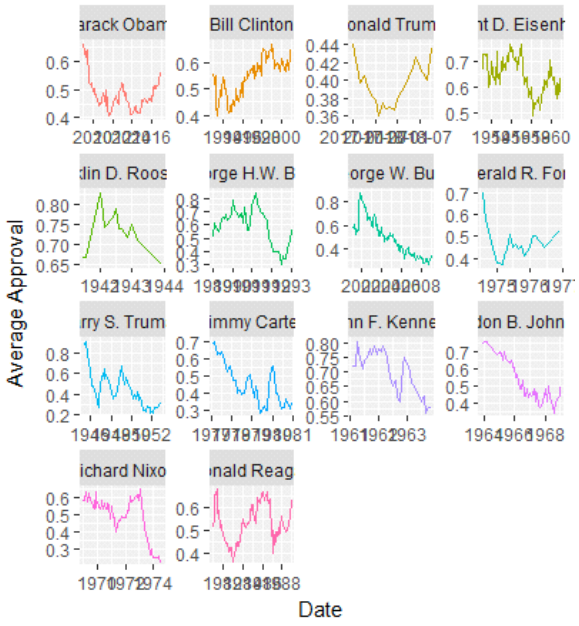
Project

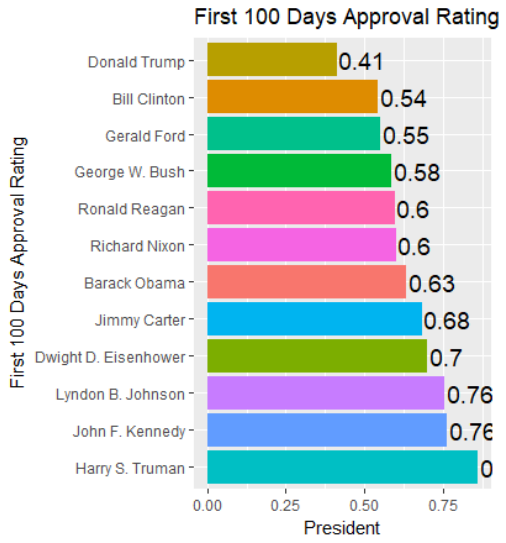
Sunday, April 7, 2019 6:26 PM

AFINN sentiment analysis output

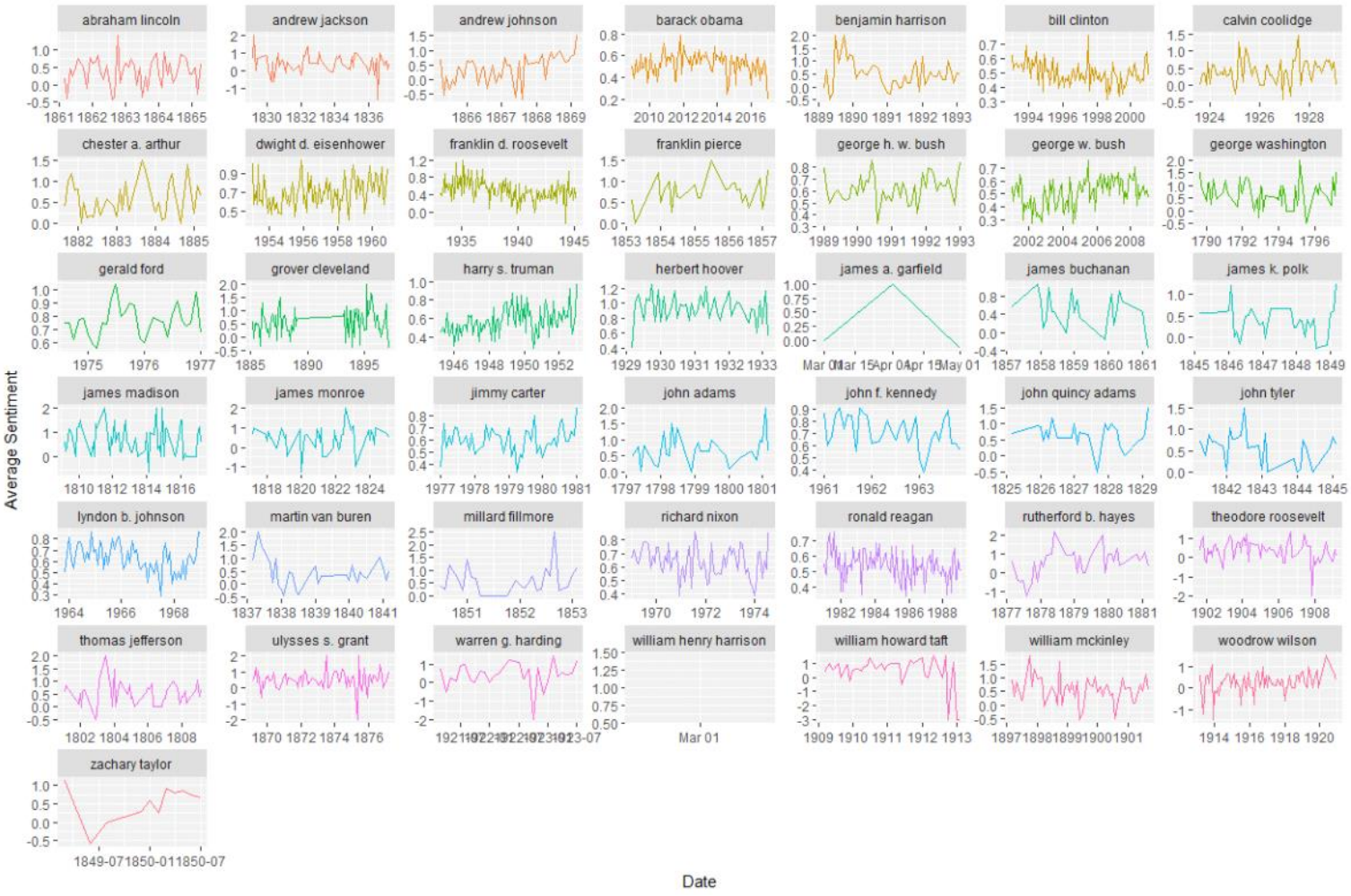
> monthly_approval



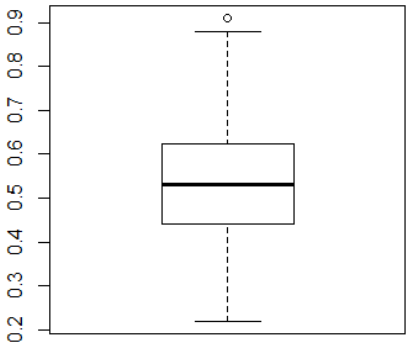
> first_100



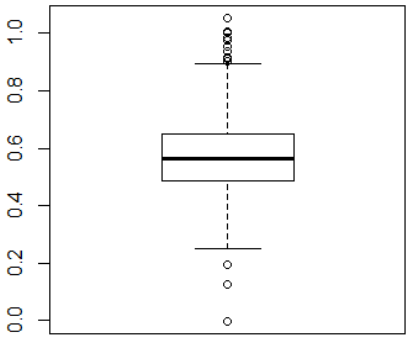
Monthly Sentiment



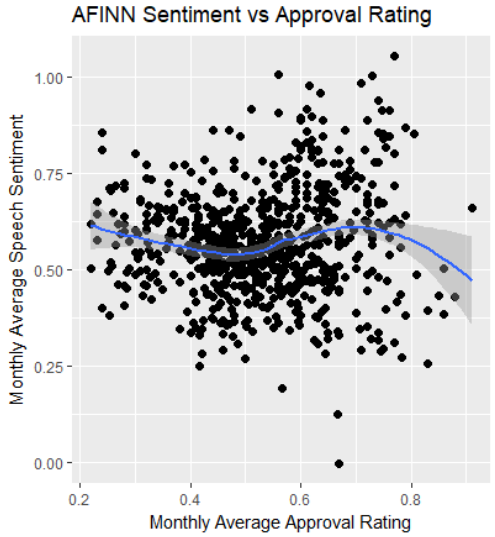

```
> # ------------------------------------------------------------------------------------------
> # Sentiment vs. Approval modeling
> # ------------------------------------------------------------------------------------------
>
> boxplot(sentiment_vs_approval$avg_approval) # Looks fine except for 1 value, probably not an outlier
```



```
> boxplot(sentiment_vs_approval$avg_sentiment) # A lot of outliers
```




```
> sentiment_vs_approval <- sentiment_vs_approval %>% filter(avg_sentiment < 80)
>
> # Plot approval vs. sentiment
```



