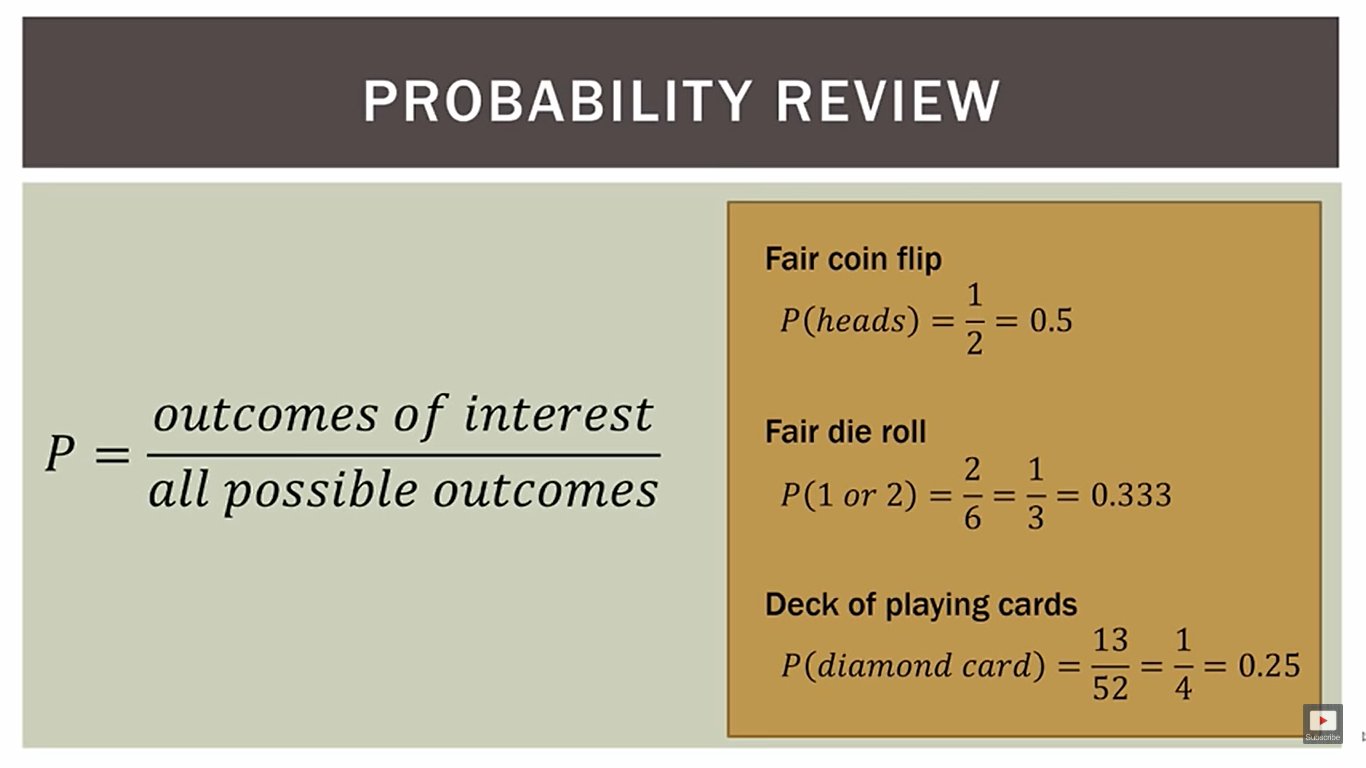
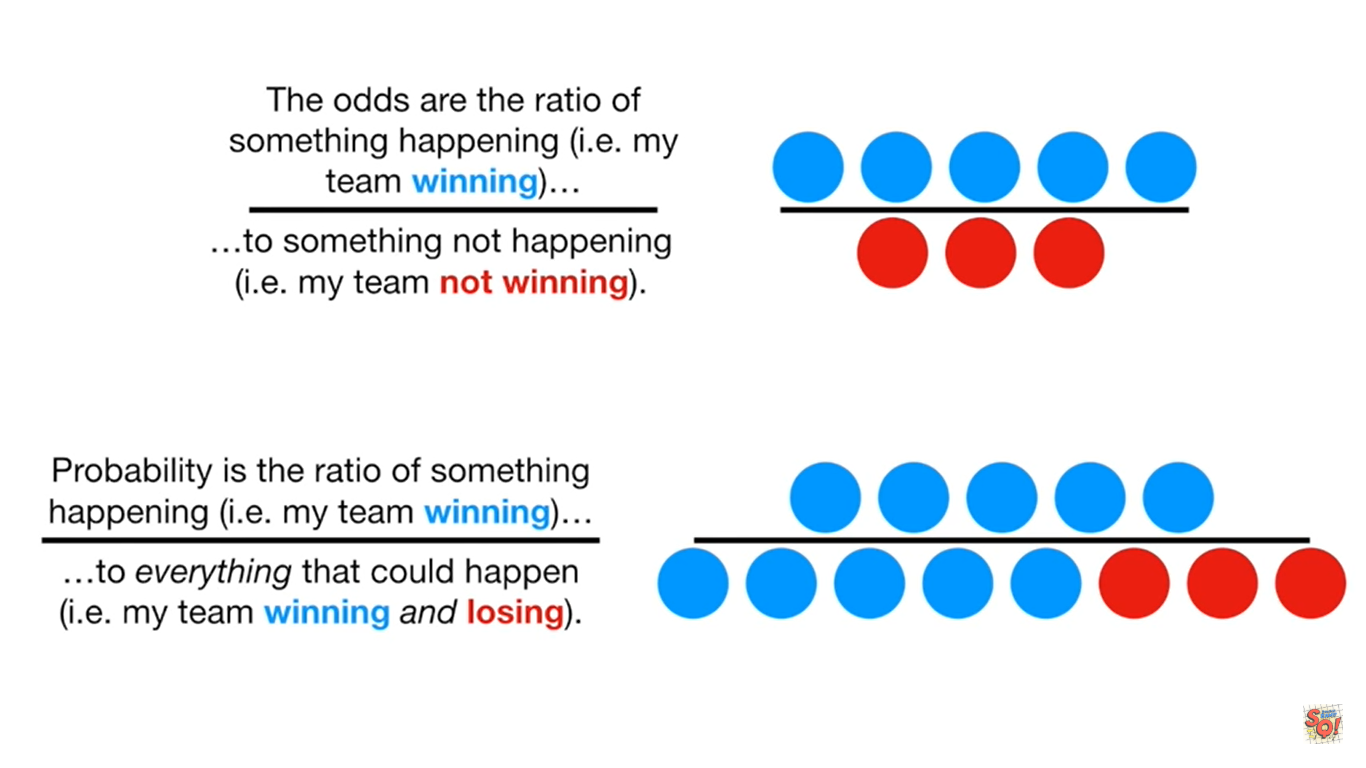
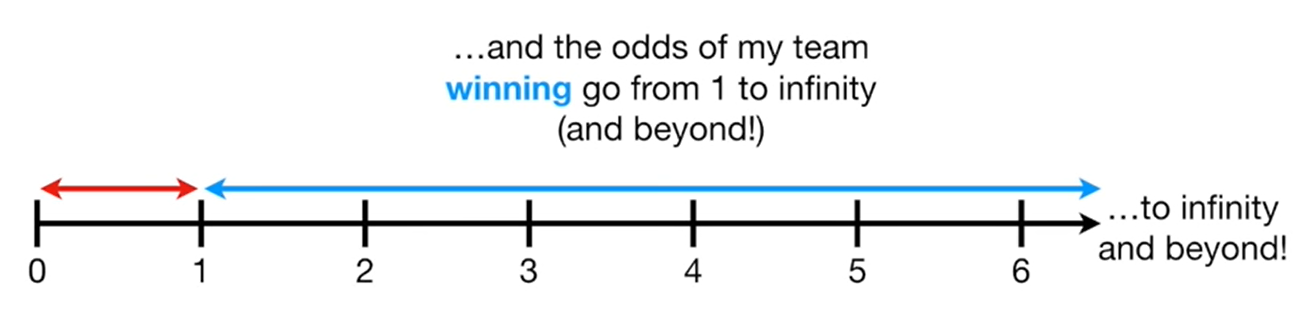
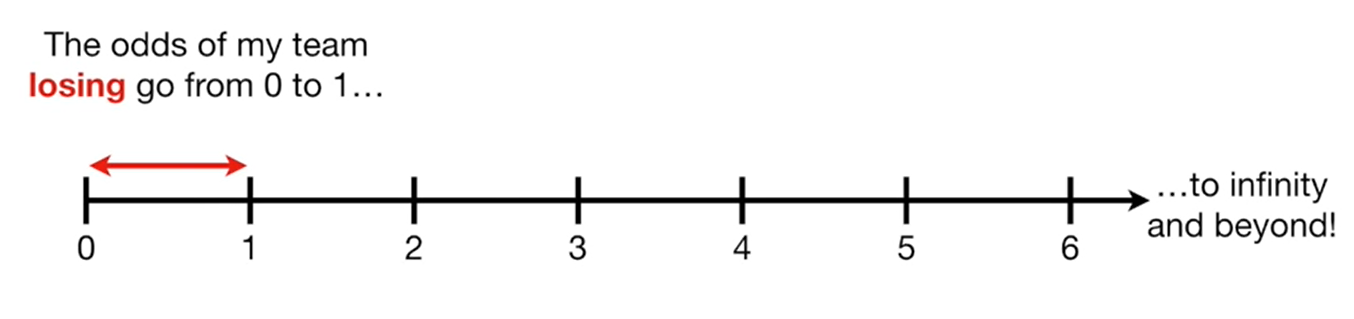
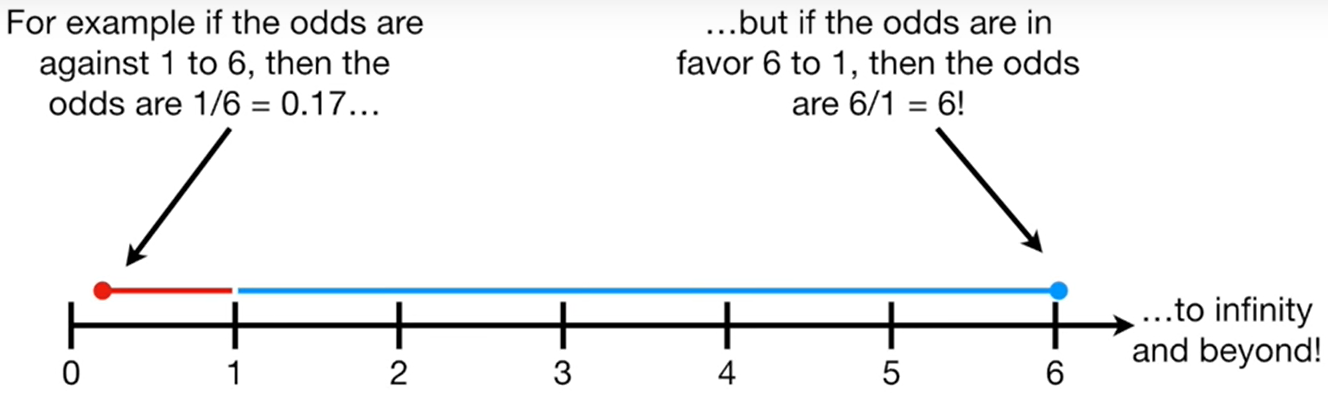
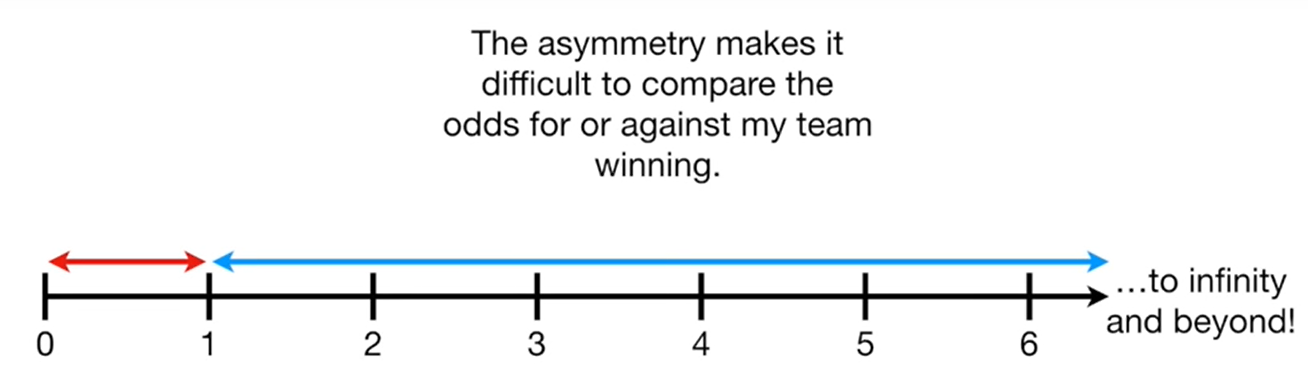
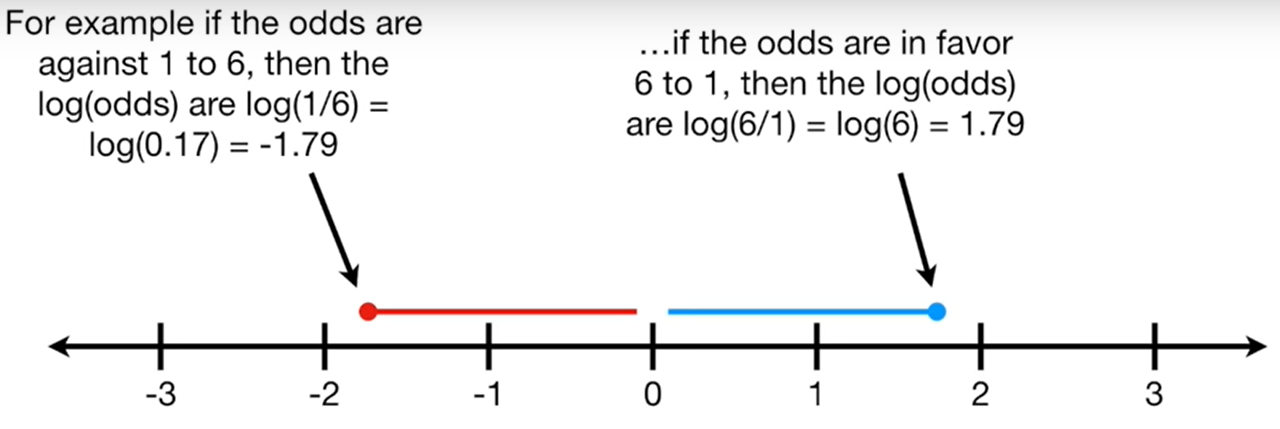
1. Import Data [excel, sockets, SQL] and Analyze,
   1. specify dependent and independent variables [Y = Classification or Continuous]
   2. Check if it is a linear or non-linear model
