09/08/22
Sehyun Cho
CAP4612

Homework 1

Tasks:

1. PCA
2. Two violin plots
3. PCA using t-SNE library
4. two violin plots of one with t-SNE value 1, the other for t-SNE value 2

- **Describe the algorithms/approaches/tools used:**
  - **a. <u>What it is or What it does</u>**
    Algorithms used: PCA algorithm and t-SNE algorithm. Both of algorithms are designed to reduce dimensionality.
    Tools used: Jupyter notebook, sklearn library, seaborn library, panda library, pyplot from matplotlib library

  - **b. <u>How it does & Application</u>**
    1. Read csv file by using pandas library, and filtered to have only important dataset, excluding first and last column which contains 'Ensembl_ID' and 'Class' data.

Figure 1 – Reading csv

	Ensembl_ID	ENSG00000005206.15	ENSG00000083622.8	ENSG00000088970.14	ENSG00000099869.7	ENSG00000100181.20	ENSG00000104691.13	ENSG00000115934.11
0	TCGA-3Z-A93Z-01A	3.390813	0.0	2.918265	0.014832	0.341984	2.194036	0.000000
1	TCGA-6D-AA2E-01A	3.144547	0.0	1.961410	0.047186	1.677598	2.605298	0.000000
2	TCGA-A3-3306-01A	2.484817	0.0	2.896470	0.000000	0.087972	3.176764	0.000000
3	TCGA-A3-3307-01A	2.789058	0.0	2.439171	0.022316	0.502293	2.679842	0.000000
4	TCGA-A3-3308-01A	3.258763	0.0	1.941660	0.050283	0.098625	2.841588	0.000000

Figure 2 – Filter the data, remove first and last column

	ENSG00000005206.15	ENSG00000083622.8	ENSG00000088970.14	ENSG00000099869.7	ENSG00000100181.20	ENSG00000104691.13	ENSG00000115934.11
0	3.390813	0.0	2.918265	0.014832	0.341984	2.194036	0.000000
1	3.144547	0.0	1.961410	0.047186	1.677598	2.605298	0.000000
2	2.484817	0.0	2.896470	0.000000	0.087972	3.176764	0.000000
3	2.789058	0.0	2.439171	0.022316	0.502293	2.679842	0.000000
4	3.258763	0.0	1.941660	0.050283	0.098625	2.841588	0.000000
...	...	...	...	...	...	...	...
2524	1.996951	0.0	1.451191	0.000000	2.138038	2.059462	0.000000
2525	2.570807	0.0	2.205505	0.000000	2.323751	2.717458	0.000000
2526	3.022679	0.0	2.595927	0.000000	1.972459	2.871708	0.020087
2527	3.139110	0.0	2.005856	0.000000	1.802198	2.570089	0.000000
2528	3.014305	0.0	2.382532	0.000000	3.221511	2.775717	0.000000

2. Call the PCA algorithm to reduce large amounts of data from csv file, reduce to two dimensional.

Figure 3 – Acquire PC1 and PC2 table by using PCA algorithm

	PC1	PC2
0	-17.864791	15.904503
1	-8.682791	9.634450
2	-20.860309	16.121578
3	-25.767978	19.649743
4	-21.025951	13.430290
...	...	...
2524	-2.416017	-3.437234
2525	3.469989	4.820638
2526	5.365098	9.652960
2527	6.445645	10.172874
2528	6.489701	9.307550

2529 rows × 2 columns

3. Acquire PC1 and PC2 by using PCA algorithm, concatenating PC1,PC2 table with ‘Class’ table.

Figure 4 – Concatenating ‘Class’ column into PC1 and PC2 table

	PC1	PC2	Class
0	-17.864791	15.904503	KIRC
1	-8.682791	9.634450	KIRC
2	-20.860309	16.121578	KIRC
3	-25.767978	19.649743	KIRC
4	-21.025951	13.430290	KIRC
...	...	...	...
2524	-2.416017	-3.437234	THCA
2525	3.469989	4.820638	THCA
2526	5.365098	9.652960	THCA
2527	6.445645	10.172874	THCA
2528	6.489701	9.307550	THCA

2529 rows × 3 columns

4. By using pyplot library, create a scatter plot graph to visualize by classes.
5. By using seaborn library, create Violin plots for ‘PC1’ and ‘PC2’.
6. Call the t-SNE algorithm for the dataset, and fit the data.
7. Collect data into the data frames each of ‘tsne_1’ and ‘tsne_2’.
8. By using seaborn library, create scatter plot graph. ‘tsne_1’ for x-axis, ‘tsne_2’ for y-axis.
9. By using seaborn library, create Violin plots for ‘tsne_1’ and ‘tsne_2’.

