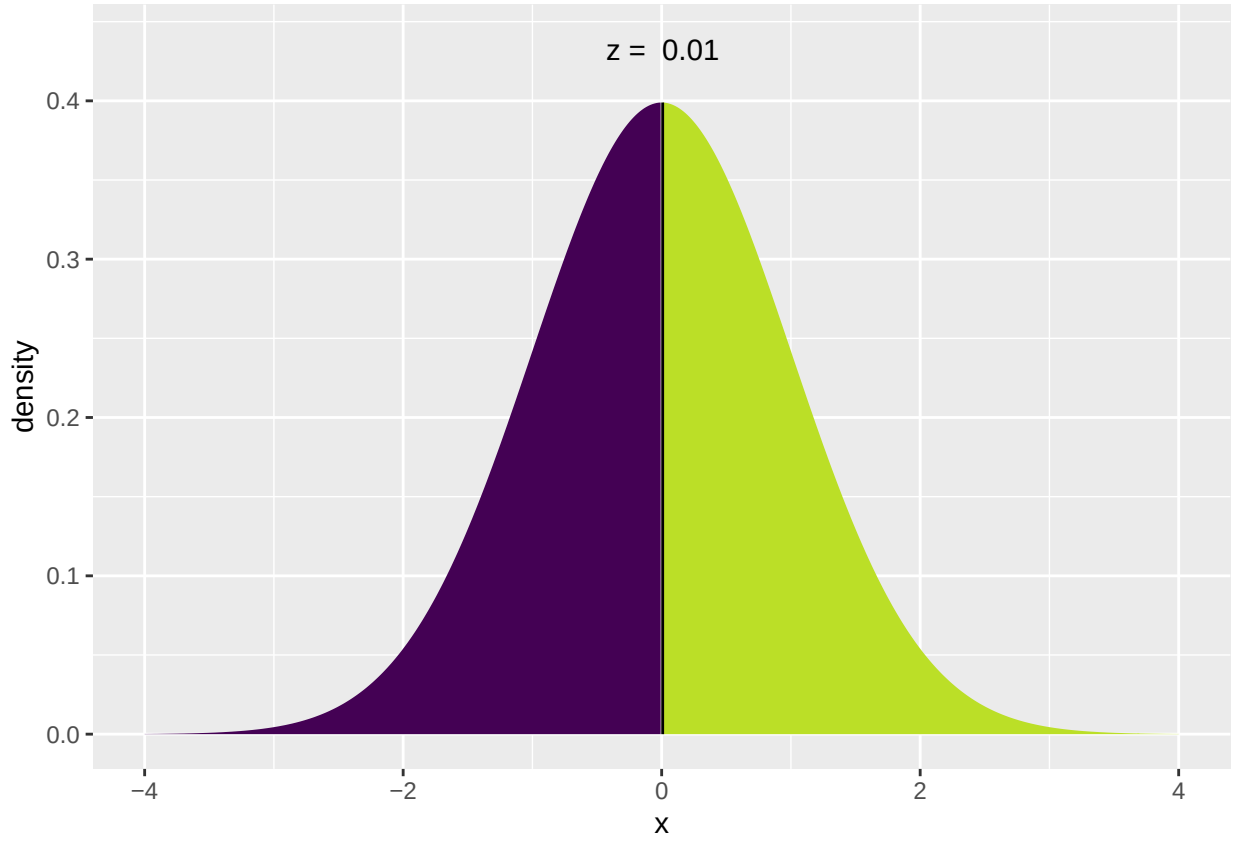
```
#proportion of positive sentiment videos preoutbreak from sample
p_hat1 = 48/82
```

```
Australia_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                      12
## 2 TRUE                       17
```

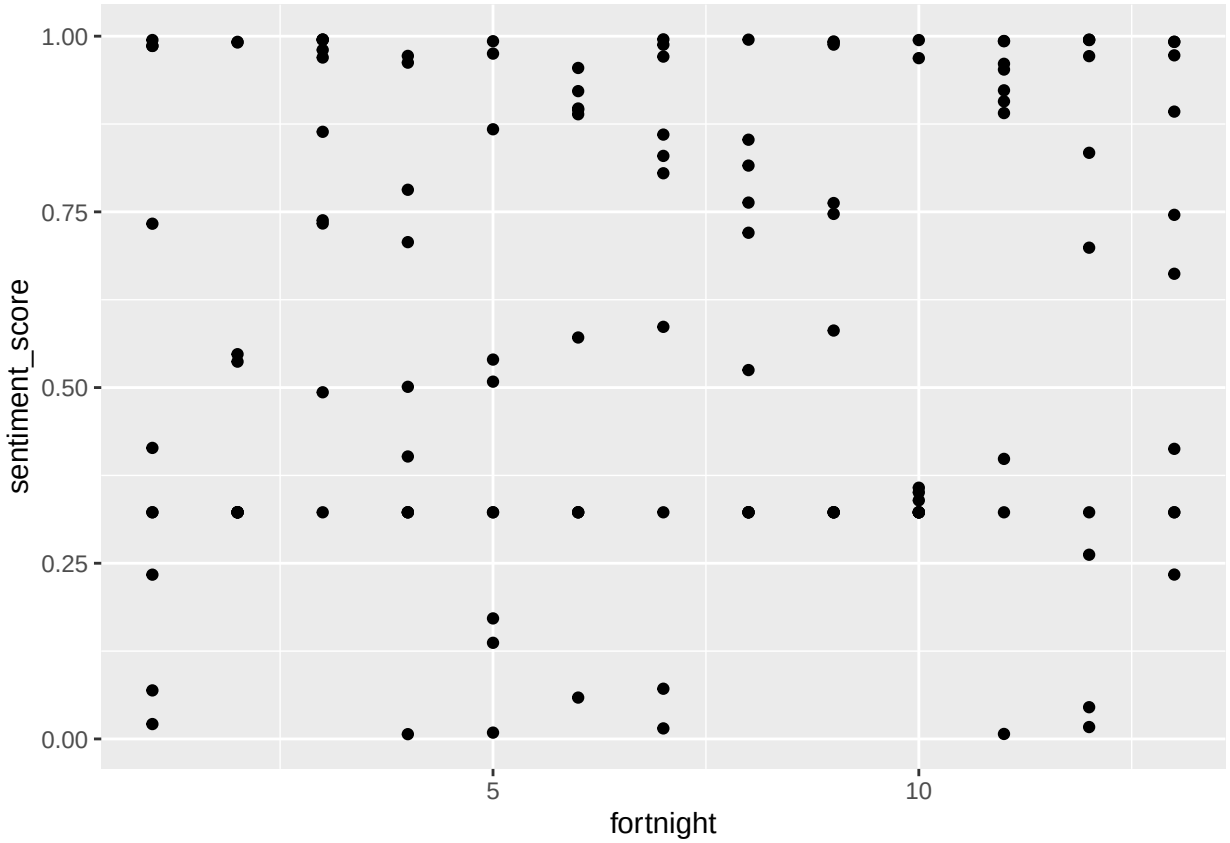
```
#proportion of positive sentiment videos postoutbreak from sample
p_hat2 = 17/29

p_hat = (48+17)/(82+29)

sd <- sqrt((((p_hat)*(1-p_hat))/82)+(((p_hat)*(1-p_hat))/29))
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 <= 0.007902) = P(Z <= 0.007902) = 0.5032
##  P(X >  0.007902) = P(Z >  0.007902) = 0.4968
##
```

