Ass3

Yukta Chavan

2023-03-15

```
pulitzer<- read.csv("E:/taming/pulitzer.csv")
pulitzer
```

```
##                         newspaper circ_2004 circ_2013 change_0413 prizes_9014
## 1                       USA Today   2192098   1674306        -24%           3
## 2             Wall Street Journal   2101017   2378827         13%          51
## 3                  New York Times   1119027   1865318         67%         118
## 4               Los Angeles Times    983727    653868        -34%          86
## 5                 Washington Post    760034    474767        -38%         101
## 6             New York Daily News    712671    516165        -28%           7
## 7                   New York Post    642844    500521        -22%           1
## 8                 Chicago Tribune    603315    414930        -31%          39
## 9           San Jose Mercury News    558874    583998          4%           7
## 10                        Newsday    553117    377744        -32%          19
## 11              Houston Chronicle    549300    360251        -34%           6
## 12            Dallas Morning News    528379    409265        -23%          18
## 13        San Francisco Chronicle    499008    218987        -56%          10
## 14               Arizona Republic    466926    293640        -37%           8
## 15              Chicago Sun-Times    453757    470548          4%           3
## 16                   Boston Globe    446241    245572        -45%          42
## 17   Atlanta Journal Constitution    409873    231094        -44%           7
## 18             Newark Star Ledger    395000    340778        -14%           9
## 19             Detroit Free Press    379304    209652        -45%          13
## 20       Minneapolis Star Tribune    377058    301345        -20%           9
## 21          Philadelphia Inquirer    376454    306831        -18%          33
## 22         Cleveland Plain Dealer    367528    311605        -15%          12
## 23        San Diego Union-Tribune    355771    250678        -30%           3
## 24                Tampa Bay Times    348502    340260         -2%          22
## 25                    Denver Post    340168    416676         22%          10
## 26            Rocky Mountain News    340007         0       -100%           6
## 27                      Oregonian    339169    228909        -33%          18
## 28                   Miami Herald    325032    147130        -55%          25
## 29         Orange County Register    310001    356165         15%           6
## 30                 Sacramento Bee    303841    200802        -34%           9
## 31        St. Louis Post-Dispatch    281198    167199        -41%           8
## 32                  Baltimore Sun    277947    177054        -36%          14
## 33               Kansas City Star    275747    189283        -31%           3
## 34                   Detroit News    271465    115643        -57%           5
## 35               Orlando Sentinel    269269    161070        -40%           8
## 36     South Florida Sun-Sentinel    268297    163728        -39%           2
## 37     New Orleans Times-Picayune    262008         0       -100%           9
## 38            Columbus Dispatch    259127    137148        -47%           2
## 39            Indianapolis Star    253778    156850        -38%           2
## 40     San Antonio Express-News    246057    139005        -44%           1
## 41      Pittsburgh Post-Gazette    242514    180433        -26%           4
## 42   Milwaukee Journal Sentinel    241605    198469        -18%          11
## 43                Tampa Tribune    238877    191477        -20%           1
## 44      Fort Woth Star-Telegram    237318    188593        -21%           2
## 45                Boston Herald    236899     95929        -60%           1
```

##Question 1(a):

```
pulitzer_1<- pulitzer %>%
  mutate(change_0413 = str_replace(change_0413, "%", "") %>% as.integer())
pulitzer_1
```

```
##                         newspaper circ_2004 circ_2013 change_0413 prizes_9014
## 1                       USA Today   2192098   1674306         -24           3
## 2             Wall Street Journal   2101017   2378827          13          51
## 3                  New York Times   1119027   1865318          67         118
## 4               Los Angeles Times    983727    653868         -34          86
## 5                 Washington Post    760034    474767         -38         101
## 6             New York Daily News    712671    516165         -28           7
## 7                   New York Post    642844    500521         -22           1
## 8                 Chicago Tribune    603315    414930         -31          39
## 9           San Jose Mercury News    558874    583998           4           7
## 10                        Newsday    553117    377744         -32          19
## 11              Houston Chronicle    549300    360251         -34           6
## 12            Dallas Morning News    528379    409265         -23          18
## 13        San Francisco Chronicle    499008    218987         -56          10
## 14               Arizona Republic    466926    293640         -37           8
## 15              Chicago Sun-Times    453757    470548           4           3
## 16                   Boston Globe    446241    245572         -45          42
## 17 Atlanta Journal Constitution    409873    231094         -44           7
## 18             Newark Star Ledger    395000    340778         -14           9
## 19             Detroit Free Press    379304    209652         -45          13
## 20       Minneapolis Star Tribune    377058    301345         -20           9
## 21          Philadelphia Inquirer    376454    306831         -18          33
## 22         Cleveland Plain Dealer    367528    311605         -15          12
## 23        San Diego Union-Tribune    355771    250678         -30           3
## 24                Tampa Bay Times    348502    340260          -2          22
## 25                    Denver Post    340168    416676          22          10
## 26             Rocky Mountain News    340007         0        -100           6
## 27                      Oregonian    339169    228909         -33          18
## 28                   Miami Herald    325032    147130         -55          25
## 29         Orange County Register    310001    356165          15           6
## 30                 Sacramento Bee    303841    200802         -34           9
## 31        St. Louis Post-Dispatch    281198    167199         -41           8
## 32                  Baltimore Sun    277947    177054         -36          14
## 33               Kansas City Star    275747    189283         -31           3
## 34                   Detroit News    271465    115643         -57           5
## 35               Orlando Sentinel    269269    161070         -40           8
## 36     South Florida Sun-Sentinel    268297    163728         -39           2
## 37     New Orleans Times-Picayune    262008         0        -100           9
## 38            Columbus Dispatch    259127    137148         -47           2
## 39            Indianapolis Star    253778    156850         -38           2
## 40     San Antonio Express-News    246057    139005         -44           1
## 41      Pittsburgh Post-Gazette    242514    180433         -26           4
## 42   Milwaukee Journal Sentinel    241605    198469         -18          11
## 43                Tampa Tribune    238877    191477         -20           1
## 44      Fort Woth Star-Telegram    237318    188593         -21           2
## 45                Boston Herald    236899     95929         -60           1
```

##Question 1(b):

