Coursework

Miles Almond, Alex Rowe, Edward Horne

04/04/2020

Question 1

We will start by simply summarising the data, giving us general quantiles, lengths, and classes of each column in the dataframe:

```
summary(BigBangTheory)
```

```
##       ID               Title             Director            Writer
##  Length:231         Length:231         Length:231         Length:231
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    AirDate              Rating          Leonard          Sheldon
##  Length:231         Min.   :6.800   Min.   : 10.00   Min.   : 20.00
##  Class :character   1st Qu.:7.700   1st Qu.: 31.00   1st Qu.: 40.50
##  Mode  :character   Median :8.100   Median : 40.00   Median : 50.00
##                     Mean   :8.036   Mean   : 43.37   Mean   : 51.54
##                     3rd Qu.:8.350   3rd Qu.: 54.00   3rd Qu.: 62.00
##                     Max.   :9.200   Max.   :124.00   Max.   :105.00
##      Penny           Howard           Raj           Leslie
##  Min.   : 0.00   Min.   : 1.00   Min.   : 0.0   Min.   : 0.0000
##  1st Qu.:23.00   1st Qu.:16.00   1st Qu.:12.0   1st Qu.: 0.0000
##  Median :31.00   Median :24.00   Median :18.0   Median : 0.0000
##  Mean   :33.29   Mean   :25.77   Mean   :20.5   Mean   : 0.4632
##  3rd Qu.:42.00   3rd Qu.:34.00   3rd Qu.:28.0   3rd Qu.: 0.0000
##  Max.   :75.00   Max.   :67.00   Max.   :71.0   Max.   :32.0000
##    Bernadette         Amy          Stuart           Emily
##  Min.   : 0.00   Min.   : 0.0   Min.   : 0.000   Min.   : 0.00
##  1st Qu.: 0.00   1st Qu.: 0.0   1st Qu.: 0.000   1st Qu.: 0.00
##  Median :11.00   Median :15.0   Median : 0.000   Median : 0.00
##  Mean   :11.61   Mean   :15.2   Mean   : 3.169   Mean   : 0.71
##  3rd Qu.:20.00   3rd Qu.:25.0   3rd Qu.: 3.000   3rd Qu.: 0.00
##  Max.   :46.00   Max.   :63.0   Max.   :36.000   Max.   :23.00
##       Mary              Zack              Bert             Janine
##  Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000
##  1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000
##  Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 0.0000
##  Mean   : 0.2641   Mean   : 0.5758   Mean   : 0.4199   Mean   : 0.1342
##  3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000
##  Max.   :33.0000   Max.   :47.0000   Max.   :25.0000   Max.   :25.0000
##       Wil
##  Min.   : 0.0000
##  1st Qu.: 0.0000
##  Median : 0.0000
##  Mean   : 0.5368
##  3rd Qu.: 0.0000
##  Max.   :23.0000
```

Next we will use the ‘skim’ function, which separates the columns by type, gives us some nice graphics for the character lines, and tells us more about the minimum, maximum, and uniqueness of the character columns.

```
skim(BigBangTheory)
```

Data summary

Name	BigBangTheory
Number of rows	231
Number of columns	21

Column type frequency:

character	5
numeric	16

Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
ID	0	1	3	5	0	231	0
Title	0	1	5	41	0	231	0
Director	0	1	8	16	0	10	0
Writer	0	1	10	65	0	160	0
AirDate	0	1	8	10	0	228	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Rating	0	1	8.04	0.46	6.8	7.7	8.1	8.35	9.2	▁▆▇▇▁
Leonard	0	1	43.37	18.38	10.00	31.00	40.00	54.00	124.00	▆▇▃▁▁

Sheldon	0	1	51.54	15.87	20.0	40.5	50.0	62.0	105.0	▃▇▆▂▁
Penny	0	1	33.29	14.33	0.0	23.0	31.0	42.0	75.0	▂▇▇▂▁
Howard	0	1	25.77	12.29	1.0	16.0	24.0	34.0	67.0	▃▇▆▂▁
Raj	0	1	20.50	12.55	0.0	12.0	18.0	28.0	71.0	▇▇▃▁▁
Leslie	0	1	0.46	3.21	0.0	0.0	0.0	0.00	32.0	▇▁▁▁▁
Bernadette	0	1	11.61	10.65	0.0	0.0	11.0	20.00	46.0	▇▅▃▁▁
Amy	0	1	15.20	14.52	0.0	0.0	15.0	25.00	63.0	▇▅▃▁▁
Stuart	0	1	3.17	6.46	0.0	0.0	0.0	3.00	36.0	▇▁▁▁▁
Emily	0	1	0.71	3.00	0.0	0.0	0.0	0.00	23.0	▇▁▁▁▁
Mary	0	1	0.26	2.53	0.0	0.0	0.0	0.00	33.0	▇▁▁▁▁
Zack	0	1	0.58	3.89	0.0	0.0	0.0	0.00	47.0	▇▁▁▁▁
Bert	0	1	0.42	2.68	0.0	0.0	0.0	0.00	25.0	▇▁▁▁▁
Janine	0	1	0.13	1.69	0.0	0.0	0.0	0.00	25.0	▇▁▁▁▁
Wil	0	1	0.54	2.79	0.0	0.0	0.0	0.00	23.0	▇▁▁▁▁