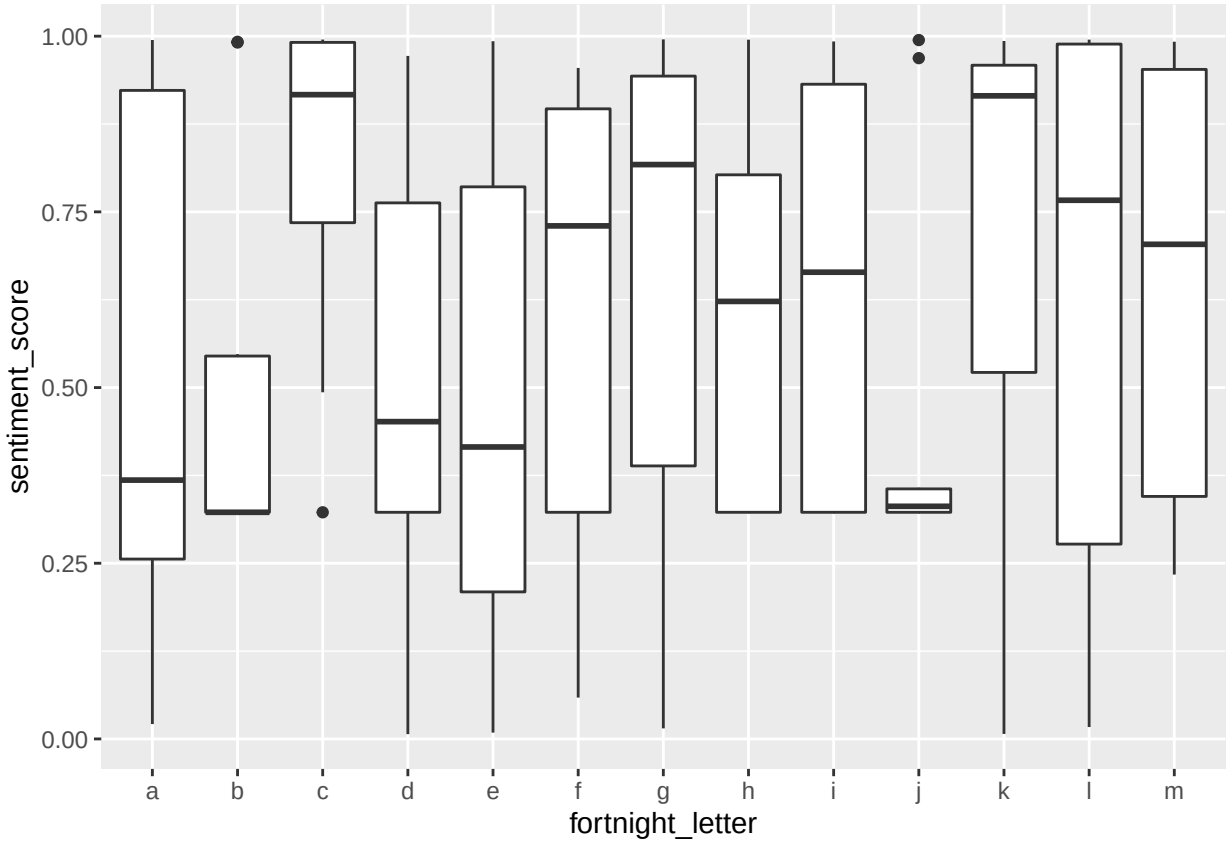



```
## [1] 0.993695
```

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



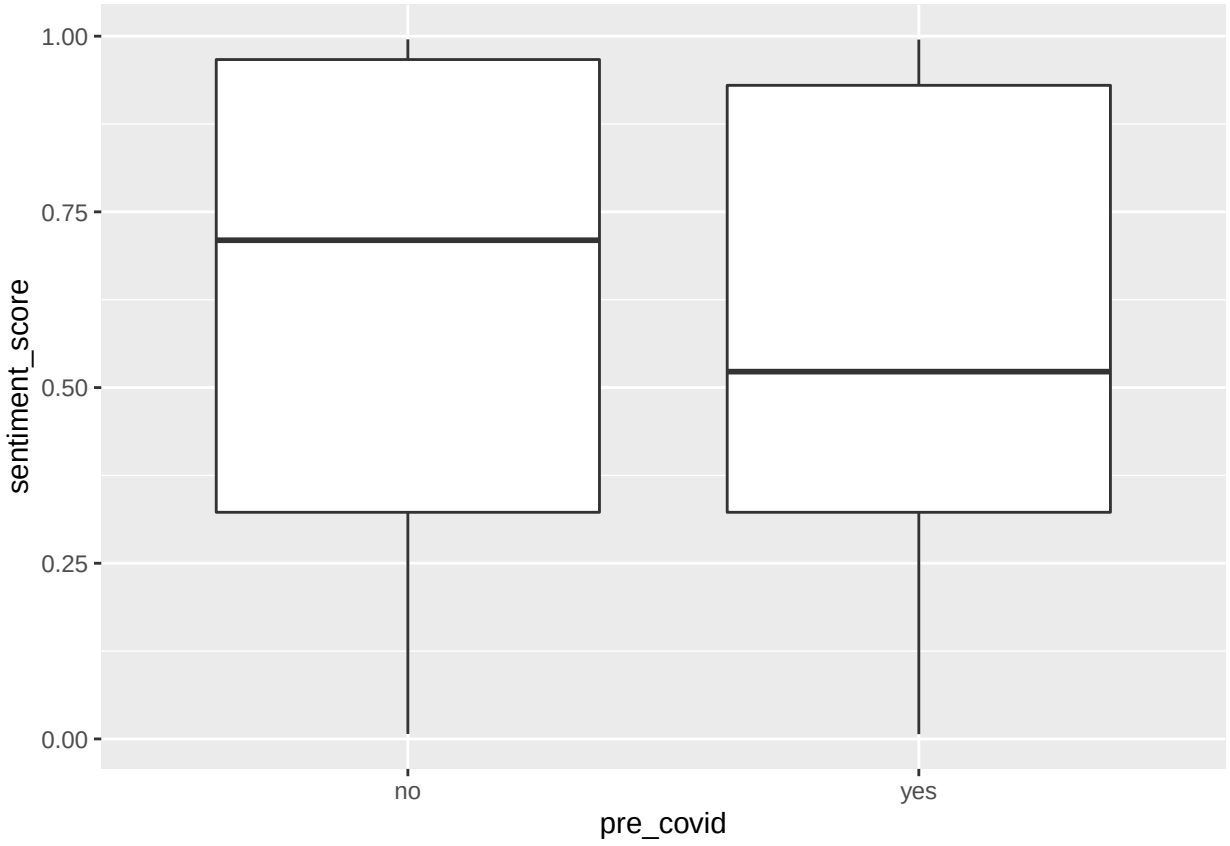

```
ggplot(Australia_analysis_how_to) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




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

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.508
##  2         2                   0.500
##  3         3                   0.809
##  4         4                   0.530
##  5         5                   0.485
##  6         6                   0.616
##  7         7                   0.644
##  8         8                   0.596
##  9         9                   0.635
## 10        10                   0.462
## 11        11                   0.735
## 12        12                   0.614
## 13        13                   0.655
```

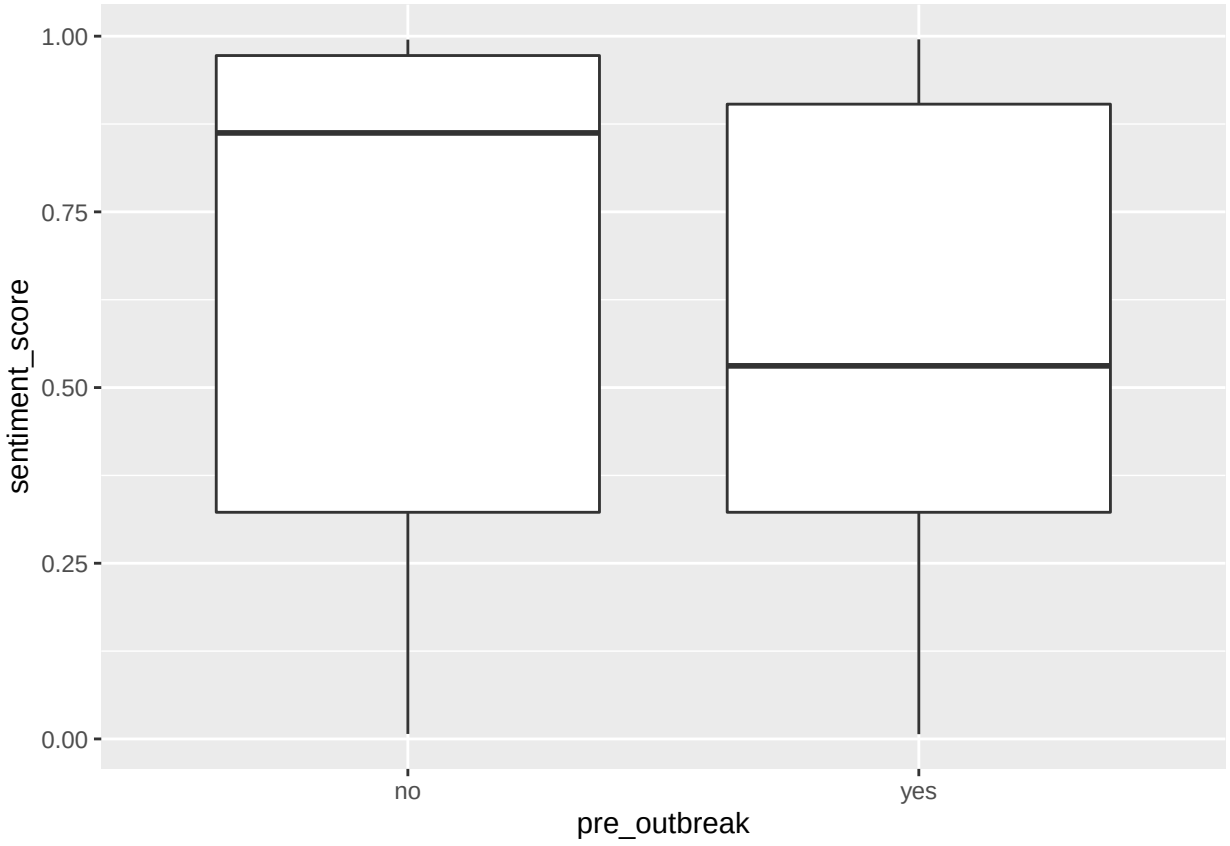
```
ggplot(Australia_analysis_how_to) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




```
Australia_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                          0.620
## 2 yes                         0.575
```

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




```
Australia_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                             0.668
## 2 yes                            0.579
```

```
#precovid how-to
count(Australia_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

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

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

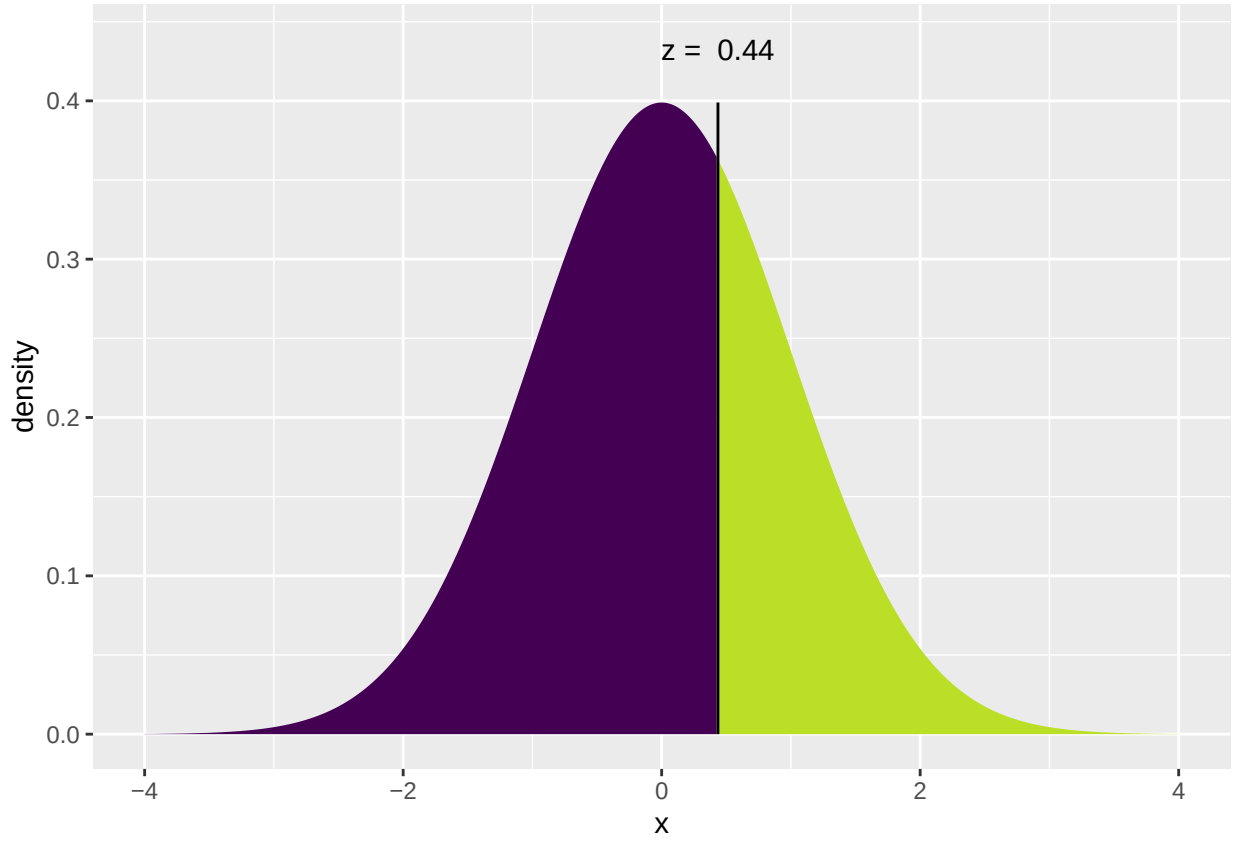
```
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      28
## 2 TRUE                       32
```

```
#proportion of positive sentiment videos precovid from sample
p_hat1 = 32/60

Australia_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                      30
## 2 TRUE                       40
```

```
#proportion of positive sentiment videos postcovid from sample
p_hat2 = 40/70

p_hat = (32+40)/(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 <= 0.4356) = P(Z <= 0.4356) = 0.6684
##  P(X >  0.4356) = P(Z >  0.4356) = 0.3316
##
```




```
## [1] 0.6631278
```

```
#outbreak how-to
count(Australia_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

Australia_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                      47
## 2 TRUE                       53
```

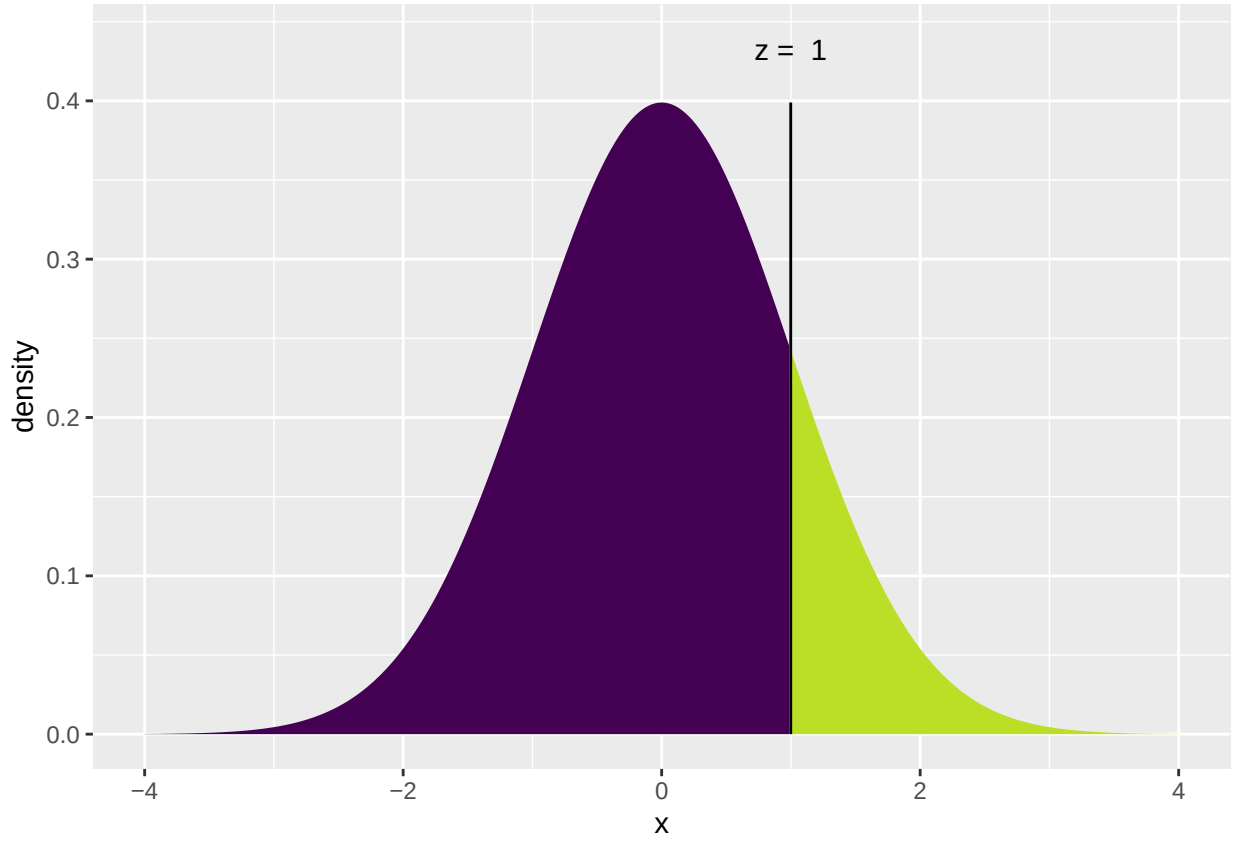
```
#proportion of positive sentiment videos preoutbreak from sample
p_hat1 = 53/100
```

```r
Australia_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                      11
## 2 TRUE                       19
```

