Data Analysis Assignment

Nelson Corrocher

July 31, 2016

Preparing the Environment

```
v <- read.csv("LIWC2015 Results.csv")
library(ggplot2)
```

Exploratory Analysis

Generate basic statistics about the dataset using functions like "summary()", execute "mean()" and "sd()" on the Wordcount and WPS columns

```
## Variables Summary:
```

```
##                Filename  DecadePublished YearPublished
##  1887-Dewey.txt     : 1   Min.   :1880    Min.   :1887
##  1891-Ladd.txt      : 1   1st Qu.:1910    1st Qu.:1919
##  1892-Davis.txt     : 1   Median :1940    Median :1947
##  1892-Jimmy.txt     : 1   Mean   :1940    Mean   :1945
##  1893-Baldwin.txt   : 1   3rd Qu.:1970    3rd Qu.:1974
##  1897-Scripture.txt: 1   Max.   :1990    Max.   :1997
##  (Other)            :60
##                Author          WC             Analytic         Clout
##  Angell            : 1   Min.   : 34496   Min.   :82.97   Min.   :39.39
##  Baldwin           : 1   1st Qu.:114816   1st Qu.:90.16   1st Qu.:54.43
##  Bartley           : 1   Median :180679   Median :92.33   Median :62.09
##  Boring et al      : 1   Mean   :183570   Mean   :91.96   Mean   :61.34
##  Bourne_&_Ekstrand: 1   3rd Qu.:227584   3rd Qu.:94.53   3rd Qu.:68.60
##  Calkins           : 1   Max.   :394580   Max.   :97.37   Max.   :78.44
##  (Other)           :60
##    Authentic           Tone             WPS            Sixltr
##  Min.   :18.19   Min.   :19.92   Min.   :14.16   Min.   :21.87
##  1st Qu.:23.91   1st Qu.:27.48   1st Qu.:18.14   1st Qu.:25.64
##  Median :28.04   Median :33.73   Median :19.75   Median :27.32
##  Mean   :28.35   Mean   :34.60   Mean   :20.25   Mean   :27.10
##  3rd Qu.:31.94   3rd Qu.:41.42   3rd Qu.:21.84   3rd Qu.:28.33
##  Max.   :48.25   Max.   :61.85   Max.   :28.92   Max.   :31.58
##
##       Dic          function.        pronoun           ppron
##  Min.   :77.30   Min.   :42.62   Min.   : 5.250   Min.   :1.440
##  1st Qu.:80.67   1st Qu.:46.92   1st Qu.: 7.537   1st Qu.:2.748
##  Median :82.03   Median :49.26   Median : 8.000   Median :3.295
##  Mean   :81.99   Mean   :49.11   Mean   : 8.218   Mean   :3.171
##  3rd Qu.:83.32   3rd Qu.:51.56   3rd Qu.: 9.043   3rd Qu.:3.610
##  Max.   :86.54   Max.   :55.00   Max.   :11.100   Max.   :6.090
##
```

```
##        i                we              you             shehe       
##  Min.   :0.0100   Min.   :0.3000   Min.   :0.0000   Min.   :0.1300  
##  1st Qu.:0.1200   1st Qu.:0.5875   1st Qu.:0.0525   1st Qu.:0.5225  
##  Median :0.1700   Median :0.8500   Median :0.1300   Median :0.8100  
##  Mean   :0.2398   Mean   :0.9714   Mean   :0.2998   Mean   :0.8409  
##  3rd Qu.:0.3175   3rd Qu.:1.1525   3rd Qu.:0.4350   3rd Qu.:1.1175  
##  Max.   :0.9600   Max.   :3.1900   Max.   :2.0700   Max.   :2.0600  
##                                                                     
##       they            ipron          article            prep       
##  Min.   :0.3500   Min.   :3.410   Min.   : 7.410   Min.   :13.30  
##  1st Qu.:0.6525   1st Qu.:4.405   1st Qu.: 8.970   1st Qu.:14.85  
##  Median :0.8000   Median :4.780   Median : 9.820   Median :15.72  
##  Mean   :0.8191   Mean   :5.045   Mean   : 9.872   Mean   :15.67  
##  3rd Qu.:0.9475   3rd Qu.:5.423   3rd Qu.:10.675   3rd Qu.:16.34  
##  Max.   :1.5600   Max.   :8.080   Max.   :12.690   Max.   :18.09  
##                                                                   
##     auxverb          adverb           conj           negate      
##  Min.   :5.620   Min.   :2.550   Min.   :4.520   Min.   :0.5900  
##  1st Qu.:6.688   1st Qu.:3.300   1st Qu.:5.730   1st Qu.:0.8025  
##  Median :7.075   Median :3.520   Median :6.135   Median :0.9050  
##  Mean   :7.097   Mean   :3.544   Mean   :6.081   Mean   :0.9335  
##  3rd Qu.:7.558   3rd Qu.:3.788   3rd Qu.:6.470   3rd Qu.:1.0075  
##  Max.   :8.490   Max.   :4.410   Max.   :7.660   Max.   :1.6300  
##                                                                  
##       verb            adj           compare         interrog      
##  Min.   : 9.28   Min.   :4.830   Min.   :2.370   Min.   :0.990  
##  1st Qu.:10.88   1st Qu.:5.412   1st Qu.:2.922   1st Qu.:1.360  
##  Median :11.32   Median :5.665   Median :3.200   Median :1.530  
##  Mean   :11.41   Mean   :5.729   Mean   :3.160   Mean   :1.552  
##  3rd Qu.:11.84   3rd Qu.:6.065   3rd Qu.:3.360   3rd Qu.:1.705  
##  Max.   :13.22   Max.   :6.660   Max.   :3.790   Max.   :2.200  
##                                                                 
##      number          quant           affect          posemo      
##  Min.   :0.860   Min.   :2.160   Min.   :2.250   Min.   :1.410  
##  1st Qu.:1.640   1st Qu.:2.590   1st Qu.:3.487   1st Qu.:1.985  
##  Median :2.435   Median :2.690   Median :4.105   Median :2.160  
##  Mean   :2.331   Mean   :2.742   Mean   :4.039   Mean   :2.190  
##  3rd Qu.:2.888   3rd Qu.:2.913   3rd Qu.:4.470   3rd Qu.:2.348  
##  Max.   :5.360   Max.   :3.550   Max.   :5.750   Max.   :3.220  
##                                                                 
##      negemo           anx             anger            sad        
##  Min.   :0.790   Min.   :0.1600   Min.   :0.0800   Min.   :0.1600  
##  1st Qu.:1.340   1st Qu.:0.3500   1st Qu.:0.2100   1st Qu.:0.2600  
##  Median :1.675   Median :0.4700   Median :0.3400   Median :0.3100  
##  Mean   :1.705   Mean   :0.4667   Mean   :0.3271   Mean   :0.3188  
##  3rd Qu.:2.098   3rd Qu.:0.5875   3rd Qu.:0.4100   3rd Qu.:0.3600  
##  Max.   :2.770   Max.   :0.7800   Max.   :0.6400   Max.   :0.4800  
##                                                                    
##      social           family           friend           female      
##  Min.   : 3.130   Min.   :0.0100   Min.   :0.0300   Min.   :0.0100  
##  1st Qu.: 5.232   1st Qu.:0.0400   1st Qu.:0.0725   1st Qu.:0.0600  
##  Median : 6.795   Median :0.1800   Median :0.0900   Median :0.2300  
##  Mean   : 6.478   Mean   :0.1820   Mean   :0.1008   Mean   :0.2427  
##  3rd Qu.: 7.918   3rd Qu.:0.2875   3rd Qu.:0.1300   3rd Qu.:0.4125
```

```