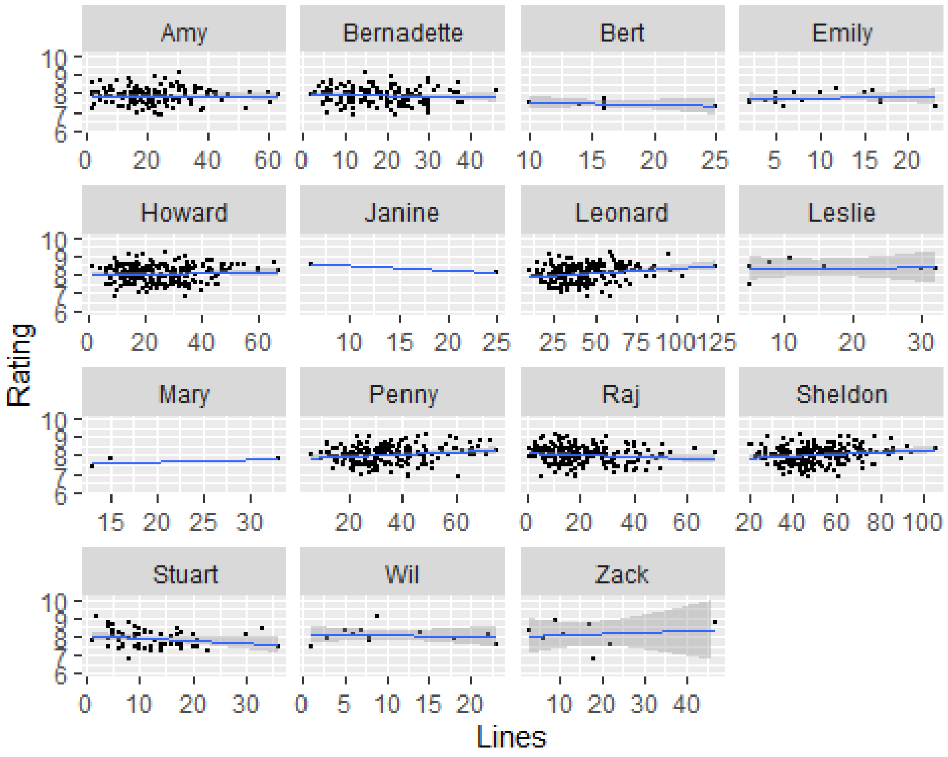
The ‘describeBy’ function allows us to summarise data by a group. In this case we have taken the 5 main characters and the rating (all numerical), and summarised the data by season:

```
BigBangTheory %>% select(1,6:11) %>%
separate(ID,sep="_",into=c("Season","Episode")) -> BigBangTheory2
BigBangTheory2[1] <- as.numeric(BigBangTheory2$Season)
BigBangTheory2 <- select(BigBangTheory2,1,3:8)
describeBy(select(BigBangTheory2,2:7),BigBangTheory2$Season,skew=FALSE)

##
##  Descriptive statistics by group
## group: 1
##          vars  n  mean    sd  min   max range   se
## Rating     1 17  8.31  0.19  7.9   8.7   0.8 0.05
## Leonard    2 17 70.24 21.44 37.0 124.0  87.0 5.20
## Sheldon    3 17 66.18 13.86 50.0 105.0  55.0 3.36
## Penny      4 17 39.12 17.37  6.0  68.0  62.0 4.21
## Howard     5 17 24.47 12.17  6.0  55.0  49.0 2.95
## Raj        6 17 14.41 16.06  2.0  63.0  61.0 3.90
## ------------------------------------------------------------
## group: 2
##         vars  n  mean    sd  min  max range   se
## Rating     1 23  8.41  0.31  7.9  9.2   1.3 0.07
## Leonard    2 23 56.74 14.02 33.0 86.0  53.0 2.92
## Sheldon    3 23 60.74 18.31 29.0 92.0  63.0 3.82
## Penny      4 23 41.43 16.49 19.0 75.0  56.0 3.44
## Howard     5 23 27.35 13.74  9.0 60.0  51.0 2.87
## Raj        6 23 15.57 11.01  3.0 50.0  47.0 2.30
## ------------------------------------------------------------
## group: 3
##         vars  n  mean    sd  min  max range   se
## Rating     1 23  8.40  0.31  7.7  9.1   1.4 0.07
## Leonard    2 23 49.87 16.51 28.0 95.0  67.0 3.44
## Sheldon    3 23 58.61 12.89 35.0 83.0  48.0 2.69
## Penny      4 23 38.04 16.10 17.0 71.0  54.0 3.36
## Howard     5 23 28.57 12.80  8.0 50.0  42.0 2.67
## Raj        6 23 24.65 16.10  0.0 71.0  71.0 3.36
## ------------------------------------------------------------
## group: 4
##         vars  n  mean    sd  min  max range   se
## Rating     1 24  8.25  0.28  7.8  8.8     1 0.06
## Leonard    2 24 46.88 21.59 12.0 96.0    84 4.41
## Sheldon    3 24 57.04 21.06 21.0 94.0    73 4.30
## Penny      4 24 31.46 17.22  0.0 73.0    73 3.51
## Howard     5 24 26.25 16.16  1.0 67.0    66 3.30
## Raj        6 24 20.58 10.85  1.0 40.0    39 2.22
## ------------------------------------------------------------
## group: 5
##         vars  n  mean    sd  min  max range   se
## Rating     1 24  8.15  0.30  7.5  8.7   1.2 0.06
## Leonard    2 24 