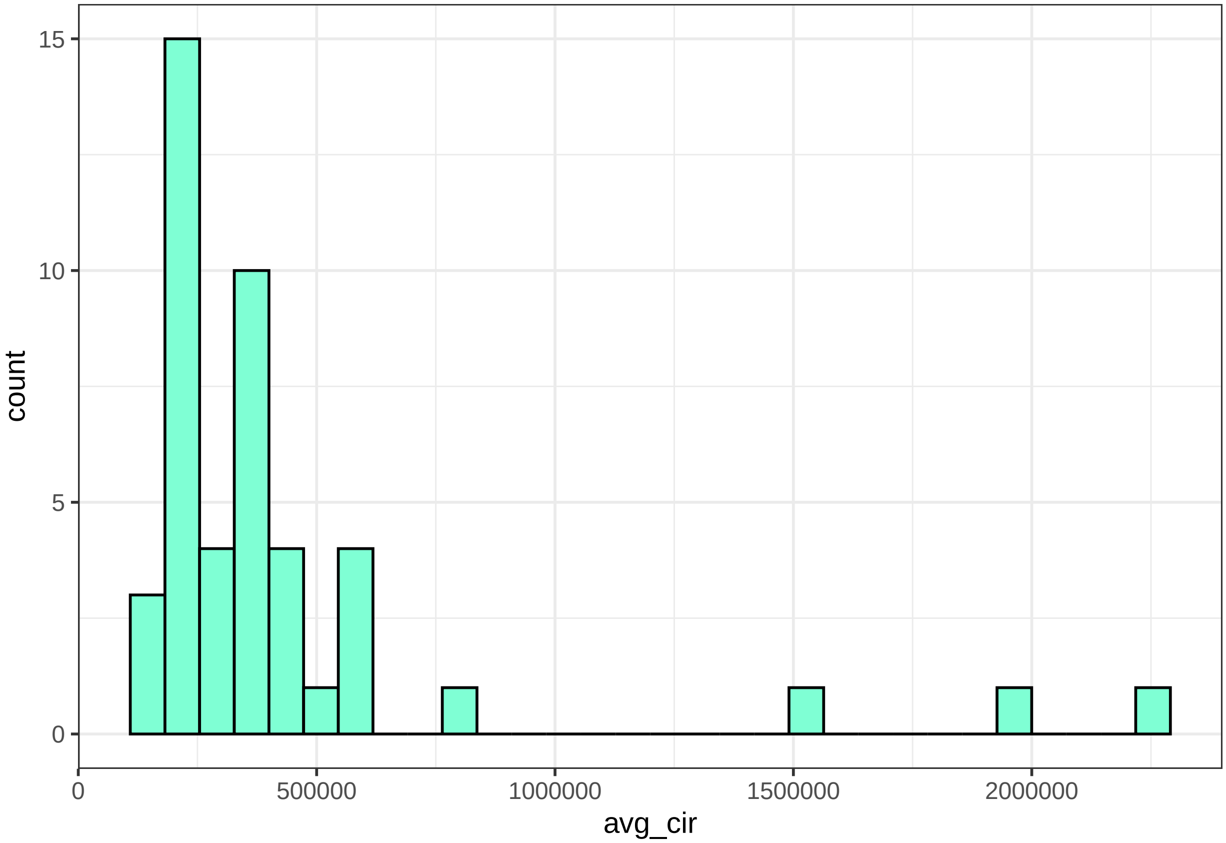
```
pulitzer_1 <- pulitzer_1 %>% mutate(avg_cir = (circ_2004+circ_2013)/2)
pulitzer_1
```

```
##                       newspaper circ_2004 circ_2013 change_0413 prizes_9014
## 1                     USA Today   2192098   1674306         -24           3
## 2           Wall Street Journal   2101017   2378827          13          51
## 3                New York Times   1119027   1865318          67         118
## 4             Los Angeles Times    983727    653868         -34          86
## 5               Washington Post    760034    474767         -38         101
## 6           New York Daily News    712671    516165         -28           7
## 7                 New York Post    642844    500521         -22           1
## 8               Chicago Tribune    603315    414930         -31          39
## 9         San Jose Mercury News    558874    583998           4           7
## 10                      Newsday    553117    377744         -32          19
## 11            Houston Chronicle    549300    360251         -34           6
## 12          Dallas Morning News    528379    409265         -23          18
## 13      San Francisco Chronicle    499008    218987         -56          10
## 14             Arizona Republic    466926    293640         -37           8
## 15            Chicago Sun-Times    453757    470548           4           3
## 16                 Boston Globe    446241    245572         -45          42
## 17 Atlanta Journal Constitution    409873    231094         -44           7
## 18           Newark Star Ledger    395000    340778         -14           9
## 19           Detroit Free Press    379304    209652         -45          13
## 20     Minneapolis Star Tribune    377058    301345         -20           9
## 21        Philadelphia Inquirer    376454    306831         -18          33
## 22       Cleveland Plain Dealer    367528    311605         -15          12
## 23      San Diego Union-Tribune    355771    250678         -30           3
## 24              Tampa Bay Times    348502    340260          -2          22
## 25                  Denver Post    340168    416676          22          10
## 26          Rocky Mountain News    340007         0        -100           6
## 27                    Oregonian    339169    228909         -33          18
## 28                 Miami Herald    325032    147130         -55          25
## 29       Orange County Register    310001    356165          15           6
## 30               Sacramento Bee    303841    200802         -34           9
## 31      St. Louis Post-Dispatch    281198    167199         -41           8
## 32                Baltimore Sun    277947    177054         -36          14
## 33             Kansas City Star    275747    189283         -31           3
## 34                 Detroit News    271465    115643         -57           5
## 35             Orlando Sentinel    269269    161070         -40           8
## 36   South Florida Sun-Sentinel    268297    163728         -39           2
## 37   New Orleans Times-Picayune    262008         0        -100           9
## 38            Columbus Dispatch    259127    137148         -47           2
## 39            Indianapolis Star    253778    156850         -38           2
## 40     San Antonio Express-News    246057    139005         -44           1
## 41      Pittsburgh Post-Gazette    242514    180433         -26           4
## 42   Milwaukee Journal Sentinel    241605    198469         -18          11
## 43                Tampa Tribune    238877    191477         -20           1
## 44      Fort Woth Star-Telegram    237318    188593         -21           2
## 45                Boston Herald    236899     95929         -60           1
##      avg_cir
## 1  1933202.0
## 2  2239922.0
## 3  1492172.5
## 4   818797.5
## 5   617400.5
## 6   614418.0
## 7   571682.5
## 8   509122.5
## 9   571436.0
## 10  465430.5
## 11  454775.5
## 12  468822.0
## 13  358997.5
## 14  380283.0
## 15  462152.5
## 16  345906.5
## 17  320483.5
## 18  367889.0
## 19  294478.0
## 20  339201.5
## 21  341642.5
## 22  339566.5
## 23  303224.5
## 24  344381.0
## 25  378422.0
## 26  170003.5
## 27  284039.0
## 28  236081.0
## 29  333083.0
## 30  252321.5
## 31  224198.5
## 32  227500.5
## 33  232515.0
## 34  193554.0
## 35  215169.5
## 36  216012.5
## 37  131004.0
## 38  198137.5
## 39  205314.0
## 40  192531.0
## 41  211473.5
## 42  220037.0
## 43  215177.0
## 44  212955.5
## 45  166414.0
```