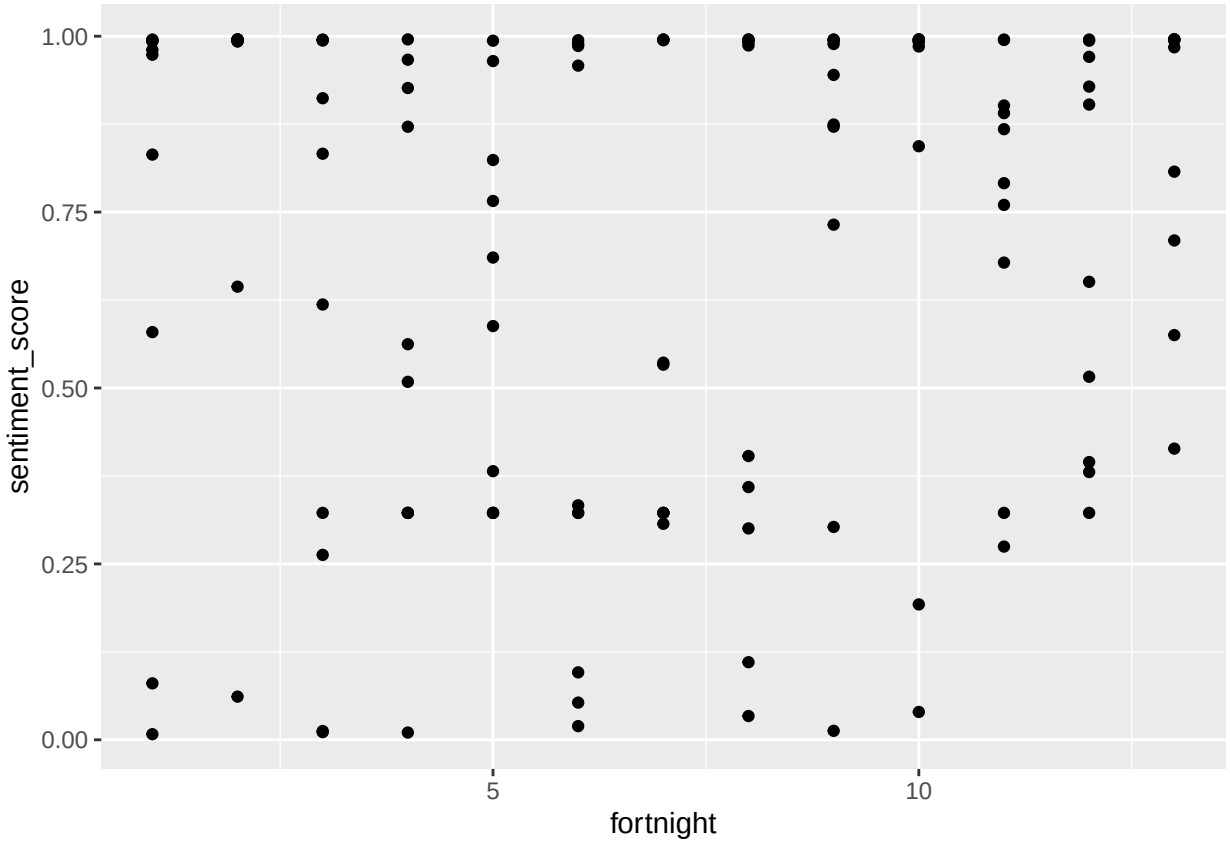
```r
#proportion of positive sentiment videos postoutbreak from sample
p_hat2 = 19/30

p_hat = (53+19)/(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 <= 0.9986) = P(Z <= 0.9986) = 0.841
##  P(X >  0.9986) = P(Z >  0.9986) = 0.159
##
```

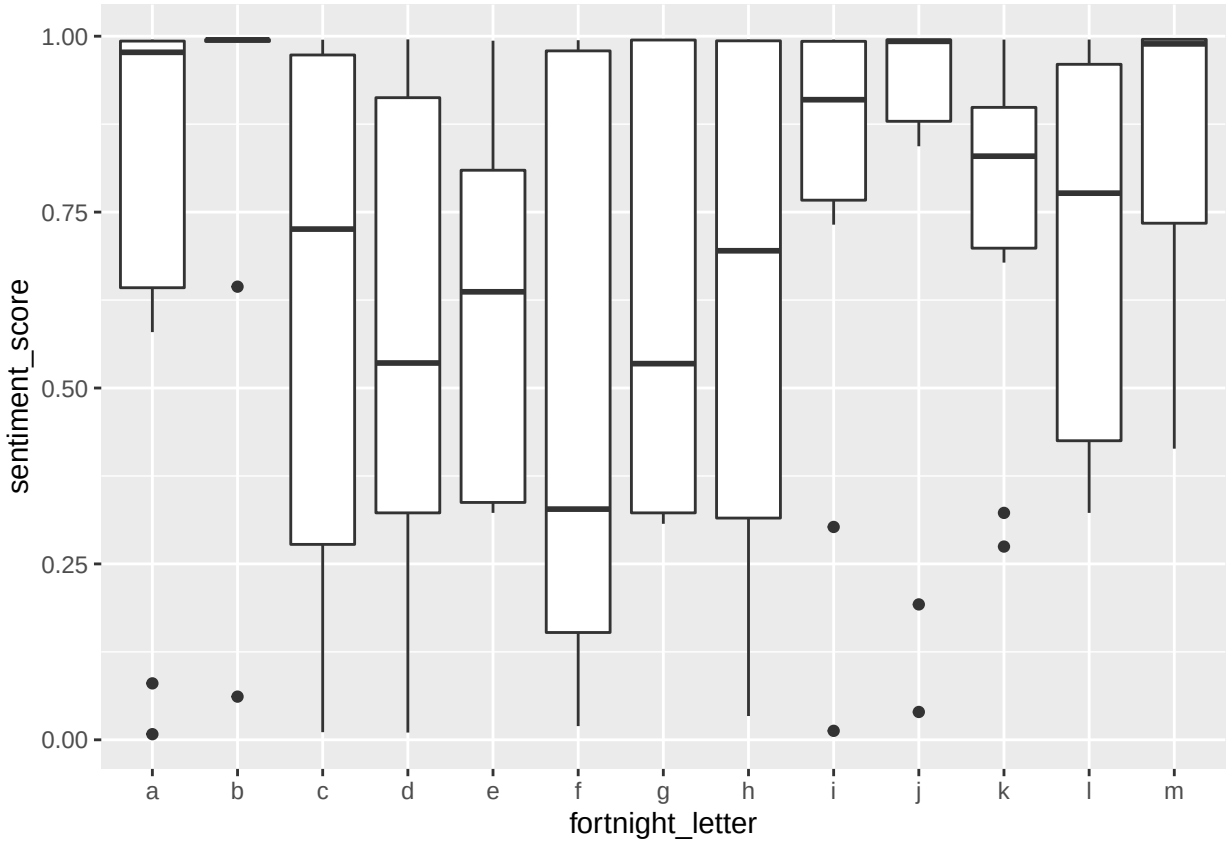



```
## [1] 0.3179875
```

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



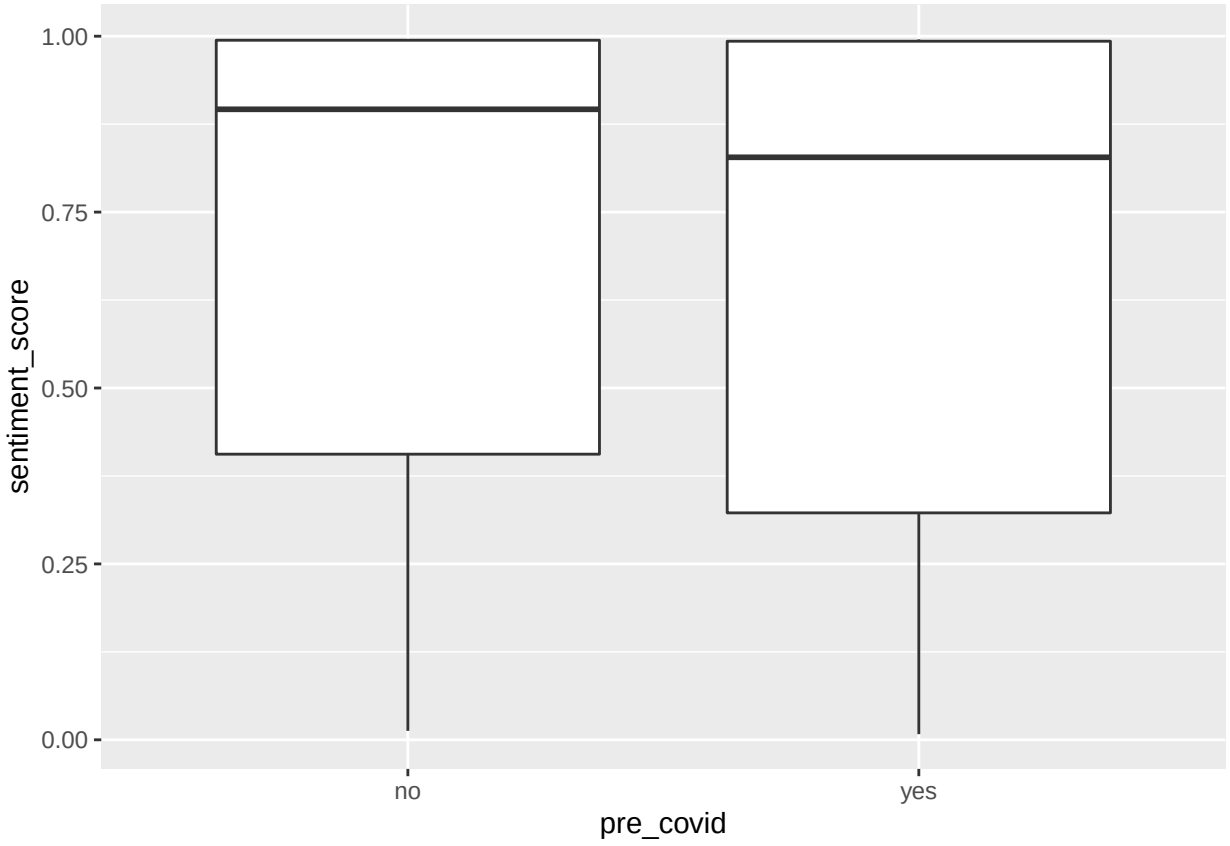

```
ggplot(Australia_analysis_education) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




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

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.743
##  2         2                   0.866
##  3         3                   0.596
##  4         4                   0.581
##  5         5                   0.617
##  6         6                   0.507
##  7         7                   0.632
##  8         8                   0.617
##  9         9                   0.771
## 10        10                   0.803
## 11        11                   0.748
## 12        12                   0.706
## 13        13                   0.847
```