##  Max.   :10.540   Max.   :0.5000   Max.   :0.2000   Max.   :0.8200
##
##       male           cogproc         insight          cause
##  Min.   :0.2200   Min.   :11.98   Min.   :1.980   Min.   :1.780
##  1st Qu.:0.5950   1st Qu.:13.56   1st Qu.:3.473   1st Qu.:2.518
##  Median :0.8700   Median :14.36   Median :3.930   Median :2.830
##  Mean   :0.9303   Mean   :14.46   Mean   :3.888   Mean   :2.854
##  3rd Qu.:1.2350   3rd Qu.:15.04   3rd Qu.:4.235   3rd Qu.:3.140
##  Max.   :2.2900   Max.   :18.59   Max.   :5.910   Max.   :4.630
##
##     discrep          tentat         certain          differ
##  Min.   :0.580   Min.   :2.570   Min.   :0.820   Min.   :2.800
##  1st Qu.:1.040   1st Qu.:3.180   1st Qu.:1.262   1st Qu.:3.320
##  Median :1.120   Median :3.420   Median :1.555   Median :3.490
##  Mean   :1.168   Mean   :3.429   Mean   :1.648   Mean   :3.608
##  3rd Qu.:1.317   3rd Qu.:3.700   3rd Qu.:1.980   3rd Qu.:3.917
##  Max.   :1.930   Max.   :4.320   Max.   :3.010   Max.   :5.220
##
##     percept           see             hear             feel
##  Min.   :1.950   Min.   :0.750   Min.   :0.2400   Min.   :0.2600
##  1st Qu.:2.562   1st Qu.:1.020   1st Qu.:0.3600   1st Qu.:0.4025
##  Median :2.940   Median :1.175   Median :0.4500   Median :0.5200
##  Mean   :3.157   Mean   :1.223   Mean   :0.4474   Mean   :0.6541
##  3rd Qu.:3.658   3rd Qu.:1.367   3rd Qu.:0.5275   3rd Qu.:0.7175
##  Max.   :5.090   Max.   :2.120   Max.   :0.6400   Max.   :1.7400
##
##       bio             body            health           sexual
##  Min.   :0.850   Min.   :0.3000   Min.   :0.2400   Min.   :0.0100
##  1st Qu.:1.995   1st Qu.:0.8425   1st Qu.:0.6225   1st Qu.:0.0700
##  Median :2.340   Median :1.1350   Median :0.7450   Median :0.1200
##  Mean   :2.438   Mean   :1.1852   Mean   :0.8439   Mean   :0.1488
##  3rd Qu.:2.812   3rd Qu.:1.4000   3rd Qu.:1.0375   3rd Qu.:0.1975
##  Max.   :3.870   Max.   :2.8300   Max.   :1.6700   Max.   :0.4800
##
##      ingest           drives       affiliation       achieve
##  Min.   :0.2100   Min.   :4.010   Min.   :0.830   Min.   :0.770
##  1st Qu.:0.2725   1st Qu.:5.518   1st Qu.:1.530   1st Qu.:1.232
##  Median :0.3650   Median :6.305   Median :1.770   Median :1.565
##  Mean   :0.3820   Mean   :6.305   Mean   :1.840   Mean   :1.538
##  3rd Qu.:0.4600   3rd Qu.:7.100   3rd Qu.:2.103   3rd Qu.:1.790
##  Max.   :0.7900   Max.   :8.870   Max.   :3.800   Max.   :2.330
##
##      power           reward            risk          focuspast
##  Min.   :1.320   Min.   :0.4200   Min.   :0.2400   Min.   :1.310
##  1st Qu.:1.752   1st Qu.:0.6825   1st Qu.:0.4625   1st Qu.:1.958
##  Median :2.095   Median :0.8500   Median :0.5850   Median :2.415
##  Mean   :2.138   Mean   :0.8458   Mean   :0.5752   Mean   :2.407
##  3rd Qu.:2.502   3rd Qu.:1.0125   3rd Qu.:0.6975   3rd Qu.:2.745
##  Max.   :3.240   Max.   :1.3000   Max.   :0.9100   Max.   :4.970
##
##   focuspresent    focusfuture        relativ          motion
##  Min.   :5.500   Min.   :0.5900   Min.   : 9.12   Min.   :0.890
##  1st Qu.:7.133   1st Qu.:0.8850   1st Qu.:11.36   1st Qu.:1.590
##  Median :7.575   Median :1.0150   Median :11.89   Median :1.740
```

```
