1. **Attività preliminare**
   - **1.1. Analisi esplorativa della variabile di risposta**
   - **1.2. Data Cleaning e data trasformation**
     - **1.2.1. Trasformare in factor le variabili categoriche**
     - **1.2.2. Eliminare le proxy (**correlazione tra le X**).** Risolvere problemi di collinearità
     - **1.2.3. Controllare la corretta assegnazione del dato INTEGER NUMERIC STRING ecc**
   - **1.3. Ricerca degli outliers (**si evidenziano per le y molto elevate in alto**) e high levarage point (**si evidenziano per x molto grandi a desra).
   - **1.4. Analisi grafica delle relazioni tra le variabili**
2. **Model selection**
   - **2.1. Modello di regressione lineare (**applico il modello, faccio la predizione e verifico con il calcolo di MSE, e intervalli di confidenza e predizione**)**
     - **2.1.1. Verificare le inter-relazioni tral X (x1*x2)**
     - 2.1.2. **Verificare andamenti non lineari (**la si individua attraverso l'analisi dei residui con il plot del modello e si vede la linea non dritta). Si risolve applciando log(x) o x^2
     - **2.1.3. Eteroschedasticità.** Verifico sulla Y con l'analisi dei residui i punti tendono ad aprirsi ad imbuto. Si risolve con log(y) o sqrt(y).
   - **2.2. Modello di classificazione (**Applico il modello, faccio la predizione e verifico con matrice di confusione**)**
     - **2.2.1. Modello di regressione logistica** (Binomial, solo con 2 categorie)
     - 2.2.2. **Modello LDA** (Nessuna correlazione tra le X **collinearità**. Quando n è piccolo e i predittori si distribuiscono in una normale è più efficace di una binomiale. Popolare quando k>2). **Condizione.** Ipotizzo una distribuzione normale per le X e **COR(x1 e x2…)=0** fai una corrplot solo per le X
       - 2.2.2.1. **ROC** graficare la tabella di confusione (solo per 2 categorie)
     - **2.2.3. Modello QDA** (L'andamento è quadratico**)**
     - 2.2.4. **Modello KNN** (approccio supervisionato e non parametrico)
       - 2.2.4.1. La KNN lavora con un approccio di verifica delle distanze euclidee in cui si riscontrano problemi se i dati sono a differenti scale. Usare la standardizzazione con scale() solo per le X is.numeric, la Y è una variabile categorica per cui con knn è tolta.
   - **2.3. Stepwise**
     - **2.3.1. Forward, n>p**
     - **2.3.2. Backward, non utilizzabile se n<p**
     - **2.3.3. mixed**
3. **Metodi di ricampionamento (**calcolo del TEST ERROR**)**
   - **3.1.1. Validation set (training set e test set)**
   - **3.1.2. Cross Validation**
     - **3.1.2.1. LOOCV K=1**
     - **3.1.2.2. K-Fold CV (K=10)**