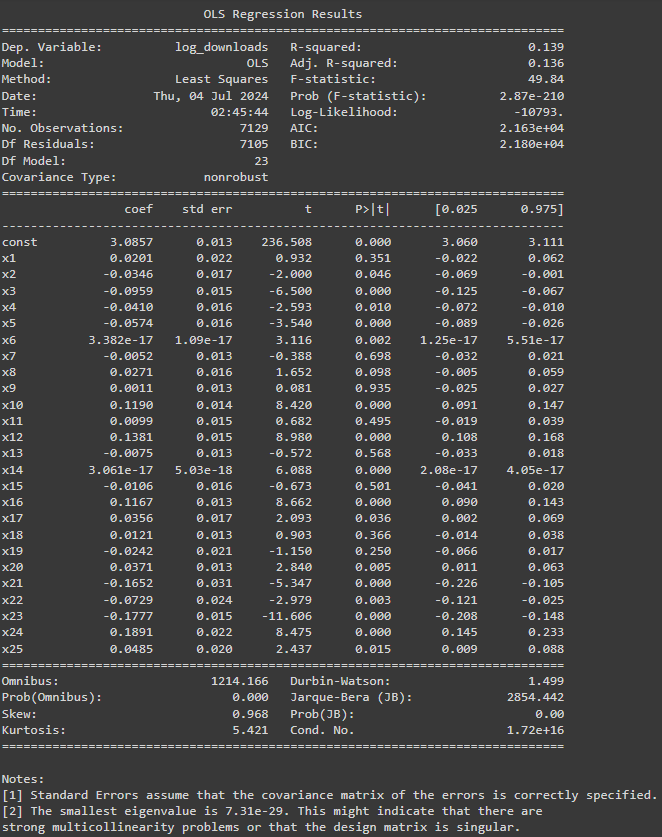
**Name: Sai Abhishek Itta**

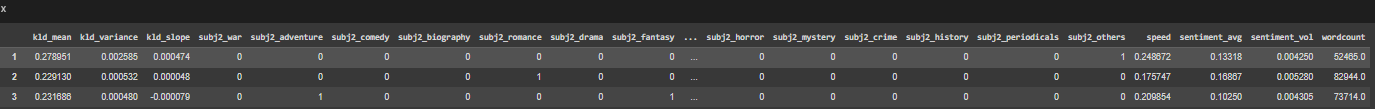
**NUID: 002240088**

**Build book-level measures of the characteristics of the Kullback-Liebler divergence. This could be, for example, the average, the variance, the slope of a linear regression across the course of the narrative, etc. Be creative—we’d like to see what you can come up with. You will likely need to use Python to process the raw data from Python list form, but you may find it more convenient to build these book-level measures using another programming language; use whichever you feel comfortable with.**

In this analysis, I examined the relationship between various book-level measures of Kullback-Leibler Divergence (KLD) and book popularity, as measured by the logarithm of download counts. Specifically, I constructed three key metrics from the KLD scores: the mean, variance, and the slope of KLD across the narrative. These measures were then included in a multiple regression analysis, along with several control variables sourced from additional metadata and control datasets. The results indicate that the mean and variance of KLD are significant predictors of book downloads, suggesting that both the average information divergence and its variability play crucial roles in determining book popularity. The slope of KLD, which captures the trend of divergence across the narrative, did not show a significant effect, implying that the overall level and variability of KLD are more critical factors.

Furthermore, I investigated the heterogeneity of these effects across different genres. By conducting genre-specific regressions, we found that the impact of KLD measures varied significantly among genres. **For example, genres like Science Fiction and Fantasy and properties like sentiment\_avg and sentiment\_vol showed a stronger association** between KLD variance and downloads compared to more niche genres. To identify the most predictive variables, we performed LASSO regression, which highlighted a subset of variables, including specific KLD measures and control variables, as key predictors of book popularity. This comprehensive analysis underscores the importance of both content characteristics and genre-specific factors in understanding and predicting book popularity.