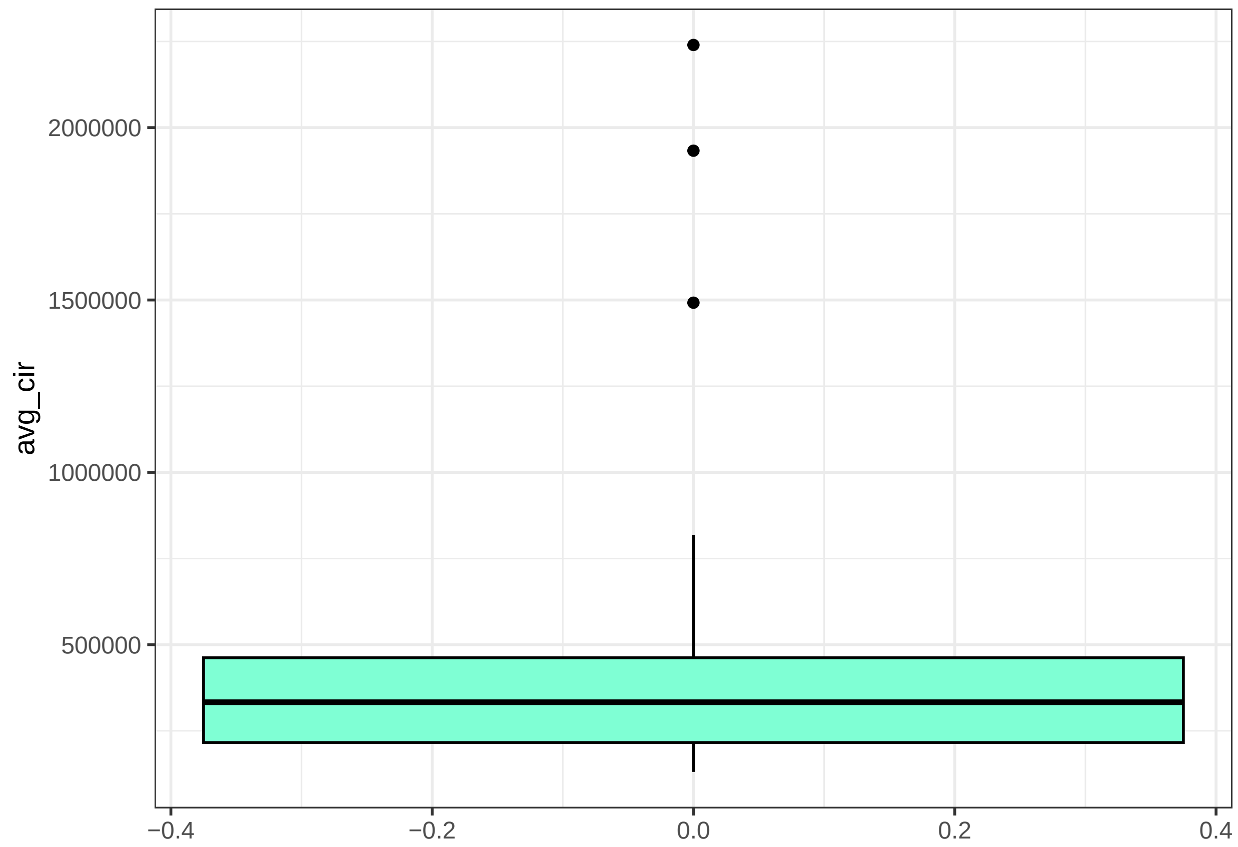
##Question 2(a)

```
ggplot(pulitzer_1, aes(x=avg_cir)) +geom_histogram(fill = "aquamarine", color = "black") +
theme_bw()
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



```
ggplot(pulitzer_1, aes(y=avg_cir)) +geom_boxplot(fill = "aquamarine", color = "black") +theme_bw()
```



```
summary(pulitzer_1$avg_cir)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  131004  216013  333083  437141  462153 2239922
```

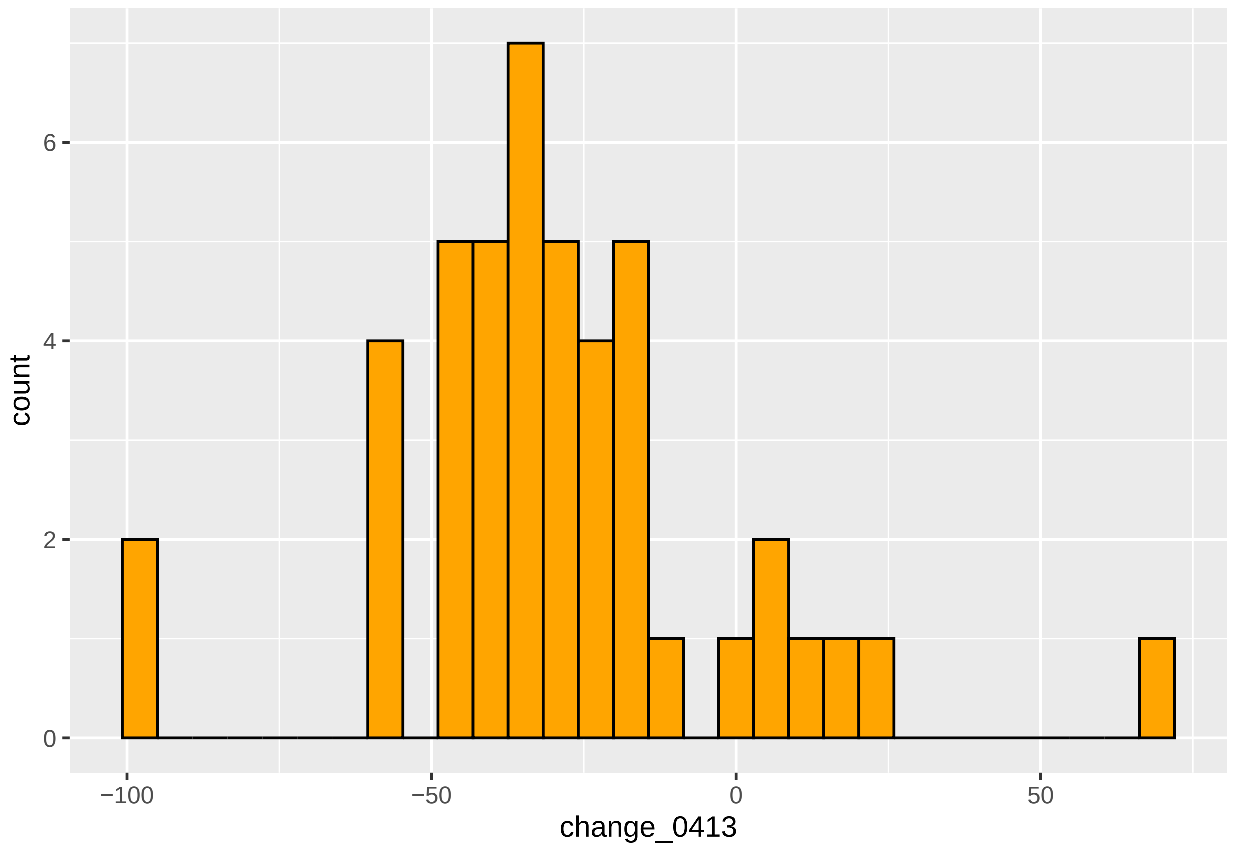
```
sd(pulitzer_1$avg_cir)
```

