Mid-term timeline project

1. **Preprocess & clean data** (anticipated duration: 1-1.5 days)
   1. Familarize with database structure (review data\_description files) (<https://docs.google.com/spreadsheets/d/1kH3Ht5K22qIHLqLddmyDAvKDSsPI9vLVCd1BoUaztyE/edit?usp=sharing>)
   2. Initial Outlier Detection and Treatment
   3. Treatment of Nulls
   4. Dummys
2. **EDA** (anticipated duration: 2 days?)
   1. Create additional hypotheses
   2. Bare minimum complete the 10 tasks outlined in the jupiter notebook
3. **Feature selection** (trees? LDA / PCA? Linear regression?) (anticipated duration: 0.5 days)
   1. Ideally we will have a good list of things we are interest in from EDA
   2. May need to loop back during the modelling phase
4. **Model selection** (anticipated duration: 2 days?)
   1. Split our data
   2. Select metrics to compare among them
   3. Spot check different type of algos
   4. Tune selected ones using grid search
   5. Feature Engineering
   6. Feature Selection/Dimensionality Reduction
   7. MODEL TYPES:
      1. Regression (target variable ARR\_DELAY)
      2. Multiclass Classification (target variables CARRIER\_DELAY, WEATHER\_DELAY, NAS\_DELAY, SECURITY\_DELAY, LATE\_AIRCRAFT\_DELAY)
      3. Binary Classification with class imbalance (target variable CANCELLED)
5. **Train model** (use K-folds) (anticipated duration: )
6. **Predict** (anticipated duration: )
7. **Compare results** from selected algos based on selected metrics, make killer **visualizations**!!! (anticipated duration: )
   1. Start ideally by/before Wednesday Evening/Thursday Morning
8. **Slide and presentation prep** (anticipated duration: )
   1. Start ideally by/before Thursday Afternoon?
   2. CONTENT:
      1. 1x Slide for each task from Exploratory Analysis
      2. Sampling Strategy(ies)
      3. Modeling Techniques
         1. Metrics for various techniques
         2. Which model is best and why
      4. Evaluation metrics
      5. Feature Importance
      6. Challenges

* Friday
  + Git prep
  + Planning
  + Understanding Scripts
  + Data Cleaning
* Saturday
  + Data Cleaning
  + EDA
* Sunday
  + EDA
* Monday
  + Review of EDA - reformulating any hypotheses
  + Modeling/Training/Evaluation/Tuning
* Tuesday
  + Modeling/Training/Evaluation/Tuning
* Wednesday
  + Modeling/Training/Evaluation/Tuning
  + Tweaks and Model Updates
  + Running Final Predictions
* Thursday
  + Data visualization
  + Code and Cleanup
  + Presentation and Submission Prep
* Friday
  + Final Touches and Presentation Prep
  + Present and Submit