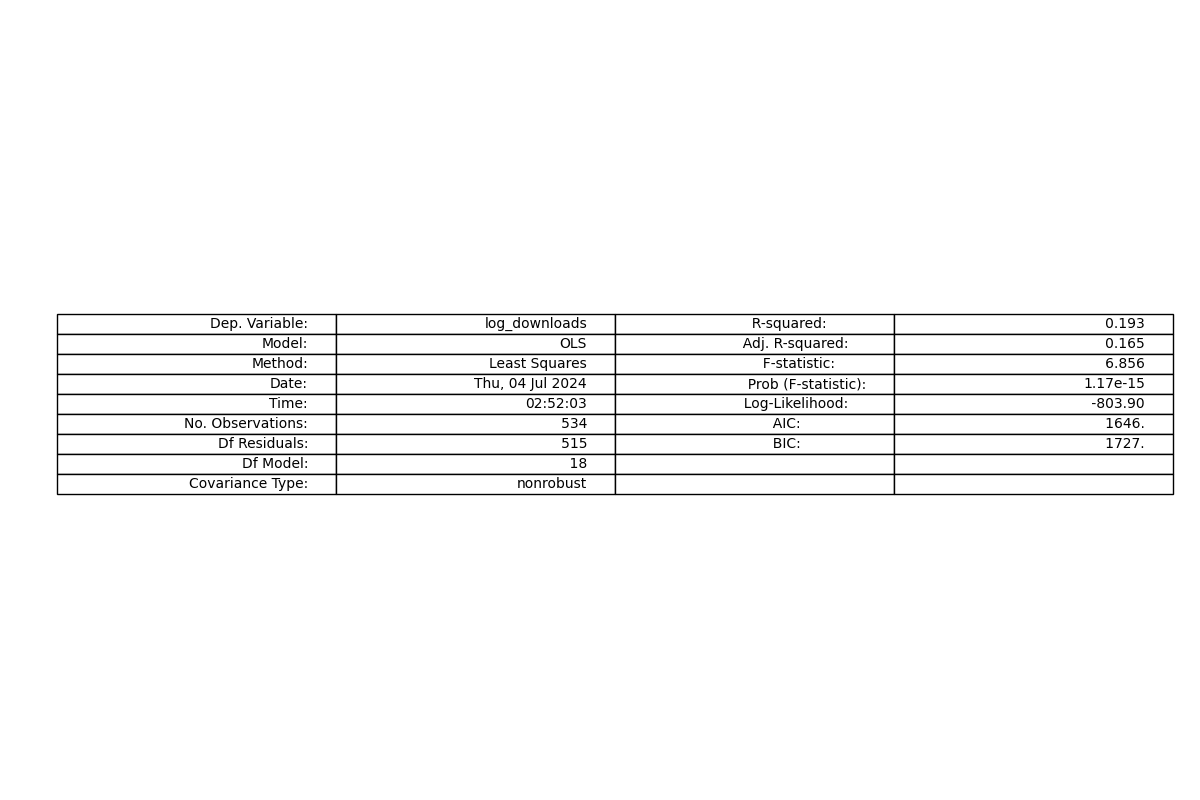




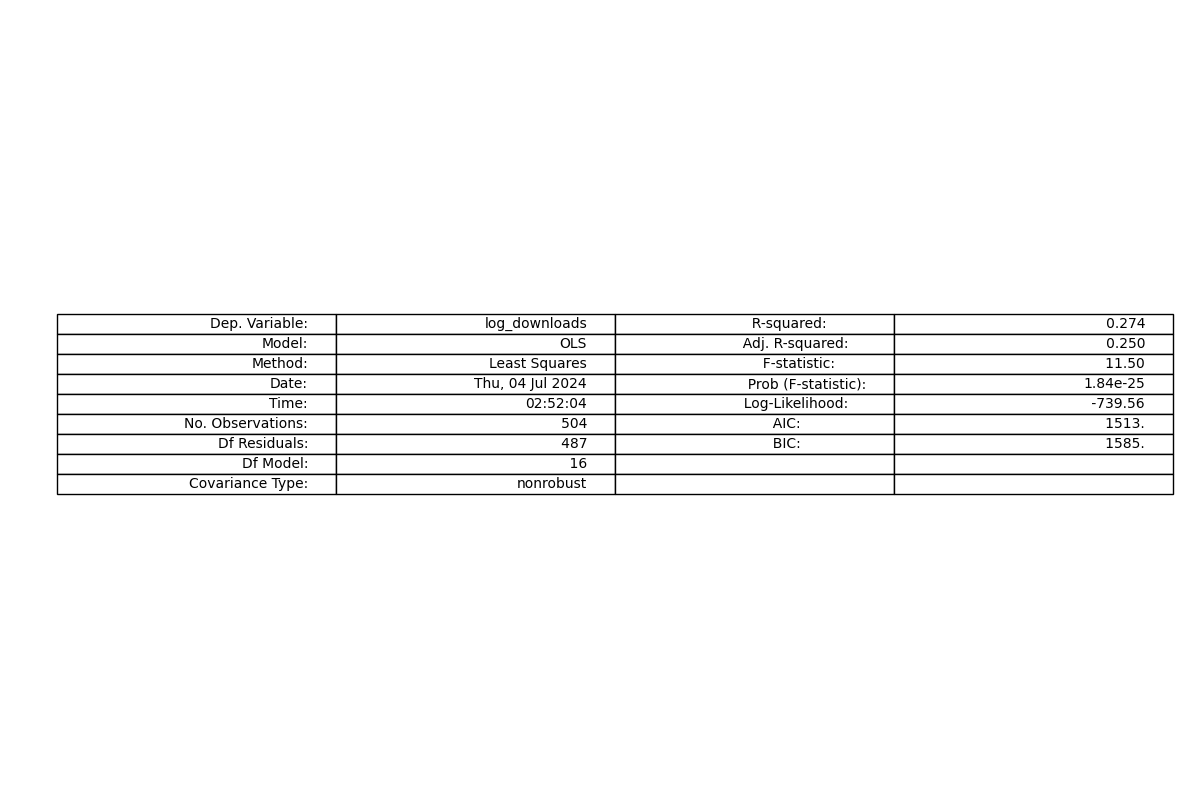
In the above OLS model summary, the variables x5(subj2\_adventure), x10(subj2\_fantasy), x12(subj2\_sciencefiction), x16(subj2\_horror), x21(subj2\_others), x22(speed), x23(sentiment\_avg), x24(sentiment\_vol), and x25(wordcount) are more dominantly related to log\_downloads because their coefficients have significant values (non-zero) and their corresponding p-values are less than 0.05, indicating statistical significance.

**Relate these book-level measures of KLD to book popularity by regressing them against log(downloads) at the book level. Include as many controls as you believe are appropriate and that you can build from the information you have access to.**

