**Explore Data:**

* Inspect data and it features
* Compute descriptive statistics
* Use data visualization to help identify significant patterns and trends in the data

**Questions to answer (& hypotheses to check):**

1. How does the title of an article affect its shares?
2. When are most articles published? (Wednesday)
3. Which channel content are users more interested in reading? (No significant difference between mean shares across channels. Distribution of articles across channels is plotted)
4. Should the content be positive, negative or neutral? (We did some stat tests and boxplot visualizations)
5. Does media information (images, videos) affect the posts’ success? (for log(Y) transform, Wilcox test gives statistically significant difference)
6. Are linear models appropriate for usage here? (LogY GLM has potential, complies with 3 assumptions of linear models)
7. What are the most important parameters that affect shares number?
8. What is the best criterion for comparing models? (Cross validation adjusted R^2, then AIC, then BIC)
9. Is there a difference in number of shares of articles, published on weekends, against articles, published on workdays? (No)
10. Cross – validation for models – do that and calculate mean adjusted R^2.
11. To the section with models’ comparison, add comparison of models based on adjusted R^2, calculated on the test set
12. Select the best model – do the demo (separate file, where we load the data, create test / train subset and demonstrate the results)

**Not done**

**Done**

**Taken (In process)**

**Tasks:**

1. Nan check + Infinity check
2. Check histograms distribution
3. Root transform to skewed cols (to fix skewness)
4. Find and plot columns with outliers (IQR analysis)
5. Do IQR filtering (?? May be not necessary 🡪 understand the reason for bimodal distribution for many features. Answer – do we need to **filter out the outliers?**)
6. Top down / down top model selection
7. Trying log transform for Y and write down the effect (tried full model with log(Y) 🡪 R^2 became better, from 0.11 to 0.17, and residual plot became better) (In this case this is a Poisson GLM model, and we should compare linear selected model with this GLM)
8. Analyze scale-location plots and residual plots of models (script #3)
9. Check the interaction effects between different variables (script #3)
10. Apply anova, ancova tests, figure it out
11. When we prepare all models (top-down selected, down-top, GLM, …) – compare all of them on validation set and present final results on the test set