```
## [1] 425701.9
```

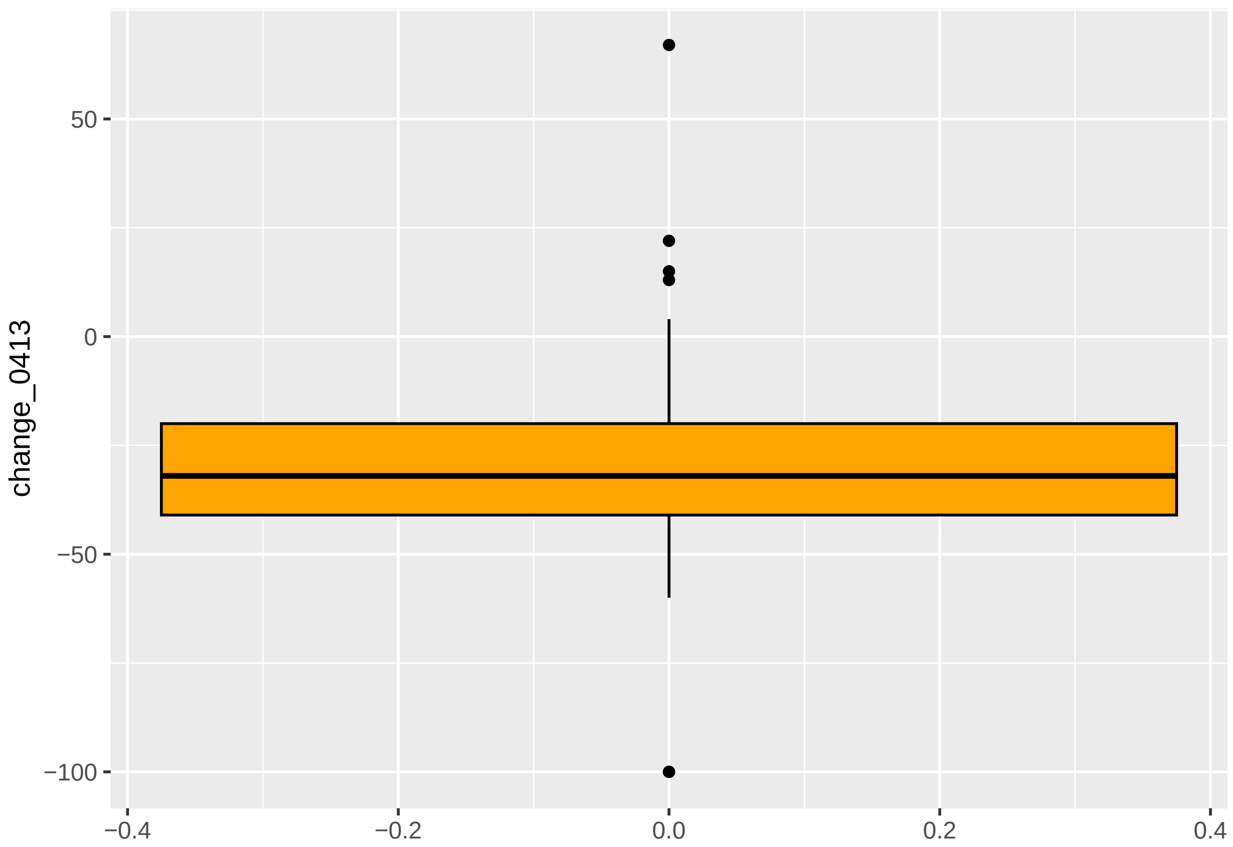
##Ans 2a- Shape: The distribution is rightly skewed and unimodal. ##Location: From box plot, median circulation is near about 299,000. ##Spread:220,000 is the interquartile range. ##Outliers: there are 4 outliers which are - 800,000, 1.5 million, 2 million and 2.25 million. ###Question 2b

```
ggplot(pulitzer_1, aes(x = change_0413)) +geom_histogram(color = "black", fill = "orange")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
ggplot(pulitzer_1, aes(y=change_0413)) +geom_boxplot(fill = "orange", color = "black")
```



```
summary(pulitzer_1$change_0413)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -100.00  -41.00  -32.00  -29.04  -20.00   67.00
```

```
sd(pulitzer_1$change_0413)
```

```
## [1] 28.08263
```

##Ans 2b-Shape: The distribution is symmetrical with slightly right skewness, and unimodal. ##Location: From the graph,-32.5% is change in the median while -29.2% is mean. ##Spread: 20.75% is IQR (from box plot or summary).sd is 27%. ##Outliers: There are 3 outliers. 1 at 67% and 2 at -100%.

##Ans 2c- As change_0413 is roughly symmetrical,it doesn't need a log transform. avg_cir can be transformed to resolve the skewness.

##Question 3(a)-

```
pu_cir <- lm(log(avg_cir) ~ prizes_9014, data=pulitzer_1)
summary(pu_cir)
```