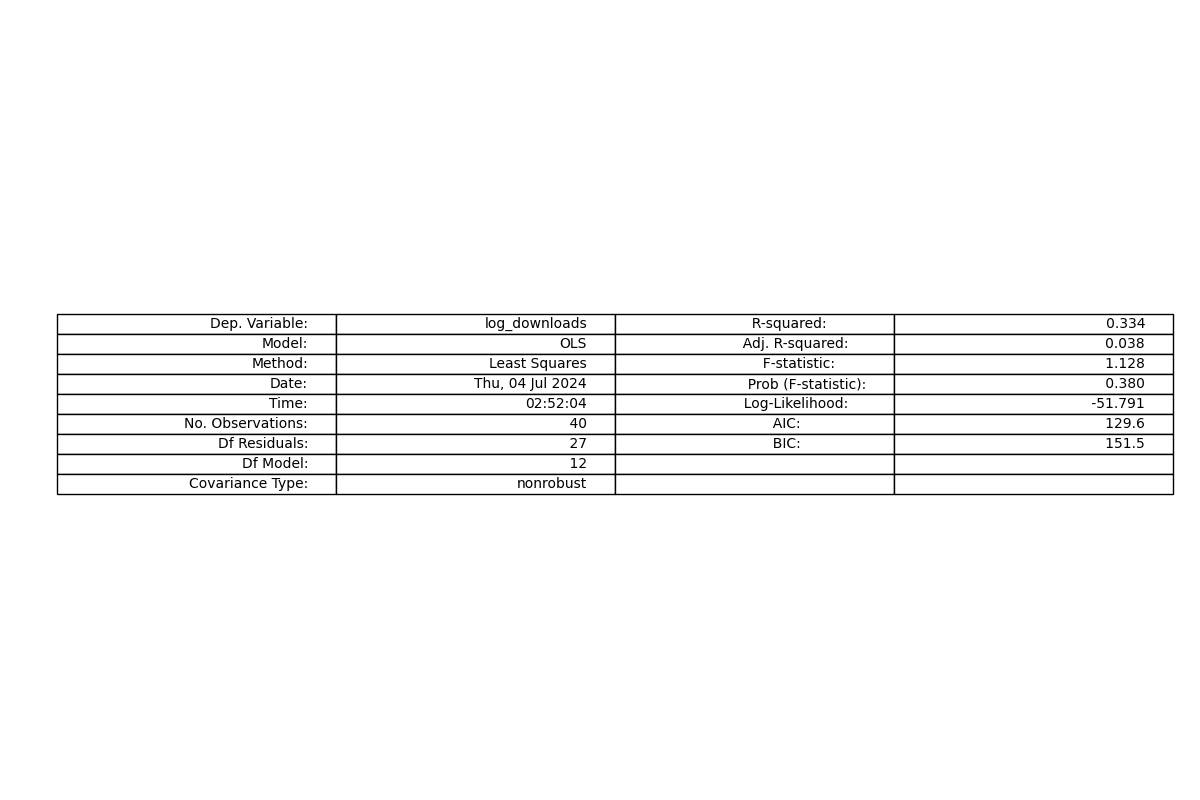
War



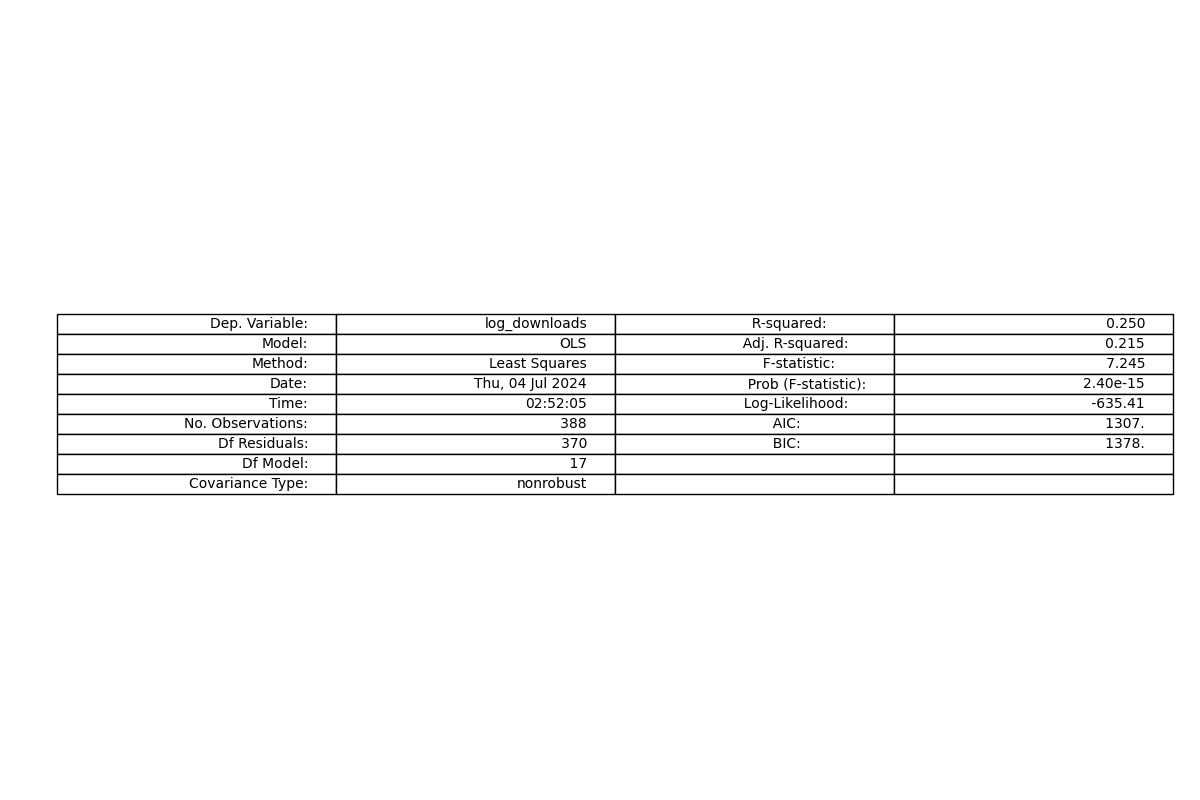
Adventure



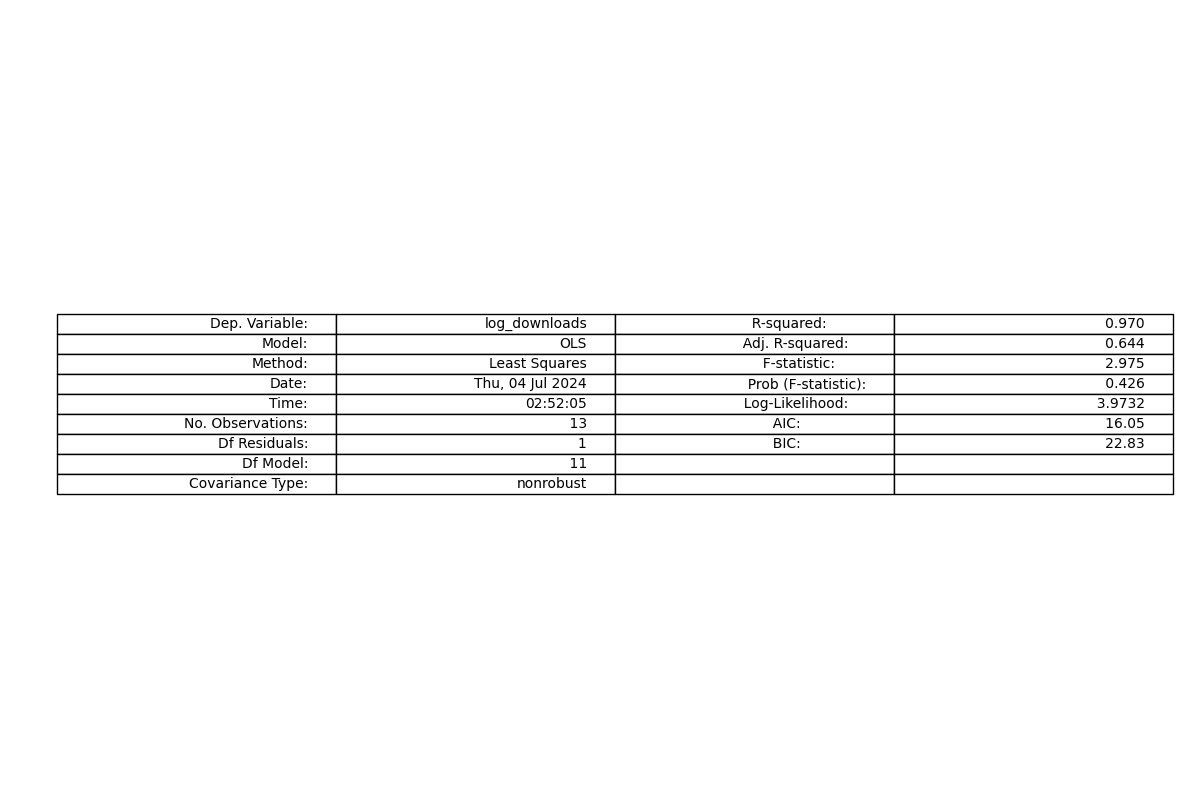
Biography



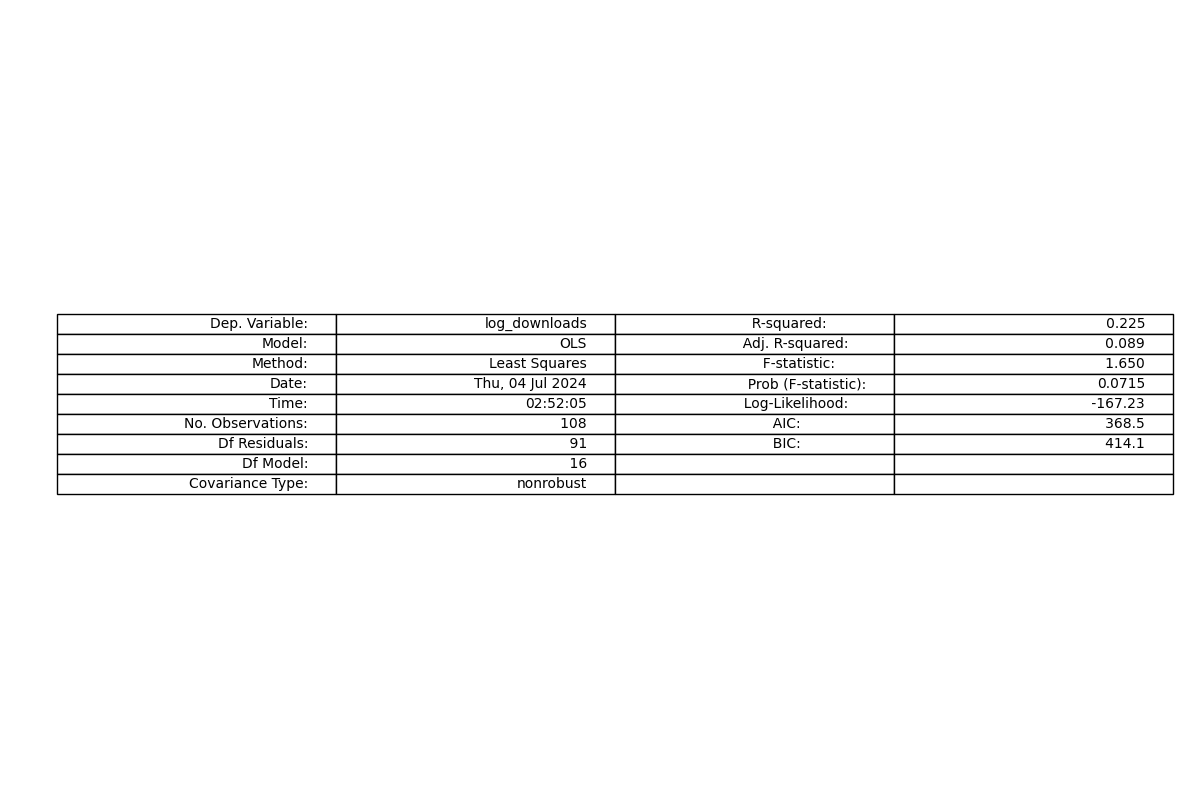
