A Statistical Analysis of the S5 Assessment Scores from the *Elm City Stories* Study

The Student's T-eam:
Noah Cline, Carissa French, William Laubscher, & Brendan Pinkerton

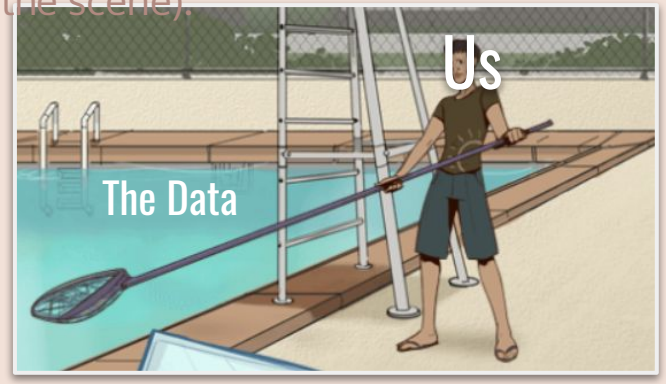
*All non-visualization Images are Screenshots from the game

Approach and Data Analysis

- We thought about the Study's purpose:
  - How did the game affect overall understanding of sexual health and substance abuse?
- We played the game and found:
  - Cyclical gameplay
  - Hard to find answers in just the play data
- Using Game Data *and* S5 Scores
  - Is there a correlation?
  - If so, how are they related?

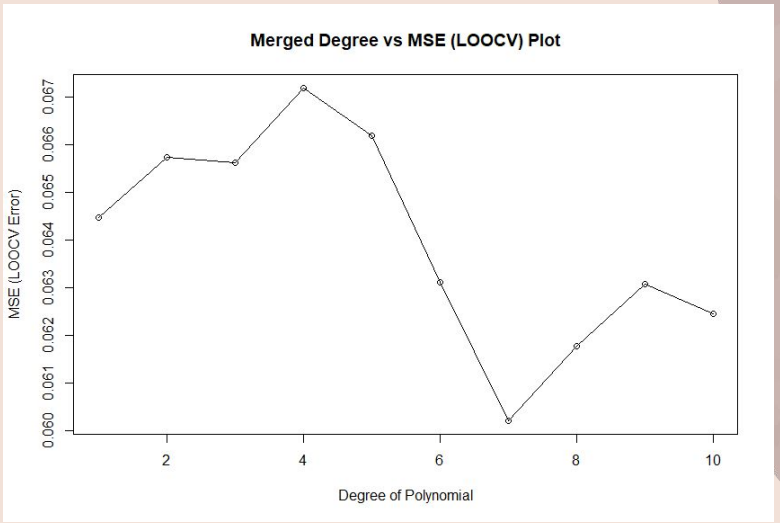
Cleaning the Data

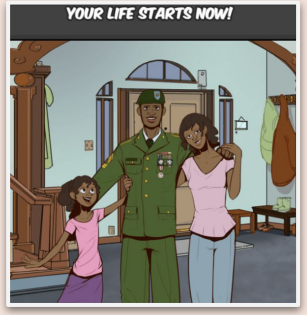
- Data Removed:
  - Players without S5 assessment score data
  - Players with total playtime determined to be an outlier as determined by the IQR method.
    - Many players with anomalous event_time_dbl data (100+ hours played)
  - Entries for Event ID's and variables specific to minigame data.
    - Entries associated with an event_id classified as "Minigame General" were left in.
  - Entries associated with event_id 207 (Panning the scene).
- Left With:
  - 32 potential variables of interest.
  - 178,999 observations.
  - All data associated 43 specific students.



Final Model

- Fit a polynomial model using the max time spent playing and total number of events which occurred during play per player to predict S5 assessment scores.
    - Use Time_Spent for max time spent per player.
    - Use Total_Events for total number of events which occurred during play.
- We selected a 7th degree polynomial model after using LOOCV to determine the polynomial model with the smallest CV error.