```
##
## Call:
## lm(formula = log(avg_cir) ~ prizes_9014, data = pulitzer_1)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8573 -0.3249 -0.1005  0.1752  1.9141
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.520712   0.092499 135.361  < 2e-16 ***
## prizes_9014  0.013288   0.003017   4.405 6.91e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.5137 on 43 degrees of freedom
## Multiple R-squared:  0.3109, Adjusted R-squared:  0.2949
## F-statistic:  19.4 on 1 and 43 DF,  p-value: 6.91e-05
```

```
exp(pu_cir$coefficients[1])
```

```
## (Intercept)
##    273953.1
```

##Slope= 0.014083, intercept=12.463142. ##Interpretation of intercept: Given a newspaper's log circulation of 12.46 at the end of a 25-year period in which it has won no Pulitzer Prizes, we expect it to have an actual circulation of 258,627, which translates to a log circulation of 0 for the newspaper. ##Interpretation of slope: If a newspaper wins 1 more Pulitzer Awards over a 25-year period, log circulation is predicted to rise by 0.0148. The correlation between Pulitzer Awards and newspaper readership is statistically significant.

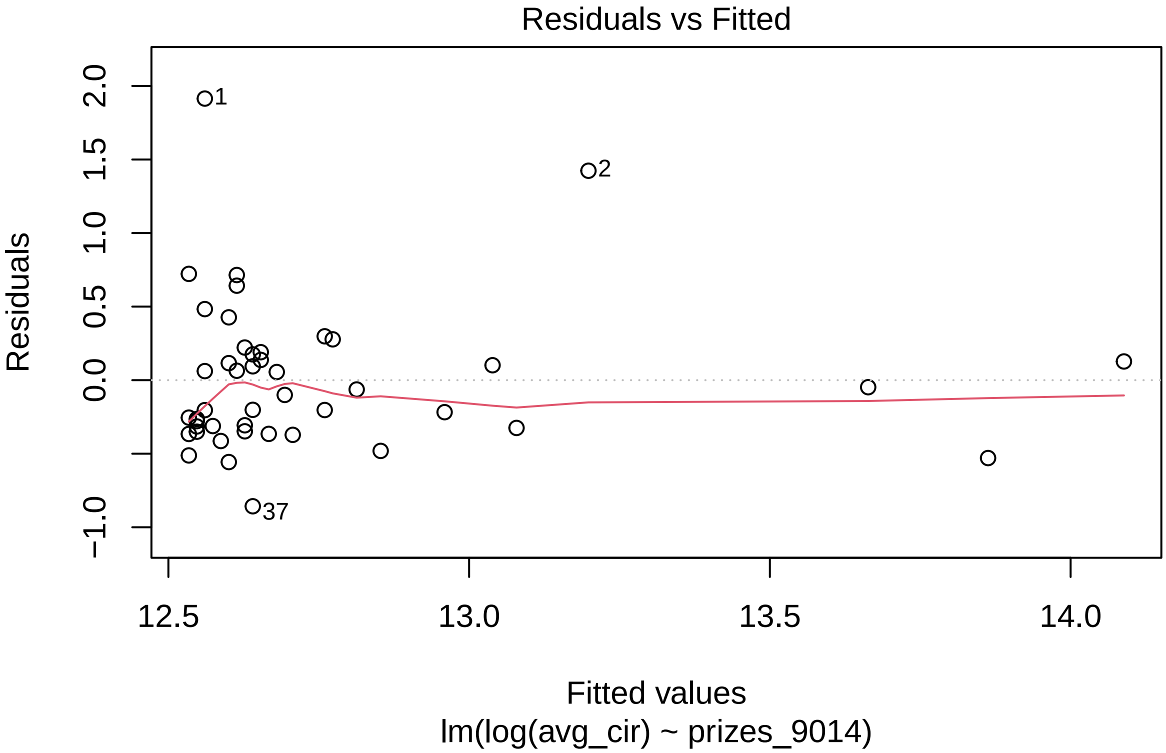
```
pu_ch <- lm(change_0413 ~ prizes_9014, data=pulitzer_1)
summary(pu_ch)
```