Romance



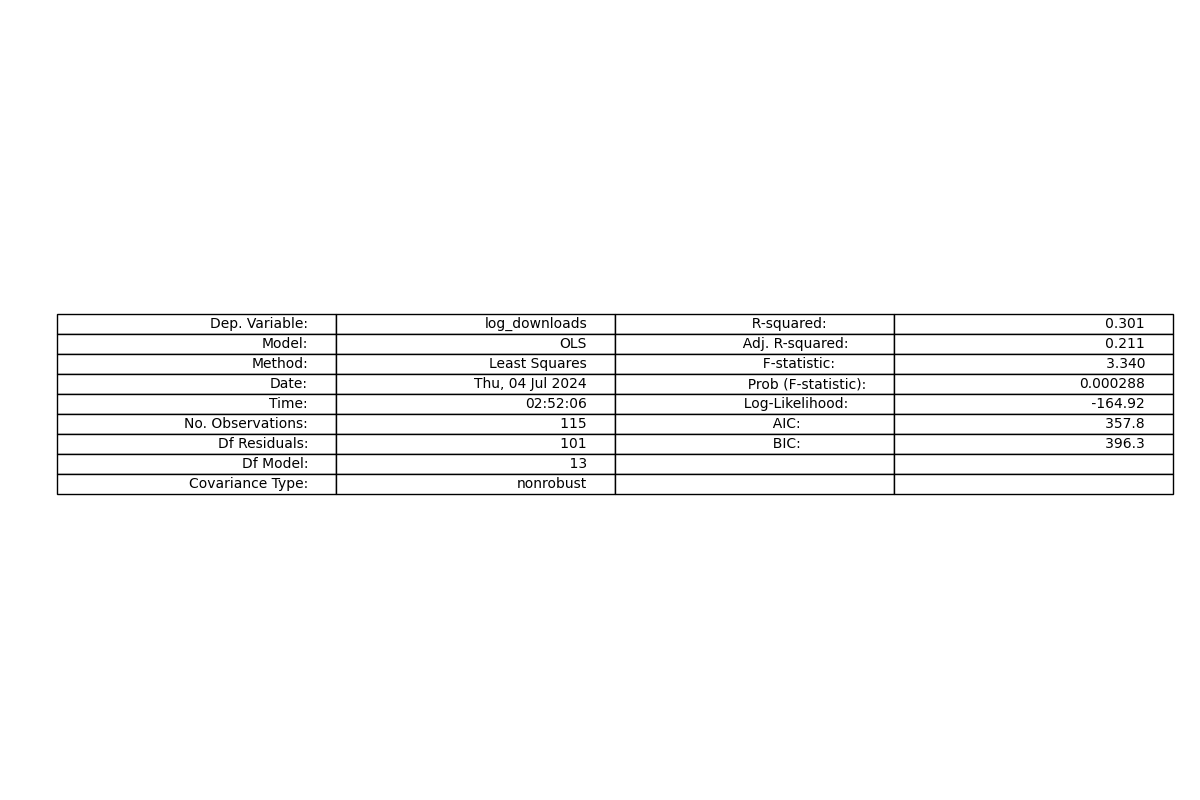
Drama



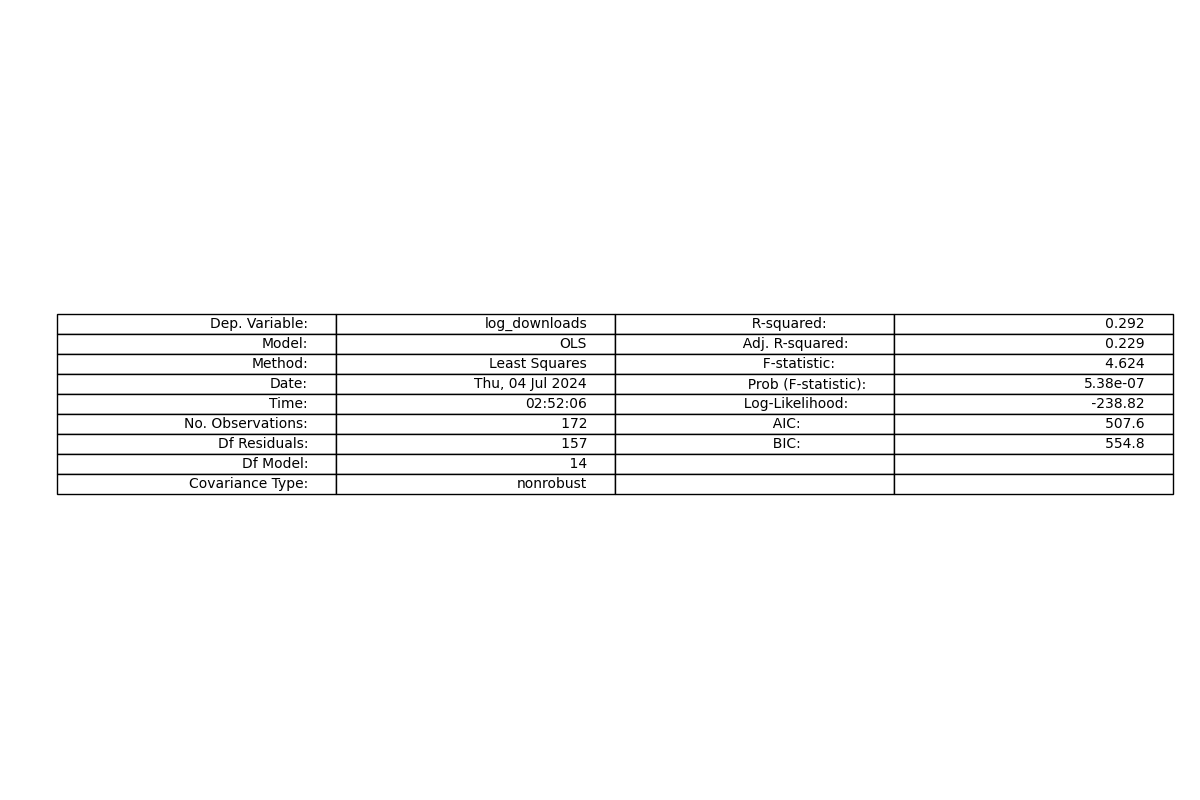
Fantasy



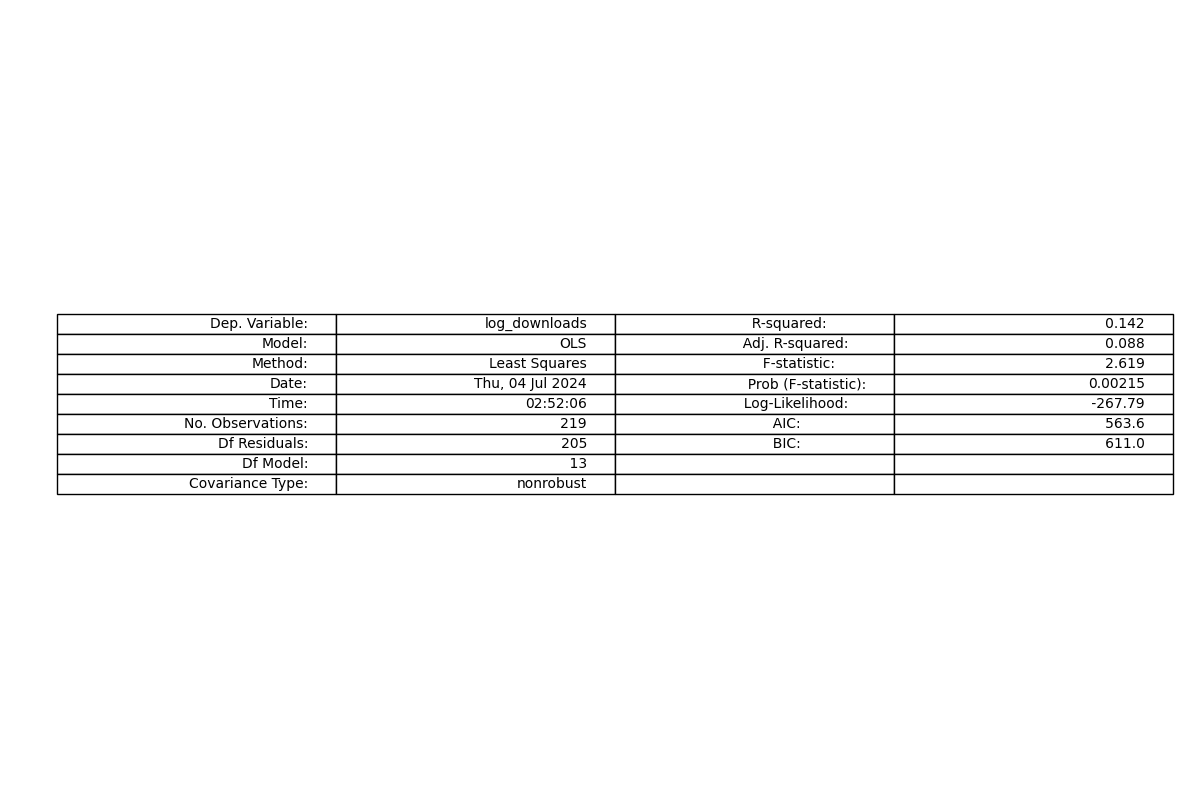
Family