39.38 18.92 15.0 85.0  70.0 3.86
## Sheldon    3 24 44.33 14.80 20.0 72.0  52.0 3.02
## Penny      4 24 28.58 15.32 10.0 66.0  56.0 3.13
## Howard     5 24 24.42 16.38  4.0 66.0  62.0 3.34
## Raj        6 24 15.42 13.16  1.0 51.0  50.0 2.69
## ------------------------------------------------------------
## group: 6
##         vars  n  mean    sd  min  max range   se
## Rating     1 24  8.13  0.24  7.8  8.6   0.8 0.05
## Leonard    2 24 33.54 12.05 13.0 56.0  43.0 2.46
## Sheldon    3 24 42.29 11.81 23.0 65.0  42.0 2.41
## Penny      4 24 27.62 10.54 13.0 48.0  35.0 2.15
## Howard     5 24 27.04 12.15 13.0 53.0  40.0 2.48
## Raj        6 24 22.04 13.29  2.0 51.0  49.0 2.71
## -------------------------------------------------------------
## group: 7
##          vars  n  mean    sd  min  max range   se
## Rating     1 24  8.04  0.36  7.6  8.9   1.3 0.07
## Leonard    2 24 37.21 12.90 14.0 67.0  53.0 2.63
## Sheldon    3 24 47.21 11.29 25.0 82.0  57.0 2.31
## Penny      4 24 34.58 11.34 10.0 51.0  41.0 2.31
## Howard     5 24 25.00 10.01  6.0 44.0  38.0 2.04
## Raj        6 24 24.17  9.27  9.0 46.0  37.0 1.89
## -------------------------------------------------------------
## group: 8
##          vars  n  mean    sd min  max range   se
## Rating     1 24  7.59  0.33   7  8.2   1.2 0.07
## Leonard    2 24 38.42  9.35  22 59.0  37.0 1.91
## Sheldon    3 24 47.54 13.07  28 78.0  50.0 2.67
## Penny      4 24 32.83 11.85  15 67.0  52.0 2.42
## Howard     5 24 25.96 10.14   9 46.0  37.0 2.07
## Raj        6 24 21.92 10.55   7 45.0  38.0 2.15
## -------------------------------------------------------------
## group: 9
##          vars  n  mean    sd  min  max range   se
## Rating     1 24  7.69  0.43  6.8  9.1   2.3 0.09
## Leonard    2 24 36.54 13.38 17.0 66.0  49.0 2.73
## Sheldon    3 24 48.00 13.47 26.0 77.0  51.0 2.75
## Penny      4 24 31.12 11.72 16.0 61.0  45.0 2.39
## Howard     5 24 24.50  9.91  8.0 45.0  37.0 2.02
## Raj        6 24 20.54  9.45  5.0 40.0  35.0 1.93
## -------------------------------------------------------------
## group: 10
##          vars  n  mean    sd  min  max range   se
## Rating     1 24  7.50  0.41  6.8  8.6   1.8 0.08
## Leonard    2 24 33.54 10.38 10.0 54.0  44.0 2.12
## Sheldon    3 24 48.38  9.70 32.0 74.0  42.0 1.98
## Penny      4 24 30.33  9.88 12.0 47.0  35.0 2.02
## Howard     5 24 23.96  8.35  9.0 36.0  27.0 1.70
## Raj        6 24 23.92 12.28  8.0 54.0  46.0 2.51
```

