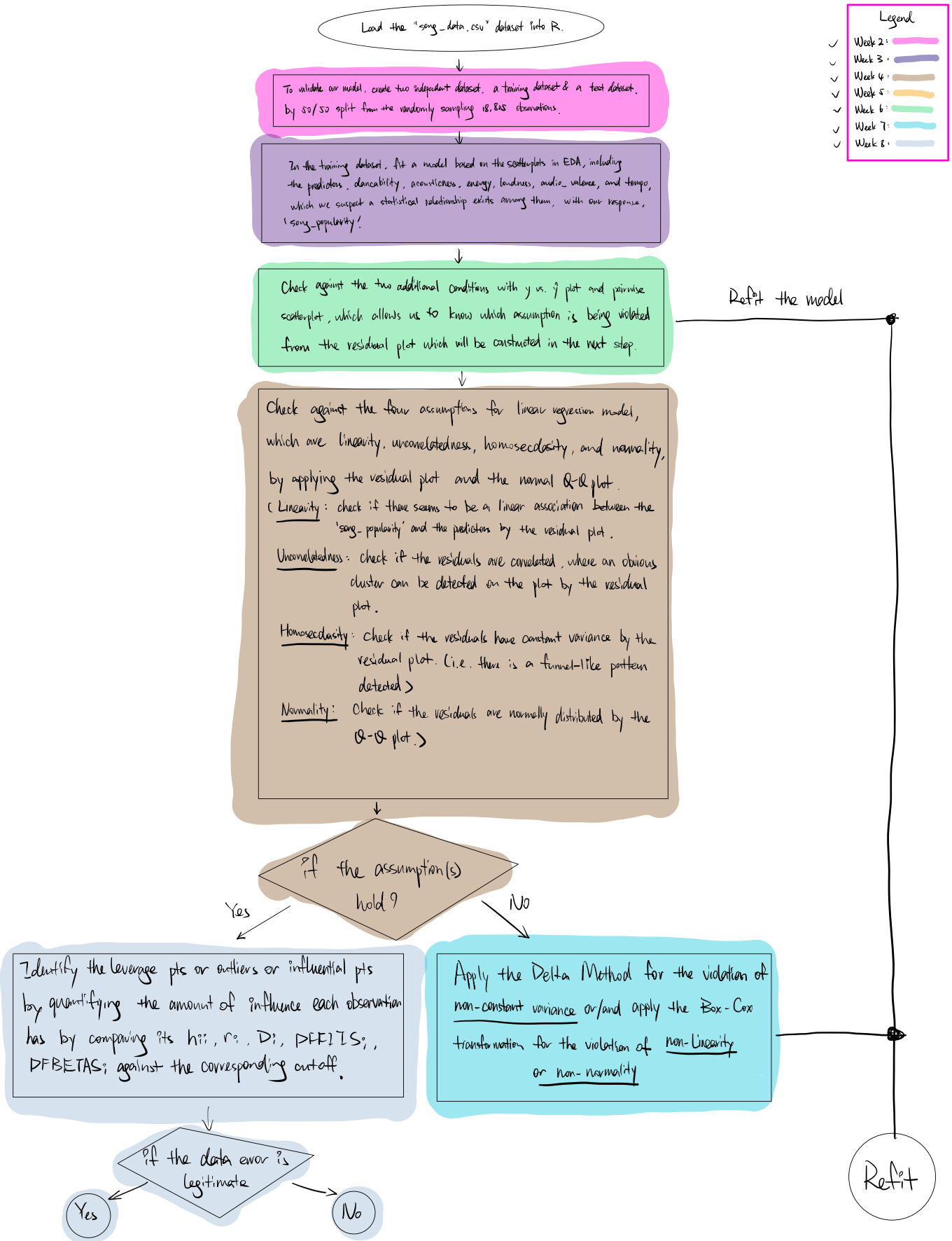
Load the "song_data.csv" dataset into R.
To validate our model, create two independent datasets, a training dataset & a test dataset, by 50/50 split from the randomly sampling 18,835 observations.
In the training dataset, fit a model based on the scatterplots in EDA, including the predictors, danceability, acousticness, energy, loudness, audio_valence, and tempo, which we suspect a statistical relationship exists among them, with our response, 'song_popularity'.
Check against the two additional conditions with y vs. ŷ plot and pairwise scatterplot, which allows us to know which assumption is being violated from the residual plot which will be constructed in the next step.
Refit the model
Check against the four assumptions for linear regression model, which are linearity, uncorrelatedness, homoscedasity, and normality, by applying the residual plot and the normal Q-Q plot.
(Linearity: check if there seems to be a linear association between the 'song_popularity' and the predictors by the residual plot.
Uncorrelatedness: check if the residuals are correlated, where an obvious cluster can be detected on the plot by the residual plot.
Homoscedasity: check if the residuals have constant variance by the residual plot. (i.e. there is a funnel-like pattern detected)
Normality: Check if the residuals are normally distributed by the Q-Q plot.)
If the assumption(s) hold?
Yes
No
Identify the leverage pts or outliers or influential pts by quantifying the amount of influence each observation has by comparing its hii, ri, Di, DFFITSi, DFBETASi against the corresponding cutoff.
Apply the Delta Method for the violation of non-constant variance or/and apply the Box-Cox transformation for the violation of non-Linearity or non-normality
If the data error is legitimate
Yes
No
Refit
Legend
Week 2:
Week 3:
Week 4:
Week 5:
Week 6:
Week 7:
Week 8:


Yes
No
note the observations as a limitation
remove it and refit the model
Refit
Apply F-test to the dataset by employing the ANOVA table and summary table to detect if there exists an overall significant linear association between the predictors and the response. Here hypothesis test is being used.
Significant association is detected ?
No
no linear association can be found between the response and the predictors.
yes
Check the t-tests against the summary table to see if each predictor is significant to the response. Here hypothesis test is being used.
Significant association is detected ?
No
remove the according predictor(s) by checking its p-value in the summary table
Refit
Yes
Compare different models by checking the partial F-test against the ANOVA table. Here, a hypothesis test is being carried out. To select a more plausible model to account for the associations between the response and the predictors by checking the adjusted R^2.
The selected plausible model is to be validated against the test dataset created at the second step in terms of its performance
Is the selected one valid ?
Valid
Invalid
Make inferences on our response in our model by establishing confidence intervals and hypothesis tests. Finally, make a discussion on our model.
Discuss the limitation of the model and any plans can be carried out in the future.