- **Describe results:**

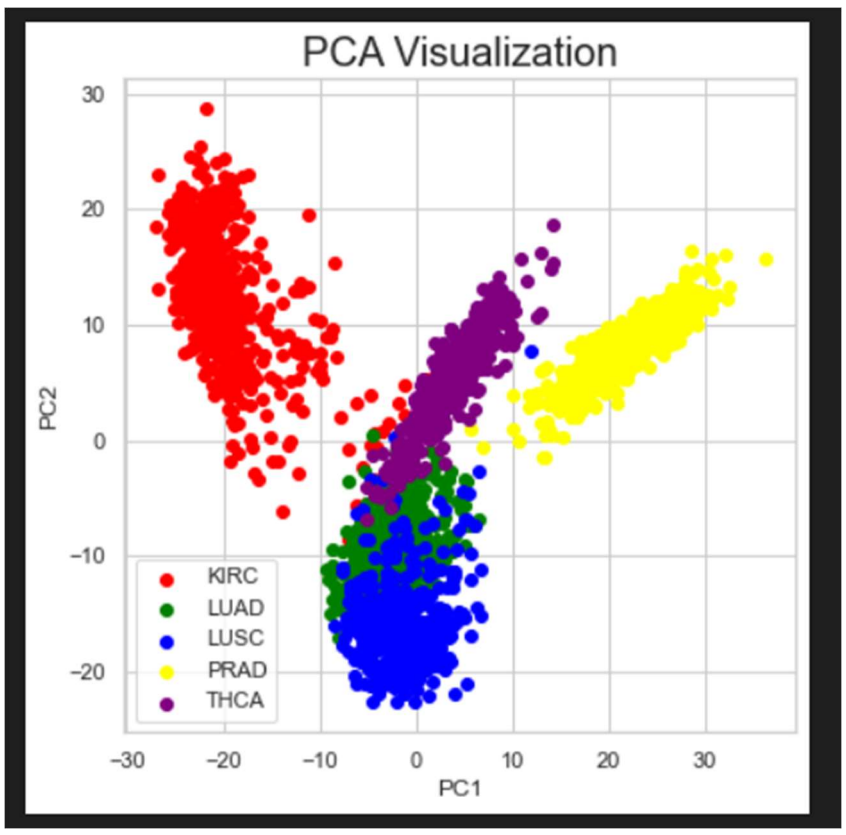

Figure 5 – PCA scatter plot graph

1. **Describe the figure and table.**
   Scatter plot for the PC1 and PC2 tables by classes.
2. **Your observation about the figure and table.**
   Could observe each class has concentration of the data points as values according to PC1 and PC2.
3. **Conclusion.**
   In conclusion, the PCA algorithm worked well compared to the expected output.