##  Mean   :7.557   Mean   :0.9985   Mean   :11.79   Mean   :1.705
##  3rd Qu.:8.000   3rd Qu.:1.1075   3rd Qu.:12.36   3rd Qu.:1.900
##  Max.   :8.890   Max.   :1.4100   Max.   :14.48   Max.   :2.480
##
##      space            time            work          leisure
##  Min.   :4.960   Min.   :2.710   Min.   :1.360   Min.   :0.2000
##  1st Qu.:6.003   1st Qu.:3.380   1st Qu.:2.252   1st Qu.:0.3500
##  Median :6.440   Median :3.600   Median :3.175   Median :0.4550
##  Mean   :6.484   Mean   :3.630   Mean   :3.056   Mean   :0.4771
##  3rd Qu.:6.938   3rd Qu.:3.928   3rd Qu.:3.938   3rd Qu.:0.5875
##  Max.   :8.160   Max.   :4.440   Max.   :5.120   Max.   :0.9100
##
##       home            money            relig            death
##  Min.   :0.0200   Min.   :0.1100   Min.   :0.0200   Min.   :0.01000
##  1st Qu.:0.0800   1st Qu.:0.2200   1st Qu.:0.0725   1st Qu.:0.04000
##  Median :0.1250   Median :0.2700   Median :0.1000   Median :0.06000
##  Mean   :0.1285   Mean   :0.2798   Mean   :0.1153   Mean   :0.06727
##  3rd Qu.:0.1800   3rd Qu.:0.3300   3rd Qu.:0.1550   3rd Qu.:0.09000
##  Max.   :0.2700   Max.   :0.5600   Max.   :0.3000   Max.   :0.18000
##
##     informal          swear           netspeak          assent
##  Min.   :0.1100   Min.   :0.00000   Min.   :0.00000   Min.   :0.01000
##  1st Qu.:0.2025   1st Qu.:0.01000   1st Qu.:0.05000   1st Qu.:0.03000
##  Median :0.2400   Median :0.01000   Median :0.10000   Median :0.04000
##  Mean   :0.2459   Mean   :0.01742   Mean   :0.09939   Mean   :0.04545
##  3rd Qu.:0.2775   3rd Qu.:0.02000   3rd Qu.:0.13000   3rd Qu.:0.06000
##  Max.   :0.5100   Max.   :0.07000   Max.   :0.37000   Max.   :0.10000
##
##      nonflu            filler             AllPunc          Period
##  Min.   :0.04000   Min.   :0.0000000   Min.   : 8.74   Min.   :3.480
##  1st Qu.:0.07000   1st Qu.:0.0000000   1st Qu.:12.19   1st Qu.:4.525
##  Median :0.09000   Median :0.0000000   Median :13.21   Median :4.965
##  Mean   :0.09333   Mean   :0.0009091   Mean   :13.46   Mean   :5.002
##  3rd Qu.:0.11000   3rd Qu.:0.0000000   3rd Qu.:14.32   3rd Qu.:5.518
##  Max.   :0.19000   Max.   :0.0200000   Max.   :18.04   Max.   :6.740
##
##      Comma            Colon             SemiC            QMark
##  Min.   :3.390   Min.   :0.0000   Min.   :0.0000   Min.   :0.0100
##  1st Qu.:4.808   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0825
##  Median :5.350   Median :0.1750   Median :0.2100   Median :0.1600
##  Mean   :5.517   Mean   :0.1671   Mean   :0.2014   Mean   :0.1921
##  3rd Qu.:5.875   3rd Qu.:0.2475   3rd Qu.:0.3300   3rd Qu.:0.2750
##  Max.   :9.730   Max.   :0.8000   Max.   :0.6900   Max.   :0.8400
##
##      Exclam             Dash            Quote           Apostro
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0100
##  1st Qu.:0.00250   1st Qu.:0.4400   1st Qu.:0.0100   1st Qu.:0.1200
##  Median :0.01000   Median :0.5900   Median :0.0400   Median :0.1800
##  Mean   :0.01818   Mean   :0.5939   Mean   :0.2323   Mean   :0.2944
##  3rd Qu.:0.03000   3rd Qu.:0.7850   3rd Qu.:0.2750   3rd Qu.:0.3275
##  Max.   :0.09000   Max.   :1.0700   Max.   :2.5500   Max.   :1.5000
##
##     Parenth           OtherP
##  Min.   :0.0400   Min.   :0.0000
```

```
##  1st Qu.:0.4425   1st Qu.:0.0225
##  Median :1.0100   Median :0.1200
##  Mean   :1.0441   Mean   :0.2029
##  3rd Qu.:1.4900   3rd Qu.:0.2675
##  Max.   :2.4800   Max.   :2.2000
##
```

```
## WPS Mean +/- SD: 20.25485 +/- 2.994393
```

```
## WC Mean +/- SD: 183570.1 +/- 85966.92
```

