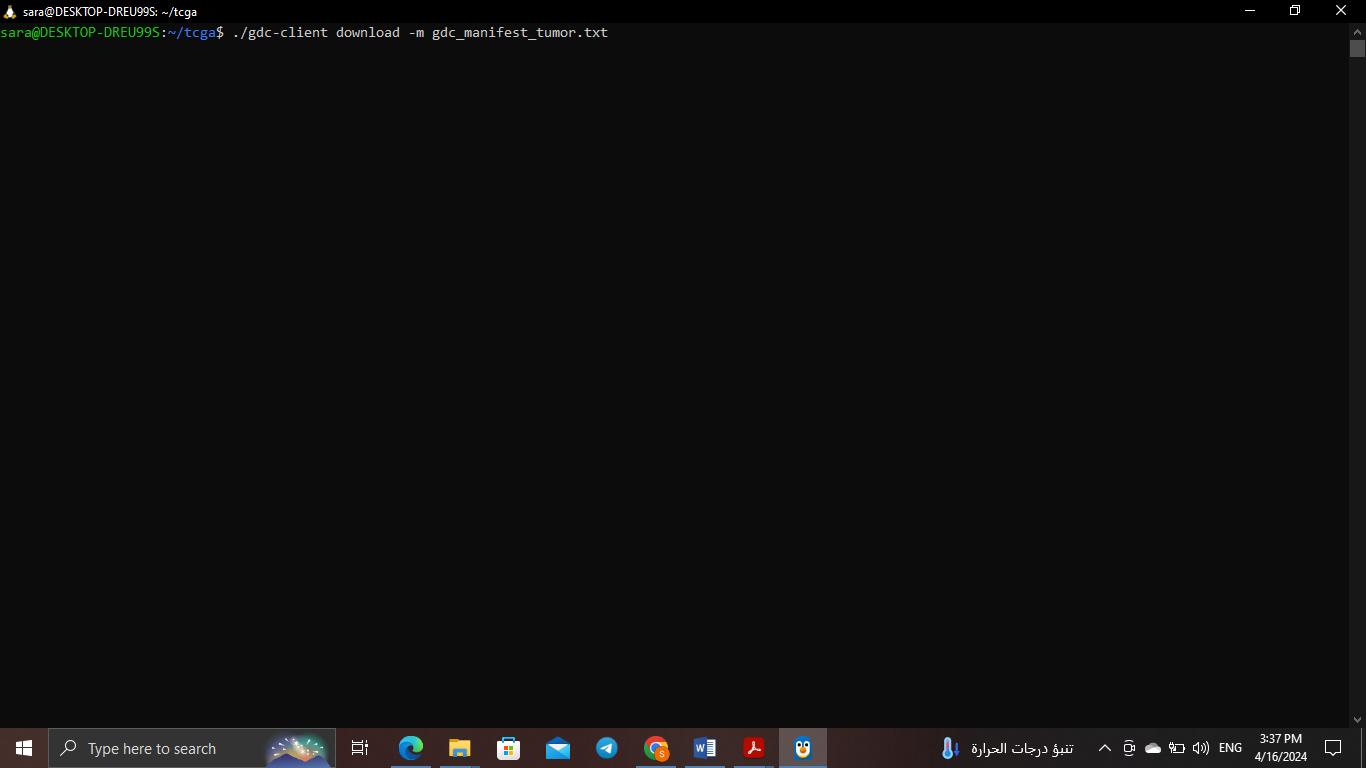
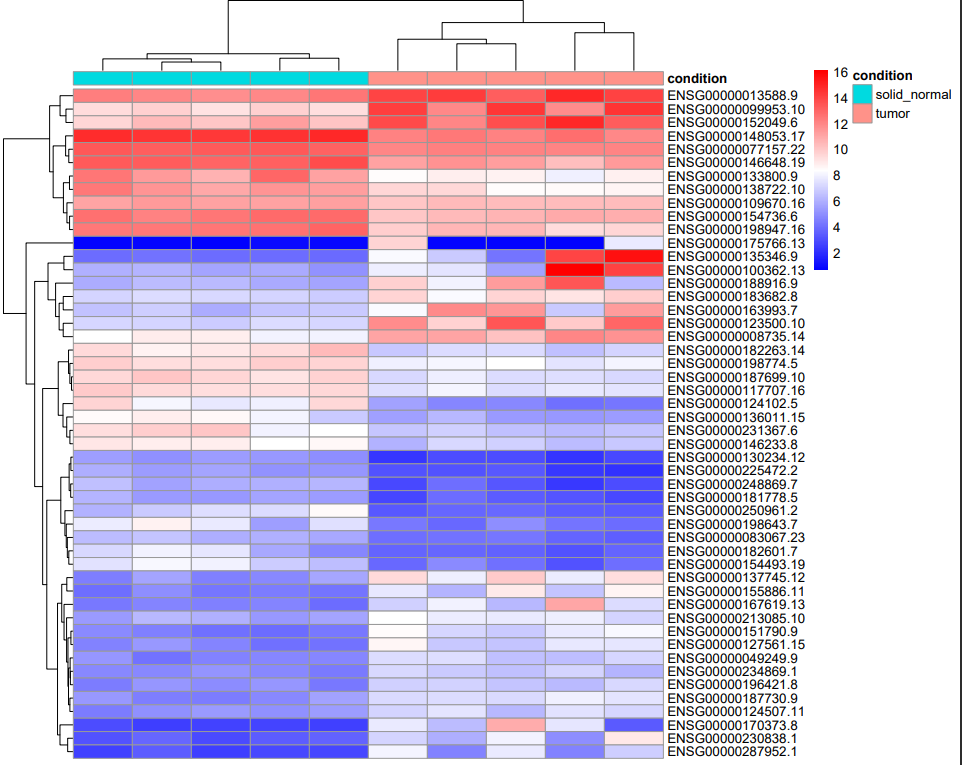
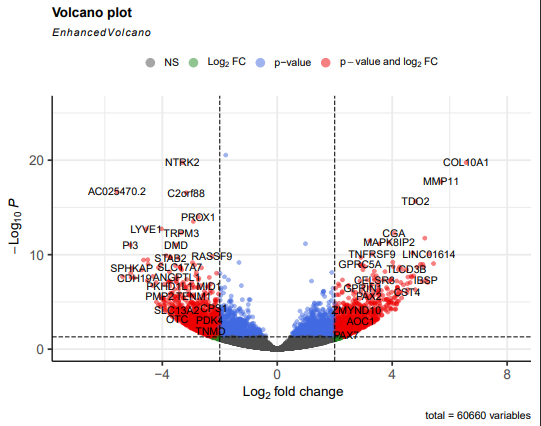
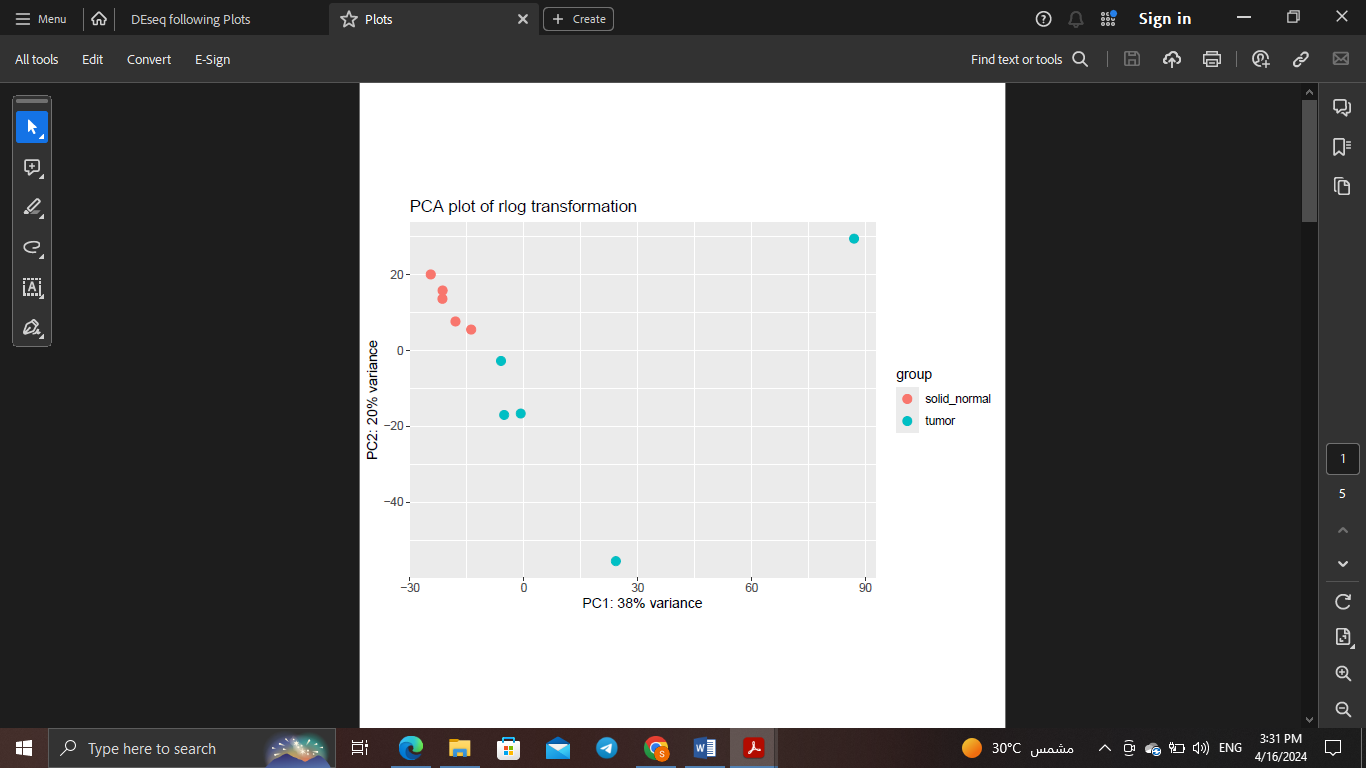
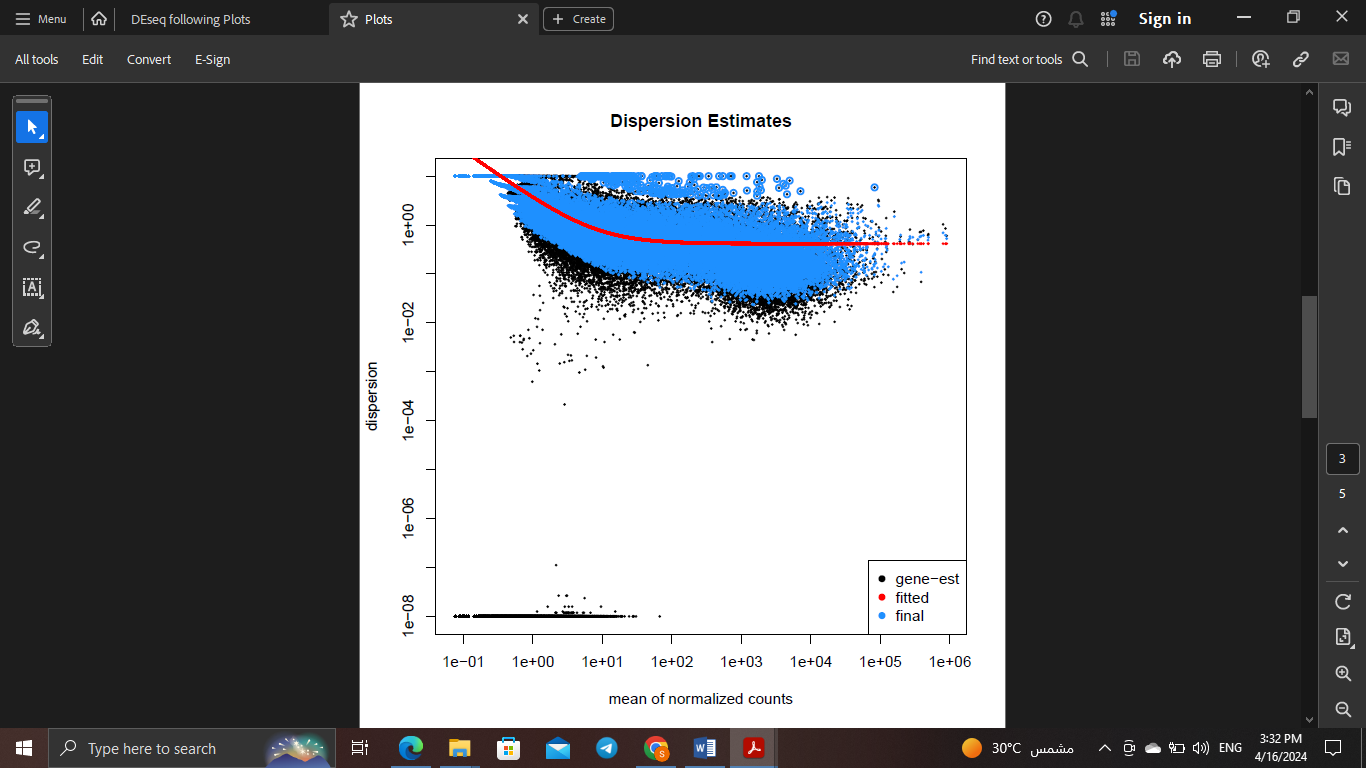
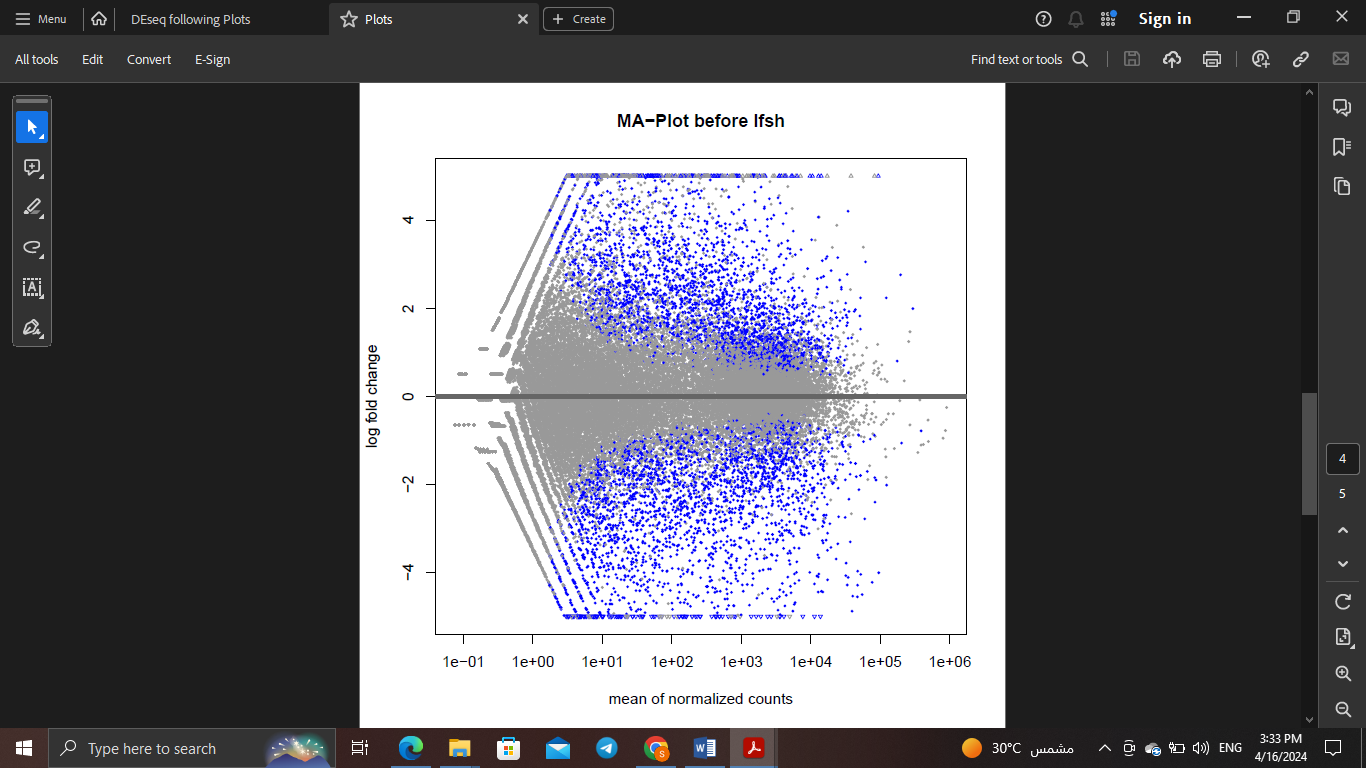
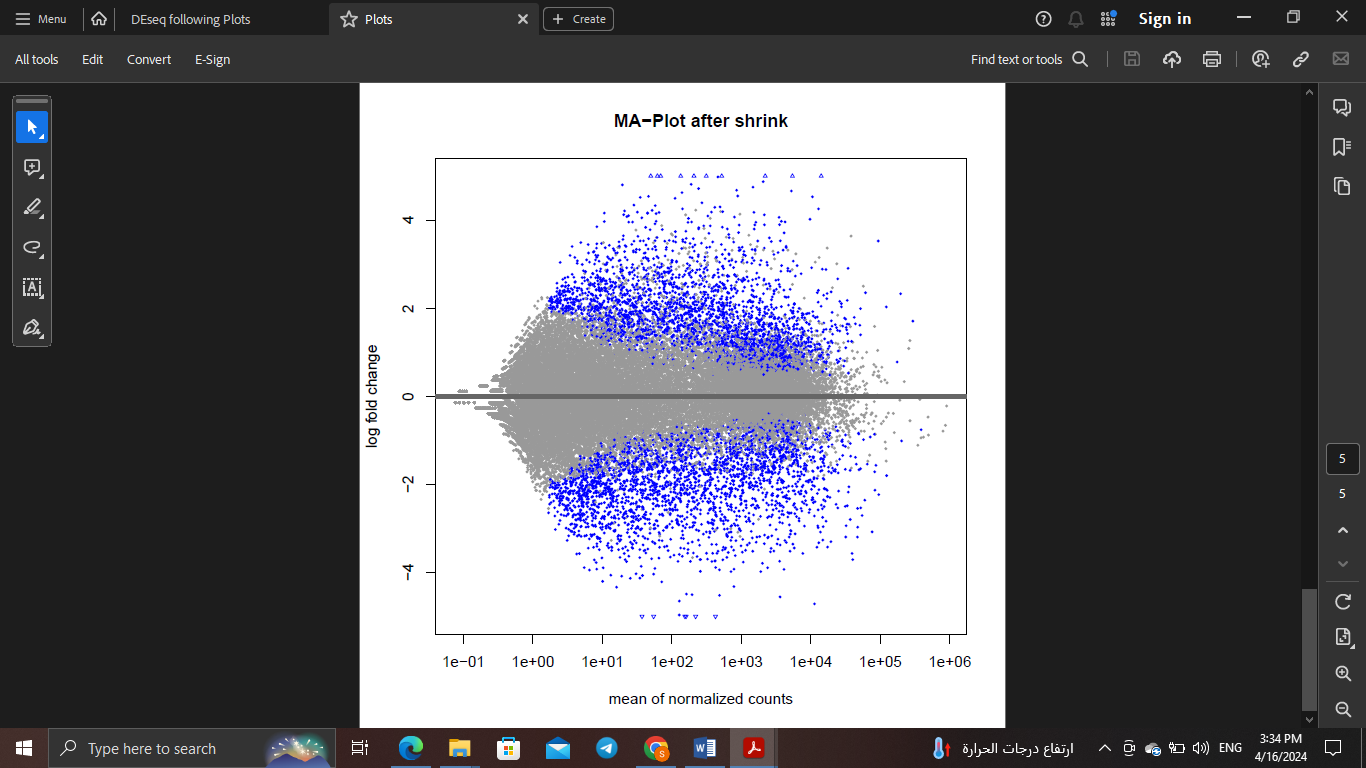
Differential Expression and Enrichment Analysis of Gene Expression

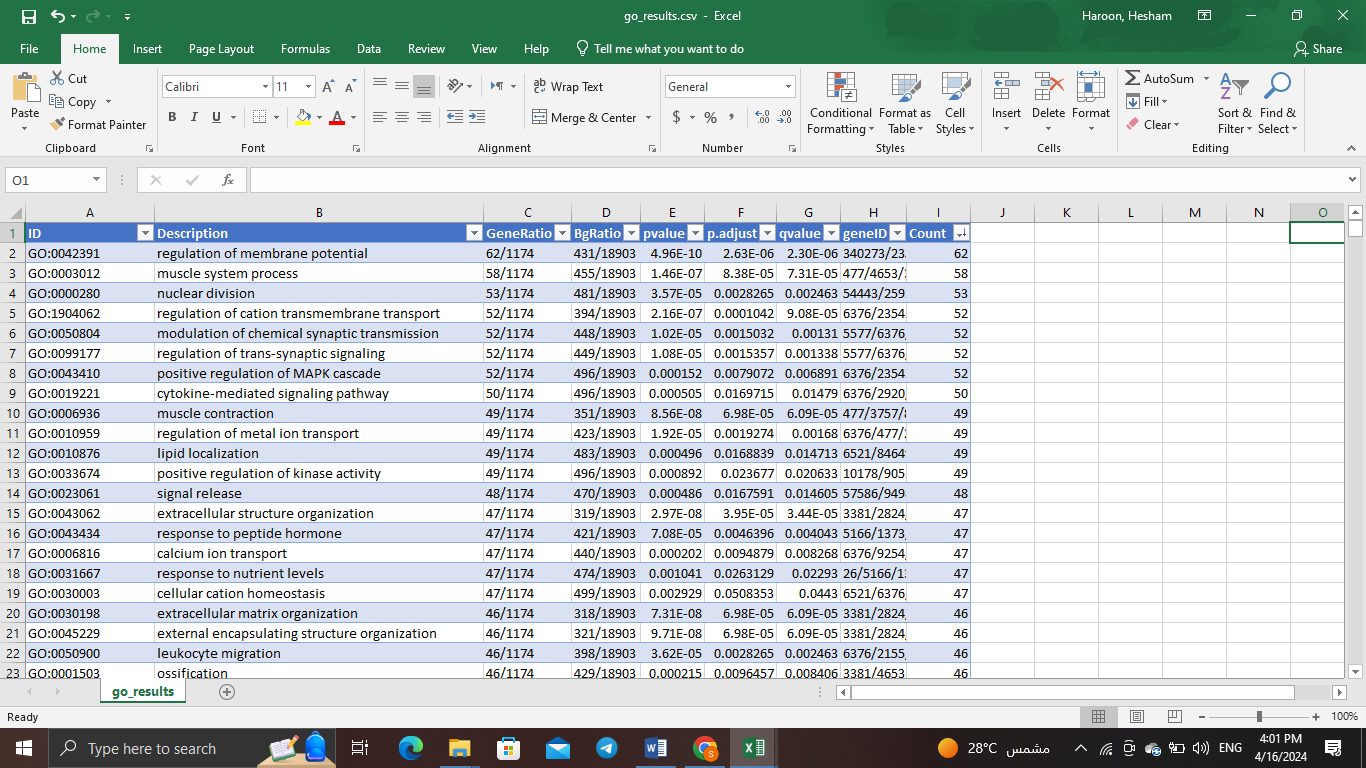
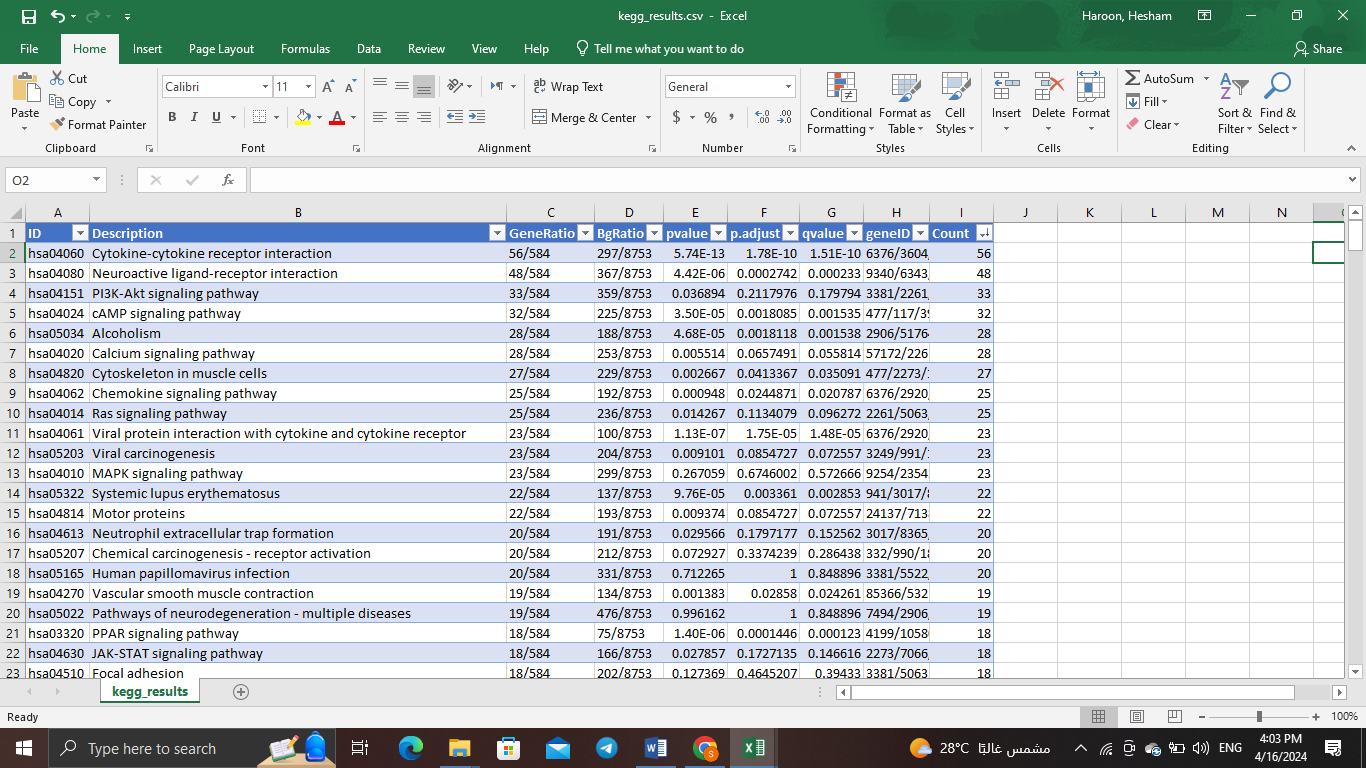
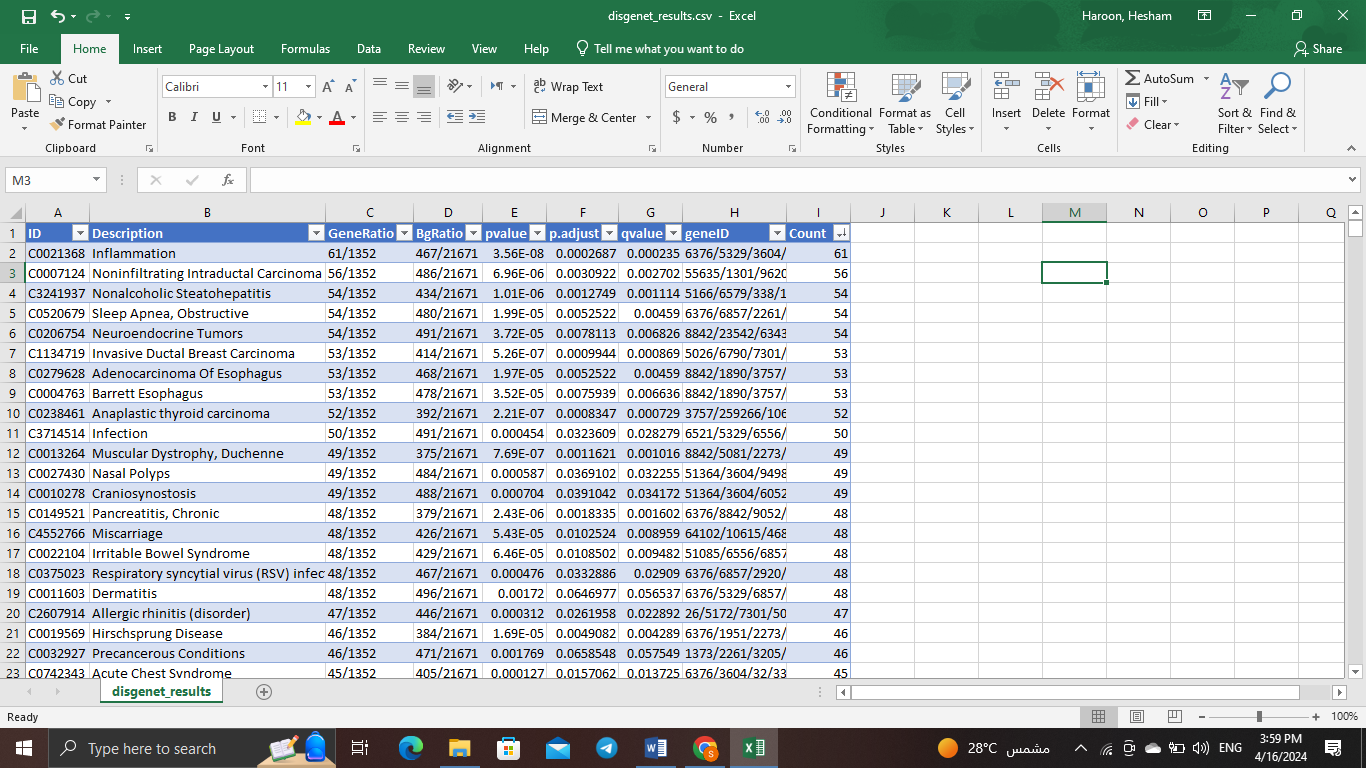
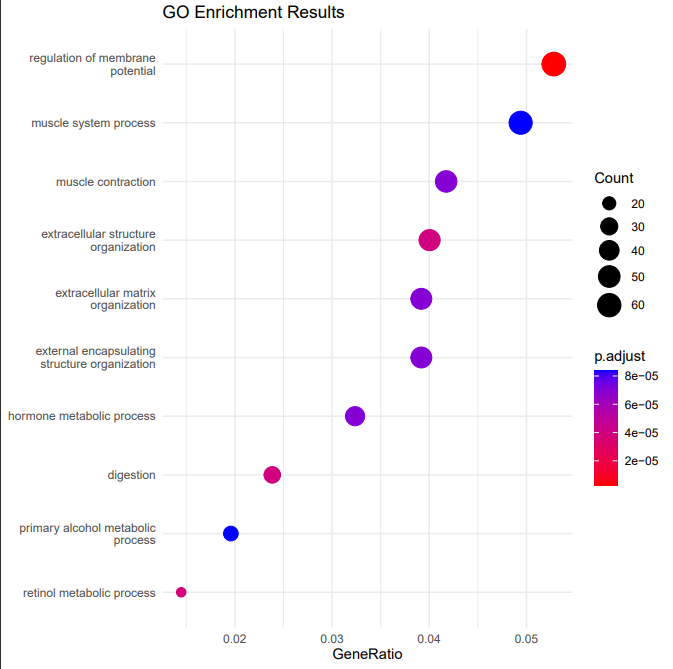
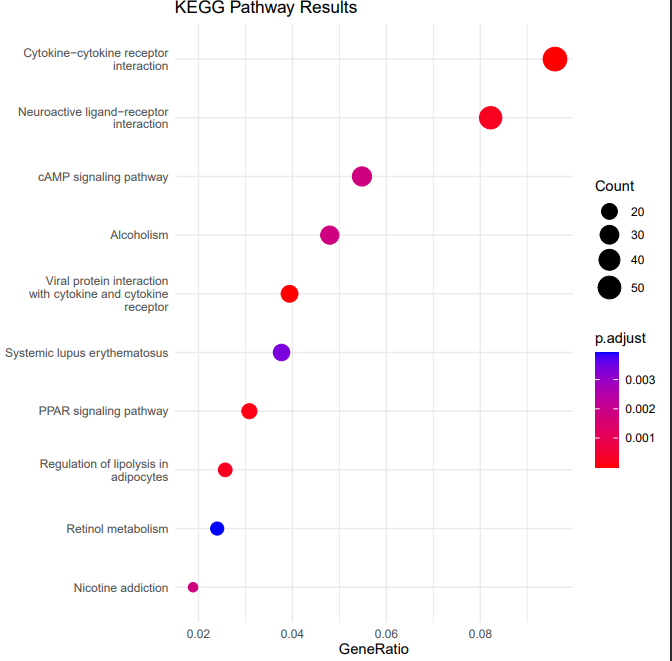
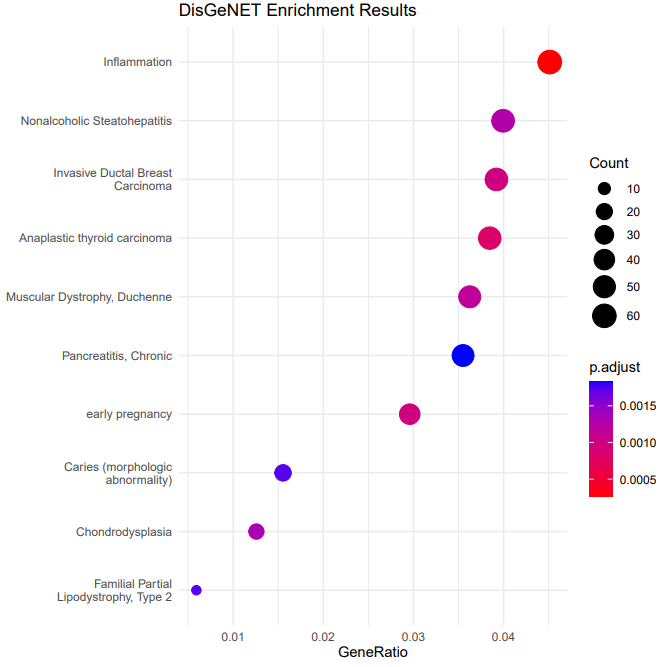
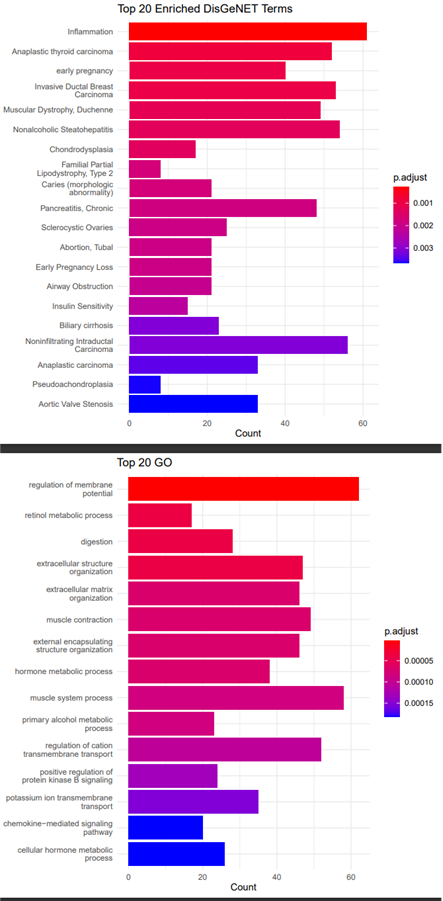
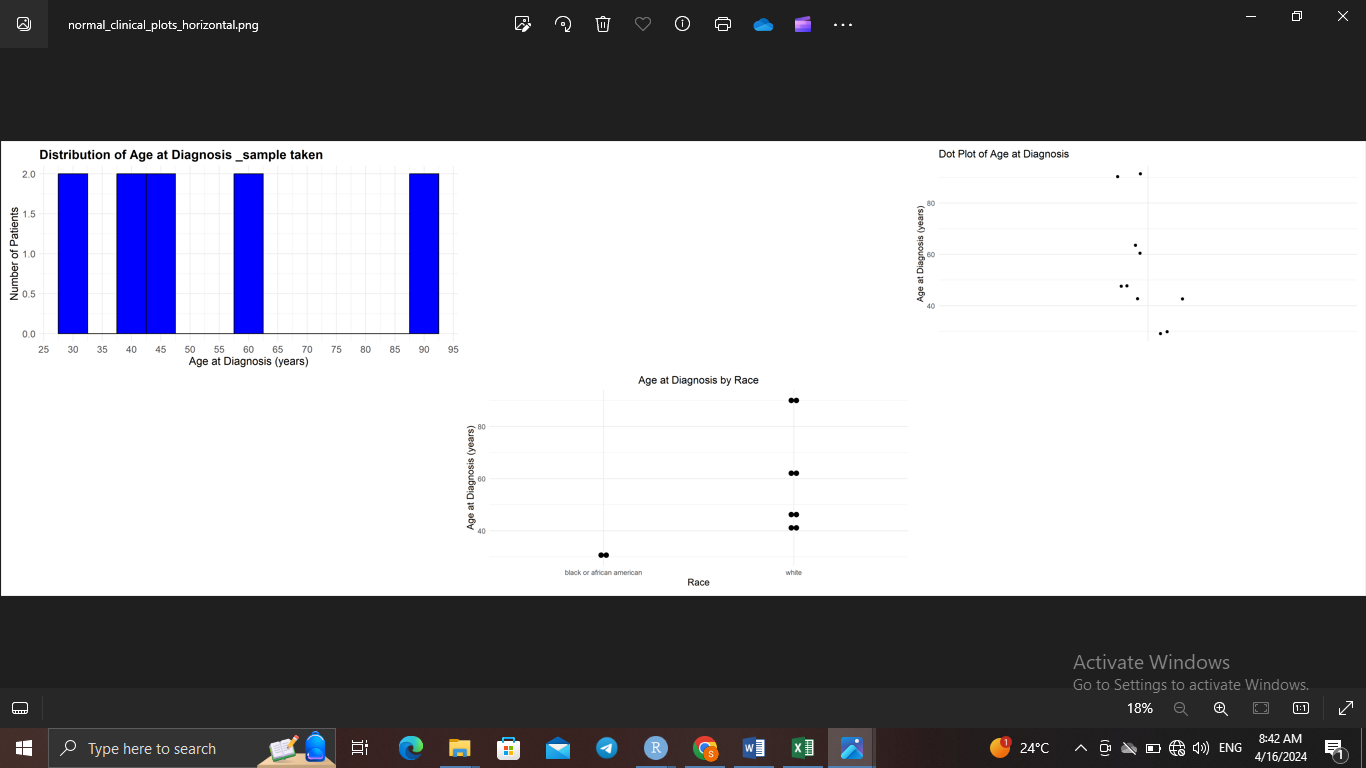
Quantification Data from TCGA

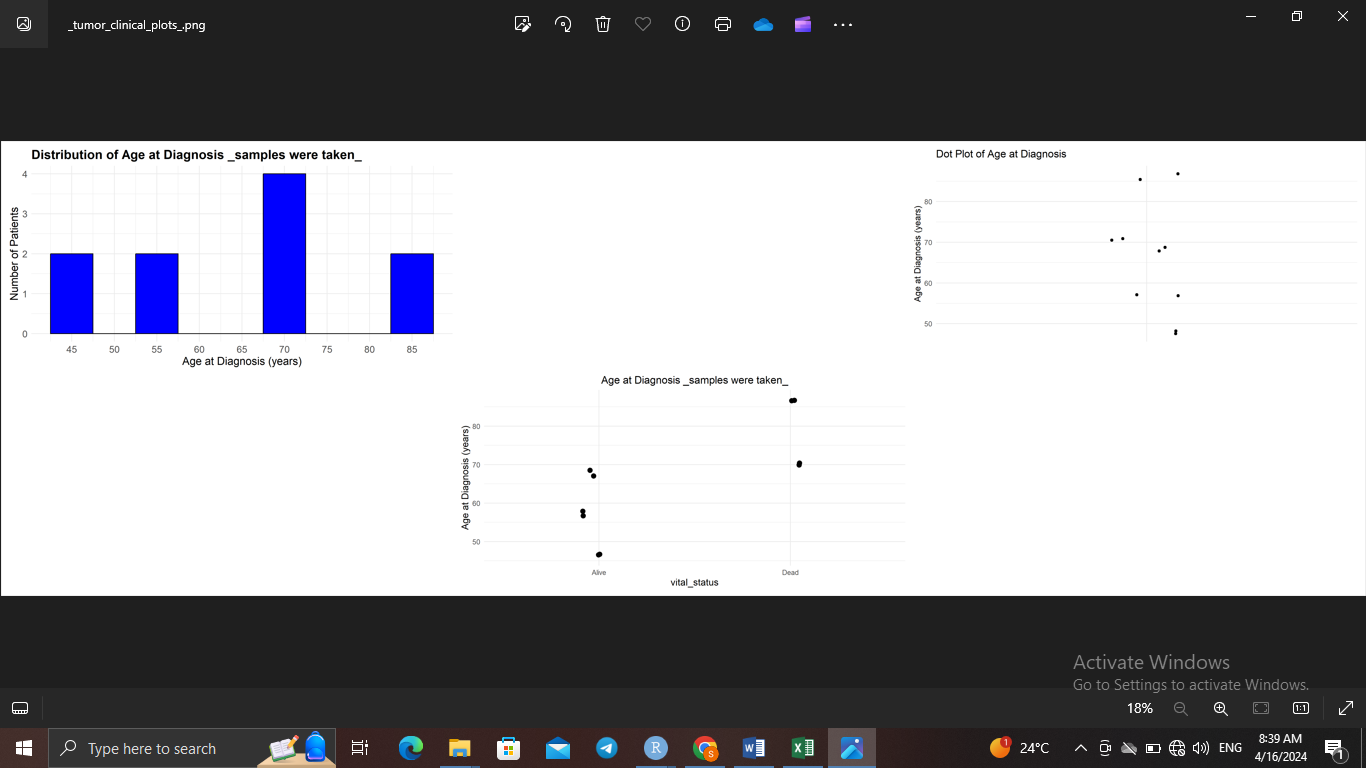
Objectives:

The primary objective of this project is to analyze RNA-Seq data from The Cancer Genome Atlas (TCGA) to identify genes differentially expressed between normal breast tissue and breast cancer samples

1. **Data Acquisition and Preprocessing:**