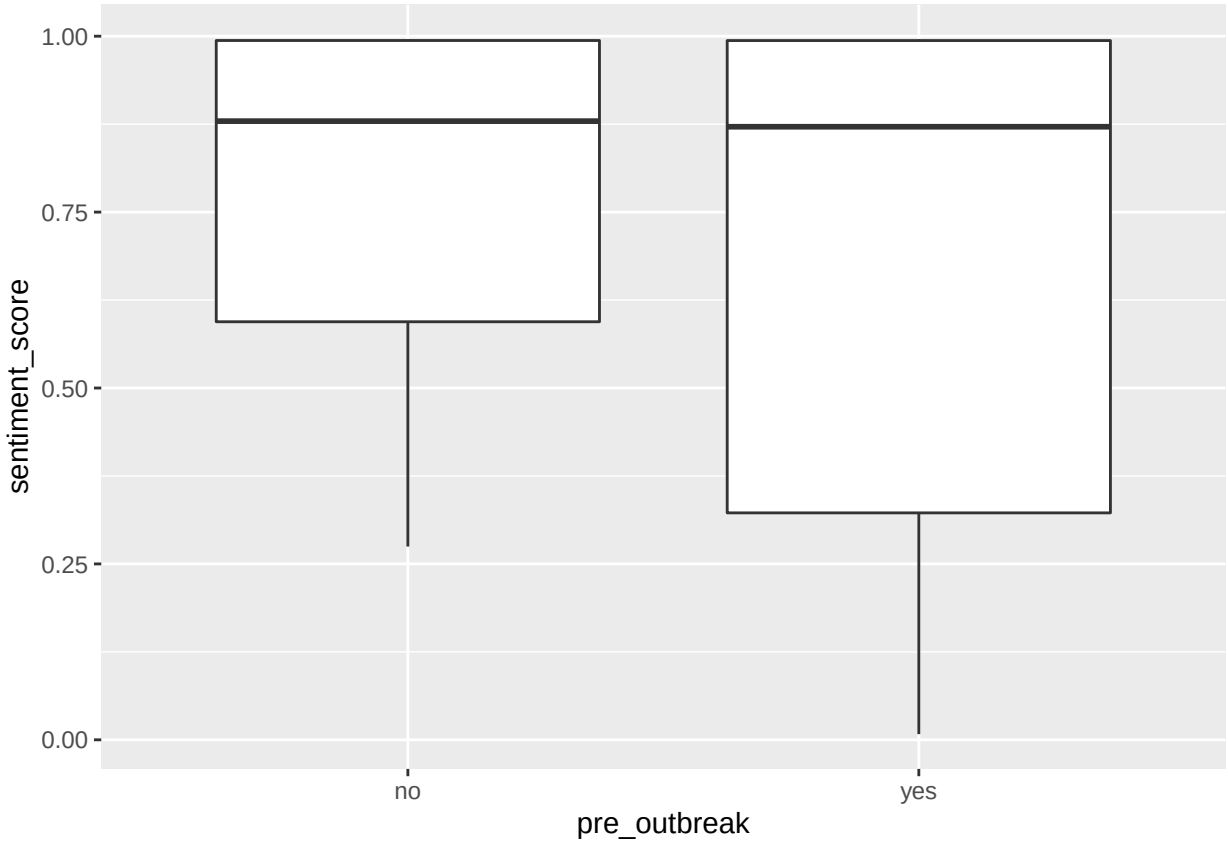
```
ggplot(Australia_analysis_education) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




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

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.732
## 2 yes                         0.652
```

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




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

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.767
## 2 yes                            0.673
```

```
#precovid education
count(Australia_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

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

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

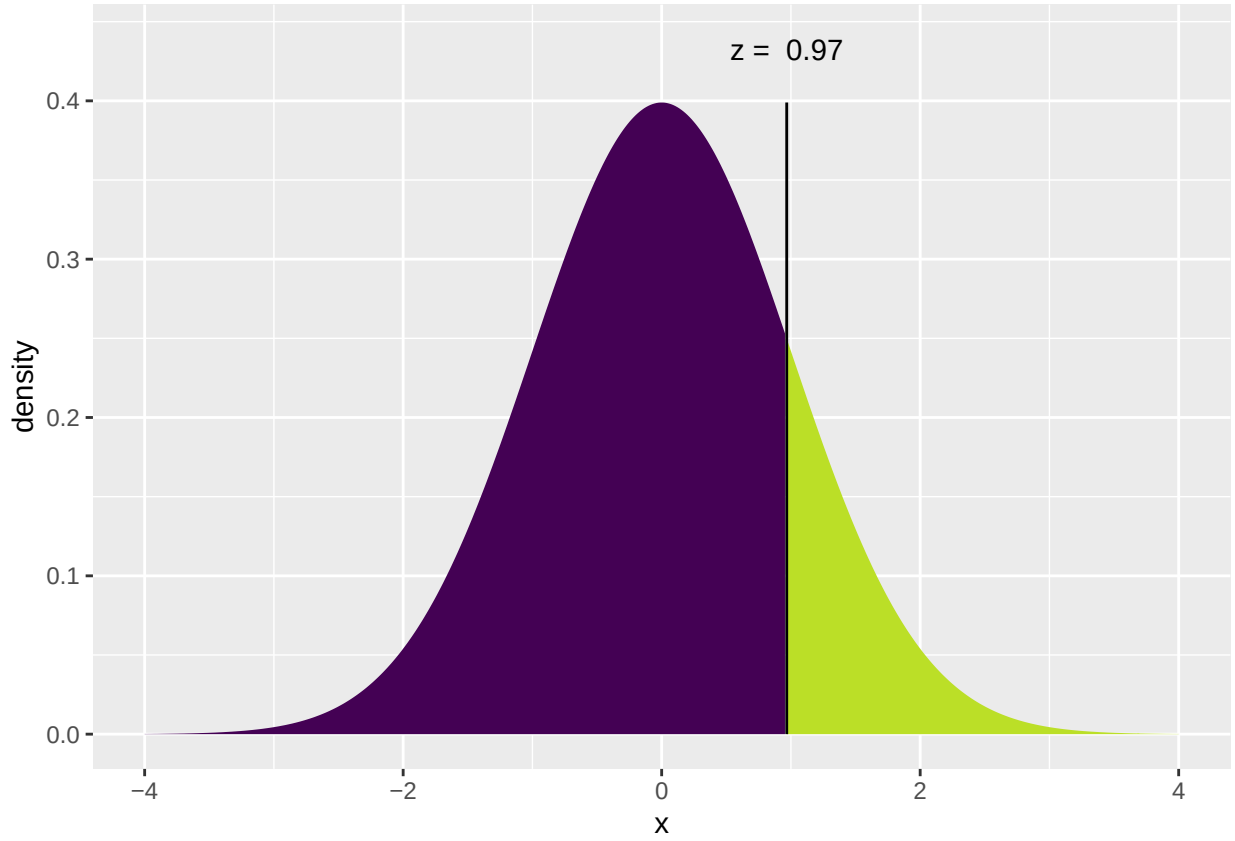
```
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      21
## 2 TRUE                       39
```

```
#proportion of positive sentiment videos precovid from sample
p_hat1 = 39/60

Australia_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                      19
## 2 TRUE                       51
```

```
#proportion of positive sentiment videos postcovid from sample
p_hat2 = 51/70

p_hat = (39+51)/(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 <= 0.9676) = P(Z <= 0.9676) = 0.8334
##  P(X >  0.9676) = P(Z >  0.9676) = 0.1666
##
```




```
## [1] 0.3332287
```

```
#outbreak education
count(Australia_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

Australia_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                      34
## 2 TRUE                       66
```

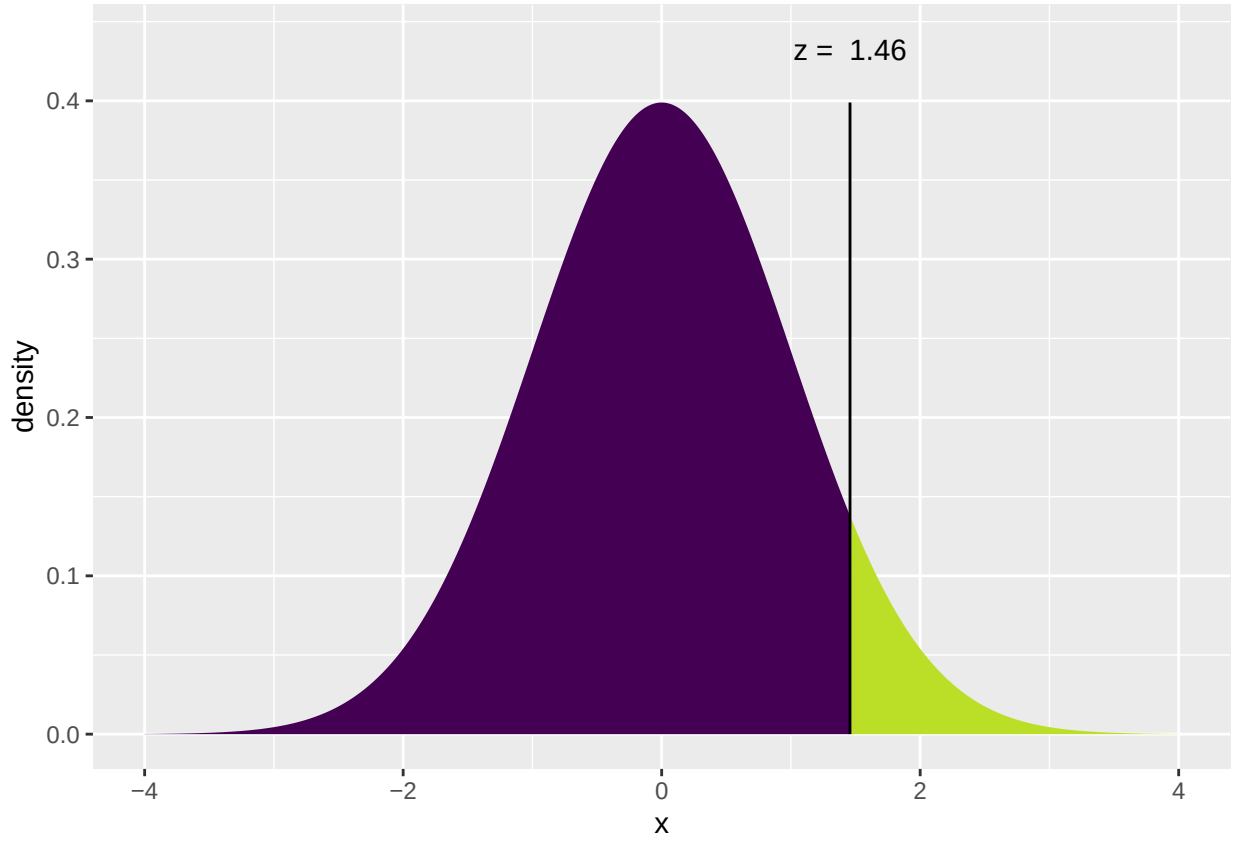
```
#proportion of positive sentiment videos preoutbreak from sample
p_hat1 = 66/100
```

```r
Australia_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                       6
## 2 TRUE                       24
```