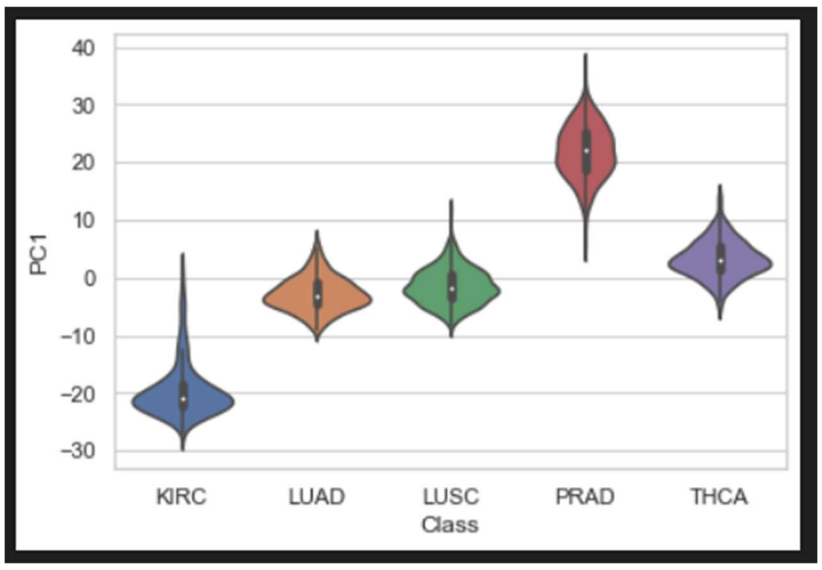

Figure 6 – Violin plot for PC1 with classes

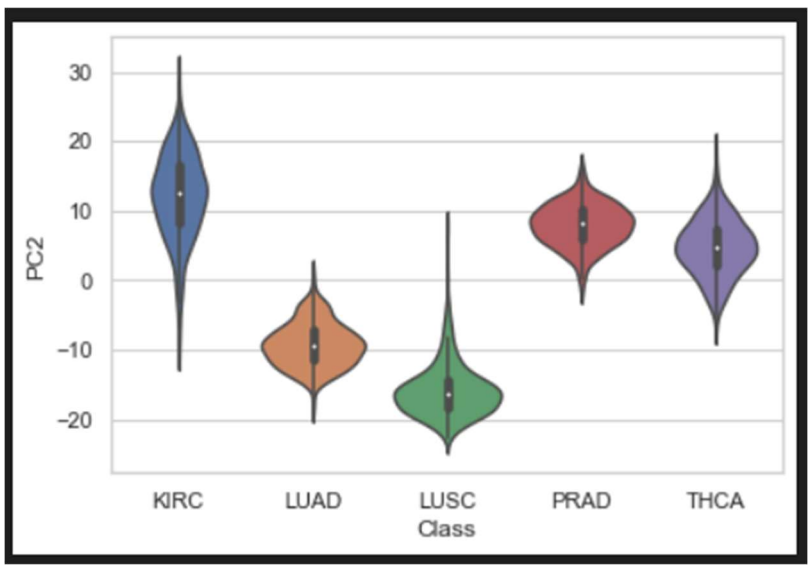

Figure 7 – Violin plot for PC2 with classes

1. **Describe the figure and table.**
   Two Violin plots created, one for PC1 and the other for PC2.
2. **Your observation about the figure and table.**
   Could observe concentration of the data as well by using violin plots and for each PC1 and PC2 filtering by class.
3. **Conclusion.**
   In conclusion, the Violin plots created as the expected output.