Find the columns with the least/greatest correlation.

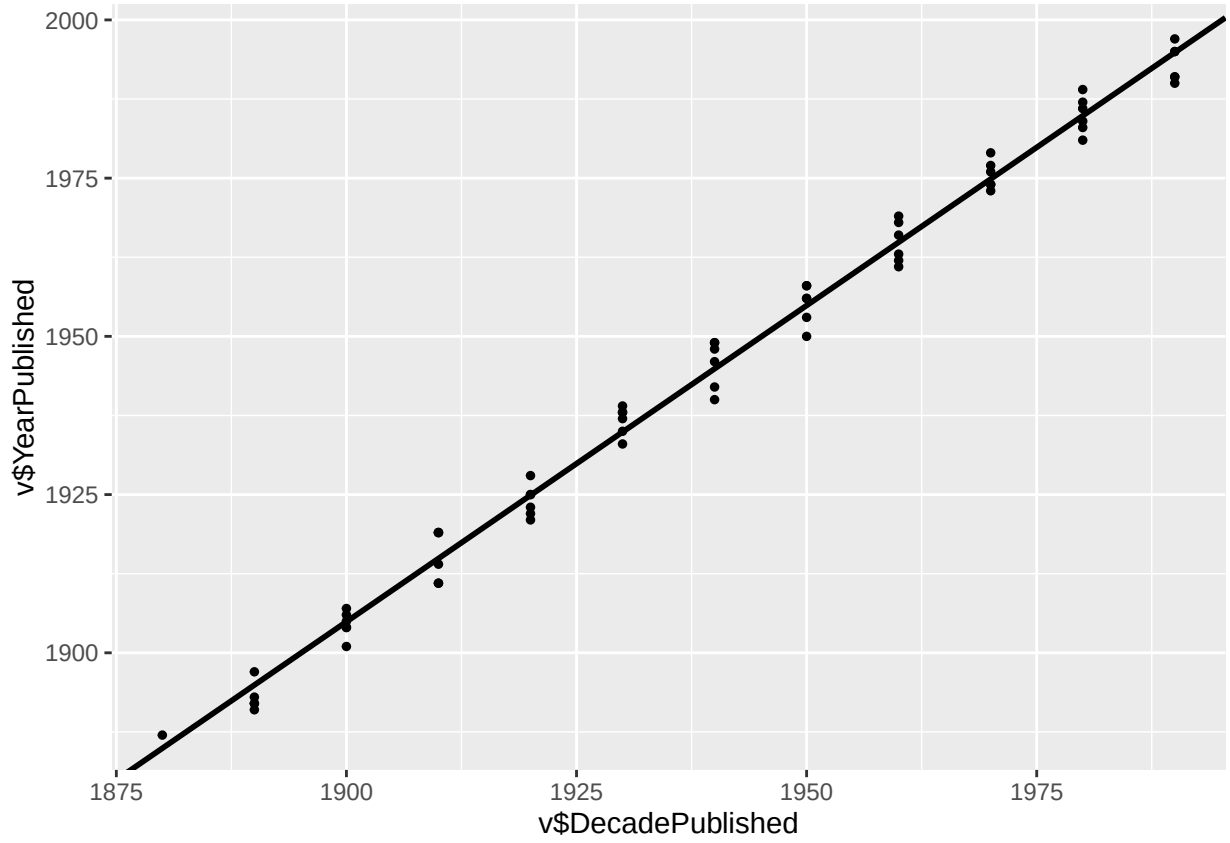
```
## Greatest Correlation: 0.9960026 ,between YearPublished and DecadePublished
```