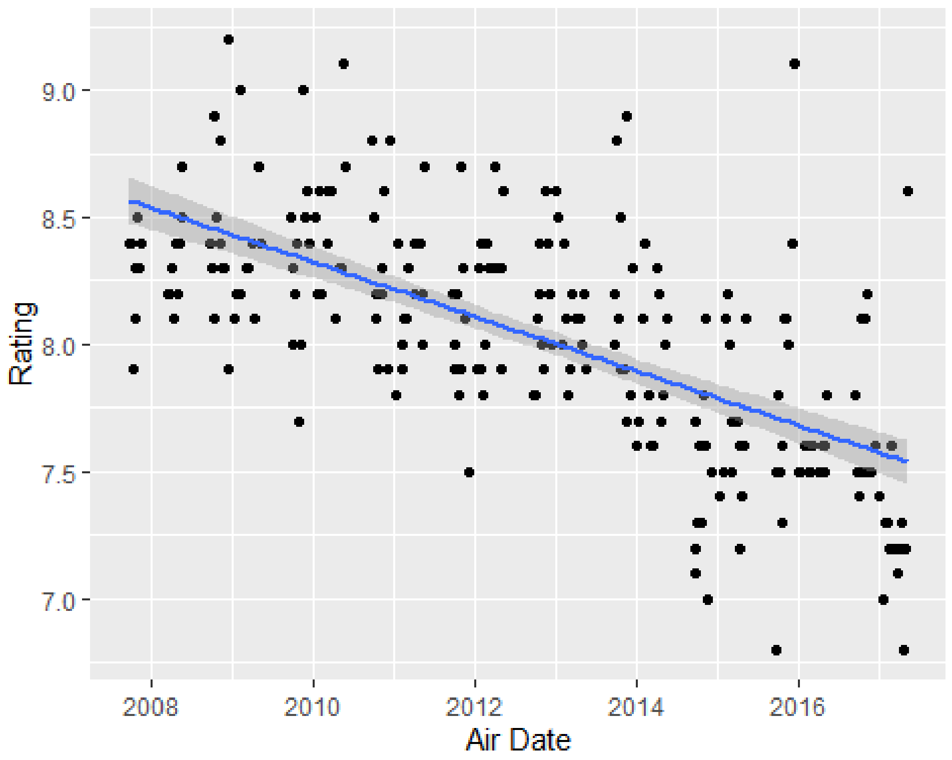
Finally, we decided to give an overview grid representation of how the ratings changed for each character in the show given how many lines they spoke. We decided to change all values of '0' in the data to NA so that when certain characters aren't in a particular episode, their data isn't skewed from an episode they had no involvement in:

```
na_if(BigBangTheory,0) %>% select_if(is.numeric) %>% gather(variable, Lines,
-Rating) %>% ggplot(aes(Lines,Rating)) + geom_point(size=0.4) +
geom_smooth(method='lm',formula='y~x',size=0.4) + facet_wrap(~variable, scale
= "free_x") + ylim(c(6,10))
```