```
##
## Call:
## lm(formula = change_0413 ~ prizes_9014, data = pulitzer_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.834 -11.073  -1.834  13.404  57.675
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -35.5915     4.7955  -7.422 3.17e-09 ***
## prizes_9014   0.3806     0.1564   2.434   0.0192 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 26.63 on 43 degrees of freedom
## Multiple R-squared:  0.1211, Adjusted R-squared:  0.1006
## F-statistic: 5.924 on 1 and 43 DF,  p-value: 0.01916
```

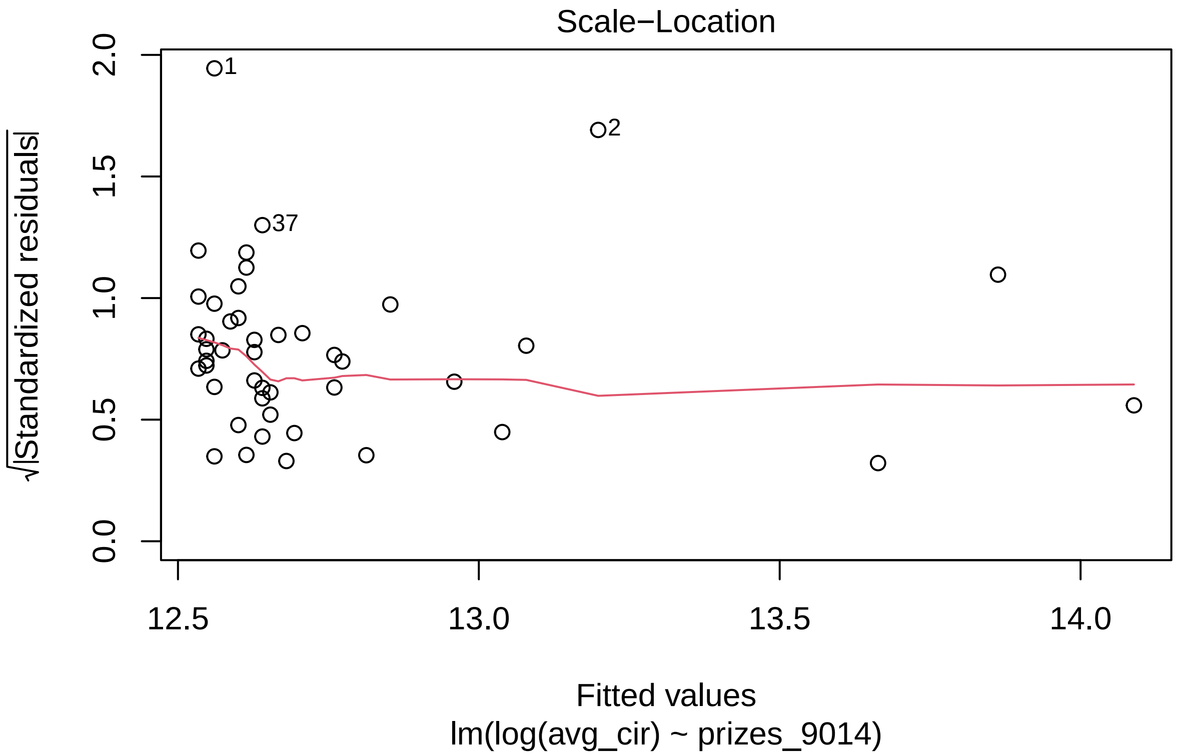
##Slope=0.3870, intercept= -35.4152 ##Interpretation of intercept: If a newspaper receives no Pulitzer Prizes over the course of a 25-year