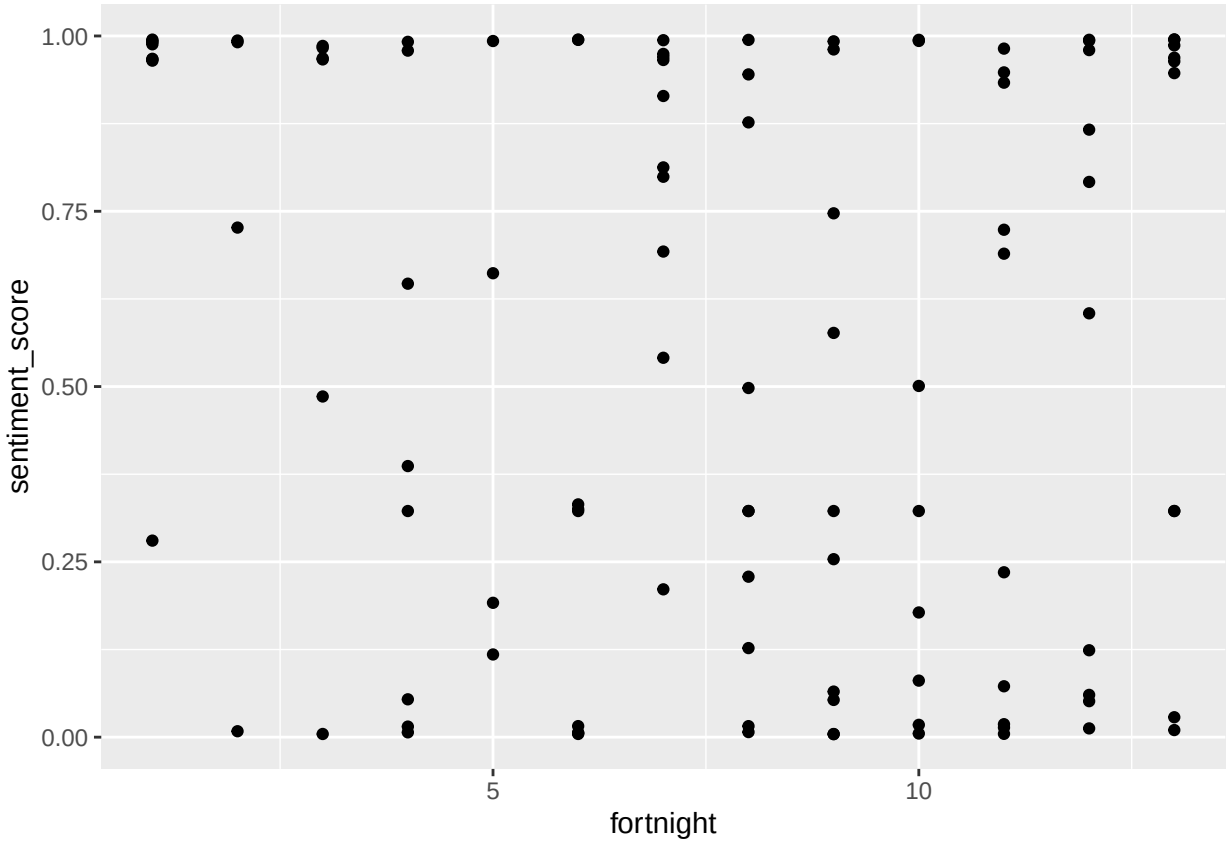
```r
#proportion of positive sentiment videos postoutbreak from sample
p_hat2 = 24/30

p_hat = (66+24)/(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 <= 1.457) = P(Z <= 1.457) = 0.9275
##  P(X >  1.457) = P(Z >  1.457) = 0.07254
##
```

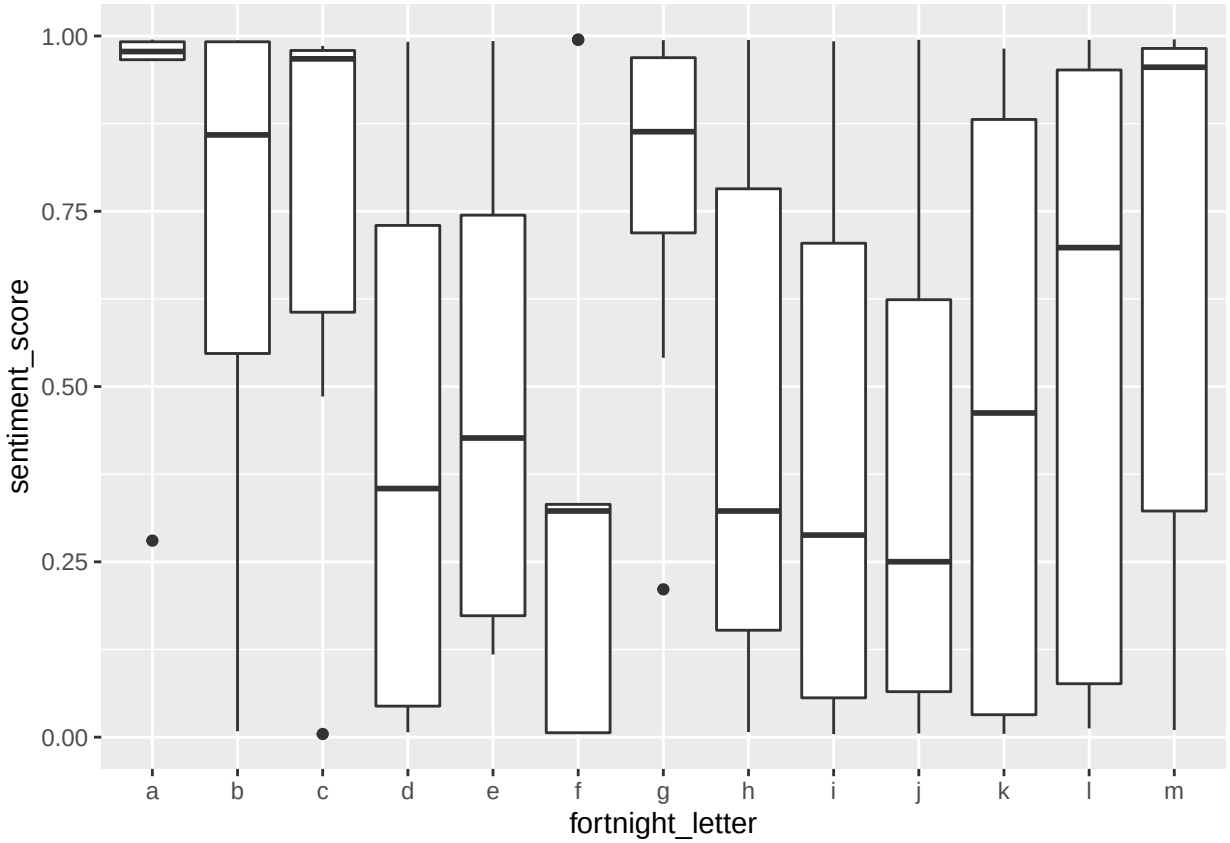



```
## [1] 0.1450705
```

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



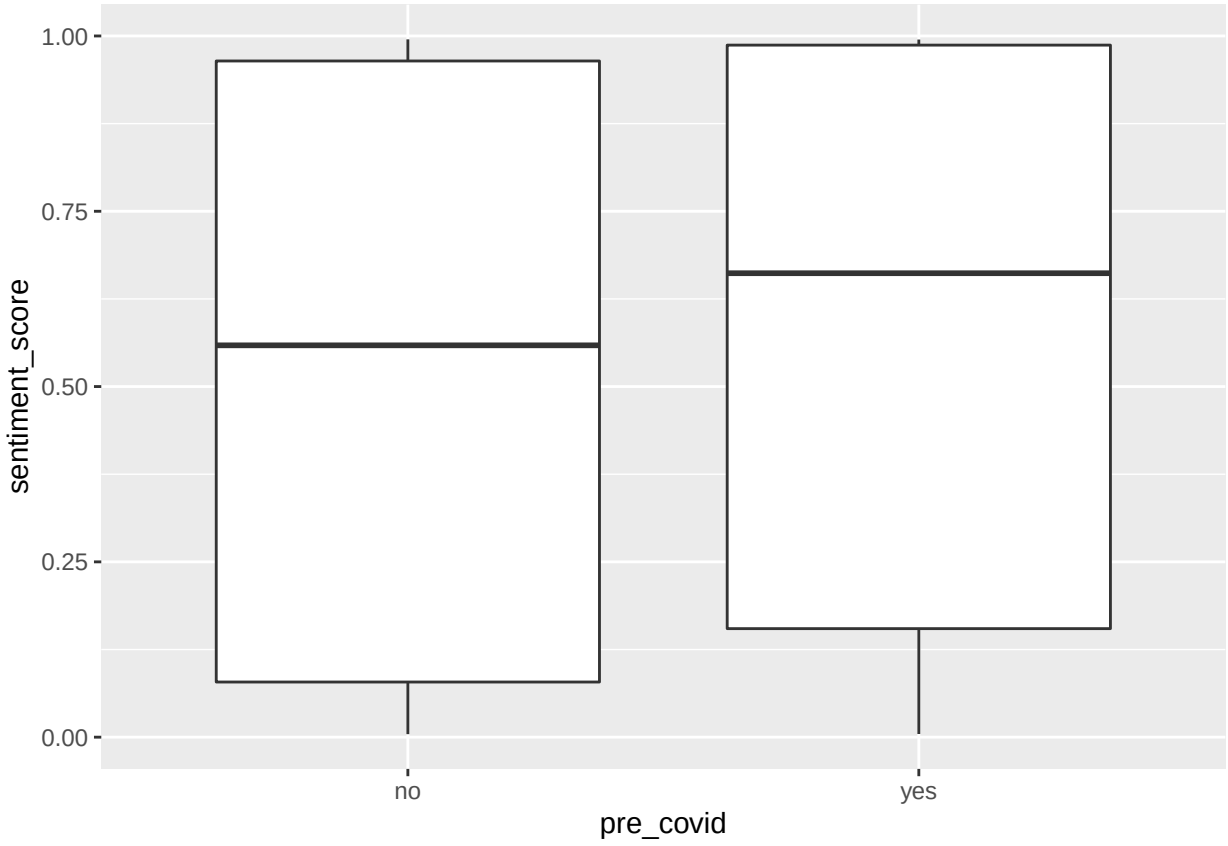

```
ggplot(Australia_analysis_science) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




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

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.893
##  2         2                   0.680
##  3         3                   0.732
##  4         4                   0.425
##  5         5                   0.491
##  6         6                   0.334
##  7         7                   0.787
##  8         8                   0.434
##  9         9                   0.400
## 10        10                   0.387
## 11        11                   0.462
## 12        12                   0.548
## 13        13                   0.654
```