Question 2

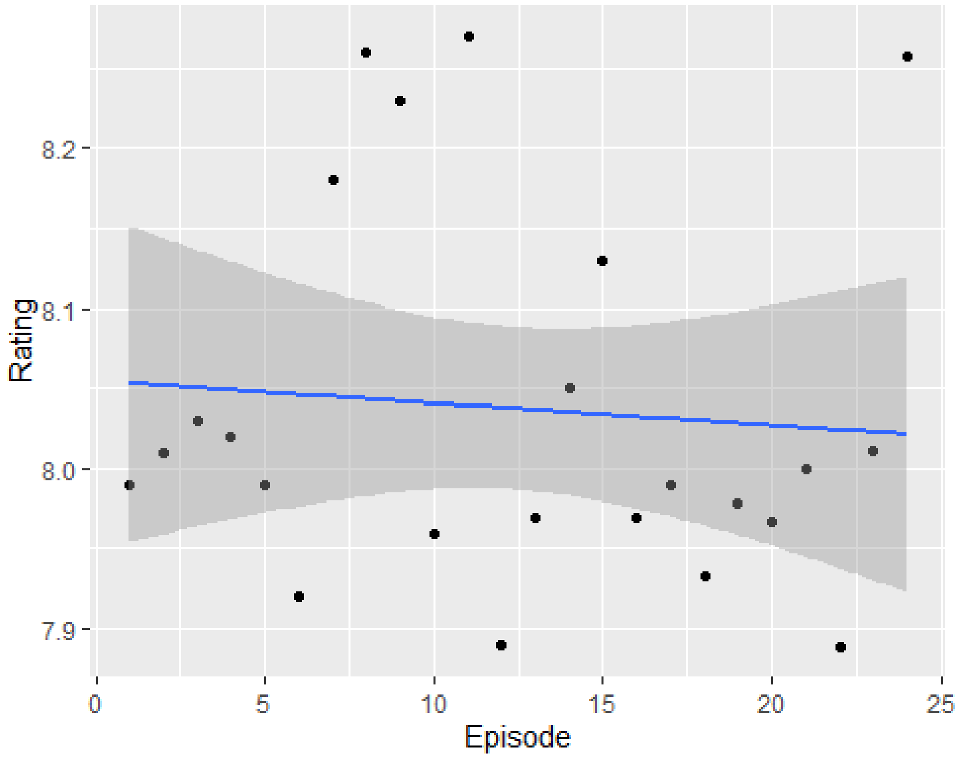
```
BigBangTheory3 = BigBangTheory
BigBangTheory3$AirDate <- as.Date( BigBangTheory3$AirDate, '%m/%d/%Y')
ggplot(BigBangTheory3, aes(AirDate,Rating)) + labs(x="Air Date") +
geom_point() + geom_smooth(method='lm',formula='y~x')
```



From this graph it is clear there is a strong negative correlation between the air date of each episode, and its corresponding rating, showing that over time the rating of the show has declined.

Question 3

```
separate(BigBangTheory,ID,sep="_",into = c("Season","Episode")) %>%
group_by(Episode) %>% summarise(Rating = mean(Rating)) -> Episodes
Episodes[1] <- sapply(Episodes[1],as.numeric)
ggplot(Episodes,aes(Episode,Rating)) + geom_point() +
geom_smooth(method='lm',formula='y~x')
```



From this graph, it can be seen that there is no real correlation between the episode number of each season, and it's corresponding rating. This is clear from the line of best fit and the error bars, as you can easily fit a straight horizontal line between these.

Question 4

We can formulate a linear model to investigate the relationship between the ratings and how much Amy and Bernadette spoke on the show.

```
lmod = lm(Rating ~ Amy + Bernadette, data = BigBangTheory)
summary(lmod)

## 
## Call:
## lm(formula = Rating ~ Amy + Bernadette, data = BigBangTheory)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.15846 -0.28482  0.01518  0.28120  1.23152 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  8.284820   0.043158 191.967  < 2e-16 ***
## Amy         -0.008395   0.002141  -3.921 0.000117 ***
## Bernadette  -0.010407   0.002920  -3.564 0.000445 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4113 on 228 degrees of freedom
## Multiple R-squared:  0.1939, Adjusted R-squared:  0.1868
## F-statistic: 27.42 on 2 and 228 DF,  p-value: 2.135e-11
```

From the data the negative coefficient estimates tell us that both Amy and Bernadette had a negative impact on the ratings over time. These also have statistically significant *p*-values at less than 0.05. However, this negative correlation is inline with the overall decrease in ratings over time hence it is difficult to tell if this was the only cause.

Question 5

To analyse if Chuck Lorre's writing had an impact on ratings first we must change the current 'Writer' variable to a factor variable which represents if Chuck Lorre served as a writer in each episode.

```
Data = BigBangTheory %>% select(Writer, Rating)
Data$Writer = gsub(".*Chuck Lorre.*", "Chuck et al", Data$Writer)
Data = as.data.frame(Data)
Data$Writer = recode_factor(Data$Writer, `Chuck et al` = "1", .default = "0")
```

Next we fit a linear model to assess the correlation between the ratings and whether Chuck Lorre was a writer.

```
lmod = lm(Rating ~ Writer, Data )
summary(lmod)

##
## Call:
## lm(formula = Rating ~ Writer, data = Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.27321 -0.30168  0.09832  0.31255  1.19832
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.07321    0.04306 187.480   <2e-16 ***
## Writer0     -0.07153    0.06000  -1.192    0.234
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4557 on 229 degrees of freedom
## Multiple R-squared:  0.00617,    Adjusted R-squared:  0.00183
## F-statistic: 1.422 on 1 and 229 DF,  p-value: 0.2344
```

From this it can be seen that while the coefficient for the variable 'Writer' is slightly negative it has a p-value of 0.234 which means that this is not a significant result and there is no evidence to say that episodes with Chuck Lorre involved in writing are any different to those without.