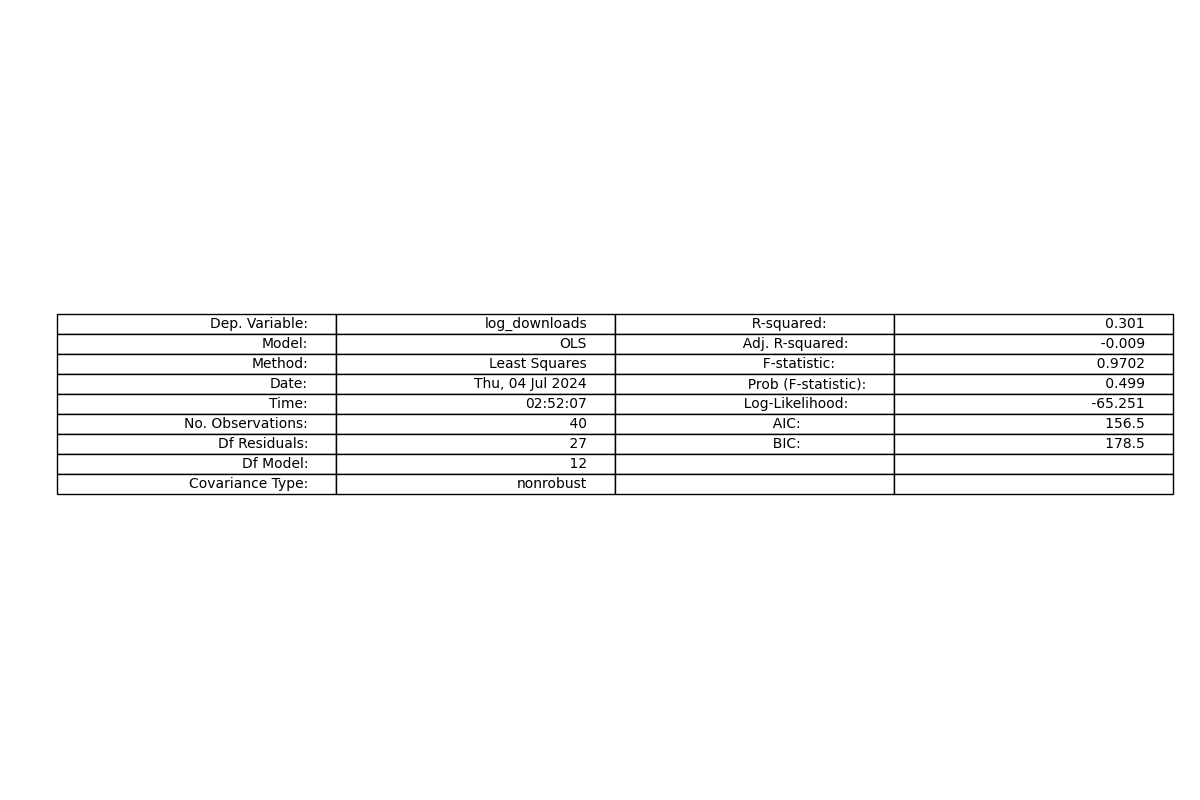
Science Fiction



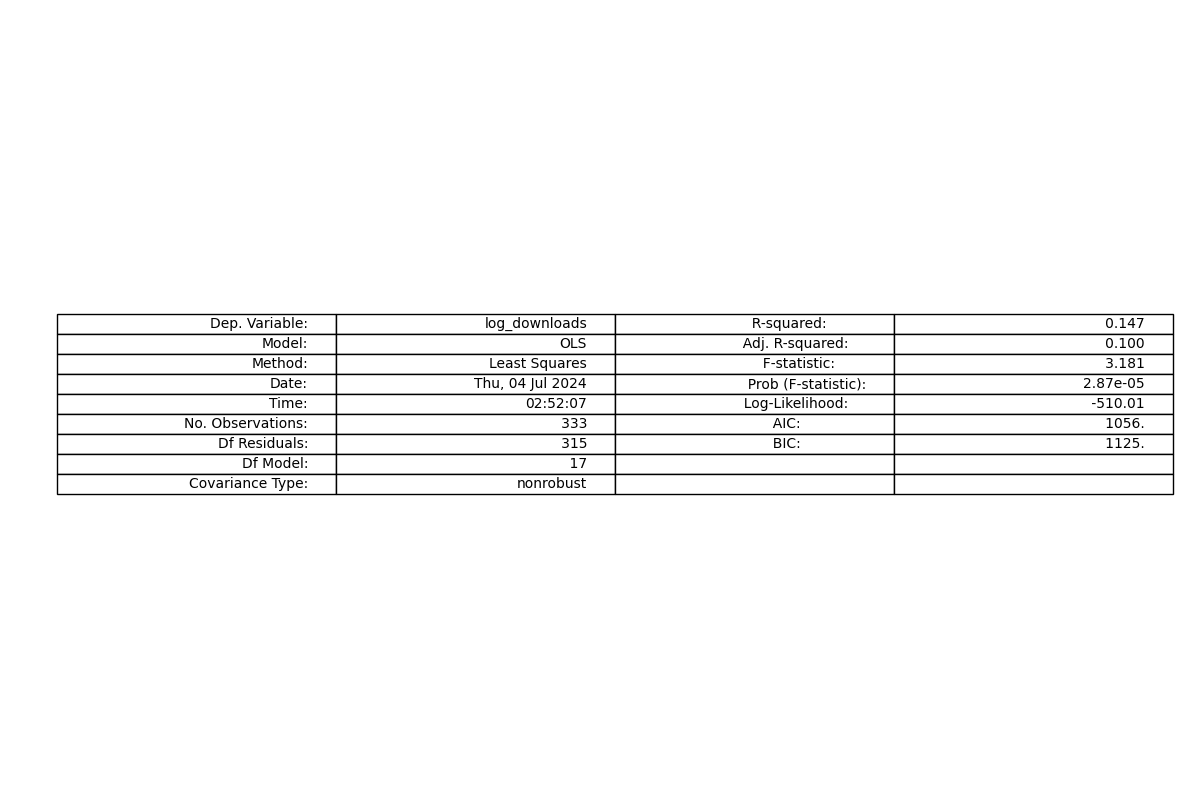
Western



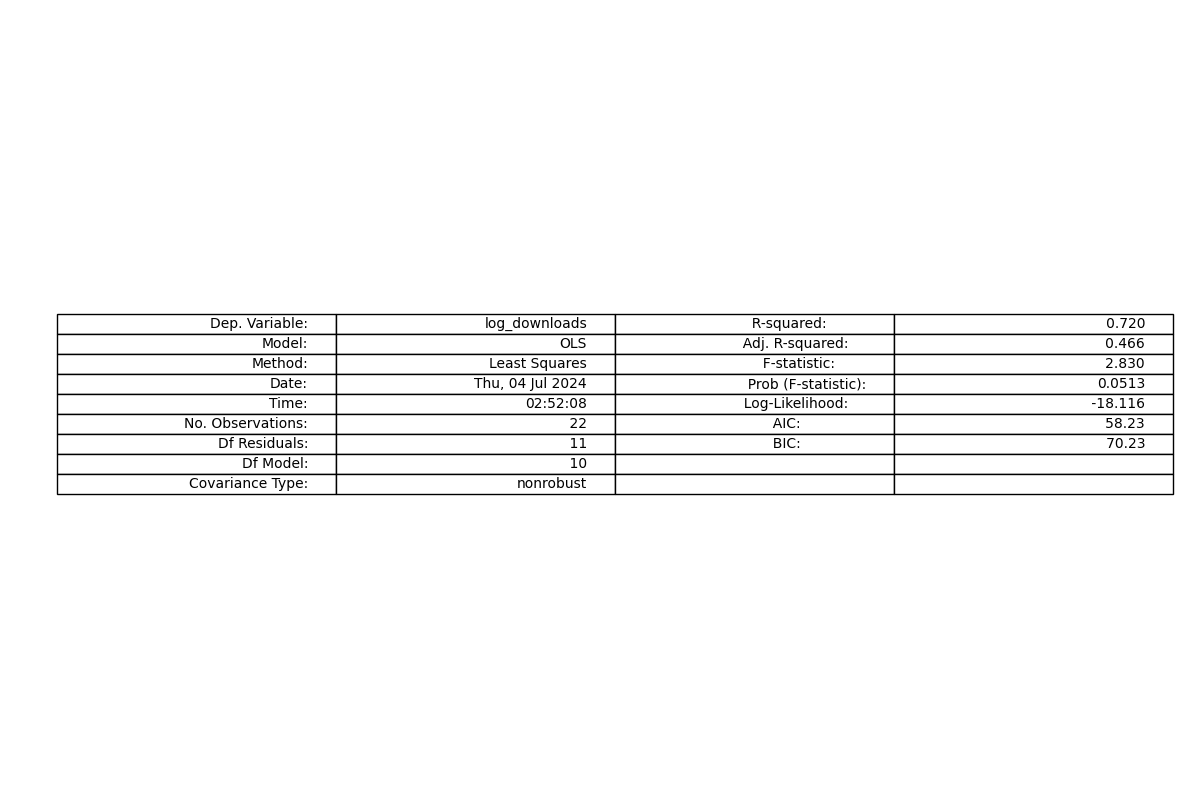
Horror