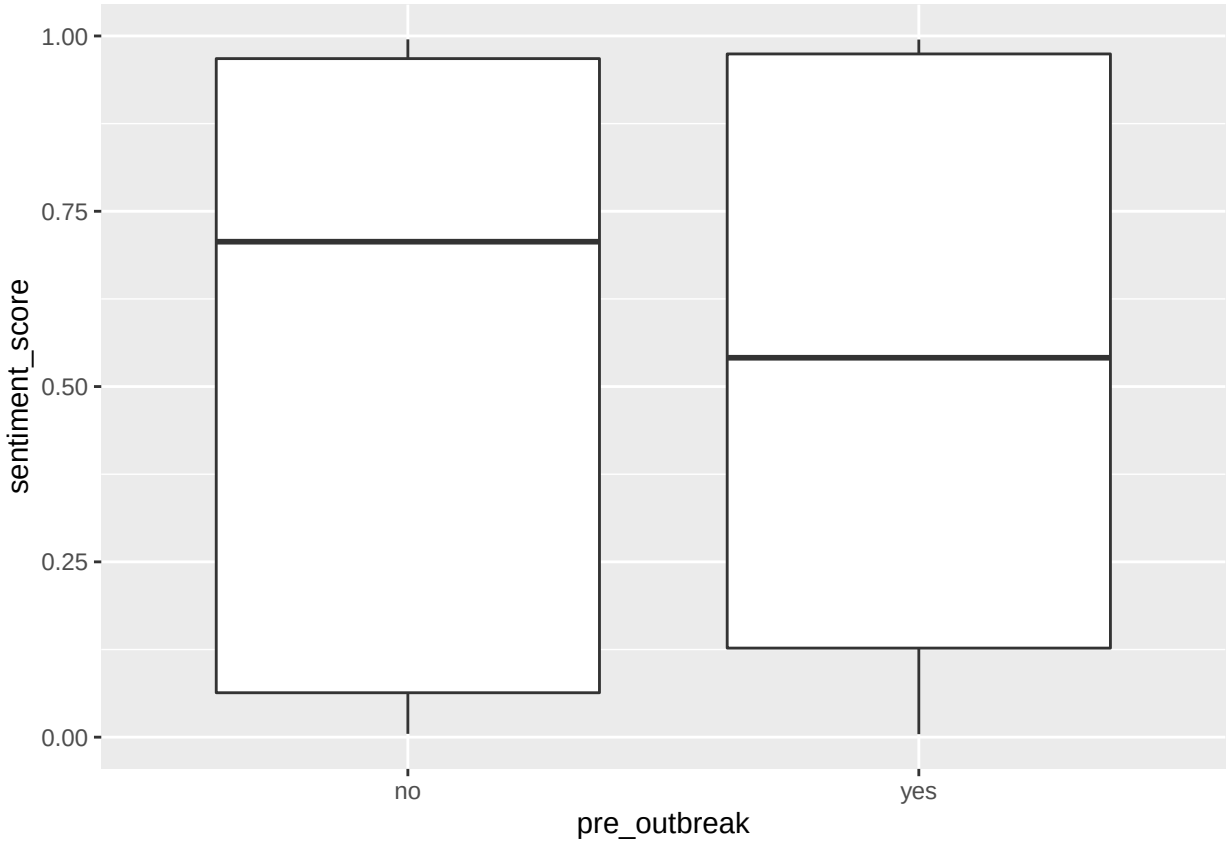
```
ggplot(Australia_analysis_science) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




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

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.529
## 2 yes                         0.580
```

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




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

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.555
## 2 yes                            0.545
```

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

```
## # A tibble: 2 x 2
##   `pre_covid == "yes"`     n
##   <lgl>                <int>
## 1 FALSE                   68
## 2 TRUE                    39
```

```
num_precovid = 39
num_postcovid = 68
num = 107

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

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

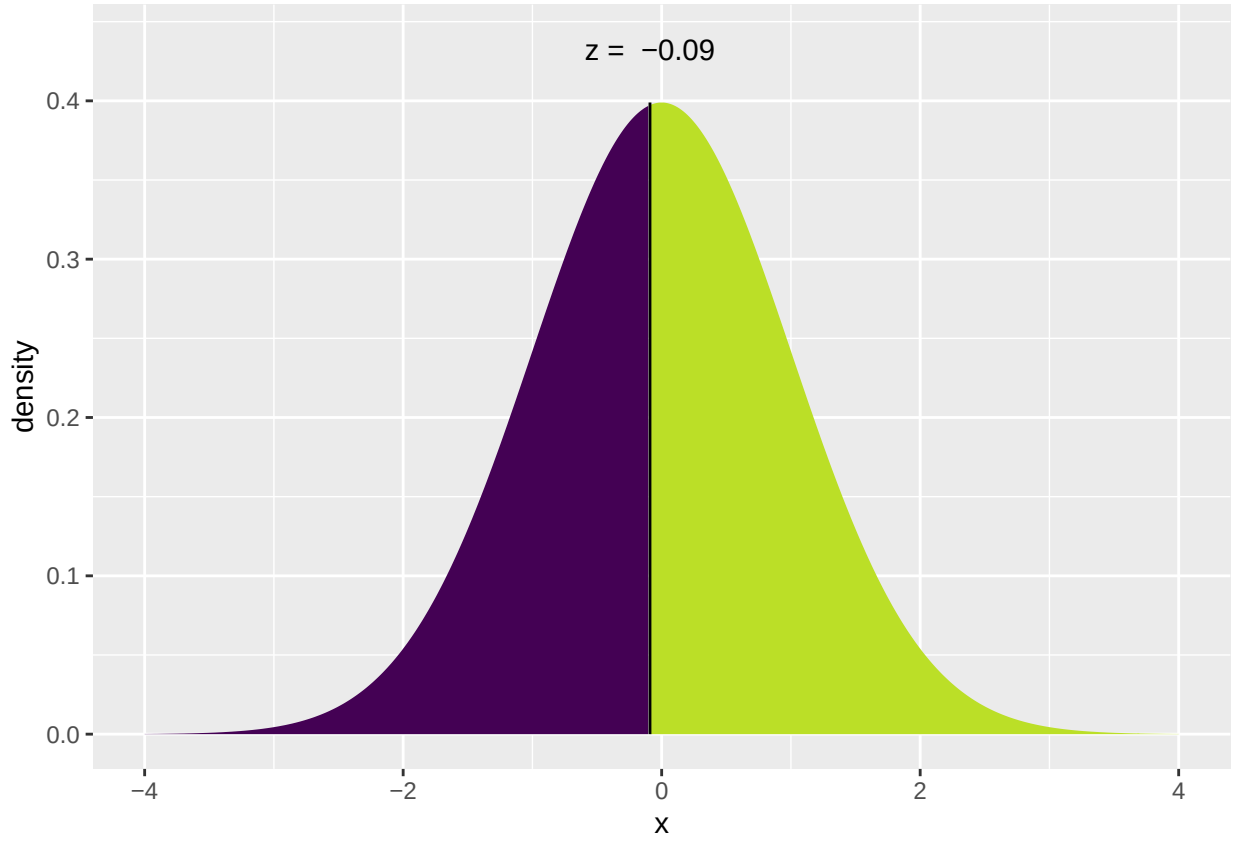
```
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      18
## 2 TRUE                       21
```

```
#proportion of positive sentiment videos precovid from sample
p_hat1 = 21/39

Australia_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                      32
## 2 TRUE                       36
```

```
#proportion of positive sentiment videos postcovid from sample
p_hat2 = 36/68

p_hat = (21+36)/(39+68)

sd <- sqrt((((p_hat)*(1-p_hat))/39)+(((p_hat)*(1-p_hat))/68))
z_score <- ((p_hat2-p_hat1)-0)/sd

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

```
##
## If X ~ N(0, 1), then
##  P(X <= -0.0903) = P(Z <= -0.0903) = 0.464
##  P(X >  -0.0903) = P(Z >  -0.0903) = 0.536
##
```




```
## [1] 0.9280478
```

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

```
## # A tibble: 2 x 2
##   `pre_outbreak == "yes"`     n
##   <lgl>                   <int>
## 1 FALSE                      30
## 2 TRUE                       77
```

```
num_preoutbreak = 77
num_postoutbreak = 30
num = 107

Australia_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                      37
## 2 TRUE                       40
```

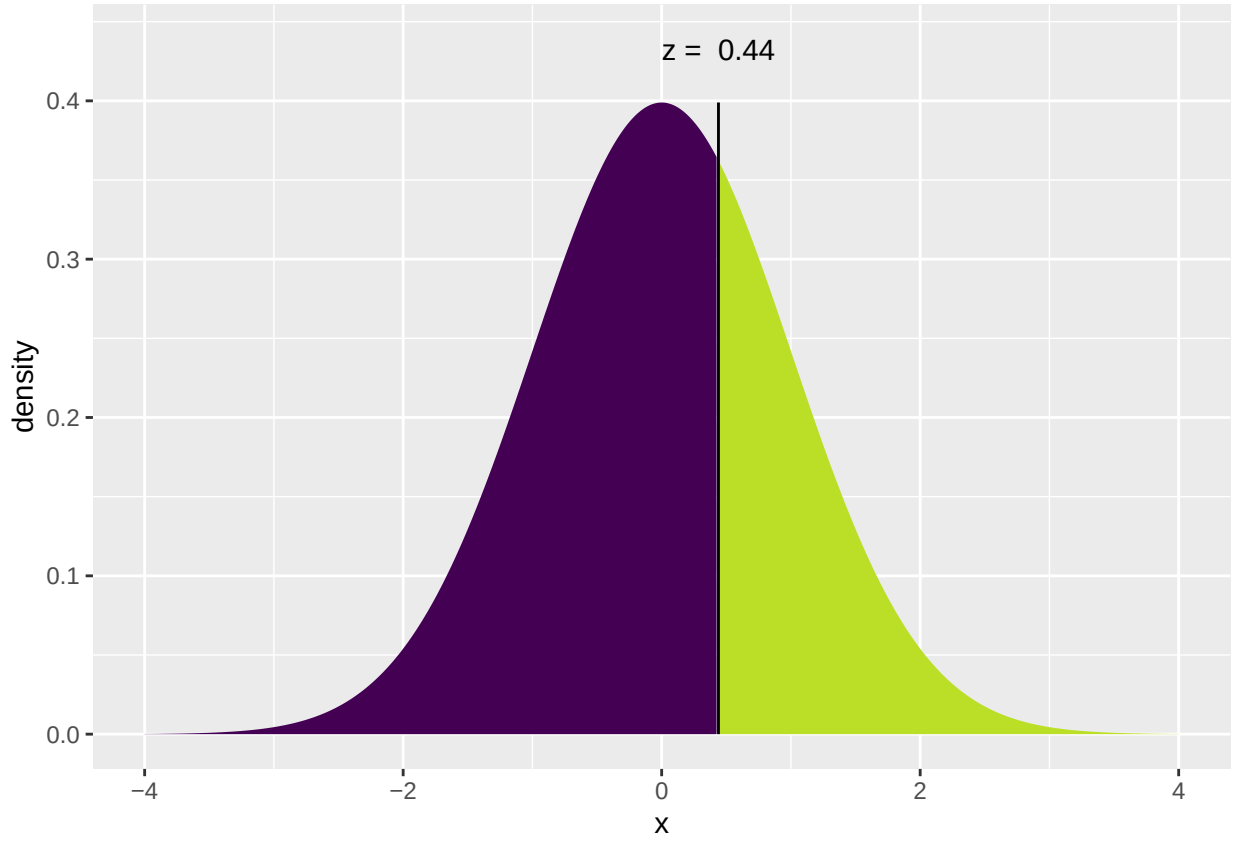
```
#proportion of positive sentiment videos preoutbreak from sample
p_hat1 = 40/77
```

```
Australia_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                      13
## 2 TRUE                       17
```

```
#proportion of positive sentiment videos postoutbreak from sample
p_hat2 = 17/30

p_hat = (40+17)/(77+30)

sd <- sqrt((((p_hat)*(1-p_hat))/77)+(((p_hat)*(1-p_hat))/30))
z_score <- ((p_hat2-p_hat1)-0)/sd