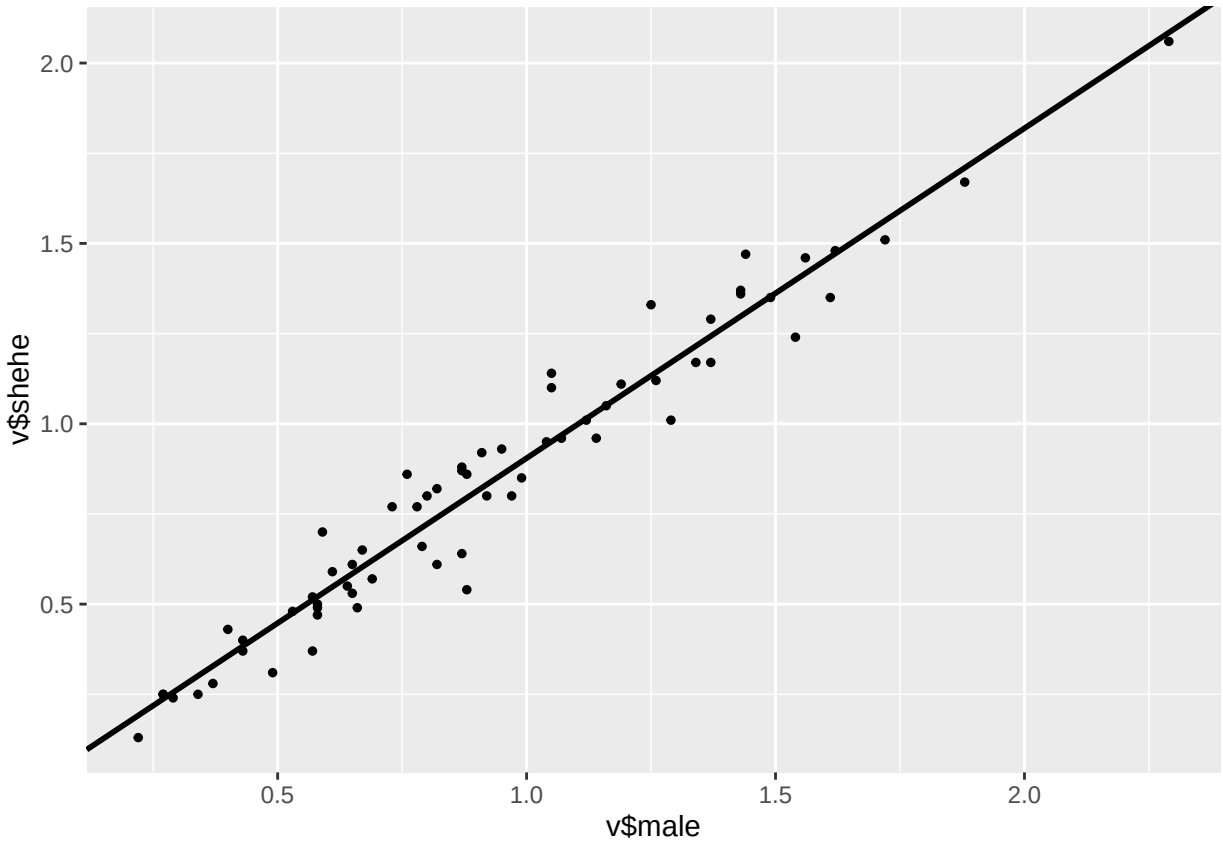
```
## Least Correlation: -0.9630485 ,between Period and WPS
```