Linear Regression models - checks Avg Diversity and Avg Sentiment score, then looks at each predictor individually

```
> all_speeches_approval_lm <- lm(avg_approval ~ avg_diversity + avg_sentiment, data = sentiment_vs_approval)
> summary(all_speeches_approval_lm)
Call:
lm(formula = avg_approval ~ avg_diversity + avg_sentiment, data = sentiment_vs_approval)
Residuals:
     Min       1Q   Median       3Q      Max
-0.32785 -0.07713  0.00629  0.08615  0.35065
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.874e-01  3.350e-02  17.536  < 2e-16 ***
avg_diversity -2.582e-04  7.036e-05  -3.669 0.000262 ***
avg_sentiment  2.990e-02  3.624e-02   0.825 0.409509
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1278 on 721 degrees of freedom
Multiple R-squared:  0.02379,         Adjusted R-squared:  0.02109
F-statistic: 8.787 on 2 and 721 DF,  p-value: 0.0001697
> rSquared <- 0.0
> for (i in seq(0.3, 1, by = 0.01)) {
+
+    # filter dataframe by approval rate
+    filtered_lm_df <- sentiment_vs_approval %>%
+       filter(avg_approval <= i)
+
+    # Fit model
+    all_speeches_approval_lm <- lm(avg_approval ~ avg_diversity + avg_sentiment, data = filtered_lm_df)
+
+    # Extract rSquared if higher than previous value
+    if (!is.na(summary(all_speeches_approval_lm)$r.squared) & summary(all_speeches_approval_lm)$r.squared >
rSquared) {
+       rSquared <- summary(all_speeches_approval_lm)$r.squared
+       approval_threshold <- i
```

```
+    }
+ }
> print(rSquared)
[1] 0.1127856
> print(approval_threshold)
[1] 0.47
> filtered_lm_df <- sentiment_vs_approval %>%
+     filter(avg_approval < approval_threshold)
```

Zooming in on just the portion of the data where the relationship seems to be "strongest"

```
> all_speeches_approval_lm1 <- lm(avg_approval ~ avg_diversity + avg_sentiment, data = filtered_lm_df)
> summary(all_speeches_approval_lm1)
Call:
lm(formula = avg_approval ~ avg_diversity + avg_sentiment, data = filtered_lm_df)
Residuals:
    Min      1Q  Median      3Q     Max
-0.24070 -0.04315  0.02247  0.04727  0.10635
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.468e-01  3.120e-02   11.12  < 2e-16 ***
avg_diversity  3.038e-04  6.252e-05    4.86 2.15e-06 ***
avg_sentiment -8.948e-02  3.857e-02   -2.32   0.0212 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06332 on 235 degrees of freedom
Multiple R-squared:  0.1256,            Adjusted R-squared:  0.1182
F-statistic: 16.88 on 2 and 235 DF,  p-value: 1.413e-07
> all_speeches_approval_lm2 <- lm(avg_approval ~ avg_diversity , data = filtered_lm_df)
> summary(all_speeches_approval_lm2)
Call:
lm(formula = avg_approval ~ avg_diversity, data = filtered_lm_df)
Residuals:
    Min      1Q  Median      3Q     Max
-0.24175 -0.04776  0.02050  0.04768  0.09486
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.892e-01  1.908e-02  15.156  < 2e-16 ***
avg_diversity 3.283e-04  6.219e-05   5.278 2.95e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0639 on 236 degrees of freedom
Multiple R-squared:  0.1056,            Adjusted R-squared:  0.1018
F-statistic: 27.86 on 1 and 236 DF,  p-value: 2.955e-07
> all_speeches_approval_lm4 <- lm(avg_approval ~ avg_sentiment, data = filtered_lm_df)
> summary(all_speeches_approval_lm4)
Call:
lm(formula = avg_approval ~ avg_sentiment, data = filtered_lm_df)
Residuals:
    Min      1Q  Median      3Q     Max
-0.17482 -0.04706  0.02179  0.05128  0.08952
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.45554    0.02276  20.018  < 2e-16 ***
avg_sentiment -0.12106    0.03980  -3.042  0.00262 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06628 on 236 degrees of freedom
Multiple R-squared:  0.03773,           Adjusted R-squared:  0.03365
F-statistic: 9.252 on 1 and 236 DF,  p-value: 0.002618
```