Now, if you only include those episodes where Chuck Lorre was the sole writer this yields similar results.

```
Data1 = BigBangTheory %>% select(Writer, Rating)
Data1$Writer = recode_factor(Data1$Writer, `Chuck Lorre` = "1",
              .default = "0")
lmod2 = lm(Rating ~ Writer, data = Data1 )
summary(lmod2)
```

```
##
## Call:
## lm(formula = Rating ~ Writer, data = Data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.23111 -0.33111  0.06889  0.26889  1.16889
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.2333     0.1862  44.228   <2e-16 ***
## Writer0      -0.2022     0.1886  -1.072    0.285
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.456 on 229 degrees of freedom
## Multiple R-squared:  0.004994,   Adjusted R-squared:  0.0006491
## F-statistic: 1.149 on 1 and 229 DF,  p-value: 0.2848
```

From this we can conclusively say that while Chuck Lorre may have been known as ‘King of Sitcoms’, his writing had no statistically significant impact on the ratings.

Question 6

To investigate if any of the episodes were wildly popular or unpopular, we first identify any outliers in the data.

We do this by using the Bonferroni correction

```
BigBangTheory5 = BigBangTheory
BigBangTheory5$AirDate <- as.Date( BigBangTheory5$AirDate, '%m/%d/%Y')
lmod3= lm(Rating ~ AirDate, BigBangTheory5)
tail(sort(abs(rstudent(lmod3))))
```

```
##       62      138      184      144      231      194
## 2.405383 2.587551 2.701349 2.936995 3.173262 4.279843
```

The critical value for declaring an outlier is computed using the Bonferroni correction, with 231 observations and 1 parameter.

```
qt(.05/(231*2),231-1-1)

## [1] -3.759152
```

We can see that only episode 194 meets the outlier criterion by exceeding 3.759 in absolute value, giving it the label of an outlier. However, we are not just looking for outliers, the studentised residuals provide the 6 most abnormally rated episodes listed below in decreasing exceptionality.

```
outliers = BigBangTheory[c(62,138,184,144,231,194), "Rating"]
outlierepisodes = BigBangTheory[c(62,138,184,144,231,194), "ID"]
studentisedr = tail(sort(abs(rstudent(lmod3))))
studentisedr = as.data.frame(studentisedr)
mydata  = as.data.frame(c(outlierepisodes,outliers,studentisedr))
names(mydata)[1] = "Season_Episode"
names(mydata)[3] = "Studentised Residual"
mydata %>% map_df(rev)

## # A tibble: 6 x 3
##   Season_Episode Rating `Studentised Residual`
##   <fct>           <dbl>                  <dbl>
## 1 9_11              9.1                   4.28
## 2 10_24             8.6                   3.17
## 3 7_9               8.9                   2.94
## 4 9_1               6.8                   2.70
## 5 7_3               8.8                   2.59
## 6 3_22              9.1                   2.41
```

The subtle aspect of this data is that while season 3 episode 22 was rated equally high at 9.1 as season 9 episode 11, the latter was a more exceptionally popular episode considering the declining trend in ratings over time. This means that despite not being the most highly rated episodes, these highly rated episodes in the final few seasons are considered more exceptional than those from earlier seasons of the show. At the other end of the spectrum, the opening episode of season 9 scored the equal lowest rating of the show's entire duration and was the most exceptionally low rated episode accounting for its release date (the other episode came later).

Question 7

To see if any characters influenced the amount of lines spoken by Raj, we begin by creating a linear model of Raj against all other characters.

```
lmod = lm(Raj ~ Leonard + Sheldon + Penny + Howard + Leslie + Bernadette +
Amy + Stuart + Emily + Mary + Zack + Bert + Janine + Wil,data =
BigBangTheory)
```

We then use the 'step' function which simplifies the model based on it's Akaike's Information Criterion value (AIC), starting with the original model and comparing against all models with one predictor less etc. until it has a model whereby the AIC value cannot be lowered.

```
lmod = step(lmod)
```

Finally, we can summarise the model while simultaneously removing coefficients where the p-value is greater than 0.05 (i.e. not statistically significant):