   3. Peak into the Dataset, glimpse(), str(), nrow(), ncol()
2. Visually Analyzing Dataset for ‘variance of variables’ by Plots [ggplot2]
3. Data Purification
   1. N/A value and Missing value Treatment
   2. Dummy Variable Creation [Created Dummy - 1]
   3. Deleting Unwanted Variables [Identity Number or 50% + empty values]
   4. Manipulation [(s|l|t)apply functions, dplyr, tidyr]
   5. Outlier Treatment [Cooks Distance]
4. Splitting Data into {Training, Testing} sets or K – Cross Validation
5. Training the Model [Applying the formula and getting the fit]
6. Analyzing the below factors and Evaluating the model
   1. Residuals
   2. Coefficients
      1. Estimate
      2. Std. Err [Lesser is better]
      3. T Value
      4. P-Value [Lesser than 0.05 reject H0, variable has significance]
   3. Residual Standard Error with ‘N - 1’ degrees of freedom [RMSE]
   4. Explained Variability or Coefficient of Variance ‘R^2’ [Greater is better]
      1. Multiple R^2
      2. Adjusted R^2 [ideally should be 76 – 85 %]
   5. F-Statistic
7. Eliminate Multicollinearity by removing variables with VIF>5
8. Final model = Run stepAIC() [‘Forward’ ‘Backward’ ‘Both’] to select most significant variables and prepare. [Lesser is better]
9. Predict based on Final Trained model
   1. Step 6 for Accuracy











Odds: Event ‘A’ Happening / Event ‘A’ NOT Happening

To minimize scale and make the set symmetricr we would do a log(odds)