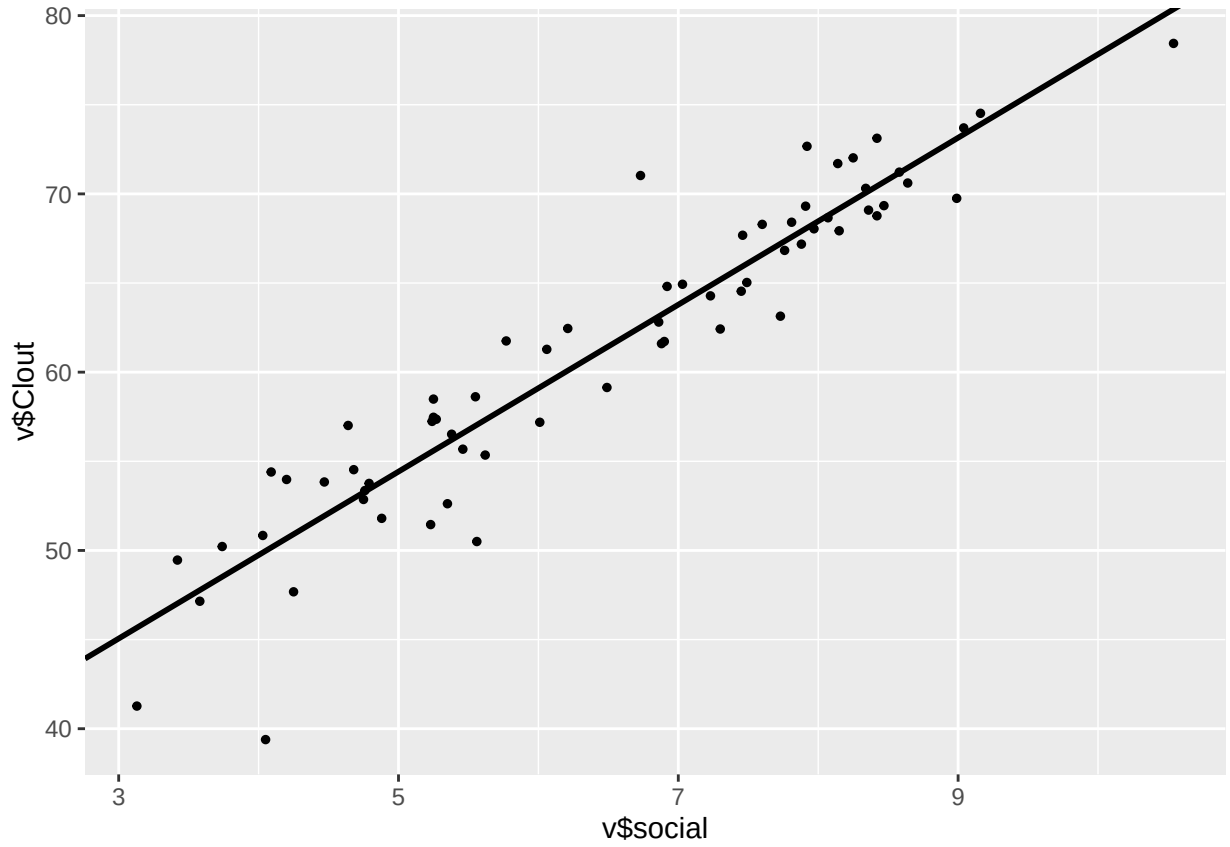
Find and include definitions for Residuals vs Fitted, Normal Q-Q, Scale-Location, Residuals vs Leverage as part of your markdown before the plots.