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

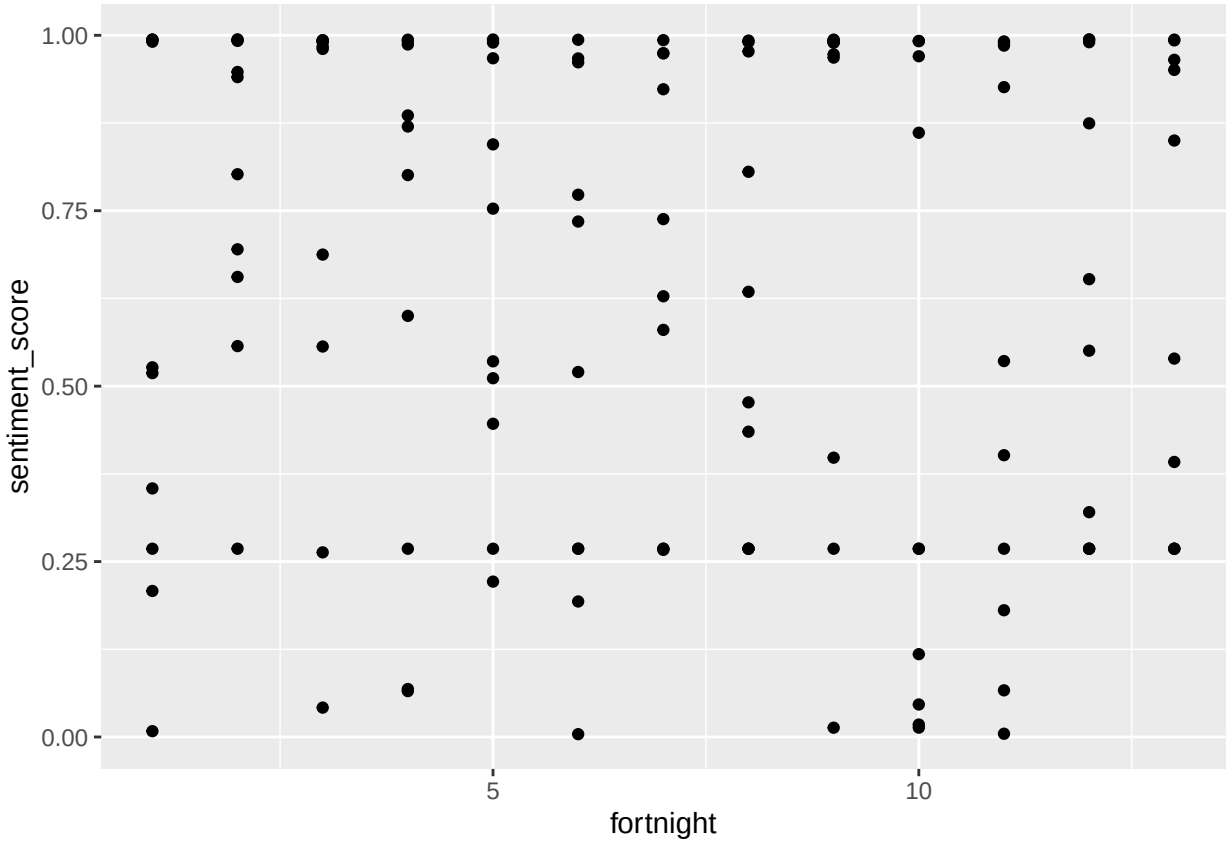
```
##
## If X ~ N(0, 1), then
##  P(X <= 0.4394) = P(Z <= 0.4394) = 0.6698
##  P(X >  0.4394) = P(Z >  0.4394) = 0.3302
##
```



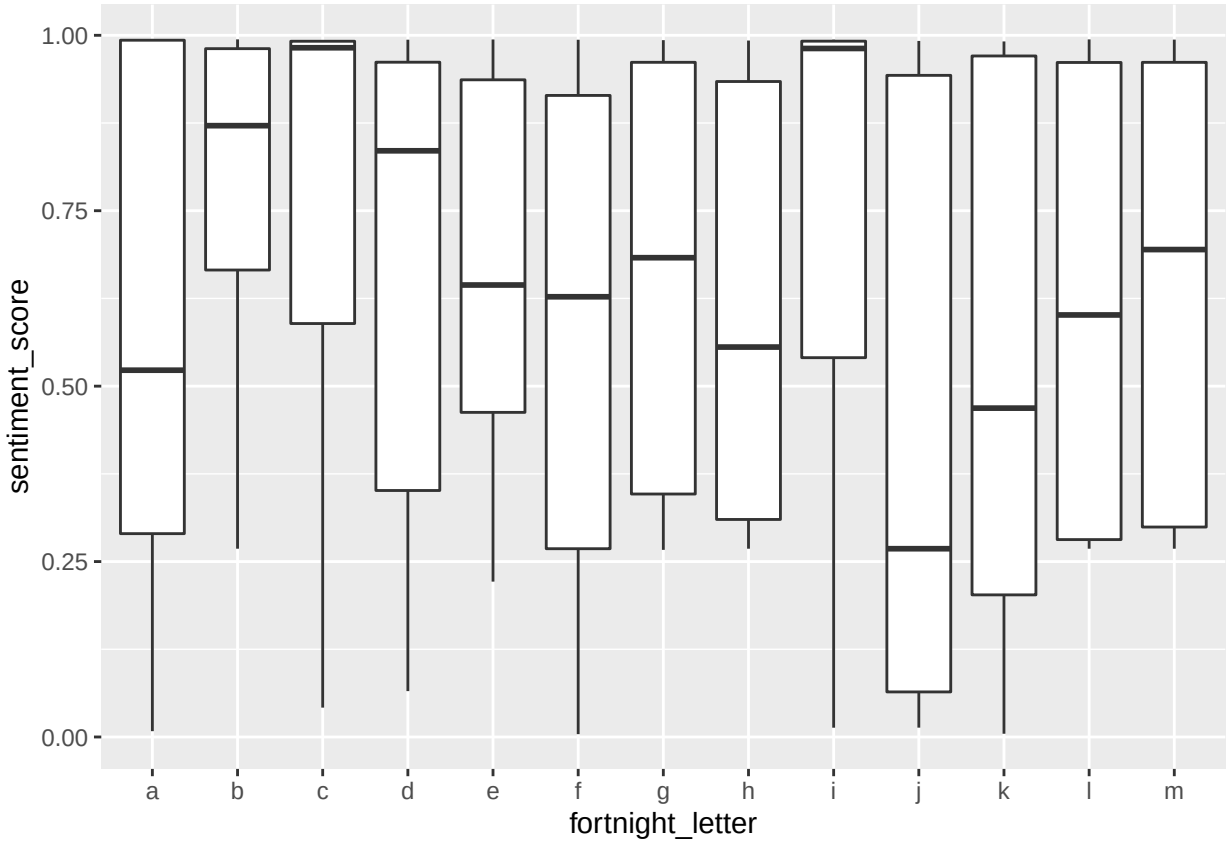

```
## [1] 0.6603498
```

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

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



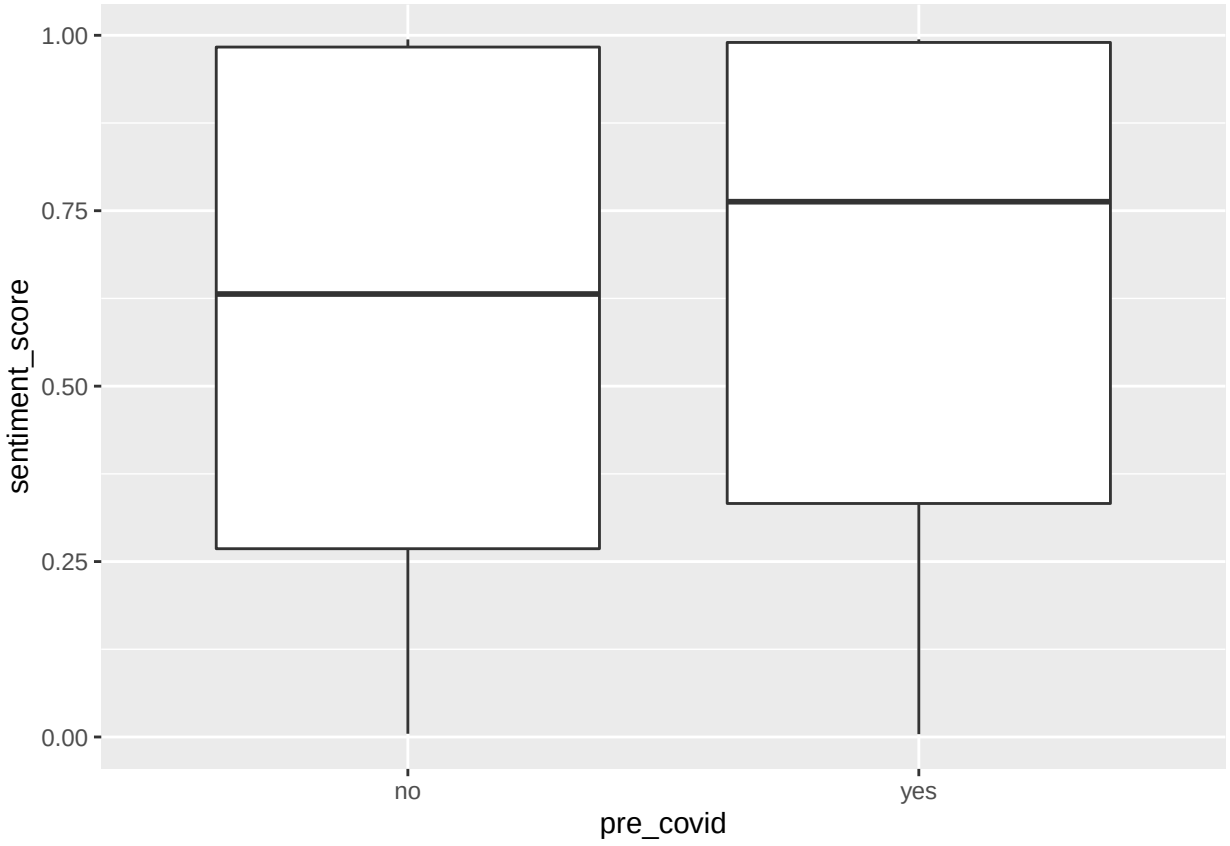

```
ggplot(Australia_analysis_all) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




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

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.586
##  2         2                   0.785
##  3         3                   0.748
##  4         4                   0.653
##  5         5                   0.653
##  6         6                   0.568
##  7         7                   0.661
##  8         8                   0.612
##  9         9                   0.758
## 10        10                   0.455
## 11        11                   0.535
## 12        12                   0.618
## 13        13                   0.649
```

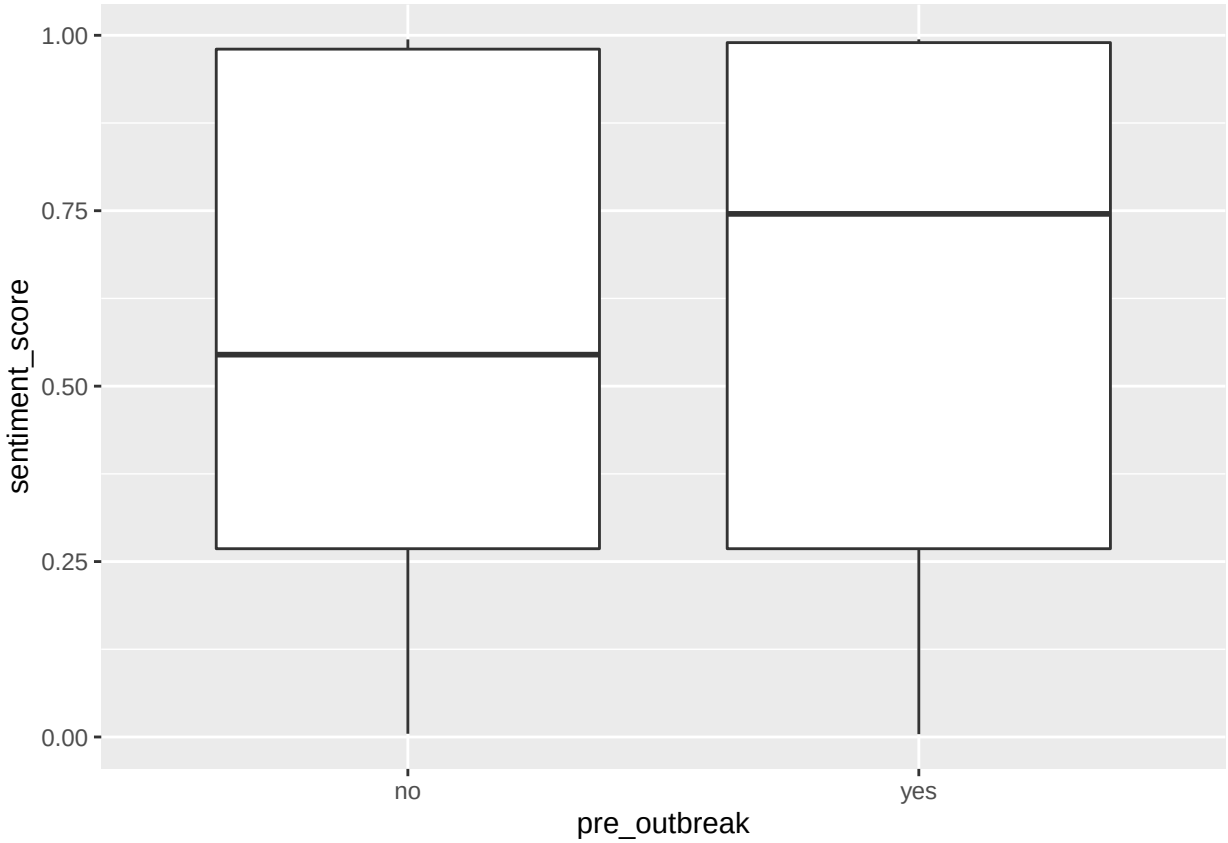
```
ggplot(Australia_analysis_all) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




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

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.613
## 2 yes                         0.665
```

