1. Exploratory
   1. Info
   2. Dtypes
   3. **Transform datatypes**
   4. **Split categorical and continuous columns into vectors**
   5. Missing values
      1. **Check reference's missing values, if we need to impute our missing values**
      2. **Categorical - if we're imputing, we should round them off - should decide based on no. of missing values in reference**
   6. Drop rows with NA's *(should we impute or just drop?)*
   7. Check rows with wrong values
      1. Age = 'R', replace with 0, change datatype to numeric
   8. Check columns with spaces/commas
      1. Work load Average/day column name
      2. Change datatype to numeric
   9. Box plots
      1. Age
      2. Work\_load\_Average
      3. Body mass index
      4. Weight
      5. Height
   10. Outlier Treatment
       1. Identify outlier in columns
       2. Replace outliers with lower and upper bound values
       3. Currently done: Age, Work\_load\_Average, Body mass index, Weight, Height
       4. To be added: other columns
       5. New box plots
   11. Correlation: Correlation value and heat map
   12. Plots:
       1. Bar charts for categorical variables
       2. Histograms for continuous values
   13. Scatterplots:
       1. For highly correlated variables
          1. Age ~ Absenteeism time (hrs)
          2. Reason for Absence ~ Absenteeism time (hrs)
          3. Day of the week ~ Absenteeism time (hrs)
       2. **Include other variables in scatterplots - grid format**
2. **Feature Selection**
   1. **Continuous variables - create dataframe**
   2. **Check for high correlation from plot - list out combinations of variables**
   3. **Should we remove any columns?**
3. **Feature Scaling**
   1. **Normality check? Histograms?**
   2. **Normalization of continuous variables**
4. **Machine Learning Models**
   1. **Option 1** 
      1. **Decision Tree (with and without PCA Dim Reduction)**
      2. **Random Forest (with and without PCA Dim Reduction)**
      3. **Linear Regression (with and without PCA Dim Reduction)**
   2. **Option 2**
      1. **Decision Tree (Classification and Regression)**
      2. **Random Forest (Classification and Regression)**
      3. **Linear Regression?**
   3. **Option 3 - our option**
5. **Results**