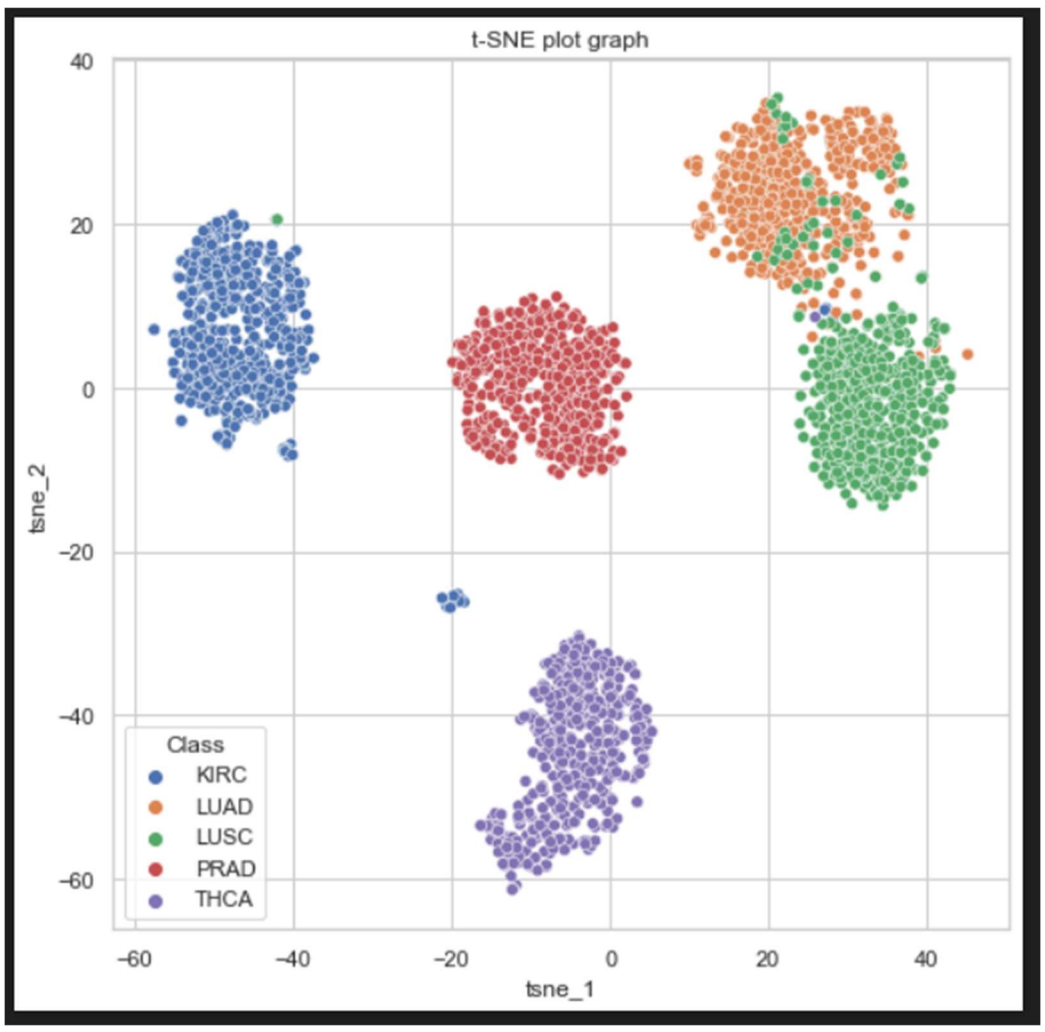

Figure 8 – t-SNE scatter plot graph

1. **Describe the figure and table.**
   t-SNE scatter plot graph for each class.
2. **Your observation about the figure and table.**
   Could observe the concentration of the data points, but some mixed data points for 'LUAD' and 'LUSC' class.
3. **Conclusion.**
   Since the t-SNE is heuristic, the result that I got is not exactly same as the expected output. Therefore, the data points seem correct how they departed.

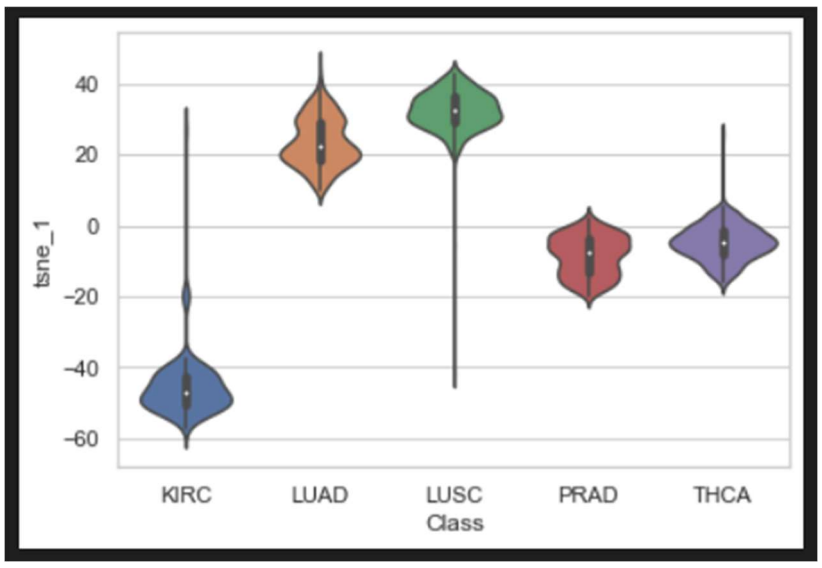

Figure 9 – Violin plot for t-SNE_1

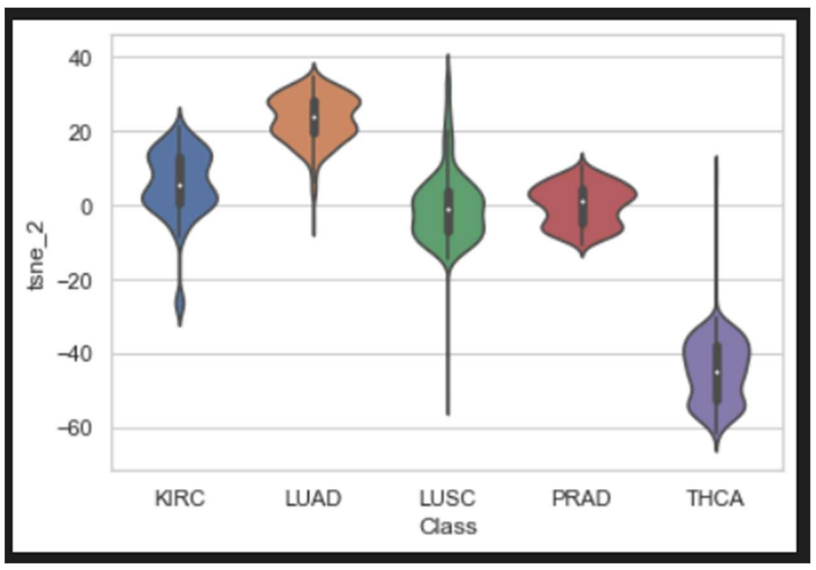

Figure 10 - Violin plot for t-SNE_2

1. **Describe the figure and table.**
   Two Violin plots created, one for t-SNE_1 and the other for t-SNE_2.
2. **Your observation about the figure and table.**
   Could observe the violin plots are correctly created compared to the scatter plot graph.
3. **Conclusion.**
   In conclusion, I could observe the difference between the PCA and t-SNE algorithms by creating scatter graph and violin graph. Especially heuristic features of the t-SNE make difference.