- **Residual vs Fit:** The plot is used to detect non-linearity, unequal error variances, and outliers.
- **Normal Q-Q:** A graphical method for comparing two probability distributions by plotting their quantiles against each other.
- **Residuals vs Leverage:** Method used to try to isolate the effects of a single variable on the residuals. The plot shows contours of equal Cook's distance
- **Scale-Location:** Ttakes the square root of the absolute residuals in order to diminish skewness.

Generate linear regression plots for the top 3 correlations you find




2.0
1.5
1.0
0.5
v$shehe
0.5
1.0
1.5
2.0
v$male



Describe 2 predictions you might be able to make about the use of language in psychology textbooks over the next decade

This question is a bit subjective because the column names are not very clear what they mean so we have to take some assumptions here. In this example, let's use three variables:

- DecadePublished
- Anger
- Sad

It would be better if we had variable that classified these book titles by subject, so it would be possible to try to predict the usage of these words only in Psychology books. Here, to linear models are going to be run: (DecadePublished ~ Anger) and (DecadePublished ~ Sad). The results will be evaulated based on what has happened to the use of the two words (sad and Anger, closely related to psychology subject) related to the time, and assuming the same trend will happen with Psychology books.

```
##
## Call:
## lm(formula = v$DecadePublished ~ v$anger)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -45.00 -17.28  -1.45  15.97  55.32
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1880.052      7.019 267.860  < 2e-16 ***
## v$anger      182.796     19.908   9.182 2.72e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.27 on 64 degrees of freedom
## Multiple R-squared:  0.5685, Adjusted R-squared:  0.5617
## F-statistic: 84.31 on 1 and 64 DF,  p-value: 2.722e-13
```

```
##
## Call:
## lm(formula = v$DecadePublished ~ v$sad)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.349 -23.261   6.261  20.743  54.268
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1870.00      14.24 131.283  < 2e-16 ***
## v$sad         219.12      43.41   5.047 3.96e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.38 on 64 degrees of freedom
## Multiple R-squared:  0.2847, Adjusted R-squared:  0.2735
## F-statistic: 25.48 on 1 and 64 DF,  p-value: 3.962e-06
```

Since the slope of of both words seems to be positive and the p-value significantly less than 0.05, one could expect that the use of both words in literature titles are going to increase.

Are there any multiply correlated columns in the dataset - where 3 or even 4 columns track together with a strong correlation?

```
summary(lm(v$WC ~ v$Sixltr +v$function. + v$ppron))
```

```
##
## Call:
## lm(formula = v$WC ~ v$Sixltr + v$function. + v$ppron)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -117001  -40252   -5870   33665  160907
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1354711     238190   5.688 3.73e-07 ***
## v$Sixltr       -9194       4291  -2.142  0.03609 *
## v$function.   -20458       3069  -6.666 8.11e-09 ***
```

```
## v$ppron        26111       9199   2.838  0.00612 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62980 on 62 degrees of freedom
## Multiple R-squared:  0.4881, Adjusted R-squared:  0.4633
## F-statistic: 19.71 on 3 and 62 DF,  p-value: 4.355e-09
```

Yes. The example above, assuming WC could be caused by Sixltr, function, ppron, these three variables would have together a strong explanation power on the results of WC. Again, the data used had no meta data to explain what its fields meant so definiting causal relation between them is very difficult. It is has value as an example, though.

Session Info

```
sessionInfo()
```

```
## R version 3.3.1 (2016-06-21)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 10586)
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] ggplot2_2.1.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.6      digest_0.6.10    plyr_1.8.4       grid_3.3.1
##  [5] gtable_0.2.0     formatR_1.4      magrittr_1.5     evaluate_0.9
##  [9] scales_0.4.0     stringi_1.1.1    rmarkdown_1.0    labeling_0.3
## [13] tools_3.3.1      stringr_1.0.0    munsell_0.4.3    yaml_2.1.13
## [17] colorspace_1.2-6 htmltools_0.3.5  knitr_1.13
```