period, we can anticipate a 35.4152% decline in circulation over the final 10 years of that time. ##Interpretation of slope: If a newspaper wins one more Pulitzer Awards during a

25-year period, its readership is predicted to increase by 0.387%. ##The change in newspaper circulation and the awarding of Pulitzer Prizes are statistically related.

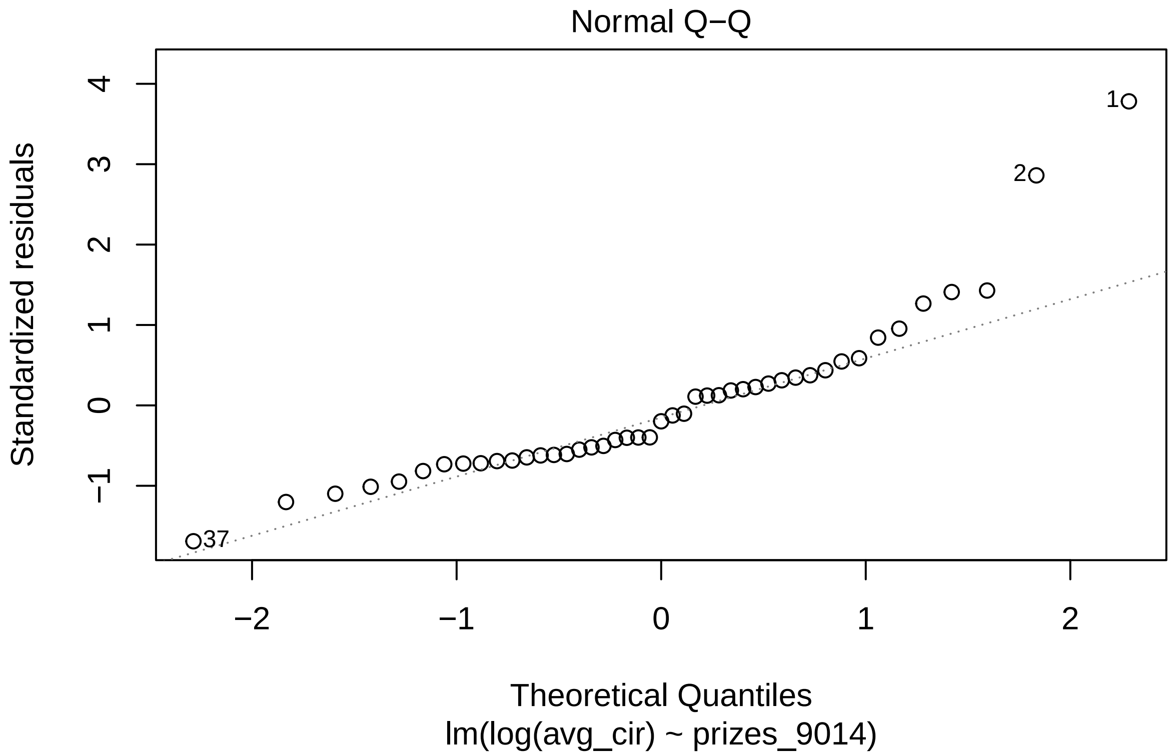
```
#linearity
plot(pu_cir, which=1)
```



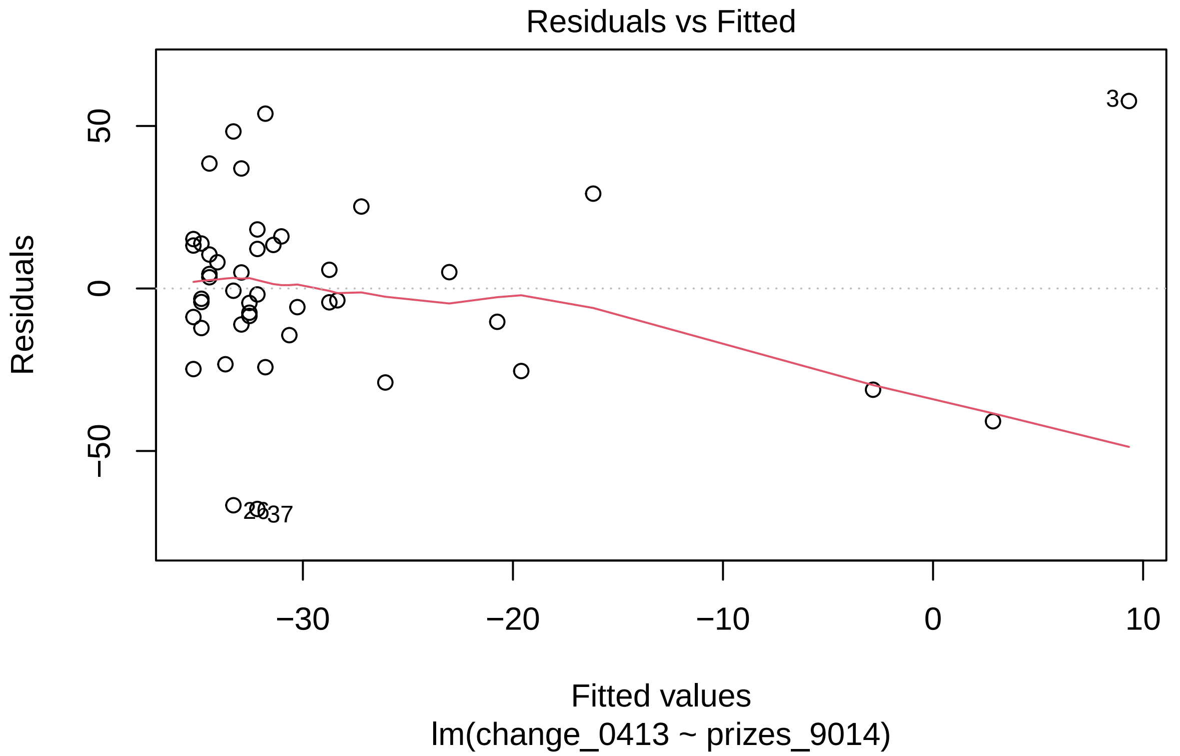
```
#homoscedasticity
plot(pu_cir, which=3)
```



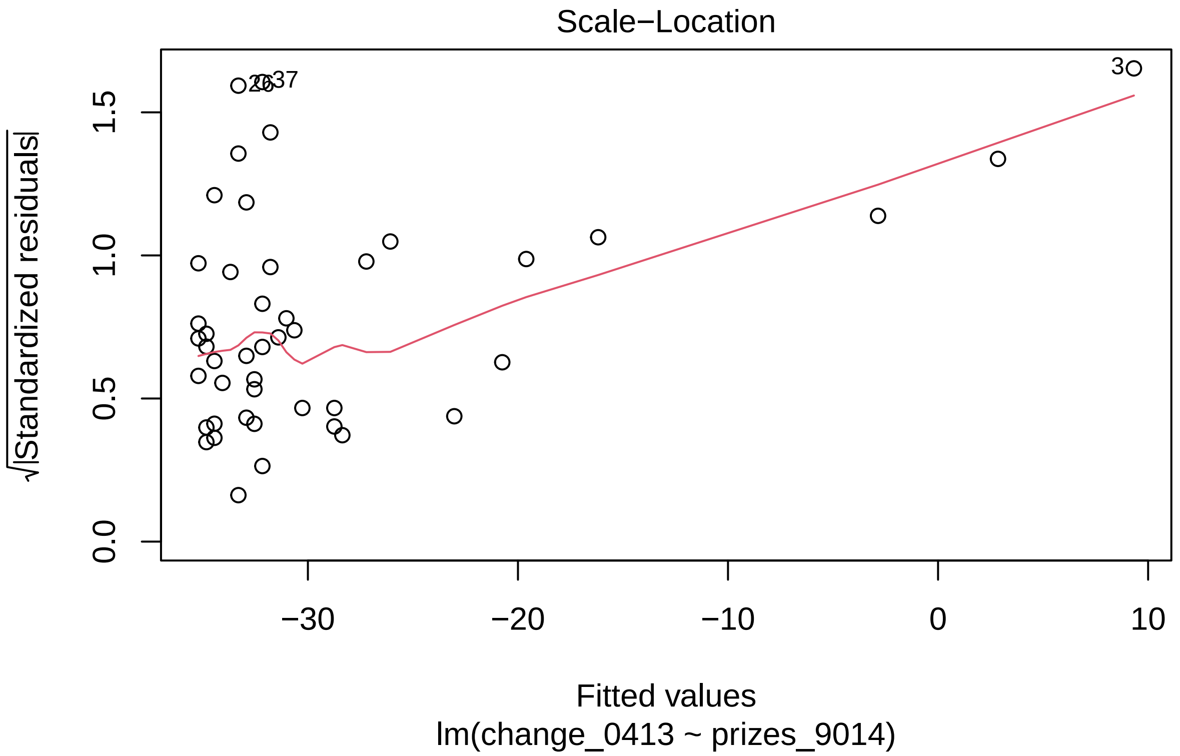
```
#normality
plot(pu_cir, which=2)
```



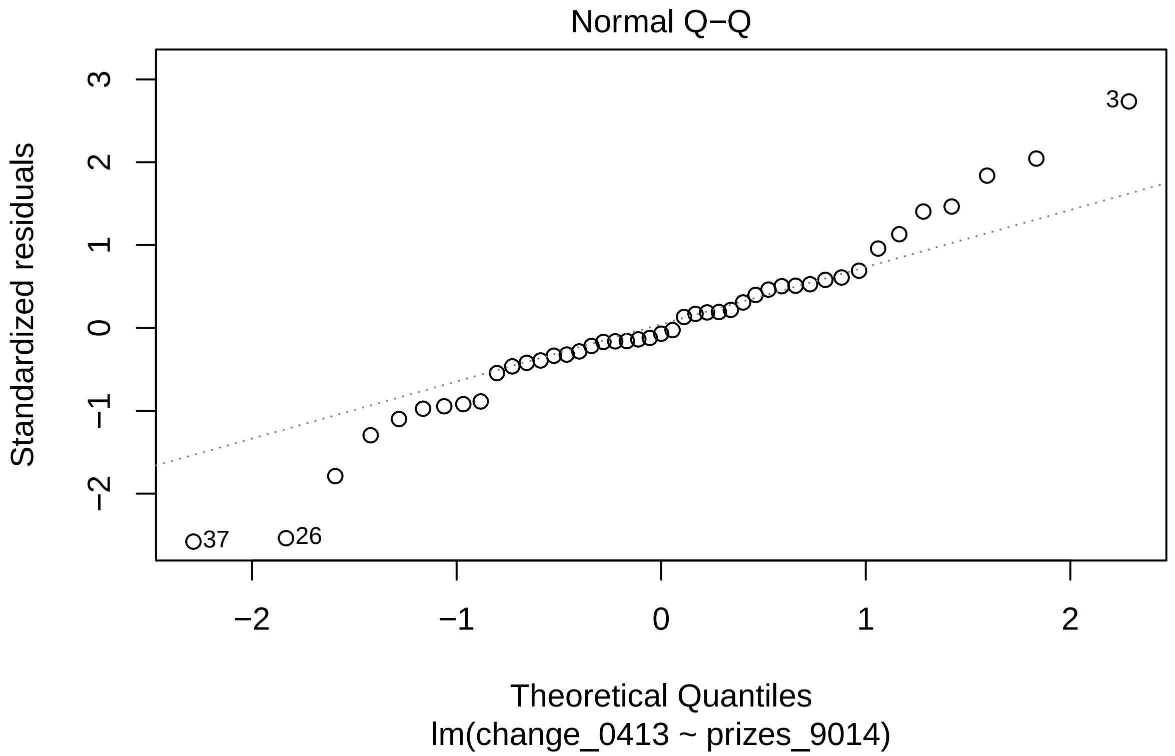
```
#linearity
plot(pu_ch, which=1)
```



