1. **Data exploration**
   1. Visualization of all TPM values for all genes as **BOXPLOT**
   2. **HEATMAP (not correlation)** all genes & CERES scores with three colours (red <0, orange =0, green >0)
   3. Maybe general distribution of CERES scores of all genes – three bars (one for negative, second for 0 and third for positive scores)- **BARPLOT**
2. **PCA Analysis** of [TPM Matrix]- dataset reduced to all genes important in given cell lines for breast cancer
   1. Visualize impact of viability of genes via Boxplot of CERES important genes
      1. Direct identification of important genes with no impact on cell viability
   2. Heat map of PCA genes (y-axis), cell samples (x-axis) and colours representing CERES scores
      1. Could help identify important genes reducing viability when knocked out
   3. Compare PCA genes to driver mutations found in literature research
3. **K-means clustering** of genes into k=2/k=3 groups based on average TPM values in all cell lines
   1. Identify gene cluster/Top 20 Genes with overexpression (high TPM values)
   2. Compare K-means cluster of overexpressed genes to PCA genes (hopefully genes overlap) – visualize with venndiagram
   3. Compare all genes from PCA/ K-means cluster and literature research
      1. Hopefully the 5 driver mutations we chose (CCND, PIK3CA, PARP, MYC, ERBB2)
4. Visualize CERES and TPM values of 5 driver mutations **BOXPLOT** 
   1. Should have negative CERES values and high TPM values

**Start with 1 driver mutation (Gene A):**

1. Generate heatmap with all genes (y-axis) & all selected cell samples (x-axis) with colours representing CERES values
   1. Identification of genes with negative values (in general without differentiating between mutated and not mutated genes)
2. Generate filter with information on the presence of gene mutation in cell line or not (for all genes) – matrix/filter with 1 for mutation and N/A for not mutated
3. Apply filter to [CERES Matrix] with all genes
4. **Pairwise correlation? Shown in Heatmap**
5. **Paired Wilcoxon signed rank Test** (non-parametric test) calculating t-values for Gene A with all other genes in
   1. Perhaps apply correction to p-value (Bonferroni correction etc.)
   2. Identify X number of second site targets (genes) based on p-value (range needs to be specified)
   3. Should correlation value coincide with t-value?
6. Check TPM values for X second site targets
   1. Overexpressed, underexpressed?
7. Could analyze type mutation in [Mutation matrix] or [Copy number matrix] if gene amplified or deleted
8. **Regression analysis**- check if CERES scores of second site targets could be used to predict CERES scores of driver mutations (Lisa Idea- but what does that show us?)
   1. Regression model
   2. F-test to verify regression model in comparison to null model of no correlation

**Repeat for other driver mutations**

**Obtain 5 sets of driver mutations and interactions with second site targets (maybe second site targets overlap?)**