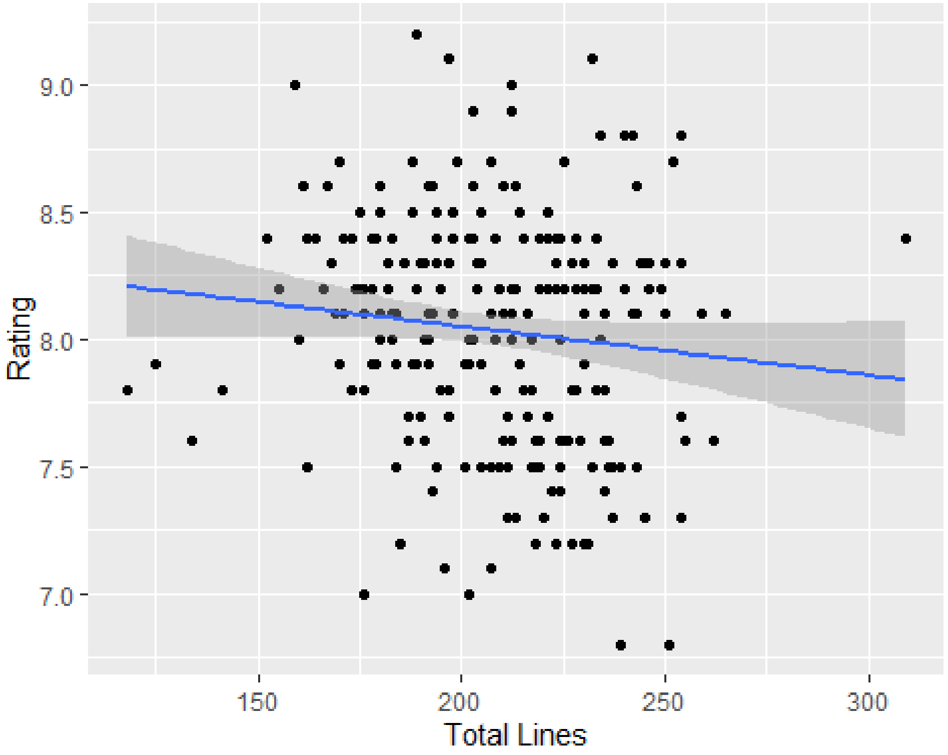
```
summary(lmod)$coefficients[-c(0,6),]

##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 38.6156254 4.65113694  8.302406 9.820281e-15
## Leonard     -0.1226268 0.04689660 -2.614834 9.535982e-03
## Sheldon     -0.1786727 0.05238581 -3.410709 7.690913e-04
## Penny       -0.1587450 0.05352909 -2.965584 3.349771e-03
## Howard       0.2229606 0.06335039  3.519482 5.240467e-04
## Bernadette  -0.3730076 0.08345076 -4.469793 1.245906e-05
## Emily        0.6394558 0.25037432  2.553999 1.131624e-02
```

This data tells us that Leonard, Sheldon, Penny, and Bernadette all have a negative effect on the amount of lines spoken by Raj (negative coefficient estimates), while Howard and Emily both have a positive effect (positive coefficient estimates). This seems plausible as Howard is Raj's best friend, and Emily is a girl Raj was dating. On the other hand, Bernadette had the worst effect on Raj's lines spoken, and this may be due to his mutism around women.

Question 8

```
BigBangTheory %>% select_if(is.numeric) %>% mutate(sum = rowSums(.[2:16]))
%>% select(Rating,sum) %>% ggplot(aes(sum,Rating)) + labs(x="Total Lines") +
geom_point() + geom_smooth(method = 'lm',formula = 'y~x')
```



We can again see that a horizontal line can be fit between the error bars in this graph, meaning there is no true correlation between the total lines of an episode and the corresponding rating.

Question 9

For the duration of season 8 Penny drastically changed her hairstyle from the character's traditional long hair to a short bob. This was heavily criticised by fans, especially during the season premiere, and could be attributed to the low ratings the season received.

In a similar scenario, Keri Russel, who plays a main character in the TV show Felicity, made a drastic change to her hair style in a comparable fashion. In this case, an article on The New York Times said that some commentators suggested her new haircut was the reason for the decline in the show's ratings. It was later shown that this was not the case and the ratings had begun to decline prior to the hair change.

"Some commentators were so upset about Ms. Russell's new style that they suspected it was affecting the show's ratings."