```
#homoscedasticity
plot(pu_ch, which=3)
```



```
#normality
plot(pu_ch, which=2)
```



##Ans-For the model predicting circulation: ###(1) Linearity appears to be reasonable; we would want to observe random scatter near zero. The residals vs. fitted plot shows almost no change in trend as we move from left to right. ###(2)The homoscedasticity appears appropriate; I hope the vertical spread does not vary as we move from left to right. The scale vs. location plot shows almost no change in trend as we move from left to right. ###(3)Despite two outliers, normality is reasonable and largely follows the trend line. ###(4)Independence doesn't appear to be warranted because all observations were made during the same dates (1990 and 2014, or 2003 and 2014 for circulation), which indicates that all values will be affected similarly by macro factors in the publishing sector and are therefore are not independent. ###for the model predicting change: ###Linearity appears reasonable; nevertheless, despite the change in the red reference line, the residuals primarily show no trend and are only affected by a few outlier points. If adequately supported, this result would also be unjustifiable. ###The scale vs. location figure shows a clear growing tendency as we move from left to right, which makes homoscedasticity appear implausible. ###Normality appears less rational than the model that predicts circulation because it mainly follows the trend line, save from at the tails. Yet, if effectively argued, it will also appear illogical. ###Independence is not justified.

##Question 4

```
direc <- tibble(prizes_9014 = c(3,25,50))
tibble(Prizes = direc$prizes_9014,
`Expected_circulation` =
predict(pu_cir, direc) %>% exp()) %>%
knitr::kable(digits = 0, format.args = list(big.mark = ","))
```

Prizes	Expected_circulation
3	285,095
25	381,900
50	532,381

###With the greatest investment in investigative journalism, the newspaper's predicted long-term circulation is at its highest. Only the circumstance in which 50 Pulitzer Prizes are won will result in an anticipated increase in circulation, in this case to 522,983 readers, compared to the existing circulation of 453,869 copies.

```r
tibble(Prizes = direc$prizes_9014,
`Expected_change (%)` =
predict(pu_ch, direc)) %>%
knitr::kable(digits = 0, format.args = list(big.mark = ","))
```

Prizes	Expected_change (%)
3	-34
25	-26
50	-17

###All tactical options result in an anticipated drop in circulation. This is distinct from the first model, in which a particular circumstance results in an anticipated rise in circulation. When the current circulation is compared to the expected circulation in the first model, the percentage change in circulation does not match (although this is less significant because the change will occur over a future decade, not necessarily from the current circulation).

```r
cir_con <- predict(pu_cir, direc, interval = "confidence", level=0.9) %>% exp()
cir_con <- tibble(Prizes = direc$prizes_9014,
`Lower bound for mean` = cir_con[,2],
`Expected circulation` = cir_con[,1],
`Upper bound for mean` = cir_con[,3] )
cir_con %>% knitr::kable(digits = 0, format.args = list(big.mark = ","))
```

Prizes	Lower bound for mean	Expected circulation	Upper bound for mean
3	245,997	285,095	330,406
25	333,782	381,900	436,954
50	431,398	532,381	657,001

###In the last 25 years, newspapers with three Pulitzer Prizes have had an average circulation that falls between 236,000 and 309,000; those with 25 Pulitzer Prizes have an average circulation that falls between 324,000 and 418,000; and those with 50 Pulitzer Prizes have an average circulation that falls between 426,000 and 642,000.So, we can state with 90% certainty that newspapers that follow each of the three mentioned strategic strategies generally have different average circulations.

```r
ch_con <- predict(pu_ch, direc, interval = "prediction", level=0.9)
ch_con <- tibble(Prizes = direc$prizes_9014,
`Lower bound for newspaper` = ch_con[,2],
`Expected change in circulation` = ch_con[,1],
`Upper bound for newspaper` = ch_con[,3] )
ch_con %>% knitr::kable(digits = 1, format.args = list(big.mark = ","))
```

Prizes	Lower bound for newspaper	Expected change in circulation	Upper bound for newspaper
3	-79.9	-34.4	11.0
25	-71.4	-26.1	19.2
50	-62.6	-16.6	29.5

with 90% certainty that the circulation of a newspaper with three Pulitzer Prizes in the last 25 years would have increased between -77.7% and 9.2%, between -69.2% and 17.7%, and between -60.2% and 28.1%.There is a lot of overlap between these prediction intervals, demonstrating that there is a wide range of potential outcomes for the Boston Sun-Herald in terms of change in circulation, independent of how many Pulitzer Awards it has received.

##Question 5(a) ###There is no proof that winning the Pulitzer Prize affects how widely anything is read. Another possibility is that larger newspapers have a higher probability of winning because the Pulitzer Prize committee is more likely to appreciate the journalism it has read. There is merely association as a result. ###We assume that the newspaper will be able to exact targets for the number of Pulitzer Prizes it will bring home. It cannot do this without the permission of the Pulitzer committee. ### The observations in the data set for the Pulitzer Prizes were made between 1990 and 2014, while those for the circulation numbers were made between 2003 and 2014. So, every macroeconomic and industry-wide variable that may have an effect on circulation figures affects everyone.

###Conclusion- ###The statistical models developed by Masthead Media are useful in identifying an association between the number of Pulitzer prizes won and the average circulation of the Boston Sun-Times. However, it is important to note that correlation does not necessarily imply causation. Therefore, it cannot be concluded that winning more Pulitzer prizes would lead to an increase in circulation.

###There are several other factors that may affect the circulation of the newspaper, such as changing reader preferences, competition from other media outlets, and shifts in advertising trends. Therefore, a comprehensive analysis of the factors that influence circulation would be necessary to make an informed decision about the newspaper's strategic direction.

###Additionally, as noted in the report, the data used in the models are from a specific time period and may not be applicable to the present or future. Therefore, ongoing monitoring and analysis of the newspaper's circulation trends and readership preferences would be necessary to adapt to changes in the market and make informed decisions about the newspaper's direction.

###In conclusion, while the statistical models developed by Masthead Media provide valuable insights into the relationship between Pulitzer prizes and circulation, they do not provide definitive answers about the direction the Boston Sun-Times should take. A more comprehensive analysis of the factors that influence circulation, ongoing monitoring of readership preferences, and adaptation to changes in the market would be necessary to make informed decisions about the newspaper's strategic direction.