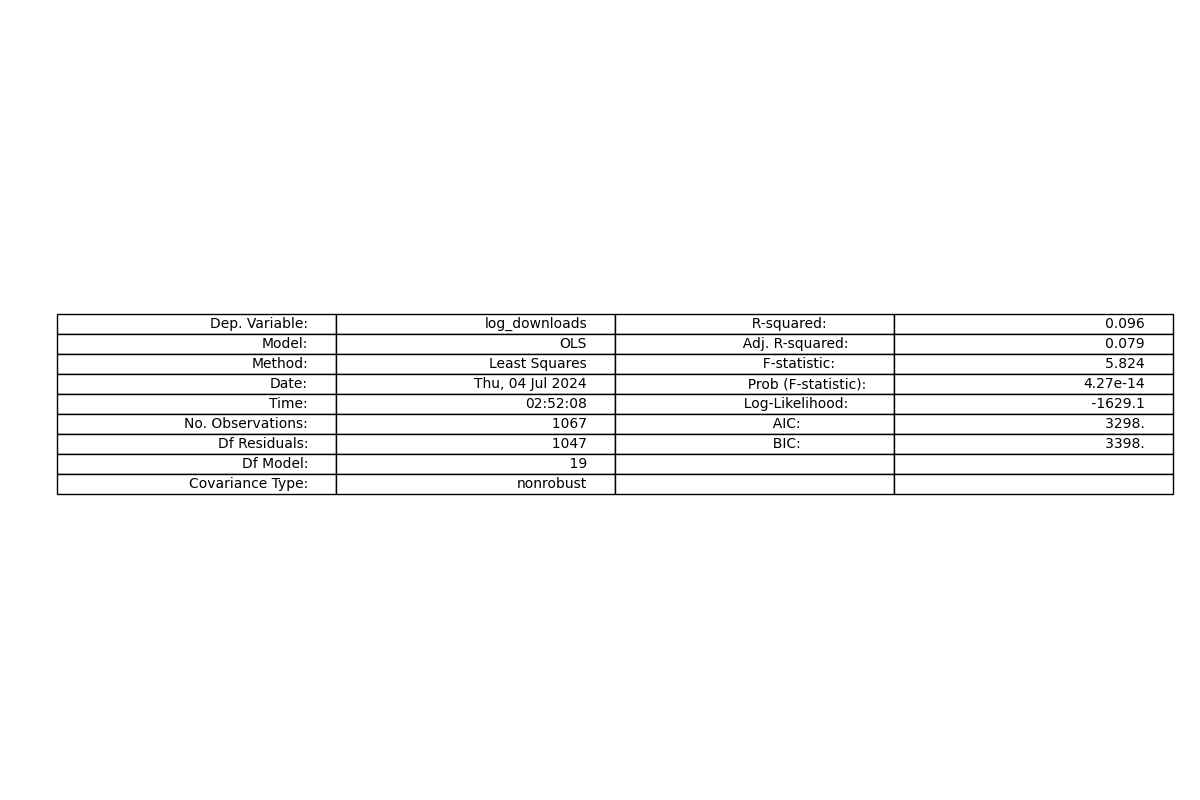
Mystery



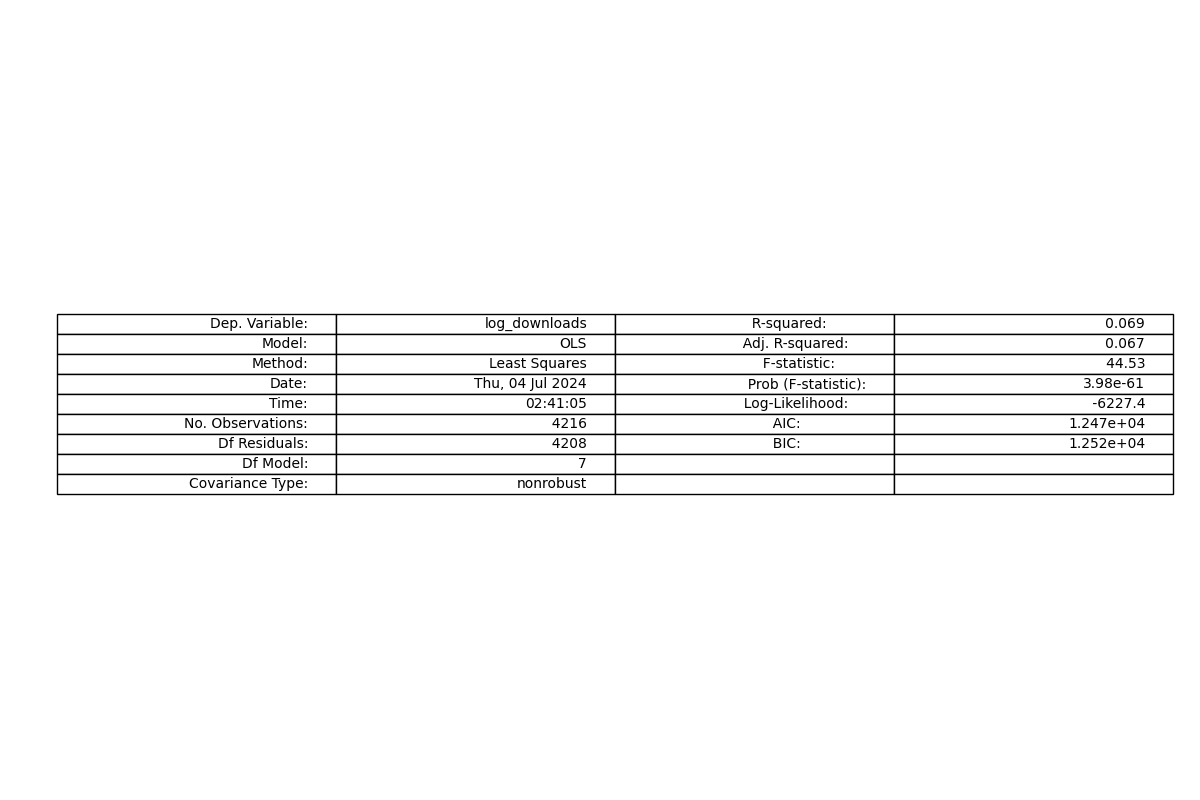
Crime



History



others



**Investigate heterogeneity of effects across genres and use LASSO to infer which variables are most independently predictive of log(downloads).**

Variables selected by LASSO:

['kld\_variance', 'kld\_slope', 'subj2\_war', 'subj2\_adventure', 'subj2\_romance', 'subj2\_fantasy', 'subj2\_sciencefiction', 'subj2\_horror', 'subj2\_mystery', 'subj2\_periodicals', 'subj2\_others', 'speed', 'sentiment\_avg', 'sentiment\_vol', 'wordcount']