R Summary Output

Residuals:

Min	1Q	Median	3Q	Max
-8.1087	-0.3953	0.3319	0.9625	3.0925

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	14.9469	0.1321	113.124	<2e-16 ***
poly(Time_Spent, 7)1	-3.1385	3.1579	-0.994	0.3219
poly(Time_Spent, 7)2	-2.2225	3.2028	-0.694	0.4888
poly(Time_Spent, 7)3	-0.1498	3.1078	-0.048	0.9616
poly(Time_Spent, 7)4	1.3001	3.7210	0.349	0.7273
poly(Time_Spent, 7)5	-6.1390	4.8108	-1.276	0.2040
poly(Time_Spent, 7)6	-2.8423	4.7248	-0.602	0.5484
poly(Time_Spent, 7)7	4.1466	3.3210	1.249	0.2138
poly(Total_Events, 7)1	4.7920	3.6376	1.317	0.1898
poly(Total_Events, 7)2	1.7069	3.2455	0.526	0.5997
poly(Total_Events, 7)3	-1.7737	4.4837	-0.396	0.6930
poly(Total_Events, 7)4	3.4269	4.2735	0.802	0.4239
poly(Total_Events, 7)5	-0.7643	4.6110	-0.166	0.8686
poly(Total_Events, 7)6	-3.9908	3.2448	-1.230	0.2207
poly(Total_Events, 7)7	4.6728	2.5118	1.860	***0.0649*** .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.671 on 145 degrees of freedom
Multiple R-squared: ***0.2214***, Adjusted R-squared: 0.1462
F-statistic: 2.945 on 14 and 145 DF, p-value: ***0.0005692***

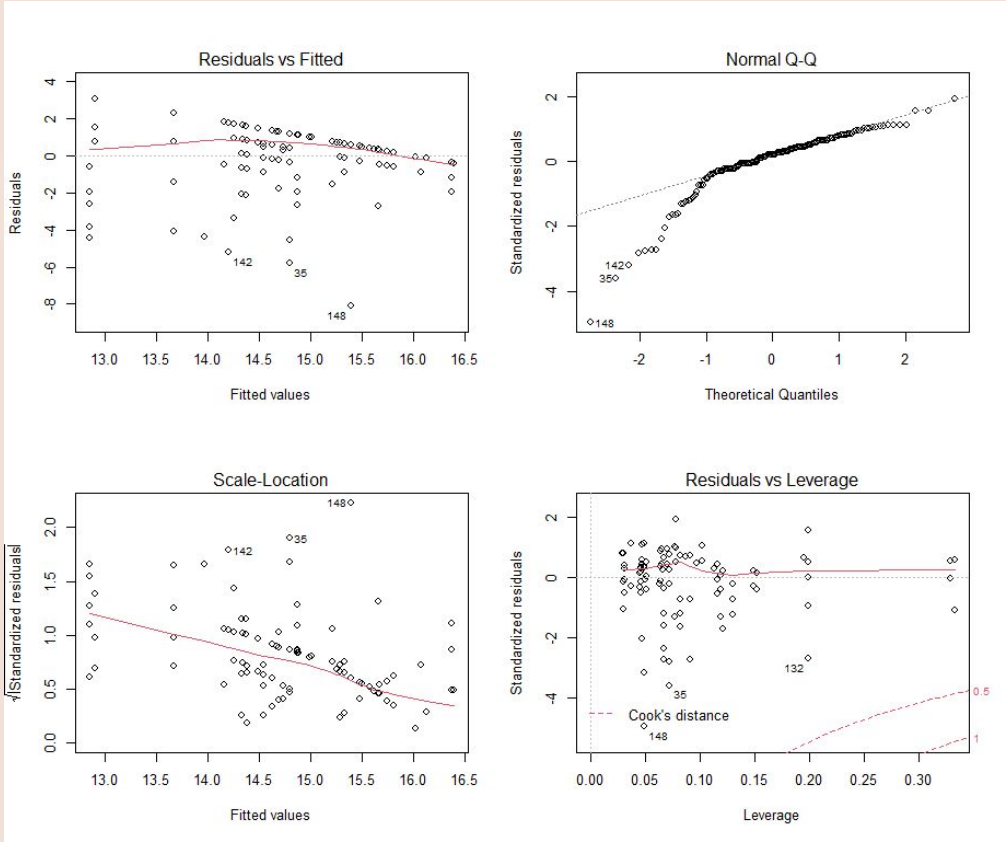
Testing Assumptions

- Shapiro-Wilk Test conducted on all variables
  - Variables are exceptionally non-normal
    - p-value ≅ 0
  - Transformations seemed to have little effect
    - Y^2, log(Y), sqrt(Y), $Y^{1/3}$, etc

- Box-Cox indicated optimal $\lambda = 2$

- Residual plots (see following slide)

Residual Graphs



And that's all, Folks