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




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

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.601
## 2 yes                            0.648
```

```
#precovid all cateogires
count(Australia_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

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

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

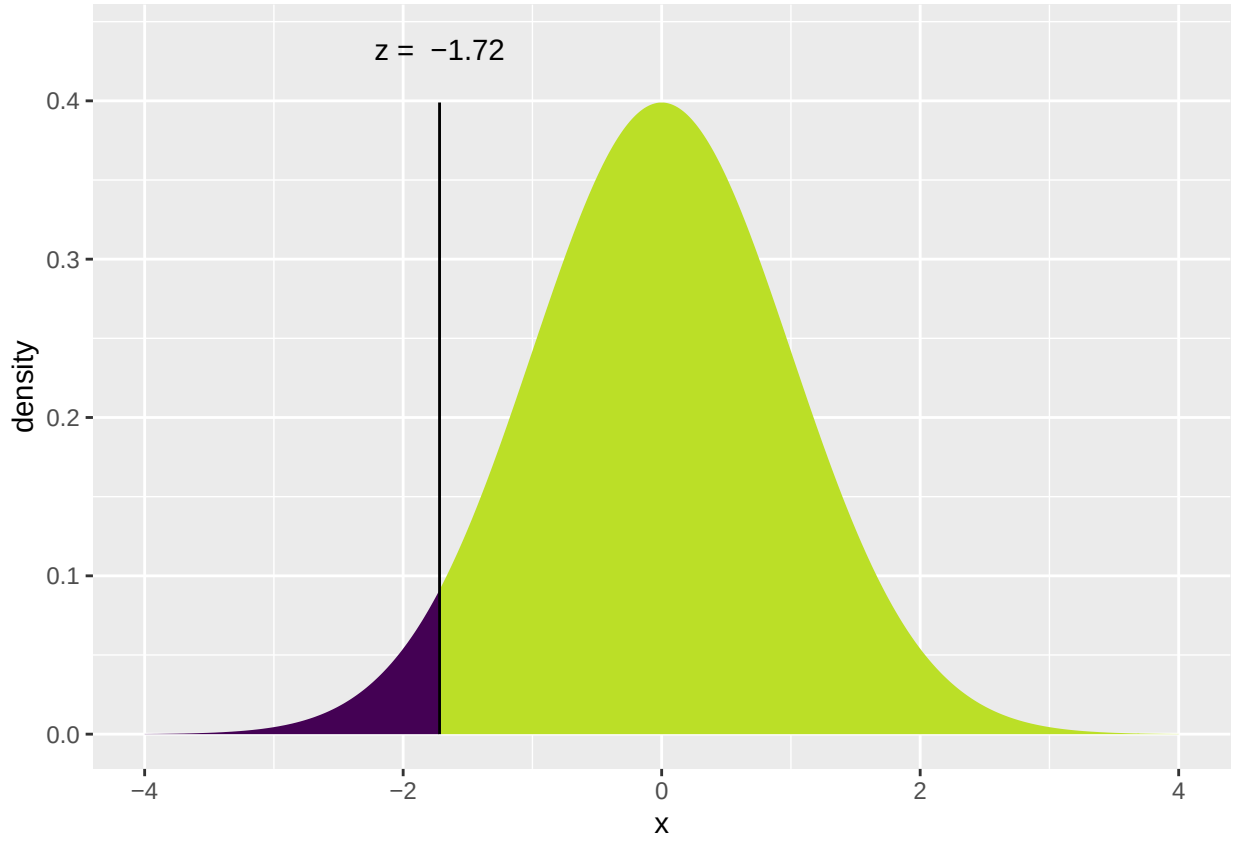
```
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      17
## 2 TRUE                       43
```

```
#proportion of positive sentiment videos precovid from sample
p_hat1 = 43/60

Australia_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                      30
## 2 TRUE                       40
```

```
#proportion of positive sentiment videos postcovid from sample
p_hat2 = 40/70

p_hat = (43+40)/(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 <= -1.718) = P(Z <= -1.718) = 0.04287
##  P(X >  -1.718) = P(Z >  -1.718) = 0.9571
##
```




```
## [1] 0.08574919
```

```
#outbreak all categories
count(Australia_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

Australia_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
```

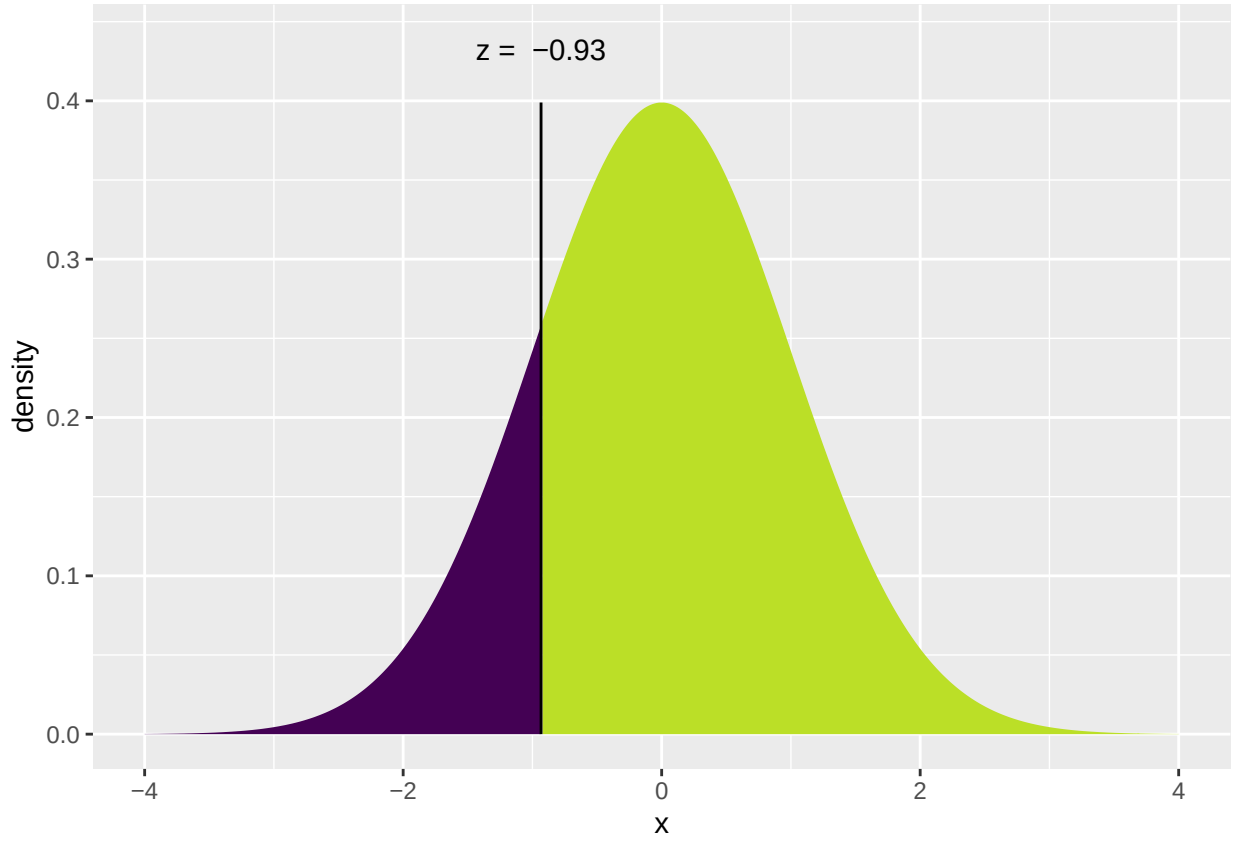
```
#proportion of positive sentiment videos preoutbreak from sample
p_hat1 = 66/100
```

```r
Australia_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                      13
## 2 TRUE                       17
```

```r
#proportion of positive sentiment videos postoutbreak from sample
p_hat2 = 17/30

p_hat = (66+17)/(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 <= -0.9332) = P(Z <= -0.9332) = 0.1754
##  P(X >  -0.9332) = P(Z >  -0.9332) = 0.8246
##
```




```
## [1] 0.3507096
```

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

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = 0.67385, df = 58.451, p-value = 0.5031
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1113696  0.2244293
## sample estimates:
##  mean in group no mean in group yes
##         0.5801963         0.5236665
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = 1.6467, df = 54.632, p-value = 0.1054
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##  -0.02978651  0.30407453
## sample estimates:
##  mean in group no mean in group yes
##          0.659269           0.522125
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = 0.29545, df = 122.39, p-value = 0.7682
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.09917629  0.13397387
## sample estimates:
##  mean in group no mean in group yes
##         0.6779775         0.6605787
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = 0.57251, df = 45.347, p-value = 0.5698
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1038963  0.1864422
## sample estimates:
##  mean in group no mean in group yes
##         0.7016933         0.6604203
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = -2.3064, df = 128, p-value = 0.0227
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.24820800 -0.01898142
## sample estimates:
##  mean in group no mean in group yes
##         0.6051667         0.7387614
```

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

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

```
## data:  sentiment_score by pre_outbreak
## t = -1.9295, df = 42.739, p-value = 0.06033
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.297438792  0.006599599
## sample estimates:
##  mean in group no mean in group yes
##         0.5549646         0.7003842
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = -0.058152, df = 118.36, p-value = 0.9537
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1291382  0.1217699
## sample estimates:
##  mean in group no mean in group yes
##         0.5873455         0.5910296
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = -0.75926, df = 46.259, p-value = 0.4515
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.21161380  0.09568422
## sample estimates:
##  mean in group no mean in group yes
##         0.5451958         0.6031606
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = -1.0337, df = 92.495, p-value = 0.304
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.20744376  0.06541624
## sample estimates:
##  mean in group no mean in group yes
##         0.5773666         0.6483803
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = 0.16818, df = 51.849, p-value = 0.8671
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1370156  0.1620812
## sample estimates:
##  mean in group no mean in group yes
##         0.6154144         0.6028816
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = 0.79522, df = 124.66, p-value = 0.428
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.06797416  0.15928481
## sample estimates:
##  mean in group no mean in group yes
##         0.6202153         0.5745600
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = 1.2487, df = 44.15, p-value = 0.2184
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.05475695  0.23315705
## sample estimates:
##  mean in group no mean in group yes
##         0.6677591         0.5785590
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = 1.328, df = 117.48, p-value = 0.1867
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.0393781  0.1997067
## sample estimates:
##  mean in group no mean in group yes
##         0.7318515         0.6516873
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = 1.6007, df = 68.605, p-value = 0.114
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.02299678  0.20961721
## sample estimates:
##  mean in group no mean in group yes
##         0.7666297         0.6733195
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = -0.62479, df = 76.604, p-value = 0.534
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2161129  0.1128905
## sample estimates:
##  mean in group no mean in group yes
##         0.5285962         0.5802074
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = 0.11212, df = 50.503, p-value = 0.9112
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1702012  0.1903311
## sample estimates:
##  mean in group no mean in group yes
##         0.5546508         0.5445858
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = -0.87407, df = 126.84, p-value = 0.3837
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.17292274  0.06696311
```

```
## sample estimates:
##  mean in group no mean in group yes
##         0.6125099         0.6654897
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = -0.65483, df = 47.741, p-value = 0.5157
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.19234172  0.09784618
## sample estimates:
##  mean in group no mean in group yes
##         0.6006177         0.6478654
```