   * The RNA-Seq data for normal and tumor samples were obtained from the TCGA-BRCA project using GDC client tool.
   * 
   * Preprocessing involved cleaning data by preparing data’s structure by removing unnecessary rows and columns for further analysis.
   * Each count matrix for each sample came with 5 columns (star counts), only one to be used, I used unstranded column as the platform that was used is illumina.
2. **Differential Expression Analysis:**
   * Created metadata according to samples’ condition
   * Prepared for deseq step by using DESeqDataSetFromMatrix
   * Normalized and ran DEseq2
   * Run rlog and the fold shrink to prepare for the investigation graphs.
   * Other graphs were used as enhancedvolcean with p=.05 and fc=2, pheatmap for the first 50 DE genes
   * Saved csv file with DE genes, threshold padj<.05 & |log2FoldChange|>2

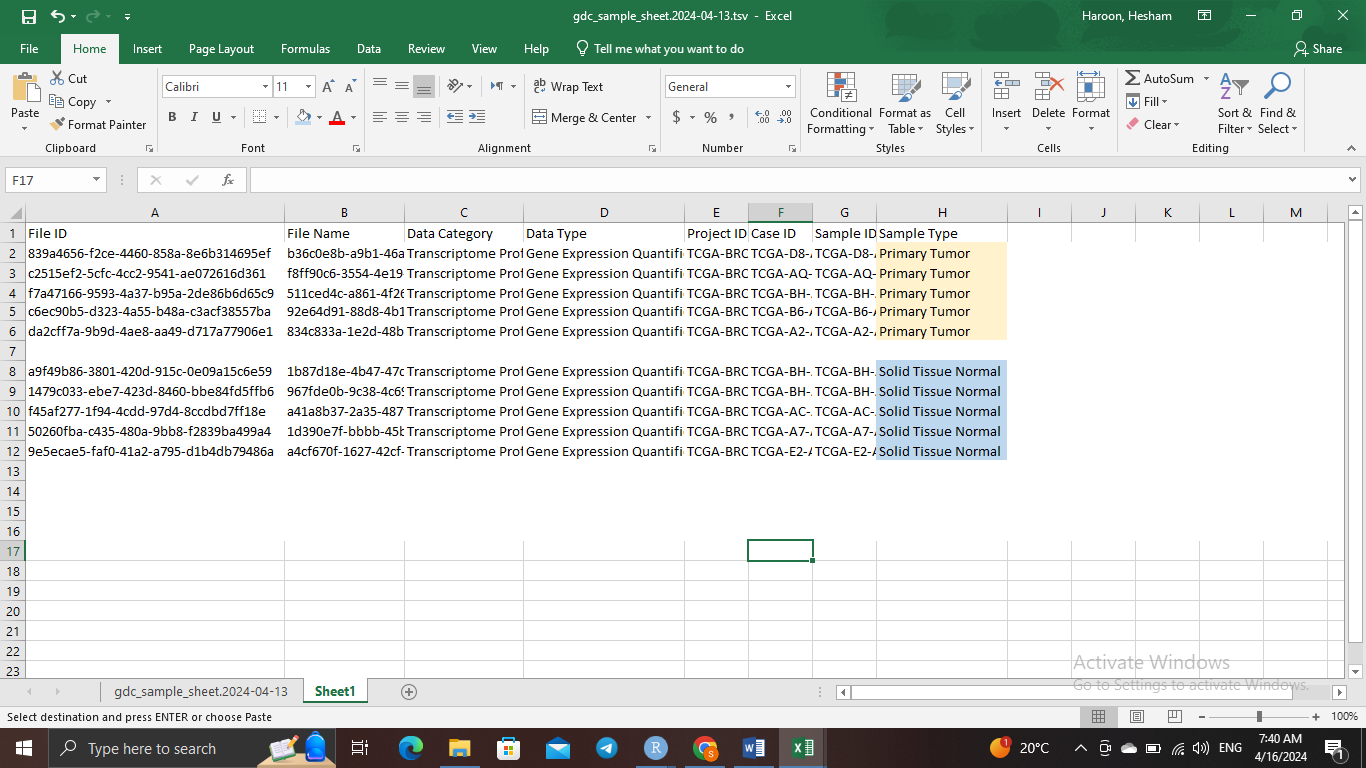
* + 
    - The heatmap visualizes gene expression levels across two conditions: solid normal tissues and tumor tissues. Each row represents a different gene, and