— New York Times

https://www.nytimes.com/2000/01/21/movies/tv-weekend-entering-the-lovelorn-zone-felicity-s-fifth-dimension.html?pagewanted=all&src=pm%20

Our question is as follows: Is there really a relationship between the change in Penny's hairstyle and the episode ratings? To check if Penny's short hair was the reason for poor ratings received in season 8 we can create a linear model using a factor variable which takes a value of 1 for the duration of season 8 when she had short hair, 0 otherwise. Combining this with the air date, which has already proven to be a significant factor, in order to avoid spurious correlation, we can assess the effect of the change of style.

```
Data2 = BigBangTheory %>% select(ID, Rating, AirDate)
Data2$ID = gsub(".*8_.*", "Short Hair", Data2$ID)
Data2 = as.data.frame(Data2)
Data2$ID = recode_factor(Data2$ID, `Short Hair` = "1", .default = "0")
Data2$AirDate <- as.Date( Data2$AirDate, '%m/%d/%Y')

lmod4 = lm(Rating ~ ID + AirDate, Data2 )
summary(lmod4)

##
## Call:
## lm(formula = Rating ~ ID + AirDate, data = Data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95951 -0.22059 -0.05068  0.18932  1.36419
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.207e+01  3.781e-01  31.926   <2e-16 ***
## ID0          2.374e-01  7.592e-02   3.126    0.002 **
## AirDate     -2.724e-04  2.261e-05 -12.050   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3377 on 228 degrees of freedom
## Multiple R-squared:  0.4567, Adjusted R-squared:  0.4519
## F-statistic: 95.82 on 2 and 228 DF,  p-value: < 2.2e-16
```

From this summary you can see that both coefficients are statistically significant with p-values < 0.01 but to some fans' surprise this model suggests that in fact Penny's change of hair style had a positive effect of increasing ratings by 0.237 compared to other seasons where she had long hair. It is most likely incorrect to suggest that Penny's hair style was the sole cause of the slight increase in ratings, but it seems sensible to conclude that it did not have a negative effect. The trend in the declination of ratings over time supersedes any apparent negative bias to ratings caused by Penny changing her hair.

Question 10 (Conclusion)

The Big Bang Theory was aired from 2007 to 2019, spanning 279 episodes across 12 seasons, it was one of Columbia Broadcasting System's (CBS) most successful TV shows, especially over its final few seasons. The primary focus of our research was assessing the behaviour of the show's ratings and trying to establish probable causes for any changes in ratings throughout the show.

As with many long running shows, The Big Bang Theory was subject to declining ratings over time; we found a strong negative correlation between the air date and the episode's ratings showing how perhaps overtime the quality of the show decreased. We looked closer into this result and checked to see if this same correlation was present within each season, however we found no such pattern, this could be due to a shorter gap between episodes than between each season. As with any show with a long run time, the makeup of the main cast is fluid and changes throughout the duration of the show. In particular, we saw Amy and Bernadette introduced as main characters as the show progressed, using the amount of lines spoken by these characters as a metric we found that there was a negative relationship between the amount spoken by both characters and the ratings, with Bernadette having a slightly larger impact than Amy. Next, we investigated how the writer of each episode may have had an impact on the ratings. Chuck Lorre served as writer on many episodes and had a great involvement in the development of the show, he is referred to as the *King of Sitcoms* so we evaluated whether he had a significant impact on the ratings. We concluded that his involvement in the writing of episodes made no difference on the ratings, this means that there was no significant difference between the ratings on episodes he was involved in writing, and those he did not. Another aspect of the show we closely examined was the character of Raj. Portrayed as a more shy character and until the end of 6th season always remained silent when Penny was present. This selective mutism is particularly noticeable around the female characters on the show, for this reason we created a model to see which characters had the greatest impact on how much he spoke. We discovered that Leonard, Sheldon, Penny, and Bernadette all have a negative effect on the amount Raj speaks, while Howard and Emily both have a positive effect. This makes a lot of sense as Howard and Raj were best friends, and Emily was one of Raj's girlfriends. On the other hand, Bernadette had the worst effect on Raj's lines spoken which is expected as she is one of the more intimidating female characters on the show. Following the theme of the lines spoken we proceeded to look into those episodes which were wordier than others, unsurprisingly we found that there was no correlation between the amount of lines spoken in each episode and the rating.

One way to establish the causes of changes to ratings is to look at episodes which were particularly unpopular or popular episodes. We found the most abnormally rated episodes by comparison to the other episodes released around a similar time. One episode which was noticeably poorly received was the opening episode of season 8, scoring a rating of 7.2 making it the 7th lowest episode. Among the top 8 lowest rated episodes 4 were from season 8, this provoked us to look into this further. A noticeable change for season 8 was the introduction of Penny's new hair style, this was widely criticised by fans and some critics blamed this for the poor ratings, hence we decided to investigate further. From statistical analysis we found that there was no significant data to suggest that this was the case, in fact we found the change of hair had a slight positive impact.

Question 11

Throughout the duration of this project we used Facebook messenger to communicate and made changes to the report in a shared google drive file.