each column corresponds to a sample. The color intensity reflects the expression level of a gene in a particular sample, with red indicating high expression and blue indicating low expression
  + 
    - Genes like COL10A1, MMP11, and TDO2 are significantly upregulated in the primary tumor samples (breast cancer)
    - Genes like PROX1, DMD, and TRPM3 are significantly downregulated in the primary tumor samples
    - Genes that are well above the threshold lines might be candidates for further investigation as potential biomarkers for cancer
  + 
    - PC1 explains 38% of the variance and effectively distinguishes between solid normal and tumor samples, indicating that the condition is the main source of variance
    - There appears to be one tumor sample that is far from the other samples along PC1. This it might represent a subtype of the tumor with a distinct gene expression profile
  + 
    - black dots represent the gene estimates of dispersion before any statistical modeling.
    - blue dots are the final estimates of dispersion for each gene after model fitting and adjustments. These should ideally be closer to the red line compared to the black dots
    - Genes with high dispersion relative to their mean expression level might indicate differential expression that is not merely a function of biological variation
  + 
  + 
    - After the log fold change (LFC) shrinkage, the points are more centered around the zero line, indicating a reduction in noise and a more conservative estimate of fold changes

1. **Functional Enrichment Analysis:**
   * Applied **clusterProfiler** for Gene Ontology (GO) enrichment analysis to explore the biological functions of differentially expressed genes.
   * 
     + Processes such as "nuclear division" align with known characteristics of cancer, such as unchecked cell division
     + terms like "regulation of membrane potential" and "muscle system process" could indicate altered cellular behaviors in tumor samples. In breast cancer, these findings could relate to how cancer cells interact with their environment
     + Enrichment in "calcium ion transport" might be associated with signaling pathways that contribute to the invasive properties of breast cancer cells (Breast cancer utilizes calcium signaling **as an advantage for survival and progression** doi: [10.1007/s12551-020-00771-9](https://doi.org/10.1007%2Fs12551-020-00771-9)**)**
     + These GO can guide future functional studies to validate the roles of these biological processes in breast cancer
   * Conducted KEGG pathway analysis to map genes to known genetic pathways potentially affected in breast cancer.
   * 
     + Pathways such as "Cytokine-cytokine receptor interaction" and "PI3K-Akt signaling pathway" are critical for cell communication and survival, and their alteration is often implicated in cancer progression
     + The enrichment of "Neuroactive ligand-receptor interaction" suggests that breast cancer cells might manipulate neurotransmitter and receptor interactions, which could affect tumor growth
     + “Cytoskeleton in muscle cells" and "Focal adhesion" pathways are linked with cell structure and migration, potentially indicating the invasive nature of tumor cells and their ability to spread to other tissues
     + Pathways associated with "Viral carcinogenesis" imply potential links between viral infections and cancer development, which could influence strategies for prevention, such as vaccination ( It was recently estimated that **a virus infection is the central cause of more than 1,400,000 cancer cases annually** doi: [10.1007/978-3-030-57362-1\_1](https://doi.org/10.1007%2F978-3-030-57362-1_1))
   * Used DisGeNET to investigate gene-disease associations and to contextualize gene functions in disease biology.
   * 
     + The top result is "Inflammation", which aligns with research suggesting that chronic inflammation may contribute to cancer development, including breast cancer.
     + There's a presence of various types of cancer beyond breast cancer, such as "Adenocarcinoma Of Esophagus" and "Invasive Ductal Breast Carcinoma", indicating common genetic underpinnings across different cancers
     + Chronic diseases like "Irritable Bowel Syndrome" and "Dermatitis" may share common genetic expressions with cancer, which can be significant for understanding long-term health implications in cancer survivors
     + Understanding diseases that share genetic backgrounds with breast cancer could open up possibilities for drug repurposing
2. **Visualization:**
   * Created visual representations such as PCA plots, volcano plots, heatmaps, histogram and dotplot using **ggplot2** to demonstrate the variance among samples and to visualize clinical and DE data.
   * 
   * 
   * 
     + The size of the dots likely represents the count of genes associated with each term, with larger dots indicating a higher count of genes.
     + he color of the dots reflects the adjusted p-value (p.adjust), with redder colors suggesting more significance
   * 
   * For clinical data after preprocessing the files
   * 
     + The age distribution suggests that the dataset includes a range of adult ages, indicating that the condition studied is not limited to a specific age group
     + The racial categorization indicates the data's demographic diversity, which is important for understanding disease prevalence and expression profiles across different populations



* + - The age distribution in tumor samples highlights a concern that middle-aged to elderly individuals may be at a higher risk or have a higher prevalence of the tumor being studied
    - The dot plot correlates the age at diagnosis with the vital status (Alive or Dead) of the individuals. This plot can reveal potential patterns between age at diagnosis and survival outcomes.

Datasets Used:

* **RNA-Seq Data:**
  + Acquired from TCGA-BRCA project using GDC Portal Client, which contains large-scale genome sequencing data.
  + Included RNA-Seq data for both normal and tumor breast tissue samples.
  + 
* **Clinical Data:**
  + Used alongside the expression data, patient-specific information into the analysis

Illustrating graphs for (